Zweidimensionale klassische und diskrete orthogonale Polynome und ihre Anwendung auf spektrale Methoden zur Lösung von hyperbolischen Erhaltungsgleichungen

Von der
Carl-Friedrich-Gauß-Fakultät
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat)

genehmigte Dissertation

von Philipp Rudolf Öffner
geboren am 27.02.1984
in Würzburg

Eingereicht am: 22.1.2015

Disputation am: 13.3.2015

1. Referentin/Referent: Prof. Dr. Thomas Sonar

2. Referentin/Referent: Prof. Dr. Klaus -J. Förster

3. Referentin/Referent: Prof. Dr. Andreas Meister

(2015)

Danksagung:

Diese Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Partielle Differentialgleichungen am Institut *Computational Mathematics* der Technischen Universität Braunschweig. Viele Menschen haben zum Gelingen dieser Arbeit beigetragen und mich unterstützt. Hierfür möchte ich mich herzlich bedanken.

Ein ganz besonderer Dank gilt meinem Doktorvater, Professor Dr. Thomas Sonar, für seine sowohl fachliche als auch menschliche Unterstützung. Er konnte mich schon zu Anfangs für das Thema begeistern und auch im Laufe des Entstehungsprozesses dieser Arbeit hatte er enormen Einfluss auf mein Wirken. Seine fachliche Beratung gab mir immer wieder neue Impulse ohne dabei meine Freiräume in der Forschung einzugrenzen. Die warmherzige Betreuung rundete das stets harmonische und freundschaftliche Arbeitsklima ab. Für all das möchte ich mich von Herzen nochmals Bedanken.

Prof. Dr. Förster von der Universität Hildesheim danke ich für die Übernahme eines Korreferats.

Prof. Dr. Meister möchte ich zum einen für die Übernahme des Korreferats danken sowie für die anregenden Diskussionen, welche wir während mehrerer Konferenzen hatten. Seine Fragen zum Thema gaben mir einen anderen Blickwinkel auf das Thema und haben mir so einen größeren Überblick ermöglicht.

Ein weiterer Punkt für das Gelingen der Arbeit war zudem das hervorragende Arbeitsklima in der Arbeitsgruppe. Sie ermöglichte mir mich schnell in Braunschweig heimisch zu fühlen und mich besser auf meine Arbeit zu konzentrieren. Ich danke hier insbesondere meinen Kollegen Marko Stautz (M.Sc., M. Ed.), Dr. Antje Vollrath und Dr. Martina Wirz. Martina Wirz möchte ich außerdem danken für das Bereitstellen des Spektrale-Differenzen-Verfahrens und ihre Erklärungen hierzu. Marko Stautz und Antje Vollrath bin ich zum Dank verpflichtet für die anregenden mathematischen Diskussionen und für die hervorragende Atmosphäre im Büro, auch wenn es manchmal mit mir nicht einfach war.

Ein weiterer Dank gilt Ella Folkers, Nina Malitzig, René Goertz, Marko Stautz und Antje Vollrath für eine detaillierte Durchsicht von Teilen einer früheren Version dieser Arbeit.

Zum Schluss möchte ich noch meinen Eltern danken. Sie haben mich mein ganzes Leben lang unterstützt und bestärkt, das zu tun, was mir Spaß bereitet. Sie halfen mir in jeder Lebenslage und waren immer für mich da. Ihnen widme ich diese Arbeit.

Inhaltsverzeichnis

1 Einleitung

Seit Ende des 18. Jahrhunderts beschäftigen sich Mathematiker mit der Theorie orthogonaler Polynome und bis heute stellt dieses Feld ein attraktives Forschungsgebiet innerhalb der Mathematik dar. Orthogonale Polynome finden Anwendung in der Zahlentheorie, der Kombinatorik, der Approximationstheorie und der mathematischen Physik. Gleichzeitig verbinden sie die verschiedenen Bereiche innerhalb der Mathematik miteinander und vergrößern dadurch den Blickwinkel auf das jeweilige Teilgebiet. Aus diesem Grund gehören klassische Resultate über orthogonale Polynome sowie die Familien von Chebyshev-, Legendre-, Laguerre- und Hermite-Polynomen zum Standardrepertoire eines jeden reinen oder angewandten Mathematikers.
Die Entwicklung des Computers gab den Impuls, die bereits existierenden Approximationsergebnisse als Grundlage für numerische Verfahren zu nutzen und nach weiteren Resultaten zu forschen. So finden orthogonale Polynome Anwendung in Verfahren zur Berechnung von Lösungen partieller Differentialgleichungen mit Hilfe spektraler Methoden. Bei spektralen Methoden wird die gesuchte Lösung u mittels einer endlichen Linearkombination von Basiselementen eines vorher gewählten Funktionenraums approximiert. Als Basiselemente sind beliebige C^∞-Funktionen zugelassen, wobei man aus Konvergenz- und Effizienzgründen Fourier-Summen von orthogonalen Polynomen bei der Approximation verwendet.
Zahlreiche Phänomene in der Natur werden durch Systeme von partiellen Differentialgleichungen beschrieben. Eine besondere Klasse hierbei sind die hyperbolischen Erhaltungsgleichungen. Mit ihnen werden beispielsweise Strömungen modelliert. Im Allgemeinen sind jedoch keine exakten Lösungen für hyperbolische Erhaltungsgleichungen bekannt, so dass man an numerischen Lösungen interessiert ist. Als zusätzliche Anforderungen an das numerische Verfahren lässt die Theorie hyperbolischer Erhaltungsgleichungen auch unstetige Lösungen zu, die auch keinesfalls eindeutig sein müssen. Die Eindeutigkeit versucht man durch das Aufstellen zusätzlicher Entropie- und Kausalitätsbedingungen zu erreichen. Allerdings hat man weiterhin das Problem der Unstetigkeit der Lösung und die numerischen Methoden müssen dies berücksichtigen. Bei Sprungunstetigkeiten variiert das Verhalten der verschiedenen numerischen Methoden abhängig von ihrer jeweiligen Konvergenzordnung. Verfahren erster Ordnung verschmieren die Unstetigkeit und glätten sie. Bei Methoden höherer Ordnung hingegen werden die Sprungunstetigkeiten scharf lokalisiert, und es bilden sich physikalisch unplausible Oszillationen, die zu Stabilitätsproblemen führen können. Daher versucht man durch Dämpfungstechniken die Oszillationen abzuschwächen und bestenfalls komplett zu entfernen.
Der Vorteil numerischer Verfahren hoher Ordnung liegt in der Möglichkeit genauerer Approximationen der Lösungen auf wesentlich gröberen Gittern als bei Verfahren gerin-

gerer Ordnung. Im Kontext der spektralen Methoden bedeutet dies die Verwendung von hohen Polynomgraden in der Approximation. Die entstehenden Oszillationen versucht man mittels modaler Filter, welche direkt auf die Koeffizienten der approximierenden Summe von Polynomen wirken, abzuschwächen.
Die Theorie spektraler Methoden ist bei Verwendung eindimensionaler Basiselemente wohlverstanden und seit langem Gegenstand der Forschung.
Ziel dieser Arbeit ist es nun, die Theorie spektraler Methoden um klassische orthogonale Polynome auf Dreiecken zu erweitern. Dabei werden wir neue Approximationsresultate beweisen, optimale modale Filter konstruieren und einen ersten Ansatz für diskrete orthogonale Polynome bei spektralen Methoden liefern. Die Ergebnisse ermöglichen uns bereits existierende numerische Lösungsmethoden für hyperbolische Erhaltungsgleichungen durch Verwendung orthogonaler Polynome auf Dreiecken und modale Filter zu erweitern und zu verbessern. Die Approximationsresultate liefern hierbei die theoretische Grundlage. In den abschließenden numerischen Testrechnungen zeigt sich, dass sich die Wahl der Basis und des Filters positiv auf die Stabilität der numerischen Methode auswirken kann. Verwendet man in einem spektralen Verfahren eine klassische orthogonale Basis, so muss man zur Berechnung der Fourier-Koeffizienten entweder ein Integral auswerten oder ein lineares Gleichungssystem lösen. In der Praxis erfolgt die Auswertung des Integrals approximativ mittels eines Quadraturverfahrens und auch das lineare Gleichungssystem wird numerisch gelöst. Nutzt man anstelle klassischer orthogonaler Polynome diskrete orthogonale Polynome, so wird zur Auswertung der Koeffizienten lediglich eine Summe gebildet. Dies vereinfacht die Berechnung wesentlich und ist exakt. Aus diesem Grund ist die Verwendung diskreter orthogonaler Polynome bei spektralen Verfahren ein vielversprechender Ansatz, der auch über diese Arbeit hinaus weiterverfolgt werden sollte. Wir liefern hierfür einen ersten Ausblick. Dabei gliedert sich die Arbeit wie folgt:
Im ersten Kapitel liefern wir die grundlegenden Definitionen und Eigenschaften hyperbolischer Erhaltungsgleichungen, bevor wir das Spektrale-Differenzen-Verfahren erklären, das wir im Zuge dieser Arbeit für allgemeine orthogonale Polynome auf Dreiecken und ihre natürlichen modalen Filter erweitern.
Im anschließenden Abschnitt geben wir einen Überblick über wichtige Eigenschaften der Jacobi-Polynome und untersuchen ihr Approximationsverhalten. Bei den Jacobi-Polynomen handelt es sich um klassische orthogonale Polynome in einer Variable. In diesem Zusammenhang erklären wir auch den Begriff der spektralen Konvergenz. Wir definieren die klassischen orthogonalen Polynome auf Dreiecksgebieten, die APK-Polynome, und referenzieren einige elementare Eigenschaften. Anders als die Jacobi-Polynome sind die APK-Polynome nicht Lösungen eines singulären Sturm-Liouville-Problems. Allerdings wird gerade die Selbstadjungiertheit des Sturm-Liouville-Operators im Beweis der spektralen Konvergenz der Jacobi-Polynome genutzt. Wir zeigen im Folgenden, dass der Differentialoperator der APK-Polynome zwar nicht selbstadjungiert, aber potentiell selbstadjungiert ist, was eine schwächere Eigenschaft darstellt. Im Anschluss nutzen wir diese Tatsache und einige von uns neu gezeigte Abschätzungen für die APK-Polynome, um das Verhalten der APK-Fourier-Koeffizienten, des Abschneidefehlers in einer gewichteten L^2-Norm und im punktweisen Sinn zu untersuchen. Gerade das Verhalten des

punktweisen Fehlers stellt hierbei ein wichtiges und neuartiges Resultat dar. Aus dieser Analyse folgt schließlich für eine C^∞-Funktion spektrale Genauigkeit bei der Approximation. Der Beweis liefert auch die theoretische Grundlage, um bei Verwendung der Polynome in unserem numerischen Verfahren überhaupt von einer spektralen Methode sprechen zu können.
Im nächsten Kapitel beschäftigen wir uns mit der Problematik der Oszillationen. Bei spektralen Methoden ist es ein übliches Vorgehen, die Oszillationen durch Verwendung modaler Filter zu reduzieren. Wir erklären in diesem Zusammenhang das Gibbs'sche Phänomen und die grundlegende Idee, welche hinter der modalen Filterung steht. Dabei rekapitulieren wir die in der Literatur bekannten modalen Filter und einige Approximationsresultate bezüglich gefilterter Reihenentwicklungen, um anschließend die gefilterte APK-Reihe erstmals zu analysieren und ihr Verhalten zu beschreiben. Wir erläutern noch die spektrale Viskositätsmethode und beweisen, dass deren Verwendung als äquivalent zu einer spektralen Methode mit modaler Filterung gesehen werden kann. Aus der Viskositätsformulierung für das spektrale Verfahren der APK-Polynome leiten wir einen modalen Filter ab. Dieser hängt direkt von einem der frei wählbaren Parameter der APK-Polynome ab und wir nutzen ihn später bei der numerischen Untersuchung. Wir diskutieren dann noch die Problematik der adaptiven Filterung und ihre Verwendung im Spektrale-Differenzen-Verfahren an, ehe wir das Resultat aus der Arbeit von Hesthaven und Kirby [36] für die gefilterte Legendre-Reihe nochmals aufgreifen und deren Beweis verbessern. Wir verallgemeinern, soweit möglich, die Teilergebnisse und spezialisieren uns erst am Ende auf die Legendre-Polynome. Eine Diskussion dieses Ergebnisses bildet den Abschluss des Kapitels.
In Kapitel 5 stellen wir schließlich die Hahn-Polynome als Beispiel diskreter orthogonaler Polynome vor, zeigen ihren Zusammenhang mit den Jacobi-Polynomen und beweisen die spektrale Konvergenz in den Koeffizienten für die Reihenentwicklung. Zudem erklären wir die Problematik bei Verwendung dieser Polynome, welche ihren Ursprung in der äquidistanten Punkteverteilung haben. Wir schlagen mögliche Lösungsansätze vor, die alle die Verwendung diskreter orthogonaler Polynome auf nicht-gleichverteilten Punkten beinhalten. Da wir grundsätzlich an Polynomen auf Dreiecksgebieten interessiert sind, erläutern wir die zweidimensionale Erweiterung der Hahn-Polynome auf einem Dreiecksgitter. Im Ganzen stellt dieses Kapitel mögliche Ansätze zur Erweiterung der Theorie spektraler Verfahren vor und dient als Grundlage für weitere Forschungsprojekte.
Im nächsten Abschnitt analysieren wir die APK-Polynome und ihre modalen Filter im Spektrale-Differenzen-Verfahren in zwei nichtlinearen Testfällen. Wir vergleichen dabei die verschiedenen APK-Polynome mit ihren modalen Filtern und diskutieren den Einfluss der Parameterwahl auf die Stabilität des Verfahrens.
Neben einem Fazit schließen wir diese Arbeit mit einem Ausblick für weitere Forschungen ab.
Diese Arbeit ist aus dem DFG-Projekt (Nummer SO 363/11-1 61411202) entstanden, in dem es um die Konstruktion und den Vergleich von dem Spektrale-Differenzen-Verfahren und Diskontinuous-Galerkin-Verfahren ging. Die Ergebnisse findet man in [55], [56], [57], [58] und [87].

2 Mathematische Grundlagen

In diesem einführenden Kapitel geben wir einen Überblick über die Theorie hyperbolischer Erhaltungsgleichungen. Dabei erläutern wir die mathematischen Grundlagen und erklären die Problematik beim Lösen dieser Differentialgleichungen.
Im Anschluss wird die für diese Arbeit relevante numerische Methode, das Spektrale-Differenzen-Verfahren, vorgestellt und erläutert. Dieses Verfahren erweitern wir im späteren Verlauf durch die Verwendung zweidimensionaler orthogonaler Polynome und spezieller Filter, um es abschließend mit numerischen Testfällen zu analysieren.
Dem Leser seien für eine ausführlichere Einführung in die Theorie hyperbolischer Erhaltungsgleichungen die Bücher [28], [48], [84] empfohlen. Bezüglich spektraler Methoden sind sicherlich das Standardwerk [11] und der Artikel [29] zu nennen. Eine Erklärung der numerischen Methode findet man in [50] und ihre genaue Implementierung in der Doktorarbeit [86].

2.1 Hyperbolische Erhaltungsgleichungen

Diese Arbeit beschäftigt sich mit orthogonalen Polynomen und deren Anwendung auf numerische Verfahren von hyperbolischen Erhaltungsgleichungen. Wir beweisen grundlegende Eigenschaften orthogonaler Polynome und verwenden sie schließlich, um eine bereits existierende numerische Lösungsmethode für hyperbolische Erhaltungsgleichungen zu erweitern. Allgemein modelliert man mit diesen Differentialgleichungen unter anderem verkehrsdynamische Prozesse, Transporte und Strömungen. Auf einige dieser Modelle werden wir in Kapitel 6 noch explizit eingehen.
Hier wird zunächst ein Überblick über die Begrifflichkeiten und mathematischen Grundlagen der Theorie hyperbolischer Erhaltungsgleichungen gegeben. Wir folgen dabei dem Buch [28] mit Ergänzungen aus [48].
Betrachtet wird eine Änderung einer Erhaltungsgröße

$$\mathbf{u} : \mathbb{R}^d \times \mathbb{R}_0^+ \longrightarrow \Omega$$

bezüglich des Raums $\mathbb{R}^d$ und Zeit $t \geq 0$. Die Funktion $\mathbf{u}$ bildet in einen **Zustandsraum** $\Omega \subset \mathbb{R}^p$ ab, weil für einige Größen, beispielsweise aus der Physik, nur bestimmte

Wertebereiche zulässig sind. Wir beschäftigen uns ausschließlich mit Systemen von hyperbolischen Erhaltungsgleichungen. Ein solches System ist gegeben durch

$$\frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial t} + \sum_{j=1}^{d} \frac{\partial \mathbf{f}_j(\mathbf{u}(\mathbf{x},t))}{\partial x_j} = \mathbf{0} \qquad \forall \mathbf{x} \in \mathbb{R}^d, t > 0, \tag{2.1}$$

wobei $\mathbf{f}_j \in (C^1(\Omega))^p$ gilt und $C^1(\Omega)$ der Raum der einmal stetig differenzierbaren Funktionen auf Ω ist. Man bezeichnet die $\mathbf{f}_j(\mathbf{u}) = (f_{1,j}, \cdots, f_{p,j})^T$ als **Flussfunktionen**.
Das System (2.1) kann man in quasilineare Form

$$\frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j(\mathbf{u}) \frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial x_j} = \mathbf{0}$$

bringen, wobei die $\mathbf{A}_j(\mathbf{u}) = \left(\frac{\partial f_{i,j}(\mathbf{u})}{\partial u_k}\right)_{1 \le i,k \le p}$ die zu $\mathbf{f}_j$ gehörigen Jacobi-Matrizen sind. Man nennt das System **hyperbolisch**, wenn für jedes $\mathbf{u}(\mathbf{x},t) \in \Omega$ und alle $\mathbf{w} = (w_1, \cdots, w_n) \in \mathbb{R}^d$ die Matrix $\mathbf{A}(\mathbf{u}, \mathbf{w}) = \sum_{j=1}^{d} w_j \mathbf{A}_j(\mathbf{u})$ reell diagonalisierbar ist.
Üblicherweise wird das System (2.1) bezüglich eines Anfangswertes $\mathbf{u}_0 : \mathbb{R}^d \to \Omega$ untersucht. Das **Anfangswertproblem (oder auch Cauchy-Problem)** für die hyperbolische Erhaltungsgleichungen ist gegeben durch

$$\begin{aligned} \frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial t} + \sum_{j=1}^{d} \frac{\partial \mathbf{f}_j(\mathbf{u}(\mathbf{x},t))}{\partial x_j} = \mathbf{0} \qquad & \forall \mathbf{x} \in \mathbb{R}^d, t > 0 \\ \mathbf{u}(\mathbf{x},t) = \mathbf{u}_0(\mathbf{x}) \quad & \forall \mathbf{x} \in \mathbb{R}^d. \end{aligned} \tag{2.2}$$

Die Fragen nach der Existenz einer Lösung $\mathbf{u} : (\mathbf{x},t) \in \mathbb{R}^d \times \mathbb{R}_0^+ \longrightarrow \mathbf{u}(\mathbf{x},t) \in \Omega$ und der Eindeutigkeit stehen dabei im Mittelpunkt[1]. Bevor wir (2.2) genauer betrachten, motivieren wir noch, warum man beim System (2.1) von Erhaltungsgleichungen und bei $\mathbf{u}$ von einer Erhaltungsgröße spricht.
Dazu sei $D \subset \mathbb{R}^d$ ein beliebiges Gebiet mit hinreichend glattem Rand und $\mathbf{n} = (\nu_1, \cdots, \nu_n)$ der äußere Normalenvektor an ∂D. Integriert man Gleichung (2.1) bezüglich des Gebietes D und verwendet den Gaußschen Integralsatz, so erhält man

$$\begin{aligned} & \frac{\mathrm{d}}{\mathrm{d}t} \int_D \mathbf{u} \, \mathrm{d}\mathbf{x} + \sum_{j=1}^{d} \int_{\partial D} \mathbf{f}_j(\mathbf{u}) \cdot \nu_j \, \mathrm{d}S = 0 \\ \Longleftrightarrow \; & \frac{\mathrm{d}}{\mathrm{d}t} \int_D \mathbf{u} \, \mathrm{d}\mathbf{x} = - \sum_{j=1}^{d} \int_{\partial D} \mathbf{f}_j(\mathbf{u}) \cdot \nu_j \, \mathrm{d}S \end{aligned}$$

[1] Man findet durchaus auch Abhandlungen in der Literatur, zum Beispiel [68], die sich mit hyperbolischen Randwertproblemen beschäftigen. Jedoch sind diese eher die Ausnahme. Hyperbolische Randwertprobleme spielen in dieser Arbeit keinerlei Rolle. Das Anfangswertproblem muss sachgemäß gestellt sein, also die Anfangsbedingungen widerspruchsfrei zum Problem.

d.h. die zeitliche Veränderung von $\int_D \mathbf{u}\,\mathrm{d}\mathbf{x}$ ist gleich dem nach außen gerichteten Fluss über den Rand.
Kommen wir dazu (2.2) genauer zu untersuchen. Man spricht von einer **klassischen Lösung** $\mathbf{u} : \mathbb{R}^d \times \mathbb{R}_0^+ \longrightarrow \Omega$ des Cauchy-Problems, wenn die Funktion $\mathbf{u} \in \left(C^1(\mathbb{R}^d \times \mathbb{R}_0^+)\right)^p$ (2.2) punktweise erfüllt. Allerdings kann man bei hyperbolischen Erhaltungsgleichungen, selbst bei sehr glatten Anfangsbedingungen, nicht garantieren, dass für jede Zeit $t > 0$ klassische Lösungen existieren. Die Lösungen können nach einer bestimmten Zeit Unstetigkeiten entwickeln. Hier kommt es zum Zusammenbruch der klassischen Lösung, was man am folgenden einfachen Beispiel erkennt.

Beispiel 2.1. Wir untersuchen eine eindimensionale skalare Erhaltungsgleichung mit Flussfunktion $f : \mathbb{R} \to \mathbb{R}$, $f \in C^1(\mathbb{R})$. Es sei folgendes Anfangswertproblem gegeben:

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} + \frac{\partial}{\partial x} f(u(x,t)) = 0 \qquad & \forall x \in \mathbb{R}, t > 0 \\ u(x,0) = u_0(x) \quad & \forall x \in \mathbb{R}. \end{aligned}$$

gegeben. Sei u klassische Lösung des Problems. Mit $a(u) := f'(u)$ erhält man

$$\frac{\partial u(x,t)}{\partial t} + a(u)\frac{\partial u(x,t)}{\partial x} = 0.$$

Man betrachtet die **Charakteristiken**[2], das heißt die Lösungen der gewöhnlichen Differentialgleichung

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = a(u(x(t),t)). \tag{2.3}$$

Wegen

$$\frac{\mathrm{d}u(x(t),t)}{\mathrm{d}t} = \frac{\partial u(x(t),t)}{\partial t} + \frac{\partial u(x(t),t)}{\partial x}\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \frac{\partial u(x(t),t)}{\partial t} + a(u(x(t),t)\frac{\partial u(x(t),t)}{\partial x} = 0$$

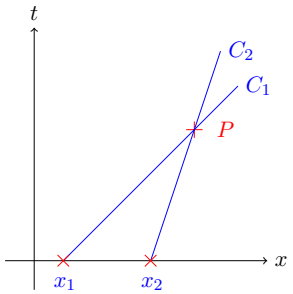
ist jede Lösung des Anfangswertproblems konstant entlang der Charakteristiken. Damit folgt unter Berücksichtigung von Gleichung (2.3), dass für Charakteristiken die Gleichung

$$x(t) = x_0 + ta(u_0(x_0))$$

gilt. Die Steigung ist dabei abhängig von den Anfangswerten u_0. Existieren nun zwei Punkte $x_1 < x_2$ mit

$$m_1 := \frac{1}{a(u_0(x_1))} < \frac{1}{a(u_0(x_2))} := m_2,$$

[2]Für eine Einführung in die Methode der Charakteristiken verweisen wir auf [23].



dann besitzen die Charakteristiken C_1 und C_2 die Steigung m_1 bzw. m_2 und schneiden sich notwendigerweise in einem Punkt P. Hier müsste die Lösung u sowohl die Werte von $u_0(x_1)$ als auch $u_0(x_2)$ annehmen, was zum Widerspruch führt. Somit kann die Differentialgleichung keine stetige Lösung in diesem Punkt besitzen, unabhängig von der Glattheit der Funktionen u_0 und f. Es bildet sich hier ein sogenannter **Stoß**.

Wir haben durch dieses einfache Beispiel gesehen, dass selbst bei glatten Anfangsbedingungen im nichtlinearen Fall $a'(u) \neq 0$ sich Unstetigkeiten nach einer gewissen Zeit entwickeln können. Dieses Verhalten macht gerade die hyperbolischen Erhaltungsgleichungen so speziell und man muss ihr Verhalten infolgedessen genauer analysieren. Der klassische Lösungsbegriff ist jedoch nicht mehr ausreichend für die Untersuchung und daher wird der Lösungsbegriff erweitert.
Sei hierfür $C_0^1(\mathbb{R}^d \times \mathbb{R}_0^+)$ der Raum der C^1-Funktionen mit kompaktem Träger in $\mathbb{R}^d \times \mathbb{R}_0^+$ und $\mathbf{u}$ Lösung von (2.2). Wir multiplizieren eine Testfunktion $\boldsymbol{\varphi} \in \left(C_0^1(\mathbb{R}^d \times \mathbb{R}_0^+)\right)^p$ an die Differentialgleichung (2.1), integrieren sowohl über den Raum als auch die Zeit und verwenden partielle Integration in den einzelnen Dimensionen. Es ergibt sich

$$\begin{aligned} 0 &= \int\limits_{\mathbb{R}^d \times \mathbb{R}_0^+} \left[\frac{\partial}{\partial t}\mathbf{u} + \sum_{j=1}^{d} \frac{\partial}{\partial x_j}\mathbf{f}_j(\mathbf{u}) \right] \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \\ &= - \int\limits_{\mathbb{R}^d \times \mathbb{R}_0^+} \left[\mathbf{u} \cdot \frac{\partial}{\partial t}\boldsymbol{\varphi} + \sum_{j=1}^{d} \mathbf{f}_j(\mathbf{u}) \cdot \frac{\partial}{\partial x_j}\boldsymbol{\varphi} \right] \mathrm{d}\mathbf{x} \, \mathrm{d}t - \int\limits_{\mathbb{R}^d} \mathbf{u}(\mathbf{x}, 0) \cdot \boldsymbol{\varphi}(\mathbf{x}, 0) \mathrm{d}\mathbf{x}. \end{aligned}$$

Dabei bezeichnet „$\cdot$“ das Skalarprodukt des $\mathbb{R}^p$. Eine klassische Lösung $\mathbf{u}$ von (2.2) erfüllt somit

$$\int\limits_{\mathbb{R}^d \times \mathbb{R}_0^+} \left[\mathbf{u} \cdot \frac{\partial}{\partial t}\boldsymbol{\varphi} + \sum_{j=1}^{d} \mathbf{f}_j(\mathbf{u}) \cdot \frac{\partial}{\partial x_j}\boldsymbol{\varphi} \right] \mathrm{d}\mathbf{x} \, \mathrm{d}t + \int\limits_{\mathbb{R}^d} \mathbf{u}_0(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{x}, 0) \mathrm{d}\mathbf{x} = 0. \tag{2.4}$$

Weiterhin muss die Funktion $\mathbf{u}$ nicht unbedingt differenzierbar sein, damit (2.4) gilt. Mit $\mathrm{L}^\infty_{loc}(\mathbb{R}^d \times \mathbb{R}_0^\infty)$ wird der Raum der lokal beschränkten messbaren Funktionen auf $\mathbb{R}^d \times \mathbb{R}_0^\infty$ bezeichnet und bereits eine Funktion $\mathbf{u} \in (\mathrm{L}^\infty_{loc}(\mathbb{R}^d \times \mathbb{R}_0^+))^p$ kann der Gleichung (2.4) genügen. Hierüber werden schließlich **schwache Lösungen** definiert.

Definition 2.2. Sei $\mathbf{u}_0 \in (\mathrm{L}^\infty_{loc}(\mathbb{R}^d))^p$. Eine Funktion $\mathbf{u} \in (\mathrm{L}^\infty_{loc}(\mathbb{R}^d \times \mathbb{R}_0^+))^p$ heißt **schwache Lösung** des Anfangswertproblem (2.2) oder **Lösung im Distributionensinn**, falls $\mathbf{u}(\mathbf{x}, t) \in \Omega$ für fast alle $(\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R}_0^+$ und die Gleichung (2.4) für alle Testfunktionen $\boldsymbol{\varphi} \in \left(C_0^1(\mathbb{R}^d \times \mathbb{R}_0^+)\right)^p$ erfüllt ist.

BEMERKUNG. Wir haben gesehen, dass jede klassische Lösung $\mathbf{u}$ auch eine schwache Lösung des Cauchy-Problems (2.2) ist. Andererseits erfüllt jede schwache Lösung (2.1) im Distributionensinn und ist die Lösung in einem Bereich stetig differenzierbar, so ist sie dort eine klassische Lösung.

Als nächstes wollen wir explizit das Verhalten der Lösungen im Bereich der Unstetigkeiten untersuchen. Wir treffen die Annahme, dass die Funktion **u stückweise glatt** ist. Dabei sprechen wir von einer stückweise glatten Funktion $\mathbf{u} : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}^p$, wenn eine endliche Anzahl an glatten, orientierbaren, n-dimensionalen Hyperflächen im $\mathbb{R}^n \times \mathbb{R}_0^+$ existiert, außerhalb derer $\mathbf{u}$ stetig differenzierbar ist und auf denen $\mathbf{u}$ eine Sprungunstetigkeit besitzt. Sei H_1 eine dieser Hyperflächen und $\mathbf{n} = (\nu_t, \nu_1, \cdots, \nu_n) \neq 0$ der Normalenvektor an H_1, dann wird mit

$$\mathbf{u}_{\pm\mathbf{n}}(t, \mathbf{x}) = \lim_{\varepsilon \to 0} \mathbf{u}((t, \mathbf{x}) \pm \varepsilon \mathbf{n})$$

der Grenzwert von $\mathbf{u}$ bezüglich der Seiten von H_1 ausgedrückt.
Für stückweise glatte Funktion kann man den folgenden Satz von Rankine-Hugoniot beweisen.

Satz 2.3. *Sei $\boldsymbol{u} : \mathbb{R}^d \times \mathbb{R}_0^+ \to \Omega$ eine stückweise glatte Funktion. Die Funktion $\boldsymbol{u}$ ist genau dann Lösung von* (2.1) *im Distributionensinn, wenn nachfolgende zwei Bedingungen erfüllt sind:*

(a) $\boldsymbol{u}$ ist klassische Lösung von (2.1) *in dem Gebiet, in welchem $\boldsymbol{u}$ stetig differenzierbar ist.*

*(b) $\boldsymbol{u}$ erfüllt die **Rankine-Hugoniot-Bedingung***

$$(\boldsymbol{u}_+ - \boldsymbol{u}_-)\nu_t + \sum_{j=1}^{d} (\boldsymbol{f}_j(\boldsymbol{u}_+) - \boldsymbol{f}_j(\boldsymbol{u}_-))\nu_j = \boldsymbol{0}$$

auf den Hyperflächen, auf denen $\boldsymbol{u}$ unstetig ist.

Beweis. [28, S.29ff] □

BEMERKUNG. In der Literatur findet man auch die Notation $[\mathbf{u}] := (\mathbf{u}_+ - \mathbf{u}_-)$ und die Rankine-Hugoniot-Bedingung (Sprungbedingung) lässt sich dann wie folgt beschreiben

$$\nu_t[\mathbf{u}] + \sum_{j=1}^{d} [\mathbf{f}_j(\mathbf{u})]\nu_j = \mathbf{0} \Longleftrightarrow s[\mathbf{u}] = \sum_{j=1}^{d} [\mathbf{f}_j(\mathbf{u})]v_j = \mathbf{0}.$$

Dabei wurde der Normalenvektor durch $\mathbf{n} = (-s, \mathbf{v})$ mit $\mathbf{v} \in \mathbb{R}^d, ||\mathbf{v}||_2 = 1$ ausgedrückt. Die Größe s nennt man **Stoßgeschwindigkeit**. Man kann die Größen s und $\mathbf{v}$ dahingehen interpretieren, dass sie Ausbreitungsgeschwindigkeit und Richtung der Unstetigkeit angeben.

Kommen wir zurück zur Definition 2.2 der schwachen Lösung. Die Lösung von Gleichung (2.4) muss - sofern sie existiert - nicht eindeutig sein.

Beispiel 2.4. Wir betrachten die nicht viskose Burgers-Gleichung

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}\left(\frac{u^2(x,t)}{2}\right) = 0$$

zu folgenden Anfangswerten

$$u_0(x) = \begin{cases} u_l & x < 0 \\ u_r & x > 0, \end{cases}$$

wobei $u_l \neq u_r$ gelte. Ein solches Anfangswertproblem heißt auch **Riemann-Problem**[3]. Man berechnet die Stoßgeschwindigkeit $s = \frac{u_+ - u_-}{2}$ und hiermit ergibt sich als *erste* schwache Lösung

$$u(x,t) = \begin{cases} u_l & x < st \\ u_r & x > st. \end{cases} \tag{2.5}$$

Des Weiteren wählt man sich einen Parameter $a \geq \max\{u_l, -u_r\}$. Durch elementare Rechnung kann man zeigen, dass die Funktionen

$$u_a(x,t) = \begin{cases} u_l & x < s_1 t \\ -a & s_1 t < x < 0 \\ a & 0 < x < s_2 t \\ u_r & s_2 t < x \end{cases} \quad \text{mit } s_1 = \frac{u_l - a}{2}, \quad s_2 = \frac{u_r + a}{2}$$

die Gleichung (2.4) erfüllen und damit auch schwache Lösungen des Anfangswertproblems sind. Die schwachen Lösungen sind dabei abhängig von dem fest gewählten Parameter a und weisen eine Sprungunstetigkeit auf. Unter der Voraussetzung $u_l \leq u_r$ erhält man mit der Methode der Charakteristiken sogar folgende stetige Lösung des Anfangswertproblems

$$u(x,t) = \begin{cases} u_l & x \leq u_l t \\ \frac{x}{t} & u_l t < x \leq u_r t \\ u_r & u_r t \leq x. \end{cases} \tag{2.6}$$

Diese Lösung wird auch als **Verdünnungswelle** bezeichnet. Die Abbildung 2.1 zeigt die schwachen Lösungen (2.5) und (2.6) der Burgers-Gleichung. Als Anfangswerte wurden $u_l = -1$ und $u_r = 1$ gewählt.

Schwache Lösungen können demnach lokal existieren, sie sind jedoch im Allgemeinen nicht eindeutig. Man benötigt eine weitere Bedingung um unter der Anzahl an schwachen Lösungen diejenige auszuwählen, welche die Kausalitätsbedingung erfüllt, also einen physikalisch plausiblen Vorgang beschreibt. Hierzu gibt es mehrere Ansätze, von denen wir zwei vorstellen wollen.

[3]Ein Riemann-Problem ist ein spezielles Anfangswertproblem (2.2). Die Anfangswerte sind dabei bis auf eine Sprungunstetigkeit konstant vorausgesetzt. Mehr zu diesem Thema findet man in der Literatur [78].

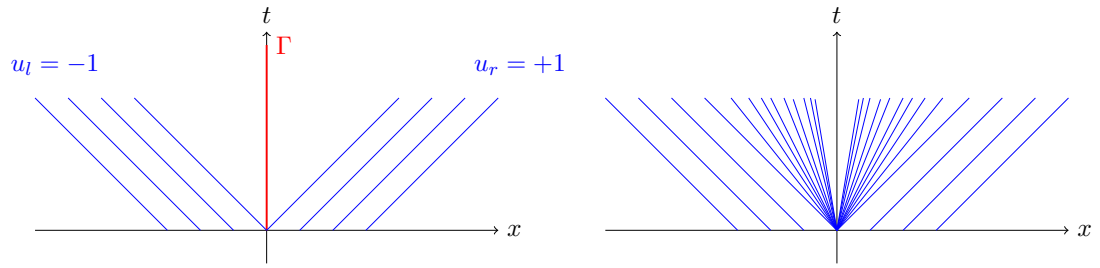


Abbildung 2.1: *Links die Lösung von* (2.5) *und rechts von* (2.6)

Den ersten Ansatz findet man auch als **Viskositätsmethode** in der Literatur wieder. Man löst nicht mehr das Cauchy-Problem (2.2), sondern addiert zu der hyperbolischen Erhaltungsgleichung (2.1) einen kleinen Viskositätsterm hinzu. Man untersucht also das parabolische System

$$\frac{\partial \mathbf{u}_\varepsilon(\mathbf{x},t)}{\partial t} + \sum_{j=1}^{d} \frac{\partial \mathbf{f}_j(\mathbf{u}_\varepsilon(\mathbf{x},t))}{\partial x_j} = \varepsilon \triangle \mathbf{u}_\varepsilon(\mathbf{x},t) \qquad \forall \mathbf{x} \in \mathbb{R}^d,\ \forall \varepsilon > 0,\ t > 0. \tag{2.7}$$

Man bestimmt die Lösungen $\mathbf{u}_\varepsilon$ und ihr Verhalten für $\varepsilon \to 0$. Die Idee dabei ist, dass die Lösungen von (2.7) gegen die relevanten Lösungen von (2.2) konvergieren.
Dieser Ansatz ist physikalisch motiviert; es beschreiben hyperbolische Erhaltungsgleichungen oft komplexere Modelle in der Natur, bei denen verschiedene Annahmen getroffen wurden, um das Modell zu vereinfachen. Die Euler-Gleichungen der Gasdynamik lassen sich aus den kompressiblen Navier-Stokes-Gleichungen ableiten, indem in den Navier-Stokes-Gleichungen Viskosität und Wärmeleitung nicht mehr berücksichtigt werden. Unter der Annahme, dass die kompressiblen Navier-Stokes-Gleichungen Lösungen für eine gegebene Nullfolge von Wärmeleitungs- und Viskositätskoeffizienten besitzen, erwartet man die Konvergenz dieser Lösungen gegen die Lösung der Euler-Gleichung.
Auch bei der Konstruktion von numerischen Verfahren ist die Viskositätsmethode von Bedeutung. Wir werden in Kapitel 4 nochmals genauer auf sie eingehen und ihren Zusammenhang mit modaler Filterung darlegen. Dabei sei jedoch jetzt schon erwähnt, dass in diesem Kontext der Viskositätsterm $\varepsilon \triangle \mathbf{u}$ komplexer sein wird.
In der Praxis scheint es wenig sinnvoll erst (2.2) zu lösen und anschließend noch die parabolische Gleichung (2.7) mit dem Grenzwertprozess zu betrachten, um aus den schwachen Lösungen von (2.2) schließlich die *relevante* Lösung auszuwählen. Auch konvergieren nicht immer die Lösungen der Viskositätgleichung (2.7) gegen die Lösung der hyperbolischen Erhaltungsgleichung (2.1). Daher sind Bedingungen von Interesse, welche direkt ein Kriterium an die schwachen Lösungen von (2.2) stellen. Man kommt zu den

Entropiebedingungen bzw. dem **Entropiekonzept**. Diese Kriterien sind ebenfalls physikalisch motiviert. Sie erhalten ihren Namen in Anlehnung an den zweiten Hauptsatz der Thermodynamik. Dieser besagt, dass sich die Entropie nicht verringern darf.
Wir stellen hierbei eine spezielle Entropiebedingung vor, welche man aus der Viskositätsmethode ableiten kann. Der Ansatz dabei ist, dass man die Entropie als mathematische Funktion bezüglich $\mathbf{u}$ definiert. Die Entropie ist bei glatten Lösungen der Erhaltungsgleichung (2.1) ebenfalls Erhaltungsgröße und darf sich beim Durchqueren von Unstetigkeiten nicht verringern. Wir beschränken uns dabei auf konvexe Entropiefunktionen und erhalten schließlich nachfolgende Definition für die Entropie und die Entropielösung.

Definition 2.5. Der Zustandsraum Ω sei konvex. Eine konvexe Funktion $\eta : \Omega \to \mathbb{R}$ wird **Entropiefunktion** für die Erhaltungsgleichung (2.1) genannt, wenn eine Funktion $\boldsymbol{\psi} : \mathbb{R}^p \to \mathbb{R}^d$ existiert, so dass die Gleichung

$$(\nabla_{\mathbf{u}}\eta)^T \mathbf{A}_j(\mathbf{u}) = (\nabla_{\mathbf{u}}\psi_j)^T, \qquad 1 \leq j \leq d$$

für alle $\mathbf{u} \in \Omega$ erfüllt ist[4]. Das Paar $(\eta, \boldsymbol{\psi})$ bezeichnet man als **Entropie-Entropiefluss-Paar**.
Eine schwache Lösung $\mathbf{u}$ des Anfangswertproblems (2.2) heißt schließlich **Entropielösung**, wenn $\mathbf{u}$ für alle Entropie-Entropiefluss-Paare $(\eta, \boldsymbol{\psi})$ von (2.1) die Entropieungleichung

$$\frac{\partial \eta}{\partial t}(\mathbf{u}) + \sum_{j=1}^{d} \frac{\partial}{\partial x_j}\psi_j(\mathbf{u}) \leq 0 \tag{2.8}$$

im Distributionensinn auf $\mathbb{R}^d \times \mathbb{R}^+$ erfüllt.

Die Ungleichung (2.8) bedeutet nichts anderes, als dass für alle Testfunktionen $\varphi \in C_0^\infty(\mathbb{R} \times \mathbb{R}^+),\ \varphi \geq 0$,

$$\int\limits_{\mathbb{R}^d \times \mathbb{R}^+} \left[\eta(\mathbf{u})\ \frac{\partial}{\partial t}\varphi + \sum_{j=1}^{d} \psi_j(\mathbf{u})\ \frac{\partial}{\partial x_j}\varphi \right]\ \mathrm{d}\mathbf{x} \geq 0 \tag{2.9}$$

gelten muss. Unter den schwachen Lösungen sucht man diejenige heraus, welche zusätzlich (2.9) erfüllt. Das Problem der Eindeutigkeit der Lösung ist trotz großer Anstrengungen in diesem Forschungsgebiet leider noch nicht geklärt. Ausschließlich für skalare Gleichungen konnte die Existenz eindeutiger Entropielösungen bewiesen werden. Für allgemeine Systeme ist dies nach wie vor eine offene Frage.
Es gibt viele weitere Entropiebedingungen, die auf unterschiedliche Art und Weise motiviert sind. Dem Leser sei hier die Literatur [28], [48] und [84] für eine detailliertere Betrachtung empfohlen. Wir beschränken uns auf die Bedingung (2.8), da man sie aus der Viskositätsmethode heraus entwickeln kann. So findet man das nachfolgende Resultat in der Literatur.

[4] Dabei bezeichnet $\nabla_{\mathbf{u}}$ den Nablaoperator bezüglich $\mathbf{u}$.

Satz 2.6. *Sei $(\eta, \boldsymbol{\psi})$ ein Entropie-Entropiefluss-Paar für die Erhaltungsgleichung* (2.1) *mit $\eta \in C^2(\mathbb{R}^p)$. Weiterhin sei $(\boldsymbol{u}_\varepsilon)_\varepsilon$ eine Familie von glatten Lösungen der parabolischen Gleichung* (2.7) *mit den Eigenschaften*

- $||\boldsymbol{u}_\varepsilon||_{(\mathrm{L}^\infty(\mathbb{R}^d \times \mathbb{R}_0^+))^p} \leq C$,
- $\lim_{\varepsilon \to 0} \boldsymbol{u}_\varepsilon = \boldsymbol{u}$ *fast überall auf* $\mathbb{R}^d \times \mathbb{R}_0^+$,

wobei C eine Konstante unabhängig von ε ist.
Dann ist $\boldsymbol{u}$ Lösung von (2.1) *im Distributionensinn und erfüllt die Entropiebedingung* (2.8).

Beweis. [28, S.42f.] □

2.2 Numerische Methoden

Die Entwicklung von Unstetigkeiten nach einem bestimmten Zeitabschnitt, selbst für glatte Anfangsbedingungen, machen die hyperbolischen Erhaltungsgleichungen so speziell und sorgen für eine gesonderte Betrachtung auch hinsichtlich der Konstruktion numerischer Verfahren. Es gibt eine ganze Reihe an unterschiedlichen Ansätzen, hyperbolische Erhaltungsgleichungen numerisch zu lösen. Hierfür verweisen wir auf die Übersichtsartikel [21] und [82]. In dieser Arbeit beschäftigen wir uns allerdings ausschließlich mit spektralen Methoden und verwenden insbesondere das **Spektrale-Differenzen-Verfahren**. Nichtsdestotrotz kann man die orthogonalen Polynome aus Kapitel drei und fünf auch bei anderen numerischen Methoden benutzen, wie beispielsweise beim diskontinuierlichen Galerkin-Verfahren aus [63]. Dabei soll ein System von Erhaltungsgleichungen

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x}, t) = -\sum_{j=1}^{d} \frac{\partial \mathbf{f}_j(\mathbf{u}(\mathbf{x}, t))}{\partial x_j} \tag{2.10}$$

für $\mathbf{x} \in G \subset \mathbb{R}^d,\ t > t_0$ für ein $t_0 \in \mathbb{R}_0^+$ und Flussfunktionen $\mathbf{f}_j$ numerisch gelöst werden. Man diskretisiert sowohl in der Zeit als auch im Raum.

2.2.1 Zeitdiskretisierung

Bei der Zeitdiskretisierung untersucht man die Gleichung (2.10) für festes $\mathbf{x}$ als gewöhnliche Differentialgleichung in der Zeit t. Aus der Literatur [9] und [34] sind hierfür Ansätze bekannt und man spricht in diesem Kontext von **Zeitintegration**.
Prinzipiell unterscheidet man bei numerischen Verfahren von gewöhnlichen Differentialgleichungen zwischen **expliziten** und **impliziten** Methoden. Bezeichnet man mit U^i die im i-ten Zeitschritt berechnete numerische Lösung, so verwenden explizite Verfahren zur

Berechnung der U^{k+1} im $(k+1)$-ten Zeitschritt ausschließlich Werte U^i von vorhergehenden Zeitschritten t_0 bis t_k. Bei impliziten Verfahren hingegen werden auch Werte U^i zu späteren Zeitpunkten verwendet. So liest sich das explizite bzw. implizite Eulerverfahren für Gleichung (2.10) wie folgt:

$$U^{k+1}(\mathbf{x}) = U^k(\mathbf{x}) - \int_{t_k}^{t_{k+1}} \left(\sum_{j=1}^{d} \frac{\partial \mathbf{f}_j(U^k(\mathbf{x}))}{\partial x_j} \right) \mathrm{d}t,$$

$$U^{k+1}(\mathbf{x}) = U^k(\mathbf{x}) - \int_{t_k}^{t_{k+1}} \left(\sum_{j=1}^{d} \frac{\partial \mathbf{f}_j(U^{k+1}(\mathbf{x}))}{\partial x_j} \right) \mathrm{d}t,$$

wobei $U^0(\mathbf{x}) = \mathbf{u}(\mathbf{x}, t_0)$ eine vorgegebene Anfangsbedingung ist. Explizite Zeitintegrationen sind im Allgemeinen leicht zu implementieren, da sie lediglich auf bereits bekannte Werte U^k zurückgreifen. Jedoch weisen sie häufig Stabilitätsprobleme auf, so dass man sehr kleine Zeitschritte verwenden muss. Die Größe der Zeitschritte kann durch die **CFL-Zahl**[5] ermittelt werden.
Bei impliziten Verfahren kann man hingegen größere Zeitschritte verwenden, da hier weniger Probleme bezüglich der Stabilität entstehen. Allerdings muss man zur Berechnung der Werte U^k stets ein Gleichungssystem lösen, was sich negativ auf den Speicherbedarf sowie die Rechenzeit auswirkt.
Die Auswahl des expliziten bzw. impliziten Zeitschrittverfahrens ist gerade in der Praxis von Interesse und ein autonomes Forschungsgebiet.
In dieser Arbeit beschränken wir uns allerdings auf ein bereits bekanntes und oft genutztes Zeitintegrationsverfahren hoher Ordnung. Wir verwenden das **Runge-Kutta Verfahren** vierter Ordnung mit geringem Speicherbedarf aus [12]. Dieses lässt sich schreiben als

$$\begin{aligned}
U^{(1)} &= U^k, \\
V^{(j)} &= A_j V^{(j-1)} + \triangle t L(U^{(j-1)}, t_k + c_j \triangle t), \qquad A_1 = 0,\ j = 1, \ldots, 5, \\
U^{(j)} &= U^{(j-1)} + B_j V^{(j)}, \qquad j = 1, \ldots, 5, \\
U^{k+1} &= U^{(5)},
\end{aligned}$$

[5] Bei der Courant-Friedrich-Levy-Zahl handelt es sich um eine dimensionslose Größe C, die sich im allgemeinen Fall aus einem Quotienten der Zeitschritte Δt_k, der Gitterlänge Δx_k und der gesuchten Größe u berechnen lässt. Die CFL-Zahl darf ein bestimmtes Maximum nicht überschreiten (im expliziten Fall häufig Eins), damit die CFL-Bedingung erfüllt ist, was ein notwendiges Kriterium für die Stabilität eines finiten Verfahren ist. Eine detaillierte Betrachtung der CFL-Zahl findet man in der Literatur [28] bzw. [78]. Dabei wird die Größe C in [28] auch geometrisch motiviert.

wobei $\triangle t = t_{k+1} - t_k$ die zugehörigen Zeitschritte, U^k die numerisch berechneten Lösungen und L den zugehörigen Wert des diskreten räumlichen Operators bezeichnet. Die Koeffizienten A_j, B_j und c_j sind

$$A_1 = 0, \qquad B_1 = \frac{1432997174477}{9575080441755}, \qquad c_1 = 0,$$

$$A_2 = -\frac{567301805773}{1357537059087}, \qquad B_2 = \frac{5161836677717}{13612068292357}, \qquad c_2 = \frac{1432997174477}{9575080441755},$$

$$A_3 = -\frac{2404267990393}{2016746695238}, \qquad B_3 = \frac{1720146321549}{2090206949498}, \qquad c_3 = \frac{2526269341429}{6820363962896},$$

$$A_4 = -\frac{3550918686646}{2091501179385}, \qquad B_4 = \frac{3134564353537}{4481467310338}, \qquad c_4 = \frac{2006345519317}{3224310063776},$$

$$A_5 = -\frac{1275806237668}{842570457699}, \qquad B_5 = \frac{2277821191437}{14882151754819}, \qquad c_5 = \frac{2802321613138}{2924317926251}.$$

2.2.2 Spektrale-Differenzen-Verfahren

Im folgenden Abschnitt beschreiben wir das **Spektrale-Differenzen-Verfahren** (kurz: SD-Verfahren) aus [86]. Dieses erweitern wir im späteren Verlauf durch allgemeinere Polynome und modale Filter. Auf die genaue Implementierung werden wir jedoch nicht eingehen, dafür verweisen wir auf die Doktorarbeit [86].
Wie der Name **spektral** schon beinhaltet, gehört dieses Verfahren zur allgemeinen Klasse **spektraler Methoden**, deren grundlegender Ansatz die Annahme ist, dass die Lösung $\mathbf{u}(\mathbf{x}, t)$ durch eine abgeschnittene Reihenentwicklung

$$\mathbf{u}(\mathbf{x}) \approx \mathbf{u}_N(\mathbf{x}) = \sum_{l=1}^{N} \hat{\boldsymbol{u}}_l \; \phi_l(\mathbf{x})$$

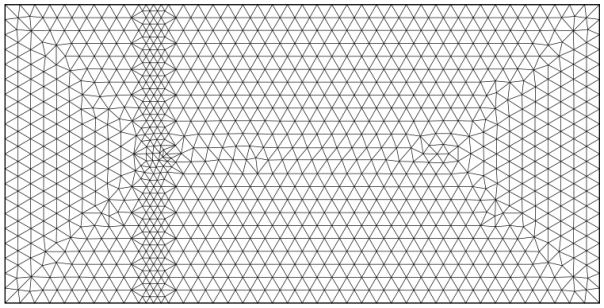
beliebig genau in einer gewählten Norm approximiert werden kann. Dabei ist $\hat{\boldsymbol{u}}_l$ der p-Vektor der Koeffizienten und $\{\phi_1, \ldots, \phi_N\}$ Basis eines Funktionenraumes. Aus Gründen der Recheneffizienz und der Stabilität wählt man prinzipiell eine orthogonale Basis bezüglich eines gewichteten Funktionenraums. Beispielsweise wählt man bei der sogenannten **Fourier-Methode** trigonometrische Funktionen, die eine Basis des Raumes $\mathrm{L}^2[0, 2\pi]$ darstellen.
Sowohl die Wahl der Basis und des Funktionenraumes als auch das Verfahren zur Berechnung der Koeffizienten spezifiziert die Methoden weiter. Mehr dazu findet man in der Literatur [11].
Wir verwenden schließlich das SD-Verfahren zur räumlichen Diskretisierung, das heißt zur Berechnung der Flussfunktionen. Hierbei ist die Idee, die Erhaltungsgleichung an

bestimmten Lösungspunkten $\mathbf{x}_j$ in jedem Zeitschritt t_k zu diskretisieren. Wir beschränken uns in dieser Arbeit auf zwei Raumdimensionen und eine konforme Triangulierung des zugrunde liegenden Gebietes G.

Definition 2.7. Eine Menge $\mathcal{T}(\overline{G}) = \{\tau_i | 1 \leq i \leq N_\mathcal{T}\}$ bestehend aus Dreiecken τ_i heißt **Triangulierung** von $\overline{G} \subset \mathbb{R}^2$, falls gilt:

- $\overline{G} = \bigcup_{i=1}^{N_\mathcal{T}} \tau_i$,
- jedes Dreieck $\tau_i \in \mathcal{T}$ ist abgeschlossen und hat ein nichtleeres Innere,
- es gilt für alle $\tau_i,\ \tau_j \in \mathcal{T}$ mit $i \neq j$: $\mathring{\tau}_i \cap \mathring{\tau}_j = \varnothing$.

Zusätzlich heißt die Triangulierung $\mathcal{T}(\overline{G})$ **konform**, wenn jede Kante eines Dreiecks $\tau_i \in \mathcal{T}$ entweder Teilmenge von ∂G oder Kante **genau** eines anderen Dreieckes $\tau_j \in \mathcal{T},\ i \neq j$ ist.



In der folgenden Graphik sehen wir eine konforme Triangulierung eines rechteckigen Gebietes mit einem **unstrukturierten** Dreiecksgitter. Dabei benutzen wir ausschließlich unstrukturierte Gitter in unseren numerischen Beispielen, da sie flexibler einsetzbar sind. Man kann mit ihnen komplexe Geometrien beschreiben und jedes Dreieck τ_i kann auf ein Standardelement transformiert werden. In unserem Fall ist das Standardelement das Einheitsdreieck

$$\mathbb{T} := \{(\xi, \eta) |\ 0 \leq \xi, \eta \leq 1,\ \xi + \eta \leq 1\}.$$

Dabei verläuft die Übertragung wie folgt:
Gegeben sei ein Dreieckselement $\tau_i \in \mathcal{T}$ mit den Eckpunkten $\mathbf{x}_j(i) = (x_j(i),\ y_j(i))$ $(j = 0, 1, 2)$ und den Richtungsvektoren $\mathbf{g}_l(i) := \mathbf{x}_l(i) - \mathbf{x}_0(i)$, $l = 1, 2$ vom Punkt $\mathbf{x}_0(i)$ nach $\mathbf{x}_l(i)$. Damit lässt sich jeder Punkt im Dreieck τ_i darstellen als

$$\mathbf{x}(i) = \mathbf{x}_0(i) + \xi \mathbf{g}_1(i) + \eta \mathbf{g}_2(i)$$

mit $(\xi, \eta) \in \mathbb{T}$. Lösen wir diese Gleichung nach (ξ, η) auf. So erhalten wir

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} x_1(i) - x_0(i) & x_2(i) - x_0(i) \\ y_1(i) - y_0(i) & y_2(i) - y_2(i) \end{pmatrix}^{-1} \begin{pmatrix} x(i) - x_0(i) \\ y(i) - y_0(i) \end{pmatrix}$$

und durch elementare Rechnung schließlich

$$\begin{pmatrix}\xi\\ \eta\end{pmatrix} = \frac{1}{2V(i)}\begin{pmatrix} y_2(i)-y_2(i) & -x_2(i)+x_0(i)\\ -y_1(i)+y_0(i) & x_1(i)-x_0(i)\end{pmatrix}\begin{pmatrix}x(i)-x_0(i)\\ y(i)-y_0(i)\end{pmatrix}. \tag{2.11}$$

Dabei bezeichnet $V(i)$ den Flächeninhalt des Dreiecks τ_i. Durch die Gleichung (2.11) ist somit eine Abbildung $T_i : \tau_i \longrightarrow \mathbb{T}$ definiert. Siehe dazu auch die Abbildung 2.2.

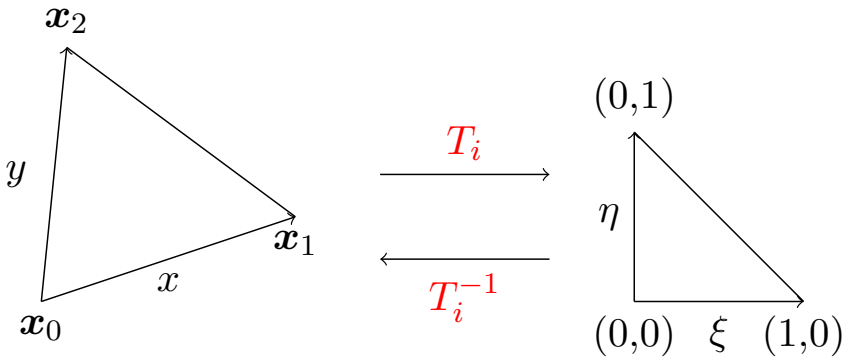


Abbildung 2.2: *Transformation eines beliebigen Dreiecks auf das Standardelement*

Um das universelle Aktualisierungsschema zu erklären, beschränken wir uns zunächst auf den skalaren Fall. Die Rekonstruktion des Flusses erfolgt in jeder Zelle τ_i und in jeder seiner Komponenten $(f_1, f_2)^T$ zur Zeit t. Man wählt N_F-Flusspunkte $\hat{\mathbf{x}}_m$, an denen der Fluss berechnet wird. Weiterhin seien $I_m \subset \mathbb{N}$ eine geeignete Indexmenge und $\mathbb{P}_{n+1}(\mathbb{T}, h)$ ist der Raum der Polynome vom Grad höchstens $n+1$ und Gewichtsfunktion h auf $\mathbb{T}$. $\{\phi_m : \mathbb{T} \to \mathbb{R} | m \in I_m\}$ ist eine orthogonale Basis von $\mathbb{P}_{n+1}(\mathbb{T}, h)$. Der Fluss $\mathbf{F} = (f_1, f_2)^T$ wird dargestellt durch

$$\mathbf{F}(u(\mathbf{x},t)) := \begin{pmatrix} f_1(\mathbf{x},t)\\ f_2(\mathbf{x},t)\end{pmatrix} = \sum_{m=1}^{N_F}\begin{pmatrix}\hat{f}_{m,1}(t)\\ \hat{f}_{m,2}(t)\end{pmatrix}\phi_m(T_i(\mathbf{x})), \tag{2.12}$$

wobei wir auf die Berechnung der Koeffizienten $\hat{f}_{m,l}(t)$ später explizit eingehen werden. Setzen wir die Gleichung in die Erhaltungsgleichung (2.10) ein, so folgt wegen der Linearität des Differentialoperators

$$u_t(\mathbf{x},t) = -\sum_{m=1}^{N_F}\begin{pmatrix}\hat{f}_{m,1}(t)\\ \hat{f}_{m,2}(t)\end{pmatrix}\cdot\nabla_{\boldsymbol{x}}\phi_m(T_i(\mathbf{x})).$$

Wir verwenden die Abbildung T_i und es ergibt sich direkt aus der Kettenregel

$$\nabla_{\boldsymbol{x}}\phi_m(T_i(\mathbf{x})) = J_{T_i}\nabla_{\boldsymbol{\xi}}\phi_m(T_i(\boldsymbol{x})),$$

wobei die Matrix J_{T_i} gegeben ist durch die Transformation (2.11)

$$J_{T_i} = \frac{1}{2V(i)} \begin{pmatrix} y_2(i) - y_2(i) & -x_2(i) + x_0(i) \\ -y_1(i) + y_0(i) & x_1(i) - x_0(i) \end{pmatrix}.$$

Daher muss man die Ableitungen der Polynome ϕ_m lediglich bezüglich des Standardelements speichern. Es entfällt die Berechnung in jedem Element τ_i. Allerdings muss man dafür die Jacobi-Matrizen für alle Dreiecke τ_i speichern, wobei in diesen nur die inneren Normalenvektoren $\mathbf{g}_1(i)$ und $\mathbf{g}_2(i)$ stehen. Man gelangt zu folgendem universellen Aktualisierungsschema:

$$u_t(\mathbf{x}, t) = -\sum_{m=1}^{N_F} \begin{pmatrix} \hat{f}_{m,1}(t) \\ \hat{f}_{m,2}(t) \end{pmatrix} \cdot J_{T_i} \nabla_{\boldsymbol{\xi}} \phi_m(T_i(\boldsymbol{x})).$$

Die zeitliche Aktualisierung erfolgt an N_u Lösungspunkten $\mathbf{x}_j$.
Sowohl die Flusspunkte $\hat{\mathbf{x}}_m$ als auch die Lösungspunkte $\mathbf{x}_j$ müssen ausschließlich im Standarddreieck bestimmt werden und können durch die Abbildung T_i^{-1} auf jedes Dreieck übertragen werden, was vor allem speichertechnische Vorteile mit sich bringt. Falls $\boldsymbol{\xi}_j = T_i(\mathbf{x}_j)$ und $\boldsymbol{\xi}_m = T_i(\hat{\mathbf{x}}_m)$ nicht übereinstimmen, wird auch u an den Flusspunkten durch

$$u(\hat{\boldsymbol{x}}_m, t) = \sum_{j=1}^{N_u} \hat{u}(\boldsymbol{x}_j, t) \phi_j(\boldsymbol{\xi}_m)$$

rekonstruiert. Die Koeffizienten $\hat{u}(\boldsymbol{x}_j, t)$ werden analog zu $\hat{f}_{m,l}$ berechnet. Schließlich wird die resultierende gewöhnliche Differentialgleichung

$$u_t(\mathbf{x}_j, t) = -\sum_{m=1}^{N_F} \begin{pmatrix} \hat{f}_{m,1}(t) \\ \hat{f}_{m,2}(t) \end{pmatrix} \cdot J_{T_i} \nabla_{\boldsymbol{\xi}} \phi_m(T_i(\boldsymbol{x}_j))$$

durch das bereits beschriebene Runge-Kutta Verfahren gelöst.
Bei Systemen erfolgen die Flussrekonstruktion und die Aktualisierung in jeder Komponente der Erhaltungsvariable, doch der Fluss wird im ganzen System berechnet.

Polynominterpolation auf dem Dreieck

Die Lage der Fluss- und der Lösungspunkte hat wesentlichen Einfluss auf das Verfahren. Daher beschäftigen wir uns auch mit der Frage nach geeigneten Interpolationspunkten $\boldsymbol{\xi}_i$ auf dem Dreieck $\mathbb{T}$. Bei Verwendung eines äquidistanten Gitters ist ein klassisches Ergebnis der Polynominterpolation in einer Raumdimension, dass das Interpolationspolynom nicht notwendigerweise den qualitativen Verlauf der zu interpolierenden Funktion wiedergibt (siehe Runge-Phänomen in der Literatur [69] oder Kapitel 5).
Man betrachtet daher bereits in einer Raumdimension nicht-äquidistante Stützstellen wie beispielsweise Gauß-Lobatto-Punkte, welche die Randpunkte des Intervalls beinhalten.

Ein Maß, um die Güte der Interpolationspunkte und somit auch der Interpolation einzuschätzen, ist die sogenannte **Lebesgue-Konstante**[6] Λ_N. Diese ist für einen Definitionsbereich D und Lagrange-Polynome $L_i,\ i = 1, \dots, N$ definiert durch $\Lambda_N = \max\limits_{x \in D} \sum\limits_{i=1}^{N} |L_i(x)|$ und sollte möglichst klein sein.
Für ein äquidistantes Gitter ergibt sich beispielsweise folgendes asymptotische Verhalten aus [71]:

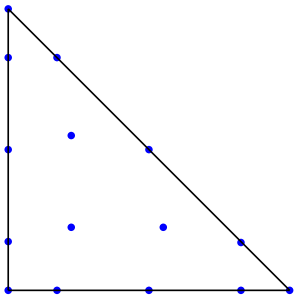
$$\Lambda_N \sim \frac{2^{N+1}}{eN \log N}, \quad N \to \infty.$$

In der Literatur wurden bereits verschiedene Punktverteilungen verwendet wie etwa die Fekete-Punkte [77], die durch Hesthaven [35] bzw. Chen und Babuška [14] konstruierten Punkte oder die warp-and-blended-Punkte von Warburton [83]. Jedoch sind die jeweiligen Berechnungen aufwendig, so dass wir in dieser Arbeit ausschließlich die zweidimensionale Erweiterung der Gauß-Lobatto-Punkte von Blyth und Pozrikidis aus [5] und [6] benutzen.

Definition 2.8. Seien $n \in \mathbb{N}$ und $\{v_0, \cdots, v_n\}$ Gauß-Lobatto-Punkte auf $[0, 1]$. Die $\frac{1}{2}(n+1)(n+2)$ **2D-Gauß-Lobatto-Punkte** auf $\mathbb{T}$ sind definiert durch die Koordinaten

$$\xi_i = \frac{1}{3}\left(1 + 2v_i - v_j - v_k\right), \quad \eta_j = \frac{1}{3}\left(1 + 2v_j - v_i - v_k\right),$$

mit $i = 0, 1, \dots, n$, $j = 0, 1, \dots, n - i$ und $k = n - i - j$.



$n = 4$

In der Abbildungen sind die **Gauss-Lobatto**-Punkte für $n = 4$ auf dem Dreieck gezeigt.
Weiterhin zeigt Tabelle 2.1, dass die Lebesgue-Konstante der 2D-Gauss-Lobatto-Punkte vergleichbar ist mit denen der anderen Punktverteilungen, so dass deren Verwendung gerechtfertigt ist.

[6] Allgemein beschränkt die Lebesgue-Konstante Λ_N den Interpolationsfehler bezüglich der Bestapproximation.

n	2D-Lobatto	Fekete	Chen and Babuška	Hesthaven	warp-and-blend
3	2.11	2.11	2.11	2.11	2.11
6	3.87	4.17	3.79	4.08	3.70
9	7.39	6.80	6.80	6.87	5.74

Tabelle 2.1: *Lebesgue-Konstante unterschiedlicher Interpolationspunkte aus* [6] *und* [83]

Berechnung der Koeffizienten

Kommen wir nun dazu, die Berechnung der Koeffizienten $\hat{f}$ bzw. $\hat{u}$ zu betrachten. Der klassische Weg ist die Berechnung als Fourier-Koeffizienten des zugrunde liegenden Basisraums $\{\phi_m\}$ mit Gewichtsfunktion h, also

$$\hat{f}_{m,l} := \frac{1}{||\phi_m||^2_{\mathrm{L}^2(\mathbb{T},h)}} \int_{\mathbb{T}} f_l(\boldsymbol{\xi},t)\phi_m(\boldsymbol{\xi})h(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}. \tag{2.13}$$

Dabei sind allerdings die f_l nicht global bekannt und man benötigt zur Auswertung des Integrals numerische Quadraturverfahren. Falls man ein Verfahren der Ordnung n erhalten möchte, muss es daher für $u \in \mathbb{P}_n(\mathbb{T},h)$ und $\mathbf{F} \in (\mathbb{P}_{n+1}(\mathbb{T},h))^2$ exakte Ergebnisse liefern. Daher müssen die Basispolynome ϕ_m aus $\mathbb{P}_{n+1}(\mathbb{T},h)$ stammen. Dies wiederum bedeutet, dass in der Gleichung (2.13) Polynome $2(n+1)$-ten Grades stehen mit der Gewichtsfunktion h. Davon ausgehend kann man das Quadraturverfahren wählen. Eine Übersicht über einige Quadraturverfahren findet sich in [15], [44] und [90].

Bemerkung. Bezüglich **diskreter** orthogonaler Polynome (Kapitel 5) ist dieser Ansatz von Interesse und wir nehmen dort nochmals Bezug darauf.

Wir verwenden schließlich nicht diesen **Projektionsansatz**, sondern betrachten die Gleichung (2.12), also folgendes Gleichungssystem

$$f_l(\mathbf{x}_j,t) = \mathcal{V}\left(\hat{f}_{m,l}(t)\right)_m \quad \forall l = 1,2,$$

wobei $\mathcal{V}$ die Vandermondsche Matrix zur Basis $\{\phi_m\}_m$ darstellt. $\mathcal{V}$ hat die Gestalt $\mathcal{V} = (\phi_m(T_i(\mathbf{x}_j)))_{m,j}$.
Das Gleichungssystem ist eindeutig lösbar für $\hat{f}_{m,l}(t)$, wenn $\mathcal{V}$ invertierbar ist. Dabei müssen die Anzahl der Interpolationspunkte $\boldsymbol{\xi}_j = T_i(\mathbf{x}_j)$ und die Anzahl der Basiselemente ϕ_m gleich sein. Die Lage der $\boldsymbol{\xi}_j$ soll weiterhin zu guten numerischen Eigenschaften von $\mathcal{V}$ führen, wie beispielsweise zu einer niedrigen Konditionszahl.
Wir verwenden hierbei erneut die Gauss-Lobatto-Punkte aus dem vorangegangen Abschnitt. Diese haben nicht nur den Vorteil einer guten Lebesgue-Konstanten, sondern ihre Konditionszahl ist verhältnismäßig klein für die Polynomfamilien, welche wir untersuchen werden. Vergleiche hierfür Tabelle 2.2.

n	1	2	3	4	5	6	7	8	9	10
κ_n	2.93	10.18	20.36	38.75	53.44	70.99	93.46	119	150	195
κ_n	9	24	270	2023	10^4	10^5	$7 \cdot 10^5$	$5 \cdot 10^6$	$3 \cdot 10^7$	$2 \cdot 10^9$

Tabelle 2.2: *Konditionszahlen der Vandermondschen Matrix bezüglich der Spaltensummmennorm in der unteren Zeile für die Lagrange-Polynome und im oberen Fall die PKD-Polynome, welche ein Spezialfall der Polynome aus Kapitel 3 sind. Die Tabelle stammt aus* [86, S.32 und S.35]

Konservativität und Flussberechnung an den Dreiecksrändern

Nachdem wir die grundlegende Berechnung erklärt haben, kommen wir auf die Konservativität zu sprechen und damit zwangsläufig auf die Flussberechnung an den Rändern der Dreiecke. Dabei betrachten wir ausschließlich hyperbolische Erhaltungsgleichungen. Diese zeichnen sich dadurch aus, dass bestimmte Größen, wie beispielsweise Masse und Impuls, erhalten bleiben. Das verwendete numerische Verfahren sollte ebenfalls diese Eigenschaften nicht verletzen und das hier vorgestellte SD-Verfahren erfüllt diese Bedingung unter gewissen Voraussetzungen. Dabei sei nochmals erwähnt, dass das SD-Verfahren den ganzen Fluss in der semidiskreten Gleichung benötigt. Die Flussberechnung im Inneren erfolgt mit Hilfe der Flussfunktionen aus den Werten von $\mathbf{u}$ in den Flusspunkten. Jedoch müssen die Flüsse am Rand jeder Zelle gesondert betrachtet und gegebenenfalls verändert werden, da sie die Kopplung zwischen zwei Zellen vornehmen müssen und die globale Konservativität nicht verletzt sein darf. Daher dürfen sich die Flüsse in Normalenrichtung zwischen zwei benachbarten Zellen nicht unterscheiden, wobei in der Implementierung selbst zwischen Kantenpunkten $\mathbf{x}_l$ und Eckpunkten $\mathbf{x}_e$ unterschieden wird und man sowohl in den Kantenpunkten als auch in den Eckpunkten einen *neuen* numerischen Fluss berechnet, so dass die Kopplung zwischen zwei Zellen erfüllt ist.
Ohne weiter ins Detail zu gehen sei dazu angemerkt, dass zur Berechnung der neuen Flussfunktion die Normalenkomponenten zwischen den Zellen durch **Riemann-Löser** bestimmt wurde. Eine detaillierte Erklärung der Implementierung und der Berechnung der Flussfunktion findet man in der Doktorarbeit von Wirz [86]. Allgemein sollte jedoch noch festgehalten werden, dass die Wahl einer geeigneten numerischen Flussfunktion ein eigenständiges Forschungsgebiet darstellt und in der Literatur [78] ausführlich behandelt wird. Die numerische Flussfunktion ersetzt im Aktualisierungsschema die jeweiligen Werte an den Randpunkten und weiterhin wurden in [86, S.30f] noch folgende zwei Lemmata bewiesen, welche die Konservativität der SD-Methode garantieren. Gerade Lemma 2.10 liefert noch einen weiteren Grund für die Wahl der Gauß-Lobatto-Punkte, da bei dieser Punktverteilung hinreichend viele Punkte auf den Dreieckskanten liegen und somit die globale Konservativität gewährleistet wird.

Lemma 2.9. *Seien* $N_u,\ N_F \in \mathbb{N},\ \{\hat{\boldsymbol{x}}_m | 1 \leq m \leq N_F\}$ *die Menge der Flusspunkte,* $\{\boldsymbol{x}_j | 1 \leq j \leq N_u\}$ *die Menge der Lösungspunkte und* L_m *die Lagrange-Polynome. Liegen die Flusspunkte auf Interpolationspunkten für ein Polynom* $(n+1)$*-ten Grades und*

die Lösungspunkte auf Quadraturpunkten für ein Quadraturverfahren n-ter Ordnung zur Berechnung des Volumenintegral, dann ist die Spektrale-Differenzen-Formulierung

$$u_t(\boldsymbol{x},t) = -\sum_{m=1}^{N_F} \begin{pmatrix} \hat{f}_{m,1}(t) \\ \hat{f}_{m,2}(t) \end{pmatrix} \cdot J_{T_i} \nabla_{\boldsymbol{\xi}} L_m(T_i(\boldsymbol{x}))$$

lokal konservativ.

Das Lemma ist bereits gültig, wenn anstelle der Lagrange-Polynome andere orthogonale Polynombasen verwendet werden. Dies wurde ebenfalls in [86, S.36] gezeigt. Für die globale Konservativität betrachtet man zwei an der Kante k benachbarte Zellen mit den jeweiligen Flüssen $\mathbf{F}_L$ in der linken bzw. $\mathbf{F}_R$ in der rechten Zelle, sowie einem Normalvektor $\mathbf{n}$ an der Kante k. Für die globale Konservativität muss der Fluss in Normalenrichtung zweier benachbarter Zellen gleich sein, das heißt

$$\int_k \mathbf{F}_L \cdot \mathrm{d}\mathbf{s} = \int_k \mathbf{F}_R \cdot \mathrm{d}\mathbf{s}.$$

Weiterhin ist $\mathbf{F} \in ((\mathbb{P}_{n+1}, h))^2$. Verwendet man eine exakte Quadratur bis zum Polynomgrad $n+1$ mit Gewichtung α_i und Stützstellen $\mathbf{x}_i$, ist dies äquivalent zu

$$\sum_i \alpha_i \mathbf{F}_L(\mathbf{x}_i) \cdot \mathbf{n} = \sum_i \alpha_i \mathbf{F}_R(\mathbf{x}_i) \cdot \mathbf{n}. \tag{2.14}$$

Ein Lemma aus [86, S.31] besagt folgendes:

Lemma 2.10. *Liegen auf jeder Kante eines Dreiecks mindestens n Flusspunkte auf Quadraturpunkten für das eindimensionale Volumenintegral entlang dieser Kante und erfüllen sie Gleichung* (2.14)*, dann ist die Spektrale-Differenzen-Methode global konservativ.*

Stabilität des SD-Verfahren

Neben der Konservativität und Konvergenz ist die Stabilität eines numerischen Verfahrens auch noch von Interesse. Dabei ist ein Verfahren **stabil**, wenn es gegenüber kleinen Störungen der Daten unempfindlich ist, das heißt vor allem, dass sich Rundungsfehler nicht aufaddieren und die numerische Lösung beschränkt bleibt, wenn die eigentliche Lösung beschränkt ist.
Das SD-Verfahren ist bereits bezüglich Stabilität von Jameson [38] und van den Abeele et al. [1] analysiert worden. Dabei beschränkt sich Jameson auf eine Raumdimension und untersucht die SD-Methode bezüglich einer Energienorm vom Sobolev-Typ, vergleiche dazu [38]. Van den Abeele et al. analysieren die SD-Methode im ein bzw. zwei dimensionalen Fall für verschiedene Flusspunktverteilungen und Ordnungen. Dabei können sie für Ordnung drei und vier keine numerische Stabilität im gleichseitigen Dreieck feststellen. Wirz kann das SD-Verfahren für PKD-Polynome auf das klassische SD-Verfahren

in Matrixschreibweise zurückführen. Es kann dann analog zu [1] vorgegangen werden, um zu zeigen, dass das SD-Verfahren auch für den Fall der PKD-Polynome Stabilitätsprobleme bei hoher Ordnungen bekommt und sicherlich auch im allgemeineren Fall der APK-Polynome.
Dadurch lässt sich auch numerisch keine geeignete CFL-Zahl bestimmen und die Wahl geeigneter Zeitschritte wird problematisch.
Wir beschränken uns erneut auf den Ansatz aus [86]. Die Stabilisierung soll durch den Einsatz von modalen Filtern realisiert werden, siehe Kapitel 4. Mit zunehmender Ordnung n wählen wir fallende Zeitschritte

$$\Delta t = \frac{C_{\text{fix}}}{(n+1)^2} \frac{h}{\lambda_m},$$

wobei C_{fix} ein fester Wert, h ein Längenmaß im Dreieck[7] (hier speziell der kürzeste Abstand vom Schwerpunkt zu einer Kante im Dreieck) und λ_m die maximale Ausbreitungsgeschwindigkeit in einer Zelle ist.

[7] Mögliche Längenmaße sind beispielsweise die kürzeste Kante, der Inkreisradius oder die kürzeste Höhe des Dreiecks.

3 Orthogonale Polynome und spektrale Konvergenz

Nachdem die mathematischen Grundlagen und das Spektrale-Differenzen-Verfahren erklärt wurden, beschäftigen wir uns in diesem Kapitel mit orthogonalen Polynomen und ihren Eigenschaften. Ziel dieses Abschnittes ist es, Satz 3.13 zu beweisen. Der Satz beinhaltet Resultate hinsichtlich des Approximationsverhaltens einer abgeschnittenen Fourier-Reihe klassischer orthogonaler Polynome auf Dreiecken. Entwickelt man eine Funktion $u \in C^\infty$, folgt aus dem Resultat insbesondere spektrale Konvergenz. Wir erweitern das Spektrale-Differenzen-Verfahren aus Kapitel 2 mit diesen Polynomfamilien. Bevor wir allerdings den Satz 3.13 beweisen, erklären wir den Begriff der **spektralen Konvergenz** und wiederholen einige elementare Eigenschaften klassischer orthogonaler Polynome. Wir beschäftigen uns zuerst mit den Jacobi-Polynomen, um anschließend die APK-Polynome einzuführen. Bei den APK-Polynomen handelt es sich um klassische orthogonale Polynome auf Dreiecken. Wir beweisen für die APK-Polynome eine Reihe von technischen Abschätzungen, die wir im Beweis von Satz 3.13 benötigen. Der Satz und sein Beweis bilden den Abschluss des Kapitels.

3.1 Orthogonale Polynome in einer Variablen

Grundlage vieler numerischer Methoden zur Approximation einer gesuchten Funktion u ist die Verwendung einer Folge von speziell gewählten Funktionen, vergleiche [11], [44] und [70]. Wie bereits in Kapitel 2 erwähnt, beschäftigen wir uns in dieser Arbeit mit spektralen Methoden. Man verwendet als Folge von Funktionen eine orthogonale Basis $\{\phi_k\}$ eines gewählten Funktionenraums. Der Approximationsfehler kann in der Norm des Raums gemessen werden. Die Approximation der Funktion u erfolgt durch die abgeschnittene Reihe

$$u \approx \sum_{k=1}^{N} \hat{u}_k \phi_k.$$

Das wohl bekannteste und am meisten untersuchte spektrale Verfahren ist die Fourier-Methode. Bei der Fourier-Methode wird eine periodische Funktion u durch ein trigo-

nometrisches Polynom[1] angenähert. Für eine 2π-periodische Funktion u lassen sich die Fourier-Koeffizienten $\hat{u}_k$ mit der Formel

$$\hat{u}_k = \frac{1}{2\pi} \int_0^{2\pi} u(x) e^{-ikx} \mathrm{d}x, \qquad \forall k \in \mathbb{N},$$

berechnen. Gerade die Approximation mit Hilfe von trigonometrischen Funktionen wurde bereits ausführlich analysiert und man findet hierzu viele Resultate in der Literatur [91].
Wenn die Funktion u unendlich oft differenzierbar ist und alle Ableitungen auch periodisch sind, dann konvergiert der Betrag der k-ten Koeffizienten $|\hat{u}_k|$ schneller gegen Null als jede Potenz k^{-j} mit $j \in \mathbb{N}$. Genau dieses charakteristische Verhalten bezeichnet man als **spektrale Konvergenz** bzw. **spektrale Genauigkeit**. Generell spricht man von einem spektralen Verfahren, wenn die Koeffizienten der abgeschnittenen Fourier-Reihenentwicklung dieses Verhalten aufweisen. Die Fourier-Methode wurde bereits zum Lösen von hyperbolischen Erhaltungsgleichungen genutzt und hinsichtlich Konvergenz, Stabilität und Existenz einer Lösung in [11], [74], [75] und weiteren Arbeiten analysiert.
Bei nichtperiodischen Fragestellung entwickelt man die Funktion u in ihre Fourier-Reihe bezüglich anderer orthogonaler Basisfunktionen $\{\phi_k\}$ als den trigonometrischen Funktionen. Jedoch kann für eine beliebige orthogonale Basis in der Regel nur **algebraische Konvergenz** der Koeffizienten $\hat{u}_k$ garantiert werden. Man spricht von algebraischer Konvergenz, wenn sich die Koeffizienten $\hat{u}_k$ wie $\mathcal{O}(k^{-j_1})$ für ein festes $j_1 \in \mathbb{N}$ verhalten. Hingegen weisen die Fourier-Koeffizienten immer spektrale Genauigkeit auf, wenn zur Entwicklung einer C^∞-Funktion u Eigenfunktionen ϕ_k eines singulären Sturm-Liouville-Problems verwendet werden oder u spezielle Randbedingungen erfüllt.
Auf dem Intervall $(-1, 1)$ ist das klassische **Sturm-Liouville-Problem** durch

$$\mathcal{L}\phi_k(x) = \lambda_k \omega(x) \phi_k(x), \qquad x \in (-1, 1), \tag{3.1}$$

mit $\mathcal{L}\phi_k := -(\tilde{p}(x)\phi_k'(x))' + \tilde{q}(x)\phi_k(x)$ und geeigneten Randbedingungen für ϕ_k gegeben. Die Koeffizientenfunktionen $\tilde{p}, \tilde{q}$ und ω erfüllen weiterhin die nachfolgenden Bedingungen:

- $\tilde{p}$ ist stetig differenzierbar, strikt positiv in $(-1, 1)$ und stetig am Rand $x = \pm 1$.
- $\tilde{q}$ ist stetig, nicht-negativ und beschränkt.
- Die Gewichtsfunktion ω ist stetig, nicht-negativ und integrierbar auf $(-1, 1)$.

Singulär wird von uns das Problem (3.1) genannt, wenn zusätzlich $\tilde{p}(\pm 1) = 0$ gilt. Wir halten uns an die Definition aus [11]. Die Autoren sprechen dabei ausschließlich von einem singulären Problem, wenn $\tilde{p}$ an beiden Randpunkten verschwindet. In der allgemeinen Sturm-Liouville-Theorie spricht man hingegen bereits von singulären Problemen,

[1] Unter einem trigonometrischen Polynom versteht man eine endliche, reelle Linearkombination trigonometrischer Funktionen. Man entwickelt somit die Funktion u in ihre Fourier-Reihe bezüglich trigonometrischer Funktionen.

wenn $\tilde{p}$ an einer der Grenzen verschwindet.
Klassische orthogonale Polynome wie Legendre-, Chebyshev- oder Jacobi-Polynome lösen genau solch ein singuläres Sturm-Liouville-Problem (3.1). In der Tabelle 3.1 sind die einzelnen Koeffizientenfunktionen $\tilde{p}$, $\tilde{q}$, die Gewichtsfunktion ω sowie die Eigenwerte λ_k aufgelistet.

	Legendre	Chebyshev	Jacobi
$\tilde{p}$	$1-x^2$	$(1-x^2)^{\frac{1}{2}}$	$(1-x)^{\alpha+1}(1+x)^{\beta+1}$
$\tilde{q}$	0	0	0
λ_k	$k(k+1)$	k^2	$k(k+\alpha+\beta+1)$
ω	1	$(1-x^2)^{-\frac{1}{2}}$	$(1-x)^{\alpha}(1+x)^{\beta}$

Tabelle 3.1: *Werte aus* [73]*, wobei* $\alpha,\ \beta > -1$ *gilt.*

Bei nichtperiodischen Funktionen u entwickelt man die Funktion u bezüglich dieser Basisfunktionen in eine Fourier-Reihe. In diesem Kontext wurden die klassischen orthogonalen Polynome bereits in verschiedenen Verfahren zum Lösen von hyperbolischen Erhaltungsgleichungen genutzt und die unterschiedlichen Verfahren hinsichtlich Konvergenz, Existenz einer Lösung und Stabilität in [11], [29], [30], [41], [52], [53] , [54] und weiteren Arbeiten untersucht. Zur Analyse verwendet man elementare Eigenschaften orthogonaler Polynome und Abschätzungen. Auch wir benötigen im späteren Verlauf der Arbeit noch einige dieser Resultate und wiederholen sie daher. Wir verweisen auf das Standardwerk [73] über orthogonale Polynome sowie die Formelsammlung [2].

Jacobi-Polynome

Bei numerischen Verfahren greift man oft auf Legendre- oder Chebyshev-Polynome zurück. Sie sind Spezialfälle der Jacobi-Polynome, die folgendermaßen definiert sind:

Definition 3.1. Es sei $n \in \mathbb{N}_0$ und $\alpha, \beta > -1$. Die **Jacobi-Polynome** $P_n^{\alpha,\beta}(x)$ sind für $x \in [-1,1]$ durch die Formel

$$P_n^{\alpha,\beta}(x) := \frac{\Gamma(\alpha+n+1)}{n!\Gamma(\alpha+\beta+n+1)} \sum_{i=0}^{n} \binom{n}{i} \frac{\Gamma(\alpha+\beta+n+i+1)}{\Gamma(\alpha+i+1)} \left(\frac{x-1}{2}\right)^i \tag{3.2}$$

oder alternativ mit Hilfe der **hypergeometrischen Funktion**[2]

$$P_n^{\alpha,\beta}(x) := \binom{n+\alpha}{n} {}_2F_1\left(-n, n+\alpha+\beta+1; \alpha+1; \frac{1-x}{2}\right)$$

gegeben.

[2] Auf diese Darstellung werden wir in Kapitel 5 explizit eingehen.

Man berechnet die Jacobi-Polynome häufig nicht durch die explizite Formel (3.2), sondern nutzt die Rodriguez-Formel. Es gilt

$$\begin{aligned} P_0^{\alpha,\beta}(x) &= 1, \\ P_1^{\alpha,\beta}(x) &= \frac{1}{2}((\alpha-\beta)+(\alpha+\beta+2)x), \\ a_{1,i}P_{i+1}^{\alpha,\beta}(x) &= a_{2,i}P_i^{\alpha,\beta}(x) - a_{3,k}P_{i-1}^{\alpha,\beta}(x), \end{aligned}$$

mit den Koeffizienten

$$\begin{aligned} a_{1,i} &= 2(i+1)(i+\alpha+\beta+1)(2i-\alpha+\beta), \\ a_{2,i} &= (2i+\alpha+\beta+1)(\alpha^2+\beta^2) - x\frac{\Gamma(2i+\alpha+\beta+3)}{\Gamma(2i+\alpha+\beta)}, \\ a_{3,i} &= 2(i+\alpha)(i+\beta)(2i+\alpha+\beta+2). \end{aligned}$$

Sei $\mathbb{P}_n([-1,1])$ der Raum der Polynome vom Höchstgrad n. Die Jacobi-Polynome vom Grad höchstens n bilden eine orthogonale Basis auf $\mathbb{P}_n([-1,1],\omega)$ mit der Gewichtsfunktion $\omega(x) = (1-x)^\alpha(1+x)^\beta$ und sind Lösung eines singulären Sturm-Liouville-Problems, wie man aus den nachfolgenden zwei Lemmata entnehmen kann.

Lemma 3.2. *Die Jacobi-Polynome vom Grad höchstens n mit festem α, β bilden eine orthogonale Basis von*

$$\mathbb{P}_n([-1,1],\omega) := \left\{ \sum_{i=0}^{n} a_i x^i \middle| a_i \in \mathbb{R},\ x \in [-1,1] \right\}$$

bezüglich des Skalarprodukts

$$\left(P_j^{\alpha,\beta}; P_k^{\alpha,\beta}\right)_{\mathrm{L}^2([-1,1],\omega)} := \int_{-1}^{1} \omega(x) P_j^{\alpha,\beta}(x) P_k^{\alpha,\beta}(x) \mathrm{d}x$$

mit der Gewichtsfunktion $\omega(x) = (1-x)^\alpha(1+x)^\beta$.
Weiterhin gilt

$$||P_k^{\alpha,\beta}||^2_{\mathrm{L}^2([-1,1],\omega)} := \left(P_k^{\alpha,\beta}; P_k^{\alpha,\beta}\right)_{\mathrm{L}^2([-1,1],\omega)} = \frac{2^{\alpha+\beta+1}\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{(2k+\alpha+\beta+1)k!\Gamma(k+\alpha+\beta+1)}.$$

Beweis. [24] □

Lemma 3.3. *Die Jacobi-Polynome $y = P_n^{\alpha,\beta}(x)$ erfüllen die lineare Differentialgleichung*

$$(1-x^2)y'' + [\beta-\alpha-(\alpha+\beta+2)x]y' + n(n+\alpha+\beta+1)y = 0,$$

was auch in der Form

$$\frac{\mathrm{d}}{\mathrm{d}x}\left\{(1-x)^{\alpha+1}(1+x)^{\beta+1}y'\right\} + n(n+\alpha+\beta+1)(1-x)^\alpha(1+x)^\beta y = 0$$

geschrieben werden kann.

Beweis. [73, S.61] □

Mit Lemma 3.2 folgt, dass die normierten Jacobi-Polynome $\frac{P_k^{\alpha,\beta}(x)}{||P_k^{\alpha,\beta}||_{\mathrm{L}^2([-1,1],\omega)}}$ eine orthonormale Basis von $\mathbb{P}_n([-1,1],\omega)$ bilden. Auch sie erfüllen die Differentialgleichung aus Lemma 3.3. Aus der Darstellung (3.2) berechnet man

$$\begin{aligned} P_n^{\alpha,\beta}(1) &= \frac{\Gamma(\alpha+n+1)}{n!\Gamma(\alpha+\beta+n+1)} \sum_{i=0}^{n} \binom{n}{i} \frac{\Gamma(\alpha+\beta+n+i+1)}{\Gamma(\alpha+i+1)} \left(\frac{0}{2}\right)^i \\ &= \frac{\Gamma(\alpha+1+n)}{n!\Gamma(\alpha+\beta+n+1)} \frac{\Gamma(\alpha+\beta+n+1)}{\Gamma(\alpha+1)} \overset{(8.1)}{=} \binom{n+\alpha}{n}. \end{aligned} \tag{3.3}$$

Im letzten Schritt haben wir die Darstellung der verallgemeinerten Binomialkoeffizienten durch die Gammafunktion benutzt. Durch lineare Transformation folgt die Identität [64, S.81f.]

$$P_n^{\alpha,\beta}(-x) = (-1)^n P_n^{\beta,\alpha}(x), \tag{3.4}$$

und wenn zusätzlich $q = \max\{\alpha,\beta\} \geq -\frac{1}{2}$ gilt, erhält man weiterhin

$$\max_{x\in[-1,1]} |P_n^{\alpha,\beta}(x)| = \binom{n+q}{n}. \tag{3.5}$$

Alle diese Eigenschaften findet man in jedem Standardwerk über orthogonale Polynome, vergleiche z.B. [73]. Das nachfolgende Resultat ist jedoch spezieller und stammt aus [32]. Wir sind nicht nur am Maximum der Jacobi-Polynome interessiert, sondern benötigen später Abschätzungen bezüglich jedes x-Wertes im Definitonsbereich.

Lemma 3.4. *Es sei* $x \in [-1,1]$, $\alpha,\beta \in \mathbb{R}_0^+$, $n \in \mathbb{N}_0$ *und*

$$g_n^{(\alpha,\beta)}(x) = \left(\frac{\Gamma(n+1)\Gamma(n+\alpha+\beta+1)}{\Gamma(n+\alpha+1)\Gamma(n+\beta+1))}\right)^{\frac{1}{2}} \left(\frac{1-x}{2}\right)^{\frac{\alpha}{2}} \left(\frac{1+x}{2}\right)^{\frac{\beta}{2}} P_n^{\alpha,\beta}(x).$$

Dann existiert eine positive Konstante $C < 12$, *so dass*

$$\left|\left(1-x^2\right)^{\frac{1}{4}} g_n^{(\alpha,\beta)}(x)\right| \leq \frac{C}{(2n+\alpha+\beta+1)^{\frac{1}{4}}} \tag{3.6}$$

für alle $x \in [-1,1]$ *gilt.*

Beweis. [32] □

Spektrale Konvergenz

Wir sind schon zu Beginn des Kapitels kurz auf spektrale Konvergenz eingegangen, was wir im folgenden Abschnitt fortsetzen wollen. Seien hierfür die ϕ_k normierte Basisfunktionen, die ein Sturm-Liouville-Problem (3.1) lösen. Nutzt man (3.1) und die

Selbstadjungiertheit des Differentialoperators, so kann man zeigen, dass für die Fourier-Koeffizienten einer Funktion u gilt:

$$\hat{u}_k = \left(\frac{1}{\lambda_k}\right) \int_{-1}^{1} \omega(x)(\mathcal{L}u)(x)\phi_k(x)\,\mathrm{d}x - \frac{1}{\lambda_k}\left[\tilde{p}(x)(\phi_k'(x)u(x) - \phi_k(x)u'(x)\right]_{-1}^{1},$$

vergleiche [11, S.283]. Die Funktion u muss dabei die Voraussetzung $u_{(1)} := \frac{1}{\omega}(\mathcal{L}u) \in \mathrm{L}^2((-1,1),\omega)$ erfüllen.
Sei weiterhin die rekursive Folge $u_{(i)} := \frac{1}{\omega}\mathcal{L}u_{(i-1)} \in \mathrm{L}^2((-1,1),\omega) \cap C^1([-1,1])$ für alle $i = 0,\ldots,m$ gegeben. Mit Hilfe der Gleichung (3.1) und der Selbstadjungiertheit des Differentialoperators $\mathcal{L}$ kann man für die Koeffizienten

$$\hat{u}_k = \left(\frac{1}{\lambda_k}\right)^m \int_{-1}^{1} \omega(x)\left(\mathcal{L}u_{(m-1)}\right)(x)\phi_k(x)\,\mathrm{d}x - \sum_{i=0}^{m-1}\left[\frac{\tilde{p}(x)}{\lambda_k^{i+1}}(\phi_k'(x)u_{(i)}(x) - \phi_k(x)u_{(i)}'(x))\right]_{-1}^{1} \tag{3.7}$$

zeigen.
Sind die ϕ_k Lösungen eines **regulären** Sturm-Liouville-Problems, so weisen die Eigenwerte ein Verhalten von $\lambda_k = \mathcal{O}(k^2)$ auf, siehe [11, S.282]. Will man für eine Funktion $u \in C^\infty([-1,1])$ spektrale Genauigkeit erreichen, so muss in der Gleichung (3.7)

$$\left[\tilde{p}(x)(\phi_k'(x)u_{(i)}(x) - \phi_k(x)u_{(i)}')(x)\right]_{-1}^{1} \equiv 0$$

für alle $i \in \mathbb{N}_0$ gelten. Man kann das garantieren, solange die Randbedingungen

$$\begin{aligned} \varsigma_1 u_{(i)}'(-1) + \chi_1 u_{(i)}(-1) &= 0, & \varsigma_1^2 + \chi_1^2 &\neq 0, \\ \varsigma_2 u_{(i)}'(1) + \chi_2 u_{(i)}(1) &= 0, & \varsigma_2^2 + \chi_2^2 &\neq 0, \end{aligned}$$

für geeignete $\varsigma_1, \varsigma_2, \chi_1, \chi_2 \in \mathbb{R}$ immer erfüllt sind. Für die Koeffizienten folgt schließlich, dass

$$\hat{u}_k = \left(\frac{1}{\lambda_k}\right)^m \int_{-1}^{1} \omega(x)\left(\mathcal{L}u_{(m-1)}\right)(x)\phi_k(x)\,\mathrm{d}x$$

gilt. Das Integral ist für alle $u_{(m)} := \frac{1}{\omega}\mathcal{L}u_{(m-1)} \in \mathrm{L}^2((-1,1),\omega)$ beschränkt. Es folgt mit der Schwarz'schen Ungleichung

$$|\hat{u}_k| \leq \frac{c_1}{k^{2m}} ||u_{(m)}||_{\mathrm{L}^2((-1,1),\omega)}$$

mit einer Konstanten $c_1 \in \mathbb{R}^+$ für alle $m \in \mathbb{N}$. Die Koeffizienten streben schneller gegen Null als jede Potenz $k^{-\tau}$ für ein fest gewähltes $\tau \in \mathbb{N}$. Es folgt spektrale Konvergenz.
Sind hingegen die Basisfunktionen ϕ_k Lösungen eines singulären Sturm-Liouville-Problems, dann ist dementsprechend $\tilde{p}(\pm 1) = 0$ für alle $i \in \mathbb{N}_0$ in (3.7) und die hintere Summe in (3.7) verschwindet. Für die Koeffizienten erhält man ebenfalls

$$\hat{u}_k = \left(\frac{1}{\lambda_k}\right)^m \int_{-1}^{1} \omega(x)\left(\mathcal{L}u_{(m-1)}\right)(x)\phi_k(x)\,\mathrm{d}x.$$

Wir betrachten die normierten Legendre-, die normierten Chebyshev- oder die normierten Jacobi-Polynome, deren Eigenwerte ebenfalls ein Verhalten von $\mathcal{O}(k^2)$ aufweisen. Es ergibt sich die Abschätzung

$$|\hat{u}_k| \leq \frac{c_2}{k^{2m}} ||u_{(m)}||_{\mathrm{L}^2((-1,1),\omega)}$$

mit einer Konstanten $c_2 \in \mathbb{R}^+$ für alle $m \in \mathbb{N}$ und insgesamt spektrale Konvergenz.
Tatsächlich handelt es sich bei den Jacobi-Polynomen um die einzigen klassischen Polynome, die ein singuläres Sturm-Liouville-Problem auf $(-1,1)$ lösen. Eine ausführlichere Betrachtung der Approximationsresultate findet man in [11, Kapitel 2 und 9].

3.2 APK-Polynome und ihre Eigenschaften

Kommen wir zu den Familien von orthogonalen Polynomen, für die wir spektrale Konvergenz am Ende dieses Abschnittes beweisen. Im ersten Abschnitt haben wir bereits festgestellt, dass man unter bestimmten Voraussetzungen an die Funktionen ϕ_k spektrale Konvergenz garantieren kann, und die Jacobi-Polynome erfüllen diese Voraussetzungen. Allerdings haben wir die Untersuchung ausschließlich in einer Variablen durchgeführt. Bei mehrdimensionalen Problemen verwendet man Tensorprodukte klassischer orthogonaler Polynome, wie beispielsweise Legendre-Legendre-Polynome oder Legendre-Chebyshev-Polynome auf Rechtecksgebieten.
Einen anderen Ansatz nutzt Dubiner in [17]. Er transformiert ein Tensorprodukt von Jacobi-Polynomen vom Rechteck $[-1,1]^2$ auf das Dreieck $\widetilde{\mathbb{T}} := \{(x,y) \in \mathbb{R}^2 | x \geq -1, y \geq -1, x+y \leq 0\}$, indem er die obere Kante $(-1,1)-(1,1)$ auf den Punkt $(-1,1)$ zusammenzieht. Dabei wird das Dreieck als Quadrat mit zwei identischen Ecken aufgefasst. Die Abbildung $\psi^{-1} : [-1,1]^2 \to \widetilde{\mathbb{T}}$ beschreibt somit die "Kollabierung" des Quadrats auf das Dreieck und ist gegeben durch

$$\begin{aligned} \psi^{-1} : [-1,1]^2 &\longrightarrow \widetilde{\mathbb{T}} \\ \begin{pmatrix} x \\ y \end{pmatrix} &\longmapsto \begin{pmatrix} \frac{(1+x)(1-y)}{2} - 1 \\ y \end{pmatrix} = \begin{pmatrix} r \\ s \end{pmatrix}. \end{aligned}$$

Hingegen bildet die Funktion

$$\begin{aligned} \psi : \widetilde{\mathbb{T}} \backslash \{-1,1\} &\longrightarrow [-1,1] \times [-1,1) \\ \begin{pmatrix} r \\ s \end{pmatrix} &\longmapsto \begin{pmatrix} \frac{2(1+r)}{1-s} - 1 \\ s \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \end{aligned}$$

das Dreieck auf das Quadrat ab, wobei die Ecke $(-1,1)$ nicht im Definitionsbereich liegt, vergleiche Abbildung 3.1. Die dadurch entstandene polynomiale Basis bezeichnet Dubiner selbst als **warped product** und nutzt sie bei der numerischen Untersuchung der Navier-Stokes-Gleichung [17]. In [63] wird spektrale Konvergenz für diese Polynomfamilie

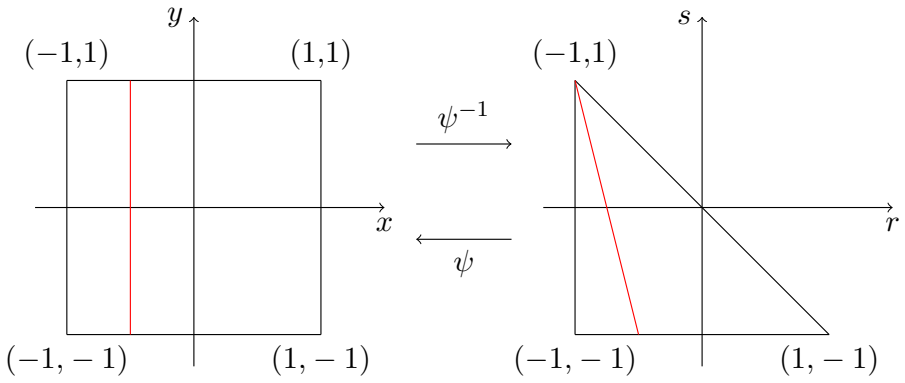


Abbildung 3.1: *Transformation ψ und ψ^{-1}*

bewiesen und [86] verwendet sie im Spektrale-Differenzen-Verfahren. Jedoch sind die von Dubiner verwendeten Polynome bereits lange bekannt und ein Spezialfall klassischer orthogonaler Polynome auf Dreiecken, die von Proriol zuerst in [66] beschrieben wurden. Diese allgemeinen Polynomfamilien führen wir nun ein. Allerdings werden wir einige Restriktionen vornehmen, die im späteren Verlauf für Rechnungen und Abschätzungen noch von Vorteil sein werden.
Sei dazu $\mathbb{T}$ das Einheitsdreieck, $h(x,y) := x^{\alpha-1}y^{\beta-1}(1-x-y)^{\gamma-\alpha-\beta}$ $(\alpha,\beta,\gamma \in \mathbb{N}, \gamma > \alpha+\beta-1)$ die Gewichtsfunktion[3]. Aus Gründen der Übersichtlichkeit werden wir zukünftig die Notationen $p := \gamma-\alpha-\beta$ und $a_l := p+\beta+2l$ verwenden.

Definition 3.5. Die Polynome $A_{m,l}(x,y),\ m,l \in \mathbb{N}_0$, definiert durch

$$A_{m,l}(x,y) := P_m^{\alpha-1,a_l}(1-2x)P_l^{p,\beta-1}\left(\frac{2y}{1-x}-1\right)(1-x)^l \tag{3.8}$$

auf $\mathbb{T}$ nennt man **Appell-Proriol-Koornwinder-Polynome** (kurz: APK-Polynome).

Bemerkung. Verwendet man anstelle des Einheitsdreiecks $\mathbb{T}$ das Dreieck $\{(x,y) \in \mathbb{R}^2 | x \geq -1, y \geq -1, x+y \leq 0\}$, so transformieren sich die APK-Polynome zu

$$\widetilde{A}_{m,l}(x,y) = P_m^{\alpha-1,2l+\gamma-\alpha}(-x)P_l^{\gamma-\alpha-\beta,\beta-1}\left(\frac{2(y+1)}{1-x}-1\right)\left(\frac{1-x}{2}\right)^l.$$

Für $\alpha=\beta=1$ und $\gamma=2$ erhält man die **PKD-Polynome**[4] vom Grad $N := m+l$.

Wir zeigen, dass die Polynome (3.8) ein System von klassischen orthogonalen Polynomen auf $\mathbb{T}$ bezüglich der Gewichtsfunktion $h(x,y)$ bilden. Die Namen Appell, Proriol und Koornwinder spiegeln die geschichtliche Entwicklung in diesem Forschungsgebiet

[3] Für den allgemeinen Fall mit $\alpha,\beta,\gamma \in \mathbb{R}^+$ und $h(x,y) := x^{\alpha-1}y^{\beta-1}(1-x-y)^{\gamma-1}$, siehe [19], [46] und [72].

[4] Dabei ist allerdings noch zu beachten, dass hier die Singularität im Eckpunkt $(1,-1)$ liegt. Durch eine Transformation kommt man schließlich zum tatsächlichen Dubiner-Fall.

wider. Appell hat Ende des 19. Jahrhunderts ein biorthogonales Polynomsystem auf dem Einheitsdreieck konstruiert [72, S.79] . Dabei können die Appell-Polynome als Verallgemeinerung der Jacobi-Polynome in zwei Variablen angesehen werden. So stehen sie auch nur orthogonal zu jedem Polynom kleineren Grades, jedoch nicht zu Polynomen gleichen Grades.
Proriol hat in seiner Arbeit [66] von 1954 die klassischen orthogonalen Polynome auf Dreiecken definiert. Diese Polynome erfüllen die Orthogonalitätsbeziehung im klassischen Sinne, das bedeutet unterschiedliche Polynome gleichen Grades stehen auch orthogonal zueinander.
Koornwinder hat schließlich einen äußerst wichtigen Übersichtsartikel [46] über die Charakterisierung von klassischen orthogonalen Polynomen auf verschiedenen Gebieten geschrieben. Seine Arbeiten haben bis heute einen enormen Einfluss im Forschungsgebiet der orthogonalen Polynome. Des Weiteren hat sich der Name des Spezialfalls der Proriol-Koornwinder-Dubiner-Polynome (PKD-Polynome) in der Literatur durchgesetzt, daher sollte man aus Gründen der Vollständigkeit auch Koornwinders Namen im allgemeineren Fall mit aufführen. In der Literatur wird auch von klassischen orthogonalen Polynomen auf Dreiecken oder von Jacobi-Polynomen auf Dreiecken gesprochen, wenn die APK-Polynome gemeint sind. Die von uns verwendete Definition 3.5 der APK-Polynome stammt aus [19] und [72].
Die ersten sechs APK-Polynome sind

$$
\begin{aligned}
A_{0,0}(x,y) &= 1,\\
A_{1,0}(x,y) &= \alpha - (1+\gamma)x,\\
A_{0,1}(x,y) &= \beta(-1+x) + (1-\alpha+\gamma)y,\\
A_{2,0}(x,y) &= \frac{x^2(\gamma+2)(\gamma+3) - 2x(\gamma+2)(\alpha+1) + (\alpha+1)\alpha}{2},\\
A_{0,2}(x,y) &= \frac{(1-x)^2\beta(\beta+1) - 2y(1-x)(\beta+1)(\gamma-\alpha+2) + y^2(\gamma-\alpha+2)(\gamma-\alpha+3)}{2},\\
A_{1,1}(x,y) &= (\alpha - (3+\gamma)x)(-\beta + \beta x + (1-\alpha+\gamma)y).
\end{aligned}
$$

Zur Vorbereitung des Beweises der spektralen Konvergenz definieren wir noch die Funktionenräume und Normen, auf denen wir arbeiten. Anschließend fassen wir bereits bekannte Eigenschaften zusammen und beweisen noch einige neue Abschätzungen.

Definition 3.6. Sei $h(x,y)$ eine Gewichtsfunktion auf dem Dreieck $\mathbb{T}$. Mit $\mathrm{L}^2(\mathbb{T},h)$ wird der Hilbertraum mit dem Skalarprodukt

$$(u;v)_{\mathrm{L}^2(\mathbb{T},h)} := \int_{\mathbb{T}} h(x,y)u(x,y)v(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

bezeichnet. Die durch das Innenprodukt induzierte Norm ist

$$||u||_{\mathrm{L}^2(\mathbb{T},h)} := \left(\int_{\mathbb{T}} h(x,y)|u(x,y)|^2\,\mathrm{d}x\,\mathrm{d}y\right)^{\frac{1}{2}}.$$

Wir bezeichnen mit $H^m(\mathbb{T}, h)$ den **gewichteten Sobolev-Raum**, definiert durch

$$H^m(\mathbb{T}, h) := \{v \in \mathrm{L}^2(\mathbb{T}, h)\text{: für jeden nichtnegativen Multiindex } \sigma \text{ mit } |\sigma| \leq m, \text{ gehören die distributionelle Ableitungen } D^\sigma v \text{ zu } \mathrm{L}^2(\mathbb{T}, h)\}.$$

Er besitzt als natürliche Norm

$$||v||_{H^m(\mathbb{T},h)} := \left(\sum_{|\sigma| \leq m} ||D^\sigma v||^2_{\mathrm{L}^2(\mathbb{T},h)} \right)^{\frac{1}{2}}.$$

Die APK-Polynome erfüllen die Orthogonalitätsbeziehung bezüglich der Gewichtsfunktion $h(x, y)$.

Lemma 3.7. *Die APK-Polynome $A_{m,l}$ sind orthogonal auf dem Dreieck $\mathbb{T}$ bezüglich des Innenprodukts*

$$(A_{m,l}; A_{n,s})_{\mathrm{L}^2(\mathbb{T},h)} = \delta_{m,n}\delta_{l,s} \frac{1}{(2l + \gamma - \alpha)(2(m + l) + \gamma)\kappa^2_{l,m}},$$

wobei $\kappa_{l,m} := \sqrt{\frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p (m)_\alpha (m+a_l)}}$ *und*

$$(\xi)_0 := 1, \quad (\xi)_j := \prod_{i=1}^{j} (\xi + i - 1) \text{ für } j \in \mathbb{N}$$

das Pochhammer-Symbol darstellt.

Beweis. Es sei

$$K(x, y) := \int_0^{1-x} \left(P_l^{p,\beta-1} \left(\frac{2y}{1-x} - 1 \right) \right) P_l^{p,\beta-1} \left(\frac{2y}{1-x} - 1 \right) y^{\beta-1} (1 - x - y)^p \, \mathrm{d}y$$

und damit gilt:

$$(A_{m,l}; A_{n,s})_{\mathrm{L}^2(\mathbb{T},h)} = \int_0^1 P_m^{\alpha-1,a_l}(1 - 2x) P_n^{\alpha-1,a_s}(1 - 2x)(1 - x)^l (1 - x)^s x^{\alpha-1} K(x, y) \, \mathrm{d}x.$$

Durch die Indextransformation

$$t = \frac{2y}{1-x} - 1, \qquad (1 - x - y) = \frac{(1-x)(1-t)}{2}, \qquad y = \frac{(1+t)(1-x)}{2},$$

erhalten wir für $K(x,y)$

$$\begin{aligned}
K(x,y) &= \int_0^{1-x} P_l^{p,\beta-1}\left(\frac{2y}{1-x}-1\right) P_s^{p,\beta-1}\left(\frac{2y}{1-x}-1\right) y^{\beta-1}(1-x-y)^p \,\mathrm{d}y \\
&= \int_{-1}^{1} P_l^{p,\beta-1}(t) P_s^{p,\beta-1}(t) \left(\frac{(1+t)(1-x)}{2}\right)^{\beta-1} \left(\frac{(1-x)(1-t)}{2}\right)^{p} \left(\frac{1-x}{2}\right) \mathrm{d}t \\
&= \left(\frac{1-x}{2}\right)^{\beta+p} \int_{-1}^{1} P_l^{p,\beta-1}(t) P_s^{p,\beta-1}(t)(1+t)^{\beta-1}(1-t)^p \,\mathrm{d}t.
\end{aligned}$$

Aus der Orthogonalität der Jacobi-Polynome (Lemma 3.2) folgt für das Integral

$$\int_{-1}^{1} P_l^{p,\beta-1}(t) P_s^{p,\beta-1}(t)(1+t)^{\beta-1}(1-t)^p \mathrm{d}t = \frac{\delta_{l,s} 2^{p+\beta}}{(2l+p+\beta)} \frac{\Gamma(l+p+1)\Gamma(l+\beta)}{\Gamma(l+1)\Gamma(l+p+\beta)}.$$

Einzig für $l=s$ ist das Integral nicht Null. Es ergibt sich durch elementare Rechnung

$$\begin{aligned}
(A_{m,l}; A_{n,s})_{\mathrm{L}^2(\mathbb{T},h)} =& \frac{\Gamma(l+p+1)\Gamma(l+\beta)}{\Gamma(l+1)\Gamma(l+p+\beta)(2l+p+\beta)} \\
& \cdot \int_0^1 P_m^{\alpha-1,a_l}(1-2x) P_n^{\alpha-1,a_l}(1-2x)(1-x)^{2l+\beta+p} x^{\alpha-1} \,\mathrm{d}x.
\end{aligned}$$

Durch die Indextransformation

$$2x-1=t, \qquad 1-x=\frac{1-t}{2}, \qquad x=\frac{1+t}{2},$$

erhält man für das Integral

$$\begin{aligned}
& \int_0^1 P_m^{\alpha-1,a_l}(1-2x) P_n^{\alpha-1,a_l}(1-2x)(1-x)^{2l+p+\beta} x^{\alpha-1} \,\mathrm{d}x \\
=& \int_{-1}^{1} P_m^{\alpha-1,a_l}(-t) P_n^{\alpha-1,a_l}(-t) \left(\frac{1-t}{2}\right)^{2l+p+\beta} \left(\frac{1+t}{2}\right)^{\alpha-1} \frac{1}{2} \,\mathrm{d}t \\
\overset{(3.4)}{=}& \left(\frac{1}{2}\right)^{2l+p+\beta+\alpha} (-1)^{m+n} \int_{-1}^{1} P_m^{a_l,\alpha-1}(t) P_n^{a_l,\alpha-1}(t) \,(1-t)^{2l+p+\beta} \,(1+t)^{\alpha-1} \,\mathrm{d}t.
\end{aligned}$$

Aus Lemma 3.2 folgt hierin

$$\int_{-1}^{1} P_m^{a_l,\alpha-1}(t) P_n^{a_l,\alpha-1}(t) \,(1-t)^{a_l} \,(1+t)^{\alpha-1} \,\mathrm{d}t = \frac{\delta_{m,n} 2^{\alpha+a_l} \Gamma(m+\alpha)\Gamma(m+a_l+1)}{\Gamma(m+1)\Gamma(m+a_l+\alpha)(2m+a_l+\alpha)}.$$

Mit diesem Resultat erhält man insgesamt für das Skalarprodukt der APK-Polynome

$$\begin{aligned}
&(A_{m,l};A_{n,s})_{\mathrm{L}^2(\mathbb{T},h)}\\
=&\delta_{m,n}\delta_{l,s}\frac{\Gamma(l+p+1)\Gamma(l+\beta)}{\Gamma(l+1)\Gamma(l+p+\beta)(2l+p+\beta)2^{a_l+\alpha}}\frac{2^{\alpha+a_l}\Gamma(m+\alpha)\Gamma(m+a_l+1)}{\Gamma(m+1)\Gamma(m+a_l+\alpha)(2m+a_l+\alpha)}\\
=&\delta_{m,n}\delta_{l,s}\frac{\Gamma(l+p+1)\Gamma(l+\beta)\Gamma(m+\alpha)\Gamma(m+a_l+1)}{\Gamma(l+p+\beta)\Gamma(l+1)(2l+\gamma-\alpha)\Gamma(m+1)\Gamma(m+a_l+\alpha)(2(m+l)+\gamma)}\\
=&\delta_{m,n}\delta_{l,s}\frac{\Gamma(l+p+1)\Gamma(l+\beta)\Gamma(m+\alpha)\Gamma(m+2l+\gamma-\alpha+1)}{(2l+\gamma-\alpha)(2(m+l)+\gamma)\Gamma(l+\gamma-\alpha)\Gamma(l+1)\Gamma(m+1)\Gamma(m+2l+\gamma)}\\
=&\delta_{m,n}\delta_{l,s}\frac{1}{(2l+\gamma-\alpha)(2(m+l)+\gamma)}\frac{\Gamma(l+p+1)}{\Gamma(l+1)}\frac{\Gamma(l+\beta)}{\Gamma(l+p+\beta)}\frac{\Gamma(m+\alpha)}{\Gamma(m+1)}\frac{\Gamma(m+a_l+1)}{\Gamma(m+a_l+\alpha)}.
\end{aligned}$$

Unter Verwendung der Rekursionsformel $\Gamma(k+p+1)=(k+p)\Gamma(k+p)$ und der Darstellung der Gammafunktion mit Hilfe der Pochhammer-Symbole

$$(x)_0=1 \quad \text{und} \quad \frac{\Gamma(x+n)}{\Gamma(x)}=(x)_n=\prod_{i=1}^{n}(x+i-1) \text{ für } n\in\mathbb{N} \tag{3.9}$$

ergibt sich

$$(A_{m,l};A_{n,s})=\delta_{m,n}\delta_{l,s}\frac{1}{(2l+\gamma-\alpha)(2(m+l)+\gamma)}\frac{(l+1)_p(m)_\alpha(m+a_l)}{(l+\beta)_p m(m+a_l)_\alpha}.$$

Mit $\kappa_{l,m}:=\sqrt{\frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p(m)_\alpha(m+a_l)}}$ folgt schließlich

$$(A_{m,l};A_{n,s})_{\mathrm{L}^2(\mathbb{T},h)}=\delta_{m,n}\delta_{l,s}\frac{1}{(2l+\gamma-\alpha)(2(m+l)+\gamma)\kappa_{l,m}^2}.$$

□

Wie wir bereits im ersten Paragraphen dieses Kapitels gesehen haben, kann man im eindimensionalen Fall spektrale Konvergenz garantieren, wenn die Basisfunktionen ϕ_k Lösungen eines singulären Sturm-Liouville-Problems sind. Hingegen sind die APK-Polynome im allgemeinen Fall keine Eigenfunktionen eines singulären Sturm-Liouville-Operators. Sie erfüllen die Eigenwertgleichung

$$DA_{m,l}=\lambda_{m,l}A_{m,l} \tag{3.10}$$

bezüglich des Differentialoperators D, der gegeben ist durch

$$D:=(x^2-x)\frac{\partial^2}{\partial x^2}+2xy\frac{\partial^2}{\partial x\partial y}+(y^2-y)\frac{\partial^2}{\partial y^2},+[(\gamma+1)x-\alpha]\frac{\partial}{\partial x}+[(\gamma+1)y-\beta]\frac{\partial}{\partial y} \tag{3.11}$$

und der Eigenwerte $\lambda_{m,l}=(m+l)(m+l+\gamma)$, siehe [19, S.46]. Eine der wichtigsten Eigenschaften des Sturm-Liouville-Operators im Beweis der spektralen Konvergenz ist die

Selbstadjungiertheit des Differentialoperators. In unserem Fall ist, wie wir noch zeigen werden, der Operator D nur im Fall $\alpha = \beta = 1$ und $\gamma = 2$ selbstadjungiert, was genau dem PKD-Fall entspricht. Tatsächlich wird in [76] gezeigt, dass die PKD-Polynome $\Phi_{m,l}$ Lösungen eines Sturm-Liouville-Problems sind. Die zugehörige Gleichung lautet

$$\mathcal{L}_{PKD}\Phi_{m,l}(v,w) = \lambda_{m,l}\Phi_{m,l}(v,w), \quad \forall (v,w) \in \widetilde{\mathbb{T}}, \tag{3.12}$$

mit dem Differentialoperator

$$\begin{aligned}\mathcal{L}_{PKD} = &- \frac{\partial}{\partial v}\left((1+v)\left[(1-v)\frac{\partial}{\partial v} - (1+w)\frac{\partial}{\partial w}\right]\right) \\ &- \frac{\partial}{\partial w}\left((1+w)\left[(1-w)\frac{\partial}{\partial w} - (1+v)\frac{\partial}{\partial v}\right]\right),\end{aligned}$$

und den Eigenwerten $\lambda_{m,l} = (m+l)(m+l+2)$. In der Arbeit [63] wird schließlich bewiesen, dass es sich bei (3.12) um ein singuläres Problem handelt, der Operator $\mathcal{L}_{PKD}$ selbstadjungiert ist und die Koeffizienten sowie der Abschneidefehler spektral gegen Null konvergieren. Mit der Transformation

$$x = \frac{1+w}{2}, \qquad y = \frac{1+v}{2}$$

und für den Fall $\alpha = \beta = 1$, $\gamma = 2$ überträgt sich der Operator D auf $\mathcal{L}_{PKD}$. Es gilt

$$\begin{aligned}D =& \left[\left(\frac{1+w}{2}\right)^2 - \frac{1+w}{2}\right]4\frac{\partial^2}{\partial^2 w} + 2(1+w)(1+v)\frac{\partial^2}{\partial w \partial v} \\ &+ \left[\left(\frac{1+v}{2}\right)^2 - \frac{1+v}{2}\right]4\frac{\partial^2}{\partial^2 v} + (3w+1)\frac{\partial}{\partial w} + (3v+1)\frac{\partial}{\partial v} \\ =&(w^2-1)\frac{\partial^2}{\partial^2 w} + 2(1+w)(1+v)\frac{\partial^2}{\partial w \partial v} + (v^2-1)\frac{\partial^2}{\partial^2 v} + (1+3w)\frac{\partial}{\partial w} + (1+3v)\frac{\partial}{\partial v} \\ =&- \frac{\partial}{\partial v}\left((1+v)\left[(1-v)\frac{\partial}{\partial v} - (1+w)\frac{\partial}{\partial w}\right]\right) \\ &- \frac{\partial}{\partial w}\left((1+w)\left[(1-w)\frac{\partial}{\partial w} - (1+v)\frac{\partial}{\partial v}\right]\right) = \mathcal{L}_{PKD}.\end{aligned}$$

Obwohl der Operator D nicht selbstadjungiert ist, kann man für die APK-Reihenentwicklung spektrale Genauigkeit der Koeffizienten und des Abschneidefehlers beweisen. So ist zwar der Operator D nicht selbstadjungiert, aber im Inneren des Dreieckes **potentiell selbstadjungiert**, vergleiche [72, S.131].

Definition 3.8. Sei $G \subset \mathbb{R}^2$ ein Gebiet. Den Differentialoperator

$$\widetilde{D}u(x,y) = a\frac{\partial^2 u(x,y)}{\partial x^2} + 2b\frac{\partial^2 u(x,y)}{\partial x \partial y} + c\frac{\partial^2 u(x,y)}{\partial y^2} + d\frac{\partial u(x,y)}{\partial x} + g\frac{\partial u(x,y)}{\partial y} \tag{3.13}$$

mit $a-g \in C^1(G)$ nennt man **potentiell selbstadjungiert** auf dem Gebiet G, wenn eine zweimal stetig differenzierbare positive Funktion $k(x,y)$ existiert, so dass der Operator

$$k\widetilde{D}u(x,y) = (ka)\frac{\partial^2 u(x,y)}{\partial x^2} + 2(kb)\frac{\partial^2 u(x,y)}{\partial x \partial y} + (kc)\frac{\partial^2 u(x,y)}{\partial y^2} + (kd)\frac{\partial u(x,y)}{\partial x} + (kg)\frac{\partial u(x,y)}{\partial y}$$

selbstadjungiert ist.

Bemerkung. Im zweidimensionalen Fall hat ein selbstadjungierter Operator die allgemeine Form

$$\tilde{D}u = \frac{\partial}{\partial x}\left(a\frac{\partial u(x,y)}{\partial x} + b\frac{\partial u(x,y)}{\partial y}\right) + \frac{\partial}{\partial y}\left(b\frac{\partial u(x,y)}{\partial x} + c\frac{\partial u(x,y)}{\partial y}\right).$$

Der Operator $\tilde{D}$ ist genau dann selbstadjungiert, wenn

$$\begin{aligned} d &= \frac{\partial a}{\partial x} + \frac{\partial b}{\partial y}, \\ g &= \frac{\partial b}{\partial x} + \frac{\partial c}{\partial y}, \end{aligned} \tag{3.14}$$

gilt bzw. potentiell selbstadjungiert, wenn eine positive $C^2(G)$-Funktion $k(x,y)$ exisitert, so dass

$$\begin{aligned} kd &= \frac{\partial(ka)}{\partial x} + \frac{\partial(kb)}{\partial y}, \\ kg &= \frac{\partial(kb)}{\partial x} + \frac{\partial(kc)}{\partial y}, \end{aligned}$$

gilt.

Differenziert man diese Gleichungen und stellt sie um, so erhält man

$$\begin{aligned} a\frac{\partial k}{\partial x} + b\frac{\partial k}{\partial y} &= k\left(d - \frac{\partial a}{\partial x} - \frac{\partial b}{\partial y}\right), \\ b\frac{\partial k}{\partial x} + c\frac{\partial k}{\partial y} &= k\left(g - \frac{\partial b}{\partial x} - \frac{\partial c}{\partial y}\right). \end{aligned}$$

Mit $\varphi := d - \frac{\partial a}{\partial x} - \frac{\partial b}{\partial y}$ und $\Psi := g - \frac{\partial b}{\partial x} - \frac{\partial c}{\partial y}$ ergibt sich

$$a\frac{\partial k}{\partial x} + b\frac{\partial k}{\partial y} = k\varphi, \qquad b\frac{\partial k}{\partial x} + c\frac{\partial k}{\partial y} = k\Psi, \tag{3.15}$$

dabei sind die Funktionen φ und Ψ wohldefiniert. Wir fassen das System (3.15) als System von Differentialgleichungen zusammen. Weiterhin nehmen wir an, dass die Funktionen φ und Ψ im Gebiet nicht gleichzeitig Null sind, das heißt

$$\varphi^2(x,y) + \Psi^2(x,y) > 0 \quad \forall (x,y) \in G. \tag{3.16}$$

Es exisitert genau dann eine Lösung für (3.15), wenn

$$\theta = ac - b^2 \neq 0 \tag{3.17}$$

gilt. Ferner führen wir noch die Funktionen $\Upsilon := \varphi c - \Psi b$ und $O := a\Psi - b\varphi$ ein. Man erhält die Lösung des Systems (3.15) unter den Bedingungen (3.16) und (3.17) aus den Gleichungen

$$\frac{1}{k}\frac{\partial k}{\partial x} = \frac{\Upsilon}{\theta}, \qquad \frac{1}{k}\frac{\partial k}{\partial y} = \frac{O}{\theta}. \tag{3.18}$$

Diese Darstellung ermöglicht es ein hinreichendes Kriterium dafür anzugeben, dass der Differentialoperator $\widetilde{D}$ potentiell selbstadjungiert ist. In [72, S.133] findet man folgenden Satz.

Satz 3.9. *Der Differentialoperator* (3.13) *ist genau dann potentiell selbstadjungiert, wenn die Bedingung*

$$\frac{\partial}{\partial x}\left(\frac{O(x,y)}{\theta(x,y)}\right) = \frac{\partial}{\partial y}\left(\frac{\Upsilon(x,y)}{\theta(x,y)}\right) \tag{3.19}$$

erfüllt ist.

Beweis. Zunächst nehmen wir an, dass der Differentialoperator (3.13) potentiell selbstadjungiert ist. Nach Definition 3.8 existiert eine positive zweimal stetig differenzierbare Funktion $k(x,y)$ auf dem Gebiet G, die das System (3.15) löst. Differenziert man weiterhin die Gleichungen (3.18), so erhält man

$$\begin{aligned}
\frac{\partial}{\partial x}\left(\frac{O(x,y)}{\theta(x,y)}\right) &= -\frac{1}{k^2(x,y)}\frac{\partial k(x,y)}{\partial x}\frac{\partial k(x,y)}{\partial y} + \frac{1}{k(x,y)}\frac{\partial^2 k(x,y)}{\partial x \partial y},\\
\frac{\partial}{\partial y}\left(\frac{\Upsilon(x,y)}{\theta(x,y)}\right) &= -\frac{1}{k^2(x,y)}\frac{\partial k(x,y)}{\partial y}\frac{\partial k(x,y)}{\partial x} + \frac{1}{k(x,y)}\frac{\partial^2 k(x,y)}{\partial x \partial y}.
\end{aligned}$$

Das wiederum beweist (3.19).
Gelte nun (3.19) auf dem Gebiet G. Betrachtet man die Gleichungen (3.18), so ergibt sich

$$\ln k(x,y) = \int_{x_0}^{x} \frac{\Upsilon(x,y)}{\theta(x,y)}\mathrm{d}x + c_1(y), \tag{3.20}$$

$$\ln k(x,y) = \int_{y_0}^{y} \frac{O(x,y)}{\theta(x,y)}\mathrm{d}y + c_2(x). \tag{3.21}$$

Zur Berechnung von $c_1(y)$ differenzieren wir (3.20) nach y und nutzen die Beziehung (3.19). Man erhält

$$\frac{1}{k(x,y)}\frac{\partial k(x,y)}{\partial y} = \frac{O(x,y)}{\theta(x,y)} = \int_{x_0}^{x} \frac{\partial}{\partial y}\left(\frac{\Upsilon(x,y)}{\theta(x,y)}\right)\mathrm{d}x + c_1'(y) = \frac{O(x,y)}{\theta(x,y)} - \frac{O(x_0,y)}{\theta(x_0,y)} + c_1'$$

und damit

$$c_1(y) = \int_{y_0}^{y} \left[\frac{O(x_0, y)}{\theta(x_0, y)} \right] \mathrm{d}y + c_3$$

mit einer Konstanten $c_3 \in \mathbb{R}$. Auf analoge Art ergibt sich

$$\frac{1}{k(x,y)} \frac{\partial k(x,y)}{\partial x} = \frac{\Upsilon(x,y)}{\theta(x,y)} = \frac{\Upsilon(x,y)}{\theta(x,y)} - \frac{\Upsilon(x,y_0)}{\theta(x,y_0)} + c_2'(x),$$

$$c_2(x) = \int_{x_0}^{x} \left[\frac{\Upsilon(x, y_0)}{\theta(x, y_0)} \right] \mathrm{d}x + c_4$$

mit einer Konstanten $c_4 \in \mathbb{R}$. Setzt man $c_1(y)$ und $c_2(x)$ in die Gleichungen (3.20) und (3.21) ein, so erhält man als Resultat eine geeignete Funktion $k(x, y)$. Aus c_1 und Gleichung (3.20) folgt

$$k(x,y) = \exp\left\{ \int_{x_0}^{x} \frac{\Upsilon(x,y)}{\theta(x,y)} \mathrm{d}x + \int_{y_0}^{y} \frac{O(x_0,y)}{\theta(x_0,y)} \mathrm{d}y + c_3 \right\}.$$

□

Mit dem Satz 3.9 sind wir nun in der Lage, den folgenden Satz zu beweisen:

Satz 3.10. *Der Differentialoperator* (3.11) *der APK-Polynome ist für jede beliebige zulässige Parameterwahl α, β und γ auf dem Inneren des Dreiecks $\mathring{\mathbb{T}}$ potentiell selbstadjungiert.*

Beweis. Wir betrachten den Differentialoperator (3.11)

$$D := (x^2 - x)\frac{\partial^2}{\partial x^2} + 2xy\frac{\partial^2}{\partial x \partial y} + (y^2 - y)\frac{\partial^2}{\partial y^2} + [(\gamma+1)x - \alpha]\frac{\partial}{\partial x} + [(\gamma+1)y - \beta]\frac{\partial}{\partial y}.$$

Es gilt

$$a(x,y) = x^2 - x, \qquad b(x,y) = xy, \qquad c(x,y) = y^2 - y,$$
$$d(x,y) = (\gamma+1)x - \alpha, \quad g(x,y) = (\gamma+1)y - \beta$$

und für die Ableitungen

$$\frac{\partial a(x,y)}{\partial x} = 2x - 1, \qquad \frac{\partial b(x,y)}{\partial x} = y, \qquad \frac{\partial b(x,y)}{y} = x, \qquad \frac{\partial c(x,y)}{\partial y} = 2y - 1,$$
$$\frac{\partial a(x,y)}{\partial x} + \frac{\partial b(x,y)}{\partial y} = 3x - 1, \qquad \frac{\partial b(x,y)}{\partial x} + \frac{\partial c(x,y)}{\partial y} = 3y - 1.$$

Es folgt aus (3.14), dass der Operator D nur im Falle $\alpha = \beta = 1$ und $\gamma = 2$ selbstadjungiert ist, was wir bereits festgestellt hatten.

Nehmen wir nun an, dass D nicht selbstadjungiert ist. Es ergeben sich für die Funktionen φ, Ψ, θ, Υ und O die folgenden Darstellungen:

$$\begin{aligned}
\varphi(x,y) &= (\gamma-2)x-\alpha+1,\\
\Psi(x,y) &= (\gamma-2)y-\beta+1,\\
\theta(x,y) &= xy(1-x-y),\\
\Upsilon(x,y) &= (1-\alpha)y^2+(1-\gamma+\beta)xy-(1-\alpha)y,\\
O(x,y) &= (1-\beta)x^2+(1-\gamma+\alpha)xy-(1-\beta)x.
\end{aligned} \tag{3.22}$$

In $\mathring{\mathbb{T}}$ ist sowohl $\theta(x,y)\neq 0$ also auch $\Upsilon^2(x,y)+O^2(x,y)>0$. Somit müssen wir nur noch die Bedingung (3.19) aus Satz 3.9 überprüfen. Differenziert man beide Seiten von (3.19), so folgt mit der Kettenregel

$$\begin{aligned}
\frac{\partial}{\partial x}\left(\frac{O(x,y)}{\theta(x,y)}\right) &= \frac{1}{\theta(x,y)}\frac{\partial O(x,y)}{\partial x}-\frac{O(x,y)}{\theta^2(x,y)}\frac{\partial\theta(x,y)}{\partial x},\\
\frac{\partial}{\partial y}\left(\frac{\Upsilon(x,y)}{\theta(x,y)}\right) &= \frac{1}{\theta(x,y)}\frac{\partial \Upsilon(x,y)}{\partial y}-\frac{\Upsilon(x,y)}{\theta^2(x,y)}\frac{\partial\theta(x,y)}{\partial y},
\end{aligned}$$

und damit ist

$$\theta(x,y)\frac{\partial O(x,y)}{\partial x}-O(x,y)\frac{\partial\theta(x,y)}{\partial x}=\theta(x,y)\frac{\partial\Upsilon(x,y)}{\partial y}-\Upsilon(x,y)\frac{\partial\theta(x,y)}{\partial y}$$

zu verifizieren.
Nutzen wir die Darstellung (3.22), so ergibt sich

$$\begin{aligned}
&\theta(x,y)\frac{\partial O(x,y)}{\partial x}-O(x,y)\frac{\partial\theta(x,y)}{\partial x}\\
&\quad= xy(1-x-y)\left(2x(1-\beta)+(1-\gamma+\alpha)y-(1-\beta)\right)\\
&\qquad-\left((1-\beta)x^2+(1-\gamma+\alpha)xy-(1-\beta)x\right)\left(y(1-x-y)-xy\right)\\
&\quad= xy(1-x-y)\left(2x(1-\beta)+(1-\gamma+\alpha)y-(1-\beta)\right)\\
&\qquad-\left((1-\beta)x^2+(1-\gamma+\alpha)xy-(1-\beta)x\right)\left(y-2xy-y^2\right)\\
&\quad= xy(1-x-y)(2x(1-\beta)+(1-\gamma+\alpha)y-(1-\beta))\\
&\qquad-xy(1-2x-y)((1-\beta)x+(1-\gamma+\alpha)y(1-\beta))\\
&\quad= xy((1-x-y)2x(1-\beta)-(1-2x-y)(1-\beta)x-x(1-\gamma+\alpha)y\\
&\qquad+x(1-\beta)+2x(1-\gamma+\alpha)y-2x(1-\beta))\\
&\quad= xy\left((1-\beta)x(2-2x-2y-1+2x+y)+xy(1-\gamma+\alpha)-x(1-\beta)\right)\\
&\quad= xy\left((1-\beta)x(1-y)+xy(1-\gamma+\alpha)-x(1-\beta)\right)\\
&\quad= x^2y^2(\alpha+\beta-\gamma)
\end{aligned}$$

und

$$\begin{aligned}
&\theta(x,y)\frac{\partial\Upsilon(x,y)}{\partial y}-\Upsilon(x,y)\frac{\partial\theta(x,y)}{\partial y}\\
&\quad= xy(1-x-y)\left(2y(1-\alpha)+(1-\gamma+\beta)x-(1-\alpha)\right)\\
&\qquad-\left((1-\alpha)y^2+(1-\gamma+\beta)xy-(1-\beta)y\right)(x-2xy-x^2).
\end{aligned}$$

Vergleicht man diese Terme mit den Termen der Rechnung zuvor, speziell denjenigen nach dem zweiten Gleichheitszeichen, so erkennt man eine ähnliche Struktur. Mit analoger Rechnung folgt

$$\theta(x,y)\frac{\partial \Upsilon(x,y)}{\partial y} - \Upsilon(x,y)\frac{\partial \theta(x,y)}{\partial y} = x^2y^2(\alpha+\beta-\gamma).$$

Die hinreichende Bedingung (3.19) ist für jede zulässige Paramterwahl α, β und γ erfüllt und nach Satz 3.9 folgt, dass der Differentialoperator (3.11) auf $\mathring{\mathbb{T}}$ potentiell selbstadjungiert ist. □

Diese Eigenschaft wird der Schlüssel im späteren Beweis der spektralen Konvergenz sein. Jedoch benötigen wir noch eine Reihe an weiteren Abschätzungen für die APK-Polynome, welche wir formulieren und beweisen.

Lemma 3.11. *Die folgende Normabschätzung gilt für alle APK-Polynome*

$$\frac{1}{||A_{m,l}||_{\mathrm{L}^2(\mathbb{T},h)}} \leq 2(m+l+\gamma)\kappa_{l,m}, \tag{3.23}$$

mit $\kappa_{l,m}$ *aus* Lemma 3.7.

Beweis. Für das Skalarprodukt gilt nach Lemma 3.7

$$(A_{m,l}; A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)} = \frac{1}{(2l+\gamma-\alpha)(2(m+l)+\gamma)}\frac{(l+1)_p}{(l+\beta)_p}\frac{(m+a_l)(m)_\alpha}{m(m+a_l)_\alpha}$$

und daraus folgt schließlich

$$\begin{aligned}
\frac{1}{||A_{m,l}||_{\mathrm{L}^2(\mathbb{T},h)}} =& \sqrt{\frac{(2l+\gamma-\alpha)(2(m+l)+\gamma)}{1}\frac{m}{(m+a_l)}\frac{(l+\beta)_p(m+a_l)_\alpha}{(l+1)_p(m)_\alpha}} \\
\leq& \sqrt{\frac{(2m+2l+\gamma)(2m+2l+\gamma)(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p(m+a_l)(m)_\alpha}} \\
=& (2m+2l+\gamma)\kappa_{l,m} \\
\leq& 2(m+l+\gamma)\kappa_{l,m}.
\end{aligned}$$

□

Lemma 3.12. *Seien* $l, m \in \mathbb{N}_0$*. Für das Innere des Dreiecks* $\mathbb{T}$*,* $(x,y) \in \mathring{\mathbb{T}}$*, bekommt man folgende APK-Abschätzung*

$$|A_{m,l}(x,y)| \leq \frac{\tilde{E}(x,y)}{(2l+\beta+p)^{\frac{1}{4}}(2(m+l)+\gamma)^{\frac{1}{4}}\kappa_{l,m}}, \tag{3.24}$$

mit

$$\tilde{E}(x,y) = \frac{E}{2(1-x-y)^{\frac{p}{2}+\frac{1}{4}}y^{\frac{1}{4}+\frac{\beta-1}{2}}x^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}}}$$

und einer positiven Konstanten $E < 144$.
Auf der Kante $[0, y]$ mit $y \in [0, 1]$ erhält man

$$|A_{m,l}(0,y)| \leq \left(\binom{m+\alpha-1}{m} \cdot \max\left\{ \binom{l+p}{l}, \binom{l+\beta-1}{l} \right\} \right). \tag{3.25}$$

Auf den Kanten $[x, 1-x]$ und $[x, 0]$ gilt

$$\begin{aligned} |A_{m,l}(x,0)| \leq & \frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}\sqrt{2}} \\ & \cdot \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m} \right)^{\frac{1}{2}} \binom{l+\beta-1}{l} \end{aligned} \tag{3.26}$$

und

$$\begin{aligned} |A_{m,l}(x,1-x)| \leq & \frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}\sqrt{2}} \\ & \cdot \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m} \right)^{\frac{1}{2}} \binom{l+p}{l} \end{aligned} \tag{3.27}$$

für alle $x \in (0,1)$ mit einer positiven Konstanten $C < 12$.
$A_{m,l}(1,0)$ besitzt den Wert

$$|A_{m,l}(1,0)| = \begin{cases} \binom{m+\gamma-\alpha}{m}, & \text{wenn } l = 0, \\ 0, & \text{sonst .} \end{cases} \tag{3.28}$$

Beweis. Die Definition der APK-Polynome 3.5 liefert die Darstellung

$$A_{m,l}(x,y) = P_m^{\alpha-1,a_l}(1-2x) P_l^{p,\beta-1}\left(\frac{2y}{1-x} - 1\right)(1-x)^l$$

als Produkt zweier Jacobi-Polynome und einem Faktor $(1-x)^l$. Für die Abschätzung der Beträge (3.24)-(3.28) müssen wir daher hauptsächlich die Jacobi-Polynome untersuchen. Als Hilfsmittel nutzen wir die Relation (3.9) der Gammafunktion und die Abschätzung (3.6) aus Lemma 3.4, mit deren Hilfe wir die Jacobi-Polynome im Inneren des Intervalls punktweise abschätzen können. Wir beginnen unsere Analyse im Inneren des Dreiecks und nutzen Ungleichung (3.6), um die einzelnen Faktoren der APK-Polynome abzuschätzen. Man erhält

$$\begin{aligned} & |P_m^{\alpha-1,a_l}(1-2x)| \\ & \overset{(3.6)}{\leq} \frac{C}{(2m+a_l+\alpha)^{\frac{1}{4}}(1-(1-2x))^{\frac{1}{4}}(1+(1-2x))^{\frac{1}{4}}} \\ & \quad \cdot \frac{2^{\frac{\alpha-1+a_l}{2}}}{(1-(1-2x))^{\frac{\alpha-1}{2}}(1+(1-2x))^{\frac{a_l}{2}}} \left(\frac{\Gamma(m+\alpha)\Gamma(m+a_l+1)}{\Gamma(m+1)\Gamma(m+\alpha+a_l)} \right)^{\frac{1}{2}} \\ & \overset{(3.9)}{=} \frac{C \cdot 2^{\frac{\alpha-1+a_l}{2}}}{(2(m+l)+\gamma)^{\frac{1}{4}}(2x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(2(1-x))^{\frac{1}{4}+\frac{a_l}{2}}} \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m} \right)^{\frac{1}{2}}, \end{aligned}$$

bzw.

$$|P_m^{\alpha-1,a_l}(1-2x)| \leq \frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{a_l}{2}}\sqrt{2}} \left(\frac{(m)_\alpha(m+a_l)}{(m+a_l)_\alpha m}\right)^{\frac{1}{2}} \tag{3.29}$$

und

$$\begin{aligned}
&\left|P_l^{p,\beta-1}\left(\frac{2y}{1-x}-1\right)(1-x)^l\right| \\
\overset{(3.6)}{\leq}& \frac{C(1-x)^l}{\left[(2l+p+\beta)\left(1-\frac{2y}{1-x}+1\right)\left(1+\frac{2y}{1-x}-1\right)\right]^{\frac{1}{4}}} \\
&\cdot\left(\frac{\Gamma(l+p+1)\Gamma(l+\beta)2^{p+\beta-1}}{\Gamma(l+1)\Gamma(l+p+\beta)\left(2-\frac{2y}{1-x}\right)^p\left(\frac{2y}{1-x}\right)^{\beta-1}}\right)^{\frac{1}{2}} \\
\overset{(3.9)}{=}& \frac{C(1-x)^l 2^{\frac{p+\beta-1}{2}}}{\left[(2l+p+\beta)\left(2\left(1-\frac{y}{1-x}\right)\left(\frac{2y}{1-x}\right)\right)\right]^{\frac{1}{4}}\left(\frac{2((1-x)-y)}{1-x}\right)^{\frac{p}{2}}\left(\frac{2y}{1-x}\right)^{\frac{\beta-1}{2}}}\left(\frac{(l+1)_p}{(l+\beta)_p}\right)^{\frac{1}{2}} \\
=& \frac{C(1-x)^{l+\frac{p+\beta-1}{2}+\frac{1}{2}}2^{\frac{p+\beta-1}{2}}}{\sqrt{2}(2l+\beta+p)^{\frac{1}{4}}2^{\frac{p+\beta-1}{2}}(1-x-y)^{\frac{1}{4}+\frac{p}{2}}y^{\frac{1}{4}+\frac{\beta-1}{2}}}\left(\frac{(l+1)_p}{(l+\beta)_p}\right)^{\frac{1}{2}} \\
=& \frac{C(1-x)^{l+\frac{p+\beta}{2}}}{\sqrt{2}(2l+\beta+p)^{\frac{1}{4}}(1-x-y)^{\frac{1}{4}+\frac{p}{2}}y^{\frac{1}{4}+\frac{\beta-1}{2}}}\left(\frac{(l+1)_p}{(l+\beta)_p}\right)^{\frac{1}{2}}.
\end{aligned} \tag{3.30}$$

Unter der Verwendung von (3.29) und (3.30) kann man für alle $m,l \in \mathbb{N}$ und alle $(x,y) \in \mathring{\mathbb{T}}$ die Werte der APK-Polynome nach oben abschätzen. Es gilt

$$\begin{aligned}
|A_{m,l}(x,y)| &= \left|P_m^{\alpha-1,a_l}(1-2x)P_l^{p,\beta-1}\left(\frac{2y}{1-x}-1\right)(1-x)^l\right| \\
&= \left|P_m^{\alpha-1,a_l}(1-2x)\right|\left|P_l^{p,\beta-1}\left(\frac{2y}{1-x}-1\right)(1-x)^l\right| \\
&\leq \frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{a_l}{2}}\sqrt{2}}\left(\frac{(m)_\alpha(m+a_l)}{(m+a_l)_\alpha m}\right)^{\frac{1}{2}} \\
&\quad\cdot\frac{C(1-x)^{l+\frac{p+\beta}{2}}}{\sqrt{2}(2l+\beta+p)^{\frac{1}{4}}(1-x-y)^{\frac{1}{4}+\frac{p}{2}}y^{\frac{1}{4}+\frac{\beta-1}{2}}}\left(\frac{(l+1)_p}{(l+\beta)_p}\right)^{\frac{1}{2}}.
\end{aligned}$$

Mit $\kappa_{l,m} := \sqrt{\frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p(m)_\alpha(m+a_l)}}$ und $E = C^2$ gilt weiterhin

$$\begin{aligned}
|A_{m,l}(x,y)| &\leq \frac{E(1-x)^{\frac{a_l}{2}}}{2(1-x-y)^{\frac{p}{2}+\frac{1}{4}}y^{\frac{\beta-1}{2}+\frac{1}{4}}x^{\frac{\alpha-1}{2}+\frac{1}{4}}(1-x)^{\frac{a_l}{2}+\frac{1}{4}}\left((2l+\beta+p)(2(m+l)+\gamma)\right)^{\frac{1}{4}}\kappa_{l,m}} \\
&= \frac{E}{2(1-x-y)^{\frac{p}{2}+\frac{1}{4}}y^{\frac{\beta-1}{2}+\frac{1}{4}}x^{\frac{\alpha-1}{2}+\frac{1}{4}}(1-x)^{\frac{1}{4}}\left((2l+\beta+p)(2(m+l)+\gamma)\right)^{\frac{1}{4}}\kappa_{l,m}}.
\end{aligned}$$

Die von l und m unabhängigen Faktoren fassen wir alle zusammen:

$$\tilde{E}(x,y) = \frac{E}{2(1-x-y)^{\frac{p}{2}+\frac{1}{4}} y^{\frac{\beta-1}{2}+\frac{1}{4}} x^{\frac{\alpha-1}{2}+\frac{1}{4}} (1-x)^{\frac{1}{4}}}.$$

Insgesamt gilt für alle $(x,y) \in \mathring{\mathbb{T}}$

$$|A_{m,l}(x,y)| \leq \frac{\tilde{E}(x,y)}{(2l+\beta+p)^{\frac{1}{4}}(2(m+l)+\gamma)^{\frac{1}{4}}\kappa_{m,l}},$$

also Ungleichung (3.24).
Als nächstes untersuchen wir das Verhalten der APK-Polynome an den Kanten.
Wir beginnen mit der Kante $[0,y]$ für $y \in [0,1]$. Es ist

$$|A_{m,l}(0,y)| = \left| P_m^{\alpha-1,a_l}(1) P_l^{p,\beta-1}(2y-1) \right| \overset{(3.3)}{=} \binom{m+\alpha-1}{m} \left| P_l^{p,\beta-1}(2y-1) \right|.$$

Für $q = \max\{p, \beta-1\} \geq -\frac{1}{2}$ nimmt das Jacobi-Polynom seinen maximalen Betragswert an einem der Randpunkte an. Mit (3.5) folgt daher

$$|A_{m,l}(0,y)| \leq \left(\binom{m+\alpha-1}{m} \cdot \max\left\{ \binom{l+p}{l}, \binom{l+\beta-1}{l} \right\} \right) \quad \forall y \in [0,1].$$

Die Abschätzung (3.29) gilt auch auf den Kanten $[x,0]$ und $[x,1-x]$ mit $x \in (0,1)$. Es ergeben sich für die Beträge der APK-Polynome mit Hilfe der Gleichungen (3.3) und (3.4)

$$\begin{aligned}
|A_{m,l}(x,0)| &= |P_m^{\alpha-1,a_l}(1-2x)(1-x)^l P_l^{p,\beta-1}(-1)| \\
&= |P_m^{\alpha-1,a_l}(1-2x)|(1-x)^l \binom{l+\beta-1}{l} \\
&\overset{(3.29)}{\leq} \frac{C(1-x)^l}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{a_l}{2}}\sqrt{2}} \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m} \right)^{\frac{1}{2}} \binom{l+\beta-1}{l} \\
&\leq \frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}\sqrt{2}} \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m} \right)^{\frac{1}{2}} \binom{l+\beta-1}{l}
\end{aligned}$$

und

$$\begin{aligned}
|A_{m,l}(x,1-x)| =& |P_m^{\alpha-1,a_l}(1-2x)(1-x)^l P_l^{p,\beta-1}(1)| \\
&\overset{(3.29)}{\leq} \frac{C(1-x)^l}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{a_l}{2}}\sqrt{2}} \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m} \right)^{\frac{1}{2}} \binom{l+p}{l} \\
&= \frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}\sqrt{2}} \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m} \right)^{\frac{1}{2}} \binom{l+p}{l}.
\end{aligned}$$

Als letztes wird der Wert der AKP-Polynome im Punkt $(1,0)$ bestimmt.
Dafür betrachten wir zuallererst die Faktoren von $P_l^{p,\beta-1}\left(\frac{2y}{1-x}-1\right)(1-x)^l$. Mit Hilfe der Reihendarstellung der Jacobi-Polynome (3.2) erhält man

$$\begin{aligned}
&P_l^{p,\beta-1}\left(\frac{2y}{1-x}-1\right)(1-x)^l \\
&\quad= \frac{(1-x)^l\Gamma(p+l+1)}{l!\Gamma(p+\beta-1+l+1)}\sum_{i=0}^{l}\binom{l}{i}\frac{\Gamma(p+\beta-1+l+i+1)}{\Gamma(p+i+1)}\left(\frac{2(y-1+x)}{2(1-x)}\right)^i \\
&\quad= \frac{\Gamma(p+l+1)}{l!\Gamma(p+\beta-1+l+1)}\sum_{i=0}^{l}\binom{l}{i}\frac{\Gamma(p+\beta-1+l+i+1)}{\Gamma(p+i+1)}(y-1+x)^i(1-x)^{l-i}.
\end{aligned}$$

Für $l \neq 0$ folgt, dass bei dem Wert $(x,y)=(1,0)$ die Summe verschwindet. Dies führt auf

$$\begin{aligned}
|A_{m,l}(1,0)| &= \left|P_m^{\alpha-1,a_l}(-1)\frac{\Gamma(p+l+1)}{l!\Gamma(p+\beta-1+l+1)}\sum_{i=0}^{l}\binom{l}{i}\frac{\Gamma(p+\beta-1+l+i+1)}{\Gamma(p+i+1)}\cdot 0^l\right| \\
&= 0.
\end{aligned}$$

Im Falle $l=0$ ergibt sich

$$|A_{m,0}(1,0)| = \left|P_m^{\alpha-1,a_l}(-1)\cdot\frac{\Gamma(p+1)}{\Gamma(p+\beta)}\frac{\Gamma(p+\beta)}{\Gamma(p+1)}\right| = \binom{m+a_l}{m}.$$

□

3.3 Approximationseigenschaften der APK-Polynome

Nicht nur das Verhalten der Koeffzienten $\hat{u}_{m,l}$ ist für uns von Interesse, sondern auch der Abschneidefehler bezüglich der gewichteten $\mathrm{L}^2(\mathbb{T},h)$-Norm sowie bezüglich des Betrages. Wir untersuchen die Approximationseigenschaften der abgeschnittenen APK-Reihe in Satz 3.13 dahingehend. Für eine C^∞-Funktion u folgert man insbesondere aus dem Resultat die spektrale Konvergenz.

Satz 3.13. *Seien $\alpha,\ \beta \in \mathbb{N}$, $p \in \mathbb{N}_0$ und sei u eine Funktion aus dem Raum $H^{2k}(\mathbb{T},h)\cap C(\mathbb{T})$, $k \in \mathbb{N}$, mit Gewichtsfunktion $h(x,y)=x^{\alpha-1}y^{\beta-1}(1-x-y)^p$. Die APK-Entwicklung von u ist gegeben durch*

$$P_N u(x,y) = \sum_{\substack{l+m\leq N\\ l,m\ \in\mathbb{N}_0}} \tilde{u}_{m,l}A_{m,l}(x,y), \quad \tilde{u}_{m,l} = \frac{(u;A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)}}{(A_{m,l};A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)}}. \tag{3.31}$$

Dann gelten folgende Abschätzungen:

$$(A_{m,l};A_{m,l})^{\frac{1}{2}}_{\mathrm{L}^2(\mathbb{T},h)}\cdot|\tilde{u}_{m,l}| = \mathcal{O}(\lambda_{m,l}^{-k}), \quad (m+l)\to\infty, \tag{3.32}$$

$$||u-P_N u||_{\mathrm{L}^2(\mathbb{T},h)} = \mathcal{O}(N^{-2k}), \quad N\to\infty. \tag{3.33}$$

Für die punktweise Abschätzung des Abschneidefehlers erhält man

(a) für $\beta - 1 < p$ *und* $1 + \frac{p+\beta}{2} \leq \frac{3}{4} + \frac{3}{4}\alpha + \frac{p}{2} < k$ *:*

$$|u(x,y) - P_N u(x,y)| = \mathcal{O}(N^{-2k+\frac{3}{2}\alpha+p+\frac{1}{2}}), \quad \forall (x,y) \in \mathbb{T}, \tag{3.34}$$

(b) für $\beta - 1 < p$ *und* $\frac{3}{4} + \frac{3}{4}\alpha + \frac{p}{2} \leq 1 + \frac{p+\beta}{2} < k$*:*

$$|u(x,y) - P_N u(x,y)| = \mathcal{O}(N^{-2k+p+\beta+\frac{3}{2}}), \quad \forall (x,y) \in \mathbb{T}, \tag{3.35}$$

(c) für $p \leq \beta - 1$ *und* $1 + \frac{p+\beta}{2} \leq \frac{1}{4} + \frac{3}{4}\alpha + \frac{\beta}{2} < k$*:*

$$|u(x,y) - P_N u(x,y)| = \mathcal{O}(N^{-2k+\frac{3}{2}\alpha+\beta-\frac{1}{2}}), \quad \forall (x,y) \in \mathbb{T}, \tag{3.36}$$

(d) für $p \leq \beta - 1$ *und* $\frac{1}{4} + \frac{3}{4}\alpha + \frac{\beta}{2} \leq 1 + \frac{p+\beta}{2} < k$*:*

$$|u(x,y) - P_N u(x,y)| = \mathcal{O}(N^{-2k+p+\beta+\frac{3}{2}}), \quad \forall (x,y) \in \mathbb{T}. \tag{3.37}$$

Bemerkung. Das Verhalten (3.33) wird bereits von Braess und Schwab in [8] gezeigt, allerdings mit einem komplett anderen Ansatz. Dabei verwenden sie Schwerpunktskoordinaten, transformieren den Differentialoperator und nutzen Symmetriebeziehungen im Dreieck.

Für den Beweis benötigen wir noch folgendes Lemma.

Lemma 3.14. *Unter den Annahmen von* **S**atz 3.13 *besitzt die Funktion* $w_u : [0,1] \to \mathbb{R}$*, definiert durch*

$$w_u(x) := \begin{cases} u(1,0), & x = 1, \\ \dfrac{\Gamma(p+\beta+1)}{\Gamma(p+1)\Gamma(\beta)} \displaystyle\int_0^{1-x} \frac{y^{\beta-1}(1-x-y)^p}{(1-x)^{p+\beta}} u(x,y)\, \mathrm{d}y, & x \neq 1, \end{cases}$$

folgende Eigenschaften:

(i) w_u *ist stetig.*

(ii) Entwickelt man w_u *in die Reihe* $\displaystyle\sum_{m=0}^{\infty} \hat{w}_{u,m} P_m^{\alpha-1,p+\beta}(1-2x)$*, dann gilt*

$$\hat{w}_{u,m} = \tilde{u}_{m,0}. \tag{3.38}$$

Dabei sei $\hat{w}_{u,m}$ *die Fourier-Entwicklung bezüglich der Jacobi-Polynome* $P_m^{\alpha-1,p+\beta}(1-2x)$ *und* $\tilde{u}$ (3.31) *aus* Satz 3.13*.*

Beweis. Die Stetigkeit von w für alle $x \in [0,1)$ folgt direkt aus der Stetigkeit von u. Wir müssen lediglich die Stetigkeit von w im Punkt $x = 1$ beweisen.
Sei $\epsilon > 0$ und $0 < |1-\xi| < \delta$, dann gilt

$$|w_u(\xi) - w_u(1)| = \left| \frac{\Gamma(p+\beta+1)}{\Gamma(p+1)\Gamma(\beta)} \int_0^{1-\xi} \frac{y^{\beta-1}(1-\xi-y)^p}{(1-\xi)^{p+\beta}} u(\xi,y)\, \mathrm{d}y - u(1,0) \right|.$$

Wir multiplizieren $u(1,0)$ mit $\frac{\Gamma(p+\beta+1)}{\Gamma(p+1)\Gamma(\beta)} \int_0^{1-\xi} \frac{y^{\beta-1}(1-\xi-y)^p}{(1-\xi)^{p+\beta}}\, \mathrm{d}y = 1$. Die Integraldarstellung[5] stammt aus den Eigenschaften der Betafunktion, vergleiche Definition 8.2. Dies führt zu

$$\begin{aligned}
|w_u(\xi) - w_u(1)| &= \left| \frac{\Gamma(p+\beta+1)}{\Gamma(p+1)\Gamma(\beta)} \int_0^{1-\xi} \frac{y^{\beta-1}(1-\xi-y)^p}{(1-\xi)^{p+\beta}} \left(u(\xi,y) - u(1,0)\right)\, \mathrm{d}y \right| \\
&\leq \frac{\Gamma(p+\beta+1)}{\Gamma(p+1)\Gamma(\beta)} \int_0^{1-\xi} \frac{y^{\beta-1}(1-\xi-y)^p}{(1-\xi)^{p+\beta}} \left|u(\xi,y) - u(1,0)\right|\, \mathrm{d}y \\
&\leq \max_{0\leq y_1 \leq 1-\xi} |u(\xi,y_1) - u(1,0)| \underbrace{\frac{\Gamma(p+\beta+1)}{\Gamma(p+1)\Gamma(\beta)} \int_0^{1-\xi} \frac{y^{\beta-1}(1-\xi-y)^p}{(1-\xi)^{p+\beta}}\, \mathrm{d}y}_{=1} \\
&= \max_{0\leq y_1 \leq 1-\xi} |u(\xi,y_1) - u(1,0)| < \epsilon.
\end{aligned}$$

Daher ist w_u auch im Punkt $x = 1$ stetig.
Wir entwickeln $w_u(x)$ in die Fourier-Reihe bezüglich der Jacobi-Polynome $P_m^{\alpha-1,p+\beta}(1-2x)$. Das zu betrachtende Intervall ist $[0,1]$ und als Gewichtsfunktion erhält man $\omega(x) = (2x)^{\alpha-1}(2-2x)^{p+\beta}$. Es ergibt sich mit Lemma 3.2

$$\begin{aligned}
\hat{w}_{u,m} =& \frac{1}{||P_m^{\alpha-1,p+\beta}||^2_{\mathrm{L}^2([-1,1],\omega)}} \int_0^1 (2x)^{\alpha-1}(2-2x)^{p+\beta} w_u(x) P_m^{\alpha-1,p+\beta}(1-2x) 2\, \mathrm{d}x \\
\overset{\text{Lemma 3.2}}{=}& \frac{(2m+\gamma)\Gamma(m+1)\Gamma(m+p+\beta+\alpha)2^{\alpha+\beta+p}}{\Gamma(m+\alpha)\Gamma(m+p+\beta+1)2^{\alpha+\beta+p}} \int_0^1 P_m^{\alpha-1,p+\beta}(1-2x) \\
&\cdot (x)^{\alpha-1}(1-x)^{p+\beta} \left(\frac{\Gamma(p+\beta+1)}{\Gamma(p+1)\Gamma(\beta)} \int_0^{1-x} \frac{y^{\beta-1}(1-x-y)^p}{(1-x)^{p+\beta}} u(x,y)\, \mathrm{d}y \right) \mathrm{d}x.
\end{aligned}$$

Mit

$$\frac{1}{(A_{m,0}, A_{m,0})_{\mathrm{L}^2(\mathbb{T},h)}} = \frac{(2m+\gamma)\Gamma(m+1)\Gamma(m+p+\beta+\alpha)\Gamma(p+\beta+1)}{\Gamma(m+\alpha)\Gamma(m+p+\beta+1)\Gamma(p+1)\Gamma(\beta)}$$

[5] Die Berechnung wurde mit Mathematica durchgeführt.

folgt:

$$\begin{aligned}\hat{w}_{u,m} =& \frac{1}{||A_{m,0}||^2_{\mathrm{L}^2(\mathbb{T},h)}} \int_0^1 \left(\int_0^{1-x} \underbrace{(x)^{\alpha-1} y^{\beta-1}(1-x-y)^p}_{=h(x,y)} u(x,y) P_m^{\alpha-1,p+\beta}(1-2x)\,\mathrm{d}y \right) \mathrm{d}x \\ =& \tilde{u}_{m,0}.\end{aligned}$$

□

Schließlich können wir nun Satz 3.13 beweisen.

Beweis. Sei $m + l \neq 0$. Dann gilt

$$(A_{m,l}; A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)} \cdot \tilde{u}_{m,l} = (u; A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)} \overset{(3.10)}{=} \left(u; \frac{DA_{m,l}}{\lambda_{m,l}} \right)_{\mathrm{L}^2(\mathbb{T},h)}.$$

Der Differentialoperator D aus (3.11) ist nach Satz 3.10 im Inneren des Dreiecks $\mathbb{T}$ potentiell selbstadjungiert. Der Rand $\partial\mathbb{T}$ hat das Maß Null und deshalb keinen Einfluss auf den Wert des Integrals[6]. Da D potentiell selbstadjungiert ist, existiert nach Definition 3.18 eine positive C^2-Funktion g, so dass gD selbstadjungiert ist. Die Funktion $\frac{1}{g}$ ist wohldefiniert und ebenfalls selbstadjungiert. Es gilt

$$\begin{aligned}\left(u; \frac{DA_{m,l}}{\lambda_{m,l}} \right)_{\mathrm{L}^2(\mathbb{T},h)} = \left(u; \frac{(gD)A_{m,l}}{g \cdot \lambda_{m,l}} \right)_{\mathrm{L}^2(\mathbb{T},h)} &= \frac{1}{\lambda_{m,l}} \left(\frac{1}{g}(gD)u; A_{m,l} \right)_{\mathrm{L}^2(\mathbb{T},h)} \\ &\overset{\text{rekursiv}}{=} \left(\frac{1}{\lambda_{m,l}} \right)^k \left(D^k u; A_{m,l} \right)_{\mathrm{L}^2(\mathbb{T},h)},\end{aligned}$$

und insgesamt hat man

$$(A_{m,l}; A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)} \cdot \tilde{u}_{m,l} = \left(\frac{1}{\lambda_{m,l}} \right)^k \left(D^k u; A_{m,l} \right)_{\mathrm{L}^2(\mathbb{T},h)} \tag{3.39}$$

gezeigt. Mit der Schwarz'schen Ungleichung auf $\mathrm{L}^2(\mathbb{T}, h)$,

$$\left(D^k u; A_{m,l} \right)_{\mathrm{L}^2(\mathbb{T},h)} \leq ||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} ||A_{m,l}||_{\mathrm{L}^2(\mathbb{T},h)},$$

erhalten wir für alle $k \in \mathbb{N}$ die Abschätzung

$$||A_{m,l}||_{\mathrm{L}^2(\mathbb{T},h)} \cdot |\tilde{u}_{m,l}| \leq \left| \frac{1}{\lambda_{m,l}} \right|^k \underbrace{||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}}_{<\infty, \text{ da } u \in H^k(\mathbb{T},h)} = \mathcal{O}(\lambda_{m,l}^{-k})$$

[6] Alternativ hätte man auch mit der Stetigkeit des Innenprodukts argumentieren können.

was (3.32) beweist.
Wir nutzen die Parseval'sche Identität und Gleichung (3.39) und erhalten

$$\begin{aligned}
||u - P_N u||^2_{\mathrm{L}^2(\mathbb{T},h)} &\overset{\text{Parseval}}{=} \sum_{l,m\ \in\mathbb{N}_0} \frac{1}{(A_{m,l};A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)}} \left|(u - P_N u; A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)}\right|^2 \\
&= \sum_{l,m\ \in\mathbb{N}_0} \frac{\left|(u;A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)} - (P_N u;A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)}\right|^2}{(A_{m,l};A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)}} \\
&= \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \tilde{u}^2_{m,l} \cdot (A_{m,l};A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)} \\
&\overset{(3.39)}{=} \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{|(D^k u;A_{m,l})_{L^2(\mathbb{T},h)}|^2}{(\lambda_{m,l})^{2k}(A_{m,l};A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)}} \\
&\leq \frac{1}{\lambda^{2k}_{N+1}} \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{|(D^k u;A_{m,l})_{L^2(\mathbb{T},h)}|^2}{(A_{m,l};A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)}} \\
&\overset{\text{Parseval}}{\leq} \frac{1}{\lambda^{2k}_{N+1}} ||D^k u||^2_{\mathrm{L}^2(\mathbb{T},h)}
\end{aligned}$$

mit $\lambda_{N+1} = \frac{1}{(N+1)(N+1+\gamma)}$. Wir verwenden die Abschätzung

$$||D^k u||^2_{\mathrm{L}^2(\mathbb{T},h)} \leq C_k ||u||_{H^{2k}(\mathbb{T},h)},$$

wobei C_k eine von u unabhängige positive Konstante ist. Insgesamt gilt

$$||u - P_N u||_{\mathrm{L}^2(\mathbb{T},h)} \leq C_k \left(\frac{1}{N^{2k}}\right) ||u||_{H^{2k}(\mathbb{T},h)} = \mathcal{O}(N^{-2k}).$$

Zum Beweis der Gleichungen (3.34)-(3.37) orientieren wir uns am Beweis der PKD-Approximationseigenschaften aus [63, S.38]. Zuerst bestätigen wir, dass

$$u(x,y) = \sum_{m,l\in N_0} \tilde{u}_{m,l} A_{m,l}(x,y) \tag{3.40}$$

für jeden Punkt $(x,y) \in \mathbb{T}$ gilt und danach beweisen wir die Gleichungen (3.34)-(3.37). Bevor wir (3.40) zeigen, geben wir noch drei elementare Abschätzungen an, die wir im Folgenden oft nutzen.
Unter der Voraussetzung $\lambda_{m,l} \neq 0$ können wir eine Abschätzung für $|\tilde{u}_{m,l}|$ mit Hilfe von Gleichung (3.39) und Abschätzung (3.23) herleiten. Es ist

$$|\tilde{u}_{m,l}| \overset{(3.39)}{=} \left(\frac{1}{\lambda_{m,l}}\right)^k \frac{\left(D^k u;A_{m,l}\right)_{\mathrm{L}^2(\mathbb{T},h)}}{(A_{m,l};A_{m,l})_{\mathrm{L}^2(\mathbb{T},h)}} \overset{(3.23)}{<} \frac{2(m+l+\gamma)\kappa_{l,m}||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}}{((m+l)(m+l+\gamma))^k}. \tag{3.41}$$

Weiterhin benötigen wir die Abschätzungen

$$\frac{(l+\beta)_p}{(l+1)_p} = \prod_{i=1}^{p}\left(\frac{l+\beta+i-1}{l+1+i-1}\right) = \prod_{i=1}^{p}\left(1+\frac{\beta-1}{l+i}\right) \leq (1+\beta-1)^p = \beta^p \tag{3.42}$$

und

$$\begin{aligned}\frac{(m+a_l)_\alpha m}{(m)_\alpha(m+a_l)} &= \frac{m\prod_{i=1}^{\alpha}(m+a_l+i-1)}{(m+a_l)\prod_{i=1}^{\alpha}(m+i-1)} = \frac{\prod_{i=2}^{\alpha}(m+a_l+i-1)}{\prod_{i=2}^{\alpha}(m+i-1)} \\ &= \frac{\prod_{i=1}^{\alpha-1}(m+a_l+i)}{\prod_{i=1}^{\alpha-1}(m+i)} = \prod_{i=1}^{\alpha-1}\left(1+\frac{a_l}{m+i}\right) \leq \left(1+\frac{a_l}{m+1}\right)^{\alpha-1}.\end{aligned} \tag{3.43}$$

Wir beginnen den Beweis von (3.40) damit, eine abgeschlossene Teilmenge $\Omega \subset \mathring{\mathbb{T}}$ zu betrachten. Wenden wir (3.24) und (3.41) an, so folgt:

$$\begin{aligned}&\sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} \sup_{(x,y)\in\Omega} |\tilde{u}_{m,l}A_{m,l}(x,y)| \leq \sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} |\tilde{u}_{m,l}| \sup_{(x,y)\in\Omega} |A_{m,l}(x,y)| \\ &\overset{(3.24)}{\leq} \sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} |\tilde{u}_{m,l}| \sup_{(x,y)\in\Omega} \left|\tilde{E}(x,y)\frac{1}{(2l+\beta+p)^{\frac{1}{4}}(2(m+l)+\gamma)^{\frac{1}{4}}\kappa_{l,m}}\right| \\ &\leq \sup_{(x,y)\in\Omega} |\tilde{E}(x,y)| \sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} |\tilde{u}_{l,m}| \frac{1}{(2l+\beta+p)^{\frac{1}{4}}(2(m+l)+\gamma)^{\frac{1}{4}}\kappa_{l,m}} \\ &\overset{(3.41)}{\leq} \sup_{(x,y)\in\Omega} |\tilde{E}(x,y)| \sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} \left(\frac{2(m+l+\gamma)\kappa_{l,m}}{((m+l)(m+l+\gamma))^k}\right. \\ &\qquad\qquad \left. \cdot \frac{||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}}{(2l+\beta+p)^{\frac{1}{4}}(2(m+l)+\gamma)^{\frac{1}{4}}\kappa_{l,m}}\right).\end{aligned}$$

Mit Hilfe der Identität

$$\sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} \left(\frac{1}{m+l}\right)^k = \sum_{i=1}^{N}\frac{i+1}{i^k} \tag{3.44}$$

und elementaren Abschätzungen erhalten wir

$$\begin{aligned}
&\sum_{\substack{0<l+m\leq N\\ l,m\ \in\widetilde{\mathbb{N}}_0}} \sup_{(x,y)\in\Omega} |\tilde{u}_{m,l}A_{m,l}(x,y)| \\
&\leq 2||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} \sup_{(x,y)\in\Omega} |\tilde{E}(x,y)| \sum_{\substack{0<l+m\leq N\\ l,m\ \in\widetilde{\mathbb{N}}_0}} \frac{1}{(m+l)^k(m+l+\gamma)^{k-\frac{3}{4}}} \\
&< 2||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} \sup_{(x,y)\in\Omega} |\tilde{E}(x,y)| \underbrace{\sum_{i\in\mathbb{N}} \frac{1}{(i)^k(i+\gamma)^{k-\frac{7}{4}}}}_{S_1}.
\end{aligned}$$

Die Reihe S_1 kann man durch die Reihen

$$\sum_{i\in\mathbb{N}} i^{\frac{7}{4}-2k} \text{ bzw. } \sum_{i\in\mathbb{N}} (i+\gamma)^{\frac{7}{4}-2k}$$

nach unten bzw. oben abschätzen. Für $k > \frac{11}{8}$ konvergieren beide Reihen und damit auch die Ausgangsreihe. Unter den Voraussetzungen an k ist $k > \frac{11}{8}$ immer erfüllt. Daher und aufgrund des Weierstraß'schen Majorantenkriteriums können wir sogar die gleichmäßige Konvergenz der Reihe garantieren und dementsprechend die Stetigkeit der Grenzfunktion auf Ω. Mit der Abschätzung

$$||u - P_N u||_{\mathrm{L}^2(\mathbb{T},h)} = \mathcal{O}(N^{-2k}) \tag{3.33}$$

und der Stetigkeit von u folgt, dass die Reihe mit der Funktion u auf Ω übereinstimmt. Da Ω beliebig gewählt war, stimmt u mit der Reihe im kompletten Inneren $\mathring{\mathbb{T}}$ des Dreiecks $\mathbb{T}$ überein.
Im Folgenden untersuchen wir noch die Kanten. Anstelle von (3.24) verwenden wir die Abschätzungen (3.25)–(3.27) für die jeweiligen Kanten. Wir beginnen mit der Kante $[0,y]$ mit $y \in [0,1]$. Hier gilt die Abschätzung (3.25). Wir nutzen die Ungleichung

$$\binom{m+\alpha-1}{m} = \frac{(m+\alpha-1)!}{(m)!(\alpha-1)!} = \frac{\prod_{i=1}^{\alpha-1}(m+i)}{(\alpha-1)!} \leq (m+\alpha-1)^{\alpha-1}, \tag{3.45}$$

bzw. die entsprechenden Ungleichungen für die jeweiligen Binomialkoeffizienten mit p oder $\beta-1$. Für den Fall $\alpha = 1$, $\beta = 1$ oder $p = 0$ gilt, dass die jeweiligen Binomialkoef-

fizienten gleich 1 sind.
Im Fall $\max\left\{\binom{l+p}{l}, \binom{l+\beta-1}{l}\right\} = \binom{l+p}{l}$ erhält man

$$\begin{aligned}
&\sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} \sup_{(0,y)\in[0,1]} |\tilde{u}_{m,l}A_{l,m}(0,y)| \leq \sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} |\tilde{u}_{m,l}| \sup_{(0,y)\in[0,1]} |A_{l,m}(0,y)| \\
&\overset{(3.25)}{\leq} \sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} |\tilde{u}_{m,l}| \left(\binom{m+\alpha-1}{m} \cdot \max\left\{\binom{l+p}{l}, \binom{l+\beta-1}{l}\right\}\right) \\
&= \sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} |\tilde{u}_{m,l}| \left(\binom{m+\alpha-1}{m} \cdot \binom{l+p}{l}\right) \\
&\overset{(3.45)}{\leq} \sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} |\tilde{u}_{m,l}| \, (m+\alpha-1)^{\alpha-1}(l+p)^p.
\end{aligned}$$

Zum Abschätzen von $\tilde{u}_{m,l}$ nutzt man (3.41). Es gilt

$$\begin{aligned}
&\sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} |\tilde{u}_{m,l}| \, (m+\alpha-1)^{\alpha-1}(l+p)^p \\
\leq &\sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} \frac{2(m+l+\gamma)\kappa_{m,l}||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}}{((m+l)(m+l+\gamma))^k}(m+\alpha-1)^{\alpha-1}(l+p)^p \\
\leq &\sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} \frac{2(m+\alpha-1)^{\alpha-1}(l+p)^p||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}}{(m+l)^k(m+l+\gamma)^{k-1}} \left(\frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p(m+a_l)(m)_\alpha}\right)^{\frac{1}{2}}.
\end{aligned}$$

Mit Hilfe der Abschätzungen (3.42) und (3.43) erhält man

$$\begin{aligned}
\leq &2^{\frac{\alpha+1}{2}}||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}\beta^{\frac{p}{2}} \sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} \frac{(m+\alpha-1)^{\alpha-1}(l+p)^p}{(m+l)^k(m+l+\gamma)^{k-1-\frac{\alpha-1}{2}}} \\
= &2^{\frac{\alpha+1}{2}}||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}\beta^{\frac{p}{2}} \sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} \frac{(m+\alpha-1)^{\alpha-1}(l+p)^p}{(m+l)^k(m+l+\gamma)^{k-1-\frac{\alpha-1}{2}}} =: \Xi_1.
\end{aligned}$$

Wegen $\alpha - 1 < \gamma$ und $p = \gamma - \alpha - \beta < \gamma$ gilt weiterhin

$$\begin{aligned}
\Xi_1 < &2^{\frac{\alpha+1}{2}}||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}\beta^{\frac{p}{2}} \sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} \frac{(l+m+\gamma)^{p+\alpha-1}}{(m+l)^k(m+l+\gamma)^{k-1-\frac{\alpha-1}{2}}} \\
\overset{(3.44)}{<} &2^{\frac{\alpha+1}{2}}||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}\beta^{\frac{p}{2}} \sum_{i\in\mathbb{N}} \frac{1}{i^k(i+\gamma)^{k-1-p-\alpha-\frac{\alpha-1}{2}}}.
\end{aligned}$$

Für die weiteren Kanten werden wir analoge Resultate erhalten. Dabei werden wir uns immer auf die hier geführte Argumentation beziehen. Wir untersuchen die Reihe

$$\sum_{i\in\mathbb{N}} \frac{(i+\gamma)^{0.5+p+\frac{3\alpha}{2}-k}}{(i)^k} \tag{3.46}$$

auf ihr Konvergenzverhalten. Grundlegende Idee ist es, den allgemeinen binomischen Lehrsatz zu benutzen. Der Satz besagt, dass für ein $\zeta \in \mathbb{R}$ und $x, y \in \mathbb{R}$

$$(x+y)^\zeta = x^\zeta \left(1+\frac{y}{x}\right)^\zeta = x^\zeta \sum_{k=0}^{\infty} \binom{\zeta}{k} \left(\frac{y}{x}\right)^k = \sum_{k=0}^{\infty} \binom{\zeta}{k} x^{\zeta-k} y^k \tag{3.47}$$

gilt und die Reihe[7] (3.47) konvergiert für alle $x, y \in \mathbb{R}$ mit $x > 0$ und $|\frac{y}{x}| < 1$. Man untersucht anstelle von Reihe (3.46) nur noch den Teil, bei dem $i > \gamma$ ist, also

$$\sum_{\substack{i\in\mathbb{N}\\ i>\gamma}} \frac{(i+\gamma)^{0.5+p+\frac{3\alpha}{2}-k}}{(i)^k}. \tag{3.48}$$

Da man lediglich eine endliche Anzahl an Summanden entfernt, unterscheidet sich das Konvergenzverhalten von (3.48) und (3.46) nicht. Wir wenden den allgemeinen binomischen Lehrsatz (3.47) auf den Teil $(\gamma + i)$ in Reihe (3.48) an und erhalten

$$\begin{aligned}\sum_{\substack{i\in\mathbb{N}\\ i>\gamma}} \frac{(i+\gamma)^{0.5+p+\frac{3\alpha}{2}-k}}{(i)^k} &= \sum_{\substack{i\in\mathbb{N}\\ i>\gamma}} \frac{(i)^{0.5+p+\frac{3\alpha}{2}-k} \sum_{n=0}^{\infty} \binom{0.5+p+\frac{3\alpha}{2}-k}{n} \left(\frac{\gamma}{i}\right)^n}{i^k} \\ = \sum_{\substack{i\in\mathbb{N}\\ i>\gamma}} \left(\frac{1}{i^{-0.5-p-\frac{3\alpha}{2}+2k}} \underbrace{\sum_{n=0}^{\infty} \binom{0.5+p+\frac{3\alpha}{2}-k}{n} \left(\frac{\gamma}{i}\right)^n}_{C_i} \right).\end{aligned}$$

C_i konvergiert für alle i. Sei nun $C_M = \max\limits_{\substack{i\in\mathbb{N}\\ i>\gamma}} \{|C_i|\}$, was man aus der Summe zieht. Man untersucht nur noch die Reihe

$$\sum_{\substack{i\in\mathbb{N}\\ i>\gamma}} \frac{1}{i^{-0.5-p-\frac{3\alpha}{2}+2k}}$$

auf Konvergenz. Die Reihe konvergiert, wenn der Exponent größer als 1 ist. Daher muss

$$2k - 0.5 - \frac{3\alpha}{2} - p > 1, \text{ das heißt } k > \left(\frac{3}{4} + \frac{3\alpha}{4} + \frac{p}{2}\right),$$

für die Konvergenz von (3.48) und damit auch für (3.46) gelten. Nach den Voraussetzung an k wird auch das immer erfüllt und entsprechend der bereits im Inneren geführten

[7] Bei dem Binomialkoeffizienten handelt es sich erneut um den verallgemeinerten Binomialkoeffizienten.

Argumentation über das Weierstraß'sche Majorantenkriterium, der gleichmäßigen Konvergenz und der Normabschätzung (3.33) folgt, dass auch auf dieser Kante die Reihe mit der Funktion u übereinstimmt.
Im Fall $\max\left\{\binom{l+p}{l}, \binom{l+\beta-1}{l}\right\} = \binom{l+\beta-1}{l}$ erhält man durch vergleichbare Umformungen und Abschätzungen die Bedingung

$$k > \frac{3}{4}\alpha + \frac{\beta}{2} + \frac{1}{4}$$

und das gleiche Ergebnis auch in diesem Fall.
Als nächstes wenden wir uns der unteren Kante des Dreiecks zu. Man wählt sich eine beliebige, aber abgeschlossene Teilmenge $\Omega_1 \subset [x, 0]$ mit $x \in (0, 1)$.
Auf Ω_1 gilt die Abschätzung (3.26). Verwenden wir diese Abschätzung mit (3.41) und (3.42), so ergibt sich die Rechnung:

$$\begin{aligned}
&\sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} \sup_{(x,0)\in\Omega_1} |\tilde{u}_{l,m}A_{l,m}(x,0)| \\
\overset{(3.26)}{\leq} &\sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} |\tilde{u}_{l,m}| \sup_{(x,0)\in\Omega_1} \left|\frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}\sqrt{2}}\right| \\
&\quad\cdot\left(\frac{(m)_\alpha(m+a_l)}{(m+a_l)_\alpha m}\right)^{\frac{1}{2}}\binom{l+\beta-1}{l} \\
\overset{(3.41)}{\leq} &\sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} \sup_{(x,0)\in\Omega_1} \left|\frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}\sqrt{2}}\right| \binom{l+\beta-1}{l} \\
&\quad\cdot\left(\frac{(m)_\alpha(m+a_l)}{(m+a_l)_\alpha m}\right)^{\frac{1}{2}}\left(\frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p(m+a_l)(m)_\alpha}\right)^{\frac{1}{2}}\frac{2(m+l+\gamma)||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}}{((m+l)(m+l+\gamma))^k} \\
\overset{(3.42)}{\leq} &\underbrace{||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}\beta^{\frac{p}{2}} \sup_{(x,0)\in\Omega_1}\left|\frac{2C}{(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}\sqrt{2}}\right|}_{C_2} \\
&\quad\cdot\sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} \frac{1}{(2(m+l)+\gamma)^{\frac{1}{4}}(m+l)^k(m+l+\gamma)^{k-1}}\binom{l+\beta-1}{l} \\
\leq &C_2 \sum_{\substack{0<l+m\leq N\\ l,m\ \in\overline{\mathbb{N}}_0}} \frac{1}{(m+l)^k(m+l+\gamma)^{k-\frac{3}{4}}}\binom{l+\beta-1}{l} =: \Xi_2.
\end{aligned}$$

Wegen

$$\binom{l+\beta-1}{l} \leq (l+\beta-1)^{\beta-1}$$

und $\gamma > \beta - 1$ erhält man

$$\Xi_2 \leq C_2 \sum_{\substack{0<l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \frac{(l+\beta-1)^{\beta-1}}{(m+l)^k(m+l+\gamma)^{k-\frac{3}{4}}} \leq C_2 \sum_{\substack{0<l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \frac{1}{(m+l)^k(m+l+\gamma)^{k+\frac{1}{4}-\beta}}$$
$$\overset{(3.44)}{\leq} C_2 \sum_{i\in\mathbb{N}} \frac{1}{i^k(i+\gamma)^{k-\frac{3}{4}-\beta}}.$$

Mit analoger Argumentation wie auf der Kante $[0, y]$ folgt schließlich die Konvergenz der Reihe, wenn

$$2k - \frac{3}{4} - \beta > 1, \text{ das heißt } k > \frac{7}{8} + \frac{\beta}{2},$$

gilt. Nach den Voraussetzungen an k ist dies garantiert und entsprechend einer analogen Argumentation folgt, dass die Funktion u mit ihrer APK-Reihe auf Ω_1 übereinstimmt. Da Ω_1 beliebig war, stimmt sogar u mit der Reihe auf der ganzen Kante $[x, 0]$ mit $x \in (0, 1)$ überein.
Als nächstes untersuchen wir die Darstellung auf der letzten Kante $[x, 1-x]$ mit $x \in (0, 1)$, bevor wir den Punkt $(1, 0)$ gesondert betrachten. Sei $\Omega_2 \subset [\tilde{x}, 1-\tilde{x}]$ mit $\tilde{x} \in [x_1, x_2] \subset (0, 1)$ und $x_1, x_2 \in (0, 1)$. Dann gilt für die APK-Polynome die Abschätzung (3.27). Durch analoges Vorgehen wie bei den vorherigen Fällen erhält man

$$\sum_{\substack{0<l+m\leq N \\ l,m\ \in\mathbb{N}_0}} |\tilde{u}_{m,l} A_{m,l}(x, 1-x)|$$
$$\overset{(3.27)}{\leq} \sum_{\substack{0<l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \sup_{(x,1-x)\in\Omega_2} \left| \frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}\sqrt{2}} \right| \binom{l+p}{l}$$
$$\cdot \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m} \right)^{\frac{1}{2}} \left(\frac{(l+\beta)_p m (m+a_l)_\alpha}{(l+1)_p (m+a_l)(m)_\alpha} \right)^{\frac{1}{2}} \frac{||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} 2(m+l+\gamma)}{((m+l)(m+l+\gamma))^k}$$
$$\leq C_3 \sum_{\substack{0<l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \frac{(l+p)^p}{(l+m)^k(l+m+\gamma)^{k-\frac{3}{4}}}$$
$$\leq C_3 \sum_{\substack{0<l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \frac{1}{(l+m)^k(l+m+\gamma)^{k-\frac{3}{4}-p}} \leq C_3 \sum_{i\in\mathbb{N}} \frac{1}{i^k(i+\gamma)^{k-\frac{7}{4}-p}}.$$

Die Reihe konvergiert genau dann, wenn

$$2k - \frac{7}{4} - p > 1, \text{ das heißt } k > \frac{11}{8} + \frac{p}{2},$$

ist und mit entsprechender Argumentation, wie in den letzten Fällen, ist somit (3.40) auf der Kante $[x, 1-x]$ mit $x \in (0, 1)$ nachgewiesen.
Somit ergeben sich an k folgende Bedingungen:

Kante	$\max\{\binom{l+p}{l}, \binom{l+\beta-1}{l}\}$	Bedingungen an k
$[0,y]$	$\binom{l+p}{l}$	$k > \frac{\gamma}{2} + \frac{\alpha}{4} - \frac{\beta}{2} + \frac{3}{4}$
$[0,y]$	$\binom{l+\beta-1}{l}$	$k > \frac{3}{4}\alpha + \frac{\beta}{2} + \frac{1}{4}$
$[x,0]$		$k > \frac{7}{8} + \frac{\beta}{2}$
$[x,1-x]$		$k > \frac{11}{8} + \frac{p}{2}$

Diese Bedingungen sind nach unseren Voraussetzungen immer erfüllt und insgesamt ist damit (3.40) für alle $(x,y) \in \mathbb{T}\backslash\{(1,0)\}$ gezeigt.
Im Punkt $(1,0)$ beweisen wir (3.40) direkt. Sei dafür $w_u : [0,1] \to \mathbb{R}$ die stetige Funktion aus Lemma 3.14. Mit Hilfe der Parseval'schen Identität folgt:

$$\int_0^1 2(2x)^{\alpha-1}(2-2x)^{p+\beta}\left[w_u(x) - \sum_{m\in\mathbb{N}_0} \hat{w}_{u,m} P_m^{\alpha-1,p+\beta}(1-2x)\right]^2 \mathrm{d}x = 0. \tag{3.49}$$

Wir untersuchen daher die Reihe $\sum_{m\in\mathbb{N}_0} \hat{w}_{u,m} P_m^{\alpha-1,p+\beta}(1-2x)$. Mit den Gleichungen (3.5), (3.38), (3.41), (3.42) und (3.43) erhält man

$$\begin{aligned}
&\sum_{0<m\le N} \left|\hat{w}_{u,m} P_m^{\alpha-1,p+\beta}(1-2x)\right| \overset{(3.38)}{=} \sum_{0<m\le N} \left|\tilde{u}_{m,0} P_m^{\alpha-1,p+\beta}(1-2x)\right| \\
&= \sum_{0<m\le N} \left|\tilde{u}_{m,0} P_m^{\alpha-1,p+\beta}(1-2x)\right| \le \sum_{0<m\le N} \left|\tilde{u}_{m,0} P_m^{\alpha-1,p+\beta}(1-2x)\right| \\
&\overset{(3.41)}{<} \sum_{0<m\le N} 2(m+\gamma)\sqrt{\frac{(\beta)_p m(m+a_0)_\alpha}{(1)_p (m+a_0)(m)_\alpha}} \frac{||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}}{m^k(m+\gamma)^k} \left|P_m^{\alpha-1,p+\beta}(1-2x)\right| \\
&\overset{\substack{(3.42)\&\\(3.43)}}{<} ||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} 2\beta^{\frac{p}{2}} \gamma^{\frac{\alpha-1}{2}} \sum_{m\in\mathbb{N}} \frac{1}{m^k(m+\gamma)^{k-1}} \left|P_m^{\alpha-1,p+\beta}(1-2x)\right| \\
&\overset{(3.5)}{\le} \begin{cases} ||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} 2\beta^{\frac{p}{2}} \gamma^{\frac{\alpha-1}{2}} \sum\limits_{m\in\mathbb{N}} \frac{1}{m^k(m+\gamma)^{k-1-p-\beta}}, & \text{für } \alpha-1 < p+\beta, \\ ||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} 2\beta^{\frac{p}{2}} \gamma^{\frac{\alpha-1}{2}} \sum\limits_{m\in\mathbb{N}} \frac{1}{m^k(m+\gamma)^{k-\alpha}}, & \text{für } \alpha-1 \ge p+\beta. \end{cases}
\end{aligned}$$

Mit der gleichen Argumentation wie zuvor folgt, dass die w-Reihe absolut konvergiert und die Grenzfunktion stetig auf $[0,1]$ ist. Wegen (3.49) stimmt die Reihe mit w überein. Insgesamt erhalten wir

$$\sum_{l,m\in\mathbb{N}_0} \tilde{u}_{m,l} A_{m,l}(1,0) = \sum_{m\in\mathbb{N}_0} \tilde{u}_{m,0} P_m^{\alpha-1,p+\beta}(-1) = w_u(1) = u(1,0).$$

Nachdem wir (3.40) bewiesen haben, kommen wir im zweiten Abschnitt des Beweises dazu, den Abschneidefehler zu berechnen. Dabei werden wir erneut das Innere, die Kanten und den Eckpunkt $(1,0)$ gesondert betrachten.
Für $(x,y) \in \mathring{\mathbb{T}}$ folgt mit den Gleichungen (3.39) und (3.40)

$$\begin{aligned} u(x,y) - P_N u(x,y) &= \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \tilde{u}_{m,l} A_{m,l}(x,y) \\ &= \int_{\mathbb{T}} h(x_1,y_1) R_N(x,y,x_1,y_1) D^k u(x_1,y_1) \mathrm{d}x_1 \mathrm{d}y_1, \end{aligned}$$

mit

$$R_N(x,y,x_1,y_1) = \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{A_{m,l}(x,y) A_{m,l}(x_1,y_1)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)} \lambda^k_{m,l}}.$$

Berechnet man die Norm von $R_N(x,y,\cdot,\cdot)$, erhält man

$$||R_N(x,y,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} = \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{A^2_{m,l}(x,y)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)} \lambda^{2k}_{m,l}} \tag{3.50}$$

und damit

$$\begin{aligned} ||R_N(x,y,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} &= \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{A^2_{m,l}(x,y)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)} \lambda^{2k}_{m,l}} \\ \overset{(3.23)}{<} & \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} 4(m+l+\gamma)^2 \kappa^2_{l,m} \frac{A^2_{m,l}(x,y)}{(m+l+\gamma)^{2k}(m+l)^{2k}} \\ \overset{(3.24)}{\leq} & 4\tilde{E}(x,y)^2 \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{\kappa^2_{l,m}}{(m+l+\gamma)^{2k-2}(m+l)^{2k}} \frac{1}{(2l+\beta+p)^{\frac{1}{2}}(2(m+l)+\gamma)^{\frac{1}{2}} \kappa^2_{l,m}} \\ = & 4\tilde{E}(x,y)^2 \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{1}{(m+l+\gamma)^{2k-2}(m+l)^{2k}} \frac{1}{(2l+\beta+p)^{\frac{1}{2}}(2(m+l)+\gamma)^{\frac{1}{2}}} \\ < & 4\tilde{E}(x,y)^2 \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{1}{(m+l+\gamma)^{2k-2}(m+l)^{2k+\frac{1}{2}}} \\ \overset{(3.44)}{<} & 4\tilde{E}(x,y)^2 \sum_{\substack{i>N \\ i\ \in\mathbb{N}_0}} \frac{1}{(i)^{4k-\frac{5}{2}}} < 4\tilde{E}(x,y)^2 \int_N^\infty \frac{1}{t^{4k-\frac{5}{2}}} \mathrm{d}t < 4\tilde{E}(x,y)^2 N^{\frac{7}{2}-4k}. \end{aligned}$$

Schließlich gilt mit der Schwarz'schen Ungleichung für jeden Punkt im Inneren des Dreiecks

$$|u(x,y) - P_N u(x,y)| \leq ||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} ||R_N(x,y,\cdot,\cdot)||_{\mathrm{L}^2(\mathbb{T},h)} = \mathcal{O}(N^{-2k+\frac{7}{4}}).$$

Im weiteren Verlauf untersuchen wir das Verhalten von $||R_N||_{\mathrm{L}^2(\mathbb{T},h)}$ auf dem Rand des Dreiecks $\mathbb{T}$.
Beginnen wir abermals mit der Kante $[0,y]$ mit $y \in [0,1]$. Es wurde bereits mit (3.25) und (3.45)

$$\max_{y\in[0,1]} |A_{m,l}(0,y)| \leq \begin{cases} (m+\alpha-1)^{\alpha-1}(l+p)^p, & \text{für } p > \beta-1, \\ (m+\alpha-1)^{\alpha-1}(l+\beta-1)^{\beta-1}, & \text{für } p \leq \beta-1, \end{cases}$$

gezeigt.

Für den Fall $p > \beta - 1$ ergibt sich damit für die Norm von R_N mit den weiteren Abschätzungen (3.23) (3.42), (3.43) und (3.44)

$$\begin{aligned} ||R_N(0,y,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} &= \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{A^2_{m,l}(x,y)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)}\lambda^{2k}_{m,l}} \\ &\overset{(3.23)}{<} \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{(m+\alpha-1)^{2\alpha-2}(l+p)^{2p}4\kappa^2_{m,l}(m+l+\gamma)^2}{((m+l)(m+l+\gamma))^{2k}} \\ &= \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{(m+\alpha-1)^{2\alpha-2}(l+p)^{2p}4}{(m+l)^{2k}(m+l+\gamma)^{2k-2}} \left(\frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p(m+a_l)(m)_\alpha}\right) \\ &\overset{(3.42)}{\leq} 4\beta^p \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{(l+p)^{2p}(m+\alpha-1)^{2\alpha-2}}{(m+l)^{2k}(m+l+\gamma)^{2k-2}} \left(\frac{m(m+a_l)_\alpha}{(m+a_l)(m)_\alpha}\right) \\ &\overset{(3.43)}{\leq} 4\beta^p \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{(l+p)^{2p}(m+\alpha-1)^{2\alpha-2}}{(m+l)^{2k}(m+l+\gamma)^{2k-2}} \left(1+\frac{a_l}{m+1}\right)^{\alpha-1}. \end{aligned}$$

Wegen $\gamma > \alpha - 1$ und $a_l + 1 < 2(m+l+\gamma)$ folgt:

$$\begin{aligned} ||R_N(0,y,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} &< 2^{\alpha+1}\beta^p \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{1}{(m+l+\gamma)^{2k-3\alpha+1-2p}(m+l)^{2k}} \\ &\overset{(3.44)}{<} 2^{\alpha+1}\beta^p \sum_{\substack{i>N \\ i\ \in\mathbb{N}}} \frac{1}{(i+\gamma)^{2k-3\alpha-2p}(i)^{2k}} =: \mathcal{R}. \end{aligned}$$

Ohne Beschränkung der Allgemeinheit nehmen wir an, dass $N > \gamma$ gilt[8]. Wir argumentieren analog über den allgemeinen binomischen Lehrsatz (3.47). Die Konvergenz für die Reihe ist gesichert, da $i > N > \gamma$ ist. Geht man zum Supremum S_2 über, so erhält man

$$\mathcal{R} < \underbrace{2^{\alpha+1}\beta^p S_2}_{C_{R_{2(1)}}} \sum_{\substack{i\in\mathbb{N} \\ i>N}} \frac{1}{(i)^{4k-3\alpha-2p}} < C_{R_{2(1)}} \int_N^\infty t^{-4k+3\alpha+2p}\mathrm{d}t < \frac{C_{R_{2(1)}}}{N^{4k-3\alpha-2p-1}}.$$

[8]Da es sich bei $\gamma \in \mathbb{N}$ um eine fest gewählte Konstante handelt und wir uns für das Approximationsverhalten für $N \to \infty$ interessieren, ist diese Annahme legitim.

Es folgt mit der Schwarz'schen Ungleichung

$$|u(0,y) - P_N u(0,y)| \leq ||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} ||R_N(0,y,\cdot,\cdot)||_{\mathrm{L}^2(\mathbb{T},h)} = \mathcal{O}(N^{-2k+\frac{3}{2}\alpha+p+\frac{1}{2}}).$$

Im zweiten Fall $p \leq \beta - 1$ kommt man durch analoges Auswerten auf

$$|u(0,y) - P_N u(0,y)| \leq \mathcal{O}(N^{-2k+\frac{3}{2}\alpha+\beta-\frac{1}{2}}).$$

Als nächste Kante betrachtet man $[x,0], x \in (0,1)$. Es ergibt sich für $||R_N||_{\mathrm{L}^2(\mathbb{T},h)}$ mit den Abschätzungen (3.23), (3.26), (3.42) und (3.44):

$$\begin{aligned}
&||R_N(x,0,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} \\
&\overset{\substack{(3.23)\&\\(3.26)}}{<} 4 \sum_{\substack{l+m>N\\ l,m\ \in\mathbb{N}_0}} \frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p (m)_\alpha (m+a_l)} \frac{1}{(m+l+\gamma)^{2k-2}(m+l)^{2k}} \binom{l+\beta-1}{l}^2 \\
&\qquad \cdot \left(\frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}} (x)^{\frac{1}{4}+\frac{\alpha-1}{2}} (1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}} \sqrt{2}} \right)^2 \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m} \right) \\
&\overset{(3.42)}{<} 2\beta^p \left(\frac{C}{(x)^{\frac{1}{4}+\frac{\alpha-1}{2}} (1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}} \right)^2 \sum_{\substack{l+m>N\\ l,m\ \in\mathbb{N}_0}} \frac{(l+\beta-1)^{2\beta-2}}{(m+l+\gamma)^{2k-\frac{3}{2}}(m+l)^{2k}} \\
&< 2\beta^p \left(\frac{C}{(x)^{\frac{1}{4}+\frac{\alpha-1}{2}} (1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}} \right)^2 \sum_{\substack{l+m>N\\ l,m\ \in\mathbb{N}_0}} \frac{1}{(m+l+\gamma)^{2k-2\beta+\frac{1}{2}}(m+l)^{2k}} \\
&\overset{(3.44)}{<} 2\beta^p \left(\frac{C}{(x)^{\frac{1}{4}+\frac{\alpha-1}{2}} (1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}} \right)^2 \sum_{\substack{i>N\\ i\ \in\mathbb{N}}} \frac{1}{(i+\gamma)^{2k-2\beta-\frac{1}{2}}(i)^{2k}} =: \mathcal{R}_1.
\end{aligned}$$

Erneut mit $N > \gamma$ und analoger Argumentation mit Hilfe des allgemeinen binomischen Lehrsatzes (3.47) gilt mit Übergang zum Supremum, dass man die Reihe nach oben hin durch

$$\mathcal{R}_1 < C_{R_3} \sum_{i>N} \frac{1}{(i)^{4k-2\beta-\frac{1}{2}}} < C_{R_3} \frac{1}{N^{4k-2\beta-\frac{3}{2}}},$$

mit einer passenden Konstanten C_{R_3} abschätzen kann. Man erhält mit der Schwarz'schen Ungleichung

$$|u(x,0) - P_N u(x,0)| \leq \mathcal{O}(N^{-2k+\frac{3}{4}+\beta}).$$

Bei der Kante $[x, 1-x], x \in (0,1)$ gehen wir analog vor:

$$
\begin{aligned}
&||R_N(x,1-x,\cdot,\cdot)||^2_{L^2(\mathbb{T},h)} \\
&\overset{(3.23)}{<} \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{A^2_{m,l}(x,1-x)}{(m+l)^{2k}(m+l+\gamma)^{2k}} 4(m+l+\gamma)^2 \frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p (m)_\alpha (m+a_l)} \\
&\overset{(3.27)}{<} \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \left(\frac{1}{(m+l)(m+l+\gamma)}\right)^{2k} 4(m+l+\gamma)^2 \frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p (m)_\alpha (m+a_l)} \\
&\qquad \cdot \left(\frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}} (x)^{\frac{1}{4}+\frac{\alpha-1}{2}} (1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}} \sqrt{2}} \left(\frac{(m)_\alpha (m+a_l)}{(m+a_l)_\alpha m}\right)^{\frac{1}{2}} \binom{l+p}{l} \right)^2 \\
&\overset{(3.42)}{<} \left(\frac{C}{(x)^{\frac{1}{4}+\frac{\alpha-1}{2}} (1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}} \right)^2 2\beta^p \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{1}{(m+l)^{2k}(m+l+\gamma)^{2k-\frac{3}{2}}} \binom{l+p}{l}^2 \\
&\overset{(3.45)}{<} \left(\frac{C}{(x)^{\frac{1}{4}+\frac{\alpha-1}{2}} (1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}} \right)^2 2\beta^p \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{(l+m+\gamma)^{2p}}{(m+l)^{2k}(m+l+\gamma)^{2k-\frac{3}{2}}} \\
&< \left(\frac{C}{(x)^{\frac{1}{4}+\frac{\alpha-1}{2}} (1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}} \right)^2 2\beta^p \sum_{\substack{l+m>N \\ l,m\ \in\mathbb{N}_0}} \frac{1}{(m+l)^{2k}(m+l+\gamma)^{2k-\frac{3}{2}-2p}} \\
&\overset{3.44}{<} \left(\frac{C}{(x)^{\frac{1}{4}+\frac{\alpha-1}{2}} (1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}} \right)^2 2\beta^p \sum_{\substack{i>N \\ i\ \in\mathbb{N}}} \frac{1}{(i)^{2k}(i+\gamma)^{2k-\frac{5}{2}-2p}} =: \mathcal{R}_2.
\end{aligned}
$$

Durch analoge Argumentation wie auf den übrigen Kanten ergibt sich

$$\mathcal{R}_2 < C_{R_4} \sum_{i>N} \frac{1}{(i)^{4k-\frac{5}{2}-2p}} < C_{R_4} \int_N^\infty \frac{1}{t^{4k-\frac{5}{2}-2p}} \mathrm{d}t < \frac{C_{R_4}}{N^{4k-\frac{7}{2}-2p}}$$

und mit der Schwarz'schen Ungleichung insgesamt

$$|u(x,1-x) - P_N u(x,1-x)| < \mathcal{O}(N^{-2k+\frac{7}{4}+p}).$$

Betrachten wir zum Schluss den Punkt $(1,0)$. Es ist

$$
\begin{aligned}
||R_N(1,0,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} &\overset{(3.28)\&\ (3.50)}{=} \sum_{m>N} \frac{|A_{m,0}(1,0)|^2}{m^{2k}(m+\gamma)^{2k} ||A_{m,0}||^2_{\mathrm{L}^2(\mathbb{T},h)}} \\
&\overset{(3.23)}{<} \sum_{m>N} 4(m+\gamma)^2 \frac{(0+\beta)_p m(m+a_0)_\alpha}{(0+1)_p (m)_\alpha (m+a_0)} \frac{\binom{m+\gamma-\alpha}{m}^2}{m^{2k}(m+\gamma)^{2k}} =: \mathcal{R}_3
\end{aligned}
$$

$$\mathcal{R}_3 \overset{(3.42)}{<} 4\left(\frac{(\beta)_p}{(1)_p}\gamma^{\alpha-1}\right)\sum_{m>N}\frac{\binom{m+\gamma-\alpha}{m}^2}{m^{2k}(m+\gamma)^{2k-2}} < 4\left(\frac{(\beta)_p}{(1)_p}\gamma^{\alpha-1}\right)\sum_{m>N}\frac{(m+\gamma)^{2\gamma-2\alpha}}{m^{2k}(m+\gamma)^{2k-2}}$$
$$< C_{R_5}\sum_{m>N}\frac{1}{m^{4k+2\alpha-2\gamma-2}} < \int_N^\infty \frac{C_{R_5}}{t^{4k+2\alpha-2\gamma-2}}\mathrm{d}t < \frac{C_{R_5}}{N^{4k+2\alpha-2\gamma-3}},$$

mit einer Konstanten C_{R_5}.
Insgesamt ergibt sich mit der Schwarz'schen Ungleichung

$$|u(1,0) - P_N u(1,0)| \leq ||D^k u||_{\mathrm{L}^2(\mathbb{T},h)}||R_N(1,0,\cdot,\cdot)||_{\mathrm{L}^2(\mathbb{T},h)} = \mathcal{O}(N^{-2k-\alpha+\gamma+\frac{3}{2}}).$$

Fassen wir schließlich die Approximationsresultate für die einzelnen Mengen zusammen, so erhält man folgende Tabelle:

Menge	$\max\{\binom{l+p}{l}, \binom{l+\beta-1}{l}\}$	$\lvert u(x,y) - P_N u(x,y)\rvert <$
$\mathring{\mathbb{T}}$		$\mathcal{O}(N^{-2k+\frac{7}{4}})$
$[0,y]$	$\binom{l+p}{l}$	$\mathcal{O}(N^{-2k+\frac{3}{2}\alpha+p+\frac{1}{2}})$
$[0,y]$	$\binom{l+\beta-1}{l}$	$\mathcal{O}(N^{-2k+\frac{3}{2}\alpha+\beta-\frac{1}{2}})$
$[x,0]$		$\mathcal{O}(N^{-2k+\frac{3}{4}+\beta})$
$[x,1-x]$		$\mathcal{O}(N^{-2k+\frac{7}{4}+p})$
$(1,0)$		$\mathcal{O}(N^{-2k-\alpha+\gamma+\frac{3}{2}})$

Tabelle 3.2: *Approximationsresultate*

Vergleicht man die Approximationsgeschwindigkeiten auf den einzelnen Abschnitten miteinander, so ist festzustellen, dass die Geschwindigkeit unter den Voraussetzungen an k auf der Kante $[0,y]$ oder im Punkt $(1,0)$ am geringsten ist. Man erhält die Gleichungen (3.34)-(3.37).

□

Bemerkung. Es ist nicht verwunderlich, dass die APK-Summe entweder auf der Kante $[0,y]$ oder im Punkt $(1,0)$ das langsamste Konvergenzverhalten der Approximation aufweist. Im Beweis von Satz 3.13 waren die Abschätzungen der APK-Polynome aus Lemma 3.12 wesentlich. Diese Abschätzungen hatten direkten Einfluss auf das Approximationsverhalten der abgeschnittenen APK-Reihe. Je größer der Wert der APK-Polynome war, desto schlechter wurde die Approximation. Warum also ist die Approximation auf der Kante $[0,y]$ oder im Punkt $(1,0)$ am schlechtesten?

Die Jacobi-Polynome nehmen ihr Maximum in einem ihrer Randpunkte an und als Produkt zweier Jacobi-Polynome folgt somit für die APK-Polynome, dass sie ihr Maximum in einer der Ecken annehmen. Daher ist dort die Approximation am schlechtesten. Trotzdem waren wir auch interessiert am Verhalten im Inneren des Dreiecks und auf den beiden übrigen Kanten, da dort bei der späteren numerischen Verwendung der APK-Polynome im Spektrale-Differenzen-Verfahren viele Gauß-Lobatto-Punkte liegen werden. Eine genauere Abschätzung für die Approximationsgeschwindigkeiten der APK-Summe auf den einzelnen Abschnitten des Dreiecks liefert uns Tabelle 3.2.
In [60] findet man den Satz 3.13 bereits veröffentlicht. Auch enthält der Artikel eine numerische Untersuchung des Fehlers bei der Approximation zweier Testfunktionen. In dieser Arbeit werden wir hingegen in Kapitel 6 die APK-Polynome direkt im Spektrale-Differenzen-Verfahren verwenden und untersuchen.

4 Modale Filter

Wie bereits im zweiten Kapitel dargelegt wurde, können Lösungen von hyperbolischen Erhaltungsgleichungen nach einer bestimmten Zeit Unstetigkeiten entwickeln. Bei der Konstruktion der numerischen Methode muss daher dieses spezielle Verhalten mitberücksichtigt werden. In diesem Abschnitt beschäftigen wir uns explizit mit einem Ansatz zur Verbesserung der numerischen Lösung, der **modalen Filterung**.
Dabei nehmen wir nochmals Bezug zur allgemeinen Problematik. Wird eine Funktion u mit Sprungunstetigkeiten durch eine abgeschnittene Fourier-Reihe approximiert, so kommt es in der Umgebung der Unstetigkeiten zu Oszillationen. In der Literatur spricht man dabei vom Gibbs'schen Phänomen [91]. Um die Oszillationen abzuschwächen bzw. bestmöglich zu entfernen, werden wir, wie in der Technik, Tiefpassfilter einsetzen. Hierfür definieren wir modale Filter mathematisch und nehmen kurz Bezug zur Elektrotechnik bzw. Signalverarbeitung. Für die gefilterte APK-Summe beweisen wir daraufhin noch einige neue Approximationsresultate, um anschließend den Zusammenhang zwischen Filterung und der **Spektralen Viskositätsmethode** (kurz: SV-Methode) zu erläutern.
Bei der SV-Methode wird ein kleiner Viskositätsterm zu der Differentialgleichung addiert. Löst man diese veränderte Differentialgleichung mit Hilfe einer spektralen Methode, so ist dies äquivalent dazu, dass man das spektrale Verfahren für die ursprüngliche hyperbolische Erhaltungsgleichung verwendet, dabei jedoch die Lösung in jedem Zeitschritt einem speziellen Exponentialfilter unterwirft. Für verschiedene Familien von APK-Polynomen erhält man dabei abhängig von einem Parameter verschiedene Filter.
Anschließend erklären wir noch, wann und in welchen Ausmaß wir überhaupt filtern und schließlich nehmen wir Stellung zum Theorem 4.2 aus [36]. Dieses trifft eine Aussage darüber, wie schnell die abgeschnittene gefilterte Legendre-Reihe gegen eine Funktion u konvergiert, wenn u nach Voraussetzung genau eine Unstetigkeitsstelle im Nullpunkt besitzt.

Gibbs'sches Phänomen

Spektrale Methoden besitzen als zugrundeliegenden Ansatz die Entwicklung der Lösung der Differentialgleichung mittels Basisfunktionenen eines Funktionenraumes. Aus Effizienzgründen werden vorrangig orthogonale Funktionen verwendet und mittels einer Fourier-Entwicklung die Funktion approximiert.
Während wir in Satz 3.13 zeigen konnten, dass die Projektion einer C^∞-Funktion in dem APK-Funktionenraum eine sehr gute punktweise Näherung darstellt (spektrale Konvergenz), weist die Entwicklung einer nur stückweise stetigen Funktion nicht solche exzel-

lenten Approximationseigenschaften auf. In diesem Kontext wird in der Literatur häufig vom **Gibbs'schen Phänomen** gesprochen. In seiner ursprünglichen Form wurde dabei eine stückweise stetige, nichtperiodische Funktion in ihre Fourier-Reihe bezüglich trigonometrischen Funktionen entwickelt und ihre Approximationseigenschaft analysiert. Dabei geht nicht nur die gleichmäßige Konvergenz der Fourier-Reihe verloren, sondern es kommt in der Nähe der Unstetigkeitsstellen zu Schwingungen, wie man am nachfolgenden Beispiel sofort sieht.

Beispiel 4.1. Wir betrachten die Funktion

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

Entwickelt man diese Funktion in ihre Fourier-Reihe bezüglich trigonometrischer Funktionen, so erhält man die Reihe $\frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin[(2n-1)x]}{2n-1}$.
Dargestellt in Abbildung 4.1 sind die Partialsummen für $N = 2,\ 20,\ 50$ Summanden. Man erkennt daran deutlich, dass die Reihe sowohl an den Rändern wie auch im Punkt

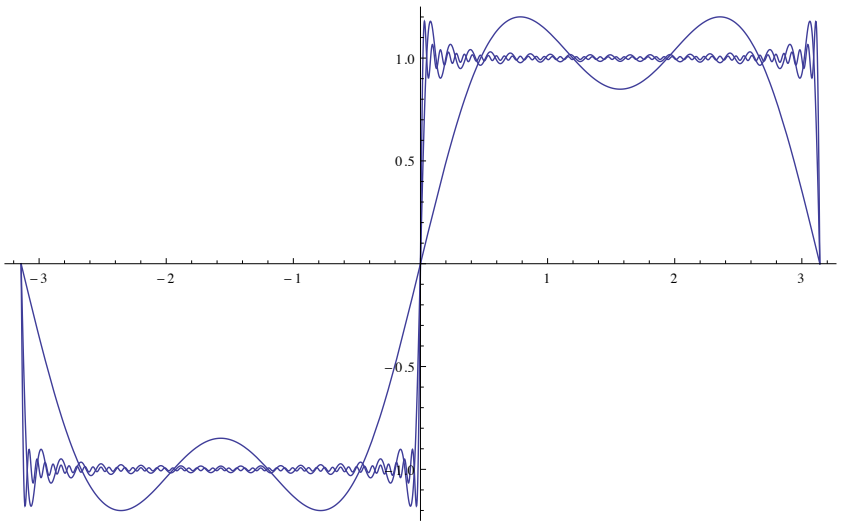


Abbildung 4.1: *Partialsummen mit* $N = 2,\ 20,\ 50$

$x = 0$ nicht gleichmäßig konvergiert. Die Oszillationen weisen für alle drei Fälle etwa die gleiche Amplitude auf. Dies ist auch nicht verwunderlich, da man durch analoge Analyse wie in [91, S.61] zeigen kann, dass die Höhe der Oszillationen circa 9% der Sprunghöhe misst. Die maximalen Über- und Unterschwingugen der Partialsummen konvergieren für $N \to \infty$ gegen den Wert $\pm\frac{2}{\pi} \int_0^{\pi} \frac{\sin t}{t} \approx \pm 1,18$.

Im Jahre 1899 veröffentlichte Gibbs das Resultat, dass die Oszillationen der Partialsummen gegen einen festen Wert konvergieren. Dabei korrigierte er eine von ihm ein Jahr zuvor getroffene Aussage, die das Abklingverhalten der Oszillationen proportional zu N^{-1} bezifferte. Allerdings war Gibbs nicht der erste, der sich mit dem Verhalten der

Fourier-Reihen von unstetigen und nicht periodischen Funktionen beschäftigte. Bereits 50 Jahre zuvor analysierte der englische Mathematiker Wilbraham diese Verhalten [85], jedoch blieb seine Arbeit lange Zeit unentdeckt, so dass sich der von Bôcher zuerst verwendete Begriff des Gibbs'schen Phänomens für den Verlust der gleichmäßigen Konvergenz der Fourier-Reihe durchsetzte [7]. Heute findet man des Öfteren vorrangig in der englischen Literatur [37] die Bezeichnung **Gibbs-Wilbraham-Phänomen**. In der Theorie spektraler Verfahren wird der Begriff auch für den globalen Genauigkeitsverlust der punktweisen Approximation verwendet, siehe [31].
Diesen Verlust, sowie den der gleichmäßigen Konvergenz und das oszillierende Verhalten, findet man nicht nur im Spezialfall bei der Entwicklung bezüglich trigonometrischer Funktionen, sondern allgemein bei der Fourier-Entwicklung einer unstetigen Funktion bezüglich anderer orthogonaler Polynome, wie beispielsweise für Chebyshev- und Legendre-Polynome ([41], [52]) bzw. allgemeiner Jacobi-Polynome [22].
Da sich bei hyperbolischen Erhaltungsgleichungen Unstetigkeiten in der Lösung entwickeln können, kommt es auch hier zum Gibbs'schen Phänomen, wenn man ein spektrales Verfahren verwendet. Dies wiederum bedeutet in unserem Falle, dass die APK-Entwicklung global keine hohe Ordnung besitzt und es zu Oszillationen in der Nähe der Unstetigkeitsstelle kommt. Mit beiden Problemen werden wir uns im nachfolgenden Abschnitt beschäftigen. Wir beginnen damit, den Umgang mit den Oszillationen genauer zu betrachten. Wir verwenden zu ihrer Abschwächung **modale Filter**.

4.1 Modale Filter

Eine der einfachsten und effektivsten Ansätze zur Reduktion des Gibbs'schen Phänomens ist die direkte Modifikation der Koeffizienten in der Reihenentwicklung. Beim Gibbs'schen Phänomen fallen die hochfrequenten Fourier-Koeffizienten nur sehr langsam ab und verursachen dadurch Oszillationen in der Nähe der Unstetigkeitsstellen. Die grundlegende Idee der modalen Filterung ist es, an die hochfrequenten Koeffizienten einen Dämpfungsfaktor/Filter direkt zu multiplizieren, um so die Schwingungen abzuschwächen bzw. bestmöglich zu entfernen. Jedoch besitzen die hochfrequenten Koeffizienten Informationen über die Unstetigkeit und ein zu starker Dämpfungsfaktor hat den Verlust dieser Informationen zur Folge, was wiederum zu einer schlechteren Approximation der Funktion führen würde. Daher ist es nötig, die Dämpfungsfaktoren genauer auf ihre Eigenschaften zu untersuchen bzw. bei der Konstruktion darauf zu achten, dass die Filter allen Anforderungen genügen.
Die ersten Abhandlungen, die sich mit der Konstruktion von passenden Dämpfungsfaktoren beschäftigten, lagen im Bereich der digitalen Signalverarbeitung, da dort häufig unstetige Funktionen durch Fourier-Summen approximiert werden und dementsprechend immer das Gibbs'sche Phänomen auftritt ([18], [79]). Ideen und Resultate aus der Signalverarbeitung wurden schließlich auf allgemeine spektrale Methoden übertragen und zur Verbesserung der Lösung für hyperbolische Erhaltungsgleichungen genutzt. Ihre Beschreibung findet man in der Literatur, vergleiche [11], [31] und [80].

Definition 4.2. Sei $p \in \mathbb{N}$ und $\sigma : [0,1] \to [0,1]$ eine $(p-1)$-mal stetig differenzierbare Funktion. Wir bezeichnen σ als **Filter der Ordnung** p, falls die Eigenschaften

$$\begin{aligned} \sigma(0) &= 1, \\ \sigma^{(k)}(0) &= 0, \qquad \forall 1 \leq k \leq p-1, \end{aligned} \tag{4.1}$$

erfüllt sind.
Weiterhin sei $N \in \mathbb{N}$, I_N eine Indexmenge und u_N die Rekonstruktion einer Funktion u in der Basis $\{\phi_k\}$, $k \in I_N$. Die Funktion $\eta : I_N \to [0,1]$ bilde jeden Index k auf $[0,1]$ ab. Man spricht von einem **modalen Filter**, wenn σ direkt auf die Koeffizienten der Reihenentwicklung wirkt, so dass die gefilterte Rekonstruktion die Form

$$u_N^\sigma(\mathbf{x}) := \sum_{k \in I_N} \sigma\left(\eta(k)\right) \hat{u}_k \phi_k(\mathbf{x})$$

besitzt.

Bemerkung. Die Funktion η dient lediglich dazu, die Indexmenge I_N abhängig von den Frequenzen bzw. dem Polynomgrad auf $[0,1]$ abzubilden. So wird üblicherweise $\eta(k) = \frac{|k|}{N}$ für die Fourier-Entwicklung bezüglich trigonometrischer Funktionen bzw. Chebyshev- und Legendre-Funktionen verwendet. In unserem Fall der APK-Polynome mit dem Polynomgrad $N = l + m$ ist $\eta\left((l,m)\right) = \frac{l+m}{N}$.

Zusätzlich zu den Eigenschaften (4.1) fordern viele Autoren, vergleiche [31] und [80], die nachfolgende zusätzliche Bedingung, dass für den Filter σ der Ordnung p

$$\sigma^{(l)}(1) = 0, \qquad 0 \leq l \leq p-1, \tag{4.2}$$

gilt und [36] verlangt dazu außerdem $\sigma \in C^p([0,1])$.
Die Bedingung (4.2) beschreibt, wie glatt der Übergang vom gefilterten Reihenabschnitt zum ungefilterten Teil ist und wird bei der Analyse des Approximationsverhalten zwischen der gefilterten und ungefilterten Reihendarstellung benötigt. Vandeven [80] nutzt dabei diese Bedingung als Erstes und kann mit ihr für eine stückweise glatte Funktion u folgendes Approximationsresultat zeigen.

Satz 4.3. *Es sei σ ein Filter der Ordnung p und $u : \mathbb{R} \to \mathbb{R}$ eine stückweise glatte, 2π-periodische Funktion. Dann gelten die folgenden Abschätzungen:*

(i) Sei $u \in C^p(\mathbb{R})$. Dann gilt

$$|u(x) - u_N^\sigma(x)| \leq C_1 \frac{1}{N^{p-\frac{1}{2}}}, \quad x \in \mathbb{R}, \tag{4.3}$$

mit einer Konstanten $C_1 \in \mathbb{R}$.

(ii) Erfüllt der Filter σ die zusätzliche Bedingung (4.2), besitzt u eine oder mehrere Sprungunstetigkeiten und ist weiterhin an der Stelle x stetig, so gilt folgende Abschätzungen

$$|u(x) - u_N^\sigma(x)| \leq C_2 \frac{1}{[d(x)]^{p-1} N^{p-1}}, \tag{4.4}$$

wobei $d(x)$ den Abstand von x zur nächstgelegenen Sprungstelle bezeichnet und $C_2 \in \mathbb{R}$ eine Konstante ist.

BEMERKUNG. Die Konstanten C_1 und C_2 sind unabhängig von x und N, jedoch abhängig von der gegebenen Funktion u sowie von den Filtern σ.

Gleichzeitig erweitert Vandeven sein Resultat auf Chebyshev-Reihen. Allgemein wurden bis jetzt analoge Resultate zu Satz 4.3 nur für spezielle Reihenentwicklung gezeigt. Die Arbeit [36] beinhaltet eine zu (4.3) vergleichbare Abschätzung für die Legendre-Entwicklung, [63] erweitert die Untersuchung auf mehrdimensionale Basisfunktionen (PKD-Basis) und zeigt ebenfalls ein zu (4.3) analoges Resultat. In [36] versuchen die Autoren ebenfalls, ein analoges Ergebnis zu der zweiten Aussage (4.4) zu beweisen, also für Funktionen u, welche nur stückweise glatt sind. Jedoch weist ihr Beweis einige Ungenauigkeiten auf, so dass wir uns am Ende dieses Kapitels nochmals explizit mit der Aussage (4.4) beschäftigen werden.
Bis heute sind die theoretischen Untersuchungen der Approximation durch modal gefilterten Fourier-Summen nicht abgeschlossen. Wir werden erstmals ein neues Teilergebnis zeigen. Wir vervollständigen das Ergebnis aus [63, S.69], indem wir es für allgemeine APK-Polynome beweisen. Die PKD-Polynome sind nur ein Spezialfall der APK-Polynome.

Satz 4.4. *Sei $u \in H^{2k}(\mathbb{T}, h) \cap C(\mathbb{T}), k \in \mathbb{N}$, $h(x,y) = x^{\alpha-1}y^{\beta-1}(1-x-y)^p$ mit $\alpha, \beta \in \mathbb{N}$ und $p \in \mathbb{N}_0$ und sei σ ein modaler Filter der Ordnung $2k-1$, mit der zusätzlichen Eigenschaft $\sigma \in C^{2k-1}([0,\varepsilon))$ in einem Teilintervall $[0,\varepsilon) \subset [0,1]$, $\varepsilon > 0$. Weiterhin sei $k > \max\left\{ \frac{3}{4} + \frac{3}{4}\alpha + \frac{p}{2}, \frac{5}{4} + \frac{p+\beta}{2}, \frac{1}{4} + \frac{3}{4}\alpha + \frac{\beta}{2} \right\}$.*
Dann gelten die punktweise Abschätzungen mit jeweiligen Konstanten $K_1 - K_6$:

(i) Für das Innere des Dreiecks $(x,y) \in \mathring{\mathbb{T}}$ gilt

$$|u(x,y) - u_N^\sigma(x,y)| \leq K_1 \frac{1}{N^{2k-\frac{7}{4}}}.$$

(ii) Auf der linken Kante $[0,y]$ mit $y \in [0,1]$ erhält man

$$|u(0,y) - u_N^\sigma(0,y)| \leq \begin{cases} K_2 \frac{1}{N^{2k-\frac{3}{2}\alpha-p-\frac{1}{2}}}, & \text{für } p > \beta - 1, \\ K_3 \frac{1}{N^{2k-\frac{3}{2}\alpha-\beta+\frac{1}{2}}}, & \text{für } p \leq \beta - 1. \end{cases}$$

(iii) Für die untere Kante $[x,0]$ mit $x \in (0,1)$ ist

$$|u(x,0) - u_N^\sigma(x,0)| \leq K_4 \frac{1}{N^{2k-\frac{3}{4}-\beta}}.$$

(iv) Für die Hypothenuse $[x,1-x]$ mit $x \in (0,1)$ gilt

$$|u(x,1-x) - u_N^\sigma(x,1-x)| \leq K_5 \frac{1}{N^{2k-\frac{7}{4}-p}}.$$

(v) Im Punkt $(1,0)$ *erhält man*

$$|u(1,0) - u_N^\sigma(1,0)| \leq K_6 \frac{1}{N^{2k+\alpha-\gamma-\frac{3}{2}}}.$$

Beweis. Mit Verwendung der potentiellen Selbstadjungiertheit des Operators D und da u hinreichend glatt ist, gilt für die Koeffizienten der APK-Reihenentwicklung die Gleichung (3.39) für $m + l > 0$. Damit, und mit der Reihendarstellung (3.40) für die Funktion u, folgt

$$|u(x,y) - u_N^\sigma(x,y)| = \left| \sum_{\substack{0 \leq l+m \leq N \\ l,m \ \in \mathbb{N}_0}} \left(1 - \sigma\left(\frac{l+m}{N}\right)\right) \tilde{u}_{m,l} A_{m,l}(x,y) + \sum_{\substack{l+m>N \\ l,m \ \in \mathbb{N}_0}} \tilde{u}_{m,l} A_{m,l}(x,y) \right|$$

$$= \left| \int_{\mathbb{T}} h(x_1,y_1) \left[S_N(x,y,x_1,y_1) + R_N(x,y,x_1,y_1)\right] D^k u(x_1,y_1) \, \mathrm{d}x_1 \, \mathrm{d}y_1 \right|,$$

wobei $S_N(x,y,x_1,y_1)$ und $R_N(x,y,x_1,y_1)$ definiert sind durch

$$S_N(x,y,x_1,y_1) = \sum_{\substack{1 \leq l+m \leq N \\ l,m \ \in \mathbb{N}_0}} \left(1 - \sigma\left(\frac{l+m}{N}\right)\right) \frac{A_{m,l}(x_1,y_1) A_{m,l}(x,y)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)} \lambda_{m,l}^k},$$

$$R_N(x,y,x_1,y_1) = \sum_{\substack{l+m>N \\ l,m \ \in \mathbb{N}_0}} \frac{A_{m,l}(x_1,y_1) A_{m,l}(x,y)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)} \lambda_{m,l}}.$$

Abschätzungen der Norm von R_N entnehmen wir dem Beweis des Satzes 3.13. Fasst man die Ergebnisse aus Tabelle 3.2 zusammen, so ergibt sich Tabelle 4.1.

Gebiet	$\max\{\binom{l+p}{l}, \binom{l+\beta-1}{l}\}$	$\|\|R_N(x,y,\cdot,\cdot)\|\|_{\mathrm{L}^2(\mathbb{T},h)} <$
$\mathring{\mathbb{T}}$		$C_{R_1} N^{-2k+\frac{7}{4}}$
$[0,y]$	$\binom{l+p}{l}$	$C_{R_{2(1)}} N^{-2k+\frac{3}{2}\alpha+p+\frac{1}{2}}$
$[0,y]$	$\binom{l+\beta-1}{l}$	$C_{R_{2(2)}} N^{-2k+\frac{3}{2}\alpha+\beta-\frac{1}{2}}$
$[x,0]$		$C_{R_3} N^{-2k+\frac{3}{4}+\beta}$
$[x,1-x]$		$C_{R_4} N^{-2k+\frac{7}{4}+p}$
$(1,0)$		$C_{R_5} N^{-2k-\alpha+\gamma+\frac{3}{2}}$

Tabelle 4.1: *Approximationstabelle*

Dabei sind $C_{R_1}, \ldots, C_{R_5} \in \mathbb{R}^+$ Konstanten.

Zum Abschätzen der S_N gehen wir analog zum Beweis von Satz 3.13 vor. Wir zeigen zuerst (i), also die Abschätzung für Punkte (x, y) im Inneren des Dreiecks. Es gilt

$$
\begin{aligned}
||S_N(x,y,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} &= \sum_{\substack{1\le l+m\le N\\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{A^2_{m,l}(x,y)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)}\lambda^{2k}_{m,l}} \\
&\overset{\substack{(3.23)\&\\(3.24)}}{\le} \sum_{\substack{1\le l+m\le N\\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{\kappa^2_{m,l}4(m+l+\gamma)^2\tilde{E}^2(x,y)}{\kappa^2_{l,m}(2l+\beta+p)^{\frac{1}{2}}(m+l)^{2k}(m+l+\gamma)^{2k}} \\
&\le 4\tilde{E}^2(x,y) \sum_{\substack{1\le l+m\le N\\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{1}{(m+l)^{2k+\frac{1}{2}}(m+l+\gamma)^{2k-2}} \\
&\overset{(3.44)}{<} 8\tilde{E}(x,y)^2 \sum_{i=1}^{N} \left(1-\sigma\left(\frac{i}{N}\right)\right)^2 \frac{1}{i^{4k-\frac{5}{2}}} \frac{N^{4k-\frac{5}{2}}}{N^{4k-\frac{5}{2}}} \\
&= 8\tilde{E}(x,y)^2 N^{-4k+\frac{7}{2}} \left(\frac{1}{N}\sum_{i=1}^{N}\left(1-\sigma\left(\frac{i}{N}\right)\right)^2\left(\frac{i}{N}\right)^{-4k+\frac{5}{2}}\right).
\end{aligned}
$$

Für $N \to \infty$ entspricht der letzte Faktor dem Integral

$$\int_0^1 (1-\sigma(\tau))^2 \tau^{\frac{5}{2}-4k} \mathrm{d}\tau.$$

Unter den gegebenen Voraussetzungen ist dieses Integral immer beschränkt. Um dies zu verifizieren entwickeln wir die Funktion σ in ihre Taylor-Reihe im Punkt $\tau_0 = 0$. Es gilt

$$\sigma(\tau) = \sum_{j=0}^{2k-1} \frac{1}{j!}\sigma^{(j)}(0)\tau^j + \mathrm{o}(\tau^{2k-1}) \quad \forall \tau \in [0,\varepsilon).$$

Wegen $\sigma(0) = 1$ und $\sigma^{(j)}(0) = 0$ für alle $j = 1, 2, \ldots, 2k-2$ reduziert sich der Ausdruck zu

$$\sigma(\tau) = 1 + \frac{1}{(2k-1)!}\sigma^{(2k-1)}(0)\tau^{2k-1} + \mathrm{o}(\tau^{2k-1})$$

für alle $\tau \in [0, \varepsilon)$. Setzen wir dies in das Integral ein, so ergibt sich

$$\int_0^{\varepsilon} \left(\frac{\sigma^{(2k-1)}(0)}{(2k-1)!}\tau^{2k-1} + \mathrm{o}(\tau^{2k-1})\right)^2 \tau^{-4k+\frac{5}{2}} \mathrm{d}\tau + \int_{\varepsilon}^1 (1-\sigma(\tau))^2 \tau^{\frac{5}{2}-4k} \mathrm{d}\tau < C^2_{I_1},$$

da jeder Teil beschränkt ist. Daher gilt

$$||S_N(x,y,\cdot,\cdot)||_{\mathrm{L}^2(\mathbb{T},h)} < \sqrt{8}\tilde{E}(x,y)N^{-2k+\frac{7}{4}}C_{I_1} = C_{S_1}N^{-2k+\frac{7}{4}},$$

und mit der Schwarz'schen Ungleichung insgesamt

$$\begin{aligned}|u(x,y)-u_N^\sigma(x,y)| \leq & ||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} \left(||R_N(x,y,\cdot,\cdot)||_{\mathrm{L}^2(\mathbb{T},h)} + ||S_N(x,y,\cdot,\cdot)||_{\mathrm{L}^2(\mathbb{T},h)}\right) \\ \leq & \underbrace{(||D^k u||_{\mathrm{L}^2(\mathbb{T},h)})(C_{R_1}+C_{S_1})}_{K_1} N^{-2k+\frac{7}{4}}.\end{aligned}$$

Zum Abschätzen der Norm S_N an der linken Kante $[0,y]$ betrachten wir zuerst den Fall $p > \beta - 1$. Es ergibt sich

$$\begin{aligned}||S_N(0,y,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} &= \sum_{\substack{1\leq l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{A^2_{m,l}(0,y)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)}\lambda^{2k}_{m,l}} \\ &\overset{\substack{(3.23)\&\\(3.25)}}{\leq} \sum_{\substack{1\leq l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{\binom{l+p}{l}^2\binom{m+\alpha-1}{m}^2 4(m+l+\gamma)^2\kappa^2_{m,l}}{(m+l)^{2k}(m+l+\gamma)^{2k}} \\ &= 4\sum_{\substack{1\leq l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{\binom{l+p}{l}^2\binom{m+\alpha-1}{m}^2}{(m+l)^{2k}(m+l+\gamma)^{2k-2}} \frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p m_\alpha (m+a_l)} \\ &\overset{\substack{(3.42)\&\\(3.43)}}{<} 4\beta^p \sum_{\substack{1\leq l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{(m+\alpha-1)^{2\alpha-2}(l+p)^{2p}}{(m+l)^{2k}(m+l+\gamma)^{2k-2}} \left(1+\frac{a_l}{m+1}\right)^{\alpha-1}.\end{aligned}$$

Mit $\gamma > \alpha - 1$ und $a_l + 1 < 2(m+l+\gamma)$ gilt weiter

$$\begin{aligned}&\leq 2^{\alpha+1}\beta^p \sum_{\substack{1\leq l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{1}{(m+l+\gamma)^{2k+1-2p-3\alpha}(m+l)^{2k}} \\ &\overset{(3.44)}{<} 2^{\alpha+1}\beta^p \sum_{i=1}^{N} \left(1-\sigma\left(\frac{i}{N}\right)\right)^2 \frac{1}{(i+\gamma)^{2k-2p-3\alpha}(i)^{2k}} \frac{N^{4k-3\alpha-2p}}{N^{4k-3\alpha-2p}} \\ &< 2^{\alpha+1}\beta^p N^{-4k+3\alpha+2p+1} \left(\frac{1}{N}\sum_{i=1}^{N} \left(1-\sigma\left(\frac{i}{N}\right)\right)^2 \left(\frac{i^{-2k}(i+\gamma)^{-2k+3\alpha+2p}}{N^{-4k+3\alpha+2p}}\right)\right).\end{aligned}$$

Wenn $\varsigma_1 := -2k+3\alpha+2p \leq 0$ ist, wird die Summe nach oben abgeschätzt, indem man im $(i+\gamma)$-ten Faktor γ einfach weglässt. Für $\varsigma_1 > 0$ gilt

$$(i+\gamma)^{\varsigma_1} = i^{\varsigma_1}\left(\frac{i+\gamma}{i}\right)^{\varsigma_1} \leq i^{\varsigma_1}(1+\gamma)^{\varsigma_1}.$$

Insgesamt erhalten wir für S_N

$$||S_N(0,y,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} < 2^{\alpha+1}\beta^p N^{-4k+3\alpha+2p+1} C_{\varsigma_1} \cdot \left(\frac{1}{N}\sum_{1\le i\le N}\left(1-\sigma\left(\frac{i}{N}\right)\right)^2\left(\frac{i}{N}\right)^{-4k+3\alpha+2p}\right),$$

mit

$$C_{\varsigma_1} = \begin{cases} 1, & \text{für} \quad \varsigma_1 \le 0, \\ (1+\gamma)^{\varsigma_1}, & \text{für} \quad \varsigma_1 > 0. \end{cases}$$

Für $N \to \infty$ konvergiert die Riemann'sche Summe gegen das Integral

$$\lim_{N\to\infty}\left(\frac{1}{N}\sum_{1\le i\le N}\left(1-\sigma\left(\frac{i}{N}\right)\right)^2\left(\frac{i}{N}\right)^{-4k+3\alpha+2p}\right) = \int_0^1 (1-\sigma(\tau))^2\tau^{-4k+3\alpha+2p}\mathrm{d}\tau.$$

Durch eine Taylor-Entwicklung und Verwendung der Filtereigenschaften folgt mit entsprechender Argumentation wie zuvor, dass das Integral immer beschränkt ist, und damit

$$||S_N(0,y,\cdot,\cdot)||_{L^2(\mathbb{T},h)} < N^{-2k+\frac{3}{2}\alpha+p+\frac{1}{2}} C_{S_2}.$$

Mit der Schwarz'schen Ungleichung folgt

$$\begin{aligned} |u(0,y) - u_N^\sigma(0,y)| \le & ||D^k u||_{\mathrm{L}^2(\mathbb{T},h)} \left(||R_N(0,y,\cdot,\cdot)||_{\mathrm{L}^2(\mathbb{T},h)} + ||S_N(0,y,\cdot,\cdot)||_{\mathrm{L}^2(\mathbb{T},h)}\right) \\ \le & \underbrace{(||D^k u||_{\mathrm{L}^2(\mathbb{T},h)})(C_{R_{2(1)}} + C_{S_2})}_{K_2} N^{-2k+\frac{3}{2}\alpha+p+\frac{1}{2}} = K_2 N^{-2k+\frac{3}{2}\alpha+p+\frac{1}{2}}. \end{aligned}$$

Sei nun $\beta - 1 \ge p$. Man erhält für S_N mit einer analogen Rechnung zum vorherigen Fall

$$\begin{aligned} ||S_N(0,y,\cdot,\cdot)||^2_{L^2(\mathbb{T},h)} &= \sum_{\substack{1\le l+m\le N \\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{A^2_{m,l}(0,y)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)}\lambda^{2k}_{m,l}} \\ &\overset{\substack{(3.23)\&\\(3.25)}}{\le} \sum_{\substack{1\le l+m\le N \\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{\binom{l+\beta-1}{l}^2\binom{m+\alpha-1}{m}^2 4(m+l+\gamma)^2\kappa^2_{m,l}}{(m+l)^{2k}(m+l+\gamma)^{2k}} \\ &< 2^{\alpha+1}\beta^p \sum_{1\le i\le N}\left(1-\sigma\left(\frac{i}{N}\right)\right)^2 \frac{1}{i^{2k}(i+\gamma)^{2k-3\alpha-2\beta+2}} \frac{N^{4k-3\alpha-2\beta+2}}{N^{4k-3\alpha-2\beta+2}} \\ &< 2^{\alpha+1}\beta^p C_{\varsigma_2} N^{-4k-1+3\alpha+2\beta}\left(\frac{1}{N}\sum_{1\le i\le N}\left(1-\sigma\left(\frac{i}{N}\right)\right)^2\left(\frac{i}{N}\right)^{-4k+3\alpha+2\beta-2}\right), \end{aligned}$$

mit $\varsigma_2 := -2k + 3\alpha + 2\beta - 2$, und

$$C_{\varsigma_2} = \begin{cases} 1, & \text{für} \quad \varsigma_2 \le 0, \\ (1+\gamma)^{\varsigma_2}, & \text{für} \quad \varsigma_2 > 0. \end{cases}$$

Betrachtet man $N \to \infty$, erhält man erneut eine Riemann'sche Summe, und mit einer analogen Argumentation wie in den beiden vorherigen Fällen folgt, dass das Integral beschränkt ist, und somit

$$||S_N(0,y,\cdot,\cdot)||_{\mathrm{L}^2(\mathbb{T},h)} < N^{-2k+\frac{3}{2}\alpha+\beta-\frac{1}{2}} C_{S_3}$$

folgt. Mit der Schwarz'schen Ungleichung und den Abschätzungen für R_N ergibt sich dann schließlich

$$|u(0,y) - u_N^\sigma(0,y)| < K_3 N^{-2k+\frac{3}{2}\alpha+\beta-\frac{1}{2}}.$$

Auf der unteren Kante $[x,0]$, der Hypothenuse $[x,1-x]$ sowie im Punkt $(1,0)$ wird analog argumentiert. Man schätzt $||S_N||_{\mathrm{L}^2(\mathbb{T},h)}$ für alle drei Fälle mit den Abschätzungen aus Kapitel 3 ab und gelangt jeweils zu einer Riemann'schen Summe. Für $N \to \infty$ konvergieren sie gegen die Integrale. Man verifiziert die Beschränktheit der Integrale, indem man die Eigenschaften des Filters σ ausnutzt und diesen im Nullpunkt in seine Taylor-Reihe entwickelt. Die meisten Terme sind 0 und man erkennt sofort, dass die Integrale in jedem der Fälle beschränkt sind. Mit den Abschätzungen für die R_N (siehe Tabelle 3.2) und der Schwarz'schen Ungleichung sind schließlich (iii)-(v) gezeigt. Im Folgenden werden wir daher nur noch die Rechnungen bezüglich der $||S_N||_{\mathrm{L}^2(\mathbb{T},h)}$ angeben und die finale Abschätzung. Für die untere Kante $[x,0]$ gilt für $||S_N||_{\mathrm{L}^2(\mathbb{T},h)}$:

$$\begin{aligned}
||S_N(x,0,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} &= \sum_{\substack{1\le l+m\le N\\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{A^2_{m,l}(x,0)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)}\lambda^{2k}_{m,l}} \\
&\overset{\substack{(3.23)\&\\(3.26)}}{\le} \sum_{\substack{1\le l+m\le N\\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{4(m+l+\gamma)^2\kappa^2_{m,l}}{(m+l)^{2k}(m+l+\gamma)^{2k}} \\
&\qquad \cdot \left(\frac{(m)_\alpha(m+a_l)}{(m+a_l)_\alpha m}\right)\binom{l+\beta-1}{l}^2 \frac{C^2}{(2(m+l)+\gamma)^{\frac{1}{2}}(x)^{\alpha-\frac{1}{2}}(1-x)^{\frac{1}{2}+\gamma-\alpha}2} \\
&\overset{(3.42)}{<} 2\beta^p \frac{C^2}{x^{\alpha-\frac{1}{2}}(1-x)^{\frac{1}{2}+\gamma-\alpha}} \sum_{\substack{1\le l+m\le N\\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{(l+\beta-1)^{2\beta-2}}{(m+l+\gamma)^{2k-\frac{3}{2}}(m+l)^{2k}} \\
&\overset{(3.44)}{<} 2\beta^p \frac{C^2}{x^{\alpha-\frac{1}{2}}(1-x)^{\frac{1}{2}+\gamma-\alpha}} \sum_{\substack{1\le l+m\le N\\ l,m\ \in\mathbb{N}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{1}{(i+\gamma)^{2k-2\beta+\frac{1}{2}}i^{2k}} \\
&< C_{S_3} \sum_{i=1}^{N} \left(1-\sigma\left(\frac{i}{N}\right)\right)^2 \frac{1}{i^{4k-2\beta-\frac{1}{2}}} \frac{N^{4k-2\beta-\frac{1}{2}}}{N^{4k-2\beta-\frac{1}{2}}} \\
&< C_{S_3} N^{-4k+2\beta+\frac{3}{2}} \left(\frac{1}{N}\sum_{i=1}^{N}\left(1-\sigma\left(\frac{i}{N}\right)\right)^2\left(\frac{i}{N}\right)^{-4k+2\beta+\frac{1}{2}}\right),
\end{aligned}$$

dabei ist die Konstante C_{S_3} von x und von der Tatsache abhängig, ob $2k-2\beta-\frac{1}{2}$ größer oder kleiner Null ist. Allgemein gilt dann

$$|u(x,0)-u_N^\sigma(x,0)| < K_4 N^{-2k+\beta+\frac{3}{4}}.$$

Für die Hypothenuse $[x,1-x]$ erhalten wir

$$\begin{aligned}
||S_N(x,1-x,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} &= \sum_{\substack{1\le l+m\le N\\ l,m\ \in\overline{\mathbb{N}}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{A^2_{m,l}(x,1-x)}{||A_{m,l}||^2_{\mathrm{L}^2(\mathbb{T},h)}\lambda^{2k}_{m,l}} \\
&\overset{\substack{(3.23)\&\\(3.26)}}{\le} \sum_{\substack{1\le l+m\le N\\ l,m\ \in\overline{\mathbb{N}}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{4(m+l+\gamma)^2}{(m+l)^{2k}(m+l+\gamma)^{2k}} \frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p (m)_\alpha (m+a_l)} \\
&\quad \cdot \left(\frac{C}{(2(m+l)+\gamma)^{\frac{1}{4}}(x)^{\frac{1}{4}+\frac{\alpha-1}{2}}(1-x)^{\frac{1}{4}+\frac{\gamma-\alpha}{2}}\sqrt{2}} \left(\frac{(m)_\alpha(m+a_l)}{(m+a_l)_\alpha m}\right)^{\frac{1}{2}} \binom{l+p}{l}\right)^2 \\
&\overset{\substack{(3.42)\&\\(3.45)}}{<} \frac{C^2}{x^{\alpha-\frac{1}{2}}(1-x)^{\frac{1}{2}+\gamma-\alpha}} 2\beta^p \sum_{\substack{1\le l+m\le N\\ l,m\ \in\overline{\mathbb{N}}_0}} \left(1-\sigma\left(\frac{l+m}{N}\right)\right)^2 \frac{(l+m+\gamma)^{2p}}{(m+l)^{2k}(m+l+\gamma)^{2k-\frac{3}{2}}} \\
&\overset{(3.44)}{<} \frac{C^2}{x^{\alpha-\frac{1}{2}}(1-x)^{\frac{1}{2}+\gamma-\alpha}} 2\beta^p \sum_{i=1}^{N} \left(1-\sigma\left(\frac{i}{N}\right)\right)^2 \frac{1}{(i)^{2k}(i+\gamma)^{2k-2p-\frac{5}{2}}} \frac{N^{4k-2p-\frac{5}{2}}}{N^{4k-2p-\frac{5}{2}}} \\
&< C_{S_4} N^{-4k+2p+\frac{7}{2}} \left(\frac{1}{N}\sum_{i=1}^{N}\left(1-\sigma\left(\frac{i}{N}\right)\right)^2 \left(\frac{i}{N}\right)^{-4k+2p+\frac{5}{2}}\right)
\end{aligned}$$

und insgesamt

$$|u(x,1-x)-u_N^\sigma(x,1-x)| < K_5 N^{-2k+\frac{7}{4}+p}.$$

Kommen wir noch zum Punkt $(1,0)$. Hier gilt

$$\begin{aligned}
||S_N(1,0,\cdot,\cdot)||^2_{\mathrm{L}^2(\mathbb{T},h)} &\overset{(3.50)}{=} \sum_{m=1}^{N} \left(1-\sigma\left(\frac{m}{N}\right)\right)^2 \frac{A^2_{m,0}(1,0)}{||A_{m,0}||^2_{\mathrm{L}^2(\mathbb{T},h)}\lambda^{2k}_{m,0}} \\
&\overset{\substack{(3.23)\&\\(3.42)}}{<} 4\left(\frac{(\beta)_p \gamma^{\alpha-1}}{(1)_p}\right) \sum_{m=1}^{N} \left(1-\sigma\left(\frac{m}{N}\right)\right)^2 \frac{\binom{m+\gamma-\alpha}{m}^2}{m^{2k}(m+\gamma)^{2k-2}} \\
&< C_5 \sum_{m=1}^{N} \left(1-\sigma\left(\frac{m}{N}\right)\right)^2 \frac{1}{m^{2k}(m+\gamma)^{2k-2-2\gamma+2\alpha}} \frac{N^{4k-2-2\gamma+2\alpha}}{N^{4k-2-2\gamma+2\alpha}} \\
&< C_{S_5} N^{-4k+3+2\gamma-2\alpha} \left(\frac{1}{N}\sum_{m=1}^{N}\left(1-\sigma\left(\frac{m}{N}\right)\right)^2 \left(\frac{m}{N}\right)^{-4k+2+2\gamma-2\alpha}\right),
\end{aligned}$$

und schließlich folgt

$$|u(1,0) - u_N^\sigma(1,0)| < K_6 N^{-2k-\alpha+\gamma+\frac{3}{2}}.$$

□

Unsere Untersuchungen sind ausschließlich für modale gefilterte Reihenentwicklungen von hinreichend glatten Funktionen gültig. Hinsichtlich des Approximationsverhaltens für Funktionen mit Sprungstellen können auch wir keine Aussage treffen. Diesbezüglich müssen noch weitere Arbeiten folgen. Gerade im Hinblick auf die Verwendung von spektralen Methoden bei hyperbolischen Erhaltungsgleichungen und der Entwicklung von Lösungen mit Sprungunstetigkeiten hat man an theoretischen Resultaten Interesse. Ergebnisse dazu können zu Kriterien führen, die es beispielsweise ermöglichen, die Auswahl des Filters, der Filterstärke und/oder der Ordnung zu optimieren, um eine genauere Lösung zu erhalten. So sind auch die Konstanten $K_1, \ldots, K_6$ aus dem Satz 4.4 von dem gewählten Filter und der Funktion u abhängig. Man findet als Beispiele modaler Filter σ folgende Funktionen in der Literatur, siehe [31]:

1) **Fejèr-Filter** erster Ordnung

$$\sigma_1(\eta) = 1 - \eta,$$

2) **Lanczos-Filter** erster Ordnung

$$\sigma_2(\eta) = \frac{\sin(\pi\eta)}{\pi\eta},$$

3) **raised-cosine-Filter** von zweiter Ordnung

$$\sigma_3(\eta) = \frac{1}{2}\left(1 + \cos(\pi\eta)\right),$$

4) **shaped-raised-cosine-Filter** von achter Ordnung

$$\sigma_4(\eta) = \sigma_3^4(\eta)(35 - 84\sigma_3(\eta) + 70\sigma_3^2(\eta) - 20\sigma_3^3(\eta)),$$

5) **Exponentialfilter** p-ter Ordnung

$$\sigma_5(\eta) = e^{-\alpha\eta^p}.$$

Die ersten vier Filter erfüllen alle die Zusatzbedingung (4.2), der Exponentialfilter jedoch nicht, da $\sigma_5(1) = e^{-\alpha}$ gilt. Um für den Exponentialfilter die Zusatzbedingung zu gewährleisten, wählt man daher üblicherweise die Filterstärke α so, dass $e^{-\alpha}$ im Bereich der Rechnergenauigkeit liegt.
Alle diese Filter finden in der digitalen Signalverarbeitung Verwendung und sind Tiefpassfilter. Sie filtern die Impulsantwort eines Signals. Dabei wurden sie in der Literatur

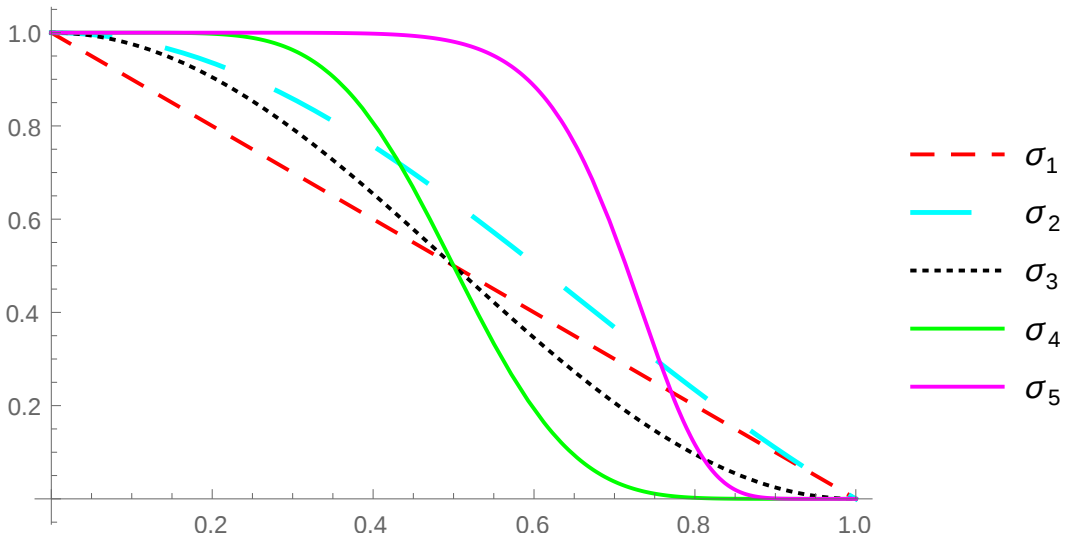


Abbildung 4.2: *Filterfunktionen* $\sigma_1, \ldots, \sigma_5$

schon auf die Anwendung bei spektralen Methoden angepasst. So findet man als Definition für den allgemeinen Lanczos-Filter folgende Funktion:

$$\tilde{\sigma}_2(\eta) = \begin{cases} \dfrac{a \sin(\pi\eta) \sin\left(\pi \frac{\eta}{a}\right)}{(\pi\eta)^2}, & \text{wenn } -a < \eta < a,\ a \neq \eta, \\ 1, & \text{für } \eta = 0, \\ 0, & \text{ansonsten,} \end{cases}$$

wobei a die Größe des Trägers bezeichnet bzw. die Frequenzbreite. Wir werden in den numerischen Untersuchungen die Filter $\sigma_1, \ldots, \sigma_5$ verwenden, um ein Gespür für den Einfluss der Filter auf die Approximation zu erhalten. Allgemein gilt: Je höher die Ordnung des Filters, desto geringer werden die niedrigfrequenten Koeffizienten verändert. Aber das reicht nicht aus, um zu entscheiden, welchen Filter man intuitiv verwenden sollte. Der nächste Abschnitt gibt allerdings ein hilfreiches Indiz.

4.1.1 Die spektrale Viskositätsmethode

Wir stellen in diesem Abschnitt einen weiteren Ansatz zur Reduzierung der Gibbs'schen Oszillation vor, der auch einen engen Zusammenhang zur modalen Filterung aufweist. Bei diesem handelt es sich um die **Spektrale Viskositätsmethode** (kurz: SV-Methode oder SSV-Methode[1]). Dabei ist die grundlegende Idee, einen sehr kleinen Viskositätsterm zur Erhaltungsgleichung zu addieren, und die so entstandene Gleichung mit einem

[1] In der englischen Literatur unterscheidet man zwischen **spectral-viscosity-method** und **super-spectral-viscosty-method**. Die Unterscheidung hängt von dem Viskositätsterm ab, welchen man addiert. Beide Methoden werden unter dem Begriff „Spektrale Viskositätsmethode" zusammengefasst.

spektralen Verfahren zu lösen [81]. Hierzu ist festzuhalten, dass Tadmor in [74] zeigt, dass spektrale Methoden nicht zwangsläufig gegen die eindeutige Entropielösung konvergieren müssen und dies im Allgemeinen auch nicht tun. Seine Überlegungen beruhen auf der Burgers-Gleichung und der Fourier-Methode.
Er untersucht dabei nicht mehr

$$\frac{\partial}{\partial t}u_N(x,t) + \frac{\partial}{\partial x}P_N f(u_N(x,t)) = 0,$$

wobei $u_N = \sum_{|k|\leq N} \hat{u}_k e^{ikx}$ und P_N die Fourier-Projektion in den Raum der trigonometrischen Funktionen ist, sondern betrachtet

$$\frac{\partial}{\partial t}u_N(x,t) + \frac{\partial}{\partial x}P_N f(u_N(x,t)) = \varepsilon_N(-1)^{p+1}\frac{\partial^p}{\partial x^p}\left[Q_N \frac{\partial^p}{\partial x^p}u_N(x,t)\right] \tag{4.5}$$

mit dem Operator Q_N, der definiert ist durch

$$Q_N u_N := \sum_{|k|\leq N} \hat{Q}_k \hat{u}_k e^{ikx} \text{ mit } \hat{Q}_k \in [0,1].$$

Die $\hat{Q}_k$ bilden eine Art Glättungsfaktor. Sie haben Einfluss auf die jeweiligen Frequenzen und wie stark diese modifiziert werden. Die niedrigen Frequenzen lässt man viskositätsfrei und entsprechend wählt man $\hat{Q}_k = 0$ für $|k| \leq m < N$, wobei der Parameter m von N abhängig ist. Der Parameter p gibt die Ordnung des Viskositätsterms an. Für $p = 1$ spricht man von **spektraler Viskosität** und für $p > 1$ von **superspektraler Viskosität**.
In Gleichung (4.5) gibt es mehrere Parameter, und ihre Wahl wird entscheidend sein für das Verhalten der Lösung von (4.5), auch hinsichtlich der Ursprungsgleichung. Tadmor beweist in [75] den nachfolgenden Satz, der das Konvergenzverhalten der SV-Methode für den skalaren eindimensionalen Fall beschreibt.

Satz 4.5. *Sei $p \in \mathbb{N}$ eine feste Filterordnung und die Filterstärke gehorche der Ungleichung*

$$C_p \leq \sum_{k=1}^{p} ||\partial_u^k f(u)||_{L^\infty} ||u_N||_{L^\infty}^{k-1}. \tag{4.6}$$

Die Parameter $\Theta < \frac{2p-1}{2p}$ und $m_N \sim N^\Theta$ seien gewählt. Für die Viskositätsstärke ε_N gelte

$$\varepsilon_N \sim 2\frac{2C_p}{N^{2p-1}},$$

und für die Viskositätskoeffizienten

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

Dann gilt:
Sind die Lösungen u_N von Gleichung (4.5) mit obiger Parameterwahl gleichmäßig beschränkt, dann konvergieren die u_N für $N \to \infty$ stark gegen die eindeutige Entropielösung der Erhaltungsgleichung

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

mit konvexer Flussfunktion f sowie 2π-periodischen Anfangsbedingungen

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

und periodischen Randbedingungen.

Weitere Arbeiten übertragen dieses Resultat auf mehrdimensionale Erhaltungsgleichungen mit periodischen Anfangsbedingungen [13] und auch auf andere spektrale Methoden, wie die Legendre-SV-Methode [54]. Dennoch sind hier noch einige offene Fragen zu klären, da für Systeme der Ansatz über kompensierte Kompaktheit [16] nicht mehr genutzt werden kann. Auch schon bei skalaren Gleichungen und der Jacobi-SV-Methode führt eine Argumentation über kompensierte Kompaktheit zu keinem Ergebnis, da man für die Konvergenzaussage eine Abschätzung des Energiefunktionals $||u_N^2||_{\mathrm{L}^2} < C$ (für ein konstantes $C \in \mathbb{R}^+$) benötigt, und diese für allgemeine Parameterwahl $\alpha, \beta > -1$ der Jacobi-Polynome nicht existiert.
Auf diesem Gebiet sind somit noch viele interessante Studien zu führen und offene Fragen zu beantworten, allerdings ist dies nicht Inhalt dieser Arbeit.
Wir gehen im Folgenden immer davon aus, dass die Spektrale Viskositätsmethode gegen die eindeutig bestimmte Entropielösung konvergiert. Was diese Methode für unsere Zwecke so interessant macht ist die Tatsache, dass die SV- bzw. SSV-Methode äquivalent zu einer spektralen Methode mit modaler Filterung ist. Wir zeigen:

Satz 4.6. *Sei $N \in \mathbb{N}$, I_N eine Indexmenge und $\{\phi_k | k \in I_N\}$ die APK-Polynome auf dem Dreieck $\mathbb{T}$. Diese erfüllen das Eigenwertproblem*

$$\mathcal{D}\phi_k = -\lambda_k \phi_k \text{ mit } \mathcal{D} = -D, \tag{4.7}$$

wobei D den Differentialoperator (3.11) bezeichnet. Sei ferner P_ϕ die Projektion auf den von $\{\phi_k\}$ aufgespannten Raum. Dann ist das Lösen der viskosen Gleichung

$$\frac{\partial}{\partial t}u_N(\boldsymbol{x},t) + \nabla \cdot P_\phi f(u_N(\boldsymbol{x},t)) = \varepsilon_N (-1)^{p+1} \mathcal{D}^p u_N(\boldsymbol{x},t) \tag{4.8}$$

äquivalent zur Multiplikation der Koeffizienten $\hat{u}_k$ mit der Funktion

$$\sigma(k) = e^{-\varepsilon_N \Delta t \lambda_k^p}$$

nach jedem Aktualisierungsschritt der nicht viskosen Gleichung.

Beweis. Die Gleichung (4.8) wird durch ein Splitting-Verfahren in zwei Schritten gelöst,

$$\frac{\partial}{\partial t}u_N(\mathbf{x},\mathbf{t}) = \varepsilon_N(-1)^{p+1}\mathcal{D}^p u_N(\mathbf{x},t) \tag{4.9}$$

und

$$\frac{\partial}{\partial t}u_N(\mathbf{x},t) + \nabla \cdot P_\phi f(u_N(\mathbf{x},t) = 0. \tag{4.10}$$

Mit $u_N(\mathbf{x},t) = \sum\limits_{k\in I_N} \hat{u}_k\phi_k(\mathbf{x})$ folgt aus (4.9)

$$\sum_{k\in I_N}\frac{\partial \hat{u}_k}{\partial t}\phi_k(\mathbf{x}) = \sum_{k\in I_N}\varepsilon_N(-1)^{p+1}\hat{u_k}\mathcal{D}^p\phi_k(\mathbf{x}) = \sum_{k\in I_N} -\varepsilon_N\hat{u}_k(t)\lambda_k^p\phi(\mathbf{x}),$$

wobei im letzten Gleichheitszeichen die Eigenwertgleichung (4.7) verwendet wurde. Vergleicht man die Koeffizienten, muss man daher allein folgende gewöhnliche Differentialgleichung

$$\frac{\partial \hat{u}_k(t)}{\partial t} = -\varepsilon_N\hat{u_k}(t)\lambda_k^p, \quad \forall k \in I_N$$

lösen. Die Lösung ist $\hat{u}_k(t) = ce^{-\varepsilon_N\lambda_k^p t}$, $c \in \mathbb{R}$. Mit $\Delta t := t^{n+1} - t^n$ und der Forderung $\hat{u}_k(t^{n+1}) = \hat{u}_k(t^n)$ für $\Delta t = 0$ folgt schließlich

$$\hat{u}_k(t^{n+1}) = e^{-\varepsilon_N\lambda_k^p(\Delta t + t^n)} = \underbrace{e^{(-\varepsilon_N\lambda_k^p\Delta t)}}_{=:\sigma(k)}\hat{u}_k(t^n).$$

□

Damit wir von $\sigma(k) = \sigma((l,m)) = e^{-\varepsilon_N(l+m)^p(l+m+\gamma)^p\Delta t}$ als modalen Filter sprechen können, erweitern wir den Exponenten mit N^{2p}, d.h.

$$\sigma\left(\frac{l+m}{N}\right) = e^{-\varepsilon_N N^{2p}\left(\frac{l+m}{N}\right)^p\left(\frac{l+m+\gamma}{N}\right)^p\Delta t} \approx e^{-\varepsilon_N N^{2p}\Delta t\left(\frac{l+m}{N}\right)^{2p}}. \tag{4.11}$$

Dabei ist $\sigma : [0,1] \to [0,1]$ und man kann σ als Exponentialfilter der Ordnung $2p$ mit Filterstärke $\alpha_i := -\varepsilon_N N^{2p}\Delta t$ auffassen. Betrachtet man (4.11) genauer, so erkennt man, dass der modale Filter vom Parameter γ abhängig ist, vor allem wenn die Polynome kleinen Grad besitzen. Für verschiedene Familien von APK-Polynomen erhält man daher auch abhängig vom Parameter γ verschiedene Filter. Dies werden wir später bei den numerischen Testfällen noch genauer analysieren, auch im Hinblick auf die Stabilität des Verfahrens. So ergeben sich durchaus andere Resultate für verschiedene APK-Familien. Mehr dazu im Kapitel 6.

Die Aussage des Satzes 4.8 kann man analog für jede orthogonale Basis $\{\phi_k\}$ tätigen, solange die ϕ_k eine Eigenwertgleichung erfüllen, da man im Beweis nur eben jene Eigenschaft benötigt.

Die Übertragung der Spektralen Viskositätsmethode und ihrer Formulierungen auf das von uns verwendete Spektrale-Differenzen-Verfahren wird analog zu [86, S.51] vollzogen. Die Transformationen T_i vom Element τ_i in das Standarddreieck $\mathbb{T}$ bewirkt dabei

keinerlei Veränderung in der Funktion (4.11). Das SD-Aktualisierungsschema lässt sich somit für eine skalare Erhaltungsgleichung in einem Element $\tau_i \in \mathcal{T}$ mit Viskositätsterm vergleichbar zu (4.8) mit dem Lösungspunkt $\boldsymbol{\xi}_j \in \mathbb{T}$ wie folgt formulieren:

$$\begin{aligned} &\frac{\partial}{\partial t} u_N(T_i^{-1}(\boldsymbol{\xi}_j, t)) + \nabla_{\boldsymbol{\xi}} \cdot \tilde{P}_N(J_{T_i})^T \tilde{\mathcal{F}}(u_N(T_i^{-1}(\boldsymbol{\xi}_j), t)) \\ =&\varepsilon_N (-1)^{p+1} \mathcal{D}^p (u_N(T_i^{-1}(\boldsymbol{\xi}_j), t)). \end{aligned} \tag{4.12}$$

Dabei ist $\tilde{\mathcal{F}}$ die entsprechende Flussfunktion, $\tilde{P}_N$ die Projektion auf den APK-Polynomraum und $\nabla_{\boldsymbol{\xi}}$ der Nabla-Operator[2] bezüglich der Raumkomponenten in $\mathbb{T}$.
Mit einem analogen Vorgehen wie im Beweis von Satz 4.6 folgt die Äquivalenz von (4.12) und dem Spektrale-Differenzen-Verfahren mit modaler Filterung. Der modale Filter ist der Exponentialfilter aus Gleichung (4.11). Der Nutzen dieser Äquivalenzbeziehung liegt darin, dass modale Filterung meist effizienter implementiert werden kann als der Viskositätsansatz, so dass wir im Code des Spektrale-Differenzen-Verfahrens den modalen Exponentialfilter (4.11) verwenden. Für die Wahl der Filterordnung und Viskositätstärke ε_N folgen wir den Ideen aus [63] und halten uns an die Abschätzungen aus Satz 4.5. Dabei sei $\varepsilon_N \sim \frac{C_p}{N^{2p-1}}$ mit einer von p abhängigen Konstanten, die als obere Schranke den Wert

$$C_p \leq \sum_{k=1}^{p} ||\partial_u^k (J_{T_i})^T \tilde{\mathcal{F}}(u)||_{\mathrm{L}^\infty} = ||(J_{T_i})^T||_\infty \sum_{k=1}^{p} ||\partial_u^k \tilde{\mathcal{F}}(u)||_{\mathrm{L}^\infty}$$

besitzt. Die Abhängigkeit von u vernachlässigen wir jedoch. Aus der Koordinatentransformation (2.11) und mit einem Längenmaß h_i des Dreiecks τ_i folgert man

$$||(J_{T_i})^T||_\infty = \frac{1}{2V_i} \max\{|x_{1,1}| + |x_{1,2}|, |x_{2,1}| + |x_{2,2}|\} \sim \frac{1}{h_i}.$$

Damit ist die Norm proportional zum Kehrwert des Längenmaßes und man wählt daher auch C_p sinnvollerweise proportional dazu, was schließlich auf

$$\varepsilon_N^i := \frac{c}{h_i N^{2p-1}} < \frac{\sum_{k=1}^{p} ||\partial_u^k \tilde{\mathcal{F}}(u)||_{\mathrm{L}^\infty}}{h_i N^{2p-1}} \tag{4.13}$$

führt, wobei $c \in \mathbb{R}$ ein vom Testfall und Filterordnung abhängiger Parameter ist.

4.1.2 Stoßindikator

Bei einer globalen Anwendung der modalen Filter kommt es zum Abfall der Ordnung des Verfahrens und damit auch der Konvergenzraten. Um diesen Verlust zu vermeiden,

[2] Es sei nochmals auf das Grundlagenkapitel 2.2.2 verwiesen. Man beachte hierbei, dass sich die Flussfunktion aus zwei Komponenten zusammensetzt, abhängig davon, ob der Punkt $\boldsymbol{\xi}_j$ auf dem Rand oder im Inneren von τ_i liegt. Weiterhin gilt $\nabla_{\mathbf{x}} = J_{T_i} \nabla_{\boldsymbol{\xi}}$ für $\mathbf{x} \in \tau_i$.

filtert man nicht in jedem Element, sondern nur in solchen, in denen Oszillationen auftreten bzw. Unstetigkeiten vermutet werden. Zum Auffinden dieser Elemente sind bereits verschiedene Indikatoren bekannt, vergleiche [4] und [65].
Wir verwenden in dieser Arbeit den koeffizientenbasierenden Indikator aus [4], der die grundlegenden Überlegungen aus [65] weiterführt. Dabei ist die Idee, das Verhältnis der höchsten zu den niedrigen Koeffizienten der APK-Reihenentwicklung in jedem Element τ_i zu vergleichen, wobei man zur Vermeidung der Division durch Null ein $\varepsilon > 0$ hinzuaddiert. Man untersucht also

$$\omega_i := \sum_{l+m=N} ||A_{m,l}||^2 (\hat{u}_{l,m})^2 \cdot \left(\sum_{l+m<N} ||A_{m,l}||^2 (\hat{u}_{l,m})^2 + \varepsilon \right)^{-1}.$$

Die Definitionen der verschiedenen Indikatoren verwenden in der Regel jeweils ω_i, dabei nutzen alle die Abschätzung

$$\omega_i := \sum_{l+m=N} ||A_{m,l}||^2 (\hat{u}_{l,m})^2 \overset{(3.32)}{\leq} N^{-2k}(N+\gamma)^{-2k} ||D(u \circ T_i)||_{\mathrm{L}^2(\mathbb{T},h)} = \mathcal{O}(N^{-4k}).$$

Der Indikator aus [4] wird schließlich definiert durch

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

und mit s_{res} wird die Filterung bezüglich des Exponentialfilters (4.11) mit dem Längenmaß (4.13) mit Hilfe der Filterstärke

$$\alpha_{N,i} = \begin{cases} s_{res} c N \frac{\Delta t}{h_i}, & s_{res} > 0.01, \\ 0, & \text{sonst}, \end{cases}$$

gesteuert. Bei schwachen Oszillationen unterhalb des Grenzwertes ist somit die Filterstärke Null. Der Filter (4.11) hat als Wert Eins und somit wird in diesem Element τ_i nicht gefiltert.

Bemerkung. In [4] wurde auch noch ein Sprungindikator eingeführt, der allerdings bei numerischen Tests einen höheren Ordnungsverlust abseits der Unstetigkeitsstelle aufweist, wie man [63] entnehmen kann.
Eine weitere Methode zum Auffinden der Elemente stellt das Kantendetektierungsverfahren mit Hilfe konjugierter Fourier-Reihen dar. Einen Überblick über diese Theorie liefert [61], und in [86] wurden sogar beide Ansätze für das SD-Verfahren bezüglich PKD-Polynomen miteinander verglichen. Dabei wurden die Koeffizienten der konjugierten Fourier-Reihe direkt aus den Koeffizienten der PKD-Polynome berechnet. Eine Erweiterung des Kantendetektierungsverfahrens auf allgemeine APK-Polynome ist hierbei ohne Probleme möglich und ist im Anhang ausgeführt, ohne jetzt explizit auf Details einzugehen.

4.2 Die Legendre-Methode

In der bisherigen Untersuchung der gefilterten APK-Reihe ist immer die Glattheit der zu entwickelnden Funktion u vorausgesetzt, siehe Satz 4.4. Jedoch soll der modale Filter in unserem numerischen Verfahren eingesetzt werden, wenn die Funktion u Sprungunstetigkeiten aufweist. Wir sind daher gerade für diesen Fall an Approximationsergebnissen der gefilterten APK-Reihe interessiert. Aus [80] ist eine Fehlerabschätzung für Reihenentwicklungen bezüglich trigonometrischer Funktionen bekannt, siehe Ungleichung (4.4). Diese wird in [36] auf Reihenentwicklungen bezüglich einer Familie orthogonaler Polynome übertragen. In [36] untersuchen die Autoren das Verhalten der abgeschnittenen, gefilterten Legendre-Reihe für den Fall, dass die zugrundeliegende Funktion u genau eine Sprungunstetigkeit aufweist. Jedoch sind der Satz und sein Beweis in der dort gegebenen Form nicht korrekt. Indem wir weitere Bedingungen an die Filterfunktion σ stellen, korrigieren wir die Aussage. Abschließend analysieren wir diese neuen Bedingungen und erläutern, warum der verwendete Ansatz nur bei Verwendung der Legendre-Polynome funktioniert.
Wir folgen der Notation und dem Aufbau der Arbeit [36]. In [36] werden Formeln und Darstellungen für die Legendre-Polynome gezeigt, wir hingegen beweisen soweit möglich den allgemeinen Fall der Jacobi-Polynome. Im späteren Verlauf beschränken wir uns auf die Gegenbauer-Polynome, um am Ende ausschließlich den Legendre-Fall zu betrachten. Soweit nicht anders vorausgesetzt, wird eine hinreichend glatte, skalare Funktion u in eine Fourier-Reihe entwickelt. Für die Filterfunktion σ gelte die Defintion aus [36] bzw. Definition 4.2 mit der Glattheitsbedingung (4.2) und $\sigma \in C^p[(0,1])$.

4.2.1 Darstellung und Eigenschaften der Jacobi-Polynome

Elementare Eigenschaften der Jacobi-Polynome wurden bereits in Kapitel 3 wiederholt. Dabei beinhaltet Lemma 3.3, dass die Jacobi-Polynome Lösungen eines singulären Sturm-Liouville-Problems sind. Sie erfüllen die Eigenwertgleichung

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x)^{\alpha+1}(1+x)^{\beta+1}\frac{\mathrm{d}P_n^{\alpha,\beta}(x)}{\mathrm{d}x}\right] &= -n(n+\alpha+\beta)(1-x)^{\alpha}(1+x)^{\beta}P_n^{\alpha,\beta}(x), \\ \iff \qquad LP_n^{\alpha,\beta}(x) &= -\lambda_n\omega(x)P_n^{\alpha,\beta}(x), \end{aligned}$$

mit dem Differentialoperator L, der Gewichtsfunkion $\omega(x) = (1-x)^{\alpha}(1+x)^{\beta}$ und dem Eigenwert $\lambda_n = n(n+\alpha+\beta+1)$. Der **Sturm-Liouville-Operator** $L_\omega := \frac{1}{\omega(x)}L$ reduziert das Problem auf

$$L_\omega P_n^{\alpha,\beta}(x) = -\lambda_n P_n^{\alpha,\beta}(x).$$

Der Differentialoperator L_ω ist selbstadjungiert und mit ihm kann man für die Fourier-Koeffizienten einer Funktion zeigen, dass

$$\hat{u} = \underbrace{\frac{1}{(P_n^{\alpha,\beta}, P_n^{\alpha,\beta})_{\mathrm{L}^2([-1,1],\omega)}}}_{:=\gamma_n} (u, P_n^{\alpha,\beta})_{\mathrm{L}^2([-1,1],\omega)} = \frac{1}{\gamma_n} \left(u, \left(\frac{-1}{\lambda_n} \right) L_\omega P_n^{\alpha,\beta} \right)_{\mathrm{L}^2([-1,1],\omega)}$$
$$= \frac{-1}{\gamma_n \lambda_n} \left(L_\omega u, P_n^{\alpha,\beta} \right)_{\mathrm{L}^2([-1,1],\omega)} \overset{\text{rekursiv}}{=} \cdots = (-1)^q \frac{1}{\lambda_n^q \gamma_n} (L_\omega^q u, P_n^{\alpha,\beta})_{\mathrm{L}^2([-1,1],\omega)}$$

gilt. Untersucht wird die abgeschnittene, gefilterte Fourier-Reihe der Jacobi-Polynome

$$F_N u_N(x) := \sum_{n=0}^{N} \sigma\left(\frac{n}{N}\right) \hat{u}_n P_n^{\alpha,\beta}(x), \qquad \text{mit } \hat{u}_n = \frac{1}{\gamma_n} \int_{-1}^{1} w(s) u(s) P_n^{\alpha,\beta}(s) \mathrm{d}s.$$

Verwendet man die Darstellung über die Fourier-Koeffizienten, so gilt

$$F_N u_N(x) = \int_{-1}^{1} w(s) u(s) K_N^0(x,s) \mathrm{d}s,$$
$$\text{mit} \quad K_N^0(x,s) = \sum_{n=0}^{N} \frac{1}{\gamma_n} \sigma\left(\frac{n}{N}\right) P_n^{\alpha,\beta}(s) P_n^{\alpha,\beta}(x).$$

Man definiert eine Folge durch

$$L_\omega K_N^{l+1}(x,s) := K_N^l(x,s), \quad \forall l = 0, 1, 2, \ldots,$$

wobei die $K_N^l(x,s)$ die Darstellung

$$K_N^l(x,s) = \sum_{n=0}^{N} \sigma\left(\frac{n}{N}\right) \left(-\frac{1}{\lambda_n} \right)^l \frac{P_n^{\alpha,\beta}(s) P_n^{\alpha,\beta}(x)}{\gamma_n}, \quad \forall l = 0, 1, 2, \ldots,$$

besitzen. Im weiteren Teil dieses Kapitels wird eine Funktion u betrachtet, die genau eine Sprungunstetigkeit in einem Punkt $x = c$ mit $c \in (-1, 1)$ aufweist. Ansonsten ist u hinreichend glatt. Für die gefilterte Fourier-Summe $F_N u_N(x)$ erhalten wir

$$F_N u_N(x) = \int_{-1}^{1} \omega(s) u_n(s) K_N^0(x,s) \mathrm{d}s = \int_{-1}^{c^-} \omega(s) u(s) L_\omega K_N^1(x,s) \mathrm{d}s + \int_{c^+}^{1} \omega(s) u(s) L_\omega K_N^1(x,s) \mathrm{d}s.$$

Betrachtet man nun nur das linke Integral, so gilt mit zweimaliger partieller Integration:

$$\int_{-1}^{c^-} \omega(s)u(s)L_\omega K_N^1(x,s)\mathrm{d}s = \int_{-1}^{c^-} u(s)\left(\frac{\partial}{\partial s}\left[(1-s)^{\alpha+1}(1+s)^{\beta+1}\frac{\partial}{\partial s}K_N^1(x,s)\right]\right)\mathrm{d}s$$

$$= u(c^-)(1-c)^{\alpha+1}(1+c)^{\beta+1}\frac{\partial}{\partial s}K_N^1(x,s)\Big|_{s=c} - \int_{-1}^{c^-}\left(\frac{\partial}{\partial s}u(s)\right)(1-s)^{\alpha+1}(1+s)^{\beta+1}\frac{\partial K_N^1(x,s)}{\partial s}\mathrm{d}s$$

$$= u(c^-)(1-c)^{\alpha+1}(1+c)^{\beta+1}\frac{\partial}{\partial s}K_N^1(x,s)\Big|_{s=c} - \int_{-1}^{c^-}\left((1-s)^{\alpha+1}(1+s)^{\beta+1}\frac{\partial u(s)}{\partial s}\right)\frac{\partial K_N^1(x,s)}{\partial s}\mathrm{d}s$$

$$= u(c^-)(1-c)^{\alpha+1}(1+c)^{\beta+1}\frac{\partial}{\partial s}K_N^1(x,s)\Big|_{s=c} - \left[(1-c^-)^{\alpha+1}(1+c^-)^{\beta+1}\left(\frac{\partial}{\partial s}u(c^-)\right)\right]K_N^1(x,c)$$
$$+ \int_{-1}^{c^-}\frac{\partial}{\partial s}\left[(1-s)^{\alpha+1}(1+s)^{\beta+1}\frac{\partial}{\partial s}u(s)\right]K_N^1(x,s)\mathrm{d}s.$$

Mit Hilfe des Sturm-Liouville-Operators folgt

$$\int_{-1}^{c^-}\frac{\partial}{\partial s}\left[(1-s)^{\alpha+1}(1+s)^{\beta+1}\frac{\partial}{\partial s}u(s)\right]K_N^1(x,s)\mathrm{d}s = \int_{-1}^{c^-}\omega(s)L_\omega u(s)K_N^1(x,s)\mathrm{d}s.$$

Analoges gilt für

$$\int_{c^+}^{1}\omega(s)u(s)L_\omega K_N^1(x,s)\mathrm{d}s = -u(c^+)(1-c)^{\alpha+1}(1+c)^{\beta+1}\frac{\partial}{\partial s}K_N^1(x,s)\Big|_{s=c}$$
$$+\left[(1-c^+)^{\alpha+1}(1+c^+)^{\beta+1}\left(\frac{\partial}{\partial s}u(c^+)\right)\right]K_N^1(x,c) + \int_{c^+}^{1}\omega(x)(L_\omega u(s)K_N^1(x,s)\mathrm{d}s.$$

Fasst man beides zusammen, so erhält man

$$F_N u_N(x) = (1-c)^{\alpha+1}(1+c)^{\beta+1}\left(u(c^-)-u(c^+)\right)\frac{\partial}{\partial s}K_N^1(x,s)\Big|_{s=c}$$
$$+\left((1-c)^{\alpha+1}(1+c)^{\beta+1}\right)\frac{\partial}{\partial s}\left[u(c^+)-u(c^-)\right]K_N^1(x,c) + \int_{-1}^{1}\omega(s)L_\omega u(s)K_N^1(x,s)\mathrm{d}s.$$

Durch mehrfache Anwendung dieser Rechenoperationen gelangt man zu der Formel

$$\begin{aligned} F_N u_N(x) = &\sum_{l=0}^{q-1}(1-c)^{\alpha+1}(1+c)^{\beta+1}\left[L_\omega^l u(c^-) - L_\omega^l u(c^+)\right]\frac{\partial}{\partial s}K_N^{l+1}(x,s)\Big|_{s=c} \\ &-\sum_{l=0}^{q-1}(1-c)^{\alpha+1}(1+c)^{\beta+1}\frac{\partial}{\partial s}\left[L_\omega^l u(c^-) - L_\omega^l u(c^+)\right]K_N^{l+1}(x,c) \\ &+\int_{-1}^{1}\omega(s)L_\omega^q u(s)K_N^q(x,s)\mathrm{d}s. \end{aligned} \tag{4.14}$$

Als nächstes beweisen wir ein Resultat bezüglich des Approximationsverhaltens der abgeschnittenen, gefilterten Jacobi-Reihe. Dabei ist u hinreichend glatt. Ein vergleichbares Ergebnis wurde bereits für die APK-Polynome gezeigt, vergleiche Satz 4.4.

Satz 4.7. *Sei $u \in H^{2q}([-1,1],\omega)$ und der Filter $\sigma(\eta) \in C^{2q}([0,\infty))$ mit den Eigenschaften*

$$\sigma(\eta) = \begin{cases} \sigma(0) = 1, & \\ \sigma(\eta) = 0, & \text{für } \eta > 1, \\ \sigma^{(l)}(0) = 0, & \text{für } l = 1,\dots,2q-1, \end{cases}$$

gegeben. Weiterhin sei $\delta = \max\{\alpha,\beta\}$ für $\alpha,\beta \in \mathbb{N}_0$ und $4q-2-2\delta \geq 0$. Dann gilt:

$$|u(x) - F_n u_N| \leq N^{1-2q+\delta}||L_\omega^q u||_{\mathrm{L}^2([-1,1],\omega)}.$$

Beweis. Sei $u(x)$ gegeben mit

$$u(x) = \int_{-1}^{1}\omega(s)u(s)G^0(x,s)\mathrm{d}s, \quad G^0(x,s) = \sum_{n=0}^{\infty}\frac{1}{\gamma_n}P_n^{\alpha,\beta}(s)P_n^{\alpha,\beta}(x).$$

Weiterhin definieren wir mit dem Differentialoperator L_ω die Folge

$$L_\omega G^{l+1} := G^l, \quad l = 0,1,2,\dots,$$

mit $u(x) \in H^{2q}([-1,1],\omega)$. Die $G^l(x,s)$ besitzen die Darstellung

$$G^l(x,s) = \sum_{n=0}^{\infty}\left(\frac{-1}{\lambda_n}\right)^l\frac{P_n^{\alpha,\beta}(s)P_n^{\alpha,\beta}(x)}{\gamma_n}.$$

Mit Gleichung (4.14) und partieller Integration folgt:

$$u(x) - F_N u_N(x) = \int_{-1}^{1}\omega(s)L_\omega^q u(s)\left[G^q(x,s) - K_N^q(x,s)\right]\mathrm{d}s. \tag{4.15}$$

Aus der Definition von G^q und K^q folgt:

$$|u(x) - F_N u_N(x)| \leq \left| \int_{-1}^{1} \mathcal{S}_N(x,s) L_\omega^q u(s) \omega(s) \mathrm{d}s \right| + \left| \int_{-1}^{1} \mathcal{R}_N(x,s) L_\omega^q u(s) \omega(s) \mathrm{d}s \right|,$$

mit

$$\mathcal{S}_N(x,s) = \sum_{n=0}^{N} \left[1 - \sigma\left(\frac{n}{N}\right) \right] \left(-\frac{1}{\lambda_n} \right)^q \frac{P_n^{\alpha,\beta}(x) P_n^{\alpha,\beta}(s)}{\gamma_n}$$

und

$$\mathcal{R}_N(x,s) = \sum_{n=N+1}^{\infty} \left(-\frac{1}{\lambda_n} \right)^q \frac{P_n^{\alpha,\beta}(x) P_n^{\alpha,\beta}(s)}{\gamma_n}.$$

Mit Hilfe der Schwarz'schen Ungleichung gilt für $\mathcal{R}_N$ die Ungleichung

$$\left| \int_{-1}^{1} \mathcal{R}_N(x,s) L_\omega^q u(s) w(s) \mathrm{d}s \right| \leq \left(\int_{-1}^{1} w(s) (\mathcal{R}_N(x,s))^2 \mathrm{d}s \right)^{\frac{1}{2}} \left(\int_{-1}^{1} w(s) (L_\omega^q u(s))^2 \mathrm{d}s \right)^{\frac{1}{2}}.$$

Analoges folgt für $\mathcal{S}_N$. Wir untersuchen

$$||\mathcal{R}_N(x,s)||^2_{\mathrm{L}^2([-1,1],\omega)} = \int_{-1}^{1} w(s) \left(\sum_{n=N+1}^{\infty} \frac{1}{\lambda_n^q \gamma_n} P_n^{\alpha,\beta}(x) P_n^{\alpha,\beta}(s) \right)^2 \mathrm{d}s.$$

Das Produkt zweier Jacobi-Polynome unterschiedlichen Grades im Skalarprodukt ist aufgrund der Orthogonalitätsbeziehung gleich Null und verschwindet aus der Rechnung. Man schätzt die Jacobi-Polynome der Variablen x mit (3.5) ab. Es ist

$$\begin{aligned}
||\mathcal{R}_N(x,s)||^2_{\mathrm{L}^2([-1,1],\omega)} &= \int_{-1}^{1} w(s) \left(\sum_{n=N+1}^{\infty} \frac{1}{\lambda_n^{2q} \gamma_n^2} \left(P_n^{\alpha,\beta}(x) \right)^2 \left(P_n^{\alpha,\beta}(s) \right)^2 \right) \mathrm{d}s \\
&= \sum_{n=N+1}^{\infty} \frac{1}{\lambda_n^{2q} \gamma_n^2} \left(P_n^{\alpha,\beta}(x) \right)^2 \int_{-1}^{1} w(s) \left(P_n^{\alpha,\beta}(s) \right)^2 \mathrm{d}s \\
&\overset{(3.5)}{\leq} \sum_{n=N+1}^{\infty} \frac{1}{\lambda_n^{2q} \gamma_n^2} \binom{n+\delta}{n}^2 \int_{-1}^{1} w(s) \left(P_n^{\alpha,\beta}(s) \right)^2 \mathrm{d}s =: \mathcal{R}_1.
\end{aligned}$$

Es ist $\delta = \max\{\alpha, \beta\}$ mit $\alpha, \beta \in \mathbb{N}_0$. Sei $\alpha \geq \beta$ (analoges für $\beta \geq \alpha$), dann gilt mit Lemma 3.2

$$\begin{aligned}\mathcal{R}_1 &\leq \sum_{n=N+1}^{\infty} \frac{1}{n^{2q}(n+\alpha+\beta+1)^{2q}} \frac{n!\Gamma(n+\alpha+\beta+1)(2n+\alpha+\beta+1)}{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)2^{\alpha+\beta+1}} \binom{n+\delta}{n}^2 \\ &\leq \sum_{n=N+1}^{\infty} \frac{1}{n^{2q}(n+\alpha+\beta+1)^{2q-1}} \frac{n!(n+\alpha+\beta)!((n+\alpha)!)^2}{(n+\alpha)!(n+\beta)!2^{\alpha+\beta}(n!\alpha!)^2} \\ &= \frac{1}{2^{\alpha+\beta}(\alpha!)^2} \sum_{n=N+1}^{\infty} \frac{1}{n^{2q}(n+\alpha+\beta+1)^{2q-1}} \frac{\Gamma(n+\alpha+\beta+1)}{\Gamma(n+\beta+1)} \frac{(n+\alpha)!}{n!} =: \mathcal{R}_2.\end{aligned}$$

Für $\alpha \geq 1$ schätzt man die hinteren Faktoren wie folgt ab:

$$\begin{aligned}\frac{\Gamma(n+\alpha+\beta+1)}{\Gamma(n+\beta+1)} &\overset{(3.9)}{=} (n+\beta+1)_\alpha = \prod_{i=1}^{\alpha}(n+\beta+i) \leq (n+\beta+\alpha)^\alpha, \\ \frac{(n+\alpha)!}{n!} &= \prod_{i=1}^{\alpha}(n+i) \leq (n+\alpha)^\alpha.\end{aligned}$$

Für $\alpha = 0$ sind die beiden Faktoren gleich Eins. Man erhält

$$\begin{aligned}\mathcal{R}_2 &\leq \frac{1}{2^{\alpha+\beta}(\alpha!)^2} \sum_{n=N+1}^{\infty} \frac{1}{n^{2q}(n+\alpha+\beta+1)^{2q-1}} (n+\beta+\alpha)^\alpha (n+\alpha)^\alpha \\ &\leq \frac{1}{2^{\alpha+\beta}(\alpha!)^2} \sum_{n=N+1}^{\infty} \frac{1}{n^{2q}(n+\alpha+\beta+1)^{2q-1-2\alpha}} \\ &\leq \frac{C_\varsigma}{2^{\alpha+\beta}(\alpha!)^2} \sum_{n=N+1}^{\infty} \frac{1}{n^{4q-2\alpha-1}} \leq \hat{C}_1 N^{-4q+2\alpha+2},\end{aligned}$$

mit den Konstanten C_ς und $\hat{C}_1 \in \mathbb{R}^+$. Für $\beta \geq \alpha$ erhält man eine analoge Abschätzung mit $||\mathcal{R}_N(x,s)||^2_{\mathrm{L}^2([-1,1],\omega)} \leq \hat{C}_2 N^{-4q+2\beta+2}$.

Für $||\mathcal{S}_N(x,s)||^2_{\mathrm{L}^2([-1,1],\omega)}$ gilt

$$\begin{aligned}||\mathcal{S}_N(x,s)||^2_{\mathrm{L}^2([-1,1],\omega)} &= \left(\sum_{n=0}^{N} \left[1-\sigma\left(\frac{n}{N}\right)\right]^2 \left(\frac{1}{\lambda_n}\right)^{2q}\right) \frac{|P_n^{\alpha,\beta}(x)|^2}{\gamma_n} \\ &\leq \frac{N|P_N^{\alpha,\beta}(x)|^2}{\gamma_N} \left(\frac{1}{\lambda_N}\right)^{2q} \left(\frac{1}{N}\sum_{n=0}^{N} \left[1-\sigma\left(\frac{n}{N}\right)\right]^2 \left(\frac{\lambda_n}{\lambda_N}\right)^{-2q}\right) =: \mathcal{S}_1.\end{aligned}$$

Unter Verwendung des bereits Gezeigten erhält man

$$\mathcal{S}_1 \leq \hat{C}_3 N^{-4q+2\delta+2} \left(\frac{1}{N}\sum_{n=0}^{N} \left[1-\sigma\left(\frac{n}{N}\right)\right]^2 \left(\frac{\lambda_n}{\lambda_N}\right)^{-2q}\right) =: \mathcal{S}_2.$$

Es gilt

$$\left(\frac{n}{N}\right)^2 \leq \frac{n(n+\alpha+\beta+1)}{N(N+\alpha+\beta+1)} \leq \tilde{C}\left(\frac{n}{N}\right)^2,$$

und damit folgt

$$\mathcal{S}_2 \leq \hat{C}_3 N^{-4q+2\delta+2}\left(\frac{1}{N}\sum_{n=0}^{N}\left[1-\sigma\left(\frac{n}{N}\right)\right]^2\left(\frac{n}{N}\right)^{-4q}\right).$$

Bei der Klammer handelt es sich um eine Riemann'sche Summe für $N \to \infty$. Sie konvergiert für $N \to \infty$ gegen das Integral

$$\int_0^1 ((1-\sigma(\eta))^2\eta^{-4q})\mathrm{d}\eta,$$

mit $\eta := \frac{n}{N}$. Durch Taylor-Entwicklung der Funktion σ im Nullpunkt und mit Hilfe der Filtereigenschaften folgt, dass das Integral

$$\left|\int_0^1 ((1-\sigma(\eta))^2\eta^{-4q})\mathrm{d}\eta\right| < \infty$$

beschränkt ist. Es gilt

$$||\mathcal{S}_N(x,s)||^2_{\mathrm{L}^2([-1,1],\omega)} \leq \hat{C}_4 N^{-4q+2+2\delta},$$

und somit folgt

$$|u(x) - F_N u_N(x)| \leq \hat{C} N^{1+\delta-2q}||L^q_\omega u||_{L^2([-1,1],\omega)},$$

mit einer Konstanten $\hat{C} \in \mathbb{R}^+$. □

Bemerkung. Im Beweis wurde die Konstante C_ς verwendet ohne dabei näher auf sie einzugehen. C_ς hängt von der Größe $2q-2\delta-1$ ab und ob dieser Wert größer oder kleiner Null ist, vergleiche dazu den Beweis von Satz 4.4.
Der Satz 4.7 beinhalten das Theorem 4.1 aus [36, S.1438]. In [36] wird ausschließlich der Legendre-Fall analysiert.

Das nächste Lemma enthält eine Reihendarstellung der Jacobi-Polynome im Nullpunkt. Zusätzlich zu unserer Grundvoraussetzung $\alpha \in \mathbb{N}_0$ schränken wir uns auf den ultrasphärischen Fall ein, das heißt, es gilt $\alpha = \beta$.

Lemma 4.8. *Es sei $m \in \mathbb{N}$ und $\alpha \in \mathbb{N}_0$. Dann gilt*

$$\frac{1}{\gamma_{2m}}P^{\alpha,\alpha}_{2m}(0) = \begin{cases} \frac{(-1)^m(2m+\alpha+1)_\alpha}{m_\alpha}\left(\sum_{q=0}^{\infty}\hat{a}_q m^{\frac{1}{2}-q}\right), & \text{für } m < \alpha, \\ (-1)^m\sum_{q=0}^{\infty}\tilde{a}_q m^{\frac{1}{2}-q}, & \text{für } m \geq \alpha, \end{cases} \tag{4.16}$$

dabei sind $\hat{a}_q$ und $\tilde{a}_q$ nicht von m abhängig.
Für ein Polynom ungeraden Grades gilt

$$\frac{1}{\gamma_{2m+1}} P^{\alpha,\alpha}_{2m+1}(0) = 0.$$

Beweis. Aus [73, Theorem 4.1, S.59] erhält man die beiden Formeln:

$$P^{\alpha,\alpha}_{2m}(x) = (-1)^m \frac{\Gamma(2m+\alpha+1)\Gamma(m+1)}{\Gamma(m+\alpha+1)\Gamma(2m+1)} P_m^{-\frac{1}{2},\alpha}(1-2x^2), \tag{4.17}$$

$$P^{\alpha,\alpha}_{2m+1}(x) = (-1)^m \frac{\Gamma(2m+\alpha+2)\Gamma(m+1)}{\Gamma(m+\alpha+1)\Gamma(2m+2)} x P_m^{\frac{1}{2},\alpha}(1-2x^2). \tag{4.18}$$

Es gilt $P^{(\alpha,\alpha)}_{2m+1}(0) = 0$ für alle $\alpha \in \mathbb{N}_0$. Dies gilt sogar für alle $-1 < \alpha \in \mathbb{R}$.
Man benutzt Formel (4.17) mit Gleichung (8.3) und erhält

$$\begin{aligned} P^{\alpha,\alpha}_{2m}(0) &= (-1)^m \frac{\Gamma(2m+\alpha+1)\Gamma(m+1)}{\Gamma(m+\alpha+1)\Gamma(2m+1)} P_m^{-\frac{1}{2},\alpha}(1) \\ &= (-1)^m \frac{\Gamma(2m+\alpha+1)\Gamma(m+1)}{\Gamma(m+\alpha+1)\Gamma(2m+1)} (-1)^m P_m^{\alpha,-\frac{1}{2}}(-1) \\ &= \frac{\Gamma(2m+\alpha+1)\Gamma(m+1)}{\Gamma(m+\alpha+1)\Gamma(2m+1)} \frac{\left(\frac{1}{2}\right)_m}{(\alpha+\frac{1}{2})_m} C_{2m}^{\alpha+\frac{1}{2}}(0), \end{aligned}$$

wobei die $C_{2m}^{\alpha+\frac{1}{2}}$ ultrasphärische Polynome[3] sind. Mit Gleichung (8.6) für $C_{2m}^{\alpha+\frac{1}{2}}(0)$ gilt

$$\begin{aligned} P^{\alpha,\alpha}_{2m}(0) =& \frac{\Gamma(2m+\alpha+1)\Gamma(m+1)}{\Gamma(m+\alpha+1)\Gamma(2m+1)} \frac{\left(\frac{1}{2}\right)_m}{(\alpha+\frac{1}{2})_m} (-1)^m \frac{\Gamma(\alpha+\frac{1}{2}+m)}{m!\Gamma(\alpha+\frac{1}{2})} \\ \overset{(3.9)}{=}& (-1)^m \frac{\Gamma(2m+\alpha+1)\Gamma(m+1)}{\Gamma(m+\alpha+1)\Gamma(2m+1)} \frac{\Gamma(m+\frac{1}{2})\Gamma(\alpha+\frac{1}{2})\Gamma(\frac{1}{2}+\alpha+m)}{\Gamma(\frac{1}{2})\Gamma(\alpha+\frac{1}{2}+m)m!\Gamma(\alpha+\frac{1}{2})} \\ =& (-1)^m \frac{\Gamma(2m+\alpha+1)\Gamma(m+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(\alpha+m+1)\Gamma(2m+1)}. \end{aligned}$$

Mit der bekannten Identität für die Gammafunktion $\Gamma(m+\frac{1}{2}) = \frac{(2m)!}{m!4^m}\sqrt{\pi}\ \forall m \in \mathbb{N}_0$, vergleiche Anhang, folgt

$$P^{\alpha,\alpha}_{2m}(0) = (-1)^m \frac{(2m)!\Gamma(2m+\alpha+1)}{2^{2m}m!\Gamma(\alpha+m+1)\Gamma(2m+1)}. \tag{4.19}$$

[3]Die ultrasphärischen Polynome oder auch Gegenbauer-Polynome sind ein Spezialfall der Jacobi-Polynome. Im Anhang 8 findet man die Definition und einige ihrer elementaren Eigenschaften aufgelistet.

Mit Lemma 3.2 für γ_{2m} erhält man insgesamt

$$\begin{aligned}\frac{P_{2m}^{\alpha,\alpha}(0)}{\gamma_{2m}} &= (-1)^m \frac{(2m)!\Gamma(2m+\alpha+1)}{2^{2m}m!\Gamma(\alpha+m+1)\Gamma(2m+1)}\frac{1}{\gamma_{2m}} \\ &= (-1)^m \frac{(4m+2\alpha+1)}{2^{2\alpha+1+2m}}\frac{(2m)!}{m!}\frac{\Gamma(2m+2\alpha+1)}{\Gamma(2m+\alpha+1)\Gamma(m+\alpha+1)} \\ &= \frac{(4m+2\alpha+1)}{2^{2\alpha+1+2m}}\frac{(2m)!}{m!}\frac{(2m+\alpha+1)_\alpha\Gamma(2m+\alpha+1)}{\Gamma(2m+\alpha+1)m!(m)_\alpha} \\ &= \frac{(-1)^m(2m+\alpha+1)_\alpha}{m_\alpha}\underbrace{\frac{1}{2^{2\alpha+1}}\frac{4m+1+2\alpha}{2^{2m}}\frac{(2m)!}{m!m!}}_{A}.\end{aligned}$$

Für $\alpha = 0$ hätten wir genau den Legendre-Fall aus [36, S.1441]) nachgewiesen. Mit Hilfe der Stirling-Formel

$$\Gamma(n+1) = n! = \sqrt{2\pi n}n^n e^{-n}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \mathcal{O}(n^{-3})\right)$$

und einem analogen Vorgehen zu [36, S.1441] für den Teil A erhält man

$$\frac{P_{2m}^{(\alpha,\alpha)}(0)}{\gamma_{2m}} = \frac{(-1)^m(2m+\alpha+1)_\alpha}{m_\alpha}\left(\sum_{q=0}^{\infty}\hat{a}_q m^{\frac{1}{2}-q}\right).$$

Somit ist der erste Teil der Formel (4.16) gezeigt.
Wir untersuchen jetzt noch den Faktor $\frac{(2m+\alpha+1)_\alpha}{m_\alpha}$ unter der Voraussetzung $m \geq \alpha$. Es gilt

$$\frac{(2m+\alpha+1)_\alpha}{m_\alpha} = \frac{\prod_{i=1}^{\alpha}(2m+\alpha+i)}{\prod_{i=1}^{\alpha}(m+i-1)} = 2^\alpha \prod_{i=1}^{\alpha}\left(1+\frac{\alpha+i}{2m}\right)\frac{m}{m+1}\cdot\ldots\cdot\frac{m}{m+\alpha-1}.$$

Mit Lemma 8.4 aus dem Anhang kann jeder Faktor $\frac{m}{m+\lambda_\alpha}$ mit $\lambda_\alpha = 1, 2, \ldots, \alpha-1$ durch eine unendlichen Reihe $\sum_{p=0}^{\infty}\hat{b}_{p,\lambda_\alpha}m^{-p}$ dargestellt werden. Jede Reihe konvergiert absolut und mittels Cauchy-Produkt kann man alle Reihen in eine absolut konvergente Reihe der Gestalt $\sum_{p=0}^{\infty}\tilde{b}_p m^{-p}$ zusammenfassen, das bedeutet:

$$\prod_{\lambda_\alpha=1}^{\alpha-1}\frac{m}{m+\lambda_\alpha} = \sum_{p=0}^{\infty}\tilde{b}_p m^{-p}.$$

Berechnet man das Cauchy-Produkt dieser absolut konvergenten Reihe mit der Reihe $\left(\sum_{q=0}^{\infty} \hat{a}_q m^{\frac{1}{2}-q}\right)$, ergibt sich

$$\begin{aligned}\frac{P_{2m}^{(\alpha,\alpha)}(0)}{\gamma_{2m}} &= (-1)^m 2^\alpha \prod_{i=1}^{\alpha}\left(1+\frac{\alpha+i}{2m}\right)\left(\sum_{p=0}^{\infty}\tilde{b}_p m^{-p}\right)\left(\sum_{q=0}^{\infty}\hat{a}_q m^{\frac{1}{2}-q}\right)\\ &= (-1)^m 2^\alpha \prod_{i=1}^{\alpha}\left(1+\frac{\alpha+i}{2m}\right)\left(\sum_{q=0}^{\infty}\tilde{c}_q m^{\frac{1}{2}-q}\right)\end{aligned}$$

mit $\tilde{c}_q = \sum_{p=0}^{q}\tilde{b}_p\hat{a}_{q-p}$.

Die Faktoren $2^\alpha \prod_{i=1}^{\alpha}\left(1+\frac{\alpha+i}{2m}\right)$ bewirken nur eine Veränderung in den Koeffizienten der Reihe, nicht jedoch im Konvergenzverhalten. Durch Umsortierung erhält man

$$\frac{1}{\gamma_{2m}}P_{2m}^{\alpha,\alpha}(0) = (-1)^m\sum_{q=0}^{\infty}\tilde{a}_q m^{\frac{1}{2}-q}.$$

□

4.2.2 Approximationsverhalten bei einer Unstetigkeitsstelle

Als nächstes kommen wir zum wichtigsten Punkt der Untersuchung. Wir analysieren das Approximationsverhalten der abgeschnittenen, gefilterten Legendre-Reihe einer Funktion u, die im Inneren des Intervalls $[-1,1]$ genau eine Sprungunstetigkeit im Punkt $x=c$ aufweist und ansonsten hinreichend glatt ist. Mit (4.14) und (4.15) folgt

$$\begin{aligned}&u(x) - F_N u_N(x)\\ =&\sum_{l=0}^{q-1}(1-c)^{\alpha+1}(1+c)^{\beta+1}\left[L_\omega^l u(c^-) - L_\omega^l u(c^+)\right]\left[\frac{\partial}{\partial s}G^{l+1}(x,s) - \frac{\partial}{\partial s}K_N^{l+1}(x,s)\right]\Big|_{s=c}\\ -&\sum_{l=0}^{q-1}(1-c)^{\alpha+1}(1+c)^{\beta+1}\frac{\partial}{\partial s}\left[L_\omega^l u(c^-) - L_\omega^l u(c^+)\right]\left[G^{l+1}(x,c) - K_N^{l+1}(x,c)\right]\\ +&\int_{-1}^{1} w(s)L_\omega^q u(s)\left[G^q(x,s) - K_N^q(x,s)\right]\mathrm{d}s.\end{aligned}$$

Man muss das Verhalten von

$$\begin{aligned}G^l(x,c) - K_N^l(x,c) =& \sum_{n=0}^{N}\left(1-\sigma\left(\frac{n}{N}\right)\right)\left(\frac{-1}{\lambda_n}\right)^l\frac{P_n^{\alpha,\beta}(c)P_n^{\alpha,\beta}(x)}{\gamma_n}\\ &+\sum_{n=N+1}^{\infty}\left(\frac{-1}{\lambda_n}\right)^l\frac{P_n^{\alpha,\beta}(c)P_n^{\alpha,\beta}(x)}{\gamma_n} = Q_N^1(x,c) + R_N^1(x,c)\end{aligned}$$

und

$$\frac{\partial}{\partial s}G^l(x,s)\Big|_{s=c} - \frac{\partial}{\partial s}K_N^l(x,s)\Big|_{s=c} = \sum_{n=0}^{N}\left(1-\sigma\left(\frac{n}{N}\right)\right)\left(\frac{-1}{\lambda_n}\right)^l \frac{P_n^{\alpha,\beta\prime}(c)P_n^{\alpha,\beta}(x)}{\gamma_n} + \sum_{n=N+1}^{\infty}\left(\frac{-1}{\lambda_n}\right)^l \frac{P_n^{\alpha,\beta\prime}(c)P_n^{\alpha,\beta}(x)}{\gamma_n} = Q_N^2(x,c) + R_N^2(x,c)$$

verstehen, um das Approximationsverhalten angeben zu können. Wir untersuchen Q_N^1, Q_N^2, R_N^1 und R_N^2 für den Legendre-Fall unter der Annahme, dass die Unstetigkeitsstelle von u im Nullpunkt liegt. Somit gilt $\alpha = \beta = 0$ und $c = 0$. Wir betrachten das Ganze in den Randpunkten $x = \pm 1$ und beginnen unsere Untersuchung mit Q_N^1 und R_N^1 im Punkt $(x,c) = (\pm 1, 0)$. Die Legendre-Polynome ungeraden Grades verschwinden im Nullpunkt, vergleiche Lemma 4.8. Daher gilt mit $2M = N$

$$\begin{aligned} Q_N^1(\pm 1,0) &= \sum_{m=0}^{M}\left(1-\sigma\left(\frac{m}{M}\right)\right)\left(\frac{-1}{\lambda_{2m}}\right)^l \frac{P_{2m}(\pm 1)P_{2m}(0)}{\gamma_{2m}} \\ &= \left(\frac{-1}{\lambda_{2M}}\right)^l \sum_{m=0}^{M}\left(1-\sigma\left(\frac{m}{M}\right)\right)\left(\frac{\lambda_{2m}}{\lambda_{2M}}\right)^{-l} \frac{P_{2m}(0)}{\gamma_{2m}} \end{aligned}$$

und

$$R_N^1(\pm 1,0) = \sum_{m=M+1}^{\infty}\left(\frac{-1}{\lambda_{2m}}\right)^l \frac{P_{2m}(\pm 1)P_{2m}(0)}{\gamma_{2m}} = \left(\frac{-1}{\lambda_{2M}}\right)^l \sum_{m=M+1}^{\infty}\left(\frac{\lambda_{2m}}{\lambda_{2M}}\right)^{-l} \frac{P_{2m}(0)}{\gamma_{2m}}.$$

Die Legendre-Polynome nehmen in den Randpunkten die Werte $P_n(\pm 1) = (\pm 1)^n$ an, siehe Gleichungen (3.3) und (3.4). Sei weiterhin $M \gg 1$, dann gilt für den Quotienten der Eigenwerte

$$\left(\frac{\lambda_{2m}}{\lambda_{2M}}\right)^{-l} = \frac{(2M)^l(2M+1)^l}{(2m)^l(2m+1)^l} = \frac{M^{2l}}{m^{2l}}\frac{m^l}{\left(m+\frac{1}{2}\right)^l} \overset{\text{Lemma 8.4}}{=} \left(\frac{m}{M}\right)^{-2l}\sum_{p=0}^{\infty}\hat{d}_p m^{-p}.$$

Mit Lemma 4.8 erhalten wir

$$\begin{aligned} \left(\frac{\lambda_{2m}}{\lambda_{2M}}\right)^{-l}\frac{P_{2m}(0)}{\gamma_{2m}} &= \left((-1)^m\sum_{q=0}^{\infty}\tilde{a}_q m^{\frac{1}{2}-q}\right)\left(\left(\frac{m}{M}\right)^{-2l}\sum_{p=0}^{\infty}\hat{d}_p m^{-p}\right) \\ &= (-1)^m\left(\frac{m}{M}\right)^{-2l}\sum_{r=0}^{\infty}\left(\sum_{p=0}^{r}\tilde{a}_{r-p}\hat{d}_p\right)m^{\frac{1}{2}-r} \qquad (4.20) \\ &= (-1)^m\left(\frac{m}{M}\right)^{-2l}\sum_{r=0}^{\infty}\hat{f}_r m^{\frac{1}{2}-r}. \end{aligned}$$

Mit $\frac{1}{\lambda_{2M}} \approx \frac{1}{(2M)^2}$ und Gleichung (4.20) ergibt sich

$$\begin{aligned} Q_N^1(\pm 1,0) &= \left(-\frac{1}{4}\right)^l \sum_{m=0}^{M} \left(1-\sigma\left(\frac{m}{M}\right)\right) \left(\frac{\lambda_{2m}}{\lambda_{2M}}\right)^{-l} \frac{P_{2m}(0)}{\gamma_{2m}} M^{-2l} \\ &= \left(-\frac{1}{4}\right)^l \sum_{m=0}^{M} \left(\left(1-\sigma\left(\frac{m}{M}\right)\right)(-1)^m \left(\frac{m}{M}\right)^{-2l} M^{-2l} \left(\sum_{r=0}^{\infty} \hat{f}_r m^{\frac{1}{2}-r}\right)\right) \\ &= \left(-\frac{1}{4}\right)^l \sum_{r=0}^{\infty} \hat{f}_r M^{-2l-r+\frac{1}{2}} \sum_{m=0}^{M} (-1)^m \left(1-\sigma\left(\frac{m}{M}\right)\right) \left(\frac{m}{M}\right)^{-2l-r+\frac{1}{2}} \end{aligned}$$

und

$$R_N^1(\pm 1,0) = \left(-\frac{1}{4}\right)^l \sum_{r=0}^{\infty} \hat{f}_r \left(\sum_{m=M+1}^{\infty} (-1)^m m^{-2l-r+\frac{1}{2}}\right).$$

Der Filter σ besitzt die Filterordnung $p > 0$ und ist zudem hinreichend glatt im Nullpunkt. Man kann σ durch seine Taylor-Entwicklung im Nullpunkt beschreiben. Es gilt mit $\eta := \frac{m}{M}$ und wegen $\sigma^{(i)}(0) = 0$ für alle $i = 1, \ldots, p-1$:

$$\sigma(\eta) = 1 + \frac{\sigma^{(p)}(\xi_m)}{(p)!}(\eta)^p \text{ mit } \xi_m \in [0,\eta].$$

Unter der Voraussetzung $2l - \frac{1}{2} > p$, gilt für jedes $r \in \mathbb{N}_0$, dass $2l + r - \frac{1}{2} - p > 0$ ist. Somit gilt

$$R_N^1(\pm 1,0) = \left(-\frac{1}{4}\right)^l \sum_{r=0}^{\infty} \hat{f}_r \underbrace{\sum_{m=M+1}^{\infty} (-1)^m m^{-2l-r+\frac{1}{2}}}_{\mathcal{O}(M^{-p-r})},$$

und die Reihe $\sum_{r=0}^{\infty} \hat{f}_r M^{-p-r}$ konvergiert aufgrund der absoluten Konvergenz der Reihe $\sum_{r=0}^{\infty} \hat{f}_r m^{\frac{1}{2}-r}$, vergleiche Gleichung (4.20).
Es ergibt sich $R_N^1(\pm 1,0) \approx \mathcal{O}(M^{-p})$.
Für die Abschätzung der Q_N^1 nutzen wir die Taylor-Entwicklung des Filters σ. Es ist

$$Q_N^1(\pm 1,0) = \left(-\frac{1}{4}\right)^l M^{-p} \sum_{r=0}^{\infty} \hat{f}_r \sum_{m=1}^{M} (-1)^m \frac{\sigma^{(p)}(\xi_m)}{m^{2l+r-\frac{1}{2}-p}}.$$

Mit einer analogen Argumentation wie zuvor folgt, dass die Reihe über r beschränkt ist. Es gilt

$$Q_N^1(\pm 1,0) \approx \mathcal{O}(M^{-p})$$

und insgesamt für den Fall $2l - \frac{1}{2} > p$

$$|Q_N^1(\pm 1,0) + R_N^1(\pm 1,0)| = \mathcal{O}(M^{-p}).$$

Bevor wir den Fall $2l - \frac{1}{2} < p$ betrachten, untersuchen wir noch $Q_N^2(\pm 1, 0) + R_N^2(\pm 1, 0)$. Wir verwenden Formel 4.5.7 aus [73, S.72]. Die Formel lautet

$$\begin{aligned}(2n+\alpha+\beta+2)(1-x^2)\frac{\mathrm{d}}{\mathrm{d}x}P_n^{\alpha,\beta}(x) &= -2(n+1)(n+\alpha+\beta+1)P_{n+1}^{\alpha,\beta}(x) \\ &\quad + (n+\alpha+\beta+1)\left((2n+\alpha+\beta+2)x - (\alpha-\beta)\right)P_n^{\alpha,\beta}(x).\end{aligned}$$

Für die Legendre-Polynome $\alpha = \beta = 0$ ergibt die Formel im Nullpunkt $x = 0$ angewandt:

$$P'_{2m-1}(0) = -2mP_{2m}(0),$$

und damit folgt

$$\begin{aligned}\frac{P'_{2m-1}(0)}{\gamma_{2m-1}} &= -\frac{(4m-1)(2m-1)!\Gamma(2m)(2m)P_{2m}(0)}{2(\Gamma(2m))^2} \\ &= -(4m-1)mP_{2m}(0) = -(4m^2-m)P_{2m}(0).\end{aligned}$$

Verwenden wir Gleichung (4.19) für $P_{2m}(0)$ und gehen analog zum Beweis von Lemma 4.8 vor, so ergibt sich

$$\frac{P'_{2m-1}(0)}{\gamma_{2m-1}} \overset{(4.19)}{=} -(4m^2-m)(-1)^m\frac{(2m)!}{2^{2m}(m!)^2} = (-1)^m\sum_{q=0}^{\infty}h_q m^{\frac{3}{2}-q},$$

und die Koeffizienten h_q sind nicht von m abhängig.
Für die Quotienten der Eigenwerte gilt

$$\left(\frac{\lambda_{2m-1}}{\lambda_{2M}}\right)^{-l} = \frac{(2M)^l(2M+1)^l}{(2m)^l(2m-1)^l} = \frac{M^{2l}}{m^{2l}}\frac{m^l}{\left(m-\frac{1}{2}\right)^l} \overset{8.4}{=} \left(\frac{m}{M}\right)^{-2l}\sum_{p=0}^{\infty}\tilde{d}_p m^{-p}.$$

Kombiniert man beides, erhält man analog zum Fall der Q_N^1 und R_N^1 eine Darstellung

$$\left(\frac{\lambda_{2m-1}}{\lambda_{2M}}\right)^{-l}\frac{P'_{2m-1}(0)}{\gamma_{2m-1}} = (-1)^m\left(\frac{m}{M}\right)^{-2l}\sum_{r=0}^{\infty}\tilde{f}_r m^{\frac{3}{2}-r},$$

welche für jedes $m \in \mathbb{N}$ absolut konvergiert.
Mit einem analogen Vorgehen wie zuvor folgt für $2l - \frac{3}{2} > p$

$$|Q_N^2(\pm 1, 0) + R_N^2(\pm 1, 0)| = \mathcal{O}(M^{-p}).$$

Nun sei $2l - \frac{1}{2} < p$ vorausgesetzt. Man betrachtet die Summe

$$\begin{aligned}Q_N^1(\pm 1, 0) + R_N^1(\pm 1, 0) &= \left(-\frac{1}{4}\right)^l\sum_{r=0}^{\infty}\hat{f}_r \\ &\quad \cdot\left[M^{-2l-r+\frac{1}{2}}\sum_{m=1}^{M}(-1)^m g\left(\frac{m}{M}\right) + \sum_{m=M+1}^{\infty}(-1)^m m^{-2l-r+\frac{1}{2}}\right],\end{aligned}$$

mit

$$g\left(\frac{m}{M}\right) := \left(1-\sigma\left(\frac{m}{M}\right)\right)\left(\frac{m}{M}\right)^{-2l-r+\frac{1}{2}}.$$

Wenn $2l+r-\frac{1}{2} > p$ gilt, ist man wieder beim bereits gezeigten Fall.
Wir fordern im Weiteren für den Filter, dass der Übergang im Punkt $x=1$ hinreichend glatt ist. Es existiert ein $n_1 \in \mathbb{N}$ mit $p-2l+\frac{3}{2} < p-2l+\frac{5}{2} < 2n_1 < p-2l+\frac{7}{2}$, so dass

$$\begin{aligned} \sigma(1) &= 0, \\ \left[\frac{\sigma(x)}{x^{2l-\frac{1}{2}}}\right]^{(2i-1)}\Bigg|_{x=1} &= 0, \qquad \forall i \in \mathbb{N} \text{ mit } 1 \le i \le n_1-1, \end{aligned} \tag{4.21}$$

gilt. Des Weiteren sei die Funktion $g \in C^{2n_1}([0,1])$ mit $n_1 \ge 2$. Wir werden diese Forderung benötigen, um Lemma 8.5 verwenden zu können. Wir betrachten für ein gerades M die Summe aus $Q_M^1 + R_N^1$ im Fall $r=0$. Solange $g \in C^{2n_1}([0,1])$ für $r>0$ gilt, kann analog argumentiert werden. Ansonsten nähern sich die Faktoren $\left(\frac{1}{m}\right)^r$ für $r>0$ schneller gegen Null an als $\left(\frac{1}{m}\right)^{r_1}$ für $r_1 < r$. Man zieht $\left(\frac{1}{m}\right)^{r_2}$ aus der Summe, bis die Voraussetzungen für g gelten. Den Faktor $\left(\frac{1}{m}\right)^{r_2}$ schätzen wir zusätzlich elementar ab, um ein äquivalentes Ergebnis wie im Fall $r=0$ zu erhalten. Wir verwenden Lemma 8.5 und Lemma 8.6. Es ergibt sich

$$\begin{aligned} &\left[M^{-2l+\frac{1}{2}}\sum_{m=1}^{M}(-1)^m g\left(\frac{m}{M}\right) + \sum_{m=M+1}^{\infty}(-1)^m m^{-2l+\frac{1}{2}}\right] \\ &= \frac{1}{2}M^{\frac{1}{2}-2l}\left(g(1)-g(0)-1\right) \\ &+ \sum_{i=1}^{n_1-1} M^{-2i+\frac{3}{2}-2l}\frac{B_{2i}}{(2i)!}(4^i-1)\left[\frac{\Gamma(2l-\frac{3}{2}+2i)}{\Gamma(2l-\frac{1}{2})} + g^{(2i-1)}(1) - g^{(2i-1)}(0)\right] \\ &+ \mathcal{O}(M^{-2l-2n_1+\frac{3}{2}}). \end{aligned} \tag{4.22}$$

Wegen $2l-\frac{1}{2} < p$ und den Eigenschaften des Filter σ gelten

$$g(1) = \frac{1-\sigma(1)}{1^{2l-\frac{1}{2}}} = 1$$

und

$$g(0) = \lim_{\eta\to 0}\frac{1-\sigma(\eta)}{\eta^{2l-\frac{1}{2}}} \overset{\text{L'Hospital}}{=} \lim_{\eta\to 0} c\frac{-\sigma^{(p)}(\eta)}{\eta^{2l-\frac{1}{2}-p}} = 0.$$

Daher ist $g(1)-g(0)-1=0$. Des Weiteren wurde

$$\underbrace{p-2l+\frac{3}{2} < p-2l+\frac{5}{2} < 2n_1 < p-2l+\frac{7}{2}}_{B}$$

vorausgesetzt. Somit ist $M^{-2n_1+\frac{3}{2}-2l}$ durch M^{-p} nach oben beschränkt.
Es muss nur noch gezeigt werden, dass unter den gegebenen Voraussetzungen

$$g^{(2i-1)}(1) - g^{(2i-1)}(0) = -\frac{\Gamma(2l-\frac{3}{2}+2i)}{\Gamma(2l-\frac{1}{2})}$$

gilt.
Wir betrachten $g^{(2i-1)}(0)$. Für alle $m \in \mathbb{N}_0$ mit $m < p-2l-\frac{1}{2}$ gilt $g^{(m)}(0) = 0$ wegen $\sigma(0) = 1$ und $\sigma^{(l)}(0) = 0$ für alle $l = 1, \dots, p-1$. Aus der Leibniz-Regel folgt

$$g^{(2i-1)}(\eta) = c_1 \frac{\sigma^{(2i-1)}(\eta)}{\eta^{2l-\frac{1}{2}}} + \cdots + c_{2i} \frac{\sigma(\eta)}{\eta^{2l-\frac{1}{2}+2i-1}}.$$

mit Konstanten $c_1, \dots, c_{2i} \in \mathbb{R}$. Um den Wert bei $\eta = 0$ zu erhalten, analysieren wir jeden einzelnen Summanden. Da $\sigma^{(2i-1)} \in C^{p-2i+1}([0,1])$ gilt, ist die Regel von L'Hospital nur noch $p-2i+1$ mal anwendbar für alle $i \in 1, 2, \dots, n_1 - 1$. Wir betrachten einen der Terme für $i = n_1 - 1$. Nach Voraussetzung gilt $2n_1 < p + \frac{7}{2} - 2l \iff 2l - \frac{1}{2} - p + 2n_1 - 3 < 0$ und es ergibt sich

$$\lim_{\eta\to 0} c_1 \frac{\sigma^{(2n_1-3)}(\eta)}{\eta^{2l-\frac{1}{2}}} \overset{\text{L'Hospital}}{=} \lim_{\eta\to 0} \tilde{c_1} \frac{\sigma^{(p)}(\eta)}{\eta^{2l-\frac{1}{2}-p+2n_1-3}} = 0,$$

Für den Summanden

$$\lim_{\eta\to 0} c_{2n_1-2} \frac{\sigma(\eta)}{\eta^{2l-\frac{7}{2}+2n_1}} \overset{\text{L'Hospital}}{=} \lim_{\eta\to 0} \tilde{c}_{2n_1-2} \frac{\sigma^{(p)}(\eta)}{\eta^{2l-\frac{7}{2}+2i-p}} = 0$$

und alle weiteren erhält man das gleiche Ergebnis.
Für $p > 2l - \frac{1}{2}$ gilt daher $g^{(2i-1)}(0) = 0$ für alle $i = 1, \dots, n_1 - 1$.

Induktiv folgt

$$\left.\frac{\mathrm{d}^{2i-1}\eta^{-2l+\frac{1}{2}}}{\mathrm{d}\eta^{p-1}}\right|_{\eta=1} = -\frac{\Gamma(2l-\frac{3}{2}+2i)}{\Gamma(2l-\frac{1}{2})}.$$

Durch Verwendung von $\sigma^{(2i-1)}(1) = 0$ für alle $i = 1, \dots, n_1 - 1$ und der Leibniz-Regel der Differentiation von Produkten ergibt sich

$$\left.\frac{\mathrm{d}^{2i-1}\sigma(\eta)\eta^{-2l+\frac{1}{2}}}{\mathrm{d}\eta^{2i-1}}\right|_{\eta=1} = 0, \quad \forall i = 1, \cdots, n_1 - 1.$$

Wir haben gezeigt, dass in Gleichung (4.22), sowohl der erste Summand sowie die komplette Summe den Wert Null besitzt. Insgesamt erhalten wir

$$|Q_N^1(\pm 1, 0) + R_N^1(\pm 1, 0)| = \mathcal{O}(M^{-p}).$$

Studieren wir das Verhalten von $Q_N^2 + R_N^2$. Dabei wurde bereits der Fall $2l - \frac{3}{2} > p$ analysiert. Gelte jetzt $2l - \frac{3}{2} < p$. Es ergibt sich

$$\left(\frac{\lambda_{2m-1}}{\lambda_{2M}}\right)^{-l} \frac{P'_{2m-1}(0)}{\gamma_{2m-1}} = (-1)^m \left(\frac{m}{M}\right)^{-2l} \sum_{r=0}^{\infty} \tilde{f}_r m^{\frac{3}{2}-r}$$

und damit

$$Q_N^2(\pm 1, 0) + R_N^2(\pm 1, 0) = \left(\frac{1}{4}\right)^l \sum_{r=0}^{\infty} \tilde{f}_r \cdot \left[M^{-2l-r+\frac{3}{2}} \sum_{m=1}^{M} (-1)^m \tilde{g}\left(\frac{m}{M}\right) + \sum_{m=M+1}^{\infty} (-1)^m m^{-2l-r+\frac{3}{2}} \right],$$

wobei $\tilde{g}(\eta) = \frac{1-\sigma(\eta)}{\eta^{2l-\frac{3}{2}}}$ ist. Wir betrachten das Ganze für $r = 0$ mit einer analogen Argumentation wie im Fall $Q_N^1 + R_N^1$, dabei sei $\tilde{g} \in C^{2n_1}[0, 1]$ mit $n_1 \geq 2$. Es ergibt sich

$$M^{-2l+\frac{3}{2}} \sum_{m=1}^{M} (-1)^m \tilde{g}\left(\frac{m}{M}\right) + \sum_{m=M+1}^{\infty} (-1)^m m^{-2l+\frac{3}{2}} = \frac{1}{2} M^{-2l+\frac{3}{2}} (\tilde{g}(1) - \tilde{g}(0) - 1)$$
$$+ \sum_{i=1}^{n_1-1} M^{-2i+\frac{5}{2}-2l} \frac{B_{2i}}{(2i)!} (4^i - 1) \left[\frac{\Gamma(2l - \frac{5}{3} + 2i)}{\Gamma(2l - \frac{3}{2})} + g^{(2i-1)}(1) - g^{(2i-1)}(0) \right] + \mathcal{O}(M^{-2l-2n_1+\frac{5}{2}}).$$

Mit Hilfe der Filtereigenschaften folgen

$$\tilde{g}(1) = \frac{1 - \sigma(1)}{1^{2l-\frac{3}{2}}} = 1$$

und

$$\tilde{g}(0) = \lim_{\eta \to 0} \frac{1 - \sigma(\eta)}{\eta^{2l-\frac{3}{2}}} \overset{\text{L'Hospital}}{=} \lim_{\eta \to 0} \hat{c} \frac{\sigma^{(p)}(\eta)}{\eta^{2l-\frac{3}{2}-p}} = 0.$$

Aufgrund den Voraussetzungen an n_1 ist $M^{-2l-2n_1+\frac{5}{2}}$ nach oben durch M^{-p} beschränkt. Für die Betrachtung der Faktoren $g^{(2i-1)}$ gehen wir analog zum Fall $Q_N^1 + R_N^1$ vor. Die Leibniz-Regel ergibt

$$g^{(2i-1)}(\eta) = \hat{c}_1 \frac{\sigma^{(2i-1)}(\eta)}{\eta^{2l-\frac{3}{2}}} + \cdots + \hat{c}_{2i} \frac{\sigma(\eta)}{\eta^{2l-\frac{3}{2}+2i-1}}, \tag{4.23}$$

und für einen einzelnen Summanden mit $i = n_1 - 1$

$$\lim_{\eta \to 0} \hat{c}_1 \frac{\sigma^{(2i-1)}(\eta)}{\eta^{2l-\frac{3}{2}}} \overset{\text{L'Hospital}}{=} \lim_{\eta \to 0} \check{c}_1 \frac{\sigma^{(p)}(\eta)}{\eta^{2l-\frac{3}{2}+2n_1-3-p}} = 0,$$

da $2l - \frac{3}{2} + 2n_1 - 3 - 2 < 0 \iff 2n_1 < p - 2l + \frac{9}{2}$ ist. $2n_1$ wurde sogar als echt kleiner $p - 2l - \frac{7}{2}$ vorausgesetzt. Mit einer analoger Rechnung für die restlichen Terme aus (4.23)

folgt, dass $\tilde{g}^{(2i-1)}(0) = 0$ für alle $i = 1, \ldots, n_1 - 1$ gilt.
Induktiv folgt erneut

$$\left.\frac{\mathrm{d}^{2i-1}\eta^{-2l+\frac{3}{2}}}{\mathrm{d}\eta^{p-1}}\right|_{\eta=1} = -\frac{\Gamma(2l-\frac{5}{2}+2i)}{\Gamma(2l-\frac{3}{2})},$$

und aufgrund der Voraussetzung $\sigma^{(2i-1)}(1) = 0$ für alle $i = 1, \ldots, n_1 - 1$ erhält man

$$\left.\frac{\mathrm{d}^{2i-1}\sigma(\eta)\eta^{-2l+\frac{3}{2}}}{\mathrm{d}\eta^{2i-1}}\right|_{\eta=1} = 0 \quad \forall i = 1, \cdots, n_1 - 1.$$

Insgesamt ist somit

$$|Q_N^2(\pm 1, 0) + R_N^2(\pm 1, 0)| = M^{-p}$$

gezeigt.

Wir haben das Verhalten von Q_N^1, Q_N^2, R_N^1 und R_N^2 vollständig erklärt. Bevor wir jetzt

$$u(\pm 1) - F_N u_N(\pm 1)$$

analysieren, definieren wir zunächst noch eine Größe, welches uns ein Maß für die Glattheit der Funktion u liefert.

Definition 4.9. Man nennt

$$k_0 := \inf\{k \in \mathbb{N} : |L_\omega^k u(c^-) - L_w^k u(c^+)| \neq 0\}$$

das **Regularitätsmaß** (measure of regularity).
Weiterhin ist die **gebrochene Sobolev-Norm** gegeben durch

$$|||u|||_{2q} := \left(||u||^2_{H^{2q}([-1,c^-[,\omega)} + ||u||^2_{H^{2q}(]c+,1],\omega)}\right)^{\frac{1}{2}}.$$

Im Fall der Legendre-Polynome ist bekanntlich die Gewichtsfunktion $\omega(x) \equiv 1$. Untersucht man jetzt Gleichung (4.14) in den Punkten $x = \pm 1$ mit einer Unstetigkeit der Funktion u im Nullpunkt $c = 0$, so gilt

$$\begin{aligned}
&u(\pm 1) - F_N u_N(\pm 1) \\
&= \sum_{l=0}^{q-1} \left[L^l u(0^-) - L^l u(0^+)\right] \left[\frac{\partial}{\partial s} G^{l+1}(\pm 1, s) - \frac{\partial}{\partial s} K_N^{l+1}(\pm 1, s)\right]\Big|_{s=0} \\
&- \sum_{l=0}^{q-1} \frac{\partial}{\partial s} \left[L^l u(0^-) - L^l u(0^+)\right] \left[G^{l+1}(\pm 1, 0) - K_N^{l+1}(\pm 1, 0)\right] \\
&+ \int_{-1}^{1} L^q u(s) \left[G^q(\pm 1, s) - K_N^q(\pm 1, s)\right] \mathrm{d}s.
\end{aligned} \tag{4.24}$$

Ist das Regularitätsmaß k_0 größer oder gleich $2q$, so sind die Summen in Gleichung (4.24) gleich 0. Es ist somit nur noch das Verhalten von

$$\int_{-1}^{1} L^q u(s) \left[G^q(\pm 1, s) - K_N^q(\pm 1, s)\right] \, \mathrm{d}s$$

zu analysieren. Man spaltet es auf und verwendet Satz 4.7 der Jacobi-Polynome für eine hinreichend glatte Funktion. Es gilt

$$\int_{-1}^{0} L^q u(s) \left[G^q(\pm 1, s) - K_N^q(\pm 1, s)\right] \mathrm{d}s + \int_{0}^{1} L^q u(s) \left[G^q(\pm 1, s) - K_N^q(\pm 1, s)\right] \mathrm{d}s \\ \leq C|||u|||_{2q} N^{1-2q}.$$

Mit $N = 2M$ und $p = 2q$ folgt für das Approximationsverhalten

$$|u(\pm 1) - F_N u_N(\pm 1)| \leq \mathcal{O}(M^{1-p}). \tag{4.25}$$

Ist nun $k_0 < 2q$, so existiert ein q_1, ab dem alle Summanden in (4.24) ungleich Null sind. Aus unserer Untersuchung von $|Q_N^1 + R_N^1|$ bzw. $|Q_N^2 + R_N^2|$ weisen diese ein Verhalten von $\mathcal{O}(M^{-2q})$ auf. Fasst man sie erneut zusammen, erhält man ein Approximationsverhalten von $\mathcal{O}(M^{1-2q})$ und mit $2q = p$ folgt entsprechend (4.25).

Unter Berücksichtigung aller Voraussetzungen ergibt sich der folgende Satz aus [36].

Satz 4.10. *Die Funktion u habe das Regularitätsmaß k_0 und besitzt außschließlich im Punkt $c = 0$ eine Unstetigkeitsstelle. Ansonsten sei u hinreichend glatt und N gerade. Weiterhin sei ein Filter $\sigma(\eta)$ der Ordnung p, $p > 1$ gegeben mit den Zusatzbedingungen (4.21) an den Filter sowie $g(\eta)$, $\tilde{g}(\eta) \in C^{2n_1}([0, 1])$. Dann gilt die Abschätzung*

$$|u(\pm 1) - F_N u_N(\pm 1)| \leq \hat{C}|||u|||_p M^{1-p}$$

mit einer Konstanten $\hat{C} \in \mathbb{R}^+$ und $2M = N$.

Bemerkung. Der Satz 4.10 ist äquivalent zu dem Theorem 4.2 aus [36]. Hier stellen die Autoren noch die abschließende Vermutung auf, dass man das Resultat auch für eine Unstetigkeitsstelle $c \in (-1, 1)$ und beliebige Punkte $x \in [-1, 1]$ mit $x \neq c$ erweitern könnte. Die Konstante $\hat{C}$ sollte dann vom Betrag $|x - c|$ abhängig sein. Das wäre schließlich das äquivalente Ergebnis zu (4.4). Der Ansatz führt jedoch zu keinem Ergebnis, da die Legendre-Polynome nicht translationsinvariant sind. Es verändert sich die komplette Untersuchung, wenn die Unstetigkeitesstelle nicht mehr im Nullpunkt liegt. Bei der Analyse müssten zusätzlich die Polynome ungeraden Grades mitbetrachtet werden. Nur im Falle $c = 0$ verschwinden nach Lemma 4.8 alle ultrasphärischen Polynome ungeraden Grades. Demzufolge würde man analoge Lemmata zu 8.5 und 8.6 für die Darstellung der Polynome ungeraden Grades benötigen. Das ist auch einer der Gründe, warum eine Übertragung dieses Ansatzes auf allgemeine Jacobi-Polynome nicht möglich ist, da nur

bei gleicher Parameterwahl $\alpha = \beta$ Polynome ungeraden Grades im Nullpunkt verschwinden. Versucht man den Satz 4.25 ausschließlich auf Gegenbauer-Polynome zu erweitern, wird man Fallunterscheidungen bezüglich des Parameters α und des Laufindexes m vornehmen müssen. Die Werte der ultrasphärischen Polynome in den Punkten $x = \pm 1$ sind nicht mehr $(\pm 1)^n$ wie bei den Legendre-Fall, sondern sind vom Parameter α abhängig. Diese Tatsache hat auch direkten Einfluss auf die weitere Darstellung und führt schließlich zu Problemen in Lemma 8.5 und Lemma 8.6. Man braucht analoge Resultate zu Lemma 8.5 und Lemma 8.6, wenn zusätzliche Terme abhängig vom Parameter α in der Reihe enthalten sind. Man benötigt ein besseres Verständnis der auftretenden unendlichen Reihen und der Koeffizienten.
Kommen wir nochmals auf Satz 4.10 und seinem Beweis zu sprechen. Die Autoren machen sich in [36] keinerlei Gedanken darüber, welche Eigenschaften und Voraussetzungen sie an den Filter σ und an die einzelnen Funktionen stellen müssen, damit alle ihre Rechnungen bzw. Ausdrücke wohldefiniert sind. Sie betrachten beispielsweise die Funktion

$$g(\eta) = \frac{1 - \sigma(\eta)}{\eta^{2l-\frac{1}{2}}}$$

mit einem Filter der Ordnung p. Für die Analyse des Approximationsverhaltens wenden sie Lemma 8.5 an. Das Lemma hat allerdings als Voraussetzung $g(\eta) \in C^{2n}([0,1])$. Im Fall $2l - \frac{1}{2} < p$ gilt das offensichtlich nicht. Bereits für $n = 1$ mit $l = 1$ und $p = 2$ folgt:

$$g(\eta) = \frac{1 - \sigma(\eta)}{\eta^{\frac{3}{2}}} \text{ und } g'(\eta) = -\frac{\sigma'(\eta)}{\eta^{\frac{3}{2}}} - \frac{\frac{3}{2}(1 - \sigma(\eta))}{\eta^{\frac{5}{2}}}.$$

Für $\eta \to 0$ gilt mit L'Hospital

$$\lim_{\eta \to 0} g'(\eta) = \lim_{\eta \to 0} -\frac{\sigma'(\eta)}{\eta^{\frac{3}{2}}} - \frac{\frac{3}{2}(1 - \sigma(\eta))}{\eta^{\frac{5}{2}}} \overset{\text{L'Hospital}}{=} \lim_{\eta \to 0} -\frac{\overbrace{\sigma^{(2)}(\eta)}^{<c}}{\underbrace{\frac{3}{2}\eta^{\frac{1}{2}}}_{\to 0}} + \frac{\frac{3}{2}\sigma'(\eta)}{\frac{5}{2}\eta^{\frac{3}{2}}} = -\infty.$$

Die Funktion $g'(\eta)$ besitzt somit einen Pol im Nullpunkt und entspricht nicht einmal $C^1([0,1])$. Die Verwendung der Formel aus Lemma 8.5 ist daher nicht zulässig. Hier müssen die Autoren genauer vorgehen und gegebenenfalls eine zusätzliche Bedingung an den Filter stellen, wie wir es auch getan haben.
Bei genauerer Betrachtung unserer Bedingungen für die Funktionen $g(\eta)$ und $\tilde{g}(\eta)$ können wir feststellen, dass die Funktionen allgemein die Bedingung $C^{2n_1}([0,1])$ nicht mehr erfüllen bzw. nur dann erfüllen, wenn die Funktion σ glatter in Null läuft, als die Ordnung des Filters es eigentlich verlangte. Man müsste $\sigma \in C^{\hat{p}}([0,1])$ mit $\hat{p} > p$, $\sigma^{(\hat{p}-1)}(0) = 0$ und $\sigma^{(\hat{p}-1)}(1) = 0$ fordern. Dies ist allerdings gleichbedeutend damit, dass der Filter σ die Ordnung $\hat{p}$ nach Definition 4.2 hat. Das Approximationsverhalten ist allerdings nur $\mathcal{O}(M^{1-p})$. Dieser Ansatz liefert also ebenfalls kein äquivalentes Ergebnis zu (4.4). Eine Lösung für das Problem könnte die Untersuchung von $Q_N^1 + R_N^1$ bzw. $Q_N^2 + R_N^2$ für

den Fall sein, indem g bzw. $\tilde{g}$ nicht mehr Element von $C^{2n_1}([0,1])$ sind. Betrachten wir $g_1(\eta) = \frac{1-\sigma(\eta)}{\eta^{2l-p-\frac{1}{2}}}$ mit $2l = p$. Dann gilt $g_1 \notin C^{2n_1}([0,1])$ mit $n_1 \geq 2$. Wir entwickeln in Q_N^1 und in R_N^1 den Filter $\sigma(\eta)$ in seine Taylor-Reihe im Nullpunkt und erhalten:

$$Q_N^1(\pm 1, 0) = \left(\frac{1}{4}\right)^l \sum_{r=0}^{\infty} \hat{f}_r M^{-p} \left[\sum_{m=1}^{M} \frac{(-1)^m \sigma^{(p)}(\xi_m)}{(p)! m^{-\frac{1}{2}+r}}\right],$$
$$R_N^1(\pm 1, 0) = \left(\frac{1}{4}\right)^l \sum_{r=0}^{\infty} \hat{f}_r \sum_{m=M+1}^{\infty} (-1)^m m^{-p-r+\frac{1}{2}}.$$

Fassen wir das zusammen, ergibt sich die Abschätzung

$$|Q_N^1(\pm 1, 0) + R_N^1(\pm 1, 0)| = \mathcal{O}(M^{-p+\frac{1}{2}}).$$

Man betrachtet analog die Fälle für $g, \tilde{g} \notin C^{2n_1}([0,1])$ und es ergibt sich allgemein ein schlechteres Approximiationsresultat als das Resultat des Satz 4.10. Welches Approximationsverhalten man schließlich erhält, hängt von der Funktion u, dem Regularitätsmaß k_0, dem Filter σ und seiner Ordnung p ab. Man benötigt eine genauere Analyse, um das tatsächliche Approximiationsverhalten anzugeben, das geht aber über den Rahmen dieser Arbeit hinaus.

5 Diskrete orthogonale Polynome

In den vorherigen Kapiteln haben wir uns ausschließlich mit klassischen orthogonalen Polynomen einer bzw. zweier Variablen beschäftigt, neue Approximationsresultate bewiesen und ihre Anwendung bei spektralen Verfahren erläutert. Neben den klassischen orthogonalen Polynomen sind die diskreten orthogonalen Polynome ein weiterer Teil der allgemeinen Theorie orthogonaler Polynome. Wir definieren die Hahn-Polynome als Beispiel diskreter orthogonaler Polynome, wiederholen einige grundlegende Eigenschaften der Hahn-Polynome und zeigen ihren Zusammenhang mit den Jacobi-Polynomen. Wir beweisen anschließend spektrale Konvergenz für die Fourier-Koeffizienten einer Funktion $u \in C^\infty$, die bezüglich der Hahn-Polynome in eine Fourier-Reihe entwickelt wurde. Bei Verwendung der Hahn-Polynome in einem spektralen Verfahren kann es zu Problemen kommen, die ihren Ursprung nicht im Gibbs'schen Phänomen, sondern im Runge-Phänomen haben. Wir diskutieren einen Ansatz, das Runge-Phänomen zu vermeiden, indem man diskrete orthogonale Polynome auf nicht gleichverteilten Gittern verwendet. Zum Abschluss geben wir noch einen Ausblick auf die zweidimensionalen Hahn-Polynome, die ihre Orthogonalitätsbeziehung auf einem Dreiecksgitter erfüllen.
Wir verwenden die heutigen Standardnotationen und Definitionen, wie man sie in dem Artikel [3] und allen neueren Lehrbüchern [45] und [62] findet. Es sei weiterhin auf den Klassiker [59] verwiesen, der eine hervorragende Einführung in die Theorie diskreter orthogonaler Polynome liefert.

5.1 Hahn-Polynome

Diskrete orthogonale Polynome besitzen in unterschiedlichen Teilbereichen der Mathematik, wie etwa in der Biomathematik [42] und der Wahrscheinlichkeitstheorie [59], eine Anwendung. Auch numerische Verfahren für partielle Differentialgleichungen verwenden diskrete orthogonale Polynome. In [25] werden sie als Alternative zur Gegenbauer-Rekonstuktion [31] untersucht und verwendet. In [26] benutzen die Autoren diskrete orthogonale Polynome in einem numerischen Verfahren zur Berechnung der Lösung u. Die Verwendung diskreter orthogonaler Polynome in einem spektralen Verfahren kann durch den Projektionsansatz[1] motiviert werden. Bei der Berechnung der Fourier-Koeffizienten $\hat{f}$ bzw. $\hat{u}$ wird stets ein Integral der Form (2.13) ausgewertet. Dies erfolgt in der Praxis mit Hilfe eines Quadraturverfahrens. Wenn man allerdings anstelle des Integrals eine

[1] Vergleiche hierzu Kapitel 2: Berechnung der Koeffizienten.

Summe als Innenprodukt hat, würde man bei der Berechnung der Koeffizienten keinen numerischen Fehler machen. Das ist nur ein Grund, warum man die Verwendung von diskreten orthogonalen Polynomen in einem spektralen Verfahren in Betracht ziehen sollte. Dieser Abschnitt soll entsprechend eine erste Grundlage für weitere Forschungstätigkeiten in diese Richtung schaffen. Wir beschränken uns auf die **Hahn-Polynome**, wiederholen einige elementare Eigenschaften und stellen ihren Zusammenhang mit den Jacobi-Polynomen dar.

5.1.1 Definition und Eigenschaften

Hahn entdeckte 1949 orthogonale Polynome als Lösungen von q-Differenzengleichungen [33]. Diese Familien orthogonaler Polynome fasst man heute unter den Namen **Hahn-Klasse** zusammen und die **Hahn-Polynome** sind ein Spezialfall. Jedoch war Hahn nicht der Erste, der mit dieser Polynomfamilie arbeitete. Bereits Chebyshev beschäftigte sich mit den Hahn-Polynomen [10] und konnte elementare Eigenschaften wie die Rodriguez-Formel zeigen. Wir beginnen mit der Definition der Hahn-Polynome aus [45].

Definition 5.1. Seien $N \in \mathbb{N}$, $-1 < \alpha, \beta \in \mathbb{R}$, das Intervall $I = [0, N]$ und die $(N+1)$-äquidistante Punkteverteilung im Intervall I gegeben. Die **Hahn-Polynome** sind im Intervall I durch die hypergeometrische Funktion

$$Q_n(x; \alpha, \beta, N) := {}_3F_2\left(-n, n+\alpha+\beta+1, -x; \alpha+1, -N; 1\right)$$

mit $n = 0,\ 1,\ \ldots,\ N$ definiert[2].

Nachfolgend zählen wir einige elementare Eigenschaften der Hahn-Polynome auf, die wir im weiteren Verlauf noch benötigen. Man findet sie in der Literatur [3], [45] und den Referenzen dort. Nachfolgend verwenden wir $Q_n(x) := Q_n(x; \alpha, \beta, N)$.

- Seien $\alpha > -1$ und $\beta > -1$, so erfüllen die Hahn-Polynome die Orthogonalitätsbeziehung

$$\begin{aligned} < Q_n(x); Q_m(x) >_w : &= \sum_{x=0}^{N} Q_n(x) Q_m(x) \binom{\alpha+x}{x} \binom{\beta+N-x}{N-x} \\ &= \frac{(-1)^n (n+\alpha+\beta+1)_{N+1} (\beta+1)_n n!}{(2n+\alpha+\beta+1)(\alpha+1)_n (-N)_n N!} \delta_{mn}, \end{aligned} \tag{5.1}$$

  für alle $m,\ n \in \mathbb{N}_0$ mit $m, n \leq N$.

- Die Rekursionsformel lautet

$$-xQ_n(x) = A_n Q_{n+1}(x) - (A_n + C_n) Q_n(x) + C_n Q_{n-1}(x) \tag{5.2}$$

[2]Die Orthogonalitätsbeziehung wird auch erfüllt von $\alpha < -N,\ \beta < -N$, siehe dazu [45].

mit

$$A_n = \frac{(n+\alpha+\beta+1)(n+\alpha+1)(N-n)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)},$$
$$C_n = \frac{n(n+\alpha+\beta+N+1)(n+\beta)}{(2n+\alpha+\beta)(2n+\alpha+\beta+1)}.$$

- Die Hahn-Polynome erfüllen die Differenzengleichung

$$\lambda_n Q_n(x) = B(x)Q_n(x+1) - [B(x)+D(x)]Q_n(x) + D(x)Q_n(x-1) \tag{5.3}$$

mit Eigenwert $\lambda_n = n(n+\alpha+\beta+1)$ und

$$B(x) = (x+\alpha+1)(x-N),$$
$$D(x) = x(x-\beta-N-1).$$

Der Unterschied in der Definition und Notation der Hahn-Polynome zwischen dem hier verwendeten Ansatz und denen in der älteren Literatur [59] genutzten, erkennt man am einfachsten an der Definition der hypergeometrischen Funktion. In [59] sind die Hahn-Polynome gegeben durch

$$h_n^{(\alpha,\beta)}(x,N) = \frac{(-1)^n}{n!}(N-n)_n(\beta+1)_n\, {}_3F_2\left(-n, n+\alpha+\beta+1, -x; \beta+1, 1-N; 1\right).$$

Sie erfüllen die Orthogonalitätsbeziehung für das Skalarprodukt

$$\sum_{x=0}^{N-1} h_n^{(\alpha,\beta)}(x,N) h_m^{(\alpha,\beta)}(x,N)\varrho(x) = 0,$$

mit $n \neq m$ und der Gewichtsfunktion

$$\varrho(x) = \frac{\Gamma(N+\alpha-x)\Gamma(\beta+1+x)}{\Gamma(x+1)\Gamma(N-x)}.$$

Betrachtet man die Gewichtsfunktion aus (5.1), so gilt

$$\omega(x) = \frac{(\alpha+1)_x(\beta+1)_{N-x}}{x!(N-x)!} = \frac{\Gamma(\alpha+1+x)\Gamma(\beta+1+N-x)}{\Gamma(x+1)\Gamma(N+1-x)\Gamma(\alpha+1)\Gamma(\beta+1)}.$$

Vergleicht man die Skalarprodukte und die Definitionen der Hahn-Polynome, so werden folgende Unterschiede deutlich:

- Im Skalarprodukt (5.1) summiert man $(N+1)$ Terme auf. In der älteren Version werden nur N Terme im Skalarprodukt summiert. Dies spielt auch bei den Gewichtsfunktionen $\omega(x)$ und $\varrho(x)$ eine Rolle.
- Die Gewichtsfunktion $\omega(x)$ beinhaltet zusätzlich den Faktor $(\Gamma(\alpha+1)\Gamma(\beta+1))^{-1}$ für eine veränderte Normierung.

- Die Parameter α und β sind in $h_n^{(\alpha,\beta)}(x,N)$ und $Q_n(x;\alpha,\beta,N)$ vertauscht.

All diese Faktoren haben Auswirkungen auf die Darstellung und Normierung der Hahn-Polynome. Im nächsten Beispiel findet man sowohl $Q_n(x)$ als auch $h_n^{(\alpha,\beta)}(x,N)$ für eine gegebene Parameterwahl.

BEISPIEL 5.2. Für $N = 20$, $\alpha = \beta = 0$ und Schrittweite $h = 1$ berechnen wir die diskreten orthogonalen Polynome $Q_n(x)$ und $h^{(0,0)}(x,N)$ bis Grad 2. Man nennt die Hahn-Polynome bei dieser Parameterwahl **diskrete Chebyshev-Polynome**. Sie sind das diskrete Gegenstück der Legendre-Polynome. Aufgrund unserer Beobachtung berechnen wir zusätzlich $h^{(0,0)}(x,21)$, um es mit den Polynomen $Q_n(x;0,0,20)$ zu vergleichen. In nachfolgender Tabelle 5.1 sind alle drei Polynome mittels Mathematica berechnet worden.

n	$Q_n(x)$	$h_n^{(0,0)}(x,20)$	$h_n^{(0,0)}(x,21)$
0	1	1	1
1	$-\frac{1}{10}x+1$	$2x-19$	$2x-20$
2	$\frac{3}{190}x^2-\frac{6}{19}x+1$	$6x^2-114x+342$	$6x^2-120x+160$

Tabelle 5.1: *Diskrete Chebyshev-Polynome*

Kommen wir zum Zusammenhang der Hahn-Polynome mit den Jacobi-Polynomen.

Satz 5.3. *Für die Hahn-Polynome gilt folgende Grenzwertbeziehung:*

$$\lim_{N\to\infty} \mathrm{Q}_n(xN;\alpha,\beta,N) = \frac{P_n^{(\alpha,\beta)}(1-2x)}{P_n^{(\alpha,\beta)}(1)},$$

mit $x \in [0,1]$.

Beweis. Wir verwenden die Darstellung der Hahn-Polynome mittels der hypergeometrischen Funktion aus Definition 5.1.

$$\begin{aligned}\lim_{N\to\infty} \mathrm{Q}_n(xN;\alpha,\beta,N) &= \lim_{N\to\infty} {}_3F_2\left(-n, n+\alpha+\beta+1, -Nx; \alpha+1, -N; 1\right)\\ &= \lim_{N\to\infty} \sum_{i=0}^{\infty} \frac{(-n)_i (n+\alpha+\beta+1)_i (-xN)_i}{(\alpha+1)_i (-N)_i i!}.\end{aligned}$$

Für $i > n$ ist $(-n)_i = 0$. Die Reihe ist für jedes $x \in [0,1]$ absolut konvergent und man kann Grenzwert und Summation vertauschen. Wir betrachten nur noch die Summanden, die von N abhängig sind. Mit der Grenzwertuntersuchung

$$\lim_{N\to\infty} \frac{(-xN)_i}{(-N)_i} = \lim_{N\to\infty} \prod_{k=1}^{i} \frac{-xN+k-1}{-N+k-1} = x^i$$

folgt

$$\lim_{N\to\infty} \mathrm{Q}_n(xN;\alpha,\beta,N) = \sum_{i=0}^{\infty} \frac{(-n)_i(n+\alpha+\beta+1)_i x^i}{(\alpha+1)_i i!} = {}_2F_1(-n,n+\alpha+\beta+1;\alpha+1;x)$$
$$= \frac{\binom{n+\alpha}{n}}{\binom{n+\alpha}{n}} {}_2F_1(-n,n+\alpha+\beta+1;\alpha+1;x) = \frac{P_n^{(\alpha,\beta)}(1-2x)}{P_n^{(\alpha,\beta)}(1)},$$

dabei nutzen wir beim letzten Gleichheitszeichen die Definition der Jacobi-Polynome 3.1 mittels der hypergeometrischen Funktion und ihren Wert bei Eins (3.3).

□

Nicht nur die Jacobi-Polynome sind Grenzwerte der Hahn-Polynome. Auch für die Meixner- und Krawtchok-Polynome findet man ähnliche Grenzwertprozesse. Eine Einordnung der Hahn-Polynome und weiterer orthogonaler Polynome (diskreter und kontinuierlicher) findet man im Askey-Schema 8.1.

5.1.2 Spektrale Konvergenz der Hahn-Polynome

Diskrete orthogonale Polynome können zur Nachbearbeitung einer numerischen Lösung u verwendet werden, vergleiche [25] und [27]. Die Verwendung diskreter orthogonaler Polynome stellt eine Alternative zur Gegenbauer-Rekonstruktion [31] dar. Bei der Gegenbauer-Rekonstruktion handelt es sich um ein Verfahren zur Reduktion des Gibbs'schen Phänomens. Die numerisch berechnete Lösung u einer partiellen Differentialgleichung wird in eine Fourier-Reihe bezüglich der Gegenbauer-Polynome entwickelt. Als neue Lösung v der Differentialgleichung verwendet man eine Partialsumme der Gegenbauer-Reihe. Unter gewissen Voraussetzungen, auf die wir nicht näher eingehen werden, kann man durch die Entwicklung der Lösung u in Gegenbauer-Polynome das in u auftretende Gibbs'sche Phänomen reduzieren bzw. bestenfalls entfernen, vergleiche [31]. In den Arbeiten [25] und [27] wird gezeigt, dass man die Rekonstruktion der Lösung u unter bestimmten Annahmen nochmals verbessern kann, falls man anstelle der ultrasphärischen Polynome diskrete orthogonale Polynome benutzt. Wir untersuchen das allgemeine Approximationsverhalten der abgeschnittenen Fourier-Reihe der Hahn-Polynome. Wir betrachten eine Funktion $u : I \to \mathbb{R}$, die auf einem kompakten Intervall $I = [a,b]$ mit $a,b \in \mathbb{R}$ definiert ist, und approximieren die Funktion u mit ihrer abgeschnittenen Hahn-Reihe

$$u(x) \approx \sum_{n=0}^{m} \hat{u}_n \tilde{Q}_n(x)$$

mit $m \leq N$. Die $\tilde{Q}_n(x)$ sind die Hahn-Polynome, die durch eine Transformation auf das Intervall $I = [a,b]$ verschoben wurden. Wir weisen spektrale Konvergenz nach, was nach Kapitel 3 bedeutet, dass für eine Funktion $u \in C^\infty(I)$ die Koeffizienten $|\hat{u}_n|$ schneller gegen Null konvergieren als jede Potenz n^{-k} für $k \in \mathbb{N}$. Es ist zu beachten, dass die

Hahn-Polynome nur bis zum Grad N definiert sind. Daher können die Koeffizienten $\hat{u}_n$ bis maximal $n = N$ berechnet werden. Für $n > N$ setzen wir die Koeffizienten $\hat{u}_n$ gleich Null. Das Abklingverhalten von $|\hat{u}_n|$ für $n \leq N$ mit einem festen N ist daher entscheidend.

Die klassischen orthogonalen Polynome einer Variablen sind Lösung eines singulären Sturm-Liouville-Problems und somit Eigenfunktionen eines selbstadjungierten Differentialoperators. Die Selbstadjungiertheit war eine wichtige Eigenschaft bei dem Nachweis der spektralen Konvergenz der Fourier-Koeffizienten, vergleiche Kapitel 3.1. Die Hahn-Polynome erfüllen nur eine Differenzengleichung (5.3) zweiter Ordnung.
Mit Hilfe der Vorwärts- und Rückwärtsdifferenzenoperatoren

$$\Delta f(x) = f(x+1) - f(x),$$
$$\nabla f(x) = f(x) - f(x-1),$$

und ihren Identitäten

$$\Delta f(x) = \nabla f(x+1), \tag{5.4}$$
$$\Delta\left[f(x)g(x)\right] = f(x)\Delta g(x) + g(x+1)\Delta f(x), \tag{5.5}$$
$$\nabla\left[f(x)g(x)\right] = f(x-1)\nabla g(x) + g(x)\nabla f(x), \tag{5.6}$$

für die Schrittweite 1 kann man (5.3) in selbstadjungierter Form schreiben

$$\Delta[-D(x)\omega(x)\nabla Q_n(x)] + \lambda_n(x)\omega(x)Q_n(x) = 0, \tag{5.7}$$

mit der Gewichtsfunktion

$$\omega(x) = \frac{\Gamma(\alpha+1+x)\Gamma(\beta+1+N-x)}{\Gamma(x+1)\Gamma(N+1-x)\Gamma(\alpha+1)\Gamma(\beta+1)}.$$

Diese erhält man durch die folgende Umformung:
Aus (5.3) folgt die Darstellung

$$-D(x)\Delta\nabla Q_n(x) + [D(x) - B(x)]\Delta Q_n(x) + \lambda_n Q_n(x) = 0.$$

Multipliziert man die Gleichung mit $\omega(x)$ und nutzt die Identität (5.4), so gilt

$$\begin{aligned} &-D(x)\omega(x)\Delta\nabla Q_n(x) + [D(x) - B(x)]\omega(x)\Delta Q_n(x) + \lambda_n\omega(x)Q_n(x) = 0 \\ \iff &-D(x)\omega(x)\Delta\nabla Q_n(x) + \nabla Q_n(x+1)[D(x) - B(x)]\omega(x) + \lambda_n\omega(x)Q_n(x) = 0. \end{aligned}$$

Weiterhin verifiziert man

$$[D(x) - B(x)]\omega(x) = \Delta\left[-D(x)\omega(x)\right]. \tag{5.8}$$

Es gilt mit (5.5):

$$\begin{aligned} &- D(x+1)\omega(x+1) + D(x)\omega(x) = [D(x) - B(x)]\omega(x) \\ \iff &D(x+1)\omega(x+1) = B(x)\omega(x) \\ \iff &\frac{\omega(x+1)}{\omega(x)} = \frac{B(x)}{D(x+1)}. \end{aligned}$$

Es gilt für den Term $\omega(N+1)=0$. Berechnet man den Term $\frac{\omega(x+1)}{\omega(x)}$, so erhält man

$$\begin{aligned}\frac{\omega(x+1)}{\omega(x)} =& \frac{\Gamma(N+1+\beta-x-1)(\alpha+1+x)\Gamma(n+1-x)}{(x+1)\Gamma(N+1+\beta-x)\Gamma(N+1-x-1)} \\ =& \frac{(\alpha+1+x)(N+1-x-1)}{(x+1)(N+\beta-x} = \frac{B(x)}{D(x+1)}.\end{aligned}$$

Damit ist Gleichung (5.8) gezeigt und es ergibt sich

$$\begin{aligned}&-D(x)\omega(x)\Delta\nabla Q_n(x) + \nabla Q_n(x+1)[D(x)-B(x)]\omega(x) + \lambda_n\omega(x)Q_n(x) = 0 \\ \iff& -D(x)\omega(x)\Delta\nabla Q_n(x) + \nabla Q_n(x+1)\Delta\left[-D(x)\omega(x)\right] + \lambda_n\omega(x)Q_n(x) = 0 \\ \overset{(5.5)}{\iff}& \Delta[-D(x)\omega(x)\nabla Q_n(x)] + \lambda_n\omega(x)Q_n(x) = 0.\end{aligned}$$

Die selbstadjungierte Form der Differenzengleichung verwenden wir im Beweis der spektralen Konvergenz. Wir schränken uns allerdings bezüglich des Definitionsbereiches I der Funktion u ein und betrachten das Ganze im kompakten Intervall $I=[0,N]$. Die Übertragung auf ein beliebiges kompaktes Intervall $[a,b]$ ist durch lineare Transformation möglich und der Beweis verläuft analog. Außerdem werden wir die normierten Hahn-Polynome verwenden, vergleiche [45, S.205], da dies die Rechnung weiter vereinfacht.

Satz 5.4. *Es seien $\alpha,\beta > -1$, $m,N\in\mathbb{N}$ mit $m\le N$, das Intervall $I=[0,N]$ mit einer $(N+1)$-äquidistanten Punkteverteilung mit Schrittweite $h=1$ und eine Funktion $u\in C^\infty([0,N])$ gegeben. Die Funktion u sei weiterhin $C^\infty([-1,1+N])$. $\tilde{Q}_n(x,\alpha,\beta,N)$ sind die in $[0,N]$ erzeugten normierten Hahn-Polynome von Grad $n\le N$. Die Hahn-Entwicklung der Funktion u bis zum Grad m ist gegeben durch*

$$P_m u(x) = \sum_{n=0}^{m} \hat{u}_n \tilde{Q}_n(x)$$

mit den Koeffizienten

$$\hat{u}_n = <\tilde{Q}_n; u>_\omega$$

und der Gewichtsfunktion

$$\omega(i) = \frac{\Gamma(\alpha+1+i)\Gamma(\beta+1+N-i)}{\Gamma(i+1)\Gamma(N+1-i)\Gamma(\alpha+1)\Gamma(\beta+1)}.$$

Für die Koeffizienten gilt

$$|\hat{u}_n| \le \frac{1}{n^{2k}} C_{u,k}$$

für alle $k\in\mathbb{N}_0$. Dabei ist $C_{u,k}$ eine von u und k abhängige Konstante.

Beweis. Es gilt für die Koeffizienten

$$\hat{u}_n = <\tilde{Q}_n; u>_\omega = \sum_{i=0}^{N} \omega(i)u(i)\tilde{Q}_n(i).$$

Mit der selbstadjungierten Form der Differenzengleichung (5.7) folgt

$$\hat{u}_n = \sum_{i=0}^{N} \omega(i) u(i) \tilde{Q}_n(i) = \frac{-1}{\lambda_n} \sum_{i=0}^{N} u(i) \Delta \left[-D(i)\omega(i)\nabla \tilde{Q}_n(i) \right].$$

Dabei sind die Summanden mit $\omega(N+1)$ und $-D(0)$ alle gleich 0. Wir verwenden partielle Summation

$$\sum_{i=0}^{N} f(i)\Delta g(i) = f(i)g(i)\Big|_0^{N+1} - \sum_{i=0}^{N} g(i+1)\Delta f(i),$$

und erhalten

$$\begin{aligned}
\hat{u}_n &= \frac{-1}{\lambda_n} \sum_{i=0}^{N} u(i) \Delta \left[-D(i)\omega(i)\nabla \tilde{Q}_n(i) \right] \\
&= \frac{-1}{\lambda_n} \left(\underbrace{-u(i)D(i)\omega(i)\nabla \tilde{Q}_n(i)\Big|_0^{N+1}}_{=0} - \sum_{i=0}^{N} \Delta u(i) (-D(i+1)\omega(i+1)\nabla \tilde{Q}_n(i+1)) \right).
\end{aligned}$$

Mit der Identität (5.4) und erneuter partieller Summation gilt

$$\begin{aligned}
\hat{u}_n &= \frac{1}{\lambda_n} \left(\sum_{i=0}^{N} \nabla \tilde{Q}_n(i+1) \left(-D(i+1)\omega(i+1)\nabla u\left(i+1\right) \right) \right) \\
&= \frac{1}{\lambda_n} \left(-\tilde{Q}_n(i) D(i)\omega(i)\nabla u(i)\Big|_{i=0}^{N+1} - \sum_{i=0}^{N} \tilde{Q}_n(i) \nabla \left[-D(i+1)\omega(i+1)\nabla u\left(i+1\right) \right] \right) \\
&= \frac{-1}{\lambda_n} \sum_{i=0}^{N} \omega(i) \tilde{Q}_n(i) \frac{1}{\omega(i)} \Delta \left[-D(i)\omega(i)\nabla u(i) \right] \\
&= \frac{-1}{\lambda_n} \left(\sum_{i=0}^{N} \omega(i) \tilde{Q}_n(i) \frac{1}{\omega(i)} \Delta \left[-D(i)\omega(i)\nabla u(i) \right] \right).
\end{aligned}$$

Sei $\mathrm{L}_{disk} := \frac{1}{\omega(i)} \Delta[-D(i)\omega(i)\nabla]$ ein diskreter Differenzenoperator. Wendet man die ganze Prozedur mit partieller Summation k-mal an, so folgt für $\hat{u}_n$:

$$\hat{u}_n = \frac{(-1)^k}{\lambda_n^k} \left(\sum_{i=0}^{N} \omega(i) \tilde{Q}_n(i) \mathrm{L}_{disk}^k u(i) \right).$$

Betrachtet man den Betrag von $|\hat{u}_n|$ und verwendet die Schwarz'sche-Ungleichung für die Summe, so erhält man

$$|\hat{u}_n| \leq \frac{1}{\lambda_n^k} ||\tilde{Q}_n||_\omega \left(\sum_{i=0}^{N} \omega(i) \left(\mathrm{L}_{disk}^k u(i) \right)^2 \right)^{\frac{1}{2}} \approx \frac{1}{n^{2k}} \left(\sum_{i=0}^{N} \omega(i) \left(\mathrm{L}_{disk}^k u(i) \right)^2 \right)^{\frac{1}{2}}.$$

Für festes N und $u \in C^\infty$ ist die Summe wohldefiniert und unabhängig von n. Das Abklingverhalten der $\hat{u}_n$ wird daher von $\frac{1}{n^{2k}}$ beschrieben für $n \leq N$. Für den Fall $n > N$ wird $\hat{u}_n$ auf Null gesetzt. □

Bemerkung. Es ist irreführend in diesem Zusammenhang von spektraler Konvergenz zu sprechen, da man alle Koeffizienten u_n mit $n > N$ auf Null gesetzt hat. Man sollte hier besser den Begriff der spektralen Genauigkeit verwenden. Tatsächlich ist man am Abklingverhalten der Koeffizienten speziell für kleine n interessiert. Man approximiert eine gesuchte Funktion u mit ihrer Fourier-Summe. Bestenfalls streben die Fourier-Koeffizienten für kleine n schon rasch gegen Null, wodurch man keine wesentliche Verbesserung der Approximation erreicht, wenn man zusätzliche Fourier-Terme zur Summe hinzuaddiert.

Dies ist ein erstes Approximationsergebnis der Partialsumme der Hahn-Reihe. Wie wir bereits angemerkt haben, bekommt man vergleichbare Ergebnisse, wenn man das Ganze auf einem beliebigen, aber kompakten Intervall $I = [a, b]$ mit $a, b \in \mathbb{R}$ betrachtet.
Gleichzeitig legitimiert es uns, bei der Verwendung von diskreten orthogonalen Polynomen in unserer numerischen Methode von einem spektralen Verfahren zu sprechen. Wir sind an weiteren Approximationsresultaten interessiert. Gerade das Verhalten des Abschneidefehlers in verschiedenen Normen, wie es in Kapitel 3 auch für die APK-Polynome untersucht wurde, sollte im Mittelpunkt für weitere Analysen stehen. Zwar konvergieren die Hahn-Polynome gegen die Jacobi-Polynome (unter bestimmten Voraussetzungen) und die Summe gegen den Integralwert, jedoch ist keinerlei Aussage über den Fehler und die Approximationsgeschwindigkeit des Fehlers in der diskreten L^2-Norm bzw. in der Maximumsnorm möglich. Selbst wenn wir die Konvergenz des Fehlers gegen Null annehmen, können wir trotzdem nicht auf spektrale Konvergenz schließen, da die beiden Prozesse nicht exponentiell verlaufen. Einen ersten Ansatz, der sich mit dem punktweisen Fehler beschäftigt, findet man in der Literatur [26]. Eine grundlegende Untersuchung fehlt hier jedoch noch.
Wir betrachten noch das folgende Beispiel, bevor wir im nächsten Abschnitt einen kleinen Ausblick über mögliche Ansätze zur Erweiterung der Theorie diskreter orthogonaler Polynome und ihrer Anwendung in spektralen Methoden geben.

Beispiel 5.5. Wir entwickeln die Funktion $f(x) = \sin \pi x$ im Intervall $[-1, 1]$ in ihre Fourier-Reihe bezüglich der Hahn-Polynome $Q_n(x; \alpha, \beta, N)$. Als Parameter werden $N = 20$, sowie einmal $\alpha = \beta = 0$ (blau, gepunktete) bzw. $\alpha = \beta = \frac{1}{2}$ (rot,gestrichelte) gewählt. Zur Approximation verwenden wir die Partialsumme bis zum Parameter $m = 5$. In der Abbildung 5.1 findet sich der Fehler $f(x) - \sum_{k=0}^{5} \hat{f}_k \tilde{Q}_k(x)$ über das Intervall verteilt. Auffällig ist das Ansteigen des Fehlers am jeweiligen Rand des Intervall. Die Rechnungen wurden alle mit Mathematica durchgeführt.

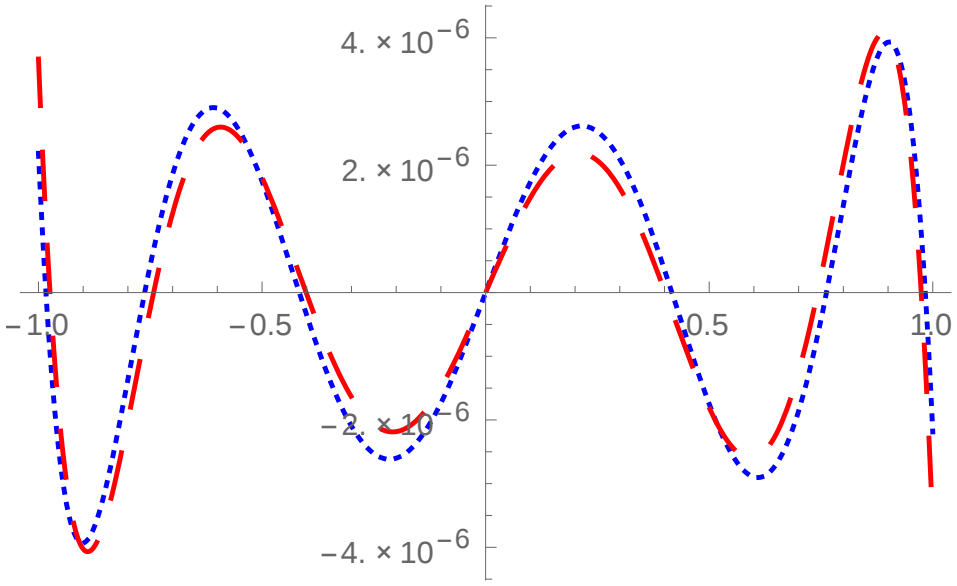


Abbildung 5.1: *Fehler der Approximation*

5.2 Erweiterung der Theorie - Ein Ausblick

5.2.1 Diskrete orthogonale Polynome auf nicht-äquidistanten Gitttern

Ein Vorteil bei der Verwendung diskreter orthogonaler Polynome in einem spektralen Verfahren liegt darin, dass man keinerlei Quadraturverfahren zur Auswertung der Integrale für die Koeffizienten bzw. der Funktionen benötigt. Man macht somit keinerlei Fehler bei der Berechnung des Skalarproduktes. Außerdem ist die Betrachtung diskreter orthogonaler Polynome eine Verallgemeinerung der kontinuierlichen Theorie orthogonaler Polynome. Dennoch muss man bei der Approximation mittels Hahn-Polynomen mehr berücksichtigen als beispielsweise bei der Verwendung von Jacobi-Polynomen. So kann man die Partialsumme maximal bis zum Grad $m \leq N$ entwickeln. Im Falle $m = N$ erhält man das Interpolationspolynom und es kann zum **Runge-Phänomen** kommen. Dabei beschreibt das Interpolationspolynom nicht mehr die approximierende Funktion u, wie man am folgenden Beispiel sieht.

Beispiel 5.6. Die Funktion $f(x) = \frac{1}{1+25x^2}$ wird mit der Partialsumme der Hahn-Reihe im Intervall $[-1, 1]$ angenähert. Wir verwenden die gleichen Parameter wie in Beispiel 5.5, m wird auf 20 gesetzt. Wir erhalten das Interpolationspolynom für eine äquidistante Punkteverteilung in $[-1, 1]$. In der Abbildung 5.2 sehen wir die Funktion $f(x)$ und die Interpolationspolynome. Es kommt hier zum klassischen Runge-Phänomen. Verwendet man die Partialsumme $m = 5$, so ergibt sich die Situation in Abbildung 5.3. Der blaue, gepunktete Graph ist die Annäherung mit den diskreten Chebyshev-Polynomen mit den Parametern $(\alpha, \beta) = (0, 0)$ und der rote, gestrichelte Graph hat als Parameter $(\alpha, \beta) = (0.5, 0.5)$. Sowohl $m = 5$ als auch $m = 20$ führen nicht dazu, dass die

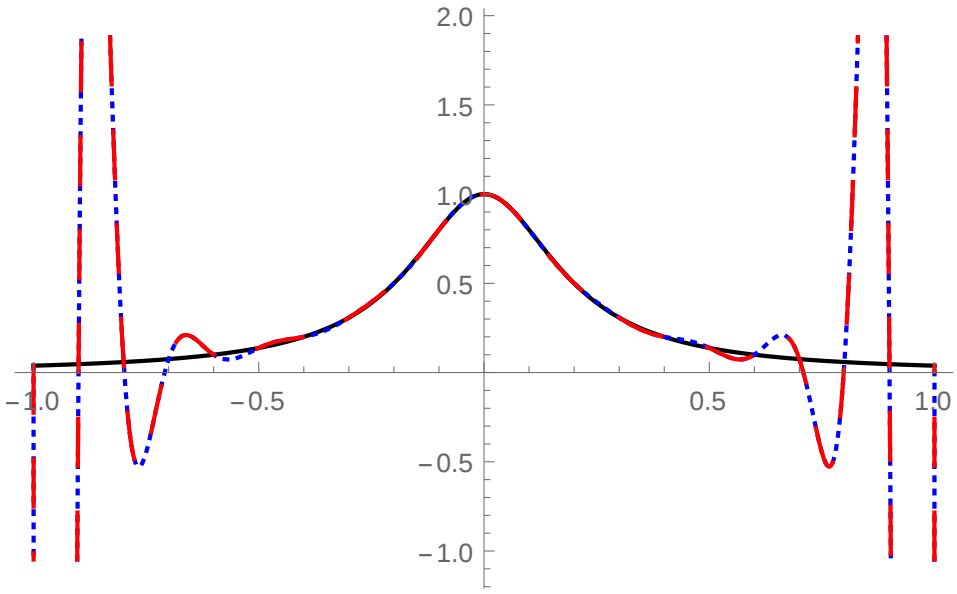


Abbildung 5.2: *Runge-Phänomen*

Funktion $f(x)$ wirklich gut beschrieben wird. Die Koeffizienten weisen jedoch spektrale Genauigkeit auf.

Bei der Approximation einer stückweise stetigen Funktion mittels Hahn-Polynomen können das Gibbs'sche Phänomen und das Runge-Phänomen einen gemeinsamen Effekt auf die Näherung haben. Dies ist ein weiteres Argument, das Verhalten des Fehlers in verschiedenen Normen zu analysieren. Die Wahl des Verhältnisses von m zu N könnte dabei von Interesse sein, wie es auch bei der Gegenbauer-Rekonstruktion der Fall ist.
Das Abschwächen der Gibbs'schen Oszillationen versuchen wir mittels Filter zu realisieren. Zur Vermeidung des Runge-Phänomens nutzt man in der Praxis nicht-äquidistante Punkteverteilungen, wie etwa Chebyshev- oder auch Gauß-Lobatto-Punkte. Die prinzipielle Idee Orthogonalität und nicht-äquidistante Punteverteilungen zu nutzen, führt schließlich zur Betrachtung **diskreter orthogonaler Polynome auf nicht gleichverteilten Gittern**. Wir stellen zwei Ansätze zur Konstruktion diskreter orthogonaler Polynome auf nicht gleichverteilten Gittern vor.

Konstruktion diskreter orthogonaler Polynome zu speziellen Punkteverteilungen

Der erste Ansatz besteht in dem Versuch sich zu einer selbst gewählten Punkteverteilung diskrete orthogonale Polynome mittels Quadraturformel zu konstruieren. In der Arbeit [20] wird das von den Autoren gezeigt. Essentiell hierfür ist der Artikel [47]. In dieser Arbeit beschreibt Koornwinder, wie man orthogonale Polynome mit der Gewichtsfunktion

$$\tilde{\omega}(x) = (1-x)^{\alpha}(1+x)^{\beta} + M\delta(x-1) + N\delta(x+1)$$

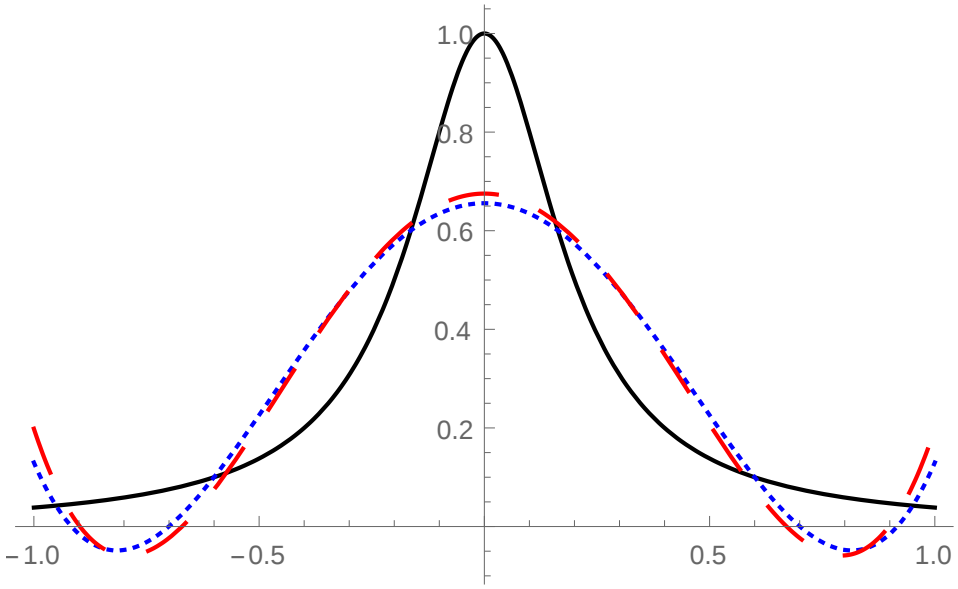


Abbildung 5.3: *Funktion und zwei Näherungen*

mittels klassischer Jacobi-Polynome ausdrückt. Die Gewichtsfunktion besteht aus der klassischen Jacobi-Gewichtsfunktion $\omega(x) = (1-x)^{\alpha}(1+x)^{\beta}$ mit $\alpha, \beta > -1$ und zwei Delta-Distributionen bei $x = -1$ und $x = 1$. Bei der Delta-Distribution handelt es sich um eine stetige lineare Abbildung von $C_0^{\infty}(A)$ nach $\mathbb{R}$. Für eine ausführliche Einführung in die Distributionentheorie verweisen wir auf das Lehrbuch [39]. Wir können hier vereinfacht annehmen, dass für eine Funktion $f \in C_0^{\infty}([-1,1])$ gilt

$$\delta(f) = \int_{-1}^{1} f(x)\delta(x)\mathrm{d}x = f(0).$$

Um das Vorgehen nun besser verständlich zu machen, fassen wir die wesentlichen Punkte des Artikel [20] nochmals stichpunktartig zusammen, ohne Rechnungen und Beweise zu wiederholen.
Ziel der Autoren ist es, zu den Gauß-Lobatto-Chebyshev-Punkten (nicht zu verwechseln mit den Gauß-Lobatto-Punkten aus Kapitel 2) oder auch extremen Chebyshev-Punkten

$$X_n = \left\{ x_k = -\cos\left(\frac{k-1}{n-1}\pi\right), \qquad k = 1, \ldots, n \right\}$$

eine Familie von diskreten orthogonalen Polynomen $P = \{p_1,\ p_2, \ldots,\ p_n\}$ mit Grad $\deg(p_k) = k-1$ zu konstruieren. Die Polynome sollen eine explizite Darstellung besitzen und die Orthogonalitätsbeziehung hinsichtlich des Innenproduktes

$$< p_i; p_j >:= \sum_{k=1}^{n} p_i(x_k)p_j(x_k) = \begin{cases} 0, & \text{für } i \neq j;\ i,\ j = 1,\ 2, \ldots,\ n \\ \rho_i \neq 0, & \text{für } i = j;\ i,\ j = 1,\ 2, \ldots,\ n. \end{cases}$$

erfüllen.
Die Autoren von [20] beginnen ihre Überlegungen mit der Definition eines Maßes μ auf $[-1,1]$. Für eine reelle Funktion f gelte

$$\int_{-1}^{1} f(x)\mathrm{d}\mu := \frac{\Gamma(\alpha+\beta+2)}{2^{\alpha+\beta+1}\Gamma(\alpha+1)\Gamma(\beta+1)} \int_{-1}^{1} f(x)(1-x)^{\alpha}(1+x)^{\beta}\mathrm{d}x + Mf(-1) + Nf(1), \tag{5.9}$$

mit $M, N \in \mathbb{R}$, $\alpha, \beta > -1$ ([47]). Die Autoren verbinden die Gleichung (5.9) mit der Gauß-Jacobi-Lobatto-Quadratur für $\alpha = \beta = -\frac{1}{2}$ in den Punkten $x_i = -\cos\left(\frac{i-1}{n-1}\pi\right)$. Es gilt

$$\frac{1}{\pi}\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x + \frac{1}{2(n-1)}\left[f(x_1)+f(x_n)\right] = \frac{1}{n-1}\sum_{i=1}^{n} f(x_i) - \frac{f^{(2n-2)}(\xi)}{2^{2n-3}(2n-2)!},$$

mit $\xi \in (-1,1)$. Mit dieser Formel werden Polynome bis zum Grad $2n-3$ exakt integriert. Wählt man sich als Innenprodukt die linke Seite davon, also

$$(f;g) := \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}}\mathrm{d}x + \frac{1}{2(n-1)}\left[f(x_1)g(x_1)+f(x_n)g(x_n)\right], \tag{5.10}$$

so gilt

$$< f;g >= (n-1)(f;g),$$

mit

$$\deg(f) + \deg(g) \leq 2n-3.$$

Eine explizite Darstellung der orthogonalen Polynome bezüglich des Innerenproduktes (5.10) kann man mit Hilfe der Arbeit [47] ableiten. Man erhält

$$p_k(x) = \frac{(n+k-2)^2}{(n-1)^2}P_{k-1}^{-\frac{1}{2},-\frac{1}{2}}(x) - \frac{(n+l-2)}{(k-1)(n-1)^2}x\frac{\mathrm{d}}{\mathrm{d}x}\left[P_{k-1}^{-\frac{1}{2},-\frac{1}{2}}(x)\right].$$

Mit bekannten Umrechnungsformeln für die Jacobi-Polynome können die Autoren von [20] eine explizite Darstellung q_k der orthogonalen Polynome folgern, und beweisen für die q_k eine 3-Term Rekursionsformel. Sie verwenden ihre Polynome q_k anschließend zum Lösen eines Minimierungsproblems und vergleichen ihr Ergebnis mit weiteren numerischen Methoden, welche ebenfalls das Minimierungsproblem lösen. Ihr Ansatz, der die Polynome q_k verwendet, ist den anderen numerischen Verfahren weit überlegen, vergleiche [20].

Das hier gezeigte Beispiel zur Konstruktion diskreter orthogonaler Polynome aus bekannten Quadraturverfahren kann als Muster genommen werden. Speziell von Vorteil ist die Auswahl der Punkteverteilungen und gegebenenfalls die Möglichkeit der Darstellung der Polynome durch Jacobi-Polynome. Allerdings ist auszuschließen, dass zu jeder

Punkteverteilung passende Quadraturformeln existieren, um so diskrete orthogonale Polynome auf diesen Punkteverteilungen zu konstruieren. Hat man dennoch einen orthogonale Polynomfamilie gefunden, handelt es sich nur um einen Spezialfall. Man ist an einer allgemeinen Theorie der diskreten orthogonalen Polynome auf nicht-äquidistanten Gittern interessiert, und kommt so zu den **q-orthogonalen Polynomen**.

q-Orthogonale Polynome

Im ersten Teil dieses Kapitels haben wir die Hahn-Polynome als Lösungen der Differenzengleichung (5.3) eingeführt. Dabei gelangte man zu jener Differenzengleichung, indem man in einer Differentialgleichung **hypergeometrischen Typs**

$$\sigma(x)y''(x) + \tau(x)y'(x) + \lambda y(x) = 0 \tag{5.11}$$

die Differentialoperatoren durch Differenzenoperatoren ersetzte. Auf einem Gitter mit konstanter Schrittweite $\Delta x = h$ ergibt sich dadurch die Näherung

$$\begin{aligned}&\frac{\sigma(x)}{h}\left[\frac{y(x+h)-2y(x)+y(x-h)}{h}\right]\\ +&\frac{\tau(x)}{2}\left[\frac{y(x+h)-y(x)}{h}+\frac{y(x)-y(x-h)}{h}\right]+\lambda y(x)=0.\end{aligned}$$

Bei den Hahn-Polynomen aus Abschnitt 5.1 waren die Funktionen

$$\begin{aligned}\sigma(x) &= -\frac{1}{2}(D(x)+B(x)), \qquad & \tau(x) &= D(x)-B(x),\\ \lambda_n &= n(n+\alpha+\beta+1),\end{aligned}$$

und die Schrittweite h war 1. Allerdings kann man Gleichung (5.11) auch auf andere Arten in eine Differenzengleichung überführen. Das zugrunde liegende Gitter kann durch eine Funktion $x(s)$ beschrieben werden, wobei die Funktion von der Variablen s abhängt. Die Schrittweite ist dementsprechend nicht mehr konstant, sondern es gilt
$\Delta x := x(s+h) - x(s)$. Das muss bei der Ersetzung der Differentialoperatoren mitberücksichtigt werden. Eine Differenzengleichung auf einer Klasse von Gittern mit variabler Schrittweite ist schließlich durch

$$\begin{aligned}&\frac{\sigma(x)}{x(s+\frac{h}{2})-x(s-\frac{h}{2})}\left[\frac{y(s+h)-y(s)}{x(s+h)-x(s)}-\frac{y(s)-y(s-h)}{x(s)-x(s-h)}\right]\\ +&\frac{\tau(x(s))}{2}\left[\frac{y(s+h)-y(s)}{x(s+h)-x(s)}+\frac{y(s)-y(s-h)}{x(s)-x(s-h)}\right]+\lambda y(s)=0\end{aligned} \tag{5.12}$$

gegeben. Es sollte klar sein, dass nicht jede Funktion $x(s)$ dazu führt, dass (5.12) vergleichbare Ergebnisse wie (5.11) liefert und die Lösungen orthogonale Polynome sind. Nur unter bestimmten Voraussetzungen an $x(s)$ kann man dies tatsächlich garantieren.

In [59], Kapitel 3 findet man diesbezüglich eine detaillierte Betrachtung. Wir fassen dabei die wichtigsten Ergebnisse zusammen[3]. Die Voraussetzungen an $x(s)$ werden erfüllt, wenn $x(s+1)+x(s)$ ein Polynom ersten Grades und $x^2(s+1)+x^2(s)$ ein Polynom zweiten Grades in $x\left(s+\frac{1}{2}\right)$ ist. Aus der ersten Bedingung folgt mit $\alpha, \beta \in \mathbb{R}$

$$\frac{1}{2}\left(x(s+1)+x(s)\right) =: \alpha x\left(s+\frac{1}{2}\right)+\beta. \tag{5.13}$$

Für $\alpha \neq 1$ besitzt (5.13) die allgemeine Lösung

$$x(s) = c_1\kappa_1^{2s} + c_2\kappa_2^{2s} + c_3,$$

wobei κ_1 und κ_2 Wurzeln der Gleichung $\kappa^2 - 2\alpha\kappa + 1 = 0$ sind, c_1, c_2 sind beliebige Funktionen der Periode $\frac{1}{2}$ und $c_3 = \frac{\beta}{1-\alpha}$. Mit c_1, c_2 als Konstanten und $\kappa_1\kappa_2 = 1$ folgt durch elementare Rechnung, dass auch die zweite Bedingung erfüllt ist.
Für $\alpha = 1$ ist die Lösung von (5.13)

$$x(s) = c_1 s^2 + c_2 s + c_3,$$

mit $c_1 = 4\beta$, c_2 und c_3 beliebige Funktionen der Periode $\frac{1}{2}$.
Wählt man nun $\kappa_1^2 = q$ und $\kappa_2^2 = \frac{1}{q}$ mit beliebigen Konstanten q, c_1, c_2 und c_3, so gelten die Voraussetzungen. Die Lösungen von (5.12) besitzen analoge Eigenschaften zu (5.11). Zum Beispiel erfüllen die Lösungen alle eine Rodriguez-Formel. Die polynomialen Lösungen y_m, die die Orthogonalitätsbeziehung bezüglich des Skalarproduktes

$$\sum_{s_i=a}^{b} y_m(s_i)y_n(s_i)\omega(s_i)\Delta x\left(s_i-\frac{1}{2}\right) = \gamma_n\delta_{m,n},$$

unter den Nebenbedingungen

$$\sigma(s)\omega(s)x^l\left(s-\frac{1}{2}\right)\Big|_{s=a}^{b+1} = 0, \quad \forall l = 0,1,\ldots \text{ und } \omega(s_i)\Delta x(s_i-\frac{1}{2}) > 0 \text{ für } a \leq s_i \leq b$$

erfüllen, nennt man **klassische diskrete orthogonale Polynome auf nicht-äquidistanten Gittern**.
Man kann die möglichen Gitterfunktionen $x(s)$ klassifizieren. Man erhält folgende sechs kanonische Formen, vergleiche [59]:

1. $x(s) = s$ mit $\alpha = 1$ und $\beta = 0$. Dies ist der Fall der Hahn-Polynome.
2. $x(s) = s(s+1)$ mit $\alpha = 1$ und $\beta = \frac{1}{4}$. Dies ist der Fall der Dual-Hahn- bzw. Racah-Polynome.

[3] Zu den Bedingungen, die bereits für σ und τ entwickelt wurden, damit Lösungen der Differenzengleichung orthogonale Polynome sind, nehmen wir hier keinerlei Bezug. Wir verweisen dafür ebenfalls auf [59].

3. $x(s) = \exp(2ws)$ mit $\alpha > 1, \alpha = \cosh w$ und $\beta = 0$.

4. $x(s) = \sinh(2ws)$ mit $\alpha > 1, \alpha = \cosh w$ und $\beta = 0$.

5. $x(s) = \cosh(2ws)$ mit $\alpha > 1, \alpha = \cosh w$ und $\beta = 0$.

6. $x(s) = \cos(2ws)$ mit $(0 < \alpha < 1)$, $\alpha = \cos w$ und $\beta = 0$.

Zu den q-orthogonalen Polynomen gelangt man, indem man beispielsweise für das dritte Gitter, $x(s) = q^s = \exp(2ws)$ mit $q = \exp(2w)$ setzt. Für $q \to 1$ $(w \to 0)$ gilt $\exp(2ws) \approx 1 - 2ws$. Das Gitter weist im Grenzwert die Eigenschaften eines linearen Gitters auf und auch die Polynome besitzen analoge Eigenschaften zu den diskreten orthogonalen Polynomen eines linearen Gitters, wie etwa den Hahn-Polynomen. Man spricht hier von **q-Analogien**.
Die führt zu den Namen **q-Hahn-Polynome** oder auch **Dual-q-Hahn-Polynome**.

Bemerkung. Einen anderen Ansatz verwendet Hahn. In seinem Paper [33] zeigt er, dass Lösungen von q-Differenzengleichungen Systeme von diskreten q-orthogonalen Polynomen liefern. Diese Lösungen besitzen analoge Eigenschafen wie die diskreten orthogonalen Polynome auf gleichverteilten Gittern. Die q-orthogonalen Polynome können über eine 3-Term Rekursionsformel berechnet werden und erfüllen auch eine Orthogonalitätsbeziehung. Hahn klassifizierte 18 Familien von orthogonalen Polynomen, die sogenannte **q-Hahn-Klasse**.

Für mehr Informationen, eine detaillierte Einführung und eine tabellarische Auflistung der q-orthogonalen Polynome empfehlen wir [45], [59] und die Internetseite [67] .
Im nächsten Abschnitt werden wir die Theorie der diskreten orthogonalen Polynome auf zwei Raumdimensionen erweitern.

5.2.2 Diskrete orthogonale Polynome in zwei Variablen

Unsere Untersuchungen in diesem Kapitel beinhalteten bis jetzt ausschließlich diskrete orthogonale Polynome in einer Variablen. Wir sind grundsätzlich an orthogonalen Polynomen in zwei Variablen interessiert, wie wir sie auch in unserem numerischen Verfahren verwenden. Prinzipiell ist die Theorie diskreter orthogonaler Polynome auf mehrere Variablen erweiterbar. So wurde der Differenzenoperator bereits für zwei [89], aber auch für mehrere Dimensionen [49] bezüglich verschiedener Gitter untersucht. Die Lösungen der Differenzengleichungen sollten dabei diskrete orthogonale Polynome in mehreren Variablen für die betrachteten Gitter sein. Wir werden hier weniger ins Detail gehen, sondern ausschließlich die Hahn-Polynome in zwei Variablen auf einem Dreieck $\mathbb{T}_N = \{(x, y) : x \geq 0,\ y \geq 0,\ x + y \leq N\}$ und Schrittweite $h = 1$ wie in [89] einführen. Wir werden noch ihre Gewichtsfunktion, die Diffferenzengleichung und eine explizite Darstellung mittels der Hahn-Polynome $Q_n(x)$ in einer Variablen angeben. Dafür benötigen wir die zweidimensionalen Differenzenoperatoren.

Für $\mathbf{x} = (x, y) \in \mathbb{R}^2$ und den Einheitsvektoren $\mathbf{e}_1 = (1, 0)$ und $\mathbf{e}_2 = (0, 1)$ gilt für die Vorwärts- und Rückwärtsdifferenzenoperatoren mit Schrittweite $h = 1$

$$\Delta_i f(\mathbf{x}) := f(\mathbf{x} + \mathbf{e}_i) - f(\mathbf{x}) \qquad \text{und} \qquad \nabla_i f(\mathbf{x}) := f(\mathbf{x}) - f(\mathbf{x} - \mathbf{e}_i).$$

Definition 5.7. Es sei $m = n - l$, $0 \leq l \leq n$ mit $m, l, n \in \mathbb{N}_0$ und $-1 < \alpha, \beta, \gamma$. Die **Hahn-Polynome in zwei Variablen** $\Phi_{l,m}(\mathbf{x}, \alpha, \beta, \gamma)$ auf dem Dreieck $\mathbb{T}_N$ sind gegeben durch

$$\Phi_{l,m}(\mathbf{x}, \alpha, \beta, \gamma) = Q_l(x; \alpha, \beta + \gamma + 2m + 1, N - l)(-N + x)_m Q_m(y; \beta, \gamma, N - x).$$

Der Grad der Polynome $\Phi_{l,m}$ ist $n = m + l$. In [43] oder [88] findet man eine ähnlich Definition, allerdings wird dort jeweils mit einer anderen Normierung gearbeitet.
Die Polynome $\Phi_{l,m}$ sind auf der Menge $\mathbb{T}_N$ orthogonal bezüglich der Gewichtsfunktion

$$W((x, y), \alpha, \beta, \gamma) = \binom{x + \alpha}{\alpha} \binom{y + \beta}{\beta} \binom{N - x - y + \gamma}{\gamma}$$

und erfüllen die Differenzengleichung

$$\begin{aligned}
&x(N - x + \beta + \gamma + 2)\Delta_1\nabla_1 u - y(x + \alpha + 1)\Delta_1\nabla_2 u \\
&-x(y + \beta + 1)\Delta_2\nabla_1 u + y(N - y + \alpha + \gamma + 2)\Delta_2\nabla_2 u \\
&+ \left[(N - x)(\alpha + 1) - x(\beta + \gamma + 2)\right]\Delta_1 u \\
&+ \left[(N - y)(\beta + 1) - y(\alpha + \gamma + 2)\right]\Delta_2 u = -n(n + \alpha + \beta + \gamma + 2)u.
\end{aligned}$$

Den Differenzenoperator kann man auch nach [89] in selbstadjungierter Form schreiben. Eine numerische Anwendung und eine Analyse des Approximationsverhaltens der mehrdimensionalen Hahn-Polynome sind nach dem heutigen Stand nicht bekannt.

Fazit

Die gezeigten Ansätze der Konstruktion diskreter orthogonaler Polynome mit Hilfe von Quadraturverfahren, q-orthogonale Polynome oder die Erweiterung auf 2d-diskrete orthogonale Polynome sollen nur einen kleinen Ausblick für weitere Forschungsarbeiten im Bereich diskreter orthogonaler Polynome und ihrer Anwendung in spektralen Verfahren liefern. Speziell beim Approximationsverhalten und der numerischen Anwendung sind noch viele ungeklärte Fragen offen. Gleichzeit sind natürlich auch zweidimensionale, diskrete orthogonale Polynome auf nicht-äquidistanten Gittern von Interesse. Auch hier ist die Theorie diskreter orthogonaler Polynome noch nicht fortgeschritten. Eine Untersuchung dieser Punkte übersteigt aber bei weitem den Rahmen dieser Arbeit, so dass wir im folgenden Abschnitt ausschließlich die numerische Untersuchung für die APK-Polynome vornehmen und ihr Verhalten beim Lösen von hyperbolischen Erhaltungsgleichungen mit Hilfe des Spektrale-Differenzen-Verfahrens analysieren.

6 Numerische Resultate

Wir präsentieren in diesem Abschnitt die numerischen Ergebnisse. Es wird das Spektrale-Differenzen-Verfahren durch die APK-Polynome erweitert und der modale Exponentialfilter wird von der Parameterwahl der APK-Polynome abhängig sein. Wir untersuchen die Spektrale-Differenzen-Methode anhand zweier Testfälle. Wir betrachten zum einen die Burgers-Gleichung mit einer Sinus-Anfangsbedingung und zum anderen die Euler-Gleichungen mit einer **Stoß-Wirbel-Interaktion**. Beide weisen Unstetigkeiten auf. Bei der Burgers-Gleichung bilden sich unter der angegebenen Anfangsbedingung Stöße und die Euler-Gleichung besitzt bereits unstetige Anfangsbedingungen. Um die dort auftretenden Gibbs'schen Oszillationen zu reduzieren und so die Stabilität zu verbessern, verwenden wir die modalen Filter aus Kapitel 4. Wir analysieren die Effizienz des Verfahrens für beide Testfälle und untersuchen die Stabilität im Hinblick auf verschiedene APK-Polynomfamilien. Es wird speziell der Parameter γ aus der Definiton der APK-Polynome entscheidend sein. Der Parameter hat direkten Einfluss auf den Exponentialfilter (4.11), vergleiche auch Kapitel 4. Wir werden für verschiedene Parameterwahlen (α, β, γ) auch andere modale Filter, wie etwa den **Lanczos-Filter** oder den **raised-cosine-Filter**, bei der Burgers-Gleichung verwenden.
Die Filterung wird über den **koeffizientenbasierenden Stoßindikator** gesteuert und als **Zeitintegrationsverfahren** dient stets das Runge-Kutta Verfahren vierter Ordnung aus Abschnitt 2.2.1. Alle Resultate in diesem Kapitel sowie die graphischen Darstellungen wurden **nicht** nachbearbeitet. Für die graphische Ausgabe wurde das Programm tecplot360 verwendet. Es wurde darauf geachtet, dass jeweils die gleichen Intervalle und die gleichen Schrittweiten in der graphischen Unterteilung zwischen den einzelnen Abbildungen benutzt wurden, um so die Ergebnisse bestmöglich vergleichen zu können.

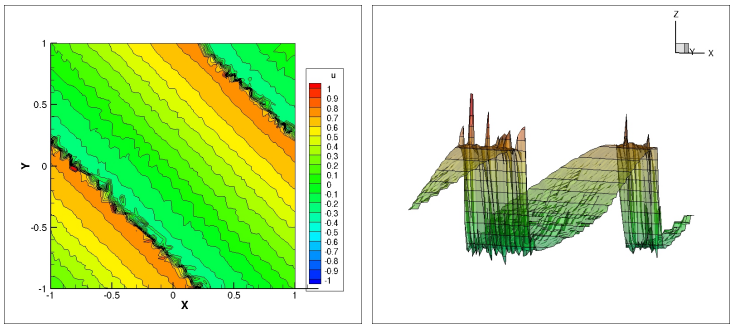
6.1 Burgers-Gleichung

In unserem ersten Test betrachten wir auf dem Gebiet $[-1, 1]^2$ die zweidimensionale Burgers-Gleichung

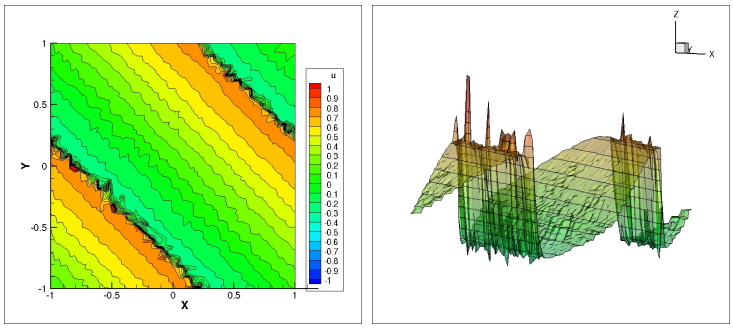
$$u_t(x, y, t) + u(x, y, t)\,(u_x(x, y, t) + u_y(x, y, t)) = 0$$

mit dem Anfangswert

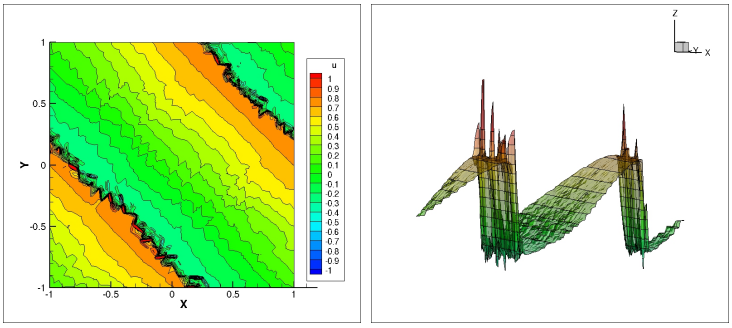
$$u_0(x, y, t) = \frac{1}{4} + \frac{1}{2}\sin(\pi(x + y))$$



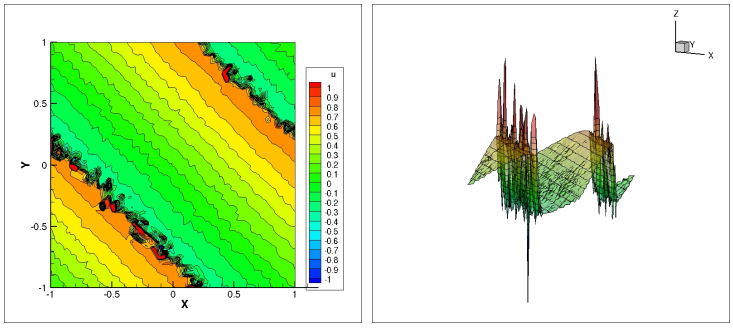
(a) Fejèr-Filter



(b) Lanczos-Filter

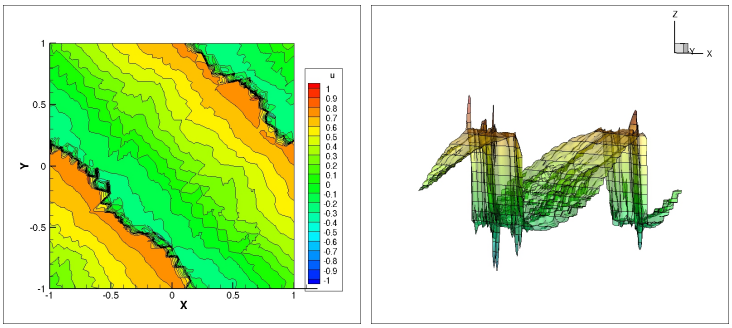


(c) raised-cosine-Filter

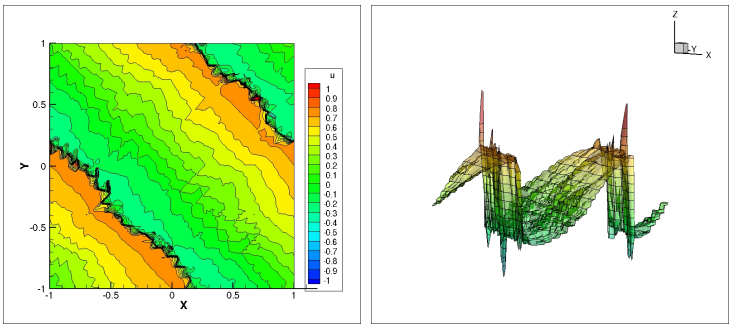


(d) shaped-raised-cosine-Filter

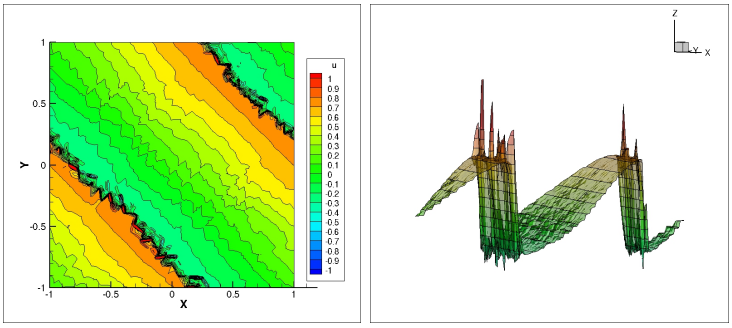
Abbildung 6.1: *Parameter* $(\alpha, \beta, \gamma) = (1, 1, 2)$ *und Ordnung* $N = 2$



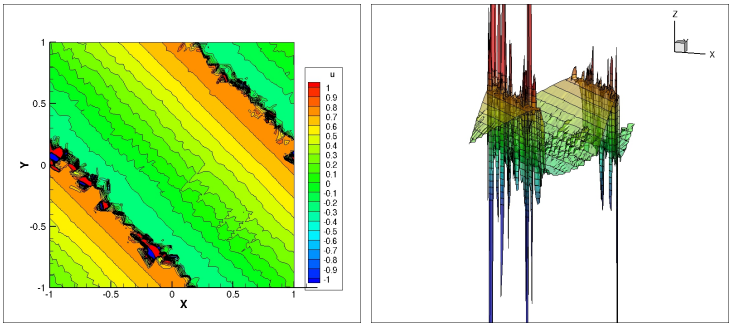
(a) Fejèr-Filter



(b) Lanczos-Filter



(c) raised-cosine-Filter



(d) shaped-raised-cosine-Filter $t = 0.4$

Abbildung 6.2: *Parameter* $(\alpha, \beta, \gamma) = (1, 2, 5)$ *und Ordnung* $N = 3$

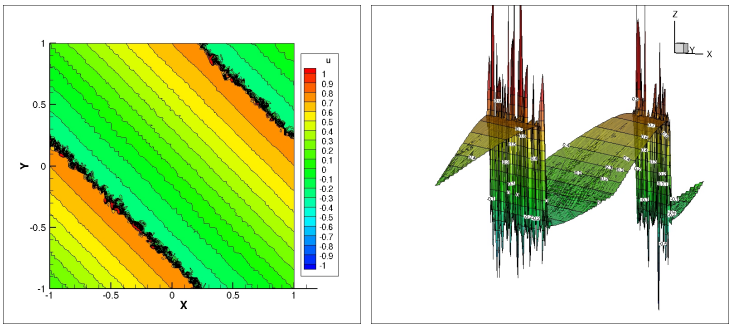
und periodischen Randbedingungen $u(-1,y,t) = u(1,y,t)$ und $u(x,-1,t) = u(x,1,t)$. Dieser Testfall ist aus [51] entnommen. Es ist bekannt, dass sich zum Zeitpunkt $t = 0.5$ zwei Unstetigkeiten bei $y = \frac{3}{2} - x$ und $y = \frac{5}{2} - x$ entwickeln. Ohne Filterung entstehen hohe Oszillationen und das Spekrale-Differenzen-Verfahren bricht zusammen. Wir verwenden die modalen Filter aus Kapitel 4 für die Parameterwahl $(\alpha,\beta,\gamma) = (1,1,2)$ und $(1,2,5)$. Bei der Implementierung des Lanczos-Filters nutzen wir die Darstellung aus der Elektrotechnik. Es gilt

$$\tilde{\sigma}(\eta) = \begin{cases} \frac{a\sin(\pi\eta)\sin\left(\pi\frac{\eta}{a}\right)}{(\pi\eta)^2}, & \text{wenn } -a < \eta < a,\ a \neq \eta, \\ 1 & \text{für } \eta = 0, \\ 0 & \text{sonst.} \end{cases}$$

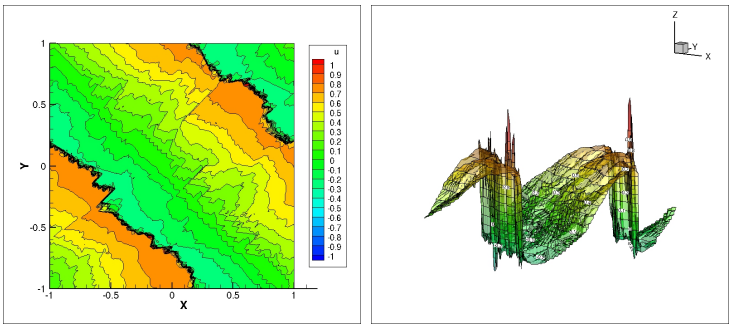
Die Bandbreite a setzen wir auf 1. Eine Analyse des Einflusses der Filterbreite auf die Stabilität wäre demnach auch interessant, allerdings wird es nicht im Rahmen dieser Arbeit gemacht.
In unseren Testrechnungen verwenden wir 1088 Dreiecke und es ergeben sich bei $t = 0.45$ die Abbildungen 6.1 und 6.2, wobei der shaped-raised-cosine-Filter für $(1,2,5)$ bei $t = 0.4$ ausgewertet wurde. Bereits hier erkennt man die hohen Ausschläge in den Oszillationen (6.2(d)). Der Maximalwert liegt bei etwa 8.9 und der Minimalwert bei -11.8. Das Verfahren bricht im Weiteren schließlich zusammen. Je höher die Ordnung N des Verfahrens ist, desto instabiler wird es. In den Abbildungen 6.3 und 6.4 sehen wir für eine Auswahl von verschiedenen APK-Familien mit ihrem Exponentialfilter

$$\sigma\left(\frac{l+m}{N}\right) = e^{-\varepsilon_N N^{2p}\left(\frac{l+m}{N}\right)^p\left(\frac{l+m+\gamma}{N}\right)^p} \approx e^{-\varepsilon_N N^{2p}\left(\frac{l+m}{N}\right)^{2p}}$$

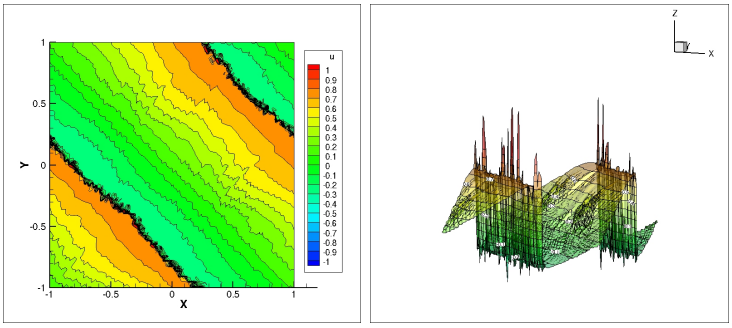
zum Zeitpunkt $t = 0.45$ die jeweilige Lösung der Burgers-Gleichung. Es wurden diesmal 4352 Dreiecke für unterschiedliche Polynomordnungen, Filterstärken und Filterordnungen verwendet. Die Filterordnung nannten wir hier $2p$, bei der Implementierung wurde jedoch mit dem exakteren Wert des Filters gearbeitet und $e^{-\varepsilon_N N^{2p}\left(\frac{l+m}{N}\right)^p\left(\frac{l+m+\gamma}{N}\right)^p}$ genutzt. Bei Erhöhung der Filterordnung wurde auch gleichzeitig die Filterstärke erhöht, um eine ausreichende Reduktion der Oszillationen zu erhalten. Ansonsten wäre es zu Stabilitätsproblemen gekommen. In den Tabellen 6.1-6.3 finden sich eine größere Auswahl an Parametervergleichen, dabei wurde der maximale und minimale Wert aufgenommen, um so ein Indiz für die Stabilität zu erhalten. Ein hoher Ausschlag der Oszillationen ist ein Indiz für das Zusammenbrechen des Verfahrens. Für $N = 2$ und $N = 3$ zeigen die Tabellen 6.1 und 6.2, dass die Oszillationen bei der Parameterwahl $(1,1,2)$ den höchsten Ausschlag besitzen, was auch die $2d$- und $3d$- Darstellungen aus den Abbildungen 6.3 und 6.4 belegen. Jedoch wird bei der Verwendung anderer Parameter $(\alpha,\beta,\gamma) \neq (1,1,2)$ der glatte Bereich *unruhiger*. Man kann das zugrundeliegende Gitter und eine Δ-Struktur erkennen. Allerdings könnte man geringe Schwankungen durch eine Nachbearbeitung der Lösung beispielsweise mittels Spines glätten.
Ein beliebiges Hochsetzen des Parameters γ, der einen direkten Einfluss auf den Filter und so auf die Stabilität besitzt, führt zu keinem adäquaten Ergebnis, da durch die Erhöhung der Filter verstärkt wird und zu viele Informationen verloren gehen. Auch führt



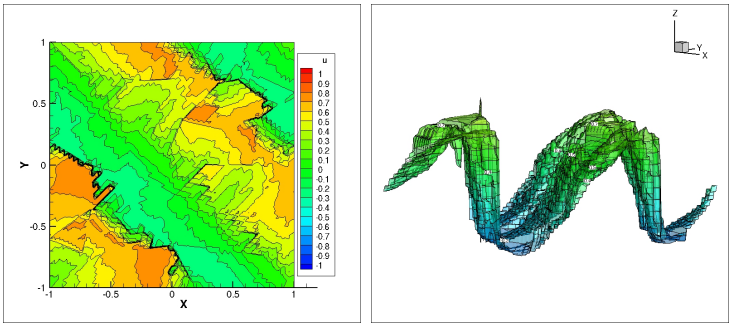
(a) $(\alpha, \beta, \gamma) = (1, 1, 2)$



(b) $(\alpha, \beta, \gamma) = (1, 1, 4)$

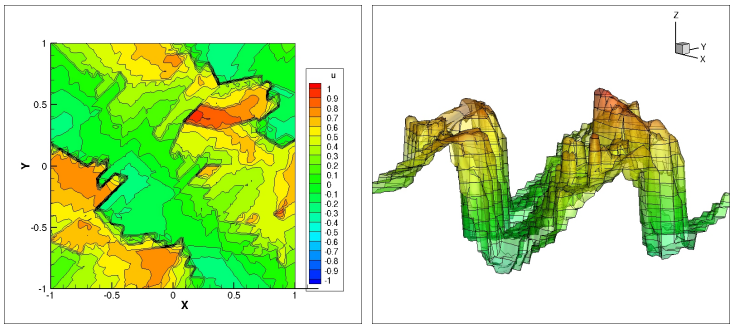


(c) $(\alpha, \beta, \gamma) = (2, 2, 4)$

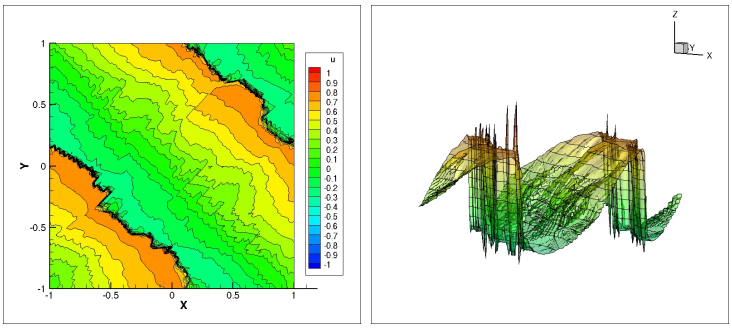


(d) $(\alpha, \beta, \gamma) = (2, 2, 8)$

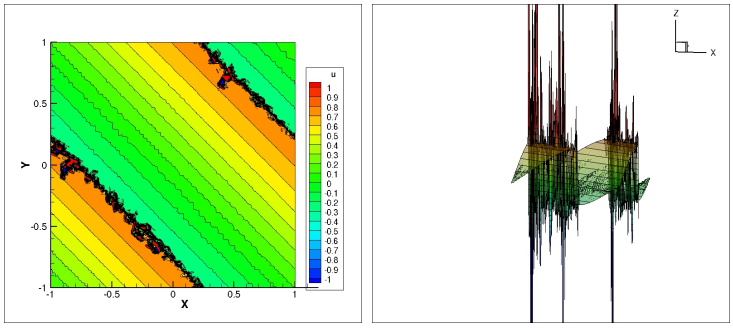
Abbildung 6.3: *Parameter* (α, β, γ)*, Polynomordnung* 2*, Filterordnung* 2*,* $c = 8$



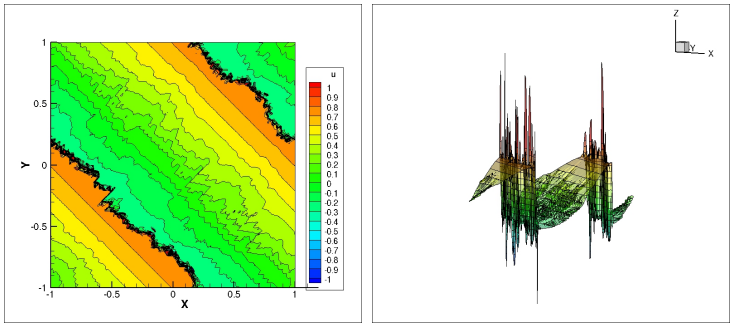
(a) $(\alpha, \beta, \gamma) = (1, 1, 20), N = 2, p = 1, c = 2$



(b) $(\alpha, \beta, \gamma) = (2, 2, 8), N = 2, p = 1, c = 2$



(c) $(\alpha, \beta, \gamma) = (1, 1, 2), N = 3, p = 2, c = 8$



(d) $(\alpha, \beta, \gamma) = (1, 1, 4), N = 3, p = 2, c = 8$

Abbildung 6.4: *Parameter* (α, β, γ)*, Polynomordnung* n*, Filterordnung* $2p$*, Filterstärke* c

(α, β, γ)	Filterstärke	p	Minimum	Maximum
$(1, 1, 2)$	2	1	-15.97	11.34
$(1, 1, 4)$	2	1	-0.82	2.92
$(2, 2, 4)$	2	1	-12.44	6.56
$(2, 2, 8)$	2	1	0.57	1.60
$(1, 2, 5)$	2	1	-1.40	3.17
$(1, 1, 20)$	2	1	-0.25	1.02
$(1, 1, 2)$	8	2	-1.22	4.55
$(1, 1, 4)$	8	2	-0.55	1.80
$(2, 2, 4)$	8	2	-0.53	2.26
$(2, 2, 8)$	8	2	-0.38	1.15
$(1, 2, 5)$	8	2	-0.46	1.42

Tabelle 6.1: *Tabelle für Polynomgrad $N = 2$, zum Zeitpunkt $t = 0.45$*

(α, β, γ)	Minimum	Maximum
$(1, 1, 2)$	-12.21	13.33
$(1, 1, 4)$	-1.50	2.9
$(2, 2, 4)$	-1.21	3.7
$(2, 2, 8)$	-1.97	1.58
$(1, 2, 5)$	-1.27	2.4

Tabelle 6.2: *Tabelle für Polynomgrad $N = 3$, zur Filterstärke 8 und Filterordnung 4*

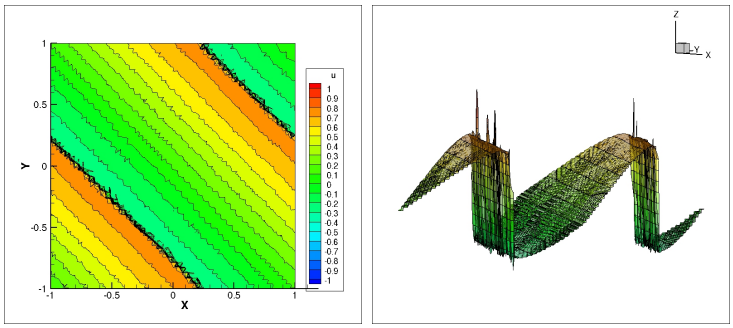
eine Erhöhung zu größeren Schwankungen im glatten Bereich der Lösung, vergleiche Abbildung 6.4(a). In Tabelle 6.3 scheint hingegen die Verwendung von $(1, 1, 2)$ optimal und auch die Abbildung 6.5 bestätigt die Vermutung, dass die berechnete Funktion eine sehr gute Approximation an die Lösung darstellt. Im glatten Bereich wirkt die Lösung sehr ruhig mit nur geringen Schwankungen, die Kanten der jeweiligen Unstetigkeiten sind sehr scharf und die Oszillationen nicht allzu hoch. Eine andere Parameterauswahl mit einem größeren γ führt dazu, dass mehr Informationen verloren gehen und die Glättung zu stark ist. Analoge Untersuchungen mit weniger Dreiecken, für verschiedene Filterstärken und Ordnungen sowie anderen Parametern liefern analoge Ergebnisse.
Insgesamt kann die Wahl der APK-Polynome und ihrer *natürlichen* Filter die Stabilität des SD-Verfahrens positiv beeinflussen. Man muss jedoch darauf achten, dass nicht zu viele Informationen verloren gehen. Es fehlt weiterhin eine analytische Untersuchung, um die Wahl der Parameter zu optimieren.

6.2 Euler-Gleichung

Bei den Euler-Gleichungen handelt es sich um ein System von nichtlinearen hyperbolischen Erhaltungsgleichungen. In der Praxis modelliert man mit den Gleichungen gas-

(α, β, γ)	Minimum	Maximum
$(1, 1, 2)$	-0.35	1.39
$(1, 1, 4)$	-0.27	0.91
$(2, 2, 4)$	-0.25	0.94
$(2, 2, 8)$	-0.34	0.91
$(1, 2, 5)$	-025	0.74

Tabelle 6.3: *Tabelle für Polynomgrad* $N = 4$*, zur Filterstärke* 20 *und Filterordnung* 2



(a) $(\alpha, \beta, \gamma) = (1, 1, 2), N = 4, p = 1, c = 20$

Abbildung 6.5: *Parameter* (α, β, γ)*, Polynomordnung* N*, Filterordnung* $2p$*, Filterstärke* c

dynamische Prozesse und leitet sie aus den physikalischen Erhaltungsätzen von Masse, Impuls und Energie ab. Anders als bei der Navier-Stokes-Gleichung, die allgemein Strömungen beschreibt, werden bei den Euler-Gleichungen weder Wärmeleitung noch Viskosität berücksichtigt. Eine detaillierte Einführung findet man in der Literatur [28]. Für den zweidimensionalen Fall ist das System durch die folgenden partiellen Differentialgleichungen

$$\frac{\partial}{\partial t}\begin{pmatrix}\rho\\ \rho u\\ \rho v\\ \rho E\end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix}\rho u\\ \rho u^2 + p\\ \rho uv\\ \rho uH\end{pmatrix} + \frac{\partial}{\partial y}\begin{pmatrix}\rho v\\ \rho uv\\ \rho v^2 + p\\ \rho uH\end{pmatrix} = \mathbf{0} \tag{6.1}$$

mit der Dichte ρ, den Geschwindigkeiten u in x- und v in y-Richtung und dem Druck p gegeben. Die Größe $E = e + \frac{u^2+v^2}{2}$ ist dabei die Energie, die sich aus der kinetischen und der spezifischen inneren Energie e zusammensetzt. H ist die Enthalpie, die in der Relation $H = E + \frac{p}{\rho}$ zur Energie E steht. Die Erhaltungsgröße ist der Vektor $\mathbf{u} = (\rho, u\rho, v\rho, \rho E)^T$. Für die Zustandsgleichung benötigt man eine weitere Gleichung, die die spezifische innere Energie e mit dem Druck p und der Dichte ρ verknüpft. Betrachtet man ein **ideales Gas**, so ist der Druck durch die Temperatur über

$$p = R\rho T$$

mit der idealen Gaskonstante R bestimmt. Für die innere Energie e ist eine sinnvolle Annahme, dass sie ausschließlich proportional zur Temperatur T ist. Man erhält

$$e = c_V T$$

mit der Wärmekapazität c_V für konstantes Volumen und spricht hierbei von einem **polytropen Gas**. Unter Verwendung der Formel ergibt sich die Zustandsgleichung für ein ideales polytropes Gas durch

$$p = \rho(\tilde{\gamma} - 1)\left(E - \frac{u^2 + v^2}{2}\right)$$

mit dem Isotropenkoeffizienten $\tilde{\gamma} = \frac{R}{c_V} + 1$.
Wir betrachten als Testfall die **Stoß-Wirbel-Interaktion** aus [40]. Auf dem Gebiet $[0, 2] \times [0, 1]$ wird die Euler-Gleichung (6.1) mit unstetigen Anfangsbedingungen betrachtet. Bei $x = 0.5$ wird ein vertikal verlaufender Stoß parallel zur y-Achse positioniert. Links besitzt das Problem die Anfangswerte

$$\mathbf{u}_0^- = (\rho, u, v, p) = (1, 1.1\sqrt{\tilde{\gamma}}, 0, 1).$$

Die rechte Seite $\mathbf{u}_0^+ = (\rho^+, u^+, v^+, p^+)$ wird mit Hilfe der Ranking-Hugeniot-Bedingung (Satz 2.3) berechnet. Man erhält

$$\begin{aligned} u^+ &= \frac{\tilde{\gamma}b - \sqrt{\gamma^2 b^2 - 2u(\tilde{\gamma}^2 - 1)d}}{u(\tilde{\gamma} + 1)}, \\ v^+ &= 0, \\ \rho^+ &= \frac{u}{u^+}, \\ p^+ &= b - uu^+, \end{aligned}$$

wobei $b = 1.21\tilde{\gamma} + 1$ und $d = \left(0.6655 + \frac{1.1}{\tilde{\gamma}-1}\right)\tilde{\gamma}^{\frac{3}{2}}$ ist.
Gleichzeitig wird bei $(x_c, y_c) = (0.25, 0.25)$ ein isotroper Wirbel positioniert, der durch die Änderungsparameter

$$\begin{aligned} (\delta u, \delta v) &= \frac{\varepsilon r}{r_c} \mathrm{e}^{\alpha(1-r^2)}(y - y_c, -(x - x_c)), \\ \delta T &= \frac{(\tilde{\gamma} - 1)\varepsilon^2 \mathrm{e}^{2\alpha(1-r^2)}}{4\alpha\tilde{\gamma}} \end{aligned}$$

mit $r = \sqrt{(x - x_c)^2 + (y - y_c)^2}$, $r_c = 0.05, \varepsilon = 0.3$ und $\alpha = 0.204$ beschrieben wird. Dadurch erfolgt auch eine Störung der Dichte und des Druckes. Die Störungen sind gegeben mit

$$\delta\rho = \delta p = (1 - \delta T)^{\frac{1}{1-\tilde{\gamma}}} - 1$$

und die Anfangsdaten links werden mit dieser Störung überlagert. Daraus abgeleitet sind die Anfangsdaten

$$\mathbf{u}_0^- = (\rho + \delta\rho, u + \delta u, v + \delta v, p + \delta p).$$

Der Wirbel bewegt sich dann nach links über den Stoß. Das Rechengitter in Abbildung 6.6 für diesen Testfall wurde bereits in Kapitel 2 bei der Triangulierung präsentiert. Beim Stoß kommt es zu einer Verfeinerung des Gitters. Als Randbedingungen setzt man

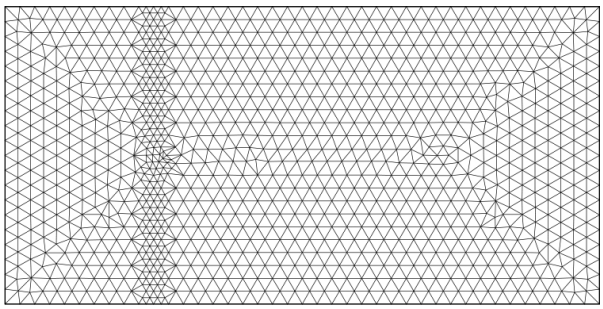


Abbildung 6.6: *Rechengitter für den Stoß-Wirbel-Testfall*

Einflussbedingungen links, Ausflussbedingungen rechts und unten und oben feste, reflektierende Wände fest.
Ohne Filterung würde das SD-Verfahren zusammenbrechen, wenn der Wirbel den Stoß erreicht. Wir nutzen verschieden APK-Familien und die zugehörigen Exponentialfilter, um dies zu verhindern. Das Gitter besteht immer aus 2122 Dreiecken. In den Graphiken ist jeweils der Impuls abgebildet. Dabei wurde eine Skalierung von 0.9 - 1.5 mit 50 Stufen bei Tecplot verwendet und keinerlei Rekonstruktion vorgenommen. Es wurden jeweils drei Testreihen zu den Polynomordnungen $N = 2, 3, 4$ mit unterschiedlichen Parametern in einer $2d$- und $3d$-Darstellung gezeichnet. Die Auswertung erfolgt zu den Zeiten 0.05/0.2/0.35/0.6/0.8 Sekunden. In den Abbildungen 6.7, 6.12 und 6.17 beginnen die jeweiligen Testreihen.
In den Tabellen 6.4 und 6.5 findet man die Anzahl der Zeitschritte bis zur jeweiligen Zeit für unterschiedliche Polynomparameter (α, β, γ) .

(α, β, γ)	0.05	0.2	0.35	0.6	0.8
$(1, 1, 2)$	991	4042	7302	12664	16985
$(1, 1, 3)$	991	4040	7263	12665	16986
$(2, 2, 4)$	889	4052	7266	12655	16966
$(1, 2, 5)$	889	4033	7240	12613	16912

Tabelle 6.4: *Anzahl der Zeitschritte für Polynomgrad* $N = 3$*, zur Filterstärke* 14 *und Filterordnung* 4

(α, β, γ)	0.05	0.2	0.35	0.6	0.8
$(1, 1, 2)$	1555	6317	11284	19621	26292
$(1, 1, 3)$	1552	6249	11071	19171	25652
$(2, 2, 4)$	1549	6343	11408	19920	26730
$(1, 2, 5)$	1549	6229	10976	18956	25350

Tabelle 6.5: *Tabelle für Polynomgrad* $N = 4$*, zur Filterstärke* 22 *und Filterordnung* 4

Betrachtet man die jeweiligen Testreihen, so fällt Folgendes auf: Im ersten Fall für die Polynomordnung $N = 2$, sind die jeweiligen Abbildungen 6.7-6.11 sehr ähnlich. Der Stoß

bei $x = 0.5$ ist scharf und bleibt stabil. Der Wirbel bewegt sich ohne Probleme über ihn hinweg. Nach 0.8 Sekunden erkennt man jeweils den Wirbel, siehe Abbildung 6.11. Einzig aus der $2d$-Darstellung in Abbildung 6.10 könnte man ableiten, dass dort die Parameterwahl $(1, 1, 3)$ optimal ist, da die Darstellung das ruhigste Verhalten auch links vom Stoß aufweist. Allerdings sind die Unterschiede zwischen den einzelnen APK-Familien so gering, dass die Schwankungen durch die meisten Nachbearbeitungsprogramme eliminiert werden können.
Bei $N = 3$ ist das anders. Der Wirbel ist schärfer erkennbar und der Stoß weist bei $x = 0.5$ für verschiedene Parameter (α, β, γ) eine unterschiedliche Anzahl und Stärke an Oszillationen auf, vergleiche Abbildung 6.12. Bei der Überquerung der Stufe kommt es in allen drei Fällen zu leichten Störungen (Abbildung 6.13 und 6.14), wobei die $2d$- und $3d$-Darstellung für den Fall $(2, 2, 4)$ am schärfsten wirkt. Auch die Anzahl und Höhe der Oszillationen ist für die Parameterwahl $(2, 2, 4)$ am geringsten. In Abbildung 6.16 kann man jeweils den Wirbel sowie die Stufe scharf erkennen. Die Anzahl der verwendeten Zeitschritte (Tabelle 6.4) weist vergleichbare Zahlen auf. So ist der Unterschied bei $t = 0.8$ gerade einmal 20 Schritte bei den hier gezeichneten Fällen (16966-16986).
Kommen wir zur letzten Testreihe für die Polynomordnung $N = 4$. Beim ersten Zeitschritt $t = 0.05$ (Abbildung 6.17) ist der Wirbel nochmals schärfer als zuvor. Man erkennt deutlich mehr Oszillationen, die jedoch geringer sind als bei $N = 3$ (Abbildung 6.12). Interessant wird dieser Fall allerdings erst, wenn der Wirbel den Stoß erreicht (Abbildungen 6.18 und 6.19). So ist bei $t = 0.2$ der Wirbel in allen Fällen noch zu erkennen, hingegen kann man bei $t = 0.35$ bei $(1, 1, 3)$ nur erahnen, wo er verläuft. Für $(2, 2, 4)$ fällt selbst das schwer. Hier sind in der $3d$-Darstellung einige Schwingungen und Verwirbelungen zu erkennen. Dieses Verhalten führt sich bei $t = 0.6$ und $t = 0.8$ (Abbildung 6.20 und 6.21) fort. Einzig bei $(1, 1, 2)$ kann man noch von einem Wirbel sprechen. Für die anderen Parameter stellt man ausschließlich ein sehr unruhiges Verhalten mit Schwingungen fest, wobei bei $(1, 1, 3)$ die Oszillationen glatt wirken. Der Grund hierfür ist leicht anzugeben. Bei Verwendung der Parameter $(1, 1, 3)$ und $(2, 2, 4)$ wird auch der jeweilige Exponentialfilter entsprechend verändert und es kommt zu einer stärkeren Glättung durch die Filter als bei $(1, 1, 2)$. Es gehen mehr Informationen verloren, in diesem Fall der Wirbel. Der Wirbel wird *weggefiltert*. Auch an den Zeitschritten kann man erkennen, dass Informationen verloren gehen. Die Rechnungen vereinfachen sich nach Überqueren der Stufe und man benötigt bei der Parameterwahl $(1, 1, 3)$ 640 Schritte weniger als bei $(1, 1, 2)$, obwohl gerade am Anfang ähnliche Zeitschritte vorhanden waren, siehe Tabelle 6.5. Die größere Anzahl an Zeitschritten für $(2, 2, 4)$ kann man durch die Tatsache erklären, dass hier die Verwirbelungen und Schwingungen zur Instabilität des Verfahrens führen. Für die Parameterwahl $(1, 2, 5)$ benötigt man beispielsweise sogar nur 25350 Schritte. Das sind circa 3.6%- Prozent weniger als im $(1, 1, 2)$-Fall. Allerdings verliert man auch hier den größten Teil der Information, das heißt den Wirbel. Das wirkt sich aber gleichzeitig positiv auf die Stabilität des Spektrale-Differenzen-Verfahrens aus, jedoch beschreibt die berechnete Lösung nicht mehr den tatsächlichen Fall, da kein Wirbel mehr in der Lösung enthalten ist.
Bei $N = 5$ und unter der Grundvoraussetzung, dass man Filterstärke und Filterordnung jeweils konstant lässt, verliert man sehr schnell die Stabilität des Verfahrens für eine

Parameterwahl mit geringem γ.
Allgemein kommt man auch in dieser Testreihe zu dem Schluss, dass die unterschiedlichen APK-Polynomfamilien und ihre Exponentialfilter durchaus Einfluss auf die Stabilität der SD-Methode und seine Genauigkeit haben. Die Wahl der Parameter (α, β, γ) ist entscheidend, sie kann sich sowohl negativ als auch positiv auf die Methode auswirken. So kann sich bei einer geschickten Parameterwahl die Stabilität des SD-Verfahrens erhöhen, jedoch muss man darauf achten, dass gleichzeitig nicht zuviele Informationen durch den Einsatz des Exponentialfilters verloren gehen, wie der Wirbel in der Euler-Testreihe für Polynomordnung $N = 4$ und Paramterwahl $(2, 2, 4)$. Die berechnete Lösung entspricht dann nicht mehr der tatsächlichen Lösung. Die Optimierung der Parameterwahl sollte dabei ein erklärtes Ziel für weitere Forschungsarbeiten sein. Es wird sicherlich eine analytische Untersuchung des gesamten Problemkreises nötig sein.

$t = 0.05$ 2d- und 3d-Plot

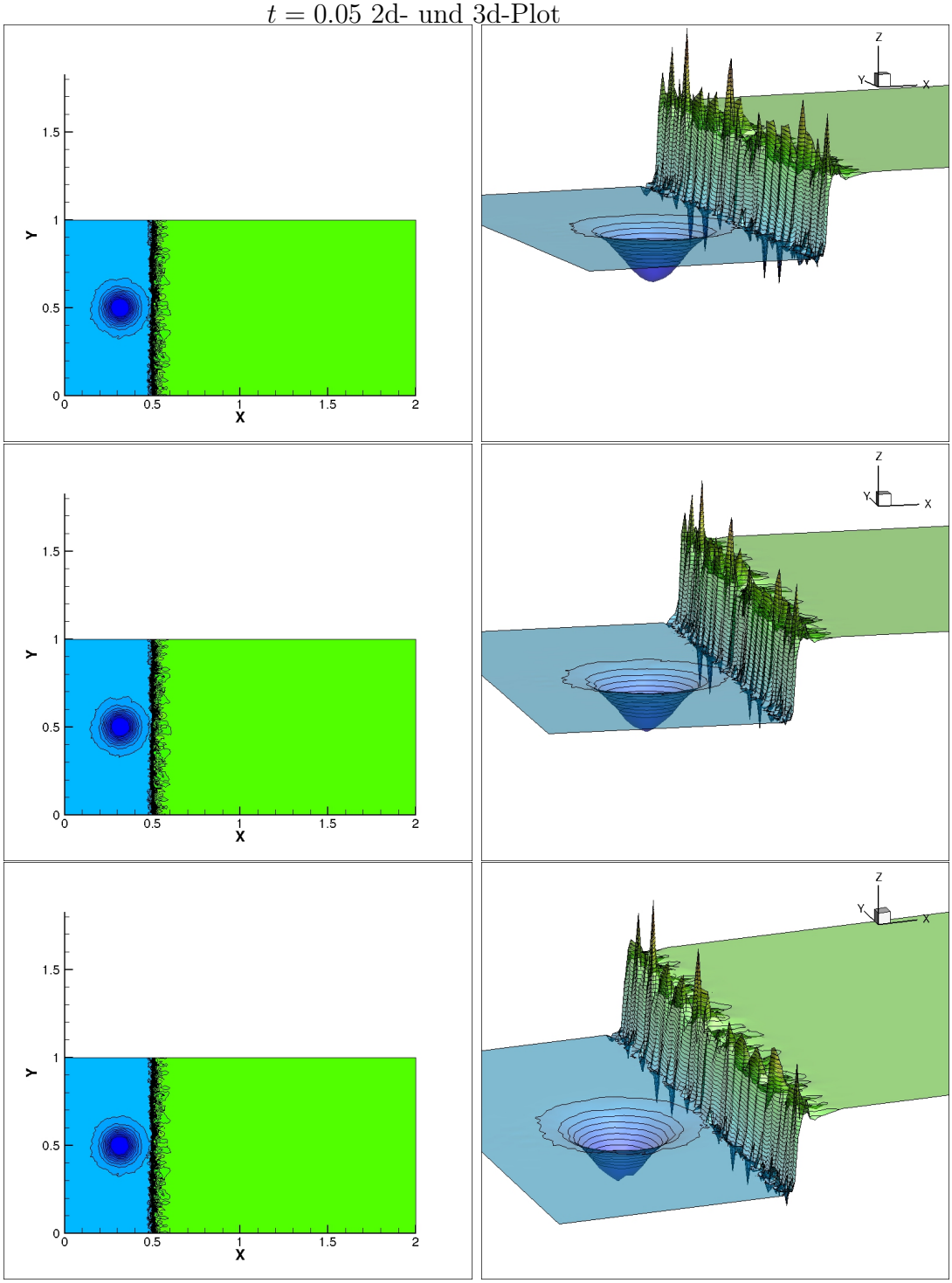


Abbildung 6.7: *APK-SDM ohne u-Rekonstruktion,* $N = 2, p = 1, c = 8$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

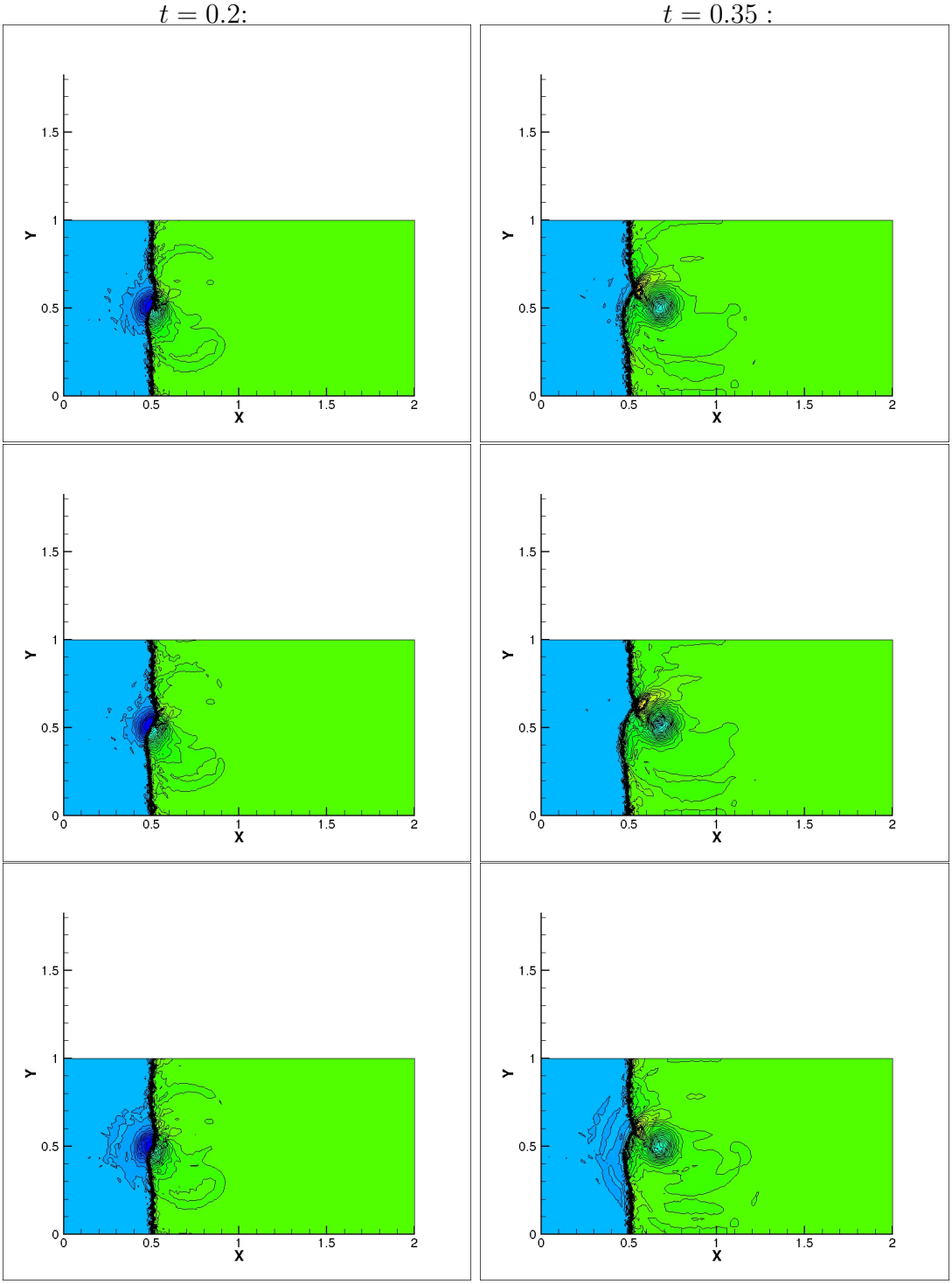


Abbildung 6.8: *APK-SDM ohne u-Rekonstruktion,* $N = 2, p = 1, c = 8$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

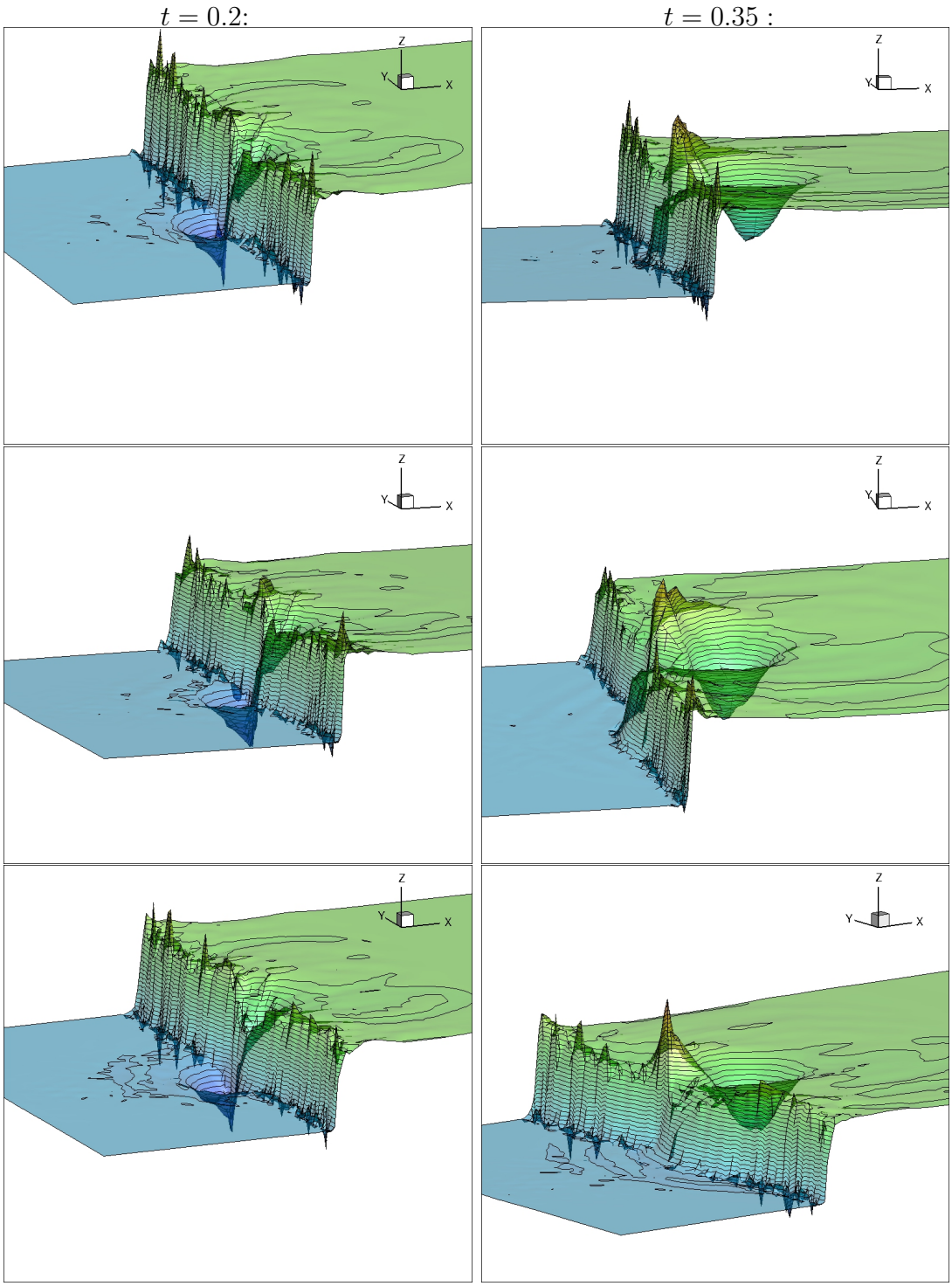


Abbildung 6.9: *APK-SDM ohne u-Rekonstruktion,* $N = 2, p = 1, c = 8$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

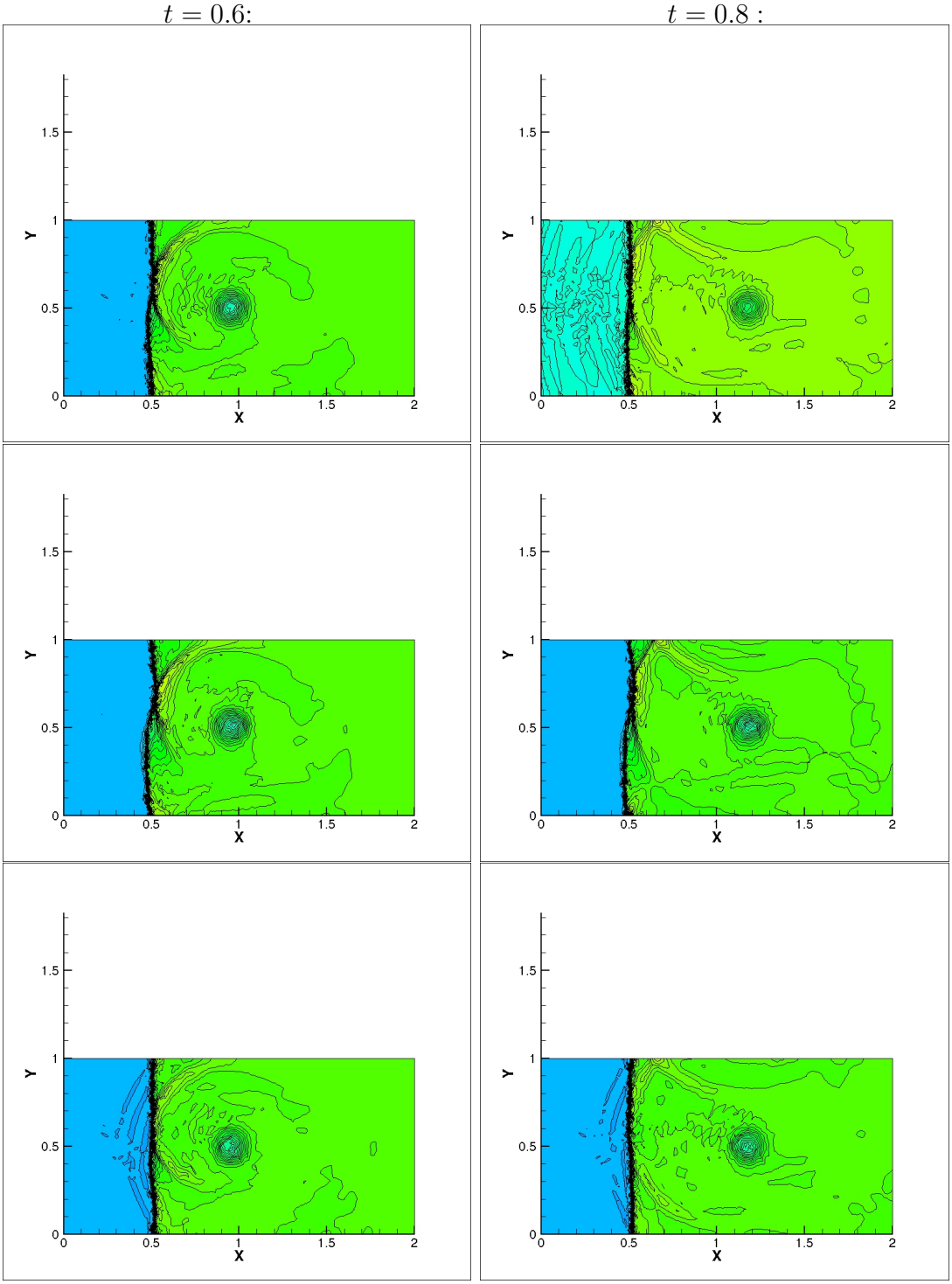


Abbildung 6.10: *APK-SDM ohne u-Rekonstruktion,* $N = 2, p = 1, c = 8$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

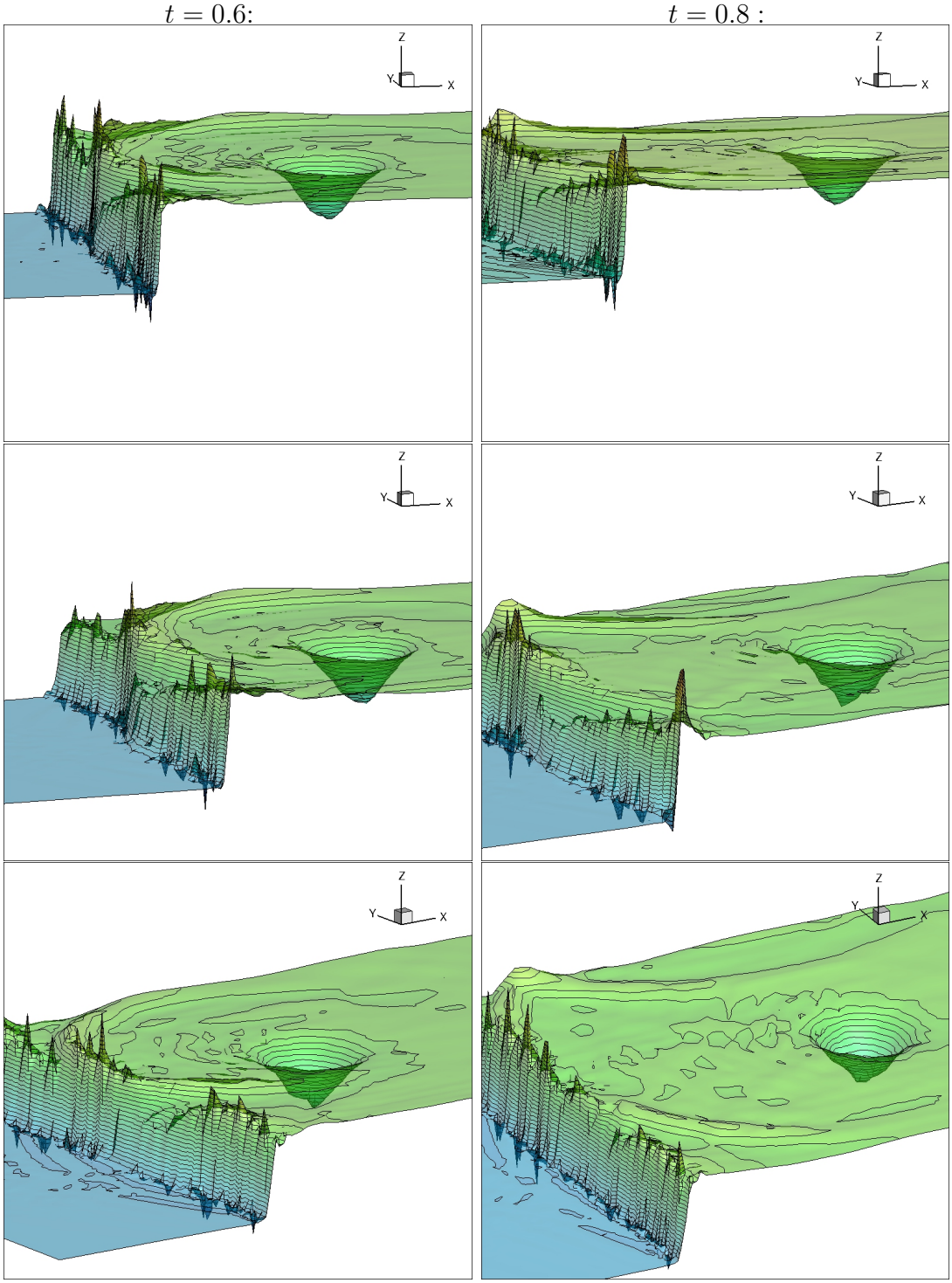


Abbildung 6.11: *APK-SDM ohne u-Rekonstruktion,* $N = 2, p = 1, c = 8$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

$t = 0.05$ 2d- und 3d-Plot

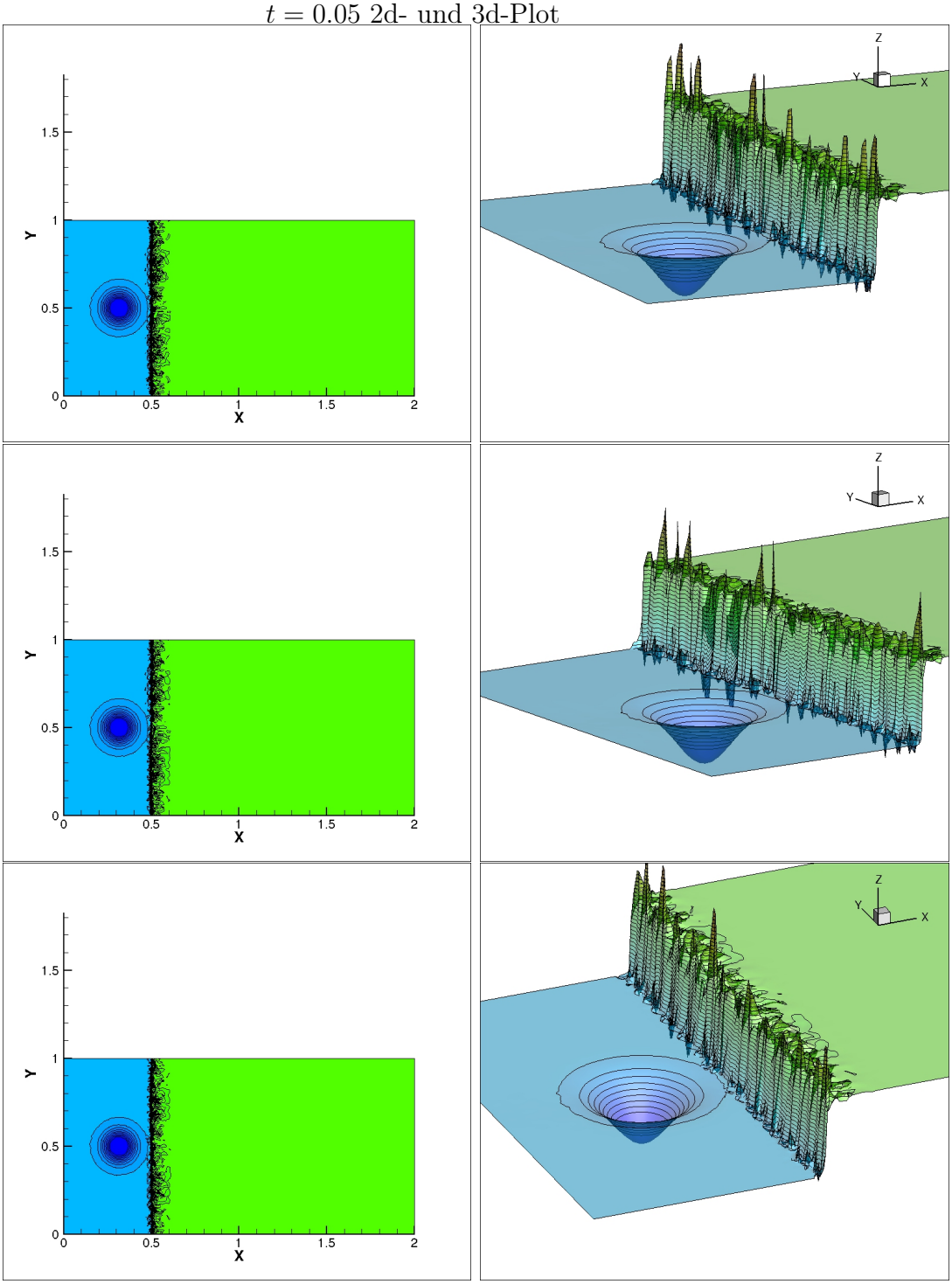


Abbildung 6.12: *APK-SDM ohne u-Rekonstruktion,* $N = 3, p = 2, c = 14$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

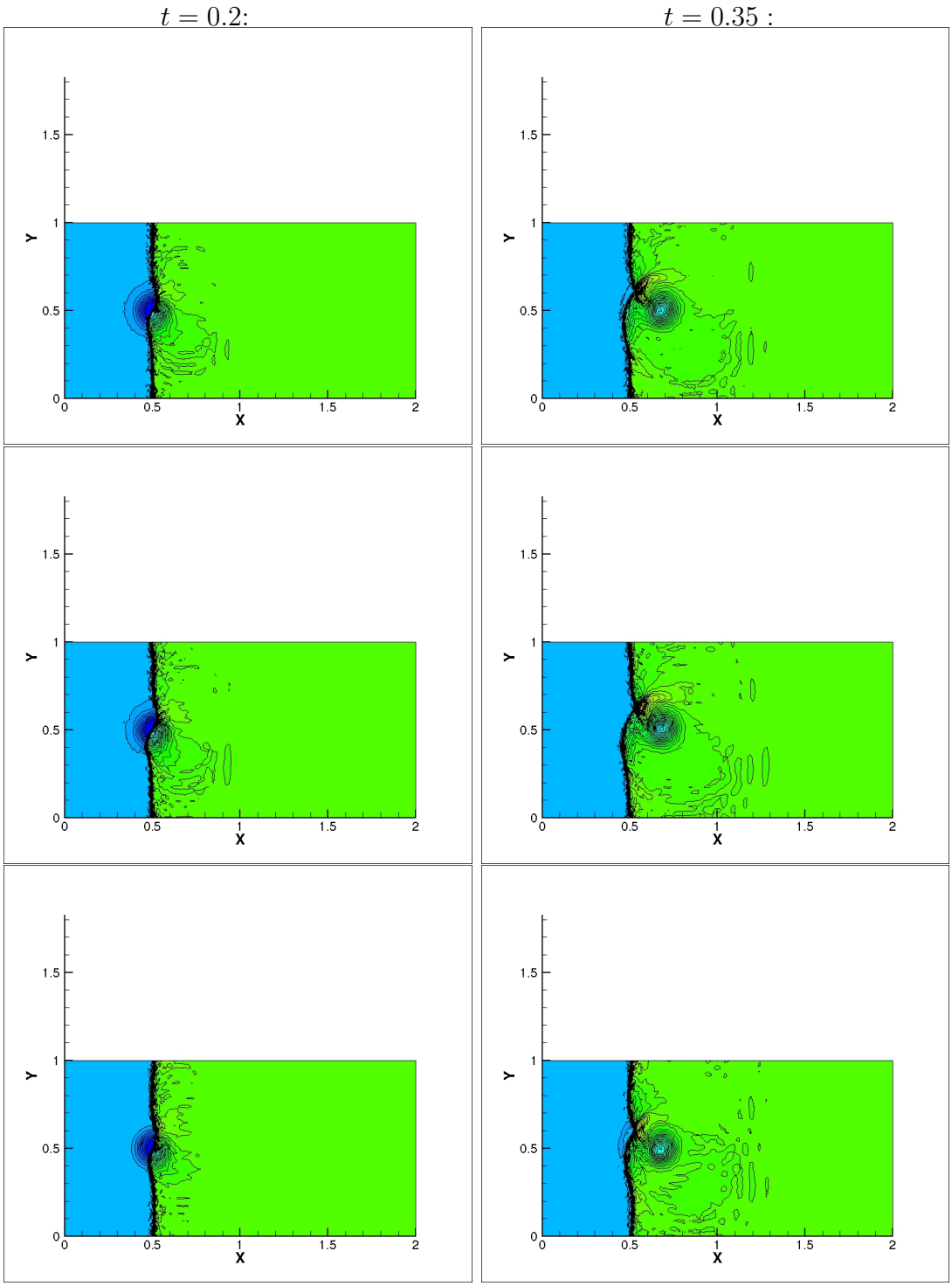


Abbildung 6.13: *APK-SDM ohne u-Rekonstruktion,* $N = 3, p = 2, c = 14$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

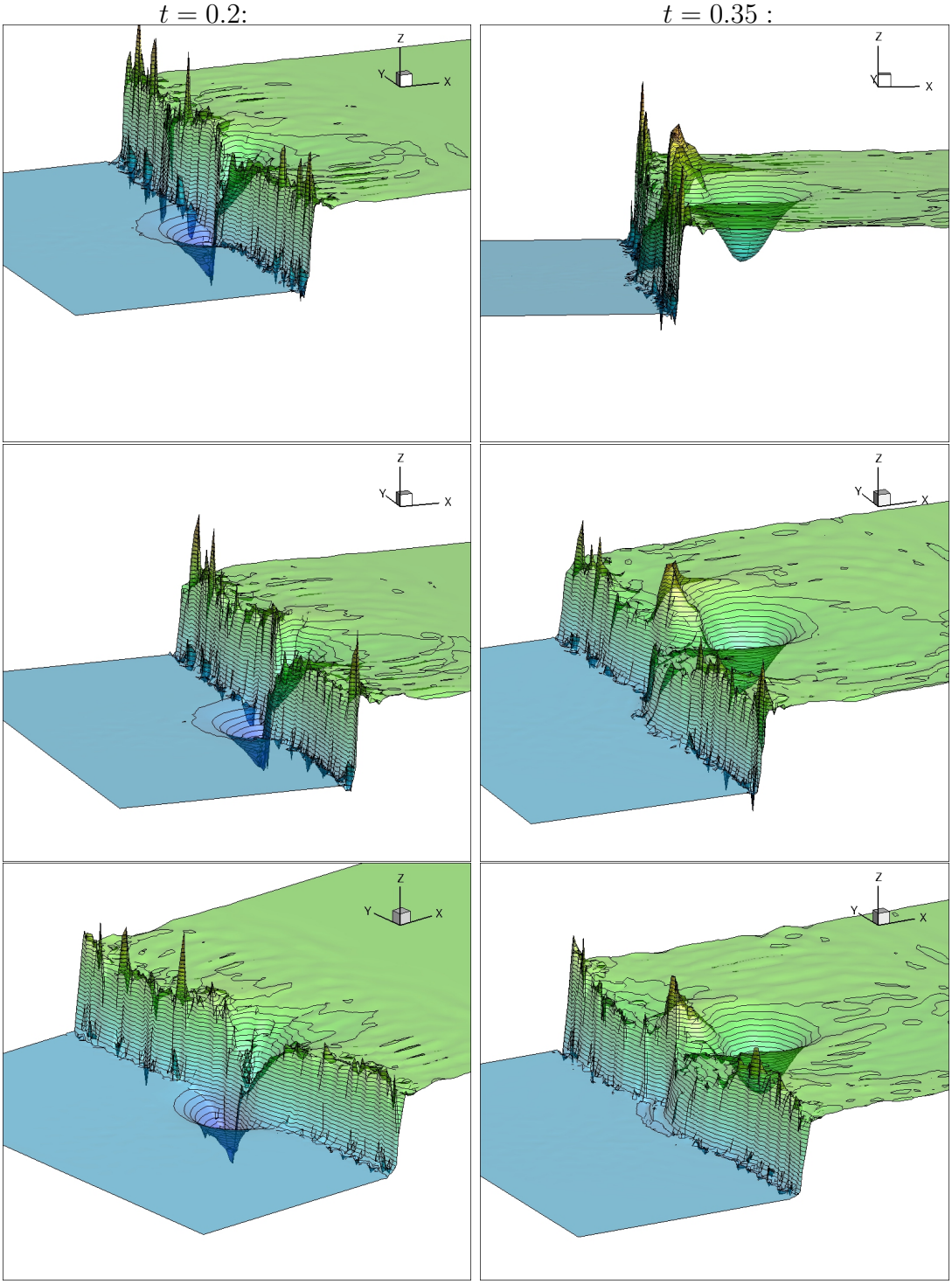


Abbildung 6.14: *APK-SDM ohne u-Rekonstruktion,* $N = 3, p = 2, c = 14$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

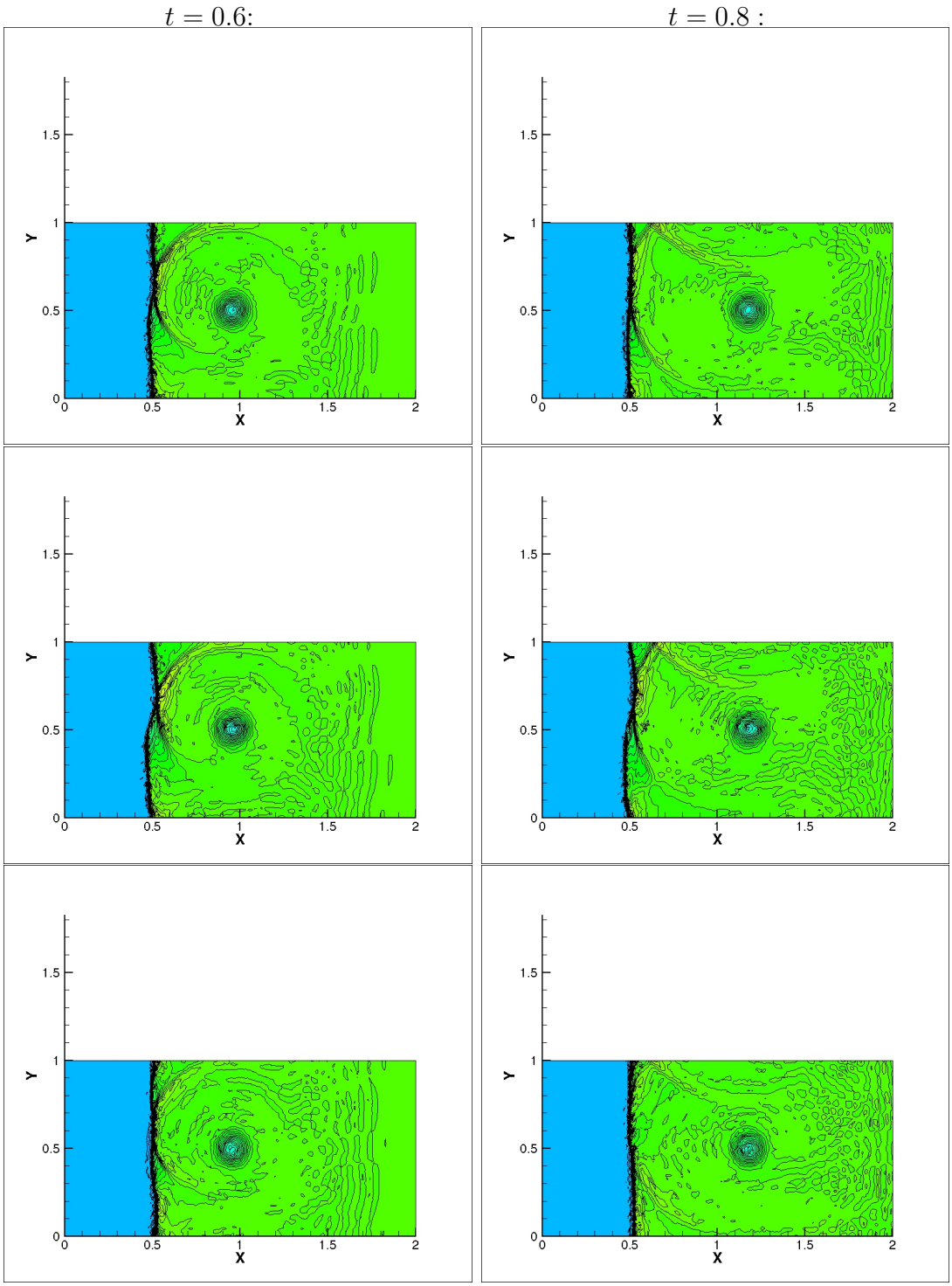


Abbildung 6.15: *APK-SDM ohne u-Rekonstruktion,* $N = 3, p = 2, c = 14$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

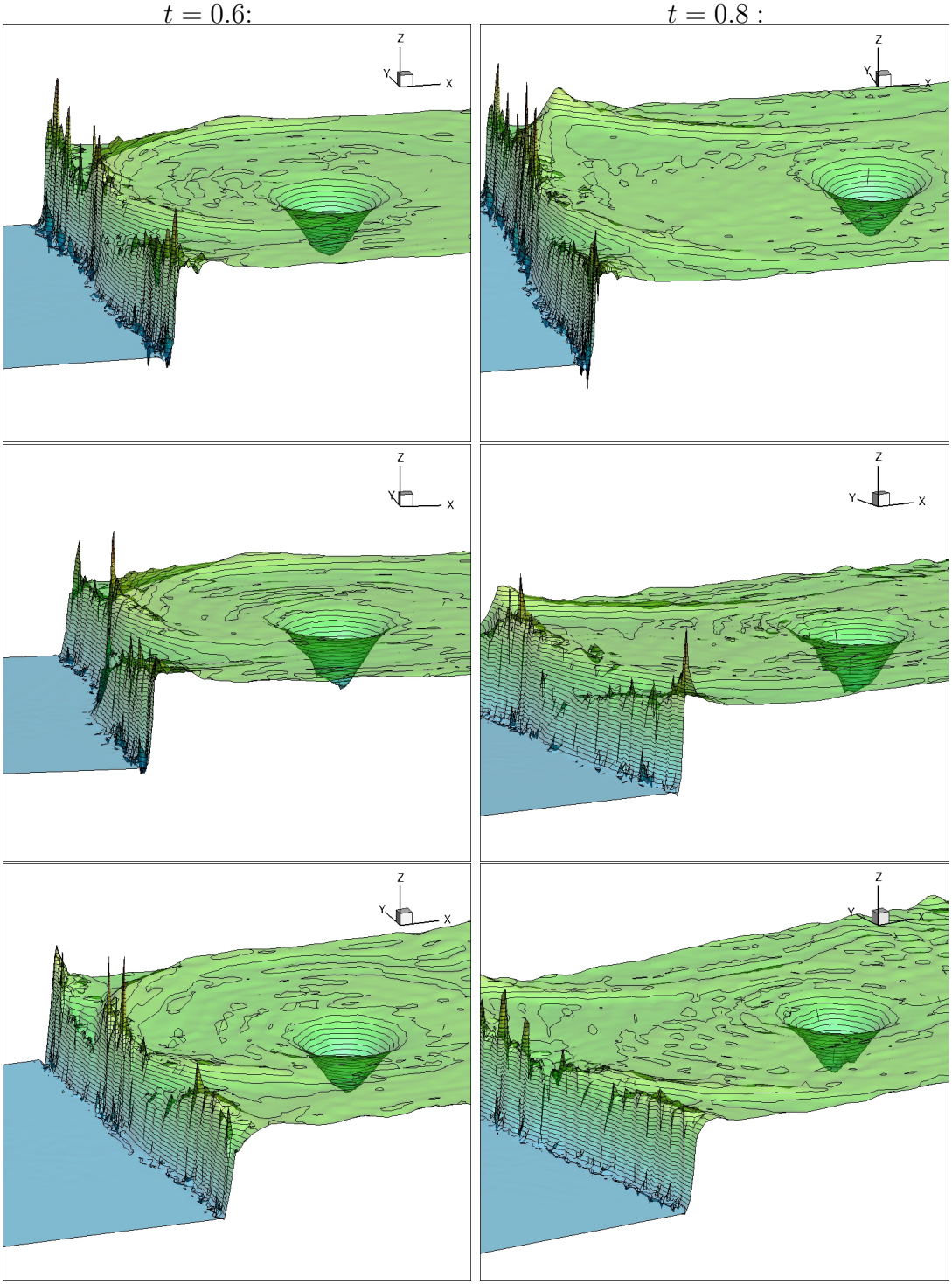


Abbildung 6.16: *APK-SDM ohne u-Rekonstruktion,* $N = 3, p = 2, c = 14$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

$t = 0.05$ 2d- und 3d-Plot

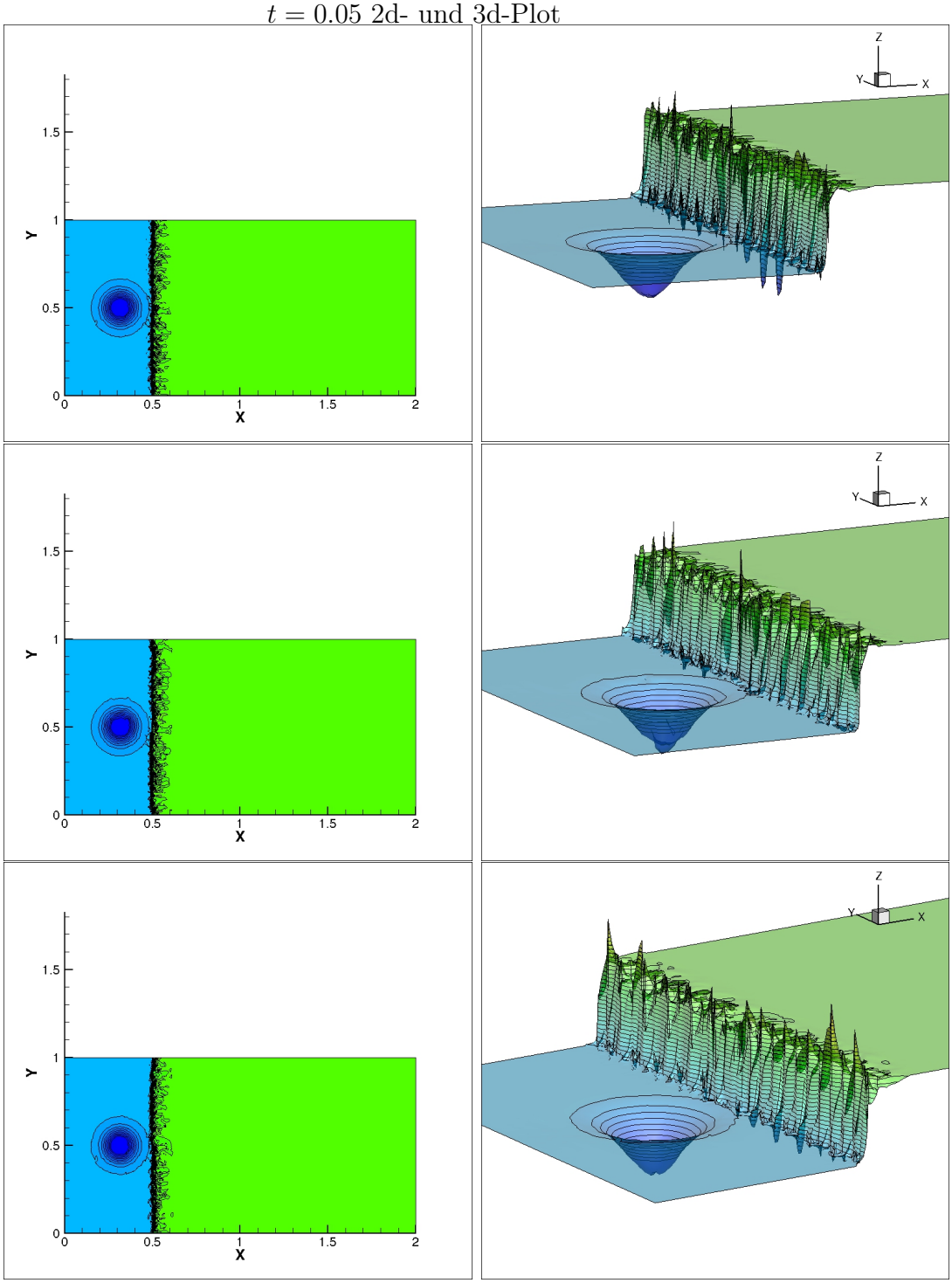


Abbildung 6.17: *APK-SDM ohne u-Rekonstruktion,* $N = 4, p = 2, c = 22$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

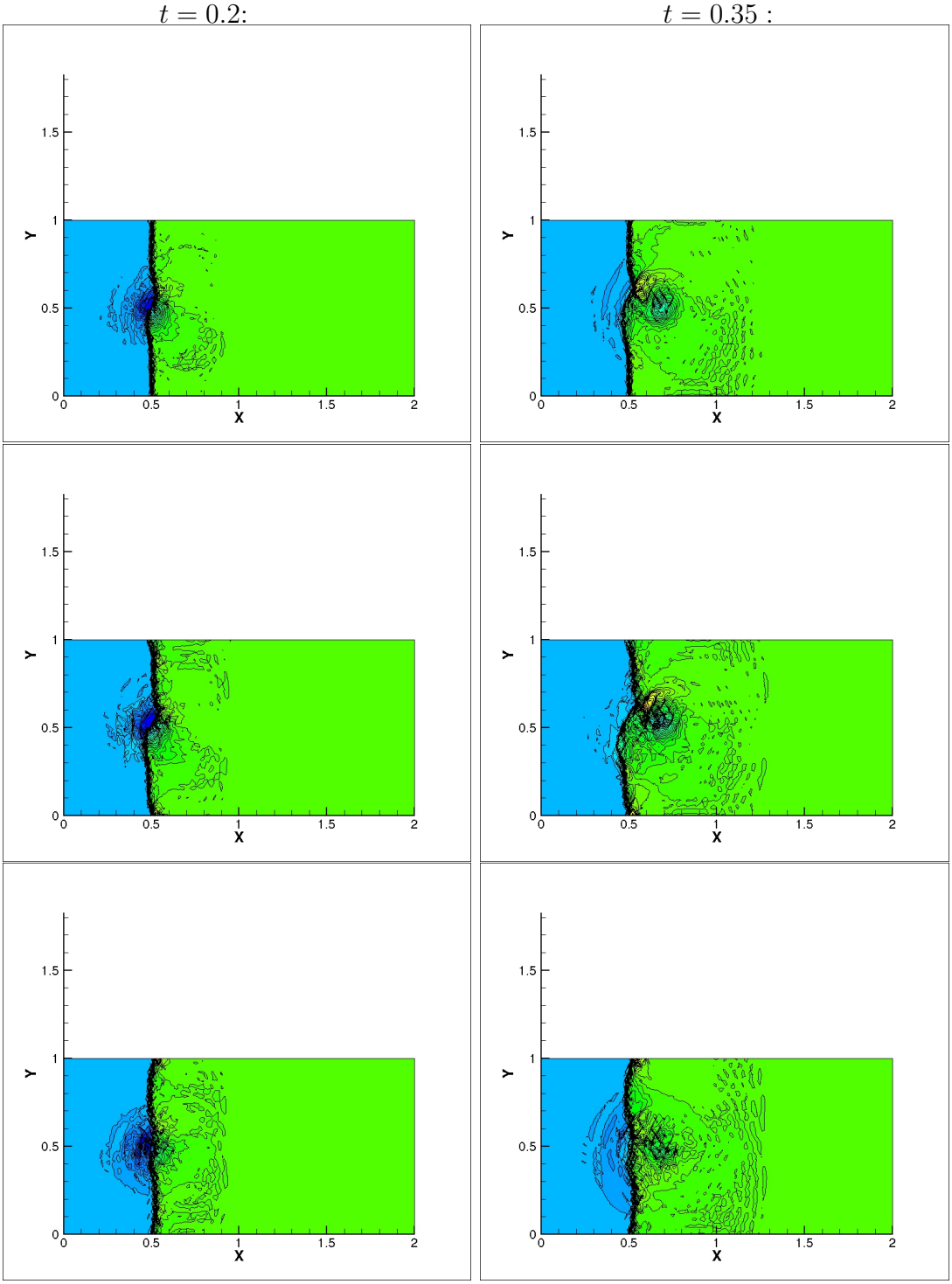


Abbildung 6.18: *APK-SDM ohne u-Rekonstruktion,* $N = 4, p = 2, c = 22$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

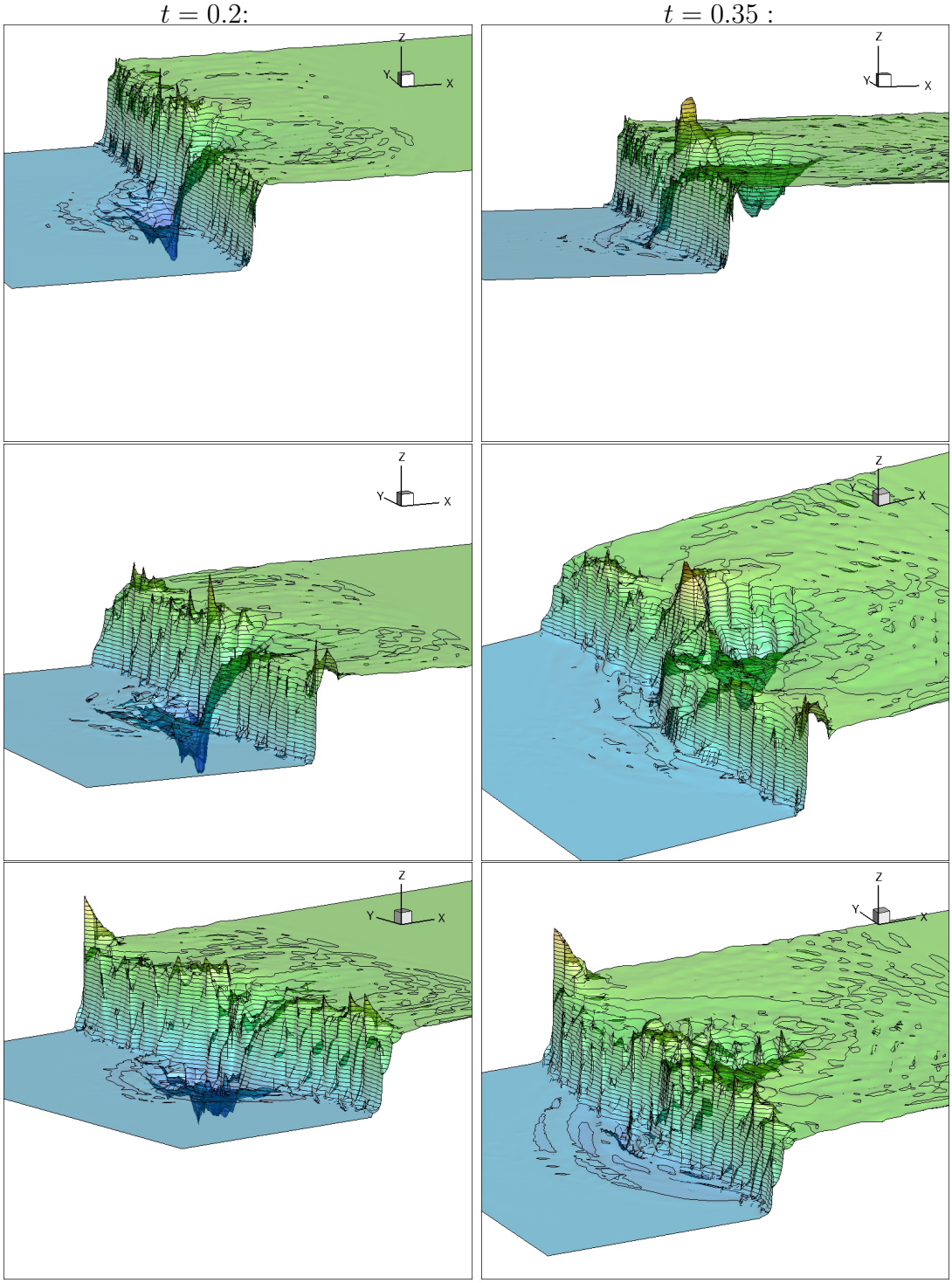


Abbildung 6.19: *APK-SDM ohne u-Rekonstruktion,* $N = 4, p = 2, c = 22$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

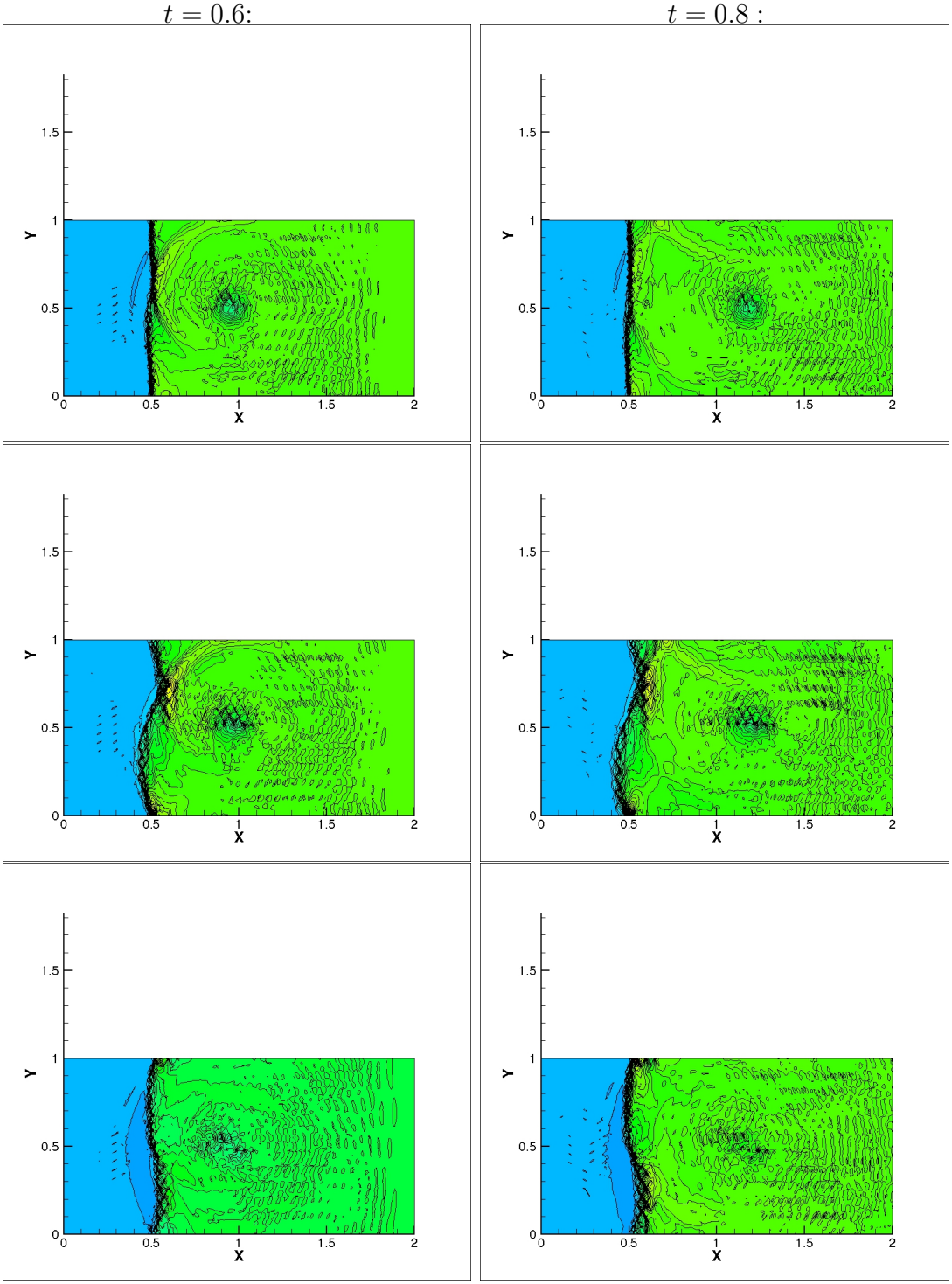


Abbildung 6.20: *APK-SDM ohne u-Rekonstruktion,* $N = 4, p = 2, c = 22$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

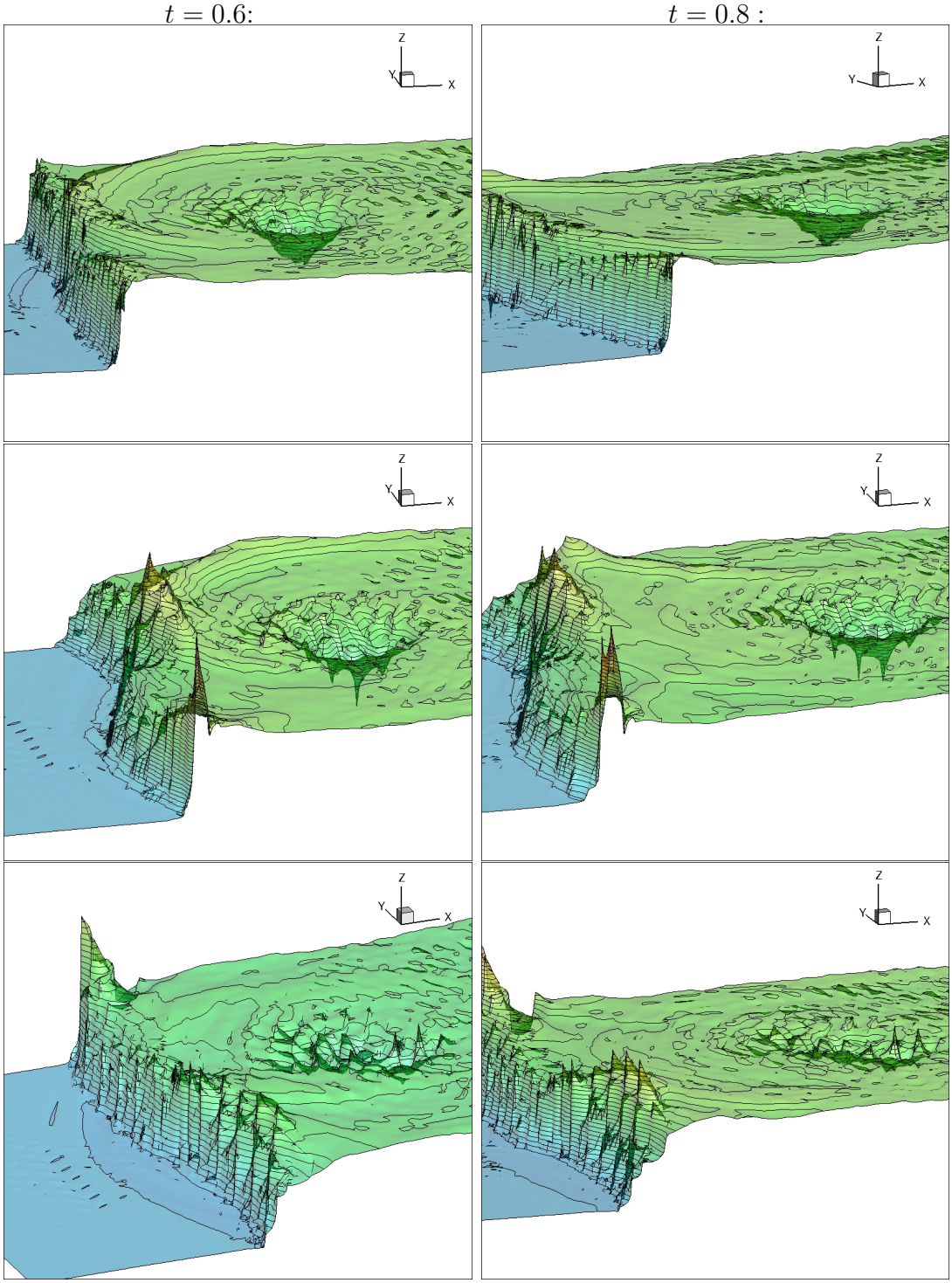


Abbildung 6.21: *APK-SDM ohne u-Rekonstruktion,* $N = 4, p = 2, c = 22$*, Parameter von oben nach unten:* $(\alpha, \beta, \gamma) = (1, 1, 2), (1, 1, 3), (2, 2, 4)$

7 Zusammenfassung und Ausblick

Diese Arbeit beschäftigt sich mit orthogonalen Polynomen und deren Anwendung bei spektralen Methoden zum numerischen Lösen von Systemen hyperbolischer Erhaltungsgleichungen.
Es wurden zunächst die APK-Polynome eingeführt. Dabei handelt es sich um klassische orthogonale Polynome auf einem Dreiecksgebiet. Für sie wurden neue Abschätzungen bewiesen, die dann zur Analyse des Approximationsverhaltens der APK-Fourier-Reihe verwendet wurden. Für die Fourier-Koeffizienten und für den Abschneidefehler wurde spektrale Konvergenz gezeigt. Der Abschneidefehler wurde sowohl in einer gewichteten L^2-Norm als auch in der Maximumsnorm betrachtet. Das Resultat lieferte die theoretische Berechtigung, bei Verwendung der APK-Polynome zur Approximation einer Funktion u mittels ihrer APK-Fourier-Summe von einem spektralen Verfahren zu sprechen.
Wie bekannt, sind unstetige Funktionen in der Theorie als Lösungen von hyperbolischer Erhaltungsgleichungen zugelassen. Berechnet man die numerische Lösung mit Hilfe eines spektralen Verfahrens hoher Ordnung, so weist die Lösung das Gibbs'sche Phänomen in der Nähe der Sprungunstetigkeiten auf. Um diese Oszillationen zu reduzieren, verwendeten wir modale Filter, welche direkt auf die Koeffizienten der Fourier-Reihe wirken. Durch diese Filter werden die hochfrequenten Anteile bearbeitet und so die Oszillationen reduziert. In diesem Zusammenhang wurde auch die gefilterte APK-Reihe untersucht. Es wurde ein Teilergebnis zur Approximation einer hinreichend glatten Funktion mittels der abgeschnittenen, gefilterten APK-Reihe gezeigt. Ein Resultat über die Approximationsgeschwindigkeit einer Funktion mit Sprungunstetigkeiten konnte hier allerdings nicht bewiesen werden. Ein solches Resultat hinsichtlich der Wahl der drei Parameter der APK-Polynome stellt jedoch ein interessantes Feld zukünftiger Untersuchungen dar und so bereitet diese Arbeit einen möglichen Ansatz für einen Beweis bezüglich der Approximation einer Funktion mit Sprungunstetigkeit vor. Dazu wurden die in der Literatur bekannten Resultate aus dem eindimensionalen Fall betrachtet. Bisher sind Ergebnisse für die Approximation mittels trigonometrischer Funktionen bzw. Legendre-Polynomen bekannt, siehe [36] und [80]. Bei der Untersuchung der Arbeit [36] wurde festgestellt, dass die Autoren die Voraussetzungen der Lemmata, welche sie in ihrem Beweis für das Verhalten der gefilterten Legendre-Reihe verwendeten, nicht überprüften. Wir zeigten, dass diese sich nicht auf einen allgemeinen Fall übertragen lassen. So wurde hier das Resultat aus [36] verbessert, um anschließend in einer kritischen Diskussion aufzuzeigen, dass eine modale Filterfunktion σ alle geforderten Voraussetzungen nur erfüllen kann, wenn sie von höherer Filterordnung ist als ursprünglich im Satz aus [36] angegeben. Daher bleibt das Ergebnis aus [80] bisher einzigartig und die weitere Untersuchung des Problems der Approximation einer Funktion mit Unstetigkeit mittels Fourier-Summen

orthogonaler Polynome stellt eine interessante Aufgabe für die weitere Forschungen dar. Es wurde weiterhin für die APK-Polynome aus der Viskositätsformulierung ein Exponentialfilter abgeleitet. Dieser modale Filter hängt von dem Parameter γ explizit ab, der einen direkten Einfluss auf die Approximation hat.
Mit den APK-Polynomen und ihren zugehörigen Exponentialfiltern wurde das Spektrale-Differenzen-Verfahren aus [86] erweitert. In zwei Testfällen wurde anschließend das erweiterte SD-Verfahren hinsichtlich seines Approximationsverhaltens und seiner Stabilität für unterschiedliche APK-Familien untersucht. Es wurden zum einen die Ergebnisse aus [63] und [86] bestätigt: Sowohl die Filterstärke als auch die Filterordnung müssen erhöht werden, wenn man den Polynomgrad erhöht. Zum anderen wurde zusätzlich festgestellt, dass auch die Parameterwahl der APK-Polynome und der Exponentialfilter sich positiv auf die Stabilität auswirken können. Je höher der Wert von γ ist, desto stärker werden die Oszillationen geglättet. Bei einem zu große Parameterwert γ gingen jedoch vorhandene Informationen verloren. Diese wurden weggefiltert. Es stelle sich zwangsläufig die Frage, ob eine optimale Parameterwahl der APK-Polynome für deren Verwendung im SD-Verfahren existiert. In weiteren Forschungsarbeiten sollte daher eine theoretische Analyse zur Beantwortung dieser Fragestellungen vorgenommen werden. Eine Stabilitätsanalyse, wie man sie bereits für die klassische Formulierung der SD-Methode in [38] oder [1] findet, könnte dabei zum Ziel führen.
Neben der Betrachtung der APK-Polynome war ein weiterer Aspekt dieser Arbeit die Erweiterung der Theorie spektraler Verfahren mittels diskreter orthogonaler Polynome, zunächst in einer Variable. Die Verwendung diskreter orthogonaler Polynome war durch die Berechnung der Koeffizienten in der Fourier-Reihe motiviert. Anders als bei den klassischen orthogonalen Polynomen muss man weder ein Integral durch ein Quadraturverfahren auswerten, noch, wie beim Interpolationsansatz, ein lineares Gleichungsystem lösen. Die Berechnung der Koeffizienten erfolgt durch Aufsummerierung an den einzelnen Punkten eines Gitters. Die Auswertung ist daher wesentlich schneller und außerdem exakt, da keinerlei numerische Fehler gemacht werden. Als Beispiel für diskrete orthogonale Polynome wurden schließlich die Hahn-Polynome in einer Variable eingeführt, die bereits in [26] bei einem numerischen Verfahren zum Lösen partieller Differentialgleichungen verwendet werden. Für die Koeffizienten der Fourier-Summe wurde spektrale Genauigkeit gezeigt. Bei der Untersuchung des Abschneidefehlers muss man jedoch noch einen weiteren Punkt beachten. Bei Verwendung äquidistanter Punkte kann es bei der Interpolation zum Runge-Phänomen kommen. Verwendet man diskrete orthogonale Polynome auf gleichverteilten Punkten, hat dieses Phänomen Einfluss auf die Approximation, was sich deutlich im Abschneidefehler wiederspiegelt. Ein möglicher Ansatz, das Runge-Phänomen zu vermeiden, besteht in der Verwendung diskreter orthogonaler Polynome auf nicht-gleichverteilten Gittern. Dieser Aspekt in der Theorie orthogonaler Polynome wurde von zwei verschiedenen Ausgangspunkten vorgestellt.
Zum Abschluss wurde nochmals eine Erweiterung auf ein Dreiecksgitter präsentiert.
Dies sind zugleich Ausblicke für weitere mögliche Forschungsprojekte, da bisher weder die diskreten orthogonalen Polynome auf nicht-äquidistanten Punkten bei einem spektralen Verfahren verwendet wurden, noch eine Untersuchung des Approximationsverhaltens für die mehrdimensionalen Hahn-Polynome existiert.

8 Anhang

Im Anhang stellen wir einige Ergänzungen vor. Wir wiederholen ausgewählte Definitionen, Umrechnungsformeln und Lemmata (meist ohne Beweis), liefern die Graphik des Askey-Schemas und abschließend berechnen wir die Fourier-Koeffizienten aus den APK-Polynomen als Erweiterung, um sie in späteren Arbeiten analysieren zu können.

Wir beginnen mit der Definionen und elementaren Umrechnungsformeln der Gammafunktion, wie man sie in jedem Lehrbuch [62] und jeder Formelsammlung [2] über spezielle Funktionen findet.

Definition 8.1. Sei $x \in \mathbb{R}^+$. Dann ist die Gammafunktion definiert durch

$$\Gamma(x) := \int_0^\infty t^{x-1} \exp(-t)\, \mathrm{d}t.$$

Eine Approximation der Gammafunktion für positive x liefert die **Stirling-Formel**:

$$\Gamma(x) = \sqrt{2\pi}\, x^{x-1/2}\, \mathrm{e}^{-x}\, \mathrm{e}^{\mu(x)} \text{ mit } 0 < \mu(x) < 1/(12x).$$

Die Gammafunktion kann zu einer meromorphen Funktion auf $\mathbb{C} \setminus \{0, -1, -2, \ldots\}$ fortsetzt werden. Es gelten folgende Darstellungs- und Umrechnungsformeln, deren detailliertere Auflistung man in [2] und [67] findet.

- $\Gamma(x+1) = x \cdot \Gamma(x)$ mit $\Gamma(1) = 1$,
- $\Gamma(x) = \lim\limits_{n\to\infty} \frac{n!\, n^x}{x(x+1)(x+2)\cdots(x+n)}$ für alle $x \in \mathbb{C} \setminus \{0, -1, -2, \ldots\}$,
- $\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin(\pi x)}$ für alle $x \in \mathbb{C} \setminus \mathbb{Z}$.
- $\Gamma(m + \frac{1}{2}) = \frac{(2m)!}{m!4^m}\sqrt{\pi}$ für $m = 0, 1, \ldots$.

Die Gammafunktion steht in direktem Zusammenhang zu der Betafunktion.

Definition 8.2. Es seien $a, b \in \mathbb{C}$ mit positivem Realteil. Dann ist die **Betafunktion** definiert durch

$$B(a,b) := \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

So wie man die Gammafunktion zu einer meromorphen Funktion auf $\mathbb{C} \setminus \{0, -1, -2, \ldots\}$ fortsetzen kann, gilt dies auch für die Betafunktion. Für $a, b, a+b \in \mathbb{C} \setminus \{0, -1, -2, \ldots\}$ gilt mit der Gammafunktion:

$$B(a,b) := \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Mit Hilfe dieser Darstellung kann ein verallgemeinerter Binomialkoeffizient für jedes $\alpha \in \mathbb{C}\setminus\{0, -1, -2, \ldots\}$ angegeben werden. Es ist

$$\binom{\alpha}{z} := \begin{cases} 0, \text{ wenn } z \text{ oder } \alpha - z \text{ negative ganze Zahlen sind,} \\ \frac{1}{(\alpha+1)B(z+1,\alpha-z+1)} = \frac{\Gamma(\alpha+1)}{\Gamma(z+1)\Gamma(\alpha-z+1)}. \end{cases} \tag{8.1}$$

Das **Pochhammer-Symbol** , oder auch **steigende Faktorielle**, ist für alle x gegeben durch

$$(x)_0 := 1, \quad (x)_n := \prod_{i=1}^{n}(x+i-1) \quad \forall n = 1, 2, 3, \ldots.$$

In der Kombinatorik hingegen wird mit dem Pochhammer-Symbol die **fallende Faktorielle**

$$(x)_0 := 1, \quad (x)_n := \prod_{i=1}^{n}(x-i+1) \quad \forall n = 1, 2, 3, \ldots$$

bezeichnet.
In dieser Arbeit verwenden wir **ausschließlich** die steigende Faktorielle, wie es auch in der Theorie der **speziellen Funktionen** üblich ist.
Das Pochhammer-Symbol kann auch mit Hilfe der Gammafunktion ausgedrückt werden:

$$(x)_n := \frac{\Gamma(x+n)}{\Gamma(x)}.$$

Es gelten außerdem:

- $(x)_n = (-1)^n(1-x-n)_n$;
- $(n+m)! = n!(n+1)_m$;
- $(-n)_i = (-1)^i i!\binom{n}{i}$.

Als nächstes beweisen wir eine Formel, die wir im Beweis der spektralen Konvergenz der APK-Polynome (Kapitel 3) benötigen.

Lemma 8.3. *Seien $l, m \in \mathbb{N}_0$ und $0 < l + m \leq N$ für $N \in \mathbb{N}$. Dann gilt*

$$\sum_{\substack{0<l+m\leq N \\ l,m\ \in\mathbb{N}_0}} \left(\frac{1}{m+l}\right)^k = \sum_{i=1}^{N} \frac{i+1}{i^k}. \tag{8.2}$$

Beweis. Wir beweisen die Formel mit vollständiger Induktion über N.
Für $N = 1$ gilt

$$\sum_{\substack{0<l+m\leq 1\\ l,m\ \in\mathbb{N}_0}} \left(\frac{1}{m+l}\right)^k = 2\sum_{l=1}^{1}\frac{1}{l^k} = 2 = \sum_{i=1}^{1}\frac{i+1}{i^k}.$$

Sei (8.2) für ein $N \in \mathbb{N}$ bereits gezeigt. Für $N+1$ ist dann

$$\begin{aligned}\sum_{\substack{0<l+m\leq N+1\\ l,m\ \in\mathbb{N}_0}} \left(\frac{1}{m+l}\right)^k &= \sum_{\substack{0<l+m\leq N\\ l,m\ \in\mathbb{N}_0}} \left(\frac{1}{m+l}\right)^k + \sum_{i=0}^{N+1}\frac{1}{(N+1)^k}\\ &\overset{I.V.}{=} \sum_{i=1}^{N}\frac{(i+1)}{i^k} + \frac{N+2}{(N+1)^k}\\ &= \sum_{i=1}^{N+1}\frac{i+1}{i^k}.\end{aligned}$$

□

Elementare Eigenschaften der **Gegenbauer-** oder auch **ultrasphärischen Polynome** wurden, wie auch die nachfolgenden Lemmata, bei der Untersuchung der Legendre-Reihe in Abschnitt 4.2 benötigt. Man findet die Hilfssätze in [36] und die letzten beiden auch in [80], dort mit den Beweisen und allen Voraussetzungen.
Bei den Gegenbauer-Polynomen handelt es sich um die Produkte aus einem Jacobi-Polynom 3.1 mit gleichen Parametern $\alpha = \beta$ und einem zusätzlichen parameterabhängigen Faktor. Sie sind definiert auf dem Intervall $[-1, 1]$ bezüglich des Parameters[1] $\lambda > -\frac{1}{2}, \lambda \neq 0$ durch

$$C_n^\lambda(x) := \frac{(2\lambda)_n}{\left(\lambda+\frac{1}{2}\right)_n} P_n^{\lambda-\frac{1}{2},\lambda-\frac{1}{2}}(x)$$

oder, äquivalent dazu,

$$P_n^{\alpha,\alpha}(x) = \frac{(\alpha+1)_n}{(2\alpha+1)_n} C_N^{\alpha+\frac{1}{2}}(x).$$

Die Gegenbauer-Polynome erfüllen die folgenden Relationen:

$$P_n^{\alpha,-\frac{1}{2}}(x) = \frac{\left(\frac{1}{2}\right)_{n+1}}{\left(\alpha+\frac{1}{2}\right)_n} C_{2n}^{\alpha+\frac{1}{2}}\left(\sqrt{\frac{x+1}{2}}\right), \tag{8.3}$$

$$C_{2n}^\alpha(x) = \frac{\Gamma(\alpha+n)n!2^{2n}}{\Gamma(\alpha)(2n)!} P_n^{\alpha-\frac{1}{2},-\frac{1}{2}}(2x^2-1), \qquad \text{mit } \alpha \neq 0, \tag{8.4}$$

[1] Für $\lambda = 0$ erhält man die Chebyshev-Polynome 1.Art T_n mit dem zusätzlichen Faktor $\frac{2}{n}$, d.h. $C_n^0 = \frac{2}{n}T_n(x)$.

$$C_{2n+1}^{\alpha}(x) = \frac{\Gamma(\alpha+n+1)n!2^{2n+1}}{\Gamma(\alpha)(2n+1)!} x P_n^{\alpha-\frac{1}{2},\frac{1}{2}}(2x^2-1), \quad \text{mit } \alpha \neq 0, \tag{8.5}$$

$$C_{2n}^{\alpha}(0) = (-1)^n \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)(n)!}. \tag{8.6}$$

Lemma 8.4. *Sei $l > 0$. Dann gilt für alle $n \neq \alpha$ und $|\alpha| < n$*

$$\left(\frac{n}{n-\alpha}\right)^l = \sum_{p=0}^{\infty} \hat{b}_p n^{-p}, \quad \hat{b}_p = \binom{p+l-1}{l-1}\alpha^p.$$

Lemma 8.5. *Seien $n, N \in \mathbb{N}$ mit $n \geq 2$ und N gerade. Mit B_{2j} werden die Bernoulli-Zahlen bezeichnet. Für jede Funktion $g \in C^{2n}([0,1])$ gilt das Folgende:*

$$\sum_{k=1}^{N}(-1)^k g\left(\frac{k}{N}\right) = \frac{1}{2}\left[g(1)-g(0)\right] + \sum_{l=1}^{N-1} N^{-2l+1}\left[\frac{B_{2l}}{(2l)!}(4^l-1)\right]\left(g^{(2l-1)}(1)-g^{(2l-1)}(0)\right) + \mathcal{O}(N^{-2n+1}).$$

Lemma 8.6. *Seien $n, N, l \in \mathbb{N}$ mit $n \geq 2$ und N gerade. Dann gilt folgende Formel:*

$$\sum_{k=N+1}^{\infty}(-1)^k k^{-2l} = \frac{1}{2}N^{-2l} + \sum_{j=1}^{n-1} N^{-2l-2j+1}\left[\frac{B_{2j}}{(2j)!}(4^j-1)\left(\frac{(2l+2j-2)!}{(2l-1)!}\right)\right] + \mathcal{O}(N^{-2l-2n+1}),$$

wobei die B_{2j} die Bernoulli-Zahlen sind.

Weiterhin ist noch das Askey-Schema (Abbildung 8.1) abgebildet, das eine Übersicht über orthogonale Polynome liefert.

Im Kapitel 4 wurde bereits einmal die Kantendektektierung mit Hilfe der konjungierten Fourier-Reihe erwähnt. In [86] wurde dieses Verfahren auf die PKD-Polynome angepasst. Dabei wurden die Fourier-Koeffizienten direkt aus den Koeffizienten der PKD-Polynome berechnet. Hier folgt die theoretische Erweiterung auf die allgemeinen APK-Polynome. Wir betrachten analog zu [86, S.75] eine Funktion u^* auf dem Einheitsquadrat $[-1,1]^2$. Die Funktion ist mit

$$u^*(x,y) = \sum_{l=0}^{N}\sum_{m=0}^{N-l} \tilde{u}_{m,l}\hat{A}_{m,l}(\hat{\psi}(x,y))$$

gegeben, wobei die $\hat{A}_{m,l}$ die normierten APK-Polynome sind. Mit $\hat{\psi}$ ist die Transformation vom Einheitsquadrat auf das Dreieck $\mathbb{T}$ gemeint. Gesucht sind die Fourier-Koeffizienten

$$\hat{f}_{\xi,\eta} = \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} u^*(x,y)e^{-\mathrm{i}\pi(x\xi+\eta y)}\mathrm{d}x\mathrm{d}y.$$

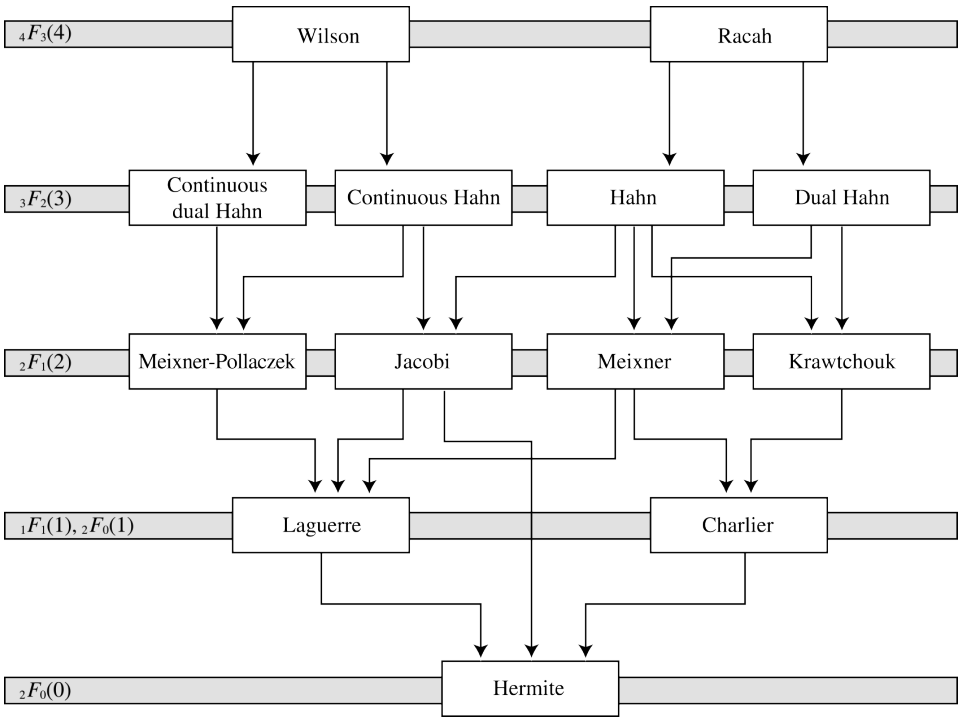


Abbildung 8.1: *Askey-Schema aus* [67]

Satz 8.7. *Seien $m, l, N \in \mathbb{N}_0$ mit $0 \leq m + l \leq N$. Weiterhin seien die Koeffizienten $\hat{u}_{m,l}$, $0 \leq m + l \leq N$ der normierten APK-Polynome $\hat{A}_{m,l}$ im Dreieck $\mathbb{T}$ gegeben[2]. Dann gilt*

$$\hat{f}_{\xi,\eta} = \sum_{l=0}^{N} \sum_{m=0}^{N-l} \hat{u}_{m,l} c_{m,l} \left(\sum_{j=0}^{l} \frac{\Gamma(p+\beta+l+j)}{\Gamma(p+j+1)2^j j!(l-j)!} E(\eta, j) \right) \cdot \left(\sum_{k=0}^{m} \frac{\Gamma(\alpha+a_l+m+k)}{\Gamma(k+a_l+1)2^k k!(m-k)!} E(\xi, k+l) \right),$$

mit

$$E(\eta, j) = \int_{-1}^{1} (y-1)^j e^{-\mathrm{i}\pi\eta y} \mathrm{d}y \text{ und}$$

$$c_{m,l} = \frac{(-1)^{m+l}\sqrt{(2l+\gamma-\alpha)(2(m+l)+\gamma)}}{2^{2+l}} \cdot \sqrt{\frac{\Gamma(l+1)\Gamma(m+1)\Gamma(a_l+1+m)\Gamma(p+l+1)}{\Gamma(l+\beta)\Gamma(m+\alpha)\Gamma(a_l+m+\alpha)\Gamma(p+l+\beta)}}.$$

[2]Durch die Transformation T_i aus Abschnitt 2.2.2 kann dies für jedes Dreieck $\tau_i \in \mathcal{T}$ erweitert werden.

Beweis. Aus der Definition der APK-Polynome 3.5 folgt mit der Transformation[3] $\hat{\psi}$ und der Normierung 3.7.

$$\hat{A}_{m,l}(\hat{\psi}(x,y)) = \sqrt{(2l+\gamma-\alpha)(2(m+l)+\gamma)}\kappa_{l,m}(-1)^m P_m^{a_l,\alpha-1}(x)\left(\frac{1-x}{2}\right)^l P_l^{p,\beta-1}(y)$$

mit $\kappa_{l,m} := \sqrt{\frac{(l+\beta)_p m(m+a_l)_\alpha}{(l+1)_p(m)_\alpha(m+a_l)}}$.
Aus der Definition der Jacobi-Polynome 3.1 erhält man

$$P_m^{a_l,\alpha-1}(x) = \frac{\Gamma(a_l+1+m)}{\Gamma(a_l+m+\alpha)}\sum_{k=0}^{m}\left(\frac{x-1}{2}\right)^k \frac{\Gamma(\alpha+a_l+m+k)}{k!(m-k)!\Gamma(k+a_l+1)}$$

und

$$P_l^{p,\beta-1}(y) = \frac{\Gamma(p+l+1)}{\Gamma(p+l+\beta)}\sum_{j=0}^{l}\frac{\Gamma(p+\beta+l+j)}{\Gamma(p+j+1)j!(l-j)!}\left(\frac{y-1}{2}\right)^j.$$

Verwenden wir die beiden Formeln, so gilt für $\hat{f}_{\xi,\eta}$:

$$\begin{aligned}
\hat{f}_{\xi,\eta} &= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} u^*(x,y)e^{-\mathrm{i}\pi(x\xi+\eta y)}\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}\sum_{l=0}^{N}\sum_{m=0}^{N-l}\tilde{u}_{m,l}\hat{A}_{m,l}(\hat{\psi}(x,y))e^{-\mathrm{i}\pi(x\xi+\eta y)}\mathrm{d}x\,\mathrm{d}y \\
&= \sum_{l=0}^{N}\sum_{m=0}^{N-l}\frac{(-1)^m\tilde{u}_{m,l}}{2^{2+l}||A_{m,l}||_{\mathrm{L}^2(\mathbb{T},h)}}\int_{-1}^{1}\int_{-1}^{1} P_m^{a_l,\alpha-1}(x)\,(1-x)^l\,P_l^{p,\beta-1}(y)e^{-\mathrm{i}\pi(x\xi+\eta y)}\mathrm{d}x\,\mathrm{d}y \\
&= \sum_{l=0}^{N}\sum_{m=0}^{N-l}\frac{(-1)^m\tilde{u}_{m,l}}{2^{2+l}||A_{m,l}||_{\mathrm{L}^2(\mathbb{T},h)}}\int_{-1}^{1} P_m^{a_l,\alpha-1}(x)\,(1-x)^l\,e^{-\mathrm{i}\pi\xi x}\mathrm{d}x\int_{-1}^{1} P_l^{p,\beta-1}(y)e^{-\mathrm{i}\pi\eta y}\mathrm{d}y \\
&= \sum_{l=0}^{N}\sum_{m=0}^{N-l}\frac{(-1)^m\tilde{u}_{m,l}}{2^{2+l}||A_{m,l}||_{\mathrm{L}^2(\mathbb{T},h)}}\frac{\Gamma(a_l+1+m)}{\Gamma(a_l+m+\alpha)}\frac{\Gamma(p+l+1)}{\Gamma(p+l+\beta)} \\
&\quad\cdot\left(\sum_{k=0}^{m}\frac{\Gamma(\alpha+a_l+m+k)}{2^k k!(m-k)!\Gamma(k+a_l+1)}\int_{-1}^{1}(x-1)^k\,(1-x)^l e^{-\mathrm{i}\pi\xi x}\mathrm{d}x\right) \\
&\quad\cdot\left(\sum_{j=0}^{l}\frac{\Gamma(p+\beta+l+j)}{2^j\Gamma(p+j+1)j!(l-j)!}\int_{-1}^{1}(y-1)^j\,e^{-\mathrm{i}\pi\eta y}\mathrm{d}y\right).
\end{aligned}$$

[3]Hierbei ist die Transformationsformel für $\hat{\psi}$ zu beachten.

Nutzen wir die Normierung aus Lemma 3.7 und

$$E(\eta,j)=\int_{-1}^{1}(y-1)^j e^{-\mathrm{i}\pi\eta y}\mathrm{d}y \text{ bzw. } E(\xi,k+l)=\int_{-1}^{1}(x-1)^{k+l}\,e^{-\mathrm{i}\pi\xi x}\mathrm{d}x,$$

vereinfacht sich der Ausdruck zu

$$\begin{aligned}\hat{f}_{\xi,\eta}=&\sum_{l=0}^{N}\sum_{m=0}^{N-l}\frac{(-1)^m\tilde{u}_{m,l}\sqrt{(2l+\gamma-\alpha)(2(m+l)+\gamma)}}{2^{2+l}}\\&\cdot\sqrt{\frac{(l+\beta)_p}{(l+1)_p}\frac{m(m+a_l)_\alpha}{(m+a_l)(m)_\alpha}\frac{\Gamma(a_l+1+m)}{\Gamma(a_l+m+\alpha)}\frac{\Gamma(p+l+1)}{\Gamma(p+l+\beta)}}\\&\cdot\left(\sum_{k=0}^{m}\frac{\Gamma(\alpha+a_l+m+k)(-1)^l E(\xi,k+l)}{2^k k!(m-k)!\Gamma(k+a_l+1)}\right)\left(\sum_{j=0}^{l}\frac{\Gamma(p+\beta+l+j)E(\eta,j)}{2^j\Gamma(p+j+1)j!(l-j)!}\right).\end{aligned}$$

Schreiben wir die Pochhammer-Symbole wieder in Gammmafunktionen, bekommen wir schließlich

$$\begin{aligned}\hat{f}_{\xi,\eta}=&\sum_{l=0}^{N}\sum_{m=0}^{N-l}\frac{(-1)^{m+l}\tilde{u}_{m,l}\sqrt{(2l+\gamma-\alpha)(2(m+l)+\gamma)}}{2^{2+l}}\\&\cdot\sqrt{\frac{\Gamma(l+1)\Gamma(m+1)}{\Gamma(l+\beta)\Gamma(m+\alpha)}\frac{\Gamma(a_l+1+m)}{\Gamma(a_l+m+\alpha)}\frac{\Gamma(p+l+1)}{\Gamma(p+l+\beta)}}\\&\cdot\left(\sum_{k=0}^{m}\frac{\Gamma(\alpha+a_l+m+k)E(\xi,k+l)}{2^k k!(m-k)!\Gamma(k+a_l+1)}\right)\left(\sum_{j=0}^{l}\frac{\Gamma(p+\beta+l+j)E(\eta,j)}{2^j\Gamma(p+j+1)j!(l-j)!}\right)\\=&\sum_{l=0}^{N}\sum_{m=0}^{N-l}\hat{u}_{m,l}c_{m,l}\left(\sum_{j=0}^{l}\frac{\Gamma(p+\beta+l+j)}{\Gamma(p+j+1)2^j j!(l-j)!}E(\eta,j)\right)\\&\cdot\left(\sum_{k=0}^{m}\frac{\Gamma(\alpha+a_l+m+k)}{\Gamma(k+a_l+1)2^k k!(m-k)!}E(\xi,k+l)\right)\end{aligned}$$

mit

$$c_{m,l}=\frac{(-1)^{m+l}\sqrt{(2l+\gamma-\alpha)(2(m+l)+\gamma)\Gamma(l+1)\Gamma(m+1)\Gamma(a_l+1+m)\Gamma(p+l+1)}}{2^{2+l}\sqrt{\Gamma(l+\beta)\Gamma(m+\alpha)\Gamma(a_l+m+\alpha)\Gamma(p+l+\beta)}}.$$

□

Für die $E(\eta,j)$ sind auch explizite Darstellungen bekannt [86, S.77] .

Literaturverzeichnis

[1] Abeele, K. Van den, Ch. Lacor und Z. J. Wang: *On the stability and accuracy of the spectral difference method*. J. Sci. Comput., 37(2):162–188, 2008.

[2] Abramowitz, M. und I.Stegun: *Handbook of Mathemaical Functions with Formulas, Graphs, and Mathematical Tables*. Nummer 55 in *Applied Mathematical*. National Bureau of Standards, 1972.

[3] Andrews, G. E. und R. Askey: *Classical orthogonal polynomials*. In: *Orthogonal polynomials and applications (Bar-le-Duc, 1984)*, Band 1171 der Reihe *Lecture Notes in Math.*, Seiten 36–62. Springer, Berlin, 1985.

[4] Barter, G. E. und D. L. Darmofal: *Shock Capturing with Higher-Order, PDE-Based Artificial Viscosity*. AIAA-2007-3823, 2007.

[5] Blyth, M. G., H. Luo und C. Pozrikidis: *A comparison of interpolation grids over the triangle or the tetrahedron*. J. Engrg. Math., 56(3):263–272, 2006.

[6] Blyth, M. G. und C. Pozrikidis: *A Lobatto interpolation grid over the triangle*. IMA J. Appl. Math., 71(1):153–169, 2006.

[7] Bôcher, M.: *Introduction to the theory of Fourier's series*. Ann. of Math. (2), 7(3):81–152, 1906.

[8] Braess, D. und Ch. Schwab: *Approximation on simplices with respect to weighted Sobolev norms*. J. Approx. Theory, 103(2):329–337, 2000.

[9] Butcher, J. C.: *The numerical analysis of ordinary differential equations*. A Wiley-Interscience Publication. John Wiley & Sons, Ltd., Chichester, 1987.

[10] Butzer, P. und F. Jongmans: *P. L. Chebyshev (1821–1894). A guide to his life and work*. J. Approx. Theory, 96(1):111–138, 1999.

[11] Canuto, C., M. Y. Hussaini, A. Quarteroni und T. A. Zang: *Spectral methods in fluid dynamics*. Springer Series in Computational Physics. Springer-Verlag, New York, 1988.

[12] Carpenter, M. und C. Kennedy: *Fourth-Order 2N-Storage Runge-Kutta Schemes*. Technical Report 109112, NASA, 1994.

[13] Chen, G. Q., Q. Du und E. Tadmor: *Spectral viscosity approximations to multidimensional scalar conservation laws*. Math. Comp., 61(204):629–643, 1993.

[14] Chen, Q. und I. Babuška: *Approximate optimal points for polynomial interpolation of real functions in an interval and in a triangle.* Comput. Methods Appl. Mech. Engrg., 128(3-4):405–417, 1995.

[15] Cools, R.: *An encyclopaedia of cubature formulas.* J. Complexity, 19(3):445–453, 2003. Numerical integration and its complexity (Oberwolfach, 2001).

[16] Dacorogna, B.: *Weak continuity and weak lower semicontinuity of nonlinear functionals*, Band 922 der Reihe *Lecture Notes in Mathematics*. Springer-Verlag, Berlin-New York, 1982.

[17] Dubiner, M.: *Spectral methods on triangles and other domains.* J. Sci. Comput., 6(4):345–390, 1991.

[18] Duchon, C. D.: *Lanczos Filtering on One and Two Dimensions.* J. of Applied Meteorology, 18:1016–1022, 1979.

[19] Dunkl, C. F. und Y. Xu: *Orthogonal polynomials of several variables*, Band 81 der Reihe *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2001.

[20] Eisinberg, A. und G. Fedele: *Discrete orthogonal polynomials on Gauss-Lobatto Chebyshev nodes.* J. Approx. Theory, 144(2):238–246, 2007.

[21] Ekaterinaris, J. A.: *High-order accurate, low numerical diffusion methods for aerodynamics.* Progress in Aerospace Sciences, 41:192–300, 2005.

[22] Emmel, L., S. M. Kaber und Y. Maday: *Padé-Jacobi filtering for spectral approximations of discontinuous solutions.* Numer. Algorithms, 33(1-4):251–264, 2003. International Conference on Numerical Algorithms, Vol. I (Marrakesh, 2001).

[23] Evans, L. C.: *Partial differential equations*, Band 19 der Reihe *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

[24] Gautschi, W.: *Orthogonal polynomials: computation and approximation.* Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2004. Oxford Science Publications.

[25] Gelb, A.: *Reconstruction of piecewise smooth functions from non-uniform grid point data.* J. Sci. Comput., 30(3):409–440, 2007.

[26] Gelb, A., R. B. Platte und W. St. Rosenthal: *The discrete orthogonal polynomial least squares method for approximation and solving partial differential equations.* Commun. Comput. Phys., 3(3):734–758, 2008.

[27] Gelb, A. und J. Tanner: *Robust reprojection methods for the resolution of the Gibbs phenomenon.* Appl. Comput. Harmon. Anal., 20(1):3–25, 2006.

[28] Godlewski, E. und P.-A. Raviart: *Hyperbolic systems of conservation laws*, Band 3/4 der Reihe *Mathématiques & Applications (Paris) [Mathematics and Applications]*. Ellipses, Paris, 1991.

[29] Gottlieb, D. und J. S. Hesthaven: *Spectral methods for hyperbolic problems*. J. Comput. Appl. Math., 128(1-2):83–131, 2001. Numerical analysis 2000, Vol. VII, Partial differential equations.

[30] Gottlieb, D. und S. A. Orszag: *Numerical analysis of spectral methods: theory and applications*. Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1977. CBMS-NSF Regional Conference Series in Applied Mathematics, No. 26.

[31] Gottlieb, D. und C.-W. Shu: *On the Gibbs phenomenon and its resolution*. SIAM Rev., 39(4):644–668, 1997.

[32] Haagerup, U. und H. Schlichtkrull: *Inequalities for Jacobi polynomials*. Ramanujan Journal, 33(2):227–246, 2014.

[33] Hahn, W.: *Über Orthogonalpolynome, die q-Differenzengleichungen genügen*. Math. Nachr., 2:4–34, 1949.

[34] Hairer, E., S. P. Nørsett und G. Wanner: *Solving ordinary differential equations. I*, Band 8 der Reihe *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Second Auflage, 1993. Nonstiff problems.

[35] Hesthaven, J. S.: *From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex*. SIAM J. Numer. Anal., 35(2):655–676, 1998.

[36] Hesthaven, J. S. und R. M. Kirby: *Filtering in Legendre spectral methods*. Math. Comp., 77(263):1425–1452, 2008.

[37] Hewitt, E. und R. Hewitt: *The Gibbs-Wilbraham phenomenon: An episode in fourier analysis*. Archive for History of Exact Sciences 20. XII, 21(2):129–160, 1979.

[38] Jameson, A.: *A proof of the stability of the spectral difference method for all orders of accuracy*. J. Sci. Comput., 45(1-3):348–358, 2010.

[39] Jantscher, L.: *Distributionen*. Walter de Gruyter, Berlin-New York, 1971. de Gruyter Lehrbuch.

[40] Jiang, G.-S. und C.-W. Shu: *Efficient Implementation of Weight ENO Schemes*. Journal of Computational Physics, 126(1):202–228, 1996.

[41] Kaber, S. M.: *A Legendre pseudospectral viscosity method*. J. Comput. Phys., 128(1):165–180, 1996.

[42] Karlin, S. und J. McGregor: *The classification of birth and death processes*. Trans. Amer. Math. Soc., 86:366–400, 1957.

[43] Karlin, S. und J. McGregor: *Linear growth birth and death processes*. J. Math. Mech., 7:643–662, 1958.

[44] Karniadakis, G. E. und S. J. Sherwin: *Spectral/hp element methods for computational fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, Second Auflage, 2005.

[45] Koekoek, R., P. A. Lesky und R. F. Swarttouw: *Hypergeometric orthogonal polynomials and their q-analogues*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2010. With a foreword by Tom H. Koornwinder.

[46] Koornwinder, T.: *Two-variable analogues of the classical orthogonal polynomials*. In: *Theory and application of special functions (Proc. Advanced Sem., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1975)*, Seiten 435–495. Math. Res. Center, Univ. Wisconsin, Publ. No. 35. Academic Press, New York, 1975.

[47] Koornwinder, T.: *Orthogonal polynomials with weight function* $(1-x)^\alpha(1+x)^\beta + M\delta(x+1) + N\delta(x-1)$. Canad. Math. Bull., 27(2):205–214, 1984.

[48] Lax, P. D.: *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*. Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1973. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 11.

[49] Lee, D. W.: *Difference equations for discrete classical multiple orthogonal polynomials*. J. Approx. Theory, 150(2):132–152, 2008.

[50] Liu, Y., M. Vinokur und Z. J. Wang: *Spectral difference method for unstructured grids. I. Basic formulation*. J. Comput. Phys., 216(2):780–801, 2006.

[51] Lui Y., M. Vinokur und Z.J. Wang: *Mulit-Dimensional Spectral Difference Method for Unstructured Grids*. NAS Technical Report, (NAS-05-009), 2005.

[52] Ma, H.: *Chebyshev-Legendre spectral viscosity method for nonlinear conservation laws*. SIAM J. Numer. Anal., 35(3):869–892 (electronic), 1998.

[53] Ma, H.: *Chebyshev-Legendre super spectral viscosity method for nonlinear conservation laws*. SIAM J. Numer. Anal., 35(3):893–908 (electronic), 1998.

[54] Maday, Y., S. M. Kaber und E. Tadmor: *Legendre pseudospectral viscosity method for nonlinear conservation laws*. SIAM J. Numer. Anal., 30(2):321–342, 1993.

[55] Meister, A., S. Ortleb und T. Sonar: *Application of spectral filtering to discontinuous Galerkin methods on triangulations*. Numer. Methods Partial Differential Equations, 28(6):1840–1868, 2012.

[56] Meister, A., S. Ortleb und T. Sonar: *New adaptive modal and DTV filtering routines for the DG method on triangular grids applied to the Euler equations*. GEM Int. J. Geomath., 3(1):17–50, 2012.

[57] Meister, A., S. Ortleb, T. Sonar und M. Wirz: *A comparison of the discontinuous-Galerkin- and spectral-difference-method on triangulations using PKD polynomials*. J. Comput. Phys., 231(23), 2012.

[58] Meister, A., S. Ortleb, T. Sonar und M. Wirz: *An extended discontinuous Galerkin and spectral difference method with modal filtering*. ZAMM Z. Angew. Math. Mech., 93(6-7):459–464, 2013.

[59] Nikiforov, A. F., S. K. Suslov und V. B. Uvarov: *Classical orthogonal polynomials of a discrete variable*. Springer Series in Computational Physics. Springer-Verlag, Berlin, 1991. Translated from the Russian.

[60] Öffner, P. und T. Sonar: *Spectral convergence for orthogonal polynomials on triangles*. Numer. Math., 124(4):701–721, 2013.

[61] Öffner, P., T. Sonar und M. Wirz: *Detecting strength and location of jump discontinuities in numerical data1-14*. Applied Mathematics 4, (12A):1–14, 2013.

[62] Olver, F., D. Lozier, Roland B. und Ch. Clark (Herausgeber): *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY, 2010. Gedruckte Version zu [67].

[63] Ortleb, S.: *Ein diskontinuierliches Galerkin-Verfahren hoher Ordnung auf Dreiecksgittern mit modaler Filterung zur Lösung hyperbolischer Erhaltungsgleichungen*. Doktorarbeit, PhD Thesis, Universität Kassel, 2011.

[64] Osilenker, B.: *Fourier series in orthogonal polynomials*. World Scientific Publishing Co., Inc., River Edge, NJ, 1999.

[65] Persson, P. O. und J. Peraire: *Sub-Cell Shock Capturing for Discontinuous Galerkin Methods*. AIAA-2006-112, 2006.

[66] Proriol, J.: *Sur une famille de polynomes à deux variables orthogonaux dans un triangle*. C. R. Acad. Sci. Paris, 245:2459–2461, 1957.

[67] *NIST Digital Library of Mathematical Functions*. "http://dlmf.nist.gov/, Release 1.0.9 of 2014-08-29". Online Zusammenfassung von [62].

[68] Sakamoto, R.: *Hyperbolic boundary value problems*. Cambridge University Press, 1982.

[69] Schaback, R. und R. Werner: *Numerische Mathematik*. Nummer 4 in *Springer-Lehrbuch*. Springer Verlag, 1991.

[70] Schwab, Ch.: *p- and hp-finite element methods*. Numerical Mathematics and Scientific Computation. The Clarendon Press, Oxford University Press, New York, 1998. Theory and applications in solid and fluid mechanics.

[71] Smith, S. J.: *Lebesgue constants in polynomial interpolation*. Ann. Math. Inform., 33:109–123, 2006.

[72] Suetin, P. K.: *Orthogonal polynomials in two variables*, Band 3 der Reihe *Analytical Methods and Special Functions*. Gordon and Breach Science Publishers, Amsterdam, 1999. Translated from the 1988 Russian original by E. V. Pankratiev [E. V. Pankrat′ev].

[73] Szegö, G.: *Orthogonal Polynomials*. American Mathematical Society, New York, 1939. American Mathematical Society Colloquium Publications, v. 23.

[74] Tadmor, .: *Convergence of spectral methods for nonlinear conservation laws*. SIAM J. Numer. Anal., 26(1):30–44, 1989.

[75] Tadmor, E.: *Super-viscosity and spectral approximations of nonlinear conservation laws*. In: *Numerical methods for fluid dynamics, 4 (Reading, 1992)*, Seiten 69–81. Oxford Univ. Press, New York, 1993.

[76] Taylor, M. A. und B. A. Wingate: *The natural function space for triangular and tetrahedral spectral elements*. Los Alamos National Laboratory- Forschungsbericht, (LA-UR-98-1711), 1998.

[77] Taylor, M. A., B. A. Wingate und R. E. Vincent: *An algorithm for computing Fekete points in the triangle*. SIAM J. Numer. Anal., 38(5):1707–1720 (electronic), 2000.

[78] Toro, E. F.: *Riemann solvers and numerical methods for fluid dynamics*. Springer-Verlag, Berlin, Second Auflage, 1999. A practical introduction.

[79] Vaidyanathan, P. P.: *Multirate Systems and Filter Banks*. Prentice Hall P T R, Upper Saddle River, New Jersey 07458, 1993.

[80] Vandeven, H.: *Family of spectral filters for discontinuous problems*. J. Sci. Comput., 6(2):159–192, 1991.

[81] Von Neumann, J. und R. D. Richtmyer: *A method for the numerical calculation of hydrodynamic shocks*. J. Appl. Phys., 21:232–237, 1950.

[82] Wang, Z.J.: *High-order methods for the Euler and Navier-Stokes equations on unstructured grids*. Progress in Aerospace Sciences, 43:1–41, 2007.

[83] Warburton, T.: *An explicit construction of interpolation nodes on the simplex*. J. Engrg. Math., 56(3):247–262, 2006.

[84] Warnecke, G.: *Analytische Methoden in der Theorie der Erhaltungsgleichungen*, Band 138 der Reihe *Teubner-Texte zur Mathematik [Teubner Texts in Mathematics]*. B. G. Teubner Verlagsgesellschaft mbH, Stuttgart, 1999.

[85] Wilbraham, H.: *On a certain periodic function*. Cambridge and Dublin Mathematical Journal, 3:198–201, 1848.

[86] Wirz, M.: *Ein Spektrale-Differenzen-Verfahren mit modaler Filterung und zweidimensionaler Kantendetektierung mithilfe konjungierter Fourierreihen*. Doktorarbeit, PhD Thesis, Technische Universität Braunschweig, 2012.

[87] Wirz, M.: *Detecting edges in high order methods for hyperbolic conservation laws.* In: *High Order Nonlinear Numerical Schemes for Evolutionary PDEs*, Band 99 der Reihe *Lecture Notes in Computational Science and Engineering*, Seiten 151–168. Springer, 2013.

[88] Xu, Y.: *On discrete orthogonal polynomials of several variables.* Advances in Applied Mathematics, 33(3):615–631, 2004.

[89] Xu, Y.: *Second-order difference equations and discrete orthogonal polynomials of two variables.* Int. Math. Res. Not., (8):449–475, 2005.

[90] Xu, Y.: *On Gauss-Lobatto integration on the triangle.* SIAM J. Numer. Anal., 49(2):541–548, 2011.

[91] Zygmund, A.: *Trigonometric series. Vol. I, II.* Cambridge University Press, Cambridge-New York-Melbourne, 1977. Reprinting of the 1968 version of the second edition with Volumes I and II bound together.