Fakultät für Physik und Astronomie
Ruprecht-Karls-Universität Heidelberg

Adaptive Finite-Elemente-Methoden zur Lösung der Wellengleichung mit Anwendung in der Physik der Sonne

Diplomarbeit
im Studiengang Physik

vorgelegt von
Wolfgang Bangerth
aus Ruit auf den Fildern

1998

Adaptive Finite-Elemente-Methoden zur Lösung der Wellengleichung mit Anwendung in der Physik der Sonne

Die Diplomarbeit wurde von Wolfgang Bangerth ausgeführt am

Institut für Angewandte Mathematik der Universität Heidelberg

unter der Betreuung von

Herrn Professor Rolf Rannacher

sowie von

Herrn Professor Peter Ulmschneider

Institut für Theoretische Astrophysik

Allen, die mich hierher begleitet haben.

Zusammenfassung:
Im Rahmen dieser Arbeit werden adaptive Methoden zur numerischen Lösung der Wellengleichung in inhomogenen Medien hergeleitet und auf ein Beispiel aus der Physik der Sonnenatmosphäre angewandt. Ziel der Arbeit ist es, den Fehler der numerischen Lösung bezüglich eines beliebigen Funktionals (die Auswertungsgröße) abzuschätzen und die Schätzung zur Erzeugung von Gittern zu verwenden, die optimal an das Funktional angepaßt sind.
Vorteile und Schwierigkeiten der Methode werden präsentiert. Insbesondere wird gezeigt, daß der vorgeschlagene Ansatz in vielen Fällen effizienter als bisherige adaptive Methoden, die das Zielfunktional nicht mitberücksichtigten, ist. In den Fällen mit nichtlinearen Zielfunktionalen, in denen das Verfahren nicht funktionierte, werden theoretische Erklärungen und numerische Ergebnisse des Fehlschlagens gegeben.
Die vorgeschlagene Methode wird auf ein einfaches Modell der Physik der Sonnenatmosphäre angewandt und die Ausbreitung linearer akustischer Wellen berechnet. Der Anteil der Energie der Wellen, der den Übergang zwischen der Chromosphäre und der Korona passieren kann, wird mit guter Genauigkeit bestimmt.

Abstract:

Adaptive Finite Element Methods for the Solution of the Wave Equation and Application to Solar Physics

In this work, adaptive concepts for the numerical solution of the wave equation in inhomogeneous media are derived and applied to an example taken from the physics of the solar atmosphere. The main focus is on ways to estimate the error in the numerical solution with regard to arbitrary functionals, i.e. quantities of interest, and the use of these estimates for the generation of computational meshes best suited for the evaluation of this functional.
Advantages and difficulties of this method are presented. In particular, it is shown that the proposed approach is significantly better in many cases than previous adaptive schemes not taking into account the quantity of interest. Cases involving nonlinear functionals and in which the approach fails, are presented along with theoretical explanations and numerical evidence of the reasons for this.
The proposed methods are applied to a simple model from the physics of the solar atmosphere and the propagation of linear acoustic waves is computed. The fraction of the wave energy that passes the chromosphere–corona transition is computed to good accuracy.

Inhaltsverzeichnis

Kapitel 1

Einführung

Im Rahmen dieser Arbeit soll die numerische Lösung der Wellengleichung

$$\begin{aligned}
\rho(\mathbf{x})u_{tt}(\mathbf{x},t) - \nabla\cdot(a(\mathbf{x})\nabla u(\mathbf{x},t)) &= 0 && (\mathbf{x},t)\in\Omega\times(0,T),\\
u(\mathbf{x},0) &= u_0(\mathbf{x}) && \mathbf{x}\in\Omega,\\
u_t(\mathbf{x},0) &= v_0(\mathbf{x}) && \mathbf{x}\in\Omega,\\
u(\mathbf{x},t) &= g_D(\mathbf{x},t) && \mathbf{x}\in\Gamma_D\times(0,T), \Gamma_D\subset\partial\Omega,\\
\mathbf{n}\cdot a(\mathbf{x})\nabla u(\mathbf{x},t) &= g_N(\mathbf{x},t) && \mathbf{x}\in\Gamma_N\times(0,T), \Gamma_N=\partial\Omega-\Gamma_D,
\end{aligned} \tag{1.1}$$

und ihre Anwendung auf den Energietransport in der unteren solaren Atmosphäre untersucht werden. Dieser soll dabei durch akustische Wellen simuliert werden, die sich in Form der obigen Gleichungen als Grenzfall kleiner Störungen aus den EULER-Gleichungen der Gasdynamik ergeben; in diesem Fall ist u die Abweichung des Drucks vom Gleichgewichtswert, ρ die Dichte und a die inverse Kompressibilität des Mediums. Diese Näherung ist nicht besonders gut, da sie die wesentlichen, nichtlinearen Effekte der Schallausbreitung in dünnen, ionisierten Gasen außer acht läßt. Es ist aber ein erster Schritt in Richtung zu mehrdimensionalen Rechnungen, die es erlauben, Mechanismen der Ausbreitung in Medien mit variablen Ausbreitungskoeffizienten zu untersuchen, insbesondere Reflexion an Grenzschichten und Refraktion; diese Effekte sind in den bisher durchgeführten eindimensionalen Rechnungen nicht darstellbar, so daß es wichtig ist, ihren Einfluß abzuschätzen.

Von der numerischen Seite her soll untersucht werden, ob adaptive Techniken im Kontext der Wellengleichung einsetzbar sind und ob es möglich ist, a-posteriori Fehlerschätztechniken anzuwenden. Aufgrund der bereits in zwei Raum- und einer Zeitdimension sehr hohen Anforderungen an Rechenzeit und Speicherbedarf erscheint es unabdingbar, adaptive Techniken zu verwenden, die im Gegensatz zu gleichförmigen oder blockweise gleichförmigen Gittern die Rechenzellen dort plazieren, wo der Fehler erzeugt wird. Die Berücksichtigung der Herkunft des Fehlers unterscheidet den vorgestellten Ansatz damit insbesondere von bisherigen Ansätzen zur Gittersteuerung, die nur dort verfeinerten, wo der Fehler groß ist. Die Notwendigkeit adaptiver Techniken gilt umso mehr bei Rechnungen mit drei Raumdimensionen oder wenn eine hohe Genauigkeit gefordert ist, da die herkömmlichen Techniken dieses nicht mehr zu leisten vermögen. Es wird gezeigt, daß adaptive Gitterverfeinerung in der Lage ist, die benötigten Ressourcen erheblich zu senken und aufgrund deutlich geringeren Speicherbedarfs einige Rechnungen erst möglich zu machen.

Bezüglich der Gitterverfeinerung soll insbesondere ein Ansatz untersucht werden, der eine Schätzung des Fehler bei der Auswertung eines beliebigen Funktionals verwendet. In der Praxis ist man selten am Fehler in einer abstrakten Norm, beispielsweise der L^2- oder der Energienorm, interessiert, sondern an der Auswertung eines Funktionals der Lösung, zum Beispiel einem Punktwert, Linienintegralen oder gewichteten Integralen. Es ist klar, daß ein optimales Rechengitter auf diese Auswertegröße abgestimmt sein muß, wobei das Ziel ist, bei gleichzeitiger Minimierung der Anzahl an Freiheitsgraden den Fehler bei der Auswertung des Funktionals zu minimieren. Dies kann bedeuten, von einer durch das Gebiet laufenden Welle nur den Teil gut aufzulösen, der

schließlich auch zum zu berechnenden Funktionalwert beiträgt, während die Teile der Welle, die in andere Richtungen laufen, nicht aufgelöst werden müssen. Es wird ein Ansatz vorgestellt, der die Gitterverfeinerung automatisch, d. h. ohne Interaktion des Benutzers, nach diesem Kriterium durchführt.

Im Gegensatz zu den meisten Rechnungen der Vergangenheit erfolgt die Diskretisierung nicht durch ein Differenzen- oder Charakteristikenverfahren mit expliziter Zeitdiskretisierung, sondern mittels Finiten Elementen und impliziten Zeitschrittverfahren. Finite Elemente erlauben gegenüber Differenzenverfahren wesentlich größere Freiheiten bei der Gestaltung der Rechengitter, was unabdingbar für die Verwendung adaptiver Verfahren ist; darüberhinaus lassen sich a-posteriori-Fehlerschätzer herleiten, die Informationen über den Beitrag einer Zelle zum Gesamtfehler liefern, so daß eine gezielte Verfeinerung einzelner Zellen möglich ist. Gegenüber Charakteristikenverfahren zeichnet sie aus, daß die Herangehensweise im wesentlichen dimensionsunabhängig ist; ihre Umsetzung bietet für zwei oder mehr Raumdimensionen keine wesentlichen zusätzlichen mathematischen Schwierigkeiten gegenüber einer Raumdimension. Charakteristikenverfahren geben dagegen die kontinuierliche Lösung der verwendeten Gleichung nahezu exakt im Diskreten wieder, sind allerdings nur sehr schwer auf höhere Raumdimensionen verallgemeinerbar und Abschätzungen des Fehlers sind schwierig (vgl. [26]). Die verwendeten Finiten Elemente sind bilineare bis biquartische konforme Elemente auf Viereckgittern.

Implizite Zeitschrittverfahren werden eingesetzt, da die Verwendendung expliziter Verfahren die Einhaltung einer Bedingung („CFL-Bedingung“) verlangt, die die Zeitschrittweite an die Größe der kleinsten Zellen bindet; da die Zellgröße variabel ist und bei adaptiven Verfahren sehr klein werden kann, würde dies eine sehr kleine Zeitschrittweite nach sich ziehen, was den Rechenaufwand enorm erhöhen würde. Im Rahmen dieser Arbeit wurden zur Zeitdiskretisierung das Crank-Nicolson- sowie das Fractional-Step-θ-Verfahren verwendet.

An der Entstehung einer Arbeit wie dieser sind mehr Menschen als nur der Author beteilgt. Ich bin allen, meinen Eltern, Freunden, Betreuern und Kollegen, die mich auf diesem oft nicht einfachen Weg begleitet, geführt und mir zur Seite gestanden haben, zu großem Dank verpflichtet!

Kapitel 2

Numerik der Wellengleichung

Zur Behandlung der Wellengleichung auf dem Computer ist es nötig, die in kontinuierlichen Koordinaten geschriebenen Gleichungen zu diskretisieren. Im Rahmen dieser Arbeit soll dabei folgende Diskretisierung untersucht werden:

- **zeitlich:** Die zeitliche Ableitung wird dadurch diskretisiert, daß wir die Gleichung (1.1) in ein System von partiellen Differentialgleichungen erster Ordnung in der Zeit überführen und diese mit einem der gängigen Verfahren (θ- bzw. Crank-Nicolson- und Fractional-Step-θ-Verfahren) diskretisieren.

- **räumlich:** Anschließend wird die räumliche Ableitung mit einem Finite-Elemente-Verfahren mit stückweise polynomialen Ansatzfunktionen behandelt. Als Polynome werden die Lagrange-Interpolationspolynome der Ordnungen eins bis vier betrachtet.

Diese Reihenfolge der Diskretisierung wird gemeinhin als Rothe-Methode bezeichnet, im Gegensatz zur sogenannten Linienmethode, bei der zuerst die Ortsdiskretisierung durchgeführt wird und das dadurch entstehende System gekoppelter Differentialgleichungen in der Zeit mit einem der bekannten Verfahren zur Lösung gewöhnlicher Differentialgleichungssysteme behandelt wird.

Für gleichbleibende Gitter sind beide Methoden im wesentlichen gleich und liefern im allgemeinen vergleichbare Ergebnisse. Lediglich die Analyse des Fehlers unterscheidet sich wesentlich. Die Rothe-Methode hat aber entscheidende Vorteile bei der Verwendung ortsadaptiver Gitter, da in diesen Fällen Anzahl und Ort der Freiheitsgrade von Zeitschritt zu Zeitschritt variieren, weshalb sich die partielle Differentialgleichung nicht mehr ohne weiteres in ein System von gewöhnlichen Differentialgleichungen umschreiben läßt.

Das Problem, daß die Lösung zu einem Zeitschritt auf das Gitter des nächsten Zeitschritts transportiert werden muß, bleibt jedoch bestehen. Hier erweist sich die Verwendung hierarchisch aufgebauter Gitter als ausgesprochen vorteilhaft, da es nicht nötig ist, herauszufinden, in welcher Zelle des Gitters ein Punkt liegt, wie es nötig wäre, wenn die Gitter zu verschiedenen Zeitschritten vollständig unabhängig wären. Programme, die mit völlig unabhängigen Gittern arbeiten, verbringen oft einen Großteil der Zeit mit solchen Aufgaben (vgl. [29, Remark 2.2]). Der Transfer zwischen Gittern kann darüberhinaus beim Aufbau der rechten Seite der Gleichungen exakt durchgeführt werden (Abschnitt 4.2), d. h. es tritt kein Interpolationsfehler wie bei der Verwendung unabhängiger Gitter auf.

Um einen streng mathematischen Rahmen zu schaffen, wird die Gleichung nicht wie oben erst in der Zeit und dann im Ort diskretisiert, sondern in einem Schritt. Sofern die Zeitdiskretisierung mit einem Galerkin-Verfahren erfolgt, sind die beiden Herangehensweisen identisch. Da sich die Aufteilung in Zeit- und Ortsdiskretisierung für die Implementation auf einem Computer besser eignet, wird zuerst eine halbformale Herleitung der Diskretisierung in zwei Schritten und daran anschließend eine rigorose Ableitung der Volldiskretisierung gegeben.

In den folgenden Abschnitten werden die folgenden Notationen verwendet:

- $I_n = (t_{n-1}, t_n]$ sei das $n$te Zeitintervall der Zeitdiskretisierung und $k_n = t_n - t_{n-1}$ seine Länge;
- $I = (0, T]$ sei das gesamte interessierende Zeitintervall; offenbar ist $I = \bigcup_n I_n$;
- Die Lösungsfunktion sei $\mathbf{w} = (u, v)$. Die Lösung des in Abschnitt 2.4.2 eingeführten dualen Problems sei $\bar{\mathbf{w}} = (\bar{u}, \bar{v})$; der Überstrich kennzeichne dualen Größen;
- $\mathbf{t} = (\varphi, \psi)$ kennzeichnet im allgemeinen eine Testfunktion;
- (p, q) bezeichnet das L^2-Skalarprodukt $\int pq \; dx$; falls p und q Vektoren sind, sei das Vektorprodukt impliziert;
- $b(\mathbf{w}, \mathbf{t})$ sei die durch $b(\mathbf{w}, \mathbf{t}) = \left(\begin{pmatrix} 0 & -\rho \\ a(\mathbf{x})\nabla & 0 \end{pmatrix} \mathbf{w}, \begin{pmatrix} 1 & 0 \\ 0 & \nabla \end{pmatrix} \mathbf{t} \right)_{\Omega \times [0,T]} - (g_N, \psi)_{\Gamma_N \times [0,T]}$ definierte Bilinearform.

2.1 Zeitdiskretisierung

Die Umschreibung in ein System erster Ordnung in der Zeit geschieht folgendermaßen: wir gehen von

$$\rho(\mathbf{x}) u_{tt}(\mathbf{x}, t) - \nabla \cdot (a(\mathbf{x}, t) \nabla u(\mathbf{x}, t)) = 0 \tag{2.1}$$

aus ($a(\mathbf{x}, t)$ ist entweder eine strikt positive skalare Funktion oder eine positiv definite, symmetrische $d \times d$-Matrix, ρ ist eine strikt positive Funktion) und setzen $v = u_t$. Der Einfachheit halber sei angenommen, daß die Koeffizienten a und ρ nicht von der Zeit abhängen. Dies ist im Einklang mit der Fragestellung dieser Arbeit; die Erweiterung bezüglich der Methodik bei zeitlich variablen Koeffizienten ist in den meisten Fällen offensichtlich. Sei außerdem $\mathcal{A} = -\nabla \cdot (a\nabla)$ der positiv definite räumliche Operator, dann gilt

$$\begin{pmatrix} \rho u_t \\ \rho v_t \end{pmatrix} + \begin{pmatrix} 0 & -\rho \\ \mathcal{A} & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = 0. \tag{2.2}$$

Diese Gleichung läßt sich mit $\mathbf{w} = (u, v)$ und dem Operator $\mathcal{B} = \begin{pmatrix} 0 & -\rho \\ \mathcal{A} & 0 \end{pmatrix}$ verkürzt auch als Differentialgleichung im Hilbertraum schreiben:

$$\rho \mathbf{w}_t + \mathcal{B}\mathbf{w} = 0. \tag{2.3}$$

Die erste Gleichung wurde mit ρ multipliziert, da dann alle Massematrizen diesen Faktor enthalten. Darüberhinaus läßt sich so die numerische Stabilität erhöhen, da die einzelnen Terme von der gleichen Größenordnung sind; das ist insbesondere bei dem in Abschnitt 3 gerechneten Beispiel wichtig, wo die Koeffizienten um bis zu drei Zehnerpotenzen variieren.

Die Umschreibung in ein System hat gegenüber der direkten Lösung von (2.1) im wesentlichen zwei Vorteile:

- Zur Diskretisierung erster Zeitableitungen existiert eine Vielzahl von gut untersuchten Verfahren aus der Numerik der gewöhnlichen Differentialgleichungen. Dort werden höhere Ableitungen im allgemeinen durch Einführung zusätzlicher Variablen in ein System von Gleichungen umgeformt, so daß zur direkten Diskretisierung zweiter Ableitungen nur relativ wenige Verfahren bekannt sind.
- Gegenüber der direkten Diskretisierung (beispielsweise durch einen Ansatz ähnlich der Diskretisierung des LAPLACE-Operators) kann man eine bessere Konvergenzordnung für die zweite Variable v erwarten, analog zur Verwendung einer gemischten Diskretisierung im

Ort. Das hat Vorteile, da viele praktisch relevante Funktionale die Geschwindigkeit enthalten, beispielsweise die Energie

$$E = (\rho v, v)_\Omega + (a\nabla u, \nabla u)_\Omega,$$

oder der Energiefluß durch eine Kurve

$$\Phi = \int_C va\nabla u \cdot \mathbf{n}\, ds.$$

- Bei direkter Diskretisierung wird die Implementation schnell sehr unübersichtlich, da Daten über mehr als einen Zeitschritt transportiert werden müssen.

Dem steht als Nachteil allerdings gegenüber, daß die mit dem System assoziierte Bilinearform keine natürliche Norm induziert, so daß a priori Abschätzungen und Stabilitätsanalysen erschwert sind. Diese sind jedoch in der Literatur zu finden, vgl. beispielsweise [16] und die Referenzen darin.

Das Gleichungssystem (2.3) soll im folgenden mit einem der bekannten Verfahren in der Zeit diskretisiert werden. Dabei kennzeichnen in den folgenden Abschnitten obere Indizes (Superskripte) die Lösungen zu bestimmten Zeitpunkten, z. B. ist $\mathbf{W}^n \equiv \mathbf{W}(t_n)$, wobei $\mathbf{W}(t)$ die Lösung des semidiskreten Problems sei. Um die Notation übersichtlich zu halten, sind die Verfahren meist nur für den ersten Zeitschritt, das heißt für $\mathbf{W}^1$ und $\mathbf{W}^0$ angegeben; die Umsetzung für allgemeine $\mathbf{W}^{n+1}$ und $\mathbf{W}^n$ ist offensichtlich. Bei Gleichungen, die sich ausschließlich auf den ersten Zeitschritt beziehen, ist dies explizit angemerkt.

Wie üblich bezeichne k die Zeitschrittweite. Im allgemeinen wird $k = k_n$, d. h. variabel zu wählen sein. Da nur Einschrittverfahren beschrieben werden, sei der Einfachheit halber auf den Index verzichtet.

2.1.1 Das θ-Verfahren

Das θ-Verfahren auf (2.3) angewandt läßt sich wie folgt schreiben:

$$\rho\left(\frac{\mathbf{W}^1 - \mathbf{W}^0}{k}\right) + \theta\mathcal{B}\mathbf{W}^1 + (1-\theta)\mathcal{B}\mathbf{W}^0 = 0. \tag{2.4}$$

Durch Umschreiben[1] läßt sich die gleichzeitige Lösung zweier gekoppelter Lösungen vermeiden und man gelangt zu zwei Teilschritten:

$$\begin{aligned}(\rho + k^2\theta^2\mathcal{A})u^1 &= \rho u^0 + k\rho v^0 - k^2\theta(1-\theta)\mathcal{A}u^0\\ \rho v^1 &= \rho v^0 - k\theta\mathcal{A}u^1 - k(1-\theta)\mathcal{A}u^0\end{aligned} \tag{2.5}$$

Durch die Wahl von θ lassen sich aus (2.4) verschiedene bekannte Verfahren gewinnen:

- $\theta = 0$: Bei dieser Wahl ist (2.4) die Diskretisierung der kontinuierlichen Gleichung durch das explizite EULER-Verfahren. Als explizites Verfahren hat es den Vorteil, daß keine Differentialgleichung im Ort gelöst werden muß, das heißt $\mathbf{W}^1$ hängt explizit von $\mathbf{W}^0$ ab und kann durch Matrix-Vektor-Multiplikationen gewonnen werden. Der wesentliche Nachteil des Verfahrens liegt in der Kopplung der Zeitschrittweite an die Größe der kleinsten Zellen, die darin begründet liegt, daß aufgrund des expliziten Charakters Information pro Zeitschritt nur über höchstens eine Zelle transportiert werden kann; soll die Ausbreitungsgeschwindigkeit richtig wiedergegeben werden, so darf die Zeitschrittweite nicht zu groß gewählt werden.[2]

[1] Die Umformung erscheint auf den ersten Blick offensichtlich. Ihre Richtigkeit läßt sich aber nur in der schwachen Formulierung zeigen, da nur dann die unterschiedlichen Räume, aus denen u und v kommen, korrekt behandelt werden können; vgl. Abschnitt 2.3.3.

[2] Im Fall gleichförmiger Gitter gilt darüberhinaus die COURANT-FRIEDRICHS-LEWY-Bedingung (vgl. [13]), die den Verlust der Stabilität des Verfahrens beschreibt, falls für Zeitschritt k und Gitterweite h die Beziehung $ck \leq Ch$ verletzt ist, wobei c die Ausbreitungsgeschwindigkeit sei. Die Konstante C ist von der Ortsdiskretisierung abhängig

Letzteres ist eine Bedingung, die den Einsatz im Umfeld adaptiv verfeinerter Gitter ausschließt, da die kleinsten Zellen sehr klein werden können, wenn stark lokalisierte Eigenschaften der Lösung gut aufgelöst werden sollen.

Das explizite EULER-Verfahren ist von erster Ordnung.

- $\theta = 1$: In diesem Fall entsteht das implizite EULER-Verfahren. Es ist stark A-stabil und ebenfalls von erster Ordnung. Falls für den Koeffizienten $a(\mathbf{x}, t) = a(\mathbf{x})$ gilt, und da wir die homogene Wellengleichung betrachten, ist dieses Verfahren identisch mit dem unstetigen GALERKIN-Verfahren nullter Ordnung, DG(0).

  Das implizite EULER-Verfahren hat den bedeutenden Nachteil, daß es stark dissipativ ist, d. h. daß es die Lösung glättet und damit zu einem unphysikalischen Verlust an Energie im System führt; es kommt damit nicht als Kandidat für eine sinnvolle Implementation infrage.

- $\theta = \frac{1}{2}$: Mit dieser Wahl ergibt sich das CRANK-NICOLSON-Verfahren. Es ist von zweiter Ordnung und A-stabil, aber leider nicht stark A-stabil, was sich im Auftreten von Oszillationen bei Störungen zeigen kann. Typische Störungen sind beispielsweise eine starke Veränderung des Gitters; da die dadurch ausgelösten Oszillationen aufgrund der fehlenden Glättungseigenschaften der Wellengleichung und des Zeitschrittverfahrens nicht gedämpft werden, akkumuliert sich im Laufe der Rechnung ein „Untergrundrauschen“, das bei der Auswertung schwacher Signale stören kann. Aufgrund der Elliptizität der in jedem Zeitschritt zu lösenden Gleichungen müssen bei der Störung durch Gitteränderung Ort der Störung und Ort der Oszillationen nicht identisch miteinander sein.

  Vorteile des CRANK-NICOLSON-Verfahren sind seine höhere Ordnung und vor allem die fehlende Dissipativität, die die Energie auf ungestörten Gittern exakt erhält. Das Verfahren läßt sich als PETROV-GALERKIN-Verfahren schreiben, was in Abschnitt 2.4 zur Herleitung von Fehlerschätzern verwendet werden wird.

Aus (2.4) läßt sich auf die Energie $E^1 = \frac{1}{2}\left((a\nabla u^1, \nabla u^1)_\Omega + (\rho v^1, v^1)_\Omega\right)$ zum neuen Zeitschritt zumindest teilweise in Abhängigkeit von der Energie E^0 im letzten Zeitschritt berechnen. Nach trivialer Rechnung erhält man

$$E^1 = E^0 + \frac{k}{2}(1-2\theta)\left[\left(\nabla v^0, a\nabla u^1\right)_\Omega - \left(\nabla v^1, a\nabla u^0\right)_\Omega\right],$$

was erklärt, weshalb das CRANK-NICOLSON-Verfahren die Energie erhält.

2.1.2 Das Fractional-Step-θ-Verfahren

Für die Gleichung (2.3) schreibt sich das Fractional-Step-θ-Verfahren (im folgenden mit FS-θ-Verfahren abgekürzt) als Dreischrittverfahren:

$$\begin{aligned}
\rho\left(\frac{\mathbf{W}^\theta - \mathbf{W}^0}{k}\right) + \alpha\theta\mathcal{B}\mathbf{W}^\theta + \beta\theta\mathcal{B}\mathbf{W}^0 &= 0,\\
\rho\left(\frac{\mathbf{W}^{1-\theta} - \mathbf{W}^\theta}{k}\right) + \beta\theta'\mathcal{B}\mathbf{W}^{1-\theta} + \alpha\theta'\mathcal{B}\mathbf{W}^\theta &= 0,\\
\rho\left(\frac{\mathbf{W}^1 - \mathbf{W}^{1-\theta}}{k}\right) + \alpha\theta\mathcal{B}\mathbf{W}^1 + \beta\theta\mathcal{B}\mathbf{W}^{1-\theta} &= 0,
\end{aligned} \tag{2.6}$$

und liegt in der Größenordnung von 1.

Die CFL-Bedingung wird über die Größe der Eigenwerte der Systemmatrix hergeleitet. Für gleichförmig verfeinerte Gitter ist der Zusammenhang zwischen Gitterweite h und den Eigenwerten bekannt, nicht jedoch für lokal verfeinerte Gitter. Ein im Verhältnis zu den kleinsten Zellen zu groß gewählter Zeitschritt kann daher unter Umständen trotzdem ein stabiles Verfahren ergeben, das beschriebene Problem mit der numerischen Ausbreitungsgeschwindigkeit bleibt jedoch bestehen.

mit $\theta = 1 - \frac{1}{2}\sqrt{2}$, $\theta' = 1 - 2\theta$, $\frac{1}{2} \leq \alpha < 1$ und $\beta = 1 - \alpha$. Betrachtet man für einen Moment wieder die Ortsabhängigkeit mit, so gelten für u die Randbedingungen

$$\begin{aligned} u^0(\mathbf{x}) &= g_D(\mathbf{x}, 0), \\ u^\theta(\mathbf{x}) &= g_D(\mathbf{x}, t_1 - \theta k), \\ u^{1-\theta}(\mathbf{x}) &= g_D(\mathbf{x}, t_1 - \theta k), \\ u^1(\mathbf{x}) &= g_D(\mathbf{x}, t_1), \end{aligned} \tag{2.7}$$

für $\mathbf{x} \in \partial\Omega$. Für NEUMANN-Randbedingungen gilt analoges. Die ungewöhnlich erscheinende Auswahl der Zeitpunkte, zu denen die Randfunktion ausgewertet wird, ergibt sich formal aus der Herleitung des Verfahrens; sie hängt damit zusammen, daß u^θ und $u^{1-\theta}$ keine Approximationen zweiter Ordnung zu $u(\theta k)$ und $u(t_1 - \theta k)$ sind.

Durch Umsortieren der beiden Gleichungen analog zu oben läßt sich das Fractional-Step-θ-Verfahren (2.6) wie folgt schreiben:

$$\begin{aligned} (\rho + \alpha^2\theta^2 k^2 \mathcal{A}) u^\theta &= \rho u^0 + \theta k \rho v^0 - \alpha\beta\theta^2 k^2 \mathcal{A} u^0 \\ \rho v^\theta &= \rho v^0 - \alpha\theta k \mathcal{A} u^\theta - \beta\theta k \mathcal{A} u^0 \\ (\rho + \beta^2 \theta'^2 \mathcal{A}) u^{1-\theta} &= \rho u^\theta + \theta' k \rho v^\theta - \alpha\beta\theta'^2 \mathcal{A} u^\theta \\ \rho v^{1-\theta} &= \rho v^\theta - \beta\theta' k \mathcal{A} u^{1-\theta} - \alpha\theta' k \mathcal{A} u^\theta \\ (\rho + \alpha^2\theta^2 k^2 \mathcal{A}) u^1 &= \rho u^{1-\theta} + \theta k \rho v^{1-\theta} - \alpha\beta\theta^2 k^2 \mathcal{A} u^{1-\theta} \\ \rho v^1 &= \rho v^{1-\theta} - \alpha\theta k \mathcal{A} u^1 - \beta\theta k \mathcal{A} u^{1-\theta} \end{aligned}$$

Im allgemeinen wählt man α so, daß $\alpha\theta = \beta\theta'$, d. h. $\alpha = \frac{1-2\theta}{1-\theta}$, da dann die Systemmatrizen der drei Teilschritte identisch sind. Diese Wahl von α verträgt sich mit der obengenannten Einschränkung.

Das Verfahren ist von zweiter Ordnung und stark A-stabil. Letztere Eigenschaft zusammen mit der Tatsache, daß es nur sehr geringe Dissipativität besitzt, machen es zu einem nahezu idealen Kandidaten für das untersuchte Problem. Dagegen stehen jedoch die aufwendige Analyse sowie die Tatsache, daß es sich nicht als GALERKIN-Verfahren schreiben läßt, was für die Fehlerschätzung nötig wäre.

Das FS-θ-Verfahren wurde ursprünglich als Operator-Splitting-Schema für die Lösung der NAVIER-STOKES-Gleichungen entwickelt [8]. Analysen des Verfahrens finden sich in [28, 24], Anwendungen in [27] und den darin aufgeführten Referenzen.

2.1.3 Vergleich der Zeitschrittverfahren

Zur Wahl des Zeitschrittverfahrens wurden einige Tests durchgeführt, um die verschiedene Verfahren zu vergleichen.

Beispiel 1. Zuerst wurden auf einem regulär verfeinerten Gitter die Anfangs- und Randwerte folgendermaßen gewählt:

$$\begin{aligned} u(\mathbf{x}, 0) &= \sin(2\pi x)\ \sin(2\pi y) && \mathbf{x} \in \Omega = (0,3) \times (0,1), \\ v(\mathbf{x}, 0) &= 0 && \mathbf{x} \in \Omega, \\ u(\mathbf{x}, t) &= 0 && (\mathbf{x}, t) \in \partial\Omega \times (0, T). \end{aligned}$$

Die Koeffizienten ρ und a waren dabei jeweils konstant eins. Die Lösung ist die $(6,2)$-Eigenschwingung des Gebietes,

$$u(\mathbf{x}, t) = \sin(2\sqrt{2}\pi t)\ \sin(2\pi x)\ \sin(2\pi y).$$

Die Gesamtenergie sollte den konstanten Wert $E = 6\pi^2$ haben.[3] Im allgemeinen ist die Wahl von Eigenmoden nicht repräsentativ für die Auswahl von Verfahren, da sie zu viele spezielle Eigenschaften haben; insbesondere kann die projizierte Lösung Eigenvektor der Systemmatrizen sein.

[3] Der genaue Wert der Energie spielt hier keine wirkliche Rolle, da man eigentlich nur mit der Energie in der auf das Gitter projizierten echten Lösung vergleichen darf.

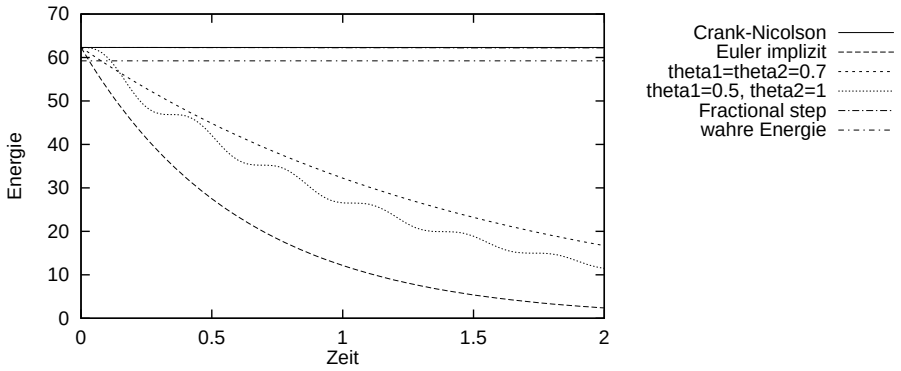

Abbildung 2.1: *Vergleich der Energie bei verschiedenen Zeitschrittverfahren bei regulär verfeinertem Gitter und glatter Lösung (Beispiel 1).*

Da es in diesem ersten Beispiel aber nur darum geht, einige der Verfahren auszusortieren, ist die Wahl zu rechtfertigen. Der Vergleich der verbleibenden Verfahren miteinander erfolgt an einem praxisrelevanteren Beispiel.

Die Lösung ist beliebig glatt und zeigt keine örtlich veränderlichen Eigenschaften wie laufende Wellen. Die Zeitschrittweite betrug $k = 0.02$ (für das FS-θ-Verfahren wurde der Vergleichbarkeit halber $k = 0.06$ gewählt), das Gebiet wurde in 192 quadratische Zellen mit $h = 1/8$ aufgeteilt. Die Grobheit des Gitters reicht sicher nicht aus, um feinere Einzelheiten der Lösung aufzulösen, genügt aber um einige der Verfahren schon auszusortieren.

In Abbildung 2.1 ist der Verlauf der Energie in der numerischen Lösung für verschiedene Verfahren über knapp drei Perioden der Eigenschwingung aufgetragen. Man erkennt die zu erwartende Abnahme der Energie beim impliziten EULER-Verfahren aufgrund dessen Dämpfungseigenschaften und daß auch die anderen aus dem θ-Verfahren hergeleiteten Methoden außer dem CRANK-NICOLSON-Verfahren diese Eigenschaft besitzen. Die Wellenlinie entspricht einem Verfahren, bei dem die beiden Gleichungen (2.4) mit unterschiedlich gewählten Parametern θ_1 und θ_2 diskretisiert wurden; dieses Verfahren dämpft kinetische und elastische Energie unterschiedlich stark, wodurch die Schlangenlinie zustandekommt. Beim CRANK-NICOLSON-Verfahren bleibt die Energie auf mindestens sechs Stellen konstant, wenn das Gitter konstant bleibt. Die Konstanz der Energie ist unabhängig vom Verhältnis von Orts- zu Zeitgitterweite, das heißt es existiert keine störende CFL-Bedingung. Die Zeitschrittweite kann daher an der Zeitskala der aufzulösenden Phänomene orientiert werden. Die Kurve des FS-θ-Verfahrens liegt praktisch auf der des CRANK-NICOLSON-Verfahrens, allerdings verliert das Verfahren hier in jedem Zeitschritt etwa einen Anteil von $6 \cdot 10^{-5}$ der Energie. Da eine typische Simulation etwa 200 Zeitschritte benötigt, ist dieser Fehlerbeitrag im allgemeinen tolerierbar.

Beispiel 2. In einem zweiten Beispiel wurde das Problem mit folgenden Anfangs- und Randwerten gestellt:

$$\begin{aligned} u(\mathbf{x},0) &= \begin{cases} \frac{|x|}{s} & \text{für } |\mathbf{x}| < s \\ 2 - \frac{|x|}{s} & \text{für } s \le |\mathbf{x}| < 2s \\ 0 & \text{sonst} \end{cases} && \mathbf{x} \in \Omega = [-1,1]^2, \\ v(\mathbf{x},0) &= 0 && \mathbf{x} \in \Omega, \\ u(\mathbf{x},t) &= 0 && (\mathbf{x},t) \in \partial\Omega \times [0,T]. \end{aligned} \tag{2.8}$$

Die Lösung ist eine kreisförmige, aus dem Zentrum herauslaufende, nicht stetig ableitbare Welle. Im Beispiel wurde $s = 0.1$, der Koeffizient a unstetig zu $a = 4$ für $y > \frac{1}{3}$, $a = 1$ sonst, und $\rho = 1$ gewählt.

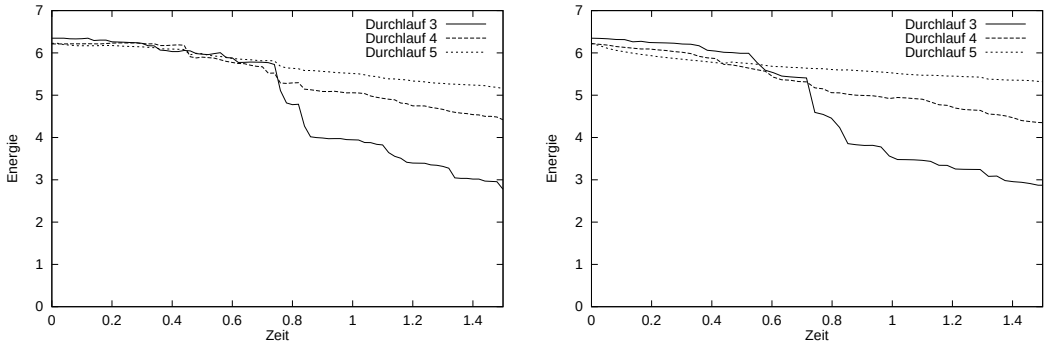

Abbildung 2.2: *Vergleich der Energie beim* Crank-Nicolson*-Verfahren (links) und beim Fractional-Step-θ-Verfahren (rechts) auf adaptiv verfeinerten Gittern. Koeffizient und Lösung sind nicht glatt (Beispiel 2).*

In Abbildung 2.2 sind Crank-Nicolson- und FS-θ-Verfahren einander gegenübergestellt, wobei die Zeitschrittweiten auf $k = 0.02$ bzw. $k = 0.03$ gesetzt wurde. Die Rechnung wurde auf adaptiv verfeinerten Gittern mit jeweils in etwa gleichen Zellzahlen bei den beiden Verfahren durchgeführt, wobei mehrere Durchläufe[4] gemacht wurden. Für den Vergleich ist im wesentlichen nur der letzte Durchlauf wichtig; die vorhergehenden geben jedoch Aufschluß über die Stabilitätseigenschaften der Zeitschrittverfahren. Als Fehlerindikator wurde ein einfacher Energieschätzer nach Kelly, Gago, Zienkiewicz und Babuška der Form

$$\eta_K^2 = Ch\|[\mathbf{n} \cdot a(\mathbf{x})\nabla u_h]\|_{\partial K}^2 \tag{2.9}$$

verwendet, vgl. [23]. Für jede Zelle wird dabei der Sprung der Konormalenableitung $[\mathbf{n} \cdot a(\mathbf{x})\nabla u_h]$ zur Nachbarzelle über den Rand integriert. Eine kurze Begründung für diesen Fehlerindikator wird in Abschnitt 2.4.4 gegeben.

Aus den schon recht glatten Kurven des letzten Durchlaufs erkennt man, daß das FS-θ-Verfahrens am Anfang zu einem stärkeren Energieverlust als das Crank-Nicolson-Verfahren führt. Die Vermutung, daß die dissipativen Eigenschaften des Verfahrens eine sichtbar stärkere Glättung von Unstetigkeiten der Lösung oder ihres Gradienten konnte nicht bestätigt werden. Die letztliche Ursache ist unbekannt.

Trotz der Tatsache, daß pro Zeitschritt drei Gleichungssysteme gelöst werden müssen, dieser Aufwand aber der pro Zeitschritt nur einmal nötigen Projektion der Lösungen des vorigen Zeitschritts sowie der Fehlerschätzung gegenübergestellt werden muß, liegt der numerische Aufwand bei den Zeitschrittweiten 0.02 bzw. 0.03 in etwa der gleichen Größenordnung.

In der Abbildung ist andererseits die bessere Stabilität des FS-θ-Verfahrens zu sehen, die sich im glatteren Verlauf der Kurve und durch die weitgehend fehlenden Sprünge in den späteren Durchläufen zeigt. Der Effekt ist besonders deutlich in der jeweils mittleren der drei Kurven zu sehen. Diese Sprünge werden durch die Störung hervorgerufen, die die Veränderung der Gitter darstellt.

Im Prinzip wäre das Fractional-Step-θ-Verfahren ein sehr guter Kandidat für das zu verwendende Zeitschrittverfahren. Es hat allerdings den Nachteil, daß es sich im Gegensatz zum Crank-Nicolson-Verfahren nicht als Galerkin-Verfahren schreiben läßt. Damit fällt auch die notwendige Galerkin-Orthogonalität weg und die Abschätzung des Fehlers mittels eines dualen Problems ist nicht mehr möglich. Da dieser Punkt das Verfahren für die Zwecke der Fehlerschätzung im wesentlichen ausschließt, wird im folgenden nur noch das Crank-Nicolson-Verfahren verwendet. Beide Verfahren zeigen bei zunehmender Gitterverfeinerung, d. h. in den späteren Durchläufen,

[4] Zur Bedeutung des Wortes „Durchlauf“: Die Verfeinerung erfolgte so, daß in einem ersten Durchlauf alle Zeitschritte nacheinander auf dem Grobgitter gerechnet wurde, im nächsten Durchlauf alle Zeitschritte auf einem einmal verfeinerten Gitter, usw. Details dazu sind in Abschnitt 4.5 zu finden.

bessere Stabilitätseigenschaften und Konvergenz gegen eine konstante Energie, so daß die Wahl des CRANK-NICOLSON-Verfahrens trotz der guten Eigenschaften des FS-θ-Verfahrens vertretbar ist.

2.2 Ortsdiskretisierung

Durch die untersuchten Zeitschrittverfahren entstanden jeweils zwei Gleichungen der Form

$$(\rho + \sigma\mathcal{A})u^1 = \alpha\rho u^0 + \beta\rho v^0 + \gamma\mathcal{A}u^0, \tag{2.10}$$

$$\rho v^1 = \rho v^0 + \nu\mathcal{A}u^1 + \lambda\mathcal{A}u^0, \tag{2.11}$$

wobei σ, α, β, γ, ν und λ vom Verfahren und von der Zeitschrittweite abhängige Konstanten sind, deren Wert hier außer dem positiven Vorzeichen von $\sigma \propto k^2$ keine Rolle spielt. Es gelte wieder $\mathcal{A} = -\nabla \cdot a(\mathbf{x})\nabla$. Die örtliche Diskretisierung dieser Gleichungen wird in der schwachen Formulierung durchgeführt, die man durch Multiplikation der Gleichungen mit einer Testfunktion ψ und Integration über das Gebiet Ω erhält. Betrachtet man Gleichung (2.10), so lautet das Problem nun:

Finde $u^1 \in W$, so daß für alle $\psi \in W$ gilt:

$$(\rho u^1, \psi) + \sigma a(u^1, \psi) = \alpha(\rho u^0, \psi) + \beta(\rho v^0, \psi) + \gamma a(u^0, \psi). \tag{2.12}$$

Die Frage der richtigen Funktionenräume ist in Kapitel 2.3 diskutiert. Es sei $(\rho u, v) \equiv \int_\Omega \rho uv \, dx$ und $a(u, v) \equiv \int_\Omega a\nabla u \nabla v \, dx$ der Term, der durch partielle Integration von $(\mathcal{A}u, v)$ entsteht. Wegen der geforderten Eigenschaften von $a(\mathbf{x})$ und $\rho(\mathbf{x})$ sind $a(\cdot,\cdot)$ und $(\rho\cdot,\cdot)$ symmetrische, positiv definite Bilinearformen. Sind inhomogene NEUMANN-Randwerte vorgeschrieben, muß die rechte Seite noch um einen Randintegralterm ergänzt werden.

Lösungen von (2.10) sind selbstverständlich auch Lösungen von (2.12), jedoch nicht notwendigerweise umgekehrt, da die variationelle Formulierung (2.12) geringere Differenzierbarkeit als (2.10) von der Lösung fordert (dafür wurde die partielle Integration durchgeführt). Die Wellengleichung garantiert nur eine Lösung, die genauso glatt ist wie Rand- und Anfangswertdaten, im allgemeinen also $u \in L^2$ mit zusätzlichen Annahmen über die Existenz einer Spur auf dem Rand und von Anfangswerten; im Rahmen dieser Arbeit sei aber in den meisten Fällen $u(\cdot, t) \in W \subset H^1$ angenommen, wobei H^1 der Raum der Funktionen sei, deren (schwacher) Gradient noch quadratintegrabel ist.

Die Diskretisierung im Ort wird nun so durchgeführt, daß wir die Lösung nicht mehr im unendlichdimensionalen Raum W suchen, sondern in einem endlichdimensionalen Raum W_h und die Testfunktionen ebenfalls aus einem endlichdimensionalen Raum T_h wählen. Die Dimension N_h der beiden Räume muß übereinstimmen. Das volldiskrete Problem lautet damit:

Finde $u_h^1 \in W_h$, so daß für alle $\psi_h \in T_h$ gilt:

$$(\rho u_h^1, \psi_h) + \sigma a(u_h^1, \psi_h) = \alpha(\rho u_h^0, \psi_h) + \beta(\rho v_h^0, \psi_h) + \gamma a(u_h^0, \psi_h). \tag{2.13}$$

Für die Ortsdiskretisierung werden hier nur Ansätze mit $W_h = T_h$ (RITZ-GALERKIN-Verfahren) untersucht; Verfahren mit $W_h \neq T_h$ heißen PETROV-GALERKIN-Verfahren und werden in dieser Arbeit für die Zeitdiskretisierung mit dem CRANK-NICOLSON-Verfahren verwendet.

Wegen der Endlichdimensionalität der Räume kann man eine endliche Basis $\{\varphi_i(x)\}_1^{N_h}$ von W_h definieren und die Lösung als $u_h^1 = \sum_i u_i^1 \varphi_i(x)$ darstellen, wobei die $\{u_i^1\}_1^{N_h}$ den Lösungsvektor bilden. Testet man nun mit einer der Funktionen $\psi_h = \varphi_j$, so erhält man die Gleichung

$$\sum_i u_i^1 (\rho\varphi_i, \varphi_j) + \sigma \sum_i u_i^1 a(\varphi_i, \varphi_j) = f_j,$$

mit der aus (2.13) entstandener rechter Seite f_j. Da es N_h linear unabhängige Testfunktionen gibt, gibt es auch genausoviele Gleichungen, so daß (2.13) äquivalent zum linearen Gleichungssystem

$$(M + \sigma A)\mathbf{u} = \mathbf{f}$$

ist, mit den Matrizen $M_{ij} = (\rho\varphi_i, \varphi_j)$ und $A_{ij} = a(\varphi_i, \varphi_j)$ und den Vektoren $\mathbf{u} = \{u_i\}$ und $\mathbf{f} = \{f_j\}$. Die Lösung dieses Gleichungssystems erfolgt mit einem der üblichen Verfahren der linearen Algebra, hier mit dem CG-Verfahren, das mit dem JACOBI- oder dem SSOR-Verfahren vorkonditioniert wird. M und A sind aufgrund der Symmetrie des L^2-Skalarprodukts und der Bilinearform $a(\cdot,\cdot)$ und wegen $W_h = T_h$ ebenfalls symmetrische Matrizen, so daß die Lösung abgesehen von der Größe des Problems keine Schwierigkeiten bereitet. Mehrgitterverfahren sollten in der Lage sein, die Gleichungssyteme erheblich effizienter zu lösen, konnten aber aus Zeitgründen im Rahmen dieser Arbeit nicht implementiert werden. Der Aufbau der Matrizen M und A erfolgt durch numerische Quadratur. Die Aufstellung der rechten Seite ist in Abschnitt 4.2 beschrieben.

Die Diskretisierung der Gleichung (2.11) erfolgt ganz analog und liefert das Gleichungssystem

$$M\mathbf{v}^1 = \mathbf{g},$$

mit dem Koeffizientenvektor $\mathbf{v}^1 = \{v_i^1\}$ und der aus (2.11) entstandenen rechten Seite $\mathbf{g} = \{g_j\}$.

Im Rahmen dieser Arbeit wird für den endlichdimensionalen Raum W_h der Raum der global stetigen Funktionen gewählt, die sich auf jedem Element K als Polynom schreiben lassen, wobei in jedem Summand die Potenz der einzelnen unabhängigen Variablen höchstens gleich r sei. Es werden Polynomgrade bis $r = 4$ verwendet. Das Gebiet wird in Viereckselemente aufgeteilt, die wegen der besseren Approximationseigenschaften gegenüber Dreiecken gewählt wurden und weil sie sich leichter auf drei Raumdimensionen verallgemeinern lassen.

In der Praxis werden immer Randwerte für u^1 und v^1 vorgegeben. Diese werden erst nach der Aufstellung der linearen Gleichungssysteme gemäß dem in Abschnitt 4.3 beschriebenen Verfahren behandelt.

2.3 Volldiskretisierung in einem Schritt

In diesem Abschnitt soll eine rigorose Ableitung der Volldiskretisierung gegeben werden, wobei keine Reihenfolge bei der Diskretisierung der Orts- und Zeitrichtungen mehr angenommen wird.

Zu lösen sei wieder das System (2.2)

$$\begin{aligned}
\begin{pmatrix}\rho u_t\\ \rho v_t\end{pmatrix} + \begin{pmatrix}0 & -\rho\\ \mathcal{A} & 0\end{pmatrix}\begin{pmatrix}u\\ v\end{pmatrix} &= 0 && \mathbf{x}\in\Omega, t\in I,\\
u(\cdot,0) &= u^0(\cdot), &&\\
v(\cdot,0) &= v^0(\cdot), &&\\
u(\mathbf{x},t) &= g_D(\mathbf{x},t) && \mathbf{x}\in\Gamma_D\subset\partial\Omega, t\in I,\\
\mathbf{n}\cdot a\nabla u(\mathbf{x},t) &= g_N(\mathbf{x},t) && \mathbf{x}\in\Gamma_N = \partial\Omega\backslash\Gamma_D, t\in I,
\end{aligned}$$

mit $\mathcal{A} = -\nabla\cdot(a(\mathbf{x})\nabla\cdot)$. Für die exakte Lösung $\mathbf{w} = (u, v)$ gelte $\mathbf{w}\in W$. Die exakte Struktur der Lösungsräume ist kompliziert und sei hier nicht erörtert; wir nehmen $W \subset H^1(I, H^1(\Omega)) \times H^1(I, L^2(\Omega))$ an.[5] Da die Lösungen im allgemeinen genauso glatt sind wie Anfangs- und Randwerte, entspricht das in etwa der Annahme stetiger Randwerte und Anfangswerte aus $H^1(\Omega)\times L^2(\Omega)$, was mit der physikalischen Anschauung in den meisten Fällen übereinstimmt.

Wir überführen die Differentialgleichung in eine schwache Form durch Multiplikation mit einer Testfunktion $\mathbf{t} = (\varphi, \psi)\in W$, Integration über $\Sigma = \Omega\times I$ und partielle Integration:

$$(\rho\mathbf{w}_t, \mathbf{t})_\Sigma + \left(\begin{pmatrix}0 & -\rho\\ a\nabla & 0\end{pmatrix}\mathbf{w}, \begin{pmatrix}1 & 0\\ 0 & \nabla\end{pmatrix}\mathbf{t}\right)_\Sigma - (g_N, \psi)_{\Gamma_N\times I} = 0, \qquad \forall\mathbf{t}\in W. \tag{2.14}$$

[5] Die Schreibweise mit HILBERT-Räumen ist dem Problem unangemessen, da nicht klar wird, daß Informations- und damit Regularitätstransport entlang von Charakteristiken stattfindet; die Regularität ist entlang von Charakteristiken höher als senkrecht dazu. Dies mathematisch exakt auszudrücken ist aber kompliziert und nicht Thema dieser Arbeit.

Wir suchen eine Approximation $\mathbf{w}_h \in \mathcal{W}_h$ zu $\mathbf{w}$ in einem endlichdimensionalen Raum $\mathcal{W}_h \subset W$, die Lösung der folgenden Gleichung sein solle:

$$(\rho \mathbf{w}_{h,t}, \mathbf{t}_h)_\Sigma + \left(\begin{pmatrix} 0 & -\rho \\ a\nabla & 0 \end{pmatrix} \mathbf{w}_h, \begin{pmatrix} 1 & 0 \\ 0 & \nabla \end{pmatrix} \mathbf{t}_h \right)_\Sigma - (g_N, \psi_h)_{\Gamma_N \times I} = 0, \qquad \forall \mathbf{t}_h \in \mathcal{T}_h. \quad (2.15)$$

Der Testraum $\mathcal{T}_h$ hat dieselbe Dimension wie $\mathcal{W}_h$, muß aber nicht notwendigerweise mit ihm übereinstimmen.

Für die Diskretisierung mit dem CRANK-NICOLSON-Verfahren in der Zeit und mit LAGRANGE-Elementen im Ort wählen wir die Räume wie folgt. Sei dazu $I = \bigcup_n I_n, I_n = (t_{n-1}, t_n]$ eine Zerlegung des Zeitintervalls und $\mathbb{T}^n = \{K\}$ eine Zerlegung von Ω in Vierecke, die die üblichen Bedingungen (*shape regularity*) erfülle. Sei $\hat{Q}^r(K)$ der Raum der Funktionen, die auf dem Einheitselement $\hat{K} = [0,1]^d$ polynomial bis zur (Bi-, Tri-)Ordnung r sind, d. h. die sich gemäß

$$\hat{u} = \sum_{\max_i \alpha_i \le r} c_\alpha \mathbf{x}^\alpha = \begin{cases} c_0 + c_{10}x + c_{01}y + c_{11}xy & \text{für } r = 1, d = 2, \\ c_0 + c_{10}x + c_{01}y + c_{11}xy + c_{20}x^2 + & \\ \quad + c_{21}x^2y + c_{12}xy^2 + c_{02}y^2 + c_{22}x^2y^2 & \text{für } r = 2, d = 2, \\ \dots & \end{cases}$$

mit einem Multiindex α und Koeffizienten c_α darstellen lassen; alternativ lassen sie sich als Produkt von Polynomen in x und y mit jeweils der Ordnung r schreiben. Sei außerdem $Q^r(\mathbb{T}^n)$ das Bild von $\hat{Q}^r$ unter der (bi-, tri-)linearen Transformation von der Einheitszelle $\hat{K}$ auf die Zellen der Triangulation $\mathbb{T}^n$; da im Rahmen dieser Arbeit keine krummlinig berandeten Gebiete betrachtet werden, ist diese Transformation ausreichend. Damit definieren wir Ansatz- und Testraum:

$$\begin{aligned} \mathcal{W}_h = \big\{ \mathbf{w}_h = (u_h, v_h) : \; & \mathbf{w}_h(\mathbf{x}, t) \text{ stetig}, \\ & \mathbf{w}_h(\cdot, t)|_{I_n} \in \left(Q^r(\mathbb{T}^n) \cup Q^r(\mathbb{T}^{n-1}) \right)^2, \\ & \mathbf{w}_h(\cdot, t_n) \in \left(Q^r(\mathbb{T}^n) \right)^2, \\ & \mathbf{w}_h(\mathbf{x}, t)|_{I_n} \text{ linear in } t, \\ & u_h = g_D \text{ auf } \Gamma_D \big\}, \end{aligned} \quad (2.16)$$

$$\begin{aligned} \mathcal{T}_h = \big\{ \mathbf{t}_h = (\varphi_h, \psi_h) : \; & \mathbf{t}_h(\cdot, t) \text{ stetig auf } \Omega, \\ & \mathbf{t}_h(\mathbf{x}, t)|_{I_n} \text{ konstant in } t, \\ & \mathbf{t}_h(\cdot, t)|_{I_n} \in (Q^r(\mathbb{T}^n))^2, \\ & \varphi_h = 0 \text{ auf } \Gamma_D \big\}. \end{aligned} \quad (2.17)$$

$\mathbf{w}_h$ muß im Innern der Zeitintervalle aus der Vereinigung zweier Finite-Elemente-Räume kommen, was weiter unten bei der Behandlung von hängenden Knoten in der Zeit erläutert.

Aus der Formulierung (2.14) könnte man annehmen, daß die Geschwindigkeit v keine Randwerte außer den natürlich in der Formulierung enthaltenen benötigt. Allerdings zeigen numerische Experimente, daß das Problem dann nicht L^2-stabil ist und Oszillationen am Rand auftreten, sobald eine Welle diesen trifft (vgl. Abbildung 2.3). Die Amplitude der Oszillationen nimmt mit kleiner werdender Gitterweite wie etwa h^{-1} zu, die Wellenlänge beträgt für lineare Elemente zwei Gitterabstände. Die fehlende L^2-Stabilität beinhaltet, daß das Problem auch in der Energienorm instabil ist, wie in Abbildung 2.4 unten zu sehen ist; die Amplitude der Instabilität in der Energie ist ebenfalls von der Ordnung h^{-1}. In der Auslenkung u sind keine Instabilitäten erkennbar.

Die Instabilitäten verschwinden, wenn man für v_h auf Γ_D die Beziehung $v_h = \partial_t g_D$ fordert, wobei die Zeitableitung mit dem gleichen Verfahren wie bei der ganzen Gleichung diskretisiert sei. Entsprechend muß (2.17) um $\psi_h = 0$ auf Γ_D erweitert werden.

Um aus der Formulierung (2.15) und den Räumen (2.16) und (2.17) ein Zeitschrittverfahren zu erhalten, wähle man $\mathbf{t}_h$ so, daß es außerhalb von I_n verschwinde und innerhalb konstant sei. Da $\mathbf{w}_h$ stetig und linear in der Zeit sein sollte, gilt mit (2.15) für alle $\mathbf{t}_h = (\varphi_h, \psi_h) \in (Q_r(\mathbb{T}^n))^2$

$$(\rho \mathbf{w}_h^n - \rho \mathbf{w}_h^{n-1}, \mathbf{t}_h)_\Omega + \frac{k_n}{2} \left(\begin{pmatrix} 0 & -\rho \\ a\nabla & 0 \end{pmatrix} (\mathbf{w}_h^n + \mathbf{w}_h^{n-1}), \begin{pmatrix} 1 & 0 \\ 0 & \nabla \end{pmatrix} \mathbf{t}_h \right)_\Omega - (g_N, \psi_h)_{\Gamma_N \times I} = 0. \quad (2.18)$$

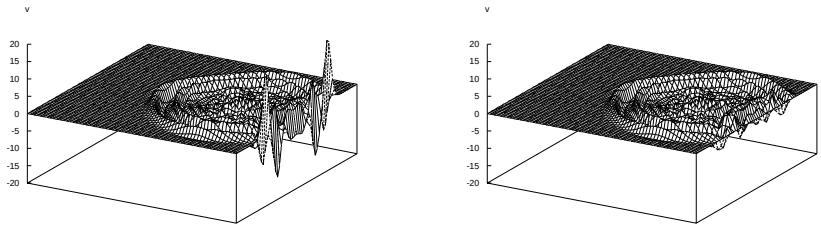

Abbildung 2.3: *Vergleich der Geschwindigkeiten v ohne (links) und mit (rechts) Behandlung der Randwerte für v.*

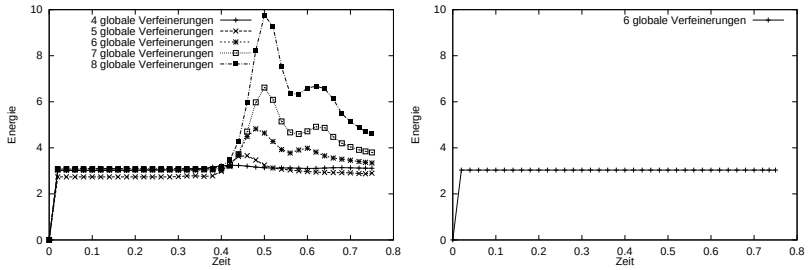

Abbildung 2.4: *Vergleich der Energieverläufe ohne (links) und mit (rechts) Behandlung der Randwerte für die Geschwindigkeit v.*

Das ist wieder die Ortsdiskretisierung mit finiten Elementen der mit dem CRANK-NICOLSON-Verfahren semidiskretisierten Gleichungen (2.4).

Da $\mathcal{T}_h \subset W$, gilt (2.14) auch für alle $\mathbf{t}_h \in \mathcal{T}_h$ und wir erhalten durch Subtraktion von (2.14) und (2.15) die GALERKIN-Orthogonalität für den Fehler $\mathbf{e} = \mathbf{w} - \mathbf{w}_h$

$$(\rho \mathbf{e}_t, \mathbf{t}_h)_\Sigma + \left(\begin{pmatrix} 0 & -\rho \\ a\nabla & 0 \end{pmatrix} \mathbf{e}, \begin{pmatrix} 1 & 0 \\ 0 & \nabla \end{pmatrix} \mathbf{t}_h \right)_\Sigma = 0, \qquad \forall \mathbf{t}_h \in \mathcal{T}_h, \tag{2.19}$$

die wir für die Fehlerschätzung benötigen werden.

2.3.1 Hängende Knoten im Ort

Der in dieser Arbeit verfolgte Ansatz zur Adaptivität verwendet Gitter, bei denen Zellen regulär verfeinert werden, ohne daß die entstehenden Knoten auf den Kanten auf der nächsten Zelle durch sogenannte „Abfangzellen“ zu regulären Knoten gemacht werden. Der Grund dafür ist, daß die Konstruktion von Abfangzellen auf Viereckgittern verhältnismäßig kompliziert ist und daß die mit dem verwendeten Verfahren entstehenden Gitter gut für Mehrgitteralgorithmen geeignet (vgl. [4, 22]); im Rahmen dieser Arbeit wurde davon allerdings kein Gebrauch gemacht.

Stattdessen belassen wir die hängenden Knoten im Gitter (vgl. Abbildung 2.5) und assoziieren mit ihnen spezielle Basisfunktionen. Dazu gibt es im wesentlichen drei Möglichkeiten:

- Nichtkonformer Ansatz: man definiert die Basisfunktionen wie üblich zellweise und nimmt die eventuelle Unstetigkeit an der Grenzfläche zwischen Zelle 1 und den Zellen 2 und 3 in Kauf. Gegegebenenfalls kann man einen Strafterm für die Unstetigkeit in die schwache Formulierung einfügen. Dieser Ansatz führt in den meisten Fällen zu einem nichtkonformen Ansatzraum.

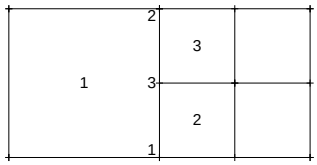

Abbildung 2.5: *Einfaches Gitter mit hängendem Knoten.*

- Assoziation einer Basisfunktion auf den Zellen 1 bis 3 zum Knoten 3. Dadurch ist ein stetiger Ansatzraum gewährleistet; auf den Zellen 2 und 3 sind die Basisfunktionen wie üblich, auf der Zelle 1 ist eine spezielle Basisfunktion zu wählen und die Basisfunktionen zu den Knoten 1 und 2 sind entsprechend zu modifizieren. Dieser Ansatz hat die Nachteile, daß die Basisfunktionen des feinen Gebiets bis in das grobe Gebiet hineinreichen, was die Implementation eines Mehrgitterverfahrens erschwert, und daß die Basisfunktionen auf einer Zelle unter Umständen von den Nachbarn abhängen, was einen erhöhten algorithmischen Aufwand bedeutet.

- Elimination der Basisfunktion zum Knoten drei. Dieser Weg hat den Vorteil, daß die Basisfunktionen weiterhin zellweise definiert werden können, wie auf regulären Gittern üblich und daß daher praktisch kein zusätzlicher algorithmischer Aufwand bei der Verwaltung des Ansatz- und Testraums nötig ist. Die Basisfunktion kann nach Aufbau der Matrizen und Vektoren eliminiert werden.

In der vorliegenden Arbeit wurde der dritte Weg gewählt; die Details der Implementation und der Elimination der Basisfunktionen zu hängenden Knoten sind in Abschnitt 4.4 zu finden.

2.3.2 Hängende Knoten in der Zeit

Um die globale Stetigkeit des Ansatzraumes zu gewährleisten, ist es nötig, hängende Knoten in der Zeit geeignet zu interpretieren. Unter einem hängenden Knoten in der Zeit sei dabei ein Knoten des örtlichen Gitters verstanden, der im vorhergehenden oder nachfolgenden Zeitschritt nicht vorhanden ist. Im wesentlichen stehen wieder die gleichen Möglichkeite wie oben offen, allerdings mit folgenden Unterschieden:

- Nichtkonformer Ansatz: in diesem Fall ist der Ansatzraum in der Zeit unstetig und man muß das Zeitschrittverfahren um Sprungterme ergänzen; das ist für die unstetigen Galerkin-Verfahren (*DG(r)*) zur Zeitdiskretisierung üblich, im Rahmen des Crank-Nicolson-Verfahrens aber nicht nötig.

- Assoziation einer Basisfunktion zu diesem Knoten. Dies ist der Weg, der hier beschritten werden soll. Er läuft darauf hinaus, daß die Basisfunktionen im Innern der Zeitintervalle als stetiger Übergang zwischen den Basisfunktionen der diskreten Zeitpunkte betrachtet werden müssen; Basisfunktionen, die auf einem der beiden Endpunkte des Intervalls keinen Knoten haben, werden auf dieser Seite als konstant Null angenommen, leben aber im Inneren und auf der anderen Seite. Das Zeitschrittverfahren behält seine Form bei und muß nicht modifiziert werden.

- Elimination der Basisfunktion. Dieser Weg ist ebenso möglich wie der obige, verlangt jedoch eine explizite Behandlung der Basisfunktionen zu hängenden Knoten in der Zeit und bedeutet damit zusätzlichen algorithmischen Aufwand. Insbesondere ist schwierig, daß auf diese Art Kopplungen zwischen mehr als nur zwei aufeinanderfolgenden Zeitschritten eingeführt werden; man muß dabei neben dem letzten auch auf den vorletzten Zeitschritt zugreifen, was den Vorteil der Umschreibung zu einem System erster Ordnung in der Zeit zunichte machen

würde. Der größere Nachteil ist jedoch, daß dieser Knoten ja extra eingefügt wurde, um eine höhere Auflösung zu erzielen, jetzt aber erst ab dem nächsten Zeitschritt zur Verfügung steht; während das bei der Verfeinerung im Ort wegen der relativ großen Verfeinerungsgebiete hinnehmbar ist, ist es hier störender und wegen der Vorteile des zweiten Weges nicht empfehlenswert.

2.3.3 Entkopplung der Gleichungen

Gleichung (2.18) enthält zwei miteinander gekoppelte, endlichdimensionale Gleichungssysteme für u_h^n und v_h^n, die man erhält, wenn man mit $\mathbf{t}_h = (\varphi_h, 0)$ und $\mathbf{t}_h = (0, \psi_h)$ testet:

$$\left(\rho u_h^n - \rho u_h^{n-1}, \varphi_h\right)_\Omega - \frac{k_n}{2}\left(\rho v_h^n + \rho v_h^{n-1}, \varphi_h\right)_\Omega = 0,$$

$$\left(\rho v_h^n - \rho v_h^{n-1}, \psi_h\right)_\Omega + \frac{k_n}{2}\left(a\nabla(u_h^n + u_h^{n-1}), \nabla\psi_h\right)_\Omega - \left(g_N, \psi_h\right)_{\Gamma_N \times I} = 0.$$

Da nach (2.17) der Raum der ψ_h die φ_h enthält, kann man die zweite Gleichung, für $\psi_h = \varphi_h$ betrachtet, nach $(\rho v_h^n, \varphi_h)_\Omega$ auflösen und das Resultat in die erste Gleichung einsetzen. Man erhält dann zwei entkoppelte, nacheinander lösbare Gleichungen:

$$\begin{aligned}\left(\rho u_h^n, \varphi\right)_\Omega + \frac{k^2}{4}\left(a\nabla u_h^n, \nabla\varphi\right) &= \left(\rho u_h^{n-1}, \varphi\right)_\Omega + k_n\left(\rho v_h^{n-1}, \varphi\right)_\Omega \\ &\quad - \frac{k^2}{4}\left(a\nabla u_h^{n-1}, \nabla\varphi\right) + \frac{k}{2}\left(g_N, \varphi\right)_{\Gamma_N \times I_n}, \\ \left(\rho v_h^n, \psi_h\right)_\Omega &= \left(\rho v_h^{n-1}, \psi_h\right)_\Omega - \frac{k_n}{2}\left(a\nabla(u_h^n + u_h^{n-1}), \nabla\psi_h\right)_\Omega + \left(g_N, \psi_h\right)_{\Gamma_N \times I}.\end{aligned}$$

Diese Umordnung ist analog zur Bildung des SCHUR-Komplements eines Block-Gleichungssystems. Die erste Gleichung ist eine HELMHOLTZ-Gleichung mit „gutartigem“ Vorzeichen; sie verursacht den meisten numerischen Aufwand beim Lösen. Die zweite Gleichung ist nur noch eine L^2-Projektion, die mit geringem Aufwand zu lösen ist.

Verwendet man die ursprüngliche Definition der Ansatz- und Testräume, so ist klar, daß es nicht möglich ist, die erste in die zweite Gleichung einzusetzen, da der Raum der φ_h kleiner als der der ψ_h ist. Fordert man die zusätzlichen Randbedingungen, so sind die beiden Räume gleich und die umgekehrte Einsetzung ist möglich; sie bringt aber außer einer eingesparten Matrix-Vektor-Multiplikation pro Zeitschritt keinen Vorteil.

Die beschriebene Entkopplung der beiden Gleichungen ist für das allgemeine θ-Verfahren und für das Fractional-Step-θ-Verfahren ganz analog möglich.

2.4 Fehlerkontrolle

In diesem Abschnitt soll ein a posteriori Fehlerschätzer für den Fehler zwischen der exakten und der numerischen Lösung bezüglich eines beliebigen Funktionals hergeleitet werden. Dieser Schätzer soll ausschließlich die numerische Lösung verwenden, so daß er den Fehler wirklich schätzen kann und nicht wie die klassischen Fehlerindikatoren aufgrund der Abhängigkeit von der unbekannten exakten Lösung den Fehler nur anzeigen kann.

Aufgrund der Ergebnisse aus 2.1 kommen unter den betrachteten Zeitschrittverfahren nur das CRANK-NICOLSON- und das Fractional-Step-θ-Verfahren in Frage. Da die Herleitung des Fehlerschätzers darauf beruht, daß sich die Diskretisierung der Gleichung als GALERKIN-Verfahren schreiben läßt, kann auf diesem Weg nur ein Schätzer für das CRANK-NICOLSON-Verfahren gefunden werden.

2.4.1 Vorbemerkungen und Notationen

In der Praxis interessiert an einer gefundenen oder berechneten Lösung oft nicht die Lösung selbst, sondern eine daraus berechnete Größe. Im Kontext der Wellengleichung könnten das beispielsweise

die Energie zu einem Zeitpunkt T

$$E(u) = \frac{1}{2}\left[(\rho u_t(\cdot,T), u_t(\cdot,T))_\Omega + a\,(u(\cdot,T), u(\cdot,T))_\Omega\right],$$

der integrierte Energiefluß durch eine Kurve $\mathcal{C}$

$$\Phi(u) = \int_0^T \int_{\mathcal{C}} u_t(a\nabla u)\cdot\mathbf{n}\; ds\; dt,$$

oder eine beliebige andere Größe sein. Die numerisch zu kontrollierende Quantität ist dann der Fehler in dieser Größe; es ist also beispielsweise zu garantieren, daß $|E(u)-E(u_h)| < Tol$, das heißt daß die aus der numerischen Lösung u_h berechnete Energie $E(u_h)$ um nicht mehr als Tol von der echten Energie $E(u)$ abweicht. Da wir die exakte Lösung u und damit $E(u)$ im allgemeinen aber nicht kennen, erscheint die Forderung nicht verifizierbar. Das im folgenden vorgestellte Verfahren wird aber zeigen, daß es trotzdem selbst dann möglich ist, eine Schätzung des Fehlers zu erhalten, wenn nur u_h bekannt ist.

Der Mechanismus zur Aufstellung des Fehlerschätzers verlangt, daß die zu kontrollierende Fehlergröße ein lineares Funktional $J(\cdot)$ der Lösung $\mathbf{w} = (u, v)$ ist. Sei $\mathcal{J}(\cdot)$ die interessierende Größe (also z. B. $\mathcal{J} = E$ oder $\mathcal{J} = \Phi$), so wäre $J(\mathbf{e}) = \mathcal{J}(\mathbf{w}) - \mathcal{J}(\mathbf{w}_h)$. Im Zusammenhang der Wellengleichung sind allerdings die meisten Größen quadratisch in der Lösung, weshalb eine Linearisierung um die Lösung vorgenommen wird. Das Fehlerfunktional $\tilde{J}(\cdot)$ muß dazu so gewählt werden, daß $\tilde{J}(\mathbf{e}) \approx \mathcal{J}(\mathbf{w}) - \mathcal{J}(\mathbf{w}_h)$. Für die beiden obigen Beispiele sind die Funktionale mit $\mathbf{t} = (\varphi, \psi)$ und $\mathbf{w} = (u, v)$ beispielsweise durch

$$\begin{aligned}\tilde{J}^E_{\mathbf{w}}(\mathbf{t}) &= \left\{(\rho v, \psi)_\Omega + (a\nabla u, \nabla\varphi)_\Omega\right\}_{t=T},\\ \tilde{J}^\Phi_{\mathbf{w}}(\mathbf{t}) &= \int_0^T \int_{\mathcal{C}} \left(v(a\nabla\varphi)\cdot\mathbf{n} + \psi(a\nabla u)\cdot\mathbf{n}\right)\; ds\; dt,\end{aligned}$$

gegeben, da

$$\begin{aligned}\tilde{J}^E_{\mathbf{w}}(\mathbf{e}) &= E(\mathbf{w}) - E(\mathbf{w}_h) + \frac{1}{2}\Big\{(a\nabla(u-u_h), \nabla(u-u_h))_\Omega + (\rho(v-v_h), v-v_h)_\Omega\Big\}_{t=T},\\ \tilde{J}^\Phi_{\mathbf{w}}(\mathbf{e}) &= \Phi(\mathbf{w}) - \Phi(\mathbf{w}_h) + \int_0^T \int_{\mathcal{C}} (v-v_h)(a\nabla(u-u_h))\cdot\mathbf{n}\; ds\; dt.\end{aligned}$$

In der Praxis hat die Linearisierung statt um $\mathbf{w}$ natürlich um $\mathbf{w}_h$ zu geschehen, was aber nichts an der Tatsache ändert, daß für $\mathbf{e} \to 0$ auch $\tilde{J}(\mathbf{e}) \to J(\mathbf{w}) - J(\mathbf{w}_h)$. Eine genauere Betrachtung der Linearisierung ist in Abschnitt 2.4.6 zu finden. Die vom Algorithmus zu garantierende Größe ist mit diesem Ansatz $\tilde{J} \le Tol$, in der Hoffnung, daß dann auch für den Fehler in der ursprünglichen Größe $\mathcal{J}(\mathbf{w}) - \mathcal{J}(\mathbf{w}_h) \le Tol$ gilt.

2.4.2 Exakte Fehlerdarstellung

Sei $\bar{\mathbf{w}} = (\bar{u}, \bar{v})$ die Lösung des dualen Problems

$$(\rho\mathbf{t}_t, \bar{\mathbf{w}})_{\Omega\times I} + b(\mathbf{t}, \bar{\mathbf{w}}) = J(\mathbf{t}) \qquad \forall\mathbf{t}\in W, \tag{2.20}$$

wobei $J(\cdot)$ wieder das Zielfunktional sei (bzw. $\tilde{J}$, falls die Auswertungsgröße nichtlinear ist). Da J linear in $\mathbf{t}$ und aus physikalischen Gründen ein beschränktes Funktional sein sollte, erhält man etwas Einblick in die Struktur der dualen Lösung, wenn man beachtet, daß sich J nach dem Rieszschen Darstellungssatz durch $J(\mathbf{t}) = (\mathbf{j}, \mathbf{t})_{\Omega\times I} = (j_1, t_1)_{\Omega\times I} + (j_2, t_2)_{\Omega\times I}$ mit einem $\mathbf{j}(\mathbf{x}, t)$ aus dem zum Lösungsraum dualen Raum darstellen läßt. Durch partielle Integration und einige Umformungen findet man, daß $\bar{\mathbf{w}}$ die Gleichungen

$$\begin{aligned}-\rho\bar{u}_t - \nabla\cdot a\nabla\bar{v} &= j_1,\\ -\rho\bar{v}_t - \rho\bar{u} &= j_2,\end{aligned}$$

oder nach ineinandereinsetzen

$$\rho\bar{v}_{tt} - \nabla \cdot a\nabla\bar{v} = j_1 - j_{2,t} \tag{2.21}$$

erfüllt.[6] Diese Wellengleichung ist in der Zeit rückwärts zu lösen; ihre Lösung beschreibt, mit welchen Gewichten ein Raum-Zeit-Punkt zum Ergebnis $J(\mathbf{w})$ beiträgt. End- und Randwerte für $\bar{\mathbf{w}}$ sind durch die unter Umständen singulären Gewichte $\mathbf{j}$ gegeben; man beachte auch die Vertauschung der Rollen von $\bar{u}$ und $\bar{v}$.

Aufgrund des dualen Problems gilt für den Fehler in der Zielgröße

$$J(\mathbf{e}) = (\rho\mathbf{e}_t, \bar{\mathbf{w}})_{\Omega\times I} + b(\mathbf{e}, \bar{\mathbf{w}}).$$

Mit der Galerkin-Orthogonalität (2.19) läßt sich ein beliebiges $\bar{\mathbf{w}}_h = (\bar{u}_h, \bar{v}_h) \in \mathcal{T}_h$ einschieben:[7]

$$\begin{aligned}
J(\mathbf{e}) &= (\rho\mathbf{e}_t, \bar{\mathbf{w}} - \bar{\mathbf{w}}_h)_{\Omega\times I} + b(\mathbf{e}, \bar{\mathbf{w}} - \bar{\mathbf{w}}_h) \\
&= \underbrace{(\rho\mathbf{w}_t, \bar{\mathbf{w}} - \bar{\mathbf{w}}_h)_{\Omega\times I} + b(\mathbf{w}, \bar{\mathbf{w}} - \bar{\mathbf{w}}_h)}_{=0,\text{ wegen (2.14)}} - (\rho\mathbf{w}_{h,t}, \bar{\mathbf{w}} - \bar{\mathbf{w}}_h)_{\Omega\times I} - b(\mathbf{w}_h, \bar{\mathbf{w}} - \bar{\mathbf{w}}_h) \\
&= \sum_{n=1}^{N} \sum_{K\in\mathbb{T}^n} \tilde{\mathcal{E}}_{K,n} \\
\tilde{\mathcal{E}}_{K,n} &= \Big\{ - (\rho\mathbf{w}_{h,t}, \bar{\mathbf{w}} - \bar{\mathbf{w}}_h)_{K\times I_n} + (\rho v_h, \bar{u} - \bar{u}_h)_{K\times I_n} \\
&\qquad - (a\nabla u_h, \nabla(\bar{v} - \bar{v}_h))_{K\times I_n} + (g_N, \bar{v} - \bar{v}_h)_{(\Gamma_N\cap\partial K)\times I_n} \Big\}.
\end{aligned}$$

Um diese Fehleridentität auch als Indikator zur Gitterverfeinerung verwenden zu können, muß der Fehler besser lokalisiert werden; dies geschieht durch partielle Integration des dritten Terms, so daß wir mit der Definition eines Sprungterms

$$[\mathbf{n} \cdot a\nabla u_h] = \begin{cases} \mathbf{n} \cdot a(\nabla u_h|_K - \nabla u_h|_J) & \text{für } \gamma \subset \partial K, \gamma \not\subset \partial\Omega, \\ 2\mathbf{n} \cdot a\nabla u_h|_K & \text{für } \gamma \subset \partial K, \gamma \subset \Gamma_D, \\ 2(\mathbf{n} \cdot a\nabla u_h|_K - g_N) & \text{für } \gamma \subset \partial K, \gamma \subset \Gamma_N, \end{cases}$$

wobei J das zu K benachbarte Element mit gemeinsamer Seite γ sei, schreiben können:

$$\begin{aligned}
J(e) &= \sum_{n=1}^{N} \sum_{K\in\mathbb{T}^n} \mathcal{E}_{K,n} \\
\mathcal{E}_{K,n} &= - (\rho\mathbf{w}_{h,t}, \bar{\mathbf{w}} - \bar{\mathbf{w}}_h)_{K\times I_n} + (\rho v_h, \bar{u} - \bar{u}_h)_{K\times I_n} \\
&\quad + (\nabla \cdot a\nabla u_h, \bar{v} - \bar{v}_h)_{K\times I_n} - \frac{1}{2} ([\mathbf{n} \cdot a\nabla u_h], \bar{v} - \bar{v}_h)_{\partial K\times I_n}.
\end{aligned} \tag{2.22}$$

Beachtet man, daß sowohl $\bar{v}$ als auch $\bar{v}_h$ auf Γ_D Nullrandwerte haben, dann kann man die zweite Zeile der Definition des Sprungterms auch Null setzen.

Der Grund für die partielle Integration liegt darin, daß wir $|\mathcal{E}_{K,n}|$ als Indikator für die Verfeinerung der Zellen verwenden wollen.[8] Während die Darstellungen in den geschweiften Klammern oben und (2.22) natürlich summiert den selben Wert ergeben, sind die Beiträge der einzelnen Zellen unterschiedlich und insbesondere von verschiedener Konvergenzordnung, was man sieht, wenn

[6] Die Gleichung ist im schwachen Sinn zu verstehen, da die $\mathbf{j}$ im allgemeinen keine starke Lösung zulassen. Die differentielle Form wurde nur zur Illustration der Art des Problems gewählt.

[7] Man beachte, daß $\bar{\mathbf{w}}_h$ bisher in keiner Beziehung zu $\bar{\mathbf{w}}$, der exakten dualen Lösung, stehen muß. Insbesondere sei nicht impliziert, daß $\bar{\mathbf{w}}_h$ durch die numerische Lösung eines Problems entstanden sein muß.

[8] Man braucht die Betragsstriche, da man ansonsten nicht weiß wie man mit Zellen verfahren soll, bei denen die Fehlerbeiträge unterschiedliche Vorzeichen haben.

man den obigen Indikator $\tilde{\mathcal{E}}_{K,n}$ beispielweise für eine im Innern von Ω gelegene Zelle K partiell integriert:

$$\tilde{\mathcal{E}}_{K,n} = \mathcal{E}_{K,n} - \frac{1}{2} \left(\mathbf{n} \cdot a(\nabla u_h|_K + \nabla u_h|_J), \bar{v} - \bar{v}_h \right)_{\partial K \times I_n} ,$$

wobei J wieder die jeweils zu K benachbarte Zelle sei. Im Grenzfall $h \to 0$ geht der Sprungterm in $\mathcal{E}_{K,n}$ gegen Null, wogegen der erste Faktor im letzten Term der letzten Gleichung gegen etwas Konstantes geht.

Die Darstellung (2.22) ist noch nicht sehr hilfreich, da die exakte duale Lösung $\bar{\mathbf{w}}$ unbekannt ist. Um (2.22) doch noch auswerten zu können, gibt es jedoch verschiedene Möglichkeiten (für statische Probleme vergleiche [6]), die im folgenden diskutiert werden sollen.

2.4.3 Auswertung mit höherem Ansatzgrad

Ersetzt man die exakte duale Lösung $\bar{\mathbf{w}}$ durch eine Approximation, die man aus der numerischen Lösung des dualen Problems erhält, so läßt sich die Fehleridentität näherungsweise auswerten. Es reicht dabei nicht aus, die duale Lösung mit dem gleichen Verfahren wie das primale Problem zu berechnen, da sich dann der Fehler durch Wahl von $\bar{\mathbf{w}}_h = \bar{\mathbf{w}}_h^{(r)}$ zu Null auswerten ließe (vorausgesetzt, $\bar{\mathbf{w}}_h^{(r)}$ liegt in $\mathcal{T}_\langle$), was natürlich keinen sinnvollen Fehlerschätzer ergäbe. Approximiert man aber $\bar{\mathbf{w}}$ durch eine numerische Lösung $\bar{\mathbf{w}}_h^{(r+1)}$, die mit einem um eins höheren Polynomgrad auf jeder Zelle, aber gleicher Ordnung in der Zeit und auf dem gleichen Orts-Zeit-Gitter erhalten wurde, so läßt sich die Fehleridentität (2.22) näherungsweise auswerten. Für $\bar{\mathbf{w}}_h$ setzen wir

$$\bar{\mathbf{w}}_\mathbf{h}|_{I_n} = \frac{1}{2} \mathcal{I}_h^{(r)} \left(\bar{\mathbf{w}}_h^{(r+1)}(t_{n-1}) + \bar{\mathbf{w}}_h^{(r+1)}(t_n) \right) ,$$

d. h. die Interpolation auf den Ansatzraum des primalen Problems. Man beachte, daß $\bar{\mathbf{w}}_h$ auf jedem Intervall I_n in der Zeit konstant sein muß, da $\bar{\mathbf{w}}_h \in \mathcal{T}_h$. Die Wahl der Interpolation anstatt der Projektion auf den niedrigeren Polynomraum ist nicht optimal; das gilt vor allem für $r = 1$, da in diesem Fall $\left(\bar{\mathbf{w}}_h^{(r+1)} - \bar{\mathbf{w}}_h \right)(\cdot, t)$ für festes t auf jeder Zelle entweder eine positive oder negative Funktion ist und wir somit Weghebungseffekte durch wechselnde Vorzeichen auf einer Zelle (wie wir sie beispielsweise durch eine Projektion erhalten könnten) nicht nutzen können. Der Grund für die Wahl der Interpolation liegt darin, daß $\bar{\mathbf{w}}_h$ stetig im Ort sein muß und daß wir das durch lokale Interpolation auf jeder Zelle erreichen können, nicht jedoch durch eine Projektion. Es wurde kein Versuch unternommen, durch globale oder patchweise Projektion die Verwendung des Fehlerschätzers zur Gittersteuerung noch zu optimieren.

Zur Auswertung des Fehlerschätzers beachten wir, daß $\mathbf{w}_{h,t}$ und $\bar{\mathbf{w}}_h$ auf jedem Intervall konstant in der Zeit sind. $\bar{\mathbf{w}}_h^{(r+1)}$ wird als auf jedem Zeitintervall als linear in der Zeit angenommen (das entspricht der Auswertung des Zeitintegrals mit der Trapezregel); diese Annahme ist sogar exakt, falls die $\bar{\mathbf{w}}_h^{(r+1)}$ durch Zeitdiskretisierung mit dem Crank-Nicolson-Verfahren erhalten wurde. Die gleiche Annahme gelte für g_N. Man erhält dann für die einzelnen Terme von (2.22) die folgenden Ausdrücke (die Superskripte $(r+1)$ an $\bar{\mathbf{w}}_h^{(r+1)}$ seien im folgenden weggelassen, um Platz für die den Zeitschritt bezeichnenden Superskripte zu haben; alle $\bar{\mathbf{w}}_h$ auf der rechten Seite

sind in Wirklichkeit $\bar{\mathbf{w}}_h^{(r+1)}$):

$$
\begin{aligned}
-\left(\rho \mathbf{w}_{h,t}, \bar{\mathbf{w}} - \bar{\mathbf{w}}_h\right)_{K\times I_n} &= -\int_{I_n} \left(\rho \frac{\mathbf{w}_h^n - \mathbf{w}_h^{n-1}}{k}, \bar{\mathbf{w}}_h^{(r+1)}(t) - \bar{\mathbf{w}}_h\right)_K dt \\
&\approx -\left(\rho(\mathbf{w}_h^n - \mathbf{w}_h^{n-1}), \frac{1}{2}(\mathbf{1} - \mathcal{I}_h^{(r)})(\bar{\mathbf{w}}_h^n + \bar{\mathbf{w}}_h^{n-1})\right)_K, \\
\left(\rho v_h, \bar{u} - \bar{u}_h\right)_{K\times I_n} &\approx k\Bigg\{\frac{1}{4}\left(\rho(v_h^n + v_h^{n-1}), (\mathbf{1} - \mathcal{I}_h^{(r)})(\bar{u}_h^n + \bar{u}_h^{n-1})\right)_K \\
&\qquad + \frac{1}{12}\left(\rho(v_h^n - v_h^{n-1}), \bar{u}_h^n - \bar{u}_h^{n-1}\right)_K\Bigg\}, \\
\left(\nabla\cdot a\nabla u_h, \bar{v} - \bar{v}_h\right)_{K\times I_n} &\approx k\Bigg\{\frac{1}{4}\left(\nabla\cdot a\nabla(u_h^n + u_h^{n-1}), (\mathbf{1} - \mathcal{I}_h^{(r)})(\bar{v}_h^n + \bar{v}_h^{n-1})\right)_K \\
&\qquad + \frac{1}{12}\left(\nabla\cdot a\nabla(u_h^n - u_h^{n-1}), \bar{v}_h^n - \bar{v}_h^{n-1}\right)_K\Bigg\} \\
&= k\Bigg\{\frac{1}{4}\left(a\Delta(u_h^n + u_h^{n-1}), (\mathbf{1} - \mathcal{I}_h^{(r)})(\bar{v}_h^n + \bar{v}_h^{n-1})\right)_K \\
&\qquad + \frac{1}{12}\left(a\Delta(u_h^n - u_h^{n-1}), \bar{v}_h^n - \bar{v}_h^{n-1}\right)_K\Bigg\} \\
&\quad + k\Bigg\{\frac{1}{4}\left((\nabla a)\cdot\nabla(u_h^n + u_h^{n-1}), (\mathbf{1} - \mathcal{I}_h^{(r)})(\bar{v}_h^n + \bar{v}_h^{n-1})\right)_K \\
&\qquad + \frac{1}{12}\left((\nabla a)\cdot\nabla(u_h^n - u_h^{n-1}), \bar{v}_h^n - \bar{v}_h^{n-1}\right)_K\Bigg\}, \\
-\frac{1}{2}\left([\mathbf{n}\cdot a\nabla u_h], \bar{v} - \bar{v}_h\right)_{\partial K\times I_n} &\approx -\frac{k}{2}\Bigg\{\frac{1}{4}\left([\mathbf{n}\cdot a\nabla(u_h^n + u_h^{n-1})], (\mathbf{1} - \mathcal{I}_h^{(r)})(\bar{v}_h^n + \bar{v}_h^{n-1})\right)_{\partial K} \\
&\qquad + \frac{1}{12}\left([\mathbf{n}\cdot a\nabla(u_h^n - u_h^{n-1})], \bar{v}_h^n - \bar{v}_h^{n-1}\right)_{\partial K}\Bigg\}.
\end{aligned}
$$

Die einzelnen Terme werden mit einer Quadraturformel der Ordnung $r+2$ ausgewertet. Der dritte Term ist für allgemeine Gitter etwas kompliziert, da bei der Transformation der zweiten Ableitung auf die Einheitszelle Ableitungen der JACOBI-Matrix und des Koeffizienten auftreten; diese lassen sich aber a priori als Funktion der Eckpunkte der Zellen angeben, so daß die Auswertung ohne numerische Differentation möglich ist.

Die Auswertung der einzelnen Terme ist auch dann möglich, wenn die Zelle K zwischen t_{n-1} und t_n vergröbert oder verfeinert wurde; das Verfahren ist analog dem in Abschnitt 4.2 beschriebenen, jedoch komplizierter. Insbesondere ist die Auswertung bei Vergröberung programmtechnisch ausgesprochen aufwendig, was an der Anwendung der Interpolation $I_h^{(r)}\bar{\mathbf{w}}^{n-1}$ liegt; da die Interpolation auf den Testraum $Q^r(\mathbb{T}^n)$ durchgeführt werden muß, kann sie bei Vergröberung nicht auf den kleinen Zellen des alten Gitters geschehen, auf denen $\bar{\mathbf{w}}^{n-1}$ definiert ist, sondern nur auf der entsprechenden, gröberen Zelle des neuen Gitters und ist somit *nichtlokal*, was die Programmierung erheblich kompliziert. Zu beachten ist auch, daß im Falle der Vergröberung die Gradienten im Innern einer Raum-Zeit-Zelle $K\times I_n$, $K\in\mathbb{T}^n$, unstetig sein können, so daß zusätzliche Sprungterme auftreten.

Die Verfeinerung von Zellen ist dagegen einfach zu behandeln, da die Auswertung der Interpolation aufgrund der Einbettung der Räume auf den jeweils kleinsten Zellen durchgeführt werden kann.

Aufgrund der geschilderten Probleme mit dem Term $I_h^{(r)}\bar{\mathbf{w}}^{n-1}$ erscheint es naheliegend, statt der obigen Wahl von $\mathbf{w}_h$ die folgende Interpolation zu nehmen:

$$\bar{\mathbf{w}}_\mathbf{h}|_{I_n} = \mathcal{I}_h^{(r)}\bar{\mathbf{w}}_h^{(r+1)}(t_n).$$

Dadurch lassen sich die geschilderten Probleme umgehen, allerdings um den Preis, daß der Fehlerschätzer lokal nicht die maximale Ordnung zeigt (es fehlt eine k-Potenz), so daß die erzeugten Gitter unter Umständen nicht optimal sind.

2.4.4 Abschätzung mit dem Bramble-Hilbert-Lemma

Einfache Abschätzung

Schreibt man (2.22) etwas um, so erhält man

$$\begin{aligned}\mathcal{E}_{K,n} = -\big(\underbrace{\rho u_{h,t} - \rho v_h}_{r_1}, \bar{u} - \bar{u}_h\big)_{K\times I_n} - \big(\underbrace{\rho v_{h,t} - \nabla\cdot a\nabla u_h}_{r_2}, \bar{v} - \bar{v}_h\big)_{K\times I_n} \\ - \frac{1}{2}\left([\mathbf{n}\cdot a\nabla u_h], \bar{v} - \bar{v}_h\right)_{\partial K\times I_n}\end{aligned} \tag{2.23}$$

Die ersten beiden Terme sind Residuenterme, der letzte mißt die Glattheit der numerischen Lösung. Dies schätzt man nun mittels der Cauchy-Schwarzschen Ungleichung ab:

$$\begin{aligned}|J(e)| &\le \sum_{n=1}^{N}\sum_{K\in\mathbb{T}^n}|\mathcal{E}_{K,n}|,\\ |\mathcal{E}_{K,n}| \;&\le\; \|r_1\|_{K\times I_n}\|\bar{u}-\bar{u}_h\|_{K\times I_n} + \|r_2\|_{K\times I_n}\|\bar{v}-\bar{v}_h\|_{K\times I_n}\\ &\quad +\tfrac{1}{2}\|[\mathbf{n}\cdot a\nabla\bar{u}]\|_{\partial K\times I_n}\|\bar{v}-\bar{v}_h\|_{\partial K\times I_n}.\end{aligned}$$

Wir wollen den Fehler so scharf wie möglich abschätzen und versuchen daher, ein $\bar{\mathbf{w}}_h \in \mathcal{T}_h$ nahe an der exakten dualen Lösung $\bar{\mathbf{w}}$ zu wählen, also zum Beispiel die Interpolation $\bar{\mathbf{w}}_h|_{I_n} = \mathcal{I}_h^{(r)}\bar{\mathbf{w}}(t_{n-1})$ oder die Projektion $\bar{\mathbf{w}}_h = \Pi_{\mathcal{T}_h}\bar{\mathbf{w}}$.

Im Falle der Interpolation lassen sich die Terme mit $\bar{\mathbf{w}} - \bar{\mathbf{w}}_h$ mit dem Bramble-Hilbert-Lemma und einer Erweiterung für Polynome mit unterschiedlichem Grad in Orts- und Zeitrichtung (siehe zum Beispiel [7] und [19]) wie folgt abschätzen:[9]

$$\begin{aligned}\|\bar{u}-\bar{u}_h\|_{K\times I_n} &\le C\left(k_n\|\bar{u}_t\|_{K\times I_n} + k_n^{r+1}\left\|\partial_t^{r+1}\bar{u}\right\|_{K\times I_n} + h_K^{r+1}\left\|\nabla^{r+1}\bar{u}\right\|_{K\times I_n}\right),\\ \|\bar{v}-\bar{v}_h\|_{K\times I_n} &\le C\left(k_n\|\bar{v}_t\|_{K\times I_n} + k_n^{r+1}\left\|\partial_t^{r+1}\bar{v}\right\|_{K\times I_n} + h_K^{r+1}\left\|\nabla^{r+1}\bar{v}\right\|_{K\times I_n}\right),\\ \|\bar{v}-\bar{v}_h\|_{\partial K\times I_n} &\le Ch_K^{-\frac{1}{2}}\left(\|\bar{v}-\bar{v}_h\|_{K\times I_n} + h_K\|\nabla(\bar{v}-\bar{v}_h)\|_{K\times I_n}\right)\\ &\le Ch_K^{-\frac{1}{2}}\left(k_n\|\bar{v}_t\|_{K\times I_n} + k_nh_K\|\nabla\bar{v}_t\|_{K\times I_n}\right).\end{aligned}$$

Hinreichende Glattheit von $\bar{\mathbf{w}}$, z. B. $\bar{\mathbf{w}} \in H^{r+1}(\Omega\times[0,T])$ wurde hier vorausgesetzt; in der Praxis ist allerdings oft weniger gegeben, insbesondere bei lokalisierten Fehlerfunktionalen (Kantenintegrale, Punktauswertungen). Die Abschätzungen bleiben im wesentlichen gültig, wenn man die nicht mehr existierenden Ableitungen durch niedrigere Ableitungen ersetzt und den Exponenten der h- und k-Potenzen entsprechend erniedrigt. Da in der numerischen Praxis die exakte duale Lösung $\bar{\mathbf{w}}$ ohnehin durch eine auf numerischem Weg erhaltene, sowie Ableitungen durch Differenzenquotienten ersetzt werden, verlieren die obigen Abschätzungen ihren Wert nicht und es sei an dieser Stelle auf die Unübersichtlichkeit der Verallgemeinerung auf fehlende Regularität verzichtet.

[9] Man verwende für die ersten beiden Ungleichungen [19, Satz 2.3.7] mit $\mu = (0, r, r)$. μ bezeichnet die Polynomgrade des diskreten Testraums $\mathcal{T}_h$ in der Zeit- und den beiden Ortsrichtungen. Mit der üblichen *shape regularity*-Bedingung an das Gitter kann man die Ableitungen in x- und y-Richtung zusammenfassen und erhält nur noch Terme mit h_K statt h_x und h_y. Für den dritten Term verwende man den Spursatz.

Unter der Annahme, daß die Abschätzungen in der obigen Form gültig sind, läßt sich ein Fehlerschätzer η gemäß

$$\begin{aligned} \eta &= C \sum_{n=1}^{N} \sum_{K \in \mathbb{T}^n} \eta_{K,n} \\ \eta_{K,n} &= h_K^d k_n \left(\rho_{K,n}^{1,k} \omega_{K,n}^{1,k} + \rho_{K,n}^{1,h} \omega_{K,n}^{1,h} + \rho_{K,n}^{2,k} \omega_{K,n}^{2,k} + \rho_{K,n}^{2,h} \omega_{K,n}^{2,h} + \rho_{K,n}^{3} \omega_{K,n}^{3} \right), \end{aligned} \tag{2.24}$$

mit den Residuen und Gewichten

$$\begin{aligned} \rho_{K,n}^{1,k} &= h_K^{-d/2} k_n^{1/2} \left\| r_1 \right\|_{K \times I_n}, \\ \omega_{K,n}^{1,k} &= h_K^{-d/2} k_n^{-1/2} \left(\left\| \bar{u}_t \right\|_{K \times I_n} + k_n^r \left\| \partial_t^{r+1} \bar{u} \right\|_{K \times I_n} \right), \\ \\ \rho_{K,n}^{1,h} &= h_K^{r+1} k_n^{-1} \rho_{K,n}^{1,k} = h_K^{r+1-d/2} k_n^{-1/2} \left\| r_1 \right\|_{K \times I_n}, \\ \omega_{K,n}^{1,h} &= h_K^{-d/2} k_n^{-1/2} \left\| \nabla^{r+1} \bar{u} \right\|_{K \times I_n}, \\ \\ \rho_{K,n}^{2,k} &= h_K^{-d/2} k_n^{1/2} \left\| r_2 \right\|_{K \times I_n}, \\ \omega_{K,n}^{2,k} &= h_K^{-d/2} k_n^{-1/2} \left(\left\| \bar{v}_t \right\|_{K \times I_n} + k_n^r \left\| \partial_t^{r+1} \bar{v} \right\|_{K \times I_n} \right), \\ \\ \rho_{K,n}^{2,h} &= h_K^{r+1} k_n^{-1} \rho_{K,n}^{2,k} = h_K^{r+1-d/2} k_n^{-1/2} \left\| r_2 \right\|_{K \times I_n}, \\ \omega_{K,n}^{2,h} &= h_K^{-d/2} k_n^{-1/2} \left\| \nabla^{r+1} \bar{v} \right\|_{K \times I_n}, \\ \\ \rho_{K,n}^{3} &= \frac{1}{2} h_K^{-d/2} k_n^{1/2} \left\| [\mathbf{n} \cdot a \nabla \bar{u}] \right\|_{\partial K \times I_n}, \\ \omega_{K,n}^{3} &= h_K^{-d/2} k_n^{-1/2} \left(\left\| \bar{v}_t \right\|_{K \times I_n} + h_K \left\| \nabla \bar{v}_t \right\|_{K \times I_n} \right), \end{aligned}$$

definieren; d bezeichne die Raumdimension. Es ist aus der Herleitung offensichtlich, daß $|J(\mathbf{e})| \leq \eta$, d. h. daß wir η nutzen können um die Größe des Fehlers nach oben hin zu garantieren. Die Vorfakten der Gewichte $\omega_{K,n}^i$ wurden so gewählt, daß sie für $h_K, k_n \to 0$ gegen kontinuierliche Funktionen gehen; der gemeinsame Faktor $h_K^d k_n$ wurde aus den Gewichten und Residuen gezogen, damit die Summe im erwähnten Grenzfall gegen ein Integral geht.

Durch Aufsummation erhält man eine globale Ordnung von $\mathcal{O}(k + h^{r+1})$. Das ist offensichtlich suboptimal, da die Konvergenzordnung des Crank-Nicolson-Verfahrens $\mathcal{O}(k^2)$ ist; das Problem liegt hier darin, daß wir die Bramble-Hilbert-Abschätzung auf Elemente der Testräume T und $\mathcal{T}_h$ angewendet haben, während die quadratische k-Ordnung im Raum $\mathcal{W}_h$ liegt. Ein Ansatz, die Ordnung des Fehlerschätzers der des Problems anzupassen, wird im nächsten Abschnitt skizziert.

Für Elemente mit ungeradem Ansatzgrad ist bekannt, daß die Terme mit dem Zellresiduum, d. h. ρ^1 und ρ^2 gegenüber dem Kantenterm ρ^3 von höherer Ordnung sind und daher im Limes vernachlässigt werden können [30]. Verwendet man noch eine a priori Stabilitätsabschätzung der dualen Lösung der Form $\omega^3 \leq C_S(k, u)$, was bei der Abschätzung des Energiefehlers im allgemeinen möglich ist, so erhält man einen Schätzer der Form (2.9), was seine ad-hoc Verwendung plausibel macht.

Getrennte Abschätzung für Zeit- und Ortsgitterweite

Man kann das eben beschriebene Problem der falschen Konvergenzordnung umgehen, wenn man statt der Interpolation $\bar{\mathbf{w}}_h = \mathcal{I}_h^{(r)} \bar{\mathbf{w}}$ eine Interpolation mit Projektionscharakter wählt. Dieser Ansatz sei hier nur kurz skizziert, Details sind in [6] für statische und in [19] für die Anwendung auf das Crank-Nicolson-Verfahren zu finden.

Durch diese Wahl von $\bar{\mathbf{w}}_h$ läßt sich daneben auch eine Trennung der h- und k-Potenzen in den Beiträgen zu $\eta_{K,n}$ erreichen, vergleiche [7, 19], womit es möglich ist, die Fehler, die durch die Orts- bzw. Zeitdiskretisierung entstanden, zu trennen. Dadurch ist es möglich, lokale Verfeinerung in Ort und Zeit unabhängig voneinander adaptiv durchzuführen, je nachdem, welcher Teil des Fehlerschätzers den größeren Beitrag liefert.

Zur Herleitung eines entsprechenden Fehlerschätzers verwenden wir die Interpolation mit Projektionseigenschaften $\tilde{\mathcal{I}}_{\mathcal{T}}$ aus [19], die folgende Eigenschaften hat:[10]

- Es gelte $\tilde{\mathcal{I}}_{\mathcal{T}} = \tilde{\mathcal{I}}_h^{(1)} P_{I_n}^0$, wobei $P_{I_n}^0$ die Projektion in der Zeit auf stückweise konstante Funktionen und $\tilde{\mathcal{I}}_h^{(1)}$ die im folgenden beschriebene Interpolation im Ort ist.

- Zur Definition der Interpolation im Ort unterteilen wir das Gebiet Ω so in Makrozellen $\tilde{K}$, daß jede Makrozelle genau einmal verfeinert werden muß, um zum feinsten Gitter zu kommen. Jeder Zelle K läßt sich so eindeutig eine Makrozelle $\tilde{K}$ zuordnen, die sozusagen ihre Mutterzelle ist.

- Der Einfachheit halber beschränken wir uns in der folgenden Beschreibung auf zwei Raumdimensionen. Jede Makrozelle besteht dort aus vier Kindzellen und umfaßt somit 9 Knoten. Die Interpolation ist nun so definiert, daß sie in den vier Eckpunkten den Punktwert der zu interpolierenden Funktion auswertet, den vier Seitenmitten der Makrozelle den Mittelwert der beiden umliegenden Eckpunkte zuordnet und dem Knoten in der Mitte einen solchen Wert zuweist, daß die Differenz zwischen einer beliebigen Funktion und ihrer Interpolierenden den Mittelwert Null hat:

  $$\left(u - \tilde{\mathcal{I}}_h^{(1)} u, \varphi\right)_{\tilde{K}} = 0 \qquad \forall \varphi \in P_0(\tilde{K}). \tag{2.25}$$

  Diese letzte Eigenschaft verleiht der Interpolation einen Projektionscharakter, da der Fehler senkrecht auf dem Raum der Funktionen steht, die auf jeder Makrozelle konstante Werte haben.

  Im Inneren jeder der vier Zellen der Makrozelle ist die Funktion die bilineare Funktion mit den oben definierten Werten in den Knoten.

- Die Erweiterung auf beliebige Raumdimensionen ist offensichtlich; aufgrund der Konstruktion über Makrozellen paßt sich die Interpolation natürlich in das Umfeld adaptiv verfeinerter Gitter ein.

Für weitere Details siehe [19, Abschnitte 2.4.1 bis 2.4.3]. Die Wahl dieser relativ komplizierten Interpolation ist nur für die Abschätung relevant, sie muß nicht tatsächlich ausgewertet werden. Diese Interpolation wurde so gewählt, daß sie die Vorteile der Interpolation (lokale Fehlerabschätzungen) und der Projektion (der Fehler steht senkrecht auf bestimmten Räumen) in sich vereint.

Mit der so definierten Interpolation $\tilde{\mathcal{I}}_{\mathcal{T}}$ läßt sich nun wieder ein Fehlerschätzer herleiten, wie am ersten Zellresiduenterm in (2.23) demonstriert sei. Dazu spalten wir zuerst in einen Orts- und einen Zeitanteil auf:

$$\left(r_1, \bar{u} - \tilde{\mathcal{I}}_{\mathcal{T}} \bar{u}\right)_{K \times I_n} = \left(r_1, \bar{u} - P_{I_n}^0 \bar{u}\right)_{K \times I_n} + \left(r_1, P_{I_n}^0 \bar{u} - \tilde{\mathcal{I}}_{\mathcal{T}} \bar{u}\right)_{K \times I_n}.$$

Die beiden Terme werden nun nacheinander behandelt. Im ersten Term beachten wir daß der Fehler der Projektion $P_{I_n}^0$ senkrecht auf dem Raum $P_0(I_n)$ steht und die Stabilität der Projektion

[10] In [19] ist die Interpolation für beliebige Polynomgrade des Testraums beschrieben. Da wir hier aber nur an einem Fehlerschätzer für das Crank-Nicolson-Verfahren interessiert sind, gilt das folgende nur für den Spezialfall in der Zeit stückweise konstanter Testfunktionen. Da in der zitierten Arbeit nur bilineare Elemente untersucht werden, seien diese auch im folgenden verwendet; die Erweiterung auf höhere Ansatzgrade ist offensichtlich, aber technisch, und sei hier deshalb unterlassen.

$\left\|P^0_{I_n}\bar{u}\right\|_{I_n} \leq C\left\|\bar{u}\right\|_{I_n}$:

$$\begin{aligned}
\left(r_1, \bar{u} - P^0_{I_n}\bar{u}\right)_{K\times I_n} &= \left(r_1, \bar{u} - P^0_{I_n}\bar{u}\right)_{K\times I_n} \\
&= \left(r_1 - P^0_{I_n} r_1, \bar{u} - P^0_{I_n}\bar{u}\right)_{K\times I_n} \\
\left|\left(r_1, \bar{u} - P^0_{I_n}\bar{u}\right)_{K\times I_n}\right| &\leq \int_K \left\|r_1 - P^0_{I_n} r_1\right\|_{I_n} \left\|\bar{u} - P^0_{I_n}\bar{u}\right\|_{I_n} dx \\
&\leq C\int_K \min\left\{\left\|r_1\right\|_{I_n}, k_n\left\|\partial_t r_1\right\|_{I_n}\right\} k_n\left\|\partial_t\bar{u}\right\|_{I_n} dx \\
&\leq Ck_n \min\left\{\left\|r_1\right\|_{K\times I_n}, k_n\left\|\partial_t r_1\right\|_{K\times I_n}\right\}\left\|\partial_t\bar{u}\right\|_{K\times I_n}.
\end{aligned}$$

Für den zeiten Summanden oben gilt einerseits

$$\left|\left(r_1, P^0_{I_n}\bar{u} - \tilde{\mathcal{I}}_{\mathcal{T}}\bar{u}\right)_{K\times I_n}\right| \leq \int_{I_n}\left\|r_1\right\|_K \left\|(\mathbf{1} - \tilde{\mathcal{I}}^{(1)}_h)P^0_{I_n}\bar{u}\right\|_K dt,$$

andererseits aber aufgrund der Projektionseigenschaft (2.25):

$$\begin{aligned}
\sum_{K\in\tilde{K}}\left(r_1, P^0_{I_n}\bar{u} - \tilde{\mathcal{I}}_{\mathcal{T}}\bar{u}\right)_{K\times I_n} &= \left(r_1, P^0_{I_n}\bar{u} - \tilde{\mathcal{I}}_{\mathcal{T}}\bar{u}\right)_{\tilde{K}\times I_n} \\
&= \left(r_1, (\mathbf{1} - \tilde{\mathcal{I}}^{(1)}_h)P^0_{I_n}\bar{u}\right)_{\tilde{K}\times I_n} \\
&= \left(r_1 - \tilde{r}_1, (\mathbf{1} - \tilde{\mathcal{I}}^{(1)}_h)P^0_{I_n}\bar{u}\right)_{\tilde{K}\times I_n} \\
&= \int_{I_n}\left(r_1 - \tilde{r}_1, (\mathbf{1} - \tilde{\mathcal{I}}^{(1)}_h)P^0_{I_n}\bar{u}\right)_{\tilde{K}} dt,
\end{aligned}$$

wobei $\tilde{r}_1$ ein beliebiges Element aus $P_0(\tilde{K}\times I_n)$ sei. Wir wollen $\tilde{r}_1$ als Interpolation von r_1 auf diesen Raum auffassen; da dieser jedoch nicht auf K, sondern auf $\tilde{K}$ definiert ist, kann man $\tilde{r}_1$ nicht lokal aus r_1 berechen und es kann auf K keine Abschätzungen der Norm von $r_1 - \tilde{r}_1$ geben; das Integrationsgebiet mußte daher auf den ganzen Patch erweitert werden.

Nimmt man nun eine näherungsweise Gleichverteilung der Fehler auf den einzelnen Zellen eines Patches an, d. h.

$$\left(r_1, P^0_{I_n}\bar{u} - \tilde{\mathcal{I}}_{\mathcal{T}}\bar{u}\right)_{K\times I_n} \approx 2^{-d}\sum_{K\in\tilde{K}}\left(r_1, P^0_{I_n}\bar{u} - \tilde{\mathcal{I}}_{\mathcal{T}}\bar{u}\right)_{K\times I_n},$$

so gilt

$$\left|\left(r_1, P^0_{I_n}\bar{u} - \tilde{\mathcal{I}}_{\mathcal{T}}\bar{u}\right)_{K\times I_n}\right| \leq \int_{I_n}\min\left\{\left\|r_1\right\|_K, 2^{-d}\left\|r_1 - \tilde{r}_1\right\|_{\tilde{K}}\right\}\left\|(\mathbf{1} - \tilde{\mathcal{I}}^{(1)}_h)P^0_{I_n}\bar{u}\right\|_K dt,$$

und wir erhalten mit dem Bramble-Hilbert-Lemma

$$\begin{aligned}
\left\|r_1 - \tilde{r}_1\right\|_{\tilde{K}} &\leq Ch_{\tilde{K}}\left\|\nabla r_1\right\|_{\tilde{K}}, \\
\left\|(\mathbf{1} - \tilde{\mathcal{I}}^{(1)}_h)P^0_{I_n}\bar{u}\right\|_K &\leq Ch^2_{\tilde{K}}\left\|\nabla^2\bar{u}\right\|_{\tilde{K}}.
\end{aligned}$$

Damit gilt insgesamt

$$\begin{aligned}
\left|\left(r_1, \bar{u} - \tilde{\mathcal{I}}_{\mathcal{T}}\bar{u}\right)_{K\times I_n}\right| \leq\ & C_1k_n\left(\min\left\{\left\|r_1\right\|_{K\times I_n}, k_n\left\|\partial_t r_1\right\|_{K\times I_n}\right\}\left\|\partial_t\bar{u}\right\|_{K\times I_n}\right) \\
& + C_2h^2_{\tilde{K}}\left(\min\left\{\left\|r_1\right\|_{K\times I_n}, h_{\tilde{K}}\left\|\nabla r_1\right\|_{\tilde{K}\times I_n}\right\}\left\|\nabla^2\bar{u}\right\|_{\tilde{K}\times I_n}\right).
\end{aligned}$$

Die Auswertung der anderen Terme in (2.23) erfolgt analog; da man außer im Summanden mit den Sprüngen auf den Zellkanten, bei dem durch die Verwendung des Spursatzes eine Vermischung

der Potenzen verursacht wird, jeweils zwei Terme bekommt, von denen einer nur k-, der andere nur h-Potenzen enthält, läßt sich durch Vergleich feststellen, ob eine Verfeinerung des Orts- oder des Zeitgitters die Genauigkeit mehr erhöhen wird.

Insgesamt erhält man als Fehlerschätzer die folgenden Ausdrücke:

$$|J(\mathbf{e})| \leq C \sum_n \sum_{K \in \mathbb{T}^n} \eta_{K,n},$$

$$\eta_{K,n} = h_K^d k_n \sum_{s=1}^{3} \rho_{K,n}^{s,k} \omega_{K,n}^{s,k} + \rho_{K,n}^{s,h} \omega_{K,n}^{s,h}, \tag{2.26}$$

mit den Residuen und Gewichten

$$\begin{aligned}
\rho_{K,n}^{1,k} &= h_K^{-d/2} k_n^{1/2} \quad \min\left\{ \|r_1\|_{K\times I_n}, k_n \|\partial_t r_1\|_{K\times I_n} \right\}, \\
\omega_{K,n}^{1,k} &= h_K^{-d/2} k_n^{-1/2} \|\partial_t \bar{u}\|_{K\times I_n}, \\
\\
\rho_{K,n}^{1,h} &= h_K^{2-d/2} k_n^{-1/2} \quad \min\left\{ \|r_1\|_{K\times I_n}, h_{\bar{K}} \|\nabla r_1\|_{\bar{K}\times I_n} \right\}, \\
\omega_{K,n}^{1,h} &= h_K^{-d/2} k_n^{-1/2} \left\|\nabla^2 \bar{u}\right\|_{\bar{K}\times I_n} \\
\\
\rho_{K,n}^{2,k} &= h_K^{-d/2} k_n^{1/2} \quad \min\left\{ \|r_2\|_{K\times I_n}, k_n \|\partial_t r_2\|_{K\times I_n} \right\}, \\
\omega_{K,n}^{2,k} &= h_K^{-d/2} k_n^{-1/2} \|\partial_t \bar{v}\|_{K\times I_n}, \\
\\
\rho_{K,n}^{2,h} &= h_K^{2-d/2} k_n^{-1/2} \quad \min\left\{ \|r_2\|_{K\times I_n}, h_{\bar{K}} \|\nabla r_2\|_{\bar{K}\times I_n} \right\}, \\
\omega_{K,n}^{2,h} &= h_K^{-d/2} k_n^{-1/2} \left\|\nabla^2 \bar{v}\right\|_{\bar{K}\times I_n}, \\
\\
\rho_{K,n}^{3,k} &= h_K^{-d/2-1/2} k_n^{1/2} \quad \min\left\{ \|[\mathbf{n}\cdot a\nabla u_h]\|_{\partial K\times I_n}, k_n \|[\partial_t \mathbf{n}\cdot a\nabla u_h]\|_{\partial K\times I_n} \right\}, \\
\omega_{K,n}^{3,k} &= h_K^{-d/2} k_n^{-1/2} \left(\|\partial_t \bar{v}\|_{K\times I_n} + h_K \|\partial_t \nabla \bar{v}\|_{K\times I_n} \right), \\
\\
\rho_{K,n}^{3,h} &= \frac{1}{2} h_K^{3/2-d/2} k_n^{-1/2} \|[\mathbf{n}\cdot a\nabla u_h]\|_{\partial K\times I_n}, \\
\omega_{K,n}^{3,h} &= \omega_{K,n}^{2,h}.
\end{aligned}$$

Die Skalierung der Gewichte ω und der ganzen Summe wurden wieder wie oben gewählt, d. h. so, daß im Limes $h, k \to 0$ die ω gegen eine kontinuierliche Funktion und die Summe gegen ein Integral gehen. Bei der Zusammenfassung von h-Potenzen wurde kein Unterschied zwischen h_K und $h_{\bar{K}}$ gemacht, da sich die beiden nur um einen Faktor zwei unterscheiden; diese Konstante wurde in die ohnehin unbekannte Konstante C im Fehlerschätzer absorbiert. Eine Implementation in einem Programm wird diesen Unterschied jedoch beachten.

Bei der Herleitung dieses Fehlerschätzers, der praktisch unverändert aus [19] übernommen werden konnte, ist noch nicht berücksichtigt, daß bei Vergöberung auch im Inneren der Raum-Zeit-Zellen $K \times I_n$ unstetige Gradienten auftreten können. In diesen Fällen sind schon in (2.23) zusätzliche Spungterme einzuführen und die Unstetigkeitsmannigfaltigkeiten aus den Gebietsintegralen herauszunehmen.

In bisherigen Arbeiten wurden diese Modifikationen nicht vorgenommen oder übersehen; in der Praxis wird sich dieser Mangel jedoch meist nicht wesentlich bemerkbar machen, da die Vergröberung immer nur recht wenige Zellen betrifft und darüberhinaus auch nur die, auf denen die Lösung im wesentlichen glatt ist, d. h. auf denen die Sprungterme klein sind. Die Implementation dieser zusätzlichen Sprungterme in einem Computerprogramm wird, ebenso wie die Verwendung der nichtlokalen Interpolation bei der Auswertung der exakten Fehleridentität, wie sie in Abschnitt 2.4.3 dargestellt wurde, erhebliche Schwierigkeiten bereiten.

2.4.5 Vergleich der beiden Wege der Auswertung

Im Programm, mit dem die Methoden dieser Arbeit umgesetzt wurden, wurde der Weg der Auswertung des Fehlerschätzers mit einer genauer gerechneten dualen Funktion gewählt. Dieser hat gegenüber der Abschätzung mit dem Bramble-Hilbert-Lemma die folgenden Vorteile:

- Im Prinzip sollte es möglich sein, die Fehleridentität (2.22) quantitativ praktisch exakt auszuwerten (die Genauigkeit des Fehlerschätzer sollte in etwa der selben Größenordnung liegen wie der Fehler auf einem einmal verfeinerten Gitter). Im Gegensatz dazu sind bei der Abschätzung mit dem Bramble-Hilbert-Lemma die Interpolationskonstanten noch unbekannt. Durch einige analytische Überlegungen kann man zeigen, daß sie für quadratische Zellen zwischen 0.1 und 1 liegen; durch Probieren in diesem Bereich können sie dann so gewählt werden, daß sie bei einfachen Problemen, bei denen die exakte Lösung bekannt ist, den Effizienzindex des Fehlerschätzers (d. h. das Verhältnis von geschätztem zu echtem Fehler) auf Werte nahe eins bringen. Allerdings ist unklar, ob das dann auch bei realen Problemen gilt, bei denen verzerrte Zellen, einspringende Ecken usw. vorkommen.

- Damit einhergehend ist die Möglichkeit, das Vorzeichen des Fehlers anzugeben. In einigen praxisrelevanten Fällen kann das von Vorteil sein, beispielsweise wenn mit einer Simulation gezeigt werden soll, daß ein vorgegebener Maximalwert nicht überschritten wird. In der Praxis gibt der Fehlerschätzer das Vorzeichen des Fehlers bereits auf recht groben Gittern immer richtig ist.

Den angeführten Vorteilen stehen jedoch die folgenden Nachteile gegenüber:

- In der Praxis hat es sich gezeigt, daß scharfe Fehlerwerte nur mit großem Aufwand zu erreichen sind. Dazu tragen sowohl mathematische Schwierigkeiten, insbesondere die Linearisierung nichtlinearer Fehlerfunktionale als auch praktische Gründe bei, zu letzterem vor allem die Implementation von Interpolationen und Vergleichen über mehrere Zeitschritte mit unterschiedlichen Gittern.

- Auf den ersten Blick erschien die Implementation des Fehlerschätzers auf diese Art einfacher. Der Grund dafür ist, daß bei der Auswertung der Normen in den Gewichten $\omega^i_{K,n}$ Ableitungen auftreten, die höher als der Ansatzgrad sind, so daß sie durch Differenzenquotienten auf dem K umgebenden Patch von Zellen approximiert werden müssen; besonders störend sind in diesem Zusammenhang die hohen Zeitableitung bei der Verwendung von Elementen mit höherem Ansatzgrad im Ort.

  Diese Einschätzung muß jedoch, zumindest für zeitabhängige Probleme, aufgrund der Nichtlokalität der Interpolation revidiert werden, die allein für einen erheblichen Teil der Komplexität der Implementation verantwortlich ist.

- Schließlich ist der ganz erheblich höhere numerische Aufwand zu beachten. Die Berechnung der dualen Lösung mit einem quadratischen statt einem linearen Ansatz benötigt rund vier Mal so viel Rechenzeit und -speicher und dominiert damit die verwendeten Ressourcen für das primale Problem um ein Mehrfaches. Dieses Verhältnis bessert sich zwar etwas für höhere Ansatzgrade, allerdings ist nicht klar, ob es immer ausreicht, die duale Lösung dann noch mit nur um eins höherem Ansatzgrad zu rechnen. Der Grund ist der mit wachsendem r kleiner werdende Abstand zwischen $Q^r(\mathbb{T})$ und $Q^{r+1}(\mathbb{T})$ im Vergleich zum Abstand zwischen $Q^r(\mathbb{T})$ und dem kontinuierlichen Lösungsraum; wird das Verhältnis der Abstände zu klein, wird die Ersetzung

  $$(\mathbf{1} - \mathcal{I}_h^{(r)})\bar{\mathbf{w}} \qquad \longrightarrow \qquad (\mathbf{1} - \mathcal{I}_h^{(r)})\bar{\mathbf{w}}_h$$

  fragwürdig.

  Wie in Kapitel 5 gezeigt, ist der Aufwand in einigen Fällen trotzdem nur in etwa der selben Größenordnung wie bei der Verwendung eines einfachen Energiefehlerschätzers oder bei globaler Verfeinerung, wofür man immerhin eine Aussage über die Größe des Fehler bekommt.

Der Aufwand ließe sich aber erheblich senken, wenn man die duale Lösung mit dem gleichen Ansatzgrad rechnen würde wie das primale Problem. In der Praxis ließe sich der zusätzliche Aufwand wohl kaum rechtfertigen.

Abschließend muß festgestellt werden, daß die Auswertung mit einer mit höherem Ansatzgrad gerechneten Lösung die Erwartung nicht erfüllen konnte. Für zukünftige Probleme scheint die Verwendung des mit dem BRAMBLE-HILBERT-Lemma hergeleiteten Fehlerschätzers daher sinnvoller. Dafür sprechen im wesentlichen die folgenden Argumente:

- In der Praxis hat es sich erwiesen, daß es oft nur wichtig ist, die Größenordnung des Fehlers zu kennen; in kritischen Fällen wird man ohnehin ein Mehrfaches des geschätzten oder berechneten Fehlers als Sicherheitsmarge nehmen, so daß eine Ungenauigkeit durch die Wahl der Interpolationskonstanten in vielen Fällen akzeptabel ist.

- Für Probleme aus dem Umfeld der Wellengleichung, aber auch für viele andere Probleme ohne Dämpfung oder mit starken Nichtlinearitäten ist der Aufwand zur Lösung mit „vernünftiger“ Genauigkeit an der Grenze des mit der heutigen Rechnertechnik Erreichbaren. In diesen Fällen lassen sich die erheblich höheren Ressourcen, die hier zur Bestimmung der dualen Lösung verwendet wurden, weder rechtfertigen noch in vielen Fällen bereitstellen.

  Ein für die Praxis wesentlich wichtigeres Kriterium ist, ob ein Problem durch die Verwendung eines Fehlerschätzers auf eine höhere Genauigkeit bzw. überhaupt gelöst werden kann. Die Erfahrung bei der Arbeit am Programm hat gezeigt, daß die erzeugten Gitter sich kaum unterschieden, wenn statt dem beschriebenen Schätzer eine nur teilweise implementierte Version (zum Beispiel nur die Gebietsresiduen ohne die Sprungterme) verwendet wurde. Die exaktere Auswertung des Fehlerschätzers wird daher nur unwesentlich effizientere Gitter produzieren können als bei der Verwendung des mit dem BRAMBLE-HILBERT-Lemmas hergeleiteten Schätzers. Dagegen zeigt die Erfahrung, daß die Verwendung eines dualen Problems eine teilweise drastische Reduktion der benötigten Zellen bewirkte, wofür hauptsächlich die Gewichtung eines Raum-Zeit-Punkts mit seinem Beitrag zum Endergebnis verantwortlich ist; in vielen Fällen ist bei hyperbolischen Gleichungen dieses Gewicht Null, namentlich wenn der Punkt außerhalt des Einflußgebiets des Zielfunktionals liegt.

- Letztlich ist der Aufwand zur Implementation des in dieser Arbeit verfolgten Weges, wenigstens für zeitabhängige Probleme, mindestens genauso hoch, wahrscheinlich sogar höher als der andere vorgeschlagene Weg.

2.4.6 Bewertung der Linearisierung nichtlinearer Zielfunktionale

Bei der Fehlerkontrolle waren wir an einer Auswertung $\mathcal{J}(\mathbf{w})$ der Lösung interessiert. $\mathcal{J}(\cdot)$ konnte dabei ein Punktwert, ein Integral, die Energie oder ein beliebiges anderes, nicht notwendigerweise lineares Funktional sein. Abzuschätzen war der Fehler

$$\mathcal{J}(\mathbf{w}) - \mathcal{J}(\mathbf{w}_h),$$

den man durch die Auswertung der numerischen anstelle der exakten Lösung macht. Dazu mußte ein *lineares* Funktional $J(\cdot)$ so gewählt werden, daß

$$J(\mathbf{e}) = J(\mathbf{w} - \mathbf{w}_h) = J(\mathbf{w}) - J(\mathbf{w}_h) = \mathcal{J}(\mathbf{w}) - \mathcal{J}(\mathbf{w}_h)$$

gilt. Die richtige Wahl dafür ist

$$J(\mathbf{w} - \mathbf{w}_h) = \left[\int_0^1 \frac{\delta \mathcal{J}}{\delta \mathbf{w}}(s\mathbf{w} + (1-s)\mathbf{w}_h) ds \right] (\mathbf{w} - \mathbf{w}_h), \tag{2.27}$$

d. h. die Differenz zweier Funktionalwerte ist die Intervallänge mal der mittleren Ableitung des Funktional im Intervall.[11] Da die exakte Lösung nicht bekannt ist, wird als rechte Seite des dualen Problems das gestörte Funktional

$$\tilde{J}(\mathbf{w}-\mathbf{w}_h) = \left[\int_0^1 \frac{\delta \mathcal{J}}{\delta \mathbf{w}}(\mathbf{w}_h)ds\right](\mathbf{w}-\mathbf{w}_h) = \frac{\delta \mathcal{J}}{\delta \mathbf{w}}(\mathbf{w}_h, \mathbf{w}-\mathbf{w}_h) \tag{2.28}$$

verwendet.[12] Die Tilde kennzeichne im folgenden Größen, die mit diesem gestörten Funktional gewonnen wurden.

Die Differenz der beiden Funktionale ist nach dem gleichen Verfahren wie oben durch

$$\begin{aligned} J(\cdot)-\tilde{J}(\cdot) &= \left[\int_0^1 \frac{\delta \mathcal{J}}{\delta \mathbf{w}}(s\mathbf{w}+(1-s)\mathbf{w}_h) - \frac{\delta \mathcal{J}}{\delta \mathbf{w}}(\mathbf{w}_h)ds\right](\cdot) \\ &= \left[\int_0^1 \left\{\int_0^1 \frac{\delta^2 \mathcal{J}}{\delta u^2}\left(r(s\mathbf{w}+(1-s)\mathbf{w}_h)+(1-r)\mathbf{w}_h\right)dr\right\}(s\mathbf{w}+(1-s)\mathbf{w}_h-\mathbf{w}_h)ds\right](\cdot) \end{aligned}$$

gegeben. Verwendet man die Linearität von $\frac{\delta^2 \mathcal{J}}{\delta u^2}$ in seinem zweiten Argument und substituiert $q = rs$ im inneren Integral, so erhält man für die Abweichung zwischen dem gesuchten Fehlerfunktional und dem aufgrund der Linearisierung fälschlicherweise erhaltenen Wert

$$\begin{aligned} \Delta J(\cdot) = J(\cdot)-\tilde{J}(\cdot) &= \left[\int_0^1 \left\{\int_0^1 \frac{\delta^2 \mathcal{J}}{\delta u^2}\left(r(s\mathbf{w}+(1-s)\mathbf{w}_h)+(1-r)\mathbf{w}_h\right)dr\right\}(\mathbf{w}-\mathbf{w}_h)s\ ds\right](\cdot) \\ &= \left[\int_0^1 \left\{\int_0^s \frac{\delta^2 \mathcal{J}}{\delta u^2}\left(q\mathbf{w}+(1-q)\mathbf{w}_h\right)dq\right\}(\mathbf{w}-\mathbf{w}_h)ds\right](\cdot) \\ &= \int_0^1 \int_0^s \frac{\delta^2 \mathcal{J}}{\delta u^2}\left(q\mathbf{w}+(1-q)\mathbf{w}_h, \mathbf{w}-\mathbf{w}_h, \cdot\right)dq\ ds. \end{aligned}$$

$\Delta J(\cdot)$ ist nun ein lineares Funktional, so daß man wie schon in Abschnitt 2.4.2 gemäß dem RIESZschen Satz ein Element $\Delta\mathbf{j}(\mathbf{w};\mathbf{x},t)$ des Dualraums mit $\Delta J(\psi) = \int \Delta\mathbf{j}(\mathbf{w},\mathbf{w}_h;\mathbf{x},t)\psi(\mathbf{x},t)\,dxdt$ assoziieren kann. Mit $J(\cdot)$ und dem gestörten Funktional $\tilde{J}(\cdot)$ sind jeweils duale Lösungen $\bar{\mathbf{w}}$ und $\tilde{\bar{\mathbf{w}}}$ verbunden, deren Differenz aufgrund der Linearität der Wellengleichung durch eine GREENsche Funktion $G^*(\mathbf{x},t;\mathbf{x}',t')$ zum dualen Operator von $\Delta\mathbf{j}(\mathbf{w};\mathbf{x},t)$ abhängt:

$$\Delta\bar{\mathbf{w}} = \bar{\mathbf{w}} - \tilde{\bar{\mathbf{w}}} = \int_\Omega \int_0^T G^*(\mathbf{x},t;\mathbf{x}',t')\Delta\mathbf{j}(\mathbf{x}',t')\ dx'\ dt'$$

mit

$$\Delta\mathbf{j}(\mathbf{x},t) = \left\{\int_0^1 \int_0^s \frac{\partial^2(\mathbf{j}(\mathbf{w})\cdot\mathbf{w})}{\partial\mathbf{w}^2}(q\mathbf{w}+(1-q)\mathbf{w}_h;\mathbf{x},t)\ dq\ ds\right\}\cdot\left(\mathbf{w}(\mathbf{x},t)-\mathbf{w}_h(\mathbf{x},t)\right),$$

wobei $\mathbf{j}$ durch $\mathcal{J}(\psi) = \int \mathbf{j}(\psi;\mathbf{x},t)\cdot\psi(\mathbf{x},t)\ dx\ dt$ definiert sei. Mit $\Delta\mathbf{j}$ gilt für die Abweichung zwischen der gewünschten Fehleridentität (2.22) und der gestörten

$$\begin{aligned} \mathcal{E}_{K,n} - \tilde{\mathcal{E}}_{K,n} = &-\left(\rho\mathbf{w}_{h,t}, \Delta\bar{\mathbf{w}}-\Delta\bar{\mathbf{w}}_h\right)_{K\times I_n} + \left(\rho v_h, \Delta\bar{u}-\Delta\bar{u}_h\right)_{K\times I_n} \\ &+ \left(\nabla\cdot a\nabla u_h, \Delta\bar{v}-\Delta\bar{v}_h\right)_{K\times I_n} - \frac{1}{2}\left([\mathbf{n}\cdot a\nabla u_h], \Delta\bar{v}-\Delta\bar{v}_h\right)_{\partial K\times I_n} \end{aligned} \tag{2.29}$$

[11] Die Ableitung eines Funktionals $F(\psi)$ an einer Stelle u_0 wird durch $\frac{\delta F}{\delta u}(u_0)(\psi)$ oder $\frac{\delta F}{\delta u}(u_0,\psi)$ bezeichnet; sie ist ein *lineares* Funktional in ψ. Ebenso ist die zweite Ableitung $\frac{\delta^2 F}{\delta u^2}(u_0,u_1,\psi)$ das in u_1 und ψ lineare Funktional der zweiten Ableitung an den Punkten u_0 und u_1; die zweite Ableitung bezieht sich dabei auf das erste Argument von $\frac{\delta F}{\delta u}(u_0,\psi)$.

[12] In einigen Fällen kann man das exakte Funktional (2.27) verwenden, selbst wenn die kontinuierliche Lösung $\mathbf{w}$ nicht bekannt ist; ein Beispiel dafür ist in Abschnitt 5.2 gegeben. Diese Fälle stellen aber in gewissem Sinne pathologische Situationen dar, in denen die Verwendung des Fehlerschätzers zur Gitterverfeinerung genauer zu diskutieren wäre; darauf sei jedoch in dieser Arbeit verzichtet.

für alle $\Delta\bar{\mathbf{w}}_h$. Die Abweichung in der Fehleridentität $\mathcal{E}_{K,n} - \tilde{\mathcal{E}}_{K,n}$ hängt daher vom Fehler $\mathbf{e} = \mathbf{w} - \mathbf{w}_h$ linear ab, falls das Zielfunktional quadratisch ist, quadratisch falls $\mathcal{J}$ kubisch ist, usw. Sie ist damit von einer Ordnung höher im Fehler als das Zielfunktional; wie in Abschnitt 3 gezeigt wird, ist der Linearisierungsfehler aber trotzdem nicht immer vernachlässigbar.

Eine weitere Abschätzung von $\mathcal{E}_{K,n} - \tilde{\mathcal{E}}_{K,n}$ wäre hier wünschenswert, ist aber schwierig, da sowohl $\mathbf{j}$ als auch G^* im allgemeinen singulär sind und man nur schwer zu Aussagen in interessanten Normen kommt. Ein in diesem Fall gangbarer Weg wäre, die Normen über die FOURIER-Transformierten dieser Größen zu gewinnen, da diese in den meisten Fällen mehr Regularität besitzen.

In der Praxis ist neben der Abschätzung des Fehlers, für die der hergeleitete lineare Zusammenhang mit dem Linearisierungsfehler gilt, wichtig, durch den Fehlerschätzer effiziente Gitter zu erzeugen. Dieser Prozeß ist leider nichtlinear, was dadurch zustande kommt, daß die Fehlerindikatoren für die einzelnen Zellen der Größe nach sortiert und die Zellen in dieser Reihenfolge verfeinert werden. Das führt gelegentlich dazu, daß Zellen verfeinert werden, die weit ab von dem Gebiet liegen, das für das Zielfunktional wichtig ist, selbst dann, wenn der Fehlerschätzer nur mäßig schlecht den tatsächlichen Fehlerbeitrag approximiert. In diesen Fällen ist das Gitter für den nächsten Durchlauf nicht besser als für den aktuellen, so daß der Prozeß aus Gitterverfeinerung und besserer Linearisierung nicht konvergiert; die erzeugten Lösungen sind häufig unbrauchbar.

Die Zielfunktionale im Kontext der Wellengleichung sind im allgemeinen entweder linear oder quadratisch in der Lösung; der lineare Fall stellt keine Probleme dar, während der quadratische Fall einer einfachen Variante der obigen Analyse zugänglich ist, da die Integrale in der Darstellung von $\Delta J(\cdot)$ berechenbar sind. Wir betrachten als Beispielfall den Energiefluß durch eine Kurve:

$$\mathcal{J}(\mathbf{w}) = \int_0^T \int_{\mathcal{C}} v(a\nabla u) \cdot \mathbf{n} \; ds \; dt.$$

Für diesen gilt mit $\mathbf{t} = (\varphi, \psi)$:

$$\frac{\delta^2 \mathcal{J}}{\delta \mathbf{w}^2}\,(q\mathbf{w} + (1-q)\mathbf{w}_h, \mathbf{w} - \mathbf{w}_h, \mathbf{t}) = \int_0^T \int_{\mathcal{C}} \{\psi(a\nabla(u - u_h)) + (v - v_h)(a\nabla\varphi)\} \cdot \mathbf{n} \; ds \; dt$$

und damit für die Abweichung des Fehlerschätzers vom Wert bei richtiger Linearisierung

$$\Delta J(\mathbf{e}) = \int_0^T \int_{\mathcal{C}} (v - v_h)(a\nabla(u - u_h)) \cdot \mathbf{n} \; ds \; dt. \tag{2.30}$$

Diese Größe wird in Abschnitt 3.4.2 näherungsweise ausgewertet; dort wird gezeigt, daß in einigen praxisrelevanten Fällen die Linearisierung so schlecht sein kann, daß der Linearisierungsfehler ΔJ in der gleichen Größenordnung wie der Wert des Fehlerfunktionals liegen kann. Das hat im allgemeinen zur Folge, daß der Prozeß aus Gitterverfeinerung und Fehlerschätzung nicht konvergiert, da die Gitterverfeinerung keine genauere Lösung des primalen Problems und damit auch keine bessere Linearisierung im nächsten Verfeinerungsschritt bewirkt.

2.4.7 Bewertung der numerischen dualen Lösung

Unabhängig davon, ob die Fehleridentität nach dem in Abschnitt 2.4.3 vorgeschlagenen Verfahren oder wie in Abschnitt 2.4.4 durch Abschätzung mit dem BRAMBLE-HILBERT-Lemma ausgewertet wird, muß die exakte duale Lösung $\bar{\mathbf{w}}$ durch eine auf numerischem Weg erhaltene Funktion ersetzt werden. Im ersten Fall war es nötig, die numerische duale Lösung mit einem höheren Ansatzgrad zu rechnen, im allgemeinen wurde er um eins größer als der Ansatzgrad des primalen Problems gewählt; im zweiten Fall ist ein beliebiger Ansatzgrad möglich, meistens wählt man jedoch den selben wie für das primale Problem.

Die Ersetzung der exakten durch eine numerische duale Lösung bewirkt einen zusätzlichen Fehler bei der Auswertung der Fehlerschätzer. Während klar ist, daß dieser Fehler von höherer Ordnung als die Fehleridentität ist, die wir gerade auswerten wollen, ist das Größenverhältnis

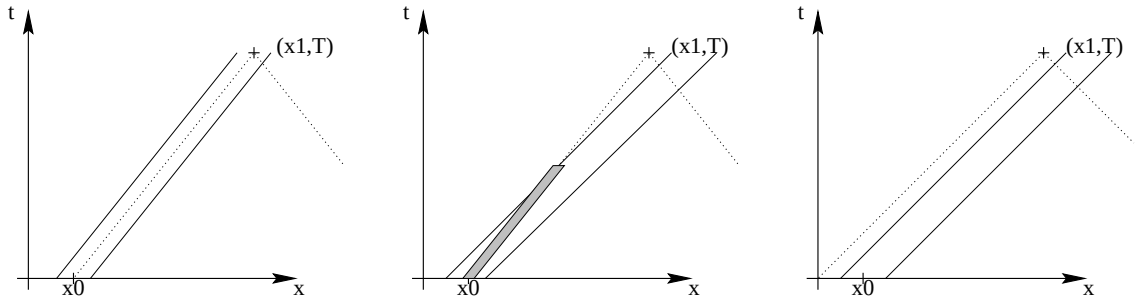

Abbildung 2.6: *Darstellung der Problematik bei der Ersetzung der exakten dualen Lösung* $\bar{\mathbf{w}}$ *durch eine numerisch gewonnene Lösung. Links exakte primale und duale Lösungen; in der Mitte exakte duale, aber numerische primale Lösung; rechts numerische primale und duale Lösung. Weitere Erläuterungen im Text.*

dieser beiden Werte nicht offensichtlich. Im allgemeinen darf die Ersetzung nicht völlig unkritisch durchgeführt werden, wie sich an folgendem Beispiel demonstrieren läßt:

Wir wollen annehmen, daß die Anfangsbedingungen eine Welle gegebener Ausdehnung von x_0 aus in eine vorgegebene Richtung loslaufen lassen. Der Einfachheit halber beschränken wir uns auf nur eine Raumdimension, so daß wir bei konstanten Koeffizienten den Weg der Welle im $x-t$-Diagramm wie in Abbildung 2.6 links mit durchgezogenen Linien darstellen können. Das Zielfunktional sei die Punktauswertung in (x_1, T); die exakte duale Lösung ist dann durch die avancierte GREENsche Funktion gegeben, deren Träger das durch die gepunkteten Linien begrenzte Gebiet ist. In höheren Dimensionen ist sie jedoch nur entlang der gepunkteten Linien von Null verschieden, was wir im folgenden Annehmen wollen.

In der Mitte der Abbildung ist der Fall dargestellt, daß die primale Lösung auf numerischem Wege gerechnet wurde, die duale Lösung aber nach wie vor exakt bekannt ist. Aufgrund des in Abschnitt 4.5.2 geschilderten Problems ist die numerische Ausbreitungsgeschwindigkeit höher als die exakte; die Steigung der Begrenzungslinien des Trägers der primalen Lösung ist daher größer als im exakten Fall. Verfeinerung wird nun dort stattfinden, wo sich die Träger von numerischer primaler und exakter dualer Lösung überlappen, d. h. im schattierten Bereich. Die Verfeinerung wird dazu führen, daß der linke Teil der Welle im nächsten Durchlauf näher an der exakten Lösung ist, also insbesondere daß die Abweichung der Ausbreitungsgeschwindigkeit geringer ist. Der rechte Teil wird nicht verfeinert und behält damit seine falsche Geschwindigkeit, was allerdings auch nicht schlimm ist, da er ohnehin nicht zum Zielfunktional beiträgt.

Es ist offensichtlich, daß bereits in diesem Fall mit exakter dualer Lösung die Konvergenz gegen die richtige primale Lösung recht langsam gehen kann, da nur etwa in der ersten Hälfte des Zeitintervalls $(0, T)$ überhaupt Verfeinerung auftritt. Im besten Fall ist die numerische Lösung dort exakt, für größere Zeiten jedoch ist die Ausbreitungsgeschwindigkeit im nächsten Durchlauf ebenso falsch wie im ersten und man braucht mehrere Schritte, bis überhaupt erst das ganze Zeitintervall verfeinert ist. Je größer die Abweichung zwischen exakter und numerischer Ausbreitungsgeschwindigkeit ist, desto mehr Schritte sind nötig, um den ganzen Weg zwischen Ursprungs- und Auswertungspunkt zu verfeinern.

Das Problem verschlimmert sich noch, wenn auch die duale Lösung numerisch gewonnen werden muß (rechter Teil der Abbildung): dann ist auch die Ausbreitungsgeschwindigkeit der dualen Lösung falsch und die Träger der beiden Funktionen überlappen überhaupt nicht mehr in Bereichen, die für das Zielfunktional relevant sind. In diesem Fall wird der Fehlerschätzer einen Wert Null zurückgeben und alle Zellen zu Zeiten vergröbern, die vor der Zeit liegen, wo sich der rechte Teil der dualen Lösung und die primale Lösung überschneiden. Es kann hier keine Konvergenz gegen die richtige Lösung geben.

Aufgrund der geschilderten Problematik ist es notwendig, die Verfeinerung mit dem dualen Fehlerschätzer erst dann zu beginnen, wenn schon gute Näherungen für die primale Lösung vor-

handen sind und wenn das Gitter gut genug zur Berechnung der dualen Lösung ist. In der Praxis zeigt es sich, daß es für letzteres nicht ausreicht, die duale Lösung auf der Schnittmenge der Träger der beiden Lösungen gut zu approximieren, sondern daß es nötig ist, das Gitter auch geeignet an die duale Lösung anzupassen (vgl. Abschnitt 4.5.1).

Daneben zeigt sich allerdings auch, daß in praxisnahen Beispielen quantitativ richtige Fehlerschätzung nicht möglich ist. Die berechneten Fehlerschätzer zeigen im allgemeinen eine nur schwache Korrelation mit dem tatsächlichen Fehler und liegen oft um ein bis zwei Größenordnungen daneben; allerdings sind die erzeugten Gitter sehr effektiv, was den verwendeten Ansatz rechtfertigt. Die schlechte Approximation des echten Fehlers durch den Fehlerschätzer ist natürlich Auswirkung der Ersetzung $\bar{\mathbf{w}} \to \bar{\mathbf{w}}_h^{(r+1)}$ und wird verstärkt durch die gegenüber kleinen Störungen instabile Auswertung der Fehleridentität, bei der ein Integral über zwei stark oszillierende Funktionen, namentlich das Residuum und $(1 - \mathcal{I}_h)\bar{\mathbf{w}}$, ausgeführt werden muß. Die Genauigkeit des Schätzers kann unter Umständen verbessert werden, wenn bei seiner Berechnung die GALERKIN-Orthogonalität nicht ausgenutzt wird, so daß die Integrale nur über das oszillierende Residuum, aber über eine glatte Funktion $\bar{\mathbf{w}}$ ausgeführt werden können; die Berechnung des Verfeinerungskriteriums hat allerdings aus den geschilderten Gründen unter Verwendung einer Interpolation und der GALERKIN-Orthogonalität zu geschehen.

Kapitel 3

Anwendung auf die solare Atmosphäre

Die im vorigen Kapitel hergeleiteten numerischen Methoden sollen in diesem Abschnitt auf ein Beispiel aus der Physik stellarer Atmosphären angewendet werden. Oberhalb der sichtbaren Oberfläche (Photosphäre) der meisten Stern befindet sich eine Chromosphäre genannte Schicht, deren Dicke bei der Sonne rund 2.200 km beträgt. In dieser bewegt sich die Temperatur zwischen 4.000 und 20.000 K. Nach außen hin direkt anschließend steigt die Temperatur in einer kleinen Übergangszone von 100-200 km Dicke auf mehrere Millionen Kelvin an und auch die Dichte ändert sich um mehrere Grösenordnungen. In dieser, Korona genannten, Schicht bleibt die Temperatur auf Dimensionen verglichen mit der Dicke der Chromosphäre weitgehend konstant.

Eine physikalisch nur teilweise verstandene Frage ist der Mechanismus der Heizung der Korona. Wegen der kalten Schicht zwischen Sonneninnerem und Korona kommen weder Wärmeleitung noch Strahlungsheizung in Frage; andererseits kann auch aus dem interstellaren Raum kein nennenswerter Energiefluß vorhanden sein. Der tatsächliche Heizungsmechanismus besteht vermutlich neben starken Magnetfeldern und magnetohydrodynamischen Effekten aus Heizung durch Schockwellen. Akustische Wellen werden dabei auf der Sonnenoberfläche durch die dort vorhandenen Konvektionsströmungen erzeugt und bewegen sich in der Chromosphäre und der anschließenden Korona nach außen; in der dünnen Gasatmosphäre bilden sie schnell starke Schocks, die in der Lage sind, Energie zu dissipieren.

Um den Energieeintrag in die Chromosphäre durch akustische Wellen abschätzen zu können, ist es wichtig sie numerisch zu simulieren. Solche Simulationen werden seit vielen Jahren erfolgreich durchgeführt, allerdings zu einem großen Teil mit nur einer Raumdimension (vgl. [33, 25, 31, 32]) und nur in wenigen Arbeiten in zwei Dimensionen (z. B. [26]). Die Erweiterung auf mehr als eine Raumdimension wird gemeinhin als wünschenswert bezeichnet, um realistische Modelle zur Ausbreitung akustischer Wellen in der solaren Atmosphäre verwenden zu können.

Die Simulation von Phänomenen in der Sonnenatmosphäre setzt voraus, daß die physikalischen Größen Temperatur und Dichte hinreichend genau bekannt sind. Da direkte Messungen nicht möglich sind, lassen sich diese nur durch die Lösung eines inversen Problems gewinnen; dazu werden Messergebnisse mit Simulationsergebnissen für ein gegebenes Modell verglichen, um das Modell zu verbessern. Solche Rechnungen sind außerordentlich aufwendig, da die zu berücksichtigenden physikalischen Effekte sehr vielfältig sind und neben den Gasgleichungen auch Strahlung, Nichtgleichgewichtseffekte durch den gerichteten Energiefluß durch Strahlung und dynamische Effekte aufgrund der Heizung durch Schockwellen beinhalten. Die Inversion der Meßdaten geschieht im allgemeinen durch *trial and error* und zusätzliche physikalische Annahmen. Umfangreiche Darstellung zu diesem Problem sind beispielsweise in [34, 15, 10] zu finden. Im Rahmen dieser Arbeit wurde ein Modell für Dichte und Temperatur aus [32] verwendet (*bKmG – broadened Kolmogorov, modified Gauss*), bei dem dynamische Effekte aufgrund akustischer Wellen mitberücksichtigt wurden; das Frequenzspektrum der das Medium heizenden Wellen folgt dabei einer Verteilung, die aus

Turbulenzmodellen und Anpassungen an die solare Umgebung hergeleitet werden kann (vgl. [31, Abschnitt 2.2]).

Der aus diesem Modell hergeleitete Verlauf der Temperatur ist in Abbildung 3.1 beispielhaft dargestellt. Man erkennt deutlich den Beginn des steilen Anstiegs an der Grenze zwischen Chromosphäre und Korona. Die Temperatur schwankt in der Korona zwischen etwa 2 und bis zu 6 Millionen Kelvin; für die Rechnungen wurde ein Anstieg innerhalb von 200 km auf eine konstante Temperatur von 2 Millionen Kelvin angenommen, wobei darauf hingewiesen sei, daß diese Temperaturwahl das Ergebnis der Rechnungen wesentlich zu beeinflussen vermag. Aus den gegebenen Daten für Temperatur T und Dichte ρ läßt sich der Koeffizient $a(\mathbf{x})$ in der Gleichung (1.1) nach folgenden Formeln herleiten:

$$c^2 = \frac{\gamma RT}{\mu},$$
$$a = \frac{c^2}{\rho}.$$

Dabei ist c die lokale Ausbreitungsgeschwindigkeit, γ der Adiabatenkoeffizient, der aufgrund der hohen Dichte zu $\frac{5}{3}$ gewählt werden kann, R die allgemeine Gaskonstante und $\mu = 1.3$ das mittlere Molgewicht des Gases.

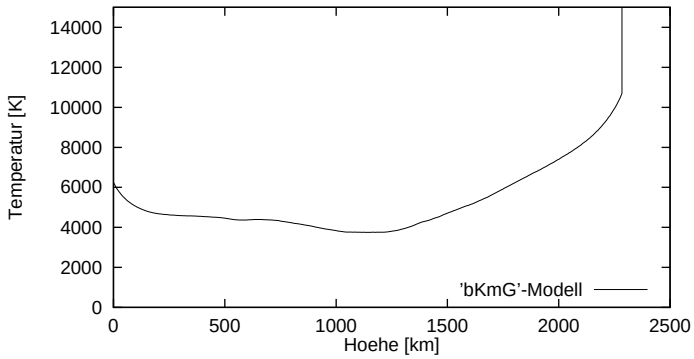

Abbildung 3.1: *Verlauf der Temperatur im bKmG-Modell (aus [32]). Nach oben hin steigt die Temperatur auf zwei Millionen Kelvin; dieser Anstieg ist nicht mehr dargestellt, um die Temperaturvariationen in der Chromosphäre noch darstellen zu können.*

Neben dem erwähnten Modell aus [32] wurden auch Rechnungen mit anderen Temperaturmodellen durchgeführt, beispielsweise Modell C aus [34] und mit den Ergebnissen aus [15] und [10]. Die Ergebnisse stimmten miteinander im Rahmen der Genauigkeit der Rechnungen weitestgehend überein. Das ist darauf zurückzuführen, daß der Energiefluß in die Korona bei linearen akustischen Wellen praktisch nur durch Reflexion an der Übergangsschicht zwischen Chromosphäre und Korona behindert wird, während die Temperaturvariationen in der Chromosphäre praktisch keinen Einfluß haben, im Gegensatz zum Fall nichtlinearer Wellen; der Temoeraturanstieg verläuft aber bei allen Modellen im wesentlichen gleich.

Wie schon in der Einleitung zu dieser Arbeit erläutert, wird eine starke Idealisierung dieser Situation angenommen, indem eine Reduktion auf nur zwei Raumdimensionen und vor allem eine Vernachlässigung aller nichtlinearer Effekte in den Gasgleichungen durchgeführt wird. Auch wurden weder Strahlungstransport noch thermische Nichtgleichgewichtsbedingungen verwendet. Die erhaltenen Ergebnisse sind daher nur sehr bedingt mit Messungen und den Ergebnissen nichtlinearer Berechnungen vergleichbar; das Ziel dieser Arbeit ist eher die Demonstration der generellen Anwendbarkeit adaptiver Methoden auf Wellenphänomene und hyperbolische Gleichungen zweiter Ordnung in der Zeit.

In den folgenden Abschnitten werden zuerst die Definition des zu rechnenden Testfalls und anschließend die verschiedenen Möglichkeiten der Auswertung der Rechnungen diskutiert. Schließlich werden die Ergebnisse der Rechnungen präsentiert.

3.1 Definition des Gebiets und der Randbedingungen

Rechengebiet und Randbedingungen werden im wesentlichen durch die physikalische Problemstellung bestimmt und sind in Abbildung 3.2 dargestellt. Als Randwertfunktion für den unteren Rand wurde

$$g_D(x,0,t) = \begin{cases} \cos\left(\frac{\pi}{2}\frac{x}{a}\right)\sin\left(\pi\frac{t}{\tau}\right) & \text{für } x < a \text{ und } t < T, \\ 0 & \text{sonst} \end{cases}$$

gewählt; dadurch läuft von der linken unteren Ecke eine Welle in das Gebiet hinein. In den Rechnungen war $a = 50km$, das heißt im Vergleich zum gesamten Gebiet sehr klein; τ wurde zu 60 Sekunden gewählt, was zu vergleichen ist mit den etwa 300 Sekunden, die eine Welle braucht um bis zur Übergangsschicht zu gelangen. Die Amplitude wurde willkürlich zu eins gewählt, da wir nur an relativen Verhältnissen interessiert sind; alle Energiewerte sind daher einheitenlos angegeben.

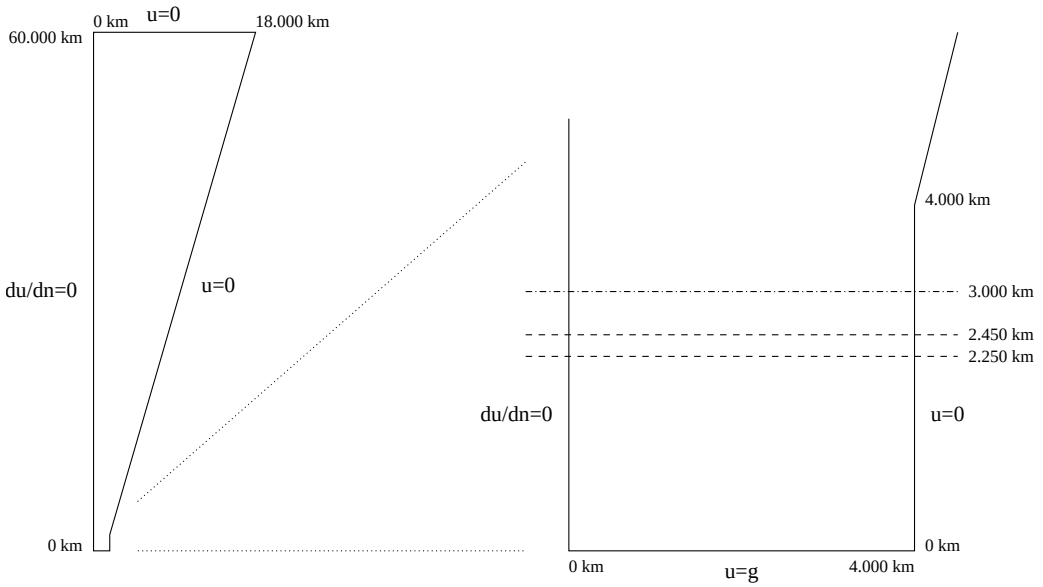

Abbildung 3.2: *Definition des Rechengebiets und der Randwerte. Links das ganze Gebiet, rechts eine Vergrößerung des unteren Bereichs. Rechts ist auch die Lage der Übergangszone zwischen etwa 2250 und 2450 km Höhe sowie die Lage der Auswertungslinie in einer Höhe von 3000 km dargestellt.*

Die Wahl der Form des Gebiets folgt im wesentlichen folgenden Überlegungen: der untere Teil des Gebiets sollte quadratisch und etwas größer als der Abstand der Sonnenoberfläche zur Übergangsschicht; die quadratische Form folgt aus der dann bestmöglichen Approximationsfähigkeit der Elemente; wäre er wesentlich größer, so würde sich das Gebiet unnötig weit nach rechts oder oben ausdehnen, was die Zellzahl erhöhen würde. Oben an das Gebiet schließt sich ein weiter „Auslaufbereich“ mit wenigen großen Zellen an, dessen Notwendigkeit im folgenden erläutert wird.

Eine optimal Behandlung des Problems würde am oberen und rechten Rand des Gebiets absorbierende Randbedingungen vorschreiben, die einer Welle gestatten, das Gebiet ohne Reflexion zu verlassen. Die Konstruktion solcher Randbedingungen ist jedoch sehr aufwendig und garantiert in den meisten Fällen keine vollständige Absorption der auftreffenden Wellen. In der Literatur finden sich im wesentlichen die folgenden Ansätze für solche Randbedingungen:

- *Darstellung in differentieller Form:* gemäß dem „klassischen“ Ansatz aus [14] lassen sich die exakten Randbedingungen als in der Zeit und im Ort nichtlokalen Integraloperator schreiben. Dieser kann als Pseudodifferentialoperator aufgefaßt und in eine TAYLOR-Reihe entwickelt werden, was eine Hierarchie von zunehmend genaueren Randbedingungen ergibt. Die ersten Glieder dieser Folge sind im folgenden für zwei Raumdimensionen und triviale Koeffizienten ($a = \rho = 1$) angegeben ($\mathbf{t}$ bezeichne einen Tangentenvektor an der Rand):

$$\begin{aligned} \frac{\partial u}{\partial \mathbf{n}} + v &= 0, \\ \frac{\partial v}{\partial \mathbf{n}} + \frac{\partial v}{\partial t} - \frac{1}{2}\frac{\partial^2 u}{\partial \mathbf{t}^2} &= 0, \\ \frac{\partial^2 v}{\partial \mathbf{n} \partial t} + \frac{\partial^2 v}{\partial t^2} - \frac{1}{4}\frac{\partial^3 u}{\partial \mathbf{t}^2 \partial \mathbf{n}} - \frac{3}{4}\frac{\partial^2 v}{\partial \mathbf{t}^2} &= 0. \end{aligned}$$

  Die Randbedingungen sind jeweils für senkrecht auftreffende Wellen ideal absorbierend, während die Absorption mit der Abweichung des Auftreffwinkels von der Senkrechten immer schlechter wird (es ist bekannt, vgl. [20], daß lokale differentielle Randbedingungen diese Eigenschaft inhärent in sich tragen und nicht für alle Einfallswinkel und Frequenzen ideal absorbierend sein können). Da die Qualität der ersten Approximation nicht sehr gut ist, die Implementation sowie die mathematische Formulierung in einem Finiten-Elemente-Ansatz mit polynomialen Ansatzfunktionen bei den folgenden Gliedern wegen der schnell höher werdenden Ableitungen aber schwierig ist, wurde auf eine Verwendung verzichtet.

- *Exakte Randbedingungen* In jüngerer Zeit wurden exakte absorbierende Randbedingungen für akustische, elastische und elektromagnetische Wellengleichungen vorgeschlagen (vgl. [18, 17]), die in der Zeit lokal sind und nur erste Ableitungen enthalten. Sie sind jedoch nichtlokal im Ort (es ist eine Entwicklung der Lösung nach Kugelfunktionen durchzuführen) und ihr größter Nachteil besteht darin, daß sie nur auf der Oberfläche einer Kugel gelten; eine Verallgemeinerung auf beliebige Gebiete erscheint außer in den Fällen ausgeschlossen, wo die Eigenfunktionen des LAPLACE-Operators auf dem Rand des Gebiets bekannt sind.

- *„Nichtreflektierende Randbedingungen“:* unter diesem Namen wird in der Ingenieursliteratur (vgl. zum Beispiel [11]) die Technik geführt, um das Gebiet herum eine oder mehrere Schichten von Zellen zu legen, in denen der Gleichung ein Dämpfungsterm hinzugefügt wird. In diesem Bereich wird jede einlaufende Welle exponentiell gedämpft. Um eine gute Absorption zu erreichen ist jedoch eine mit kleiner werdender Wellenlänge größere Schicht von Zellen notwendig; darüberhinaus ist die Dämpfung zwar exponentiell, jedoch nicht vollständig, so daß ein kleiner Teil der Welle trotzdem reflektiert wird.

Alle drei vorgeschlagenen Verfahren fordern eine Umformulierung des ursprünglichen Problems sowie erheblichen Implementationsaufwand. Es wurde in Übereinstimmung mit der Fragestellung die folgenden Lösung verfolgt: da reflektierte Wellen nur dann ein Problem darstellen, wenn sie Energie über die Auswertungslinie in 3000 km Höhe transportieren, ist lediglich Reflexion am oberen Rand problematisch. Reflexion am rechten Rand unterhalb der Übergangszone ist erwünscht, da sie das Gebiet virtuell nach rechts fortsetzt und somit selbst die Wellen später auf die Übergangszone treffen, deren Weg anfangs zu flach verlief um noch vor dem rechten Rand bis in eine Höhe von 2250 km zu gelangen. Oberhalb der Übergangszone ist die Reflexion vom rechten Rand nicht relevant, da die Wellen von unten kommen und somit auch wieder schräg nach oben reflektiert werden, so daß kein Energiefluß über die Auswertungslinie zurück zu befürchten ist. Aus diesen beiden Gründen wurden rechts reflektierende Randbedingungen gesetzt; die Wahl von DIRICHLET- gegenüber NEUMANN-Werten ist hierbei willkürlich.

Um die allein unerwünschte Reflexion vom oberen Rand zu verhindern, wurde dieser durch Plazierung einiger nach oben hin größer werdender Zellen weit nach oben verlagert; der Abstand vom unteren Rand wurde dabei so gewählt, daß die Wellen innerhalb der Simulationszeit das Gebiet nicht von unten nach oben und zurück durchwandern können. Da der Abstand aufgrund

der im oberen Bereich sehr hohen Ausbreitungsgeschwindigkeit recht groß sein müßte, wurde eine Modifikation der nichtreflektierenden Randbedingungen verwendet. Dazu wird die Ausbreitungsgeschwindigkeit nach oben hin, von der tatsächlichen Physik abweichend, verringert; findet diese Verringerung auf einer Längenskala statt, die groß gegenüber typischen Wellenlängen ist, so ist keine Reflexion von Wellen zu erwarten, diese werden lediglich abgebremst. Auch beeinflußt diese Modifikation die im Gebiet vorhandene Energie nicht, was die Auswertung der Energieflüsse vereinfacht. In der Summe erlaubt dieser Ansatz, am oberen Rand reflektierende Randbedingungen zu verwenden; die Wahl von DIRICHLET-Bedingungen ist wieder willkürlich.

Als Randbedingung für den linken Rand wurden homogene NEUMANN-Werte gesetzt, da das Problem symmetrisch ist und somit nur die Hälfte der Welle gerechnet werden muß. Um nach oben hin größer werdende Zellen verwenden zu können ohne dabei deren Seitenverhältnis zu verschlechtern, wurde ein nach oben hin breiter werdendes Gebiet verwendet. Da die Ausdehnung des Gebiets klein gegenüber dem Durchmesser der Sonne ist, ist es nicht nötig, die Wellengleichung in Zylinder- statt kartesischen Koordinaten zu lösen.

3.2 Auswertung der Rechnungen

Ziel der Rechnungen war es, den Anteil der Energie einer Welle zu bestimmen, der die Grenzschicht passieren kann. Dazu gibt es im wesentlichen drei Möglichkeiten:

- *Direkte Auswertung des Energieflusses:* da der Energiestrom durch

  $$\mathbf{j} = va\nabla u$$

  gegeben ist, läßt sich die gesamte, durch eine Kurve C fließende Energiemenge durch Aufintegration erhalten:

  $$\mathcal{J}(\mathbf{w}) = \int_0^T \int_C v(\mathbf{x},t)\,(a(\mathbf{x})\nabla u(\mathbf{x},t))\cdot\mathbf{n}\;ds\;dt.$$

- *Energiemenge oberhalb der Auswertungslinie:* da ursprünglich keine Energie im Gebiet war und aufgrund der gewählten Randbedingungen auch kein Energieverlust durch den Rand möglich ist, muß die zum Endzeitpunkt oberhalb der Auswertungslinie vorhandene Energie gleich der durch die Linie hindurchgetretene Energie sein:

  $$\mathcal{J}(\mathbf{w}) = E_{y>y_0}(T) = \int_{\Omega\cap\{y>y_0\}} \frac{1}{2}\rho(\mathbf{x})v(\mathbf{x},T)^2 + \frac{1}{2}a(\mathbf{x})\,(\nabla u(\mathbf{x},T))^2\;\;d^dx.$$

  Wie in Abschnitt 2.1.3 beschrieben, ist die Energie bei Verwendung veränderlicher Gitter numerischen Störungen unterlegen. Dies trifft wegen der im oberen Teil des Gebiets sehr großen Zellen bei der Auswertung dieses Funktionals besonders zu; Abbildung 3.3 zeigt beispielhaft den Verlauf der Energie oberhalb der Auswertungslinie. Da wegen der Stufen der Wert der Energie am Endzeitpunkt nicht sehr zuverlässig ist, wurde $\mathcal{J}(\mathbf{w})$ mit Hand aus dem Zeitverlauf bestimmt, indem das Maximum des Verlaufs genommen wurde; es wurde nicht versucht, den weiteren Verlauf ohne die Sprünge zu antizipieren und so eine „bessere“ Schätzung zu erhalten. Zumindest für die feineren Gitter gibt es im allgemeinen nur noch einen oder zwei klar erkennbare Sprünge in einem Bereich, wo die Kurve schon sehr flach ist, so daß die Auswertung mit Hand recht zuverlässig erscheint.

- *Energiedifferenz unterhalb der Auswertungslinie:* ebenso muß der Energiefluß durch die Linie gleich der Differenz der Energien sein, die nach Einbringung der Welle und zum Endzeitpunkt im unteren Teilgebiet vorhanden sind:

  $$\mathcal{J}(\mathbf{w}) = E(t=\tau) - E_{y<y_0}(T).$$

Diese Auswertung ist im allgemeinen sehr viel schlechter, da die Energiedifferenz nur wenige Prozente der Gesamtenergiemenge ausmacht und daher sehr viel anfälliger ist gegenüber numerischen Störungen der Gesamtenergie. Typischerweise gehen in den ersten Verfeinerungsschritten mehrere zehn Prozent der Energie durch numerische Probleme verloren; diese werden dann irrtümlich dem Ergebnis dieser Auswertung zugerechnet. Erst bei Rechnungen mit extrem vielen Freiheitsgraden kommt der numerische Energieverlust bei Rechnungen mit stark schwankenden Koeffizienten in die Größenordnung von unter einem Prozent, was jedoch immer noch erheblich ist gegenüber den Werten dieses Funktionals.

Die Auswertung läßt sich jedoch näherungsweise um den beschriebenen Fehler korrigieren, wenn man die Abnahme der Gesamtenergie zwischen $t = \tau$ und $t = T$ von $\mathcal{J}(\mathbf{w})$ abzieht; diese Abnahme muß auf numerische Effekte zurückzuführen sein. Die Gültigkeit dieser Korrektur setzt voraus, daß der numerische Verlust ausschließlich im unteren Teil des Rechengebiets stattfindet; da dort der bei weitem überwiegende Teil der Gesamtenergie konzentriert ist, ist diese Annahme recht gut erfüllt, so daß wir

$$\mathcal{J}(\mathbf{w}) = (E(t = \tau) - E_{y<y_0}(T)) - (E(T) - E(t = \tau))$$

als Auswertegröße ansehen können.

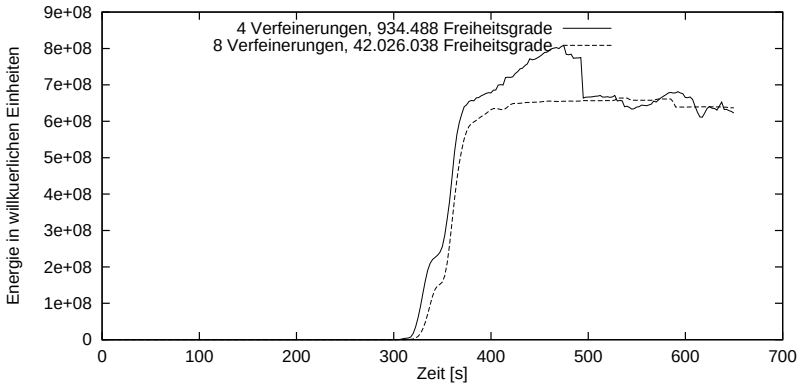

Abbildung 3.3: *Verlauf der Energie oberhalb der Auswertungslinie bei verschiedenen Verfeinerungsstufen. Man erkennt die durch numerische Instabilitäten hervorgerufenen Stufen, sowie die höhere Ausbreitungsgeschwindigkeit auf gröberen Gittern. Die Verfeinerung geschah durch den Energiefehlerindikator, die Ergebnisse sind aber vergleichbar bei der Verwendung des an diese Auswertung angepaßten dualen Funktionals.*

Bei exakter Rechnung würden alle drei Auswertungen das selbe Ergebnis liefern. In der Praxis zeigt sich jedoch, daß die ersten beiden Funktionale deutlich besser gegen einheitliche Werte konvergieren.

Die Auswertung des Energieflusses wurde in einer Höhe von $y_0 = 3000$ km durchgeführt. Exakter wäre eine Auswertung direkt oberhalb der Übergangszone, d. h. in einer Höhe von rund 2500 km, aus Praktikabilitätsgründen wurde diese Linie jedoch nach oben verschoben; numerische Tests zeigen, daß die Veränderung von y_0 keinen Einfluß auf das Ergebnis hat, was dadurch verständlich wird, daß die Ausbreitungsgeschwindigkeit im Bereich dazwischen so hoch ist, daß Energie, die die Übergangszone passiert, praktisch sofort weiter nach oben transportiert wird. Eine Auswertung bei $y_0 = 4000$ km ergibt keine wesentlich anderen Ergebnisse, obwohl man aufgrund der einspringenden Ecke und der oberhalb der Linie nicht mehr quadratischen Zellen eine schlechtere Konvergenz erwarten könnte.

3.3 Adaptivität

Bei der Durchführung der Rechnungen zeigte sich, daß ohne die Verwendung adaptiver Gitter keine verwertbaren Ergebnisse erzeugt werden konnten, da die dazu nötige Anzahl an Gitterzellen wesentlich größer war, als es Rechenzeit- und Speicherbeschränkungen erlaubten. Selbst die adaptiven Rechnungen benötigten mehrere Zehnmillionen Freiheitsgrade,[1] um Ergebnisse zu liefern, deren relative Genauigkeit bei vielleicht 2 oder 3 Prozent liegt.

Es wurden vergleichende Rechnungen durchgeführt, wobei als Verfeinerungskriterien einerseits der „traditionelle" Fehlerindikator (2.9), im folgenden kurz „Energiefehlerindikator" genannt,[2] andererseits der im letzten Kapitel vorgestellte Fehlerschätzer auf der Basis des dualen Problems Verwendung fand. Für die Rechnungen mit dem dualen Problem wurden die folgenden Fehlerfunktionale gewählt, die den drei Auswertungsmöglichkeiten oben entsprechen; da alle drei Funktionale nichtlinear sind, muß die in Abschnitt 2.4.1 erläuterte Linearisierung verwendet werden:

- *Direkte Auswertung des Energieflusses:*

$$\tilde{J}_{\mathbf{w}}(\mathbf{t}) = \int_0^T \int (v(x, y_0, t)(a\nabla\varphi) \cdot \mathbf{n} + \psi(a\nabla u(x, y_0, t)) \cdot \mathbf{n}) \; dx.$$

- *Energiemenge oberhalb der Auswertungslinie:*

$$\tilde{J}_{\mathbf{w}}(\mathbf{t}) = (\rho v(\cdot, T), \psi(\cdot, T))_{\Omega \cap \{y > y_0\}} + (a\nabla u(\cdot, T), \nabla\varphi(\cdot, T))_{\Omega \cap \{y > y_0\}}.$$

- *Energiedifferenz unterhalb der Auswertungslinie:*

$$\begin{aligned}\tilde{J}_{\mathbf{w}}(\mathbf{t}) = &\left[(\rho v(\cdot, \tau), \psi(\cdot, \tau))_{\Omega \cap \{y < y_0\}} + (a\nabla u(\cdot, \tau), \nabla\varphi(\cdot, \tau))_{\Omega \cap \{y < y_0\}}\right] \\ &- \left[(\rho v(\cdot, T), \psi(\cdot, T))_{\Omega \cap \{y < y_0\}} + (a\nabla u(\cdot, T), \nabla\varphi(\cdot, T))_{\Omega \cap \{y < y_0\}}\right].\end{aligned}$$

  Das Funktional berücksichtigt damit beide Zeitpunkte, zu denen eine Auswertung stattfindet, gleichermaßen und mit dem entsprechenden Vorzeichen.

Daneben wurde zu Vergleichszwecken mit einem linearen Zielfunktional

$$J(\mathbf{t}) = \mathcal{J}(\mathbf{t}) = \int_0^T ((a\nabla\varphi(x, y_0, t)) \cdot \mathbf{n} + \psi(x, y_0, t)a) \; dx. \tag{3.1}$$

gerechnet. Dieses hat keine direkte physikalische Bedeutung (und hat darüberhinaus auch keine definierte Einheit), vermeidet aber die Problematik der Linearisierung. Es trägt lediglich die Information über die Ausdehnung des Einflußgebietes des Zielfunktionals in sich.

Bei den Rechnungen mit dem dualen Fehlerschätzer ist es wichtig, zuerst einige Verfeinerungszyklen mit dem Energiefehlerindikator durchzuführen. Beginnt man zu früh damit, den dualen Schätzer zur Gittersteuerung zu verwenden, so konvergiert der Prozeß nicht und die Ergebnisse sind unbrauchbar. Der Grund dafür liegt darin, daß wegen der nichtlinearen Zielfunktionale die duale Lösung von der primalen abhängt; ist diese aufgrund eines zu groben Gitters schlecht approximiert, so wird auch die numerische duale Lösung weit von der exakten entfernt sein und im allgemeinen zu Gitterverfeinerung an den falschen Stellen führen, so daß die primale Lösung im nächsten Durchlauf genauso schlecht sein wird wie im vorhergehenden. Im Endergebnis konvergiert

[1] Die Anzahl der Freiheitsgrade ist im allgemeinen über die einzelnen Zeitschritte aufsummiert angegeben; da die Zahl zwischen einzelnen Zeitschritten stark schwanken kann, ist dies die einzige vergleichbare Größe. Allerdings wurde nur die Anzahl der Freiheitsgrade für die ursprüngliche Variable u akkumuliert, berücksichtigt man auch die Freiheitsgrade in der Geschwindigkeit v, so erhält man den doppelten Wert.

[2] Diese Bezeichnung ist irreführend, weil sie suggeriert, daß es sich um die Energie in der Wellengleichung handelt. Genauer wäre, nur von der elastischen Energie zu sprechen, es sei hier allerdings die Sprechweise übernommen, die sich bei der Lösung der LAPLACE-Gleichung eingebürgert hat.

der Prozeß aus adaptiver Gitterverfeinerung und Rechnung nicht, obwohl die Zellzahl beständig zunimmt.

Um die Konvergenzprobleme weit entfernt vom Konvergenzpunkt zu illustrieren, sei ein Beispiel bei der Auswertung des Linienintegrals als dualem Funktional angeführt. Das Beispiel zeigt dabei auch die extrem schlechte Konvergenz dieses Funktionals auf nicht ausreichend feinen Gittern (weniger als ca. 5000 Freiheitsgraden im Mittel pro Zeitschritt).

In Abbildung 3.4 ist der Energiefluß durch die Auswertungslinie nach vier und nach acht Verfeinerungsschritten (mit dem Energiefehlerschätzer) dargestellt. Legt man den Verlauf nach acht Verfeinerungsschritten zugrunde, so wird die duale Lösung im wesentlichen aus drei Pulsen bestehen, die von der Linie zu den Zeiten 345s, 370s und 400s ausgehen und in der Zeit zurückpropagieren; das Gitter wird nur entlang des Weges dieser drei Pulse verfeinert werden. Finge man jedoch bereits nach vier Verfeinerungsschritten mit dem dualen Fehlerschätzer an, so besäße die duale Lösung neben den beiden markanten Pulsen einen starken Anteil, der aus dem Rauschen stammt, das ab etwa 400s einsetzt. Da dieses Rauschen in der Rückwärtsrechnung *vor* den beiden Pulsen von der Linie ausgeht, ist das Gebiet wo die duale Lösung merklich von Null verschieden ist, erheblich größer als notwendig, was die für eine gegebene Genauigkeit erforderliche Zellzahl wesentlich in die Höhe treibt. Hier kommt darüberhinaus zum Tragen, daß die duale Lösung nach $t \approx 450s$ praktisch gleich Null ist, so daß man die weitere Rechnung auf ein grobes Gitter und große Zeitschritte beschränken kann, da die Fehlerindikatoren nahe Null sind, was bei Linearisierung um den links gezeigten Energiefluß unerkannt bliebe.[3]

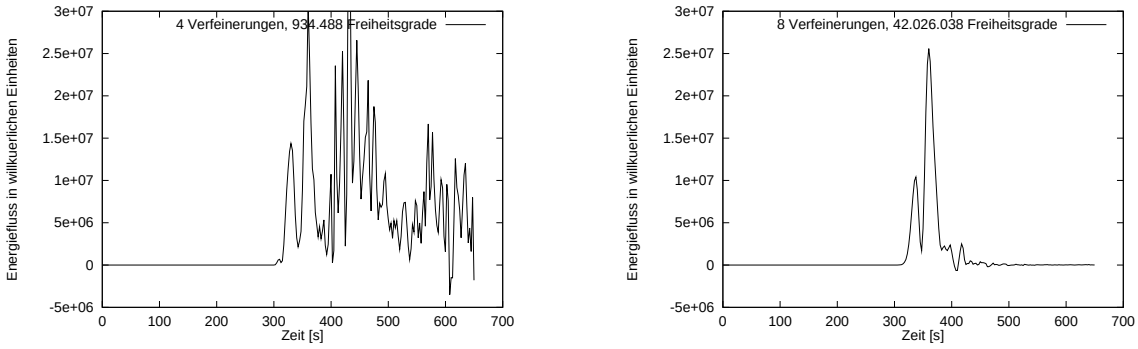

Abbildung 3.4: *Energiefluß durch die Auswertungslinie in Abhängigkeit von der Zeit bei verschiedenen Verfeinerungsstufen. Die Verfeinerung geschah durch den Energiefehlerindikator.*

3.4 Ergebnisse der Rechnungen

In den folgenden Abschnitten sind die Ergebnisse der Rechnungen bei globaler Verfeinerung, sowie bei Verfeinerung mit dem Energiefehlerindikator und dem dualen Fehlerschätzer gegenübergestellt. Es zeigt sich, daß genaue Werte für die interessierende Größe, den Anteil der von der Übergangsschicht transmittierten Energie, nur mit den Rechnungen zu erhalten waren, bei denen das Gitter mit dem Energiefehlerindikator verfeinert wurde; der vermutliche Wert für diese Größe ist 0.0270 ± 0.0015, d. h. es werden knapp drei Prozent der Energie durchgelassen. Dieser Wert liegt in einer Größenordnung, die mit einer einfachen Rechnung auch zu erwarten ist: der analytische Ausdruck für den an einer Unstetigkeit des Brechungsindex' transmittierten Energieanteils

[3] Die Verwendung dieser Information setzt voraus, daß die Gitterverfeinerung in einer Art durchgeführt wird, bei der die Fehlerindikatoren zweier Raum-Zeit-Zellen auch dann miteinander verglichen werden, wenn sie verschiedenen Zeitschritten angehören. Dies ist bei bis zu 50.000.000 Freiheitsgraden eine erhebliche programmtechnische Herausforderung. Im verwendeten Programm wurde diese Optimierung deshalb nicht verwendet, obwohl klar ist, daß sie zu einer weiteren drastischen Reduktion der Freiheitsgrade führen würde. Zu Details hierzu vergleiche Abschnitt 4.5.3.

ist durch

$$\frac{\Delta E}{E} = \frac{n}{n' + n} = \frac{1}{\frac{n'}{n} + 1}$$

gegeben, wobei n' und n die Brechnungsindizes auf den beiden Seiten der Unstetigkeit sind. Hier sind n, n' linear von der Wurzel aus der Temperatur T abhängig, so daß sich der Energiebruchteil zu

$$\frac{\Delta E}{E} \approx \frac{1}{\sqrt{\frac{T_{Korona}}{T_{Chromosphäre}}} + 1} \approx \frac{1}{20 + 1} \approx 0.048$$

ergibt. Berücksichtigt man noch den streifenden Einfall eines Teils der Welle, so ist der erhaltene Wert in der Größenordnung des erwarteten.

3.4.1 Verfeinerung mit einem Energiefehlerindikator

In den Abbildungen 3.5 und 3.6 sind die Ergebnisse der Rechnungen bei Verfeinerung mit dem Energiefehlerindikator dargestellt. Das Gitter[4] ist recht gut an die Lösung angepaßt, es folgt sowohl der transmittierten Welle mit ihrer hohen Geschwindigkeit, als auch dem nach unten zurückreflektierten Anteil.

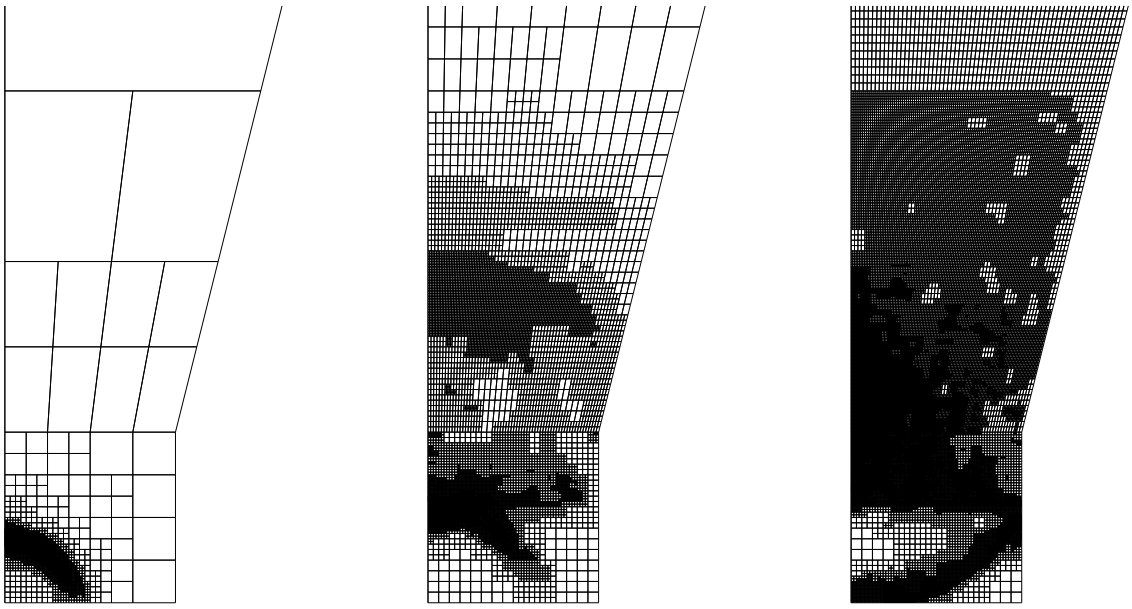

Abbildung 3.5: *Mit dem Energiefehlerindikator erzeugte Gitter zu den Zeiten t =250s, 375s und 650s.*

Die Ergebnissen der numerischen Auswertung der Rechnungen in Abbildung 3.6 lassen sich die folgenden zusammenfassen:

- Die Auswertung der Energie oberhalb der Meßlinie konvergiert recht gut gegen einen festen Wert. Dagegen ist die Auswertung der Energiedifferenz unterhalb dieser Linie sehr ungenau; das liegt daran, daß der numerische Verlust in der Gesamtenergie, der durch die variablen Gitter verursacht wird, diesem Funktional zugerechnet wird. Da sich die in den oberen Teil

[4] Es ist nur der untere Teil des Gebiets dargestellt, da nur dieser interessant ist. Der obere Teil wurde nur als Ersatz für absorbierende Randbedingungen eingeführt, so daß dort nichts wesentliches mehr passiert.

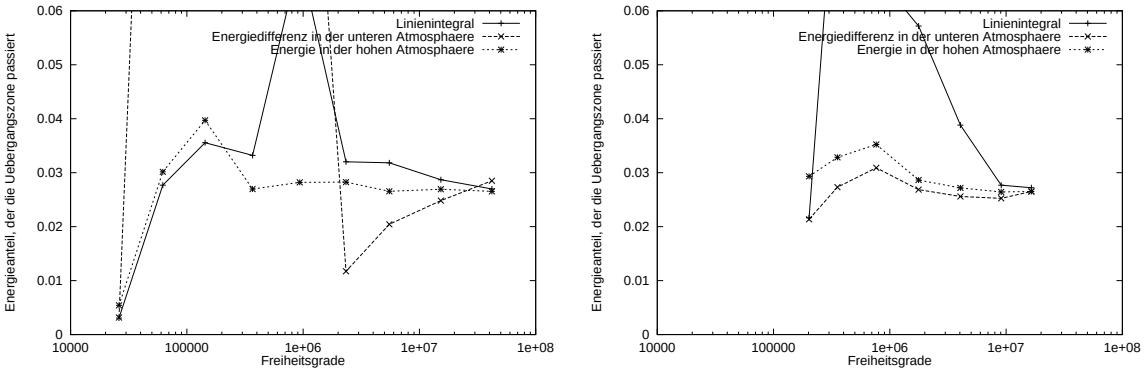

Abbildung 3.6: *Ergebnisse der verschiedenen Auswertungsmethoden bei Rechnungen mit dem Energiefehlerindikator; links bilineare Elemente, rechts bikubische. Aufgetragen ist der die Übergangsschicht passierenden Anteil der Gesamtenergie gegen die kumulierte Anzahl an Freiheitsgraden.*

des Gebiets transmittierte Energiemenge nur in der Größenordnung von zwei bis drei Prozent der Gesamtenergie bewegt, muß letztere numerisch auf wesentlich besser als ein Prozent konstant gehalten werden. Dies ist erst in den letzten Verfeinerungsschritten der Fall, so daß erst in diesen eine gute Genauigkeit bei der Auswertung dieses Funktionals gegeben war; davor ist das Gitter noch zu grob.

Andererseits ist die Auswertung der Energie oberhalb der Meßlinie sehr stabil gegen diese Art von Störungen, da auf der einen Seite ohnehin nur numerischer Energieverlust nach $t = \tau$ eine Rolle spielt, auf der anderen Seite der Energieverlust im Hauptteil der Welle, d. h. unterhalb der Linie irrelevant ist. Bewegt sich der numerische Energieverlust also in der Größenordnung von beispielsweise einigen Prozent auf einem mäßig feinen Gitter, so verfälscht er auch das Ergebnis bei diesem Funktional nur um die entsprechende Anzahl an Prozenten, während er bei der Auswertung der Energie unterhalb der Linie einen Fehler von mehreren hundert Prozent ausmachen kann.

- Die Auswertung mit dem Linienintegral konvergiert ebenfalls recht schlecht. Dies ist verständlich, beachtet man die Verläufe in Abbildung 3.4 und deckt sich mit der generellen Aussage, daß die Konvergenzordnung von Integralen über Linien oder Punktauswertungen eine schlechtere Konvergenzordnung haben als Gebietsintegrale. Andererseits ist zu beachten, daß es sich nicht eigentlich um ein Linienintegral handelt, da die Energie ja über die Linie hinwegtransportiert wird, was letztlich mit der Herleitung dieses Integrals als Divergenz der Energiemenge in einem der beiden Teilgebiete zusammenhängt; der Charakter dieses Funktionals entspricht daher eher dem eines Gebietsintegrals als dem eines Linienintegrals bei elliptischen Gleichungen.

- Aus dem rechten Teil der Abbildung 3.6 läßt sich die deutlich bessere Konvergenz der Rechnung mit bikubischen Elementen ablesen. Vor allem die Erhaltung der Gesamtenergie ist erheblich besser, so daß die Auswertung der Energiedifferenz unterhalb der Auswertungslinien bessere Ergebnisse zeigt. Die Auswertung des Linienintegrals ist jedoch ebenso unsicher wie bei linearen Elementen. Trotz der niedrigeren Anzahl von Freiheitsgraden für eine vergleichbare Genauigkeit ist der numerische Aufwand nicht geringer oder sogar höher als bei linearen Elementen, da die Anzahl der Einträge in dem Systemmatrizen größer ist, was die benötigte Anzahl an Operationen für Matrix-Vektor-Multiplikationen erhöht.

Aus den jeweils genauesten Erfgebnissen der beiden Rechnungen und den verschiedenen Wegen der Auswertung läßt sich für den Anteil der von der Übergangszone transmittierten Energie der Wert 0.0270 ± 0.0015, das heißt mit einer Genauigkeit von rund fünf Prozent bestimmen. Die

Fehlergrenzen wurden dabei symmetrisch so gewählt, daß alle sechs Datenpunkte innerhalb dieses Intervalls liegen.

3.4.2 Verfeinerung mit dem dualen Schätzer

In den Abbildungen 3.7, 3.8 und 3.9 sind die mit den drei nichtlinearen Zielfunktionalen erzeugten Gitter dargestellt. Die Gitter, die mit dem linearen Zielfunktional (3.1) erzielt wurden, ähneln denen des nichtlinearen Funktionals auf der Auswertungslinie, sind aber etwas voller. Im wesentlichen entsprechen die Gitter dem, was man aufgrund der anschaulichen Bedeutung des dualen Problems jeweils erwarten würde. Es ist dabei zu beachten, daß zuerst etliche Verfeinerungsschritte mit dem Energiefehlerindikator durchgeführt wurden, bevor auf den dualen Schätzer umgeschaltet wurde; da nur einige wenige Schritte mit letzterem gemacht wurden, bestehen noch Bereiche mit relativ feinen Zellen, die nicht zum Zielfunktional beitragen, beispielsweise ober- und unterhalb der Auswertungslinie am Endzeitpunkt in Abbildung 3.7, unterhalb der Auswertungslinie am Endzeitpunkt in Abbildung 3.8 und oberhalb jener in Abbildung 3.9. Diese Bereiche würden in den nächsten Verfeinerungsschritten aufgelöst und durch grobe Gitter ersetzt, wofür allerdings mehr Verfeinerungsschritte mit dem dualen Fehlerschätzer nötig wären.

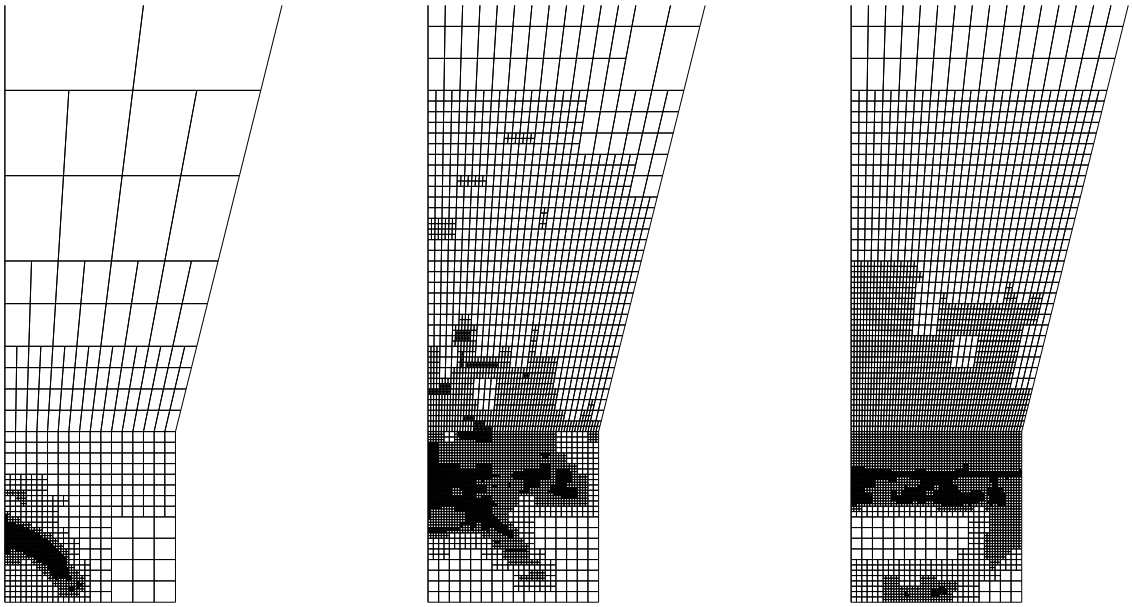

Abbildung 3.7: *Mit dem dualen Schätzer erzeugte Gitter zu den Zeiten t =250s, 375s und 650s. Als Fehlerfunktional diente der Fluß durch die Auswertungslinie.*

Während die Gitter in etwa den erwarteten entsprechen, zeigen die numerischen Ergebnisse der Auswertungen der Rechnungen enttäuschendes Verhalten. Dieses ist in den Abbildungen 3.10 und 3.11 für die vier verschiedenen Zielfunktionale dokumentiert. In jeder der Abbildungen sind alle drei Wege der Auswertung dargestellt; allerdings ist eigentlich nur die zum Zielfunktional passende angebracht, die in Abbildung 3.12 für die verschiedenen Rechnungen einander gegenübergestellt werden.

Bei den gezeigten Kurven ist zu beachten, daß die ersten Datenpunkte jeweils durch Verfeinerung mit dem Energiefehlerindikator erhalten wurden. Die entsprechende Anzahl an Schritten ist in den Abbildungen vermerkt. Nachdem auf den dualen Fehlerschätzer umgestellt wurde, verringert sich die Zellzahl häufig in den ersten Schritten, da mehr Zellen aufgelöst als verfeinert werden, so daß die Kurven rückläufiges Verhalten zeigen. Erst nach einigen wenigen Verfeinerungsschritten mit dem dualen Fehlerschätzer nimmt die Anzahl der Zellen wieder zu.

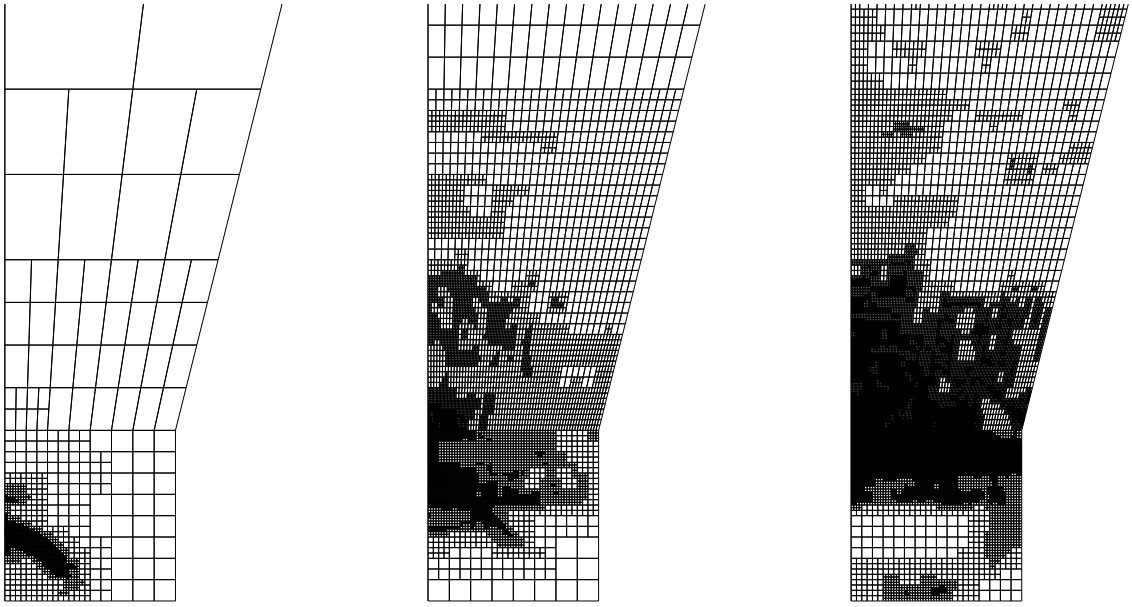

Abbildung 3.8: *Mit dem dualen Schätzer erzeugte Gitter zu den Zeiten t =250s, 375s und 650s. Als Fehlerfunktional diente die Energie oberhalb der Auswertungslinie zum Endzeitpunkt.*

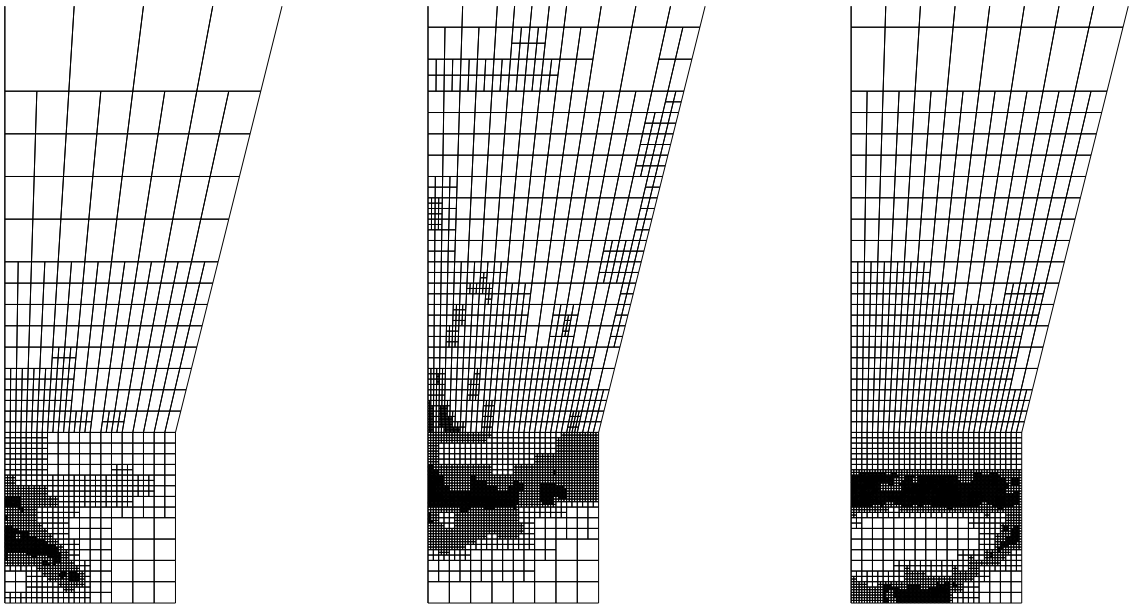

Abbildung 3.9: *Mit dem dualen Schätzer erzeugte Gitter zu den Zeiten t =250s, 375s und 650s. Als Fehlerfunktional diente die Differenz der Energien unterhalb der Auswertungslinie zu den Zeitpunkten* $t = T = 650s$ *und* $t = \tau = 60s$.

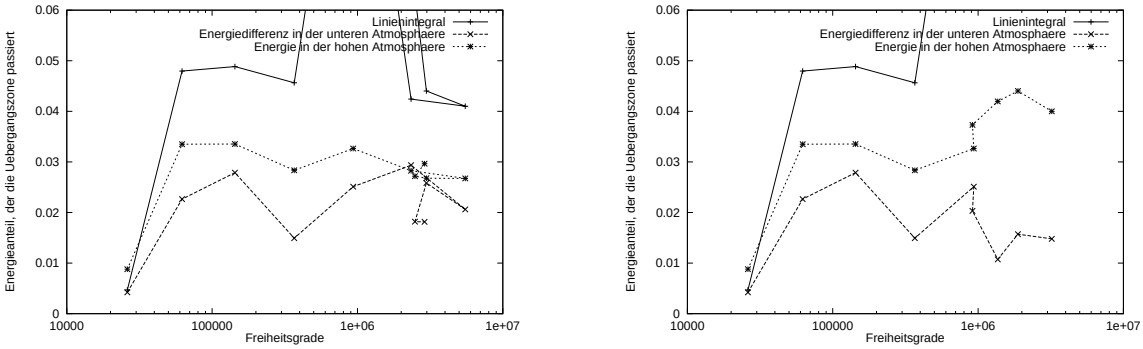

Abbildung 3.10: *Ergebnis der Auswertungen bei Verfeinerung mit dem dualen Problem. Links Verfeinerung mit dem nichtlinearen, rechts mit dem linearen Linienintegral, wobei die ersten 7 bzw. 5 Datenpunkte mit dem Energiefehlerindikator gewonnen wurden.*

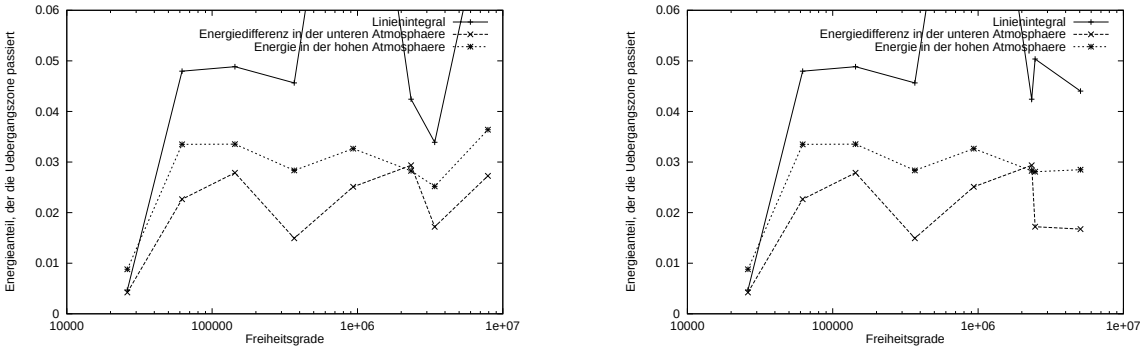

Abbildung 3.11: *Ergebnis der Auswertungen bei Verfeinerung mit dem dualen Problem. Links Verfeinerung mit dem Zielfunktional der Energie oberhalb der Auswertungslinie am Endzeitpunkt, rechts mit der Energiedifferenz unterhalb der Auswertungslinie, wobei jeweils die ersten 6 Datenpunkte mit dem Energiefehlerindikator gewonnen wurden.*

In allen Fällen zeigte die Auswertung nach dem Umschalten nicht das erwünschte konvergente Verhalten. Das aus Darstellungsgründen nicht mehr sichtbare weitere Verhalten der Auswertung des Linienintegrals bei linearen Zielfunktional divergiert und nimmt Werte größer als 0.2 an, was mit den zu erwartenden Werten von etwa 0.027 zu vergleichen ist. Zwar wurden zu wenige Verfeinerungsschritte mit dem dualen Schätzer durchgeführt, um über die Konvergenz endgültige Aussagen treffen zu können, es deuten aber andere Indizien, beispielsweise das wieder zunehmende hochfrequente Rauschen beim Energiefluß durch die Auswertungslinie nach den zwei Hauptpulsen (vgl. Abbildung 3.4), darauf hin, daß die Rechnungen auch in weiteren Verfeinerungsschritten nicht konvergieren könnten.

Zur Untersuchung, weshalb die Verfeinerung mit dem dualen Fehlerschätzer so schlecht konvergiert, wurde der in Abschnitt 2.4.6 hergeleitete Ausdruck für den Linearisierungsfehler bei der Bestimmung des Energieflusses durch die Auswertungslinie, (2.30), näherungsweise numerisch ausgewertet. Die Ergebnisse sind in Abbildung 3.13 dargestellt; dabei wurden die folgenden Bezeichnungen verwendet:

- i: Verfeinerungsstufe;
- $N^{(i)}$: Über die Zeitschritte summierte Anzahl von Freiheitsgraden;
- $\mathcal{J}(\mathbf{w}_h^{(i)}) = \Phi(\mathbf{w}_h^{(i)})$: Bei dieser Verfeinerungsstufe berechneter Energiefluß durch die Kurve;

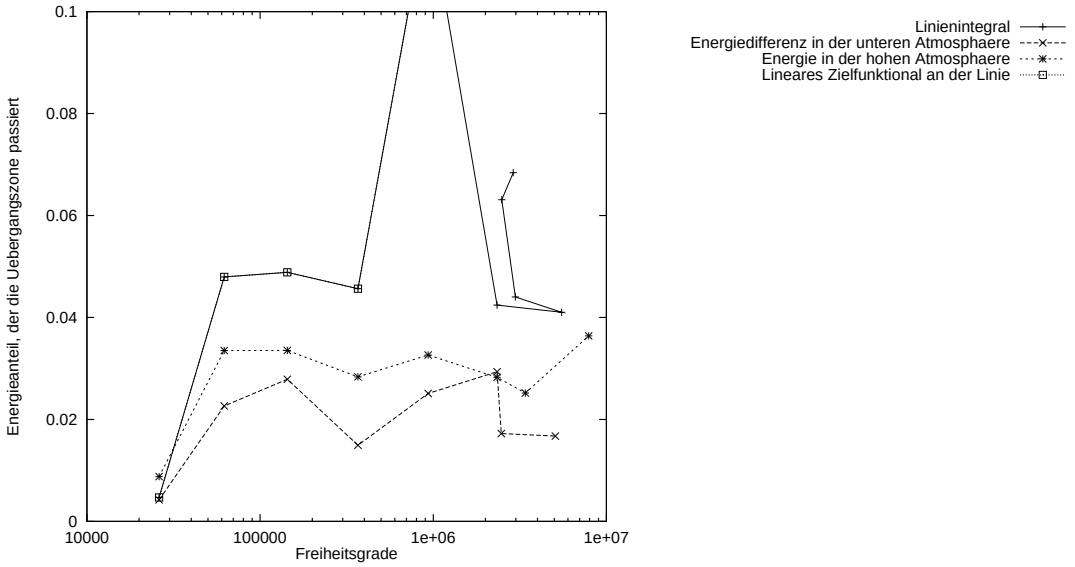

Abbildung 3.12: *Ergebnis der Auswertungen bei Verfeinerung mit dem dualen Problem. Vergleich der zum jeweiligen dualen Funktional gehörenden Auswertung.*

- $J(\mathbf{e}^{(i)}) = \mathcal{J}(\mathbf{w}) - \mathcal{J}(\mathbf{w}_h^{(i)})$: Tatsächlicher Fehler bei Annahme der Exaktheit $\mathbf{w} = \mathbf{w}_h^{(8)}$;
- $\tilde{J}(\mathbf{e}^{(i)}) = \int_0^T \int_{\mathcal{C}} \left(v_h^{(i)}(a\nabla(u - u_h^{(i)})) \cdot \mathbf{n}(v - v_h^{(i)})(a\nabla u_h^{(i)}) \cdot \mathbf{n} \right) \, ds \, dt$: Mit dem linearisierten Problem geschätzter Fehler. In einer Simulation ist die exakte Lösung nicht bekannt und diese Größe kann nur mit Hilfe des dualen Problems und den damit erhaltenen Fehlerschätzern berechnet werden; bei der Lösung der dualen Problems treten zusätzliche Fehler auf, die die Berechnung dieses exakten Fehlers oft sehr ungenau werden lassen. Hier wurde die Fehleridentität ohne Verwendung des dualen Problems ausgewertet, indem für die exakte Lösung die genaueste numerische Lösung $\mathbf{w}_h^{(8)}$ eingesetzt wurde;
- $\Delta J(\mathbf{e}^{(i)}) = J(\mathbf{e}^{(i)}) - \tilde{J}(\mathbf{e}^{(i)})$: Differenz zwischen echtem Fehler und mit dem linearisierten Funktional geschätztem Fehler;
- $\frac{|\Delta J(\mathbf{e}^{(i)})|}{|J(\mathbf{e}^{(i)})|}$: Verhältnis von Linearisierungsfehler und Wert des tatsächlichen Fehlers.

Die Daten entstammen einer Rechnung mit Verfeinerung durch den Energiefehlerindikator; der Energiefluß für zwei Verfeinerungsstufen ist in Abbildung 3.4 dargestellt.

Zur Bestimmung des Linearisierungsfehlers ist die Kenntnis der exakten primalen Lösung notwendig; da diese nicht bekannt ist, wurde für $\mathbf{w} = (u, v)$ die genaueste, bekannte Lösung $\mathbf{w}_h^{(8)} = (u_h^{(8)}, v_h^{(8)})$ eingesetzt. Es kann davon ausgegangen werden, daß diese eine gute Schätzung für die exakte Lösung darstellt, ebenso für den auf den ersten Blick nicht gut auskonvergierten Wert des Energieflusses (zweite Spalte der Tabelle), da die Unterschiede in den aufintegrierten Energieflüßsen bei verschiedenen Verfeinerungsstufen zu einem erheblichen Teil durch die Oszillationen nach den drei Hauptpulsen hervorgerufen wird; der integrierte Energiefluß nach $t = 450s$ beträgt für $i = 8$ nur noch rund ein Zehntel des Wertes bei $i = 7$ und hat nur noch einen Anteil von etwa zwei Prozent am Gesamtenergiefluß. Als Mittelwert aus den verschiedenen Auswertungen bei linearen und kubischen Elementen ergibt sich ein integrierter Energiefluß von $(6.7 \pm 0.4) \cdot 10^8$, also sehr nahe an dem in der Tabelle verzeichneten Wert, so daß die Wahl von $\mathbf{w}_h^{(8)}$ gerechtfertigt erscheint.

Aus der Tabelle geht hervor, daß der Linearisierungsfehler im Verhältnis zum tatsächlichen Fehler zwar mit zunehmender Verfeinerung kleiner wird (er sollte etwa mit der Ordnung h gehen),

i	$N^{(i)}$	$\mathcal{J}(\mathbf{w}_h^{(i)})$	$\mathcal{J}(\mathbf{w}) - \mathcal{J}(\mathbf{w}_h^{(i)})$	$\tilde{J}(\mathbf{e}^{(i)})$	$\Delta J(\mathbf{e}^{(i)})$	$\frac{\lvert\Delta J(\mathbf{e}^{(i)})\rvert}{\lvert\mathcal{J}(\mathbf{w})-\mathcal{J}(\mathbf{w}_h^{(i)})\rvert}$
0	26.100	$2.77 \cdot 10^8$	$3.97 \cdot 10^8$	$-3.38 \cdot 10^8$	$7.35 \cdot 10^8$	1.85
1	62.086	$7.16 \cdot 10^9$	$-6.48 \cdot 10^9$	$-1.57 \cdot 10^{10}$	$9.24 \cdot 10^9$	1.43
2	143.491	$5.54 \cdot 10^9$	$-4.86 \cdot 10^9$	$-1.12 \cdot 10^{10}$	$6.37 \cdot 10^9$	1.31
3	367.371	$1.66 \cdot 10^9$	$-9.84 \cdot 10^8$	$-2.90 \cdot 10^9$	$1.92 \cdot 10^9$	1.95
4	934.488	$3.01 \cdot 10^9$	$-2.34 \cdot 10^9$	$-4.86 \cdot 10^9$	$2.53 \cdot 10^9$	1.08
5	2.336.911	$1.05 \cdot 10^9$	$-3.77 \cdot 10^8$	$-8.41 \cdot 10^8$	$4.64 \cdot 10^8$	1.23
6	5.513.318	$1.02 \cdot 10^9$	$-3.46 \cdot 10^8$	$-6.87 \cdot 10^8$	$3.41 \cdot 10^8$	0.99
7	15.277.876	$7.91 \cdot 10^8$	$-1.17 \cdot 10^8$	$-2.22 \cdot 10^7$	$1.05 \cdot 10^8$	0.90
8	42.026.038	$6.74 \cdot 10^8$	–	–	–	–

Abbildung 3.13: *Bestimmung des Linearisierungsfehlers bei nichtlinearen Zielfunktionalen. Man beachte, daß die ersten drei Datensätze im wesentlichen ohne Aussage sind, da dort die Gesamtenergie um mehr als einen Faktor vier vom richtigen Wert abweicht, damit auch der Energiefluß zu groß ist und der Vergleich mit* $\mathcal{J}(\mathbf{w}) = \Phi(\mathbf{w}) \approx \Phi(\mathbf{w}_h^{(8)}) = \mathcal{J}(\mathbf{w}_h^{(8)})$ *nicht realistisch sein kann. Weitere Erläuterungen im Text.*

jedoch selbst bei sehr genauen Rechnungen immer noch in der selben Größenordnung wie der Gesamtfehler liegt. Die Linearisierung um die numerische Lösung ist daher recht schlecht, und man muß erwarten, daß die numerisch berechnete duale Lösung erheblich von der tatsächlichen abweicht. Führt man die Fehlerschätzung und Verfeinerung durch Lösung eines dualen Problems tatsächlich durch (die Daten der Tabelle wurden durch Verfeinerung mit einem Energiefehlerindikator gewonnen), so zeigen die berechneten Fehlerschätzer keinerlei Korrelation mit dem Fehler, den man durch Vergleich mit der genauesten vorhandenen Lösung erwartet, und die erzeugten Gitter zeigen Verfeinerung in Gebieten, die für die Berechnung der Zielgröße nicht relevant sind.

Ob die schlechte Linearisierung erklärt, weshalb in den gerechneten Beispielen der Prozeß aus Gitterverfeinerung und Rechnung von primalem und dualem Problem nicht konvergiert und der Ansatz mit dem dualen Fehlerschätzer fehlschlägt, ist nicht vollständig klar; insbesondere ist nicht klar, weshalb die eigentlich gut adaptierten Gitter zu keinem entsprechenden Ergebnis führen. Es kommen aber eine Reihe weiterer Möglichkeiten in Betracht, die im Rahmen dieser Arbeit jedoch nicht mehr untersucht werden konnten. Insbesondere ist hier das Problem zu nennen, daß Wellen am Übergang zu gröberen Gittern teilweise reflektiert werden. Da der duale Fehlerschätzer im Gegensatz zum Energiefehlerindikator versucht, Teile der Wellen auf groben Gittern zu berechnen, wenn sie nicht im Einflußgebiet des Zielfunktionals liegen, ist zu erwarten, daß diese unerwünschte Reflexion relevante Ausmaße annehmen kann. Sie wäre dann in der Lage, die Ergebnisse erheblich zu verfälschen. In Abschnitt 5.4 sind einige, allerdings unvollständige Untersuchungen zum Einfluß von Gitterunstetigkeiten bei einem Modellfall zu finden. Gestützt wird die These, daß Reflexion an Gitterunstetigkeiten der Grund für die schlechten Ergebnisse ist, von der Beobachtung, daß das in Abbildung 3.4 gezeigte hochfrequente Rauschen nach den zwei Hauptpulsen nach der Umschaltung vom Energiefehlerindikator auf den dualen Fehlerschätzer wieder zunimmt, was von gestreuten Wellen verursacht sein könnte; letztlich ist diese These aber nicht bewiesen und verlangt nach weiteren Untersuchungen.

3.4.3 Globale Verfeinerung

Zum Vergleich sind noch die erhaltenen Werte bei globaler Verfeinerung aufgeführt. Wie Abbildung 3.14 zu entnehmen ist, ist der Anteil der Energie, der die Übergangsschicht passiert, deutlich niedriger als bei den Ergebnisse mit adaptiver Verfeinerung. Obwohl die Kurve recht flach ist, kann man davon ausgehen, daß sie keine konvergierten Werte repräsentieren. Der Grund dafür ist der in Abbildung 3.15 dargestellte Energiefluß durch die Auswertungslinie, der mit den Ergebnisse aus Abbildung 3.4 zu vergleichen ist; die noch stark oszillierende Kurve ist zwar im Vergleich mit den vorhergehenden Verfeinerungsstufen deutlich glatter, aber nicht vergleichbar mit dem Ergeb-

nis der feinsten adaptiven Rechnung. Die Oszillation stammt, wie in Abbildung 3.4 links, von einem zu groben Gitter; der glattere Verlauf liegt in der Abwesenheit der bei adaptiven Rechnungen vorhandenen Störungen durch Gitteränderungen begründet. Besonders bemerkenswert ist der unphysikalische und damit offensichtlich falsche zeitweilige negative Energiefluß, der bei den adaptiven Rechnungen auch nicht auftritt.

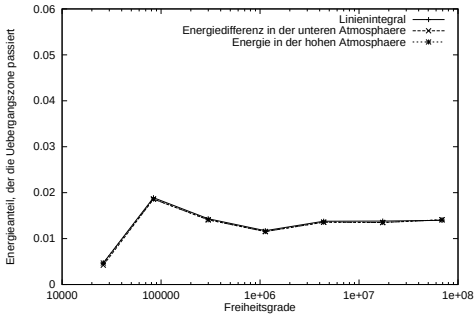

Abbildung 3.14: *Ergebnisse bei globaler Verfeinerung.*

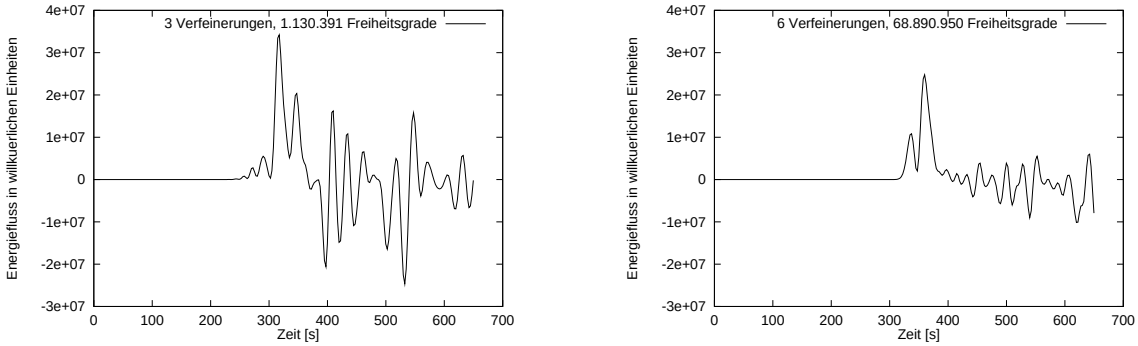

Abbildung 3.15: *Energiefluß durch die Auswertungslinie in Abhängigkeit von der Zeit bei verschiedenen Verfeinerungsstufen. Die Verfeinerung geschah durch globale Verfeinerung. Man vergleiche die Ergebnisse mit denen aus Abbildung 3.4, die mit deutlich weniger Freiheitsgraden gewonnen wurden.*

Aus den genannten Gründen muß davon ausgegangen werden, daß noch erheblich mehr als die verwendeten 68.000.000 Freiheitsgrade nötig wären, um ein vergleichbares Ergebnis wie bei den adaptiven Rechnungen zu erhalten; diese wurden freilich mit maximal rund 42.000.000 Freiheitsgraden erzielt.

Bemerkenswert an den Ergebnissen aus Abbildung 3.14 ist noch, daß die drei Auswertungsmöglichkeiten praktisch die selben Werte ergeben. Das liegt einerseits an der Energieerhaltung in Abwesenheit von Gitteränderungen (das begründet die Gleichheit der Ergebnisse der Energieauswertung ober- und unterhalb der Linie), bedeutet aber andererseits auch, daß praktisch der gesamte Energieaustausch zwischen den beiden Teilen des Gebiets durch die Auswertungslinie stattfindet. Letzteres ist angesichts der globalen Kopplung der in jedem Zeitschritt zu lösenden elliptischen Gleichung nicht selbstverständlich; insbesondere zeigen die Ergebnisse adaptiver Rechnungen, daß durch Gitteränderung Energie auch nichtlokal auf dem Gitter verteilt wird, was dazu führt, daß sich aufintegrierter Energiefluß durch die Auswertungslinie und die oberhalb dieser Linie vorhandene Energie unterscheiden.

Kapitel 4

Technische Aspekte der Implementation

In diesem Kapitel sind einige der Aspekte der Implementation des Problems kurz beschrieben, die nicht zu den üblichen Standardtechniken bei Programmierung mit finiten Elementen gehören. Dazu gehören insbesondere die Behandlung hängender Knoten, aber auch der Transfer von Finite-Elemente-Funktionen von einem Gitter zu einem anderen. Die Behandlung von hängenden Knoten folgt im wesentlichen [22, 4, 5]. Der Transfer zwischen Gittern ist parallel zu [19] implementiert. Eine ausführliche technische Referenz der DEAL.II-Bibliothek ist unter [3] zu finden.

4.1 Die deal.II-Bibliothek

Das Programm, das die in dieser Arbeit beschriebenen Methoden implementiert, basiert auf der parallel zur Arbeit entstandenen Bibliothek DEAL.II. Sie ist im wesentlichen eine Neuimplementation einiger Verfahren, die anhand der am Institut für Angewandte Mathematik in den letzten Jahren entwickelten Bibliothek DEAL entwickelt wurden, d. h. insbesondere die Verwendung von hängenden Knoten bei hierarchischen Gittern und die Verwendung von Fehlerschätzkonzepten zur adaptiven Gitterverfeinerung. Von DEAL unterscheidet sie sich jedoch in wesentlichen Punkten, darunter unter anderem:

- *Verschiedene Finite Elemente:* Es wurde eine Schnittstelle definiert, die es erlaubt nahezu beliebige Elementtypen zu implementieren. Dazu muß der Bibliothek lediglich die folgende Information bekanntgegeben werden:
  - wieviele Freiheitsgrade ein Element pro Knoten, Kante und Zelle hat;
  - wie die Interpolation entlang von Kanten zu erfolgen hat, an der Zellen unterschiedlicher Größe zusammenkommen;
  - wie die Interpolations- und Prolongationsmatrizen zwischen Zellen und ihren Kindern aussehen.

  Mit diesen Informationen kann die Bibliothek die Verteilung von Freiheitsgraden, die Berechnung von Nebenbedingungen für hängende Knoten u.ä. vornehmen, ohne über das verwendete Element genauer Bescheid zu wissen.

  Für die Numerik sind eine Reihe von Funktionen zu implementieren, die Werte, Ableitungen, lokale Massematrizen, usw. der Basisfunktionen zurückliefern. Diese Funktionen haben ebenfalls einheitliche Namen, so daß es möglich ist, Programme ohne Kenntnis des verwendeten Elements zu schreiben und die Auswahl des Elements zur Laufzeit vorzunehmen.

  Im Moment sind LAGRANGE-Elemente mit linearen bis quartischen Ansatzgraden, sowie ein spezielleres Element implementiert.

- *Dimensionsunabhängigkeit:* Die meisten Programme lassen sich dimensionsunabhängig formulieren, wenn man berücksichtigt, daß Operationen im allgemeinen auf Zellen und ihren Seiten stattfinden. Die Bibliothek unterstützt dies, indem sie neben den Datentypen Hex, Quad und Line auch logische Typen wie Cell und Face definiert. Diese werden je nach der verwendeten Dimension auf die richtigen Objekte gesetzt, so daß eine Schleife über Zellen in zwei Dimensionen über Vierecke, in drei Dimensionen über Hexaeder geht.

  Die bei weitem meisten Algorithmen in Anwenderprogrammen und der Bibliothek lassen sich auf diese Weise dimensionsunabhängig schreiben und sind in einer, zwei und drei Dimensionen ohne Änderung lauffähig. Dies vereinfacht nicht nur die Programmierung sondern verringert durch Wiederverwendung von Programmteilen auch erheblich das Risiko von Fehlern, da nicht die gleichen Algorithmen in verschiedenen Versionen für unterschiedliche Dimensionen gewartet werden müssen.

- *Moderne Programmiertechniken:* Durch die Verwendung von in den letzten Jahren entwickelten Programmiertechniken, insbesondere Templates und Iteratoren und durch die Nutzung der Standard Template Library von C++ ist die Programmierung der Bibliothek erheblich einfacher und weniger fehleranfällig; erst die Nutzung dieser Möglichkeiten hat es erlaubt, dimensionsunabhängige Algorithmen zu schreiben.

  Darüberhinaus macht die Bibliothek wesentlich mehr als DEAL von Fehlerabfragen zur Laufzeit Gebrauch, um Programmierfehler frühzeitig zu erkennen. Für Produktionsrechnungen und ausgetestete Programme kann die Fehlerprüfung ausgeschaltet werden.

4.2 Transfer von einem Gitter zum anderen

Bei der Diskretisierung ist in jedem Zeitschritt eine Gleichung des folgenden Typs zu lösen:

$$(\rho + \alpha\mathcal{A})u^1 = \rho u^0 + \beta k \rho v^0 - \gamma \mathcal{A} u^0.$$

Die Faktoren α, β und γ spielen in diesem Abschnitt keine Rolle. Mit finiten Elementen im Raum diskretisiert ergibt sich die folgende Aufgabe: *Suche* $u_h^1 \in V^1$, *so daß für alle* $\psi^1 \in V^1$

$$(\rho u_h^1, \psi^1) + \alpha a(u_h^1, \psi^1) = (\rho u_h^0, \psi^1) + \beta k(\rho v_h^0, \psi^1) - \gamma a(u_h^0, \psi^1) \tag{4.1}$$

gilt. Dabei sind $u_h^0, v_h^0 \in V^0$, und V^0 und V^1 seien die Ansatzräume auf den Gittern zum vorigen und zum aktuellen Zeitschritt.

Schreibt man das Problem in ein Gleichungssystem um, so ergibt sich

$$(M_{ij}^1 + \alpha A_{ij}^1)u_j^1 = M_{ij}^{10} u_j^0 + \beta k M_{ij}^{10} v_j^0 - \gamma A_{ij}^{10} u_j^0$$

mit den Matrizen

$$\begin{aligned} M_{ij}^1 &= (\rho\phi_i^1, \phi_j^1), \\ A_{ij}^1 &= (\nabla\phi_i^1, \nabla\phi_j^1), \\ M_{ij}^{10} &= (\rho\phi_i^1, \phi_j^0) \qquad \text{und} \\ A_{ij}^{10} &= (\nabla\phi_i^1, \nabla\phi_j^0). \end{aligned}$$

Dabei sind die $\phi_j^0 \in V^0, \phi_j^1 \in V^1$ die Basisfunktionen der jeweiligen Finite-Element-Räume. Während $M^1, A^1 \in \mathbb{R}^{n^1 \times n^1}$ quadratische Matrizen der Dimension n^1 des Ansatzraums zum neuen Zeitschritt sind, gilt $M^{10}, A^{10} \in \mathbb{R}^{n^1 \times n^0}$, mit n^0 der Dimension des Ansatzraumes zum alten Zeitschritt.

Die Aufstellung der Matrizen M^1, A^1 ist Standard und bietet keine besonderen Schwierigkeiten, wenn sie zellweise durchgeführt wird. Für die Aufstellung der anderen Matrizen spalten wir die Integration über Zellen auf:

$$M_{ij}^{10} = \sum_{K \in \mathbb{T}} (\rho\phi_i^1, \phi_j^0)_K = \sum_{K \in \mathbb{T}} M_{ij}^{10,K}$$

Dabei sei $\mathbb{T}$ eine Triangulierung des Gebiets so, daß jedes $K \in \mathbb{T}$ sowohl in $\mathbb{T}^0$ als auch in $\mathbb{T}^1$ enthalten ist, wobei $\mathbb{T}^n$ die Triangulation des Gebiets im $n$ten Zeitschritt definiere; Kinder von K seien höchstens in einem der beiden Gitter enthalten. Anschaulich ist $\mathbb{T}$ die Menge der jeweils feinsten Zellen, die in beiden Gitter noch enthalten sind. Zur Illustration fordern wir, daß sich die Zellen der beiden Gitter um höchstens ein Verfeinerungslevel unterscheiden dürfen, d. h. daß jedes $K \in \mathbb{T}$ entweder auf keinem der beiden Gitter Kinder hat oder auf genau einem der beiden Gitter genau einmal verfeinert ist. Diese Einschränkung ist hier nur der Einfachheit halber gemacht; Verfeinerungen über mehrere Stufen sind möglich und werden rekursiv mit dem unten beschriebenen Verfahren gehandhabt.

Bei nur einer Verfeinerung gibt es drei Fälle zu unterscheiden:

1. K ist auf keinem der beiden Gitter weiter verfeinert; dann ist $dim(V^1|_K) = dim(V^0|_K)$ und $M_{ij}^{10,K}$ ist die übliche zellweise Massematrix.

2. K ist auf $\mathbb{T}^0$ einmal verfeinert; dann ist $dim(V^1|_K) < dim(V^0|_K)$. Für konforme Ansätze mit einer Schachtelung der Ansatzräume gilt insbesondere $V^1|_K \subset V^0|_K$.

3. K ist auf $\mathbb{T}^1$ einmal verfeinert; dann ist $dim(V^1|_K) > dim(V^0|_K)$. Für konforme Ansätze mit einer Schachtelung der Ansatzräume gilt hier $V^1|_K \supset V^0|_K$.

Die Fälle 2 und 3 sind etwas aufwendiger. Für geschachtelte Ansatzräume ließe sich beispielsweise für den zweiten Fall die Funktion $\phi_i^1|_K$ durch die Basisfunktionen aus V^0 darstellen:

$$\phi_i^1|_K = \sum_{l \in \mathbb{L}} c_{il} \phi_l^0|_K, \tag{4.2}$$

mit $\{\phi_l^0\}_{l \in \mathbb{L}}$ der Menge aller Basisfunktionen mit Träger auf K. Die Matrix c_{il} ist unabhängig von K und konstant. Damit läßt sich $M_{ij}^{10,K}$ durch eine Summe über die Kindzellen von K schreiben:

$$\begin{aligned} M_{ij}^{10,K} &= (\rho \phi_i^1, \phi_j^0)_K \\ &= \sum_{l \in \mathbb{L}} c_{il} (\rho \phi_l^0, \phi_j^0)_K \\ &= \sum_{k \in \mathbb{K}} \sum_{l \in \mathbb{L}} c_{il} (\rho \phi_l^0, \phi_j^0)_k \\ &= \sum_{k \in \mathbb{K}} \sum_{l \in \mathbb{L}} c_{il} M_{lj}^{0,k}, \end{aligned}$$

wobei $\mathbb{K}$ die Menge der Kindzellen von K darstelle. Die Berechnung von $M_{ij}^{10,K}$ ist somit wieder auf eine Summation über zellweise Massematrizen zurückgeführt. Der umgekehrte, dritte Fall, daß K nur auf $\mathbb{T}^1$ verfeinert ist, geht ganz analog, indem man ϕ_i^0 entsprechend (4.2) durch die ϕ_l^1 ausdrückt.

Die Darstellung (4.2) setzt voraus, daß $V^1|_K \subset V^0|_K$, das heißt daß sich jede Basisfunktion durch die Basisfunktionen auf höheren Verfeinerungsleveln darstellen läßt. Das ist bei den in dieser Arbeit verwendeten LAGRANGE-Elementen immer gegeben. Bei anderen Finite-Elemente-Klassen ist die Aufstellung der rechten Seite von (4.1) nicht so einfach und im allgemeinen auch nicht exakt durchzuführen, da Quadraturen verwendet werden müssen.

4.3 Behandlung von Dirichlet-Randwerten

Zur Illustration der Behandlung von DIRICHLET-Randwerten betrachten wir beispielhaft die Gleichung

$$\begin{aligned} -\Delta u &= f && \mathbf{x} \in \Omega, \\ u &= g && \mathbf{x} \in \partial\Omega. \end{aligned}$$

Durch Wahl einer Funktion u_0 mit $u_0 = g$ auf $\partial\Omega$ läßt sich ebensogut das folgende Problem lösen: suche $\tilde{u} = u - u_0$, so daß

$$\begin{aligned} -\Delta u &= f + \Delta u_0 && \mathbf{x} \in \Omega, \\ u &= 0 && \mathbf{x} \in \partial\Omega \end{aligned}$$

gilt. Die Randwerte sind damit homogen geworden. In den meisten Fällen ist die Randfunktion g nicht durch die Spur einer Finite-Elemente-Funktion $u_{0,h}$ darstellbar, so daß man mit einer approximativen Behandlung der Randwerte auskommen muß, beispielsweise durch Verwendung einer Projektion oder Interpolation auf die Spur des Ansatzraumes, $\tilde{g} = P_{V^0} g$ oder $\tilde{g} = \mathcal{I}_{V^0} g$, und die dazugehörige Funktion $\tilde{u}_0$. In der Praxis ist die Wahl einer geeigneten Funktion $\tilde{u}_0$ jedoch unpraktisch; stattdessen gibt es zwei Ansätze zur direkten Behandlung der ursprünglichen Gleichung, die beide darauf beruhen, Systemmatrix und rechte Seite auf allen Zellen des Gebiets und damit mit allen Freiheitsgraden aufzubauen und anschließend den durch die Randbedingungen festgelegten Freiheitsgraden auf dem Rand eine spezielle Behandlung zukommen zu lassen:

Filtern: diese Methode bietet sich an, wenn die linearen Gleichungen durch iterative Verfahren gelöst werden. Die meisten iterativen Löser lassen sich als Defektkorrekturverfahren schreiben, das heißt es wird ein Anfangswert berechnet, der anschließend jeweils nur noch durch additive Korrekturen verändert wird. Das Filtern funktioniert nun so, daß man dem Startvektor die richtigen Werte für die Freiheitsgrade auf dem Rand aufzwingt und bei den Korrekturvektoren die Einträge für diese Freiheitsgrade jeweils auf Null setzt (filtern), was einer modifizierten Matrix-Vektor-Multiplikation entspricht, die der Multiplikation mit der Systemmatrix aus dem modifizierten Problem entspricht.

Dieser Ansatz wurde in DEAL gewählt. Er ist in vielen Fällen, wo nur wenige Iterationen zur Lösung gebraucht werden, bedeutend schneller als der zweite Weg, verlangt aber viel Sorgfalt bei der Implementation der linearen Algebra.

Elimination aus der Matrix: diese Methode ist bei allen Lösern möglich, insbesondere auch für direkte Löser. Sie beruht darauf, daß die betroffenen Freiheitsgrade tatsächlich aus der Matrix und der rechten Seite eliminiert werden, wodurch sich die Dimension des linearen Gleichungssystems auf die Anzahl der tatsächlich freien Freiheitsgrade reduziert. Da das Umkopieren in eine neue, kleinere Matrix und einen entsprechenden Vektor zu unpraktisch und mit zu großem Speicheraufwand verbunden wäre, geht man wie folgt vor: für jeden Freiheitsgrad auf dem Rand wird zuerst die entsprechende Zeile aus der Matrix gestrichen, lediglich der Diagonaleintrag bleibt stehen. Der Wert der rechten Seite wird so gesetzt, daß der Lösungsvektor den richtigen Wert bekommt. Da in dieser Zeile nun nur noch ein einziger Eintrag steht, ist der Wert des Freiheitsgrades tatsächlich festgelegt. Um die Matrix wieder positiv definit und vor allem symmetrisch zu machen wird nun mit dieser Zeile ein GAUSS-Eliminations-Schritt gemacht, um auch alle Einträge der entsprechenden Spalte zu Null zu machen; dabei wird auch die rechte Seite entsprechend modifiziert. Durch den Eliminationsschritt ist der Freiheitsgrad vom Rest des Gleichungssystems abgekoppelt.

Für einige Iterationsverfahren ist die Durchführung des Eliminationsschritts nicht einmal nötig, wie man sich überlegen kann. Der Allgemeinheit halber wurde aber kein Gebrauch von dieser Tatsache gemacht.

Schreibt man sich das Verfahren genau hin, dann entsprechen Matrix und Vektor abgesehen von den zusätzlichen Zeilen und Spalten genau denen, die entstünden wenn man das oben beschriebene modifizierte kontinuierliche Problem lösen würde, wobei für $\tilde{u}_0$ die Funktion gewählt wird, die nur auf der äußersten Schicht von Zellen lebt und auf dem ganzen Gebiet stetig ist (diese Funktion ist eindeutig).

Der beschriebene Weg wurde in DEAL.II gewählt. Er ist etwas allgemeiner als das Filtern, hat jedoch den Nachteil, daß der Eliminationsschritt in schwach besetzten Matrizen mit erheblichem Rechenaufwand in den verteilten Datenstrukturen verbunden ist. Vom Standpunkt der Rechenzeit lohnt er sich nur für größere Probleme, bei denen viele Iterationsschritte nötig sind, da der Aufwand für das Filtern in jedem Schritt anfällt, während die Elimination nur einmal nötig ist. In der Summe ist der Implementationsaufwand wohl geringer, da er zentral einmal bereitgestellt wird und nicht für jeden Fall die Matrix-Vektor-Multiplikation modifiziert werden muß. Darüberhinaus

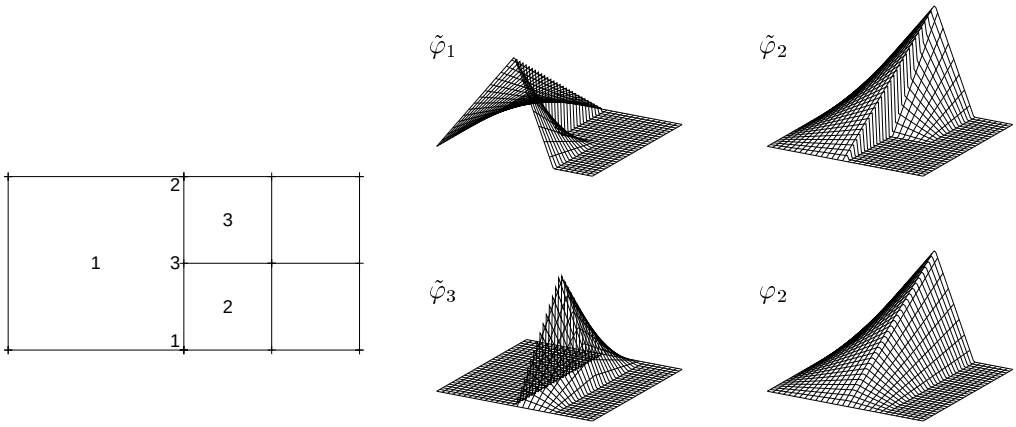

Abbildung 4.1: *Schema zur Elimination von hängenden Knoten (links). Die ‚unechten' Ansatzfunktionen $\tilde{\varphi}_1$, $\tilde{\varphi}_2$ und $\tilde{\varphi}_3$, sowie die ‚echte' Ansatzfunktion φ_2*

ist es durch die zentrale Bereitstellung möglich, eine gut optimierte Implementation anzubieten, die von der internen Eigenschaft von verschiedenen Matrizentypen Gebrauch macht, was bei Anwenderprogrammen nicht sinnvoll wäre und somit beim Filtern nicht geht.

Weder DEAL noch DEAL.II sind auf die jeweils beschriebene Methode festgelegt; es sind lediglich die im Moment implementierten Verfahren.

Die Werte der zu eliminierenden Freiheitsgrade werden gemäß der Interpolation der Randwerte auf die Restriktion des Finite-Elemente-Raums auf den Rand bestimmt. Die Interpolation wurde hier der L^2-Projektion vorgezogen, da die Berechnung der Projektion bei gekrümmten Rändern aufwendiger ist. Andererseits ist die Projektion bei unstetigen Randwerten unumgänglich, um die Konvergenzordnung zu erhalten; aufgrund der für diese Arbeit vorausgesetzten Glattheit der Randwerte ist die Interpolation aber ausreichend.

Die Behandlung von NEUMANN-Randwerten geschieht durch die Wahl einer modifizierten Bilinearform $a(\cdot,\cdot)$ und bedarf keine Manipulation der linearen Gleichungssysteme. Die meisten anderen Randbedingungen, beispielsweise absorbierende Randbedingungen, lassen sich ebenso behandeln, oder sie erlauben die Berechnung von Werten für die Freiheitsgrade auf dem Rand aus den Werten zum vorigen Zeitschritt. Im letzteren Fall lassen sich dann wieder die beiden oben angeführten Techniken verwenden.

4.4 Behandlung von hängenden Knoten im Ort

Bei der Verwendung von Gittern mit hierarchischer Verfeinerung einzelner Zellen ohne Verwendung von Abfangelementen bleiben auf den Schnittstellen zwischen verfeinerten und unverfeinerten Bereichen hängende Freiheitsgrade zurück. Um bei den verwendeten konformen Finite-Elemente-Ansätzen vernünftige Lösungen zu garantieren, muß an diesen Stellen dafür gesorgt werden, daß die Lösungsfunktion stetig bleibt.

Die Behandlung der hängenden Knoten sei an Abbildung 4.1 illustriert, wobei der Einfachheit halber von bilinearen Elementen ausgegangen sei. Die Behandlung höherer Ansatzgrade und höherer Dimensionen als zwei verläuft ganz analog, jedoch sind die Gleichungen weniger übersichtlich.

Da, wie gesagt, die Funktion am Knoten 3 stetig sein soll, muß wegen der Linearität der Ansatzfunktionen auf den Kanten der Koeffizient der Basisfunktion zum Knoten 3 genau der Mittelwert der Koeffizienten der Knoten 1 und 2 sein. Im Punkt 3 befindet sich damit kein echter

Freiheitsgrad, aus algorithmischen Gründen werden aber Matrizen und Vektoren so aufgebaut als wäre dort einer, da sich dann Matrizen und Vektoren zellweise aufbauen lassen, ohne daß schon beim Aufbau Rücksicht auf die hängenden Knoten genommen werden muß. In einem zweiten Schritt werden in den Vektoren die Einträge zur Basisfunktion 3 jeweils hälftig auf die Einträge zu den Basisfunktionen 1 und 2 verteilt. Analog werden in den Matrizen die Spalten und Zeilen zur Basisfunktion 3 auf die Zeilen und Spalten der anderen Basisfunktionen verteilt. Die zum hängenden Knoten gehörige Zeile und Spalte wird auf Null gesetzt um den unechten Freiheitsgrad vom Rest des Systems zu isolieren.

Anschaulich ist dieses Vorgehen folgendermassen verständlich: die echten Basisfunktionen φ_1 und φ_2 zu den Punkten 1 und 2 haben ihren Träger jeweils auf den Elementen 1 bis 3, wie man erwarten würde wenn der hängende Knoten 3 gar nicht da wäre. Diese Basisfunktionen lassen sich jedoch durch die Basisfunktionen $\tilde{\varphi}_1$ und $\tilde{\varphi}_2$ zu den entsprechenden Punkten plus jeweils der Hälfte der Basisfunktion $\tilde{\varphi}_3$ zum mittleren Knoten darstellen:

$$\varphi_1 = \tilde{\varphi}_1 + \frac{1}{2}\tilde{\varphi}_3,$$
$$\varphi_2 = \tilde{\varphi}_2 + \frac{1}{2}\tilde{\varphi}_3.$$

$\tilde{\varphi}_1$ und $\tilde{\varphi}_2$ haben ihre Träger auf den Zellen 1 und 2 bzw. 1 und 3, $\tilde{\varphi}_3$ lebt auf den Elementen 2 und 3. Baut man nun Matrizen und Vektoren auf den Zellen ohne Wissen über hängende Knoten auf, das heißt mit den Basisfunktionen $\tilde{\varphi}_i$, so ist klar, wie die anschließende Aufsummierung zu den Matrizen mit den echten Basisfunktionen φ_j stattzufinden hat. In höheren Dimensionen oder bei höheren Ansatzgraden lassen sich die Ansatzfunktionen auf der eingeschränkten Kante durch eine Matrix

$$\varphi_i = \sum_j c_{ij}\tilde{\varphi}_j$$

mit den unechten Freiheitsgraden auf den Teilkanten verknüpfen, so daß die Elimination aus Matrizen und Vektoren dimensions- und ansatzgradunabhängig auf rein algebraischer Basis durchgeführt werden kann.

Die Implementation der oben beschriebenen Manipulation von Matrizen und Vektoren ist im Prinzip analog zur Behandlung von Randwerten durchführbar. Es kommen wieder die beiden Möglichkeiten Filtern und direkte Manipulation in Frage, wobei in DEAL.II die direkte Elimination (*Kondensation*) gewählt wurde. Nach dem Lösen der Gleichungssysteme ist darauf zu achten, die Freiheitsgrade wieder zu verteilen, das heißt den Eintrag zum Knoten 3 auf die Hälfte der Einträge der beiden benachbarten Freiheitsgrade zu setzen.

4.5 Steuerung des Gitters

Die vernünftige Steuerung der Gitter auf den verschiedenen Zeitebenen ist ein Thema über das alleine wohl Diplomarbeiten zu schreiben wären. Im folgenden Abschnitt sollen einige der Probleme und die unternommenen Versuche zur Lösung kurz beschrieben werden. Sie sind im wesentlichen unabhängig von Zeitschrittverfahren, Verfeinerungskriterium, Fehlerschätzer und ähnlichen Parametern, so daß diese nicht bei allen Beispielen angegeben sind.

4.5.1 Auseinanderdriften der Gitter

Die Erzeugung von Gittern findet im wesentlichen wie folgt statt: auf einem gegebenen groben Gitter $\mathbb{T}_{h_0}$, das in allen Zeitschritten identisch ist, werden die Lösungen $\mathbf{w}_{h_0}$ und $\bar{\mathbf{w}}_{h_0}$ des Vorwärts- und Rückwärtsproblems berechnet. Aus diesen beiden Lösungen werden für jeden Zeitschritt die zu verfeinernden/zu vergröbernden Zellen bestimmt und daraus Gitter $\mathbb{T}^n_{h_1}$ erzeugt, die jetzt auf

den einzelnen Zeitschritten n im allgemeinen unterschiedlich sind. Auf diesen Gittern werden jetzt wieder die Lösungen u_{h_1} und z_{h_1} berechnet, daraus die Gitter $\mathbb{T}^n_{h_2}$, usw.

Im Endeffekt wird jedes Gitter $\mathbb{T}^n_{h_l}$ aus $\mathbb{T}^n_{h_{l-1}}$ und den Lösungen $\mathbf{w}^n_{h_{l-1}}$ und $\bar{\mathbf{w}}^n_{h_{l-1}}$ bestimmt. Auf diese Art besteht praktisch keine Kopplung zwischen den Gittern zu verschiedenen Zeitschritten und in der Tat unterscheiden sich die Gitter $\mathbb{T}^n_{h_l}$ und $\mathbb{T}^{n+1}_{h_l}$ oft beträchtlich. Das ist dann ein Problem, wenn sich die Verfeinerungslevel einiger Zellen zwischen den beiden Gittern zu stark ändern, da dann die Projektion der Lösung vom alten Gitter auf das neue leidet (technisch ist ein Transfer über mehrere Verfeinerungsebenen ohne weiteres möglich, wie in Abschnitt 4.2 beschrieben).

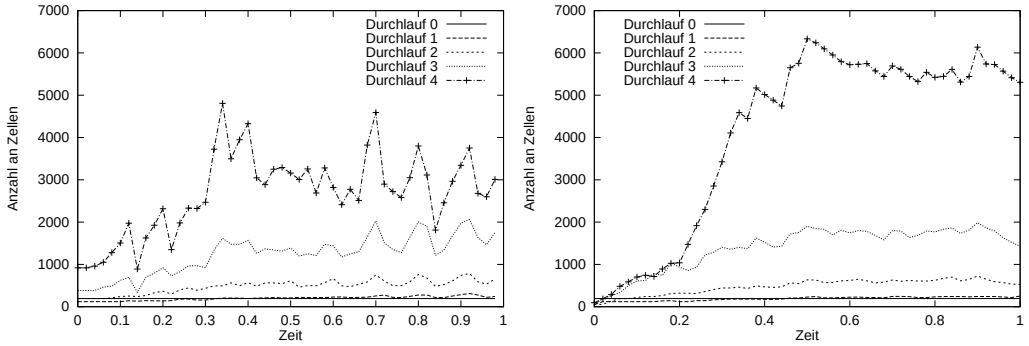

Abbildung 4.2: *Anzahl an Gitterzellen in verschiedenen Durchläufen. Die Gitter entstanden durch Verfeinerung nach einem Energiefehlerschätzer beim einem Beispiel, bei dem eine Welle vom Rand aus durch das Gebiet läuft. Da die Anfangsbedingungen konstant sind, ist die Zellzahl am Anfang sehr klein; die Randbedingungen sind ab $t = 0.4$ konstant Null, so daß sich die Welle nur verschiebt und die Zellzahl in etwa konstant bleiben sollte. Links ist nur die Anzahl der Verfeinerungslevel begrenzt, um die sich einzelne Zellen in aufeinanderfolgenden Zeitschritten unterscheiden dürfen, rechts sind alle beschriebenen Kopplungen zwischen aufeinanderfolgenden Gittern verwendet. Die absolute Zahl der Zellen ist aufgrund unterschiedlicher Verfeinerungsparameter nicht vergleichbar.*

Um die Divergenz der Gitter zu verringern, wurde versucht, die $\mathbb{T}^n_{h_l}$ für aufeinanderfolgende n möglichst eng aneinanderzukoppeln. Dazu werden jeweils zusätzliche Zellen zur Verfeinerung markiert, um zu verhindern, daß sich die Verfeinerungsstufen zweier Zellen in aufeinanderfolgenden Gittern um mehr als eins unterscheiden[1] und es wurden einige Gitterglättungsalgorithmen implementiert, um isolierte verfeinerte oder unverfeinerte Zellen und einige zusätzliche Fälle zu eliminieren. Durch diese Schritte wurde das Gitter $\mathbb{T}^n_{h_l}$ außer an $\mathbb{T}^n_{h_{l-1}}$ auch an die Gitter $\mathbb{T}^{n\pm1}_{h_l}$ gekoppelt. Das dämpft zwar die schlimmsten Divergenzen etwas, die Anzahl an Zellen in den einzelnen Gittern bleibt im wesentlichen aber unkorreliert, wie Abbildung 4.2 zeigt. Kleine Oszillationen in den ersten Durchläufen verstärken sich in den darauffolgenden, so daß Variationen der Zellzahl in aufeinanderfolgenden Zeitschritten bis zu einem Faktor zwei durchaus vorkommen, bei vier oder fünf Zeitschritten auseinanderliegenden Gittern auch Faktoren drei oder noch mehr; ähnliche Oszillationen wurden auch von HARTMANN [19] beobachtet. Ein Mechanismus scheint zu sein, daß die Kopplung die Gitter für einige Zeit einigermaßen konstant hält, bis sie sich ruckartig ändern.

Eine Lösung dieses Problems ist wohl nur durch eine stärkere Kopplung der Gitter in jedem Durchlauf möglich, zum Beispiel in dem $\mathbb{T}^n_{h_l}$ auch an $\mathbb{T}^{n\pm2}_{h_l}$ gekoppelt wird. Die Schwierigkeit hier ist algorithmischer Natur, weil kein vernünftiges Kriterium für den Abstand zweier Gitter voneinander und seine Begrenzung bekannt scheint. Zu eng dürfen die Gitter auf der anderen Seite auch nicht aneinander gekoppelt sein, da sonst die Vorteile des mitwandernden Gitters verloren

[1] Diese Bedingung läßt sich nicht immer garantieren, da das Gitter in einem Zeitschritt n vom darauffolgenden verändert werden kann, was dann aber unter Umständen Auswirkungen auf das Gitter zum Zeitschritt $n-1$ haben könnte. Solche Änderungen über mehr als einen Zeitschritt werden aus algorithmischen Gründen nicht beachtet und können dazu führen, daß sich Gitter um mehr als eine Verfeinerungsebene unterscheiden.

gehen; so führt beispielsweise die Limitierung der Differenz der Verfeinerungslevel einzelner Zellen des aktuellen Gitters mit dem vorletzten (statt nur mit dem letzten) Gitter im wesentlichen zu einem Einfrieren der Gitter, was nicht erwünscht sein kann.

Ein anderer Ansatz wäre die Verwendung geeigneterer Verfeinerungskriterien. Im Beispiel in Abbildung 4.2 wurden in jedem Zeitschritt die Zellen mit dem größten Fehlerbeitrag zur Verfeinerung markiert, die zusammen $q_r = 95$ Prozent des Fehlers ausmachen, während die Zellen zur Vergröberung markiert wurden, die die unteren $q_c = 2$ Prozent ausmachen. Dieses Verfahren ist wohl zu grob, um wirklich effizient zu arbeiten,[2] weshalb auch kompliziertere Verfahren untersucht wurden.

In Ermangelung fundierterer Ansätze wurde eine Anzahl heuristischer Regeln implementiert, die wesentlich bessere, jedoch noch nicht völlig befriedigende Ergebnisse liefert. Ein Beispiel ist in Abbildung 4.2 rechts wiedergegeben, bei dem die Zellzahl in etwa dem entspricht, was man gerne hätte; die Verläufe sind allerdings nicht in allen Fällen so gut. Die Regeln zur Gittersteuerung sind im folgenden als Pseudoprogramm angegeben:

1. Markiere die Zellen mit dem größten Fehler η_K zur Verfeinerung, die zusammen den Anteil q_r am Gesamtfehler η haben.

2. Markiere die Zellen mit dem kleinsten Fehler η_K zur Vergröberung, die zusammen den Anteil q_c am Gesamtfehler η haben.

3. Lösche die Vergöberungsflags von Zellen, bei denen nicht alle Kinder einer Zelle markiert sind.

4. Markiere zusätzliche Zellen, so daß das aktuelle und das Gitter des letzten Zeitschrittes gewissen Regularitätsforderungen genügen.

5. Markiere in diesem und im letzten Gitter zusätzliche Zellen, so daß sich die beiden Gitter um jeweils höchsten eine Verfeinerungsstufe auf jeder Zelle unterscheiden.

6. Wenn im ersten Verfeinerungsdurchlauf, dann *Ende*.

7. Zähle die Anzahlen N^n und N^{n-1} der durch die Markierungen auf diesem und auf dem letzten Gitter entstehenden Zellen.

8. Berechne eine obere und untere Schranke für die Anzahl an Zellen im aktuellen Gitter gemäß folgender Formeln:

$$N^n_{max} = N^{n-1}(1 + \delta_\uparrow), \qquad N^n_{min} = N^{n-1}(1 - \delta_\downarrow).$$

   Falls im ersten Verfeinerungsdurchlauf oder die alte Zellzahl auf dem aktuellen Gitter kleiner als 200 ist, verwende $3\delta_{\uparrow\downarrow}$ statt $\delta_{\uparrow\downarrow}$. Falls im zweiten Durchlauf oder die Zellzahl kleiner als 300, dann verwende $2\delta_{\uparrow\downarrow}$.

9. Falls $N^n > N^n_{max}$ und $N^n > 200$, dann lösche solange Verfeinerungsmarkierungen, bis $N^n < N^n_{max}$.

10. Sonst: Falls $N^n < N^n_{min}$, dann markiere solange Zellen zur Verfeinerung, bis $N^n > N^n_{min}$.

11. Führe die Schritte 3,4 und 5 erneut aus.

12. Falls nicht $N^n_{min} < N^n < N^n_{max}$, dann gehe zu 7.

[2] Im allgemeinen wird als optimal angesehen, wenn $q_r = 50$ Prozent gewählt wird. Abgesehen davon, daß dieser Wert allerdings nur für einfache statische Probleme als gut gilt, bräuchte man auch wesentlich mehr Verfeinerungsdurchläufe, um zum letztendlich brauchbaren Gitter zu gelangen, was aus Rechenzeitgründen nicht praktikabel erscheint.

Die Variation von $\delta_{\uparrow\downarrow}$ in Punkt 8 ist notwendig, um das Gitter in den ersten Verfeinerungsschritten nicht zu stark einzuschränken. Aufgrund der in Schritt 11 durchgeführten Aktionen ist nicht mehr garantiert, daß sich die Zellzahl N^n im Zielkorridor $[N^n_{min}, N^n_{max}]$ befindet, so daß im letzten Schritt eine Iteration gestartet wird; im allgemeinen sind zwei Durchläufe der Punkte 7 bis 11 ausreichend. Als geeignet für $\delta_\uparrow$ und $\delta_\downarrow$ haben sich Werte in der Gegend von 0.1 und 0.03 erwiesen. $\delta_\downarrow$ sollte kleiner als $\delta_\uparrow$ sein, da die Approximation bei einer Verkleinerung der Zellzahl leidet, während eine Vergrößerung im allgemeinen keine negativen Auswirkungen hat. $\delta_\uparrow$ sollte trotzdem nicht zu groß sein, da für das duale Problem die umgekehrte Zeitrichtung gilt und da ein zu starker Anstieg im allgemeinen eine starke Verkleinerung der Zellzahl nach sich zieht.

Zur Berechnung der Fehlergröße η_K wird entweder ein Energiefehlerindikator oder der duale Fehlerschätzer verwendet. Beide messen jedoch nur den Fehler in der primalen Lösung. Soll der duale Fehlerschätzer verwendet werden, so ist es notwendig, daß auch die duale Lösung mit hinreichender Genauigkeit berechnet wird; es wird daher der Energiefehlerindikator auch auf die duale Lösung angewendet, die zellweisen Fehlergrößen η^d_K aus diesem Prozeß geeignet skaliert, so daß der maximale Fehler pro Zelle $\max_K \eta^p_K$ aus dem primalen Problem und $\max_K \eta^d_K$ aus dem dualen Problem gleich groß sind und schließlich wird aus diesen beiden Größen eine gewichtete Summe gebildet, die als Verfeinerungskriterium verwendet wird:

$$\eta_K = \alpha\eta^p_K + \frac{\max_{K'} \eta^p_{K'}}{\max_{K'} \eta^d_{K'}}\eta^d_K$$

In der Praxis hat es sich erwiesen, daß eine geeignete Gewichtung den primalen Fehlerschätzer etwa $\alpha = 4$ bis 8 Mal so stark bewertet wie den dualen Indikator. Bei der Berechnung der Sprünge der Normalenableitung für den Energiefehlerschätzer ist beim dualen Problem die Variable $\bar{v}_h$ zu verwenden, wie man aus (2.21) erkennt.

Aus dem gleichen Grund wie oben ist es bei nichtlinearen Zielfunktionalen, bei denen eine Linearisierung um die numerisch erhaltene Lösung notwendig ist, wichtig, daß die ersten Verfeinerungsschritte mit einem Energiefehlerschätzer statt dem dualen Schätzer durchgeführt werden. Im anderen Fall kann es passieren, daß die duale Lösung aufgrund der mangelnden Genauigkeit der numerischen primalen Lösung stark von der Lösung des exakten dualen Problems abweicht. Die Verfeinerung wird dann in Gebieten durchgeführt, die für das Ergebnis nicht relevant sind, so daß die primale Lösung im nächsten Schritt genauso falsch ist wie im ersten; das Verfeinerungsverfahren konvergiert dann nicht.

Die Verwendung vorgelagerter Verfeinerungsschritte mit einem Energiefehlerindikator widerspricht aber nicht dem Ansatz, die Adaptivität mit einem dualen Problem anzugehen; der Grund ist, daß die Lösbarkeit eines Problems auf eine bestimmte Genauigkeit praktisch ausschließlich durch die von den letzten ein bis zwei Verfeinerungsdurchgängen benötigten Ressourcen (Rechenzeit und Speicher) bestimmt ist, während die davorliegenden Durchgänge weitgehend vernachlässigbar sind. Gelingt es also mit einem dualen Problem den benötigten maximalen Ressourcenbedarf zu senken, so ist die Verwendung suboptimaler Verfahren in den vorherigen adaptiven Schritten akzeptabel.

Das Verfahren zur Steuerung der Zellzahl ist in der beschriebenen Form nicht ganz befriedigend. Der Grund liegt darin, daß der Algorithmus versagt, sobald der Zeitschritt sehr klein wird, da die Abweichungen $\delta_{\uparrow\downarrow}$ eigentlich proportional zur Zeitschrittweite sein sollten, um eine Beschränkung der Änderungs*rate* zu garantieren. Problematisch dabei ist, daß der Zielkorridor für die Zellzahl nicht kleiner als etwa fünf Prozent sein sollte, da die resultierende Zellzahl bei Markierung einiger Zellen nicht genau vorhersagbar ist und es sonst nicht möglich ist, das Ziel zu erreichen. Lösen läßt sich dieses Problem nur bei Kopplung von mehr als zwei Zeitschritten aneinander.

4.5.2 h-Abhängigkeit der Dispersionsrelation

Eines der in der Praxis unangenehmsten Probleme ist die Abhängigkeit der Ausbreitungsgeschwindigkeit von der lokalen Gitterweite. Dieses Problem ist wohlbekannt und auch theoretisch gut untersucht, siehe zum Beispiel [1]. Der Effekt ist besonders ausgeprägt für lineare Elemente. Im

wesentlichen läuft die Abhängigkeit der Dispersionsrelation von der Gitterweite h darauf hinaus, daß Wellen auf gröberen Gittern schneller laufen als auf feinen und daß die Geschwindigkeit erst im Limes $h \to 0$ gegen die Ausbreitungsgeschwindigkeit des kontinuierlichen Problems geht. Auch ist die Geschwindigkeit abhängig von der Wellenlänge und davon, ob die Ausbreitung in Richtung des Gitters oder diagonal dazu verläuft. Letzteres ist in Abschnitt 5.2 numerisch untersucht.

Abbildung 4.3 illustriert die vom Verhältnis Gitterweite zu Wellenlänge abhängige Ausbreitungsgeschwindigkeit. Gerechnet wurde auf dem Gebiet $[0,3] \times [0,1]$, wobei unten und oben homogene NEUMANN-Randwerte gesetzt waren und durch Steuerung von $u(x=0,y,t) = \sin(2.5\pi t)$ für $t < 0.4$ eine Welle vom linken Rand in das Gebiet läuft. Das Signal am Ort $\mathbf{x} = (1, 0.5)$ sollte genau diesem halben Sinus entsprechen und zum Zeitpunkt $t = 1$ eintreffen. Abbildung 4.3 zeigt, daß das Signal für zu grobe Gitter zu früh eintrifft und daß seine Form durch den Vorläufer stark verzerrt ist. Um den Fehler in der Wellenausbreitungsgeschwindigkeit zu bestimmen nimmt man den Schwerpunkt der Energie des Signals:[3]

$$\langle t \rangle = \frac{\int t s(t)^2 \, dt}{\int s(t)^2 \, dt}, \qquad s(t) = u(1, 0.5, t).$$

Das Verhältnis aus der daraus berechneten Ausbreitungsgeschwindigkeit zum richtigen Wert ist in der Abbildung rechts in Abhängigkeit vom Verhältnis zwischen Wellenlänge (die hier 0.8 ist) zur Gitterweite dargestellt. Die Abweichung geht etwa mit der zweiten Potenz von h gegen Null.

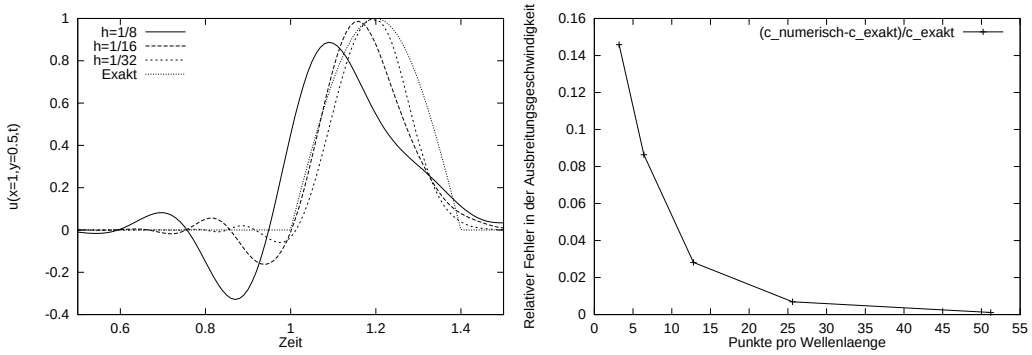

Abbildung 4.3: *Abhängigkeit der Ausbreitungsgeschwindigkeit von der Gitterweite. Links: Signal $u(x = 1, y = 0.5, t)$ für verschiedene Verfeinerungen. Rechts: Fehler in der Ausbreitungsgeschwindigkeit $c_{numerisch}$ gegenüber dem exakten Wert c_{exakt}.*

In der Ingenieurliteratur (vgl. beispielsweise [35]) findet man oft den Ratschlag, mindestens 10 bis 20 Gitterpunkte pro Wellenlänge zu verwenden. Die Ergebnisse der Abbildung stimmen damit in etwa überein, wenn man beachtet daß aufgrund der abgeschnittenen Sinuswelle auch höhere Frequenzen als $\nu = 2.5\pi$ vorhanden sind.

Im Kontext adaptiver Verfeinerung spielt der beschriebene Effekt insofern eine Rolle, als die Gitter für den zweiten Durchlauf auf der Basis der Lösung des ersten Durchlaufs bestimmt wird. Da die Welle im ersten Durchlauf aber zu schnell war, bleibt die Welle im zweiten Durchlauf nur am Anfang auf dem verfeinerten Bereich, fällt aber irgendwann hinter diesen Bereich zurück. Solange die Welle auf dem verfeinerten Bereich bleibt, halten sich die Probleme in Grenzen, da die Geschwindigkeit auf dem feineren Gitter schon recht nahe an die tatsächliche herankommt und das Gitter im dritten Durchlauf gut brauchbar ist. Sobald die Welle aber vom verfeinerten Bereich „hinten herunter gefallen“ ist, läuft sie wieder auf dem groben Bereich und damit wieder mit der falschen Geschwindigkeit; damit wird der Bereich hinter dem im ersten Durchlauf verfeinerten

[3] Diese Wahl ist relativ willkürlich, man könnte genauso gut einen der Nulldurchgänge oder eines der Maxima der Welle nehmen; da die Energie aber im allgemeinen die Meßgröße ist und darüberhinaus die Gruppengeschwindigkeit der Welle kennzeichnet, wird sie zur Definition der Ausbreitungsgeschwindigkeit verwendet.

Gebiet verfeinert, dieser Bereich bewegt sich aber immer noch zu schnell, weshalb wir im nächsten Durchlauf das gleiche Problem haben werden, wenn auch später.

Im Endeffekt benötigt man mehrere Durchläufe, bis allein der einmal verfeinerte Bereich mit dem tatsächlichen Ort der Welle übereinstimmt. Die gleichen Probleme treten auch auf höheren Verfeinerungsstufen auf, wenn auch nicht ganz so markant.

Ansätze zur Lösung dieses Problems sind einerseits, gleich auf einem relativ feinen Gitter anzufangen und bei den darauffolgenden viele Zellen zu vergröbern, andererseits am Anfang eine gewisse Anzahl von Durchläufen durchzuführen, bei denen die Verfeinerung sehr vorsichtig gehandhabt wird, um eine zu starke Verfeinerung an den Orten zu verhindern, wo die exakte Lösung, im Gegensatz zur numerischen Lösung auf zu groben Gittern, glatt ist. Bei Problemen mit kurzen Wellenlängen ist der erste Weg nicht gangbar, beispielsweise würde die Forderung nach nur 10 Gitterpunkten pro Wellenlänge bereits auf dem ersten Gitter bei dem in Abschnitt 5.3 gerechneten Beispiel 10^6 Zellen pro Zeitschritt nach sich ziehen.

Im wesentlichen verhindert dieses Problem die mehrfache Verfeinerung von Zellen mit großem Fehlerindikator in einem Durchgang, wie es bei elliptischen und parabolischen Problemen durchführbar ist (vgl. [22, 19]). Ein dritter Weg ist die Verwendung von finiten Elementen höherer Ordnung, die eine geringere Dispersion zeigen, vergleiche [1] und Abschnitt 5.2.

4.5.3 Orts-Zeit-Adaptivität

Eine optimale Strategie zur Gitterverfeinerung würde neben der Ortsgitter- auch die Zeitschrittweite adaptiv anpassen. Diese Art der Adaptivität ist ausgehend von den in Abschnitt 2.4 vorgestellten Konzepten leicht machbar und in [19] für die Wärmeleitungsgleichung auch in der Praxis gezeigt; allerdings ist die Bedeutung der Zeitschrittweitenadaption bei der Wellengleichung nicht so groß wie bei der Diffusionsgleichung, wofür im wesentlichen zwei Gründe ausschlaggebend sind:

- *Keine unterschiedlichen Zeitskalen:* Bei der Wellengleichung sind die Zeitskalen im allgemeinen konstant; insbesondere gibt es keine stationären Zustände, so daß sich eine wesentliche Vergrößerung der Zeitschrittweite meistens verbietet.

- *Kein Informationsverlust:* Bei der Wärmeleitungsgleichung unterliegt Information starker Dämpfung, so daß Zielfunktionale, die die Lösung am Endzeitpunkt auswerten, große Zeitschritte am Anfang zulassen. Da bei der Wellengleichung kein Informationsverlust vorhanden ist, verlangt eine Auswertung zu beliebiger Zeit eine möglichst exakte Rechnung zu *allen* früheren Zeitpunkten.

Aus den genannten Gründen erschien die Implementation zeitschrittadaptiver Methoden nicht so relevant. Ein wesentlich wichtigerer Punkt wäre jedoch der Vergleich von Fehlerinformation über verschiedene Zeitschritte hinweg: die Fehlerschätzer liefern einen Indikatorwert für jede Raum-Zeit-Zelle; die Entscheidung, ob eine Zelle verfeinert bzw. vergröbert wird, findet jedoch im verwendeten Programm nur durch Vergleich mit den anderen Zellen des gleichen Zeitschritts statt. Durch dieses Verfahren werden auch Zellen verfeinert, deren Indikator wesentlich kleiner ist als die Indikatoren von Zellen auf anderen Zeitschritten, die nicht verfeinert werden. Im Endeffekt vergrößert sich so die Zahl der für eine gegebene Genauigkeit benötigte Zahl an Freiheitsgraden; in Abschnitt 3.3 ist ein Beispiel angeführt, wo die Einsparung erheblich wäre.

Trotzdem wurde im Rahmen dieser Arbeit von einer Implementation dieses Quervergleichs Abstand genommen, da der Vergleich von bis zu 50.000.000 Indikatoren eine erhebliche programm- und speichertechnische Herausforderung darstellen würde. Für zukünftige Arbeiten ist diese Optimierung aber wünschenswert und erfolgversprechend.

Ein weiteres, nicht verfolgtes Ziel ist die Verwendung örtlich unterschiedlicher Zeitschrittweiten. Ebenso wie die Ortszellen lokal verfeinert werden, sollte es möglich sein, die Orts-Zeit-Zellen einzeln zu verfeinern. Dies würde es erlauben, die Orts-Zeit-Gitter nahezu optimal an das Problem anzupassen. Neben dem zu erwartenden erheblichen Implementationsaufwand ist dieser Ansatz allerdings nur dann möglich, wenn es gelingt, die verwendete Formulierung analog zu Abschnitt 2.3 in die einzelnen, dann lokalen Zeitschritte zu entkoppeln. Diese Entkopplung ist jedoch schwierig,

so daß sich in der Literatur nur wenige praxisrelevante Ansätze zur Verwendung von Raum-Zeit-Elementen finden und auch diese vorwiegend nur für eine Raumdimension (für ein solches Beispiel vergleiche [16]).

Kapitel 5

Weitere numerische Beispiele

In diesem Kapitel sollen einige weitere Beispiele untersucht werden, um Vor- und Nachteile der adaptiven Methoden gegenüber globaler Verfeinerung zu ergründen. Die meisten der Beispiele haben keine direkte Anwendung, dienen aber prototypisch als reduzierte Fälle realer Probleme. Dazu wurden Zielfunktionale ausgewählt, um ein Spektrum möglicher Anwendung abzudecken, darunter Punkt- und Linienauswertungen über die Zeit integriert und ein Raum-Zeit-Integral. Zusammen mit den Gebietsintegralen des letzten Kapitels sind somit außer Punkt- und Linienauswertungen ohne Zeitintegral alle Typen von Funktionalen abgedeckt; die Punktauswertung ist allerdings im allgemeinen kein wohldefiniertes Funktional auf dem Lösungsraum.

Als letztes Beispiel wird ein Effekt untersucht, der bei der Verwendung adaptiver Gitter zurecht oft kritisiert wird, namentlich die teilweise Reflexion von Wellen an Gitterdiskontinuitäten. Es wird gezeigt, daß die damit zusammenhängenden Probleme mit etwas Sorgfalt umgangen werden können und dann zu wesentlich besseren Ergebnissen führen als bei Verwendung von zeitlich nicht veränderlichen und nur an die Koeffizienten adaptierten Gittern.

In einigen der folgenden Abschnitten ist der Fehler gegen die Rechenzeit aufgetragen. Diese Werte können nur eine grobe Richtschnur sein, da die Programme nicht auf Geschwindigkeit optimiert wurden, vor allem aber da die Rechenzeit bei den adaptiven Rechnungen von einer Vielzahl von Einstellmöglichkeiten abhängt; dazu zählen vor allem das Startgitter, die Strategie, mit der aus den zellweisen Indikatoren bestimmt wird, welche Zellen verfeinert oder vergröbert werden sollen, und schließlich die Anzahl der zu verfeinernden oder vergröbernden Zellen. Die Erfahrung zeigt, daß durch ausgiebiges Spielen mit diesen Werten ein mehrfach geringerer Rechenaufwand möglich ist, was bei den Rechnungen mit global verfeinerten Gittern naturgemäß wegfällt. Es wurde trotzdem darauf verzichtet, hier Optimierungen vorzunehmen, da es mehr um die Demonstration der Konzepte ging; Optimierung der Laufzeit kann nur bei echten Beispielen aus der Praxis sinnvoll sein. Bei der Angabe der Rechenzeit ist zu beachten, daß bei adaptiven Rechnungen die gesamte Rechenzeit angegeben ist, inklusive der vorherigen Durchläufe, die zur Verfeinerung des Gitters benötigt wurden; diese sind bei globaler Verfeinerung natürlich nicht nötig und daher auch nicht angegeben.

Die Erfahrung zeigt, daß die Rechenzeit für jeden Verfeinerungsdurchlauf ungefähr doppelt so lang ist, wie für den vorherigen. Daraus folgt, daß adaptive Verfahren dann effektiver als globale Verfeinerung sind, wenn die Rechenzeit für das letzte Gitter unter der Hälfte der Rechenzeit auf einem global verfeinerten Gitter liegt. Dies ist für einen Energiefehlerschätzer gegeben, falls er kumuliert weniger als etwa die Hälfte der Freiheitsgrade benötigt; für den dualen Schätzer ist der Aufwand zur Lösung des dualen Problems mitzuberücksichtigen, die nötige Ersparnis an Freiheitsgraden läßt sich daher nicht so einfach angeben.

Andererseits könnte die Rechenzeit für das global verfeinerte Gitter wesentlich reduziert werden, wenn man von der dann (bei konstanter Zeitschrittweite) konstanten Systemmatrix am Anfang eine LU- oder ähnliche Zerlegung machen würde, da in jedem Zeitschritt dann nur noch ein Vorwärts-Rückwärts-Einsetzen nötig wäre. Oft ist jedoch der begrenzende Faktor bei Rechnungen mit hoher Genauigkeit der verfügbare Hauptspeicher, was die beschriebene Technik als

nicht praktikabel erscheinen läßt und die Reduktion der Zahl der Freiheitsgrade als wichtiger als die Verringerung der Rechenzeit erscheinen läßt. Typischerweise ist der Bedarf etwa 1000-1200 Byte pro Freiheitsgrad bei linearen Elementen, so daß der Speicherausbau der für diese Arbeit zugänglichen Rechner ein Maximum bei etwa 300.000-500.000 Freiheitsgraden pro Zeitschritt für jede der beiden Variablen setzt; Rechnungen dauern bei dieser Zahl an Freiheitsgraden und 200 Zeitschritten ein bis zwei Tage, was sich beim Einsatz von Mehrgittertechniken deutlich verringern lassen sollte.

5.1 Zeitintegral einer Punktauswertung

Bei der Simulation seismischer Wellen ist die Auswertung der Verschiebung oder der Geschwindigkeit an einem Punkt (z. B. an einer Erdbebenstation, um mit Meßdaten vergleichen zu können) ein häufiger Fall. In diesem Abschnitt seien adaptive und nicht-adaptive Verfahren einander gegenübergestellt, um die Genauigkeit vergleichen zu können.

Zu lösen sei die Wellengleichung mit konstanten Koeffizienten $\rho \equiv a \equiv 1$ auf dem Gebiet $[-1,1]^2$. u sei auf dem Rand Null und die Anfangswerte seien durch eine radialsymmetrische Funktion gegeben:

$$\begin{aligned} u_0(r) &= 0, \\ v_0(r) &= \theta(s-r), \end{aligned}$$

mit $s = \frac{1}{10}$ und der HEAVISIDEschen Sprungfunktion θ. Solange die Welle den Rand nicht erreicht, d. h. für $t < 0.9$, läßt sich das Problem beschreiben, als ob es in der ganzen Ebene gestellt sei.

Als Zielfunktional wurde der über das Zeitintervall $[0, \frac{1}{2}]$ integrierte Wert,

$$\mathcal{J}(\mathbf{w}) = \int_0^{\frac{1}{2}} u(0,t)\ dt,$$

gewählt; durch die spezielle Wahl des Ursprungs als Auswertungspunkt kann man den exakten Wert des Funktionals bestimmen und den exakten Fehler angeben. Während die Auswertung von u an einem Punkt zu einem einzelnen Zeitpunkt auf der Klasse der Lösungen der Wellengleichung nicht definiert ist, ist das Zeitintegral unter schwachen Regularitätsforderungen (keine Diskontinuitäten entlang Linien, deren Tangenten an ein endliches Stück der Kurve auf den Auswertungspunkt zeigen) an die Anfangsbedingungen erlaubt, so daß das Funktional hier nicht regularisiert werden muß.

Zur Bestimmung des exakten Werts des Funktionals beachten wir, daß in zwei Dimensionen die retardierte GREENsche Funktion des Problems durch

$$G_+(\mathbf{x}-\mathbf{y}, t-\tau) = \frac{1}{2\pi} \frac{\theta(t-\tau-|\mathbf{x}-\mathbf{y}|)}{\sqrt{(t-\tau)^2 - |\mathbf{x}-\mathbf{y}|^2}}$$

gegeben ist und sich die Anfangswerte als spezielle Inhomogenität schreiben lassen:

$$u_{tt} - \Delta u = -\delta'(t) u_0(\mathbf{x}) + \delta(t) v_0(\mathbf{x}).$$

Damit gilt für die Lösung am Ursprung:

$$u(0,t) = \int dy \int d\tau\ G_+(\mathbf{y}, t-\tau)(-\delta'(\tau) u_0(\mathbf{y}) + \delta(\tau) v_0(\mathbf{y})),$$

und mit den gegebenen Anfangswerten

$$\begin{aligned} u(0,t) &= \int dy\ G_+(\mathbf{y},t) v_0(\mathbf{y}) \\ &= \int_0^\infty dr\ r \frac{\theta(t-r)}{\sqrt{t^2-r^2}} v_0(r) \\ &= \int_0^{\min(s,t)} dr\ r \frac{1}{\sqrt{t^2-r^2}}. \end{aligned}$$

Dieses Integral läßt sich leicht auswerten und man erhält

$$u(0,t) = t - \theta(t-s)\sqrt{t^2 - s^2}. \tag{5.1}$$

Der exakte Wert des obigen Funktionals ist damit $\frac{1}{8} - \frac{1}{20}\sqrt{6} + \frac{1}{200}\ln(5 + 2\sqrt{6}) \approx 0.013988$.

Der zeitliche Verlauf der kontinuierlichen sowie der Verlauf der numerischen Lösung im Ursprung ist in Abbildung 5.1 dargestellt. Die numerischen Ergebnisse sind nach zwei und nach sechs Verfeinerungszyklen dargestellt, oben rechts für Verfeinerung mit einem Energiefehlerindikator, unten links mit dem dem Problem angepaßten dualen Fehlerschätzer und unten rechts bei globaler Verfeinerung auf konstant 1024 und auf 16384 Zellen, d. h. mit 4225 bzw. 66049 Freiheitsgraden pro Zeitschritt. Die Rechnung wurde mit biquadratischen Elementen für das primale und bikubischen Elementen für das duale Problem durchgeführt. Die Zeitschrittweite betrug konstant 0.005, d. h. es wurden 100 Zeitschritte durchgeführt.

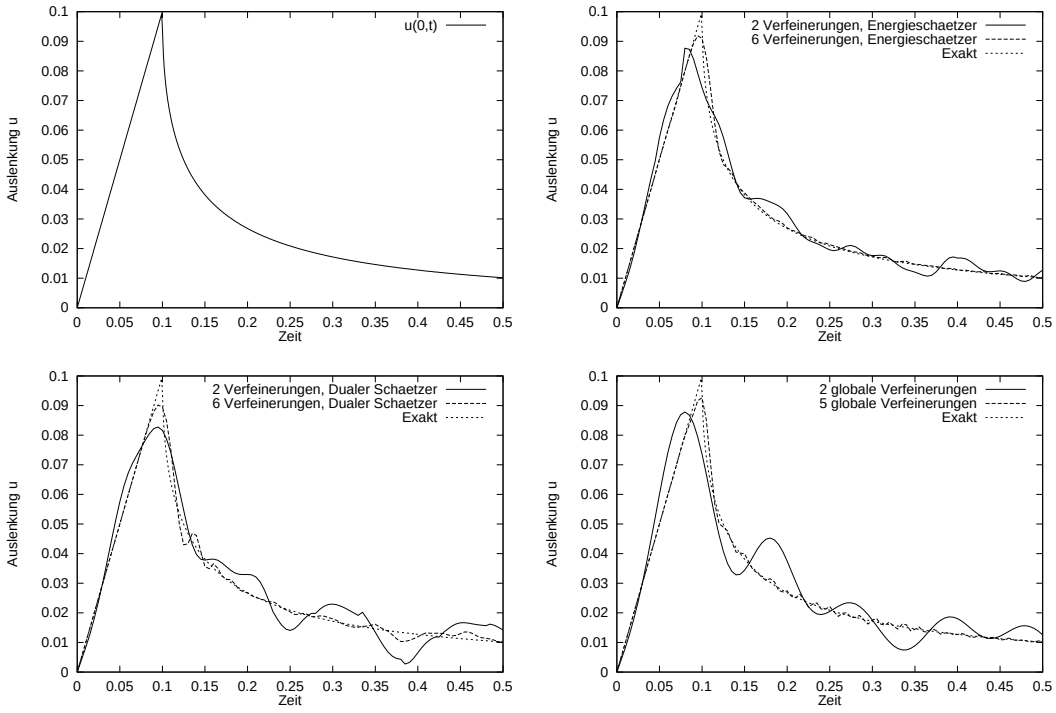

Abbildung 5.1: *Zeitlicher Verlauf der Auslenkung $u(0,t)$ am Ursprung in der exakten Lösung (links oben) sowie der numerischen Lösung nach zwei und nach fünf bzw. sechs Verfeinerungszyklen. Rechts oben Verfeinerung mit einem Energiefehlerindikator, links unten mit dem angepaßten dualen Problem, rechts unten bei globaler Verfeinerung.*

In Abbildung 5.2 sind die relativen Fehler bei der Auswertung des Funktionals gegen die Anzahl der Freiheitsgrade[1] und gegen die Rechenzeit für die verschiedenen Fehlerschätzer und globale Verfeinerung gegeneinander aufgetragen. Als Anzahl der Freiheitsgrade wurde die Summe der Zahlen in den einzelnen Zeitschritten gewählt. Der erste Wert ist bei allen drei Kurven jeweils identisch, da auf dem gleichen Grobgitter gerechnet wurde; daß der dritte Wert so niedrig liegt, ist Zufall, da der Fehler in diesem Bereich sein Vorzeichen wechselt, was aber in der logarithmischen Darstellung nicht zu sehen ist. Da die Auswertung der Fehleridentität (2.22) durch Rechnen des

[1] Wie in Kapitel 3 ist die Anzahl der Freiheitsgrade über die einzelnen Zeitschritte aufsummiert angegeben; da die Zahl zwischen einzelnen Zeitschritten stark schwanken kann, ist dies die einzige vergleichbare Größe. Allerdings wurde nur die Anzahl der Freiheitsgrade für die ursprüngliche Variable u akkumuliert, berücksichtigt man auch die Freiheitsgrade in der Geschwindigkeit v, so erhält man den doppelten Wert.

dualen Problems mit um eins erhöhtem Polynomgrad erfolgte, ist der Vorteil des dualen Schätzers gegenüber dem Energiefehlerindikator bei der Rechenzeit nicht so deutlich; das Verhältnis sollte aber klarer sein, wenn man mit dem gleichen Ansatzgrad rechnet (Abschnitt 2.4.4), da dann der Aufwand für das duale Problem nicht mehr den für das primale Problem überwiegt.

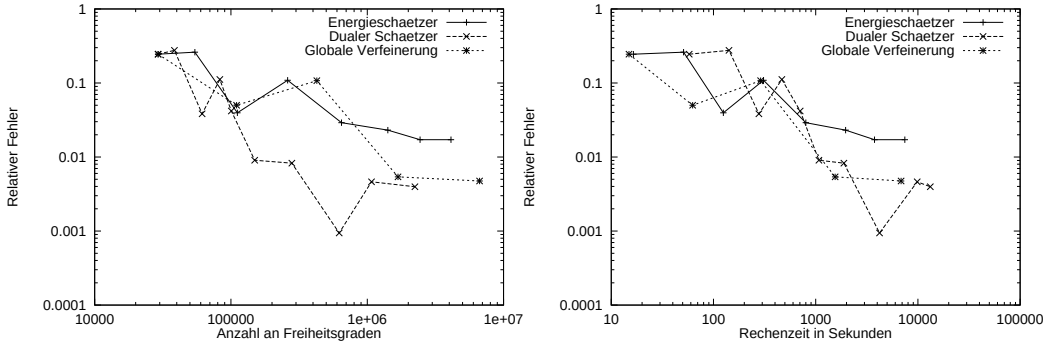

Abbildung 5.2: *Relativer Fehler in Abhängigkeit von der Zahl der Freiheitsgrade und der Rechenzeit bei verschiedenen Verfeinerungsverfahren.*

Im allgemeinen interessiert nicht das Zeitintegral über $u(0,t)$, d. h. die mittlere Abweichung vom exakten Verlauf, sondern die mittlere quadratische Abweichung

$$rms = \left(\int_I |u(0,t) - u_h(0,t)|^2 \right)^{\frac{1}{2}}.$$

Dieser Fehler ist in Abbildung 5.3 für die verschiedenen Verfeinerungskriterien dargestellt.

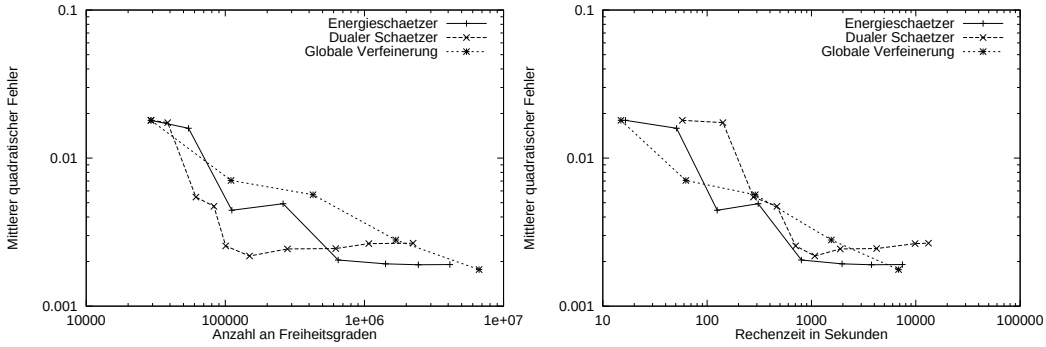

Abbildung 5.3: *Mittlerer quadratischer Fehler rms in Abhängigkeit von der Zahl der Freiheitsgrade und der Rechenzeit bei verschiedenen Verfeinerungsverfahren. Die Ergebnisse des mit dem dualen Schätzer verfeinerten Gitters sind besser als die der anderen Verfahren, obwohl des Fehlerfunktional nicht an die Fragestellung angepaßt war.*

Aus den Abbildungen geht hervor, daß der Fehler durch die Zeitdiskretisierung, die nicht verändert wurde, bei etwa 0.5-1 Prozent (Zeitintegral, Abbildung 5.2) bzw. 0.002-0.003 (rms, Abbildung 5.3) liegt. Das durch den dualen Fehlerschätzer gesteuerte Verfahren ist in der Lage, diesen Bereich mit deutlich weniger Zellen als die anderen beiden Prozesse zu erreichen. Der drittletzte, sehr niedrige Wert des dualen Schätzers in Abbildung 5.2 ist allerdings eher zufälliger Natur; dort wechselte der Fehler sein Vorzeichen.

In Abbildung 5.4 sind die vom dualen Schätzer und vom Energiefehlerindikator erzeugten Gitter

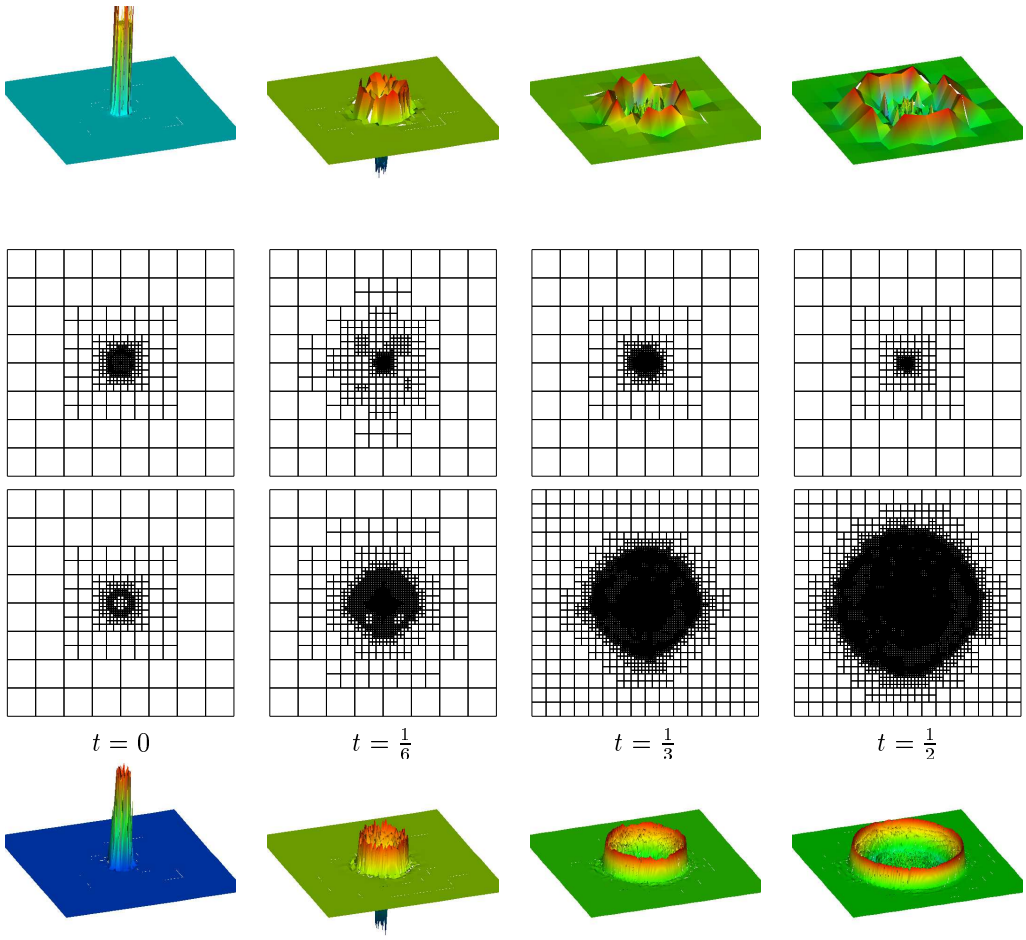

Abbildung 5.4: *Vergleich der vom dualen Schätzer (oben, nach sieben Verfeinerungsschritten) und vom Energiefehlerindikator (unten, nach sechs Verfeinerungsschritten) erzeugten Gitter und Lösungen* v_h.

einander gegenübergestellt. Es ist deutlich zu erkennen, daß der duale Schätzer nur dort verfeinert, wo es für das Endergebnis wichtig ist, während die Welle selbst nur sehr schlecht aufgelöst wird.

5.2 Winkelabhängigkeit der Ausbreitungsgeschwindigkeit

Es ist bekannt, daß die Ausbreitungsgeschwindigkeit abhängig davon ist, ob die Welle in Richtung der Gitterlinien oder diagonal dazu läuft (siehe z. B. [1]). Im allgemeinen wird dieser Effekt als „winkelabhängige Ausbreitungsgeschwindigkeit“ bezeichnet, wobei mit dem Winkel der Winkel zwischen Ausbreitungsrichtung und den Gitterlinien gemeint ist.

Der beschriebene Effekt ist besonders da störend, wo die genaue Ankunftszeit einer Welle eine Rolle spielt, beispielsweise beim Vergleich von simulierten Erdbebenwellen mit den Aufzeichnungen von Meßstationen. Die Abweichung der Dispersionsrelation vom richtigen Verlauf ist für Viereckselemente in etwa 2-5 mal kleiner als bei Dreiecken (vgl. [21]); der Effekt der winkelabhängigen Ausbreitungsgeschwindigkeiten ist für lineare Elemente besonders ausgeprägt und soll in diesem Abschnitt untersucht werden. Das Beispiel selbst ist von keinem praktischen Wert.

Zu lösen ist wieder die homogene Wellengleichung mit konstantem Koeffizienten. Als Anfangsverteilung wurde $u_0 = \theta(a-r)e^{-\left(\frac{r}{a}\right)^2}\left(1-\frac{r^2}{a^2}\right)$, $v_0 = 0$ mit $a = \frac{1}{10}$ gewählt. Die exakte Lösung ist wieder radialsymmetrisch.

Als Zielfunktional wurde

$$\begin{aligned}\mathcal{J}(u) &= \left(\int_0^T \left|\int_\Omega u(r,\phi,t)\left(\omega(r,\phi) - \omega\left(r,\phi+\frac{\pi}{4}\right)\right) dr\ d\phi\right|^2 dt\right)^{\frac{1}{2}} \\ &= \|(u,w)_\Omega\|_{L^2([0,T])}\end{aligned}$$

gewählt. Die Gewichtsfunktion sei

$$\omega(r,\phi) = \max(\cos 4\phi, 0)\ \rho\left(\frac{r-\frac{3}{4}}{a}\right), \qquad \rho(s) = (1-s^2)\theta(1-|s|),$$
$$w = \left(\omega(r,\phi) - \omega\left(r,\phi+\frac{\pi}{4}\right)\right),$$

mit $a = 0.03$. Die Gewichtsfunktion w ist in Abbildung 5.5 dargestellt. Sie ist so gewählt, daß sie die Teile der Welle, die parallel zu Gitterlinien und die die diagonal dazu laufen mit unterschiedlichen Vorzeichen gewichtet. Auf diese Weise erhält man genau die Differenz zwischen den Teilen der Welle parallel und diagonal zum Gitter; für die exakte Lösung sind die beiden Teile gleich und der Wert des Funktionals ist Null. Als Referenz für die Größenordnung der zu erwartenden Ergebnisse kann der numerisch gewonnene Wert $\|(u,|w|)_\Omega\|_{L^2([0,T])} \approx 3.22 \cdot 10^{-3}$ angesehen werden, der die beiden Teile mit dem gleichen Vorzeichen gewichtet.

Für den dualen Fehlerschätzer muß wieder ein linearisiertes Funktional $\tilde{J}(u_h;\cdot)$ gewählt werden, so daß $\tilde{J}(u_h;e) = \mathcal{J}(u)^2 - \mathcal{J}(u_h)^2$; da der erste Summand verschwindet, ist die Gleichheit trotz der Nichtlinearität hier erfüllbar und man wählt

$$J(u_h;\varphi) = \tilde{J}(u_h;\varphi) = \int_0^T (w,u_h)_\Omega\,(w,\varphi)_\Omega\,dt.$$

Der Faktor $(w,u_h)_\Omega$ in J gewichtet die Zeiten höher, bei denen eine Abweichung von der Symmetrie vorliegt. Der Fehler $\mathcal{J}(u) - \mathcal{J}(u_h) = -\mathcal{J}(u_h) = -\sqrt{\mathcal{J}(u_h)^2} = -\sqrt{-\mathcal{J}(u)^2 + \mathcal{J}(u_h)^2}$ wird durch $-\sqrt{-J(u_h;e)}$ geschätzt. Das Funktional J stimmt mit dem exakten Funktional (2.27) überein, während das nach (2.28) linearisierte Funktional genau den doppelten Wert hätte.

In Abbildung 5.6 sind die Fehler $\mathcal{J}(u_h)$ für verschiedene Verfeinerungsstrategien dargestellt. Er ist sowohl für globale Verfeinerung als auch für Verfeinerung mit dem Energiefehlerschätzer proportional zu N^{-1}, wobei N die aufsummierte Anzahl an Freiheitsgraden der ursprünglichen Variable u des primalen Problems ist. Dieser Verlauf korrespondiert ungefähr mit einer Konvergenz proportional zu h^{-2}.

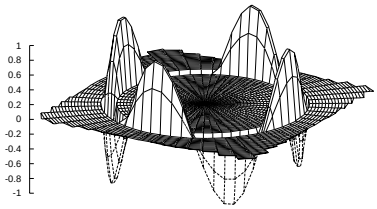

Abbildung 5.5: *Die Gewichtsfunktion w. Der besseren Darstellbarkeit halber wurde hier $a = 0.06$ gesetzt.*

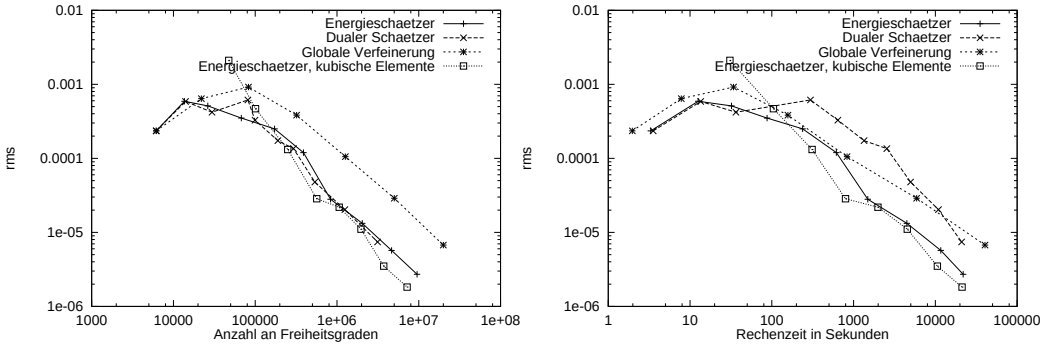

Abbildung 5.6: *Fehler in Abhängigkeit von der Zahl der Freiheitsgrade und der Rechenzeit bei verschiedenen Verfeinerungsverfahren. Man vergleiche die Werte mit dem im Text angegebenen Referenzwert* $3.22 \cdot 10^{-3}$.

Zum Vergleich ist auch der Fehler bei Verwendung von kubischen Elementen dargestellt. Sie zeigen das nach [1] erwartete Verhalten, da der gleiche Fehler bei weniger als der Hälfte der Freiheitsgrade und etwa der Hälfte der Rechenzeit erreicht wird.

In Abbildung 5.7 sind die Werte für Fehlerschätzer und tatsächlichen Fehler angegeben, wobei die Verfeinerung mit dem dualen Schätzer durchgeführt wurde. In beiden Reihen liegen die letzten drei Punkte in doppeltlogarithmischer Auftragung praktisch kollinear und man findet eine Konvergenzrate von $N^{-1.05}$ für den echten Fehler $\mathcal{J}(u_h)$ und von $N^{-0.95}$ für den Schätzer $(-J(u_h; \mathbf{e}))^{\frac{1}{2}}$.

Aufgrund der Konstruktion von J ist dieser Fall etwas pathologisch: nichtlineare Zielfunktionale müssen im allgemeinen linearisiert werden, wobei die Linearisierung theoretisch um die exakte Lösung stattfinden sollte, in Ermangelung der Kenntnis davon aber um die numerische Lösung u_h gemacht wird. Im in diesem Abschnitt behandelten Fall ist die Linearisierung um die exakte Lösung u von vornherein unmöglich, da die funktionale Ableitung an dieser Stelle Null ist und die Linearisierung ist so beschaffen, daß $J(u_h; u_h) = J(e; u_h) = J(u_h; e) = J(e, e)$. Da $e \to 0$, konvergiert auch J gegen Null und damit auch die duale Lösung $\bar{w}$. Die duale Lösung ist somit so beschaffen, daß sie das Gitter so steuert, daß der Fehler des letzten Verfeinerungszyklus' verringert wird; sie kann nicht verhindern, daß der Fehler an anderer Stelle neu entsteht. Es ist erstaunlich, daß das Gitter trotzdem gegen eine vernünftige Konfiguration konvergiert. Allerdings zeigte die Arbeit an diesem Beispiel, daß die Konvergenz sehr sensitiv von der Wahl der Verfeinerungsparameter abhängt; in einigen Fällen verfeinerte das Verfahren zwar an einigen Stellen,

N	$\mathcal{J}(u_h)$	$\sqrt{-J(u_h;e)}$	Verhältnis
81.205	$6.15 \cdot 10^{-4}$	$7.64 \cdot 10^{-4}$	1.24
99.412	$3.28 \cdot 10^{-4}$	$2.13 \cdot 10^{-4}$	0.65
189.129	$1.73 \cdot 10^{-4}$	$1.94 \cdot 10^{-4}$	1.12
293.614	$1.35 \cdot 10^{-4}$	$2.87 \cdot 10^{-4}$	2.12
535.133	$4.80 \cdot 10^{-5}$	$1.27 \cdot 10^{-4}$	2.65
1.245.983	$2.03 \cdot 10^{-5}$	$5.61 \cdot 10^{-5}$	2.76
3.134.672	$7.42 \cdot 10^{-6}$	$2.37 \cdot 10^{-5}$	3.19

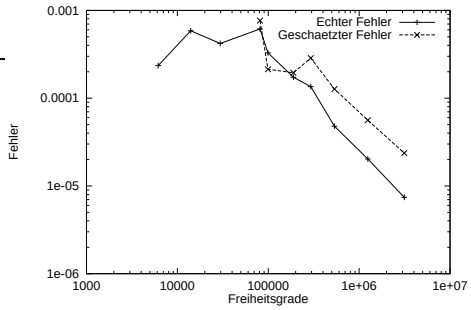

Abbildung 5.7: *Vergleich zwischen echtem und geschätztem Fehler. Man vergleiche die Werte mit dem im Text angegebenen Referenzwert* $3.22 \cdot 10^{-3}$.

vernachlässigte aber andere, auf denen der Fehler im nächsten Durchlauf beinahe genauso groß war wie im aktuellen; die Verfeinerung „lief dem Fehler immer hinterher".

Es ist noch anzumerken, daß offensichtlich bei diesem Zielfunktional die Variante nicht möglich ist, das duale Problem nur bis zu einer Gitterverfeinerungsstufe mitzurechnen, bei der sich die Gewichte $\omega^i_{K,n}$ aus Abschnitt 2.4.4, als kontinuierliche Funktion aufgefaßt,[2] nicht mehr wesentlich ändern. Dieser Vorschlag wurde gemacht, um den Aufwand zur Berechnung des Fehlerschätzers zu reduzieren; es wäre dann nur noch nötig, das duale Problem auf einem groben Gitter zu berechnen und die Gewichte abzuspeichern. Dieser Ansatz ist prinzipiell auch bei nichtlinearen Zielfunktionalen anwendbar, wenn davon ausgegangen werden kann, daß sich die duale Lösung bei genauerer Rechnung des primalen Problems nicht mehr wesentlich ändert; dies ist hier aber gerade nicht der Fall, da die duale Lösung nur die Genauigkeit der primalen Lösung widerspiegelt.

5.3 Streuung an vielen Diskontinuitäten

Als Beispiel mit variablen Koeffizienten wurde die Wellengleichung mit einem Koeffizienten $a(\mathbf{x})$ gerechnet, der viele Diskontinuitäten enthält:

$$a(\mathbf{x}) = 1 + \frac{1}{2}\left(t(x) + t\left(\frac{2x+y}{\sqrt{3}}\right)\right),$$
$$t(s) = \begin{cases} 1 & \text{wenn } \sin(3\pi s) > 0, \\ 0 & \text{sonst.} \end{cases}$$

Der Koeffizient ist in Abbildung 5.8 dargestellt. Koeffizienten dieser Art treten in der Geologie bei stark gefalteten Gesteinsschichten auf; allerdings ist die Variation des Koeffizienten im allgemeinen weit kleiner als in diesem Beispiel, dafür gibt es mehr Unstetigkeiten. Die Diskontinuitäten sind so verteilt, daß sie nicht alle auf Gitterlinien und auch nicht parallel oder diagonal dazu liegen.

Das Gebiet ist wieder $[-1,1]^2$, als Randwerte wurden oben homogene NEUMANN-Randbedingungen, überall sonst homogene DIRICHLET-Bedingungen gesetzt; letzteres ist nicht realistisch, da eigentlich absorbierende Randbedingungen gesetzt werden müßten. Die Anfangswerte wurde auf

$$u_0 = 0,$$
$$v_0 = \theta(s-r)e^{-\frac{|\mathbf{x}|^2}{s^2}}\left(1 - \frac{|\mathbf{x}|^2}{s^2}\right),$$

mit $s = 0.02$ gesetzt. Die Simulationszeit wurde relativ groß zu $T = 2$ gewählt und es wurden 300 Zeitschritte durchgeführt.

[2] Die Skalierung der Gewichte in Abschnitt 2.4.4 wurde ja gerade so durchgeführt, daß sie gegen eine kontinuierliche Funktion konvergieren.

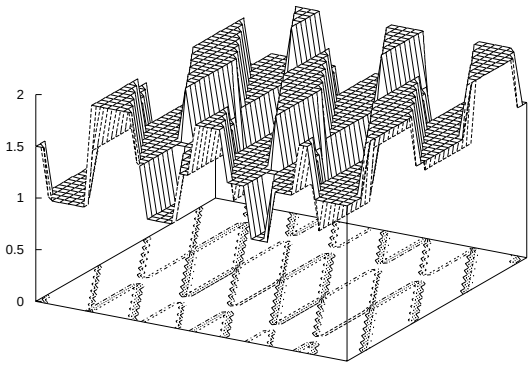

Abbildung 5.8: *Struktur des Koeffizienten mit vielen Diskontinuitäten.*

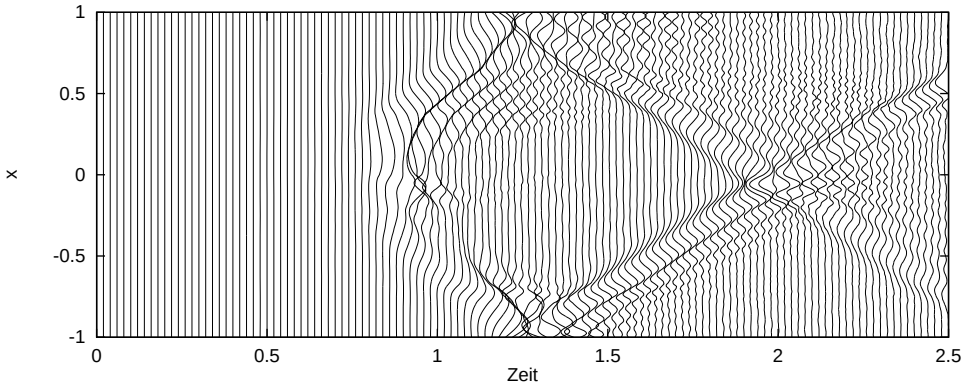

Abbildung 5.9: *Synthetisches Seismogramm der Wellenausbreitung, am oberen Rand des Gebietes abgenommen.*

Faßt man das Beispiel als die Ausbreitung einer Erdbebenwelle oder eines absichtlich eingebrachten akustischen Signals auf, so ist man daran interessiert, die Simulation $\mathbf{w}_h(\mathbf{x}, t)$ mit den Meßwerten $\mathbf{w}(\mathbf{x}, t)$ an der Oberfläche $y = 1$ zu vergleichen. In Abbildung 5.3 ist das zugehörige, sogenannte „synthetische Seismogramm“, in dem jede vertikale Linie die Auslenkung u an der oberen Grenze des Gebiets zu einem bestimmten Zeitpunkt darstellt; die Auslenkung ist nach rechts aufgetragen. Deutlich ist die hyperbolische Kurve erkennbar, die das erste Auftreffen der Welle auf den oberen Rand kennzeichnet, sowie die gekreuzten Linien, die die Verlängerungen der Hyperbel sind und von den am rechten und linken Rand reflektierten Teilen der Welle stammen, die an den oberen Rand kommen. Ebenfalls zu sehen sind die teilweise vorhandenen Reflexionen und die noch aus dem Ursprung stammenden Nachwellen der Anfangsverteilung, sowie die Störungen der Formen durch die variablen Koeffizienten.

Die in diesem Beispiel zu kontrollierende Größe wäre eigentlich

$$\mathcal{J}(\mathbf{e}) = \|\mathbf{w}(\mathbf{x}, t) - \mathbf{w}_h(\mathbf{x}, t)\|_{(\Omega \cap \{y=1\}) \times I} .$$

Da die exakte Lösung aber nicht bekannt ist, ist die Auswertung dieses Funktionals nicht möglich. Wir wählen daher ein Funktional, das wenigstens die Ortsabhängigkeit der interessierenden Größen richtig wiedergibt und linear ist, um es verhältnismäßig leicht auswerten zu können. Der Mittelwert

der Funktion

$$\mathcal{J}(\mathbf{w}) = \int_0^T \int u(x, y = 1, t)\ dx\ dt$$

ist dafür allerdings nicht gut geeignet, da unterschiedliche Ankunftszeiten und -orte einzelner Wellen keinen Unterschied im Ergebnis machen. Das kann man allerdings erreichen, indem man

$$\mathcal{J}(\mathbf{w}) = \int_0^T \int u(x, y = 1, t)\ \omega(x, t)\ dx\ dt$$

mit einem stark oszillierenden Gewichtsfaktor wählt, beispielsweise $\omega(x,t) = \sin(3\pi x)\sin\left(5\pi\frac{t}{T}\right)$. Ein zeitlich leicht verschobenes Signal führt aufgrund des dann anderen Gewichtsfaktors zu einem deutlich anderen Ergebnis $\mathcal{J}$. Die zeitliche Oszillation ist so gewählt, daß die Periode deutlich größer als die Periode der Wellen ist, andererseits aber auch wieder nicht so groß, daß die Sensitivität gegen Verschiebungen in der Größenordnung beispielsweise einer halben Periode der Wellen verloren ginge. Die zeitliche Periode der Gewichtsfunktion wurde hier 30 mal so groß wie eine typische Zeit (Radius der Anfangsverteilung durch mittlere Ausbreitungsgeschwindigkeit) gewählt, die örtliche Periode ebenfalls 30 mal so groß wie eine typische Länge (hier der Radius der Anfangsverteilung).

Die Ergebnisse der Rechnung sind in den Abbildungen 5.10 und 5.11 dargestellt. In der ersten Abbildung ist der Verlauf von $\mathcal{J}(\mathbf{w}_h)$ gegen die Anzahl an Gitterzellen widergegeben. Die ersten drei Wertepaare von dualem Fehlerschätzer und Energiefehlerindikator sind identisch, da die Verfeinerung in diesen Durchläufen jeweils nach dem Energiefehlerindikator durchgeführt wurde. Durch Extrapolation der Werte bekommt man eine Schätzung des exakten Wertes des Zielfunktionals; durch Vergleich mit diesem Wert erkennt man, daß die Rechnung mit dem dualen Problem in der Lage ist, das Funktional mit deutlich geringerer Zellzahl gut zu approximieren.

In der zweiten Grafik sind die vom Energiefehlerindikator und dem dualen Schätzer erzeugten Gitter und Lösungen gezeigt. Man erkennt deutlich, daß der duale Schätzer in der Lage ist, die Lokalisierung des Einflußgebietes zu berücksichtigen, indem nur der Bereich von Ω verfeinert wird, von dem aus überhaupt Beiträge zu $\mathcal{J}(\cdot)$ gelangen können; insbesondere werden gegen Ende der Rechnung wesentlich weniger Zellen benötigt als beim Energiefehlerindikator, der versucht, *allen* Wellen und Reflexionen zu folgen. Das vom dualen Schätzer zum Zeitpunkt $t = 0$ erzeugte Gitter ist stärker verfeinert als das des Energiefehlerindikators, da dort versucht wird, das Gitter neben der primalen auch an die duale Lösung anzupassen (vgl. Abschnitt 4.5.1).

Dualer Schätzer		Energiefehlerindikator	
N	$\mathcal{J}(\mathbf{w}_h)$	N	$J(\mathbf{w}_h)$
327.789	$-2.985 \cdot 10^{-6}$	327.789	$-2.985 \cdot 10^{-6}$
920.380	$-4.630 \cdot 10^{-6}$	920.380	$-4.630 \cdot 10^{-6}$
2.403.759	$-4.286 \cdot 10^{-6}$	2.403.759	$-4.286 \cdot 10^{-6}$
1.918.696	$-4.177 \cdot 10^{-6}$	5.640.223	$-4.385 \cdot 10^{-6}$
2.975.119	$-4.438 \cdot 10^{-6}$	10.189.837	$-4.463 \cdot 10^{-6}$
6.203.497	$-4.524 \cdot 10^{-6}$	17.912.981	$-4.521 \cdot 10^{-6}$
		41.991.779	$-4.517 \cdot 10^{-6}$

Abbildung 5.10: *Vergleich der mit dem dualen Schätzer und dem Energiefehlerindikator erhaltenen Ergebnisse $J(\mathbf{w}_h)$ in Abhängigkeit von der benötigten Zellzahl. Durch Extrapolation bekommt man als wahrscheinlichsten Wert $J(\mathbf{w}) = -4.515 \cdot 10^{-6} \ldots - 4.520 \cdot 10^{-6}$.*

5.4 Teilweise Reflexion an Gitterunstetigkeiten

Die teilweise Reflexion (*spurious reflection*) von Wellen an Gitterdiskontinuitäten ist ein in der Literatur vielfach beschriebenes und als großes Hindernis für die Verwendung adaptiver Methoden

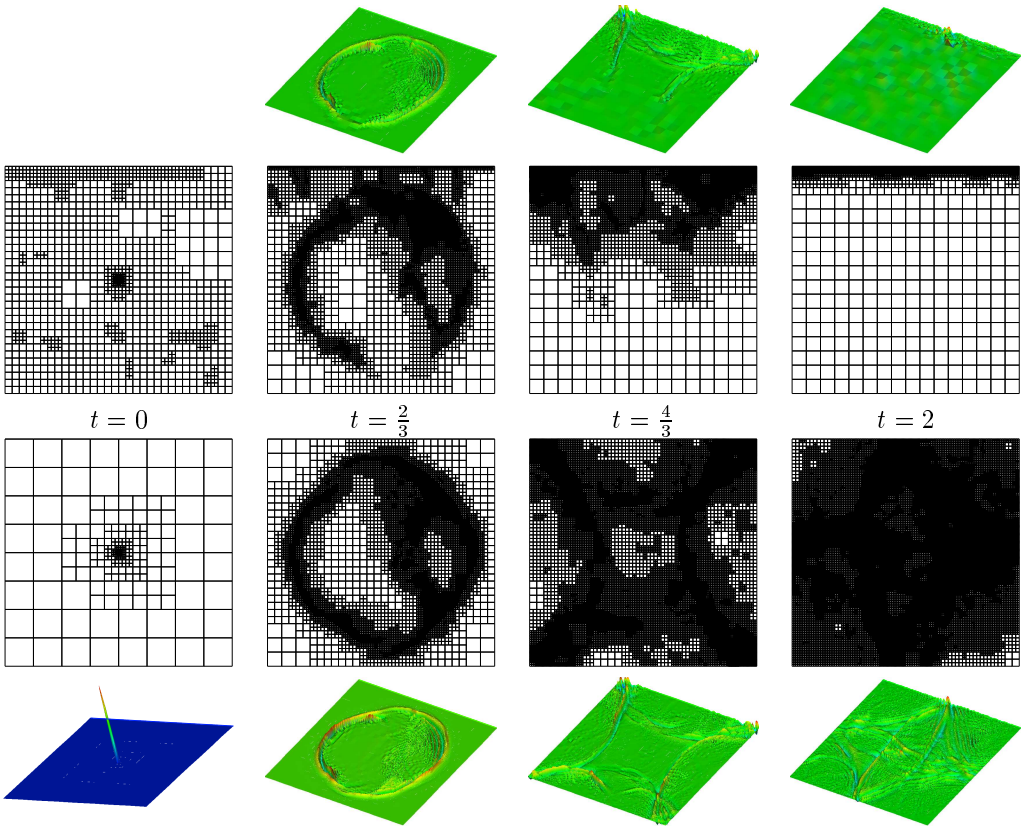

Abbildung 5.11: *Vom dualen Schätzer (oben) und vom Energiefehlerindikator (unten) erzeugten Gitter und Lösungen u zu verschiedenen Zeitpunkten. Da zum Anfangszeitpunkt $u_0 = 0$ ist, ist links unten v_0 dargestellt.*

erkanntes Problem. Um es zu untersuchen, wurden Rechnungen der Wellengleichung mit konstanten Koeffizienten $a = \rho = 1$ auf dem Gebiet $(0,3) \times (0,1)$ durchgeführt; oben und unten wurden homogene NEUMANN-Randwerte, links und rechts DIRICHLET-Randwerte vorgeschrieben, wobei die Randfunktion links zu

$$g_D(x = 0, y, t) = \begin{cases} \sin^2(5\pi t) & \text{für } t < 0.2, \\ 0 & \text{sonst} \end{cases}$$

gewählt wurde. Die Lösung ist eine von links in das Gebiet hineinlaufende Welle.

Für Rechnungen ohne Adaptivität wurde das Gitter in Abbildung 5.12 verwendet; die beiden Unstetigkeiten des Gitters liegen bei $x = 1$ und $x = 1.25$. Solche Gitter sind in der Geophysik typisch, um die Zone niedriger Ausbreitungsgeschwindigkeit nahe der Erdoberfläche aufzulösen, ohne das ganze Erdinnere mit einem feinen Gitter zu überziehen (vgl. [2]). Da die kleinste Ortsgitterweite im feinen Teil links $h = \frac{1}{64}$ betrug, wurde auch $k = \frac{1}{64}$ gewählt; das Verhältnis von halber Wellenlänge (d. h. der Länge der Welle) zu h beträgt damit etwa 13, ebenso wie das Verhältnis von halber Periodendauer zur Zeitschrittweite.

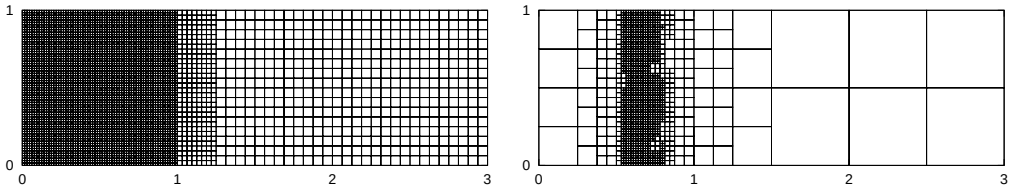

Abbildung 5.12: *Festes Gitter mit Diskontinuitäten (links) und typisches Gitter bei der Verwendung adaptiver Verfahren (rechts).*

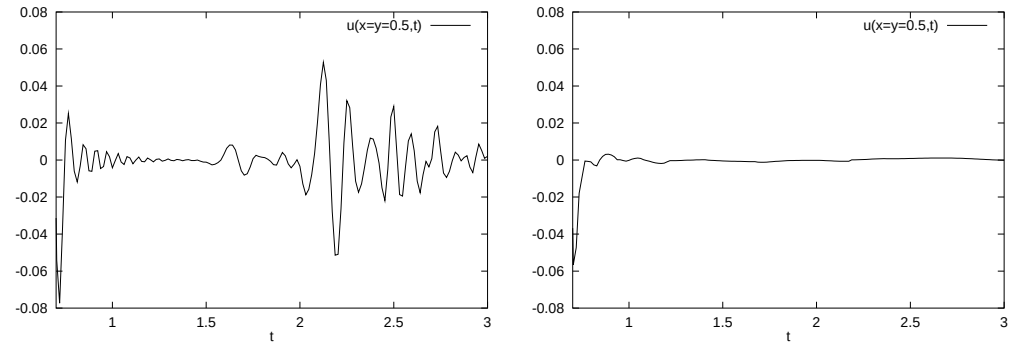

Abbildung 5.13: *Vergleich der Signale $u(x = 0.5, y = 0.5, t)$ bei festem Gitter (links) und bei Verwendung des Energiefehlerindikators (rechts). Die Amplitude des Hauptsignals ist* 1.0.

Um die Reflexion der von links gegen die Gitterunstetigkeiten laufenden Wellen zu verdeutlichen, ist in Abbildung 5.13 der Verlauf der Auslenkung $u(x = \frac{1}{2}, y = \frac{1}{2}, t)$ für $t > 0.7$ dargestellt. Die exakte Lösung ist an diesem Punkt für $t > 0.7$ identisch Null, da die einlaufende Welle eine halbe Zeiteinheit bis zu diesem Punkt benötigt und den Punkt demnach auch eine halbe Zeiteinheit nach dem Ende der Injektion vom linken Rand passiert hat. Stattdessen beobachtet man am Anfang noch die Nachläufer der Welle, die hier aber nicht interessieren, sowie zwischen $t = 1.5$ und 1.7 die Reflexion der Welle an der ersten Unstetigkeit und ab $t = 2$ die Reflexion an der zweiten Unstetigkeit; ab $t = 2.5$ ist außerdem die am linken Rand reflektierte erste Reflexion wieder am Auswertungspunkt. Die zeitliche Länge der zweiten Reflexion ist wesentlich größer als die Dauer

der ursprünglichen Welle von 0.2 Zeiteinheiten; diese Verlängerung rührt daher, daß das Hauptsignal nach Passieren der ersten und vor Erreichen der zweiten Gitterunstetigkeit bereits eine Länge von rund 0.3, nach Passieren der zweiten Diskontinuität sogar von rund 0.7 Zeiteinheiten (jeweils ohne die kleinen Nachläufer) hat. Die Amplitude der ersten Reflexion beträgt rund 0.9 Prozent, die der zweiten etwa 6 Prozent der Amplitude der ursprünglichen Wellen. Insgesamt tritt bei der Verwendung dieser festen Gitters also neben der starken Änderung der Form der primären Welle auch Reflexion mit einer nicht vernachlässigbaren Amplitude auf.

Die Ergebnisse mit Verfeinerung durch einen Energiefehlerindikator sind dagegen in den Abbildungen 5.12 und 5.13 jeweils rechts dargestellt. Der maximale Verfeinerungsgrad wurde so begrenzt, daß keine Zelle kleiner als die kleinste Zelle des festen Gitters war; insgesamt hatte diese Rechnung aufsummiert 209.296 Freiheitsgrade in der Variable u (und damit doppelt so viele insgesamt), während die Rechnung auf dem festen Gitter 958.245 benötigte. Es ist offensichtlich, daß die Reflexion an Gitterunstetigkeiten keine Rolle spielt, was auch nicht überrascht, da das Gitter ja gerade so gewählt war, daß es der Welle folgt und diese daher nie eine größere Unstetigkeit treffen sollte.

Eine sich in diesem Zusammenhang jedoch ergebende Frage ist, was bei Verwendung des dualen Schätzers zur Gitterverfeinerung passiert. Bei diesem wird das Gitter nicht ausschließlich nach der Lösung, sondern auch nach der Zielgröße gesteuert, und die in den letzten Abschnitten und Kapiteln gegebenen Beispiele zeigen, daß es dabei normal ist, daß einige Wellen nicht aufgelöst werden und also irgendwann in ein Gebiet mit gröberem Gitter laufen. Dabei sollten auch Reflexionen entstehen, die in das Einflußgebiet des Zielfunktionals zurücklaufen und das Ergebnis der Rechnung verfälschen. Man könnte dies vergleichen mit der *error pollution* bei elliptischen Differentialgleichungen, bei der auch aufgrund nicht optimal angepaßter Gitter Fehler weitab vom Ursprungsort des Problems, beispielsweise einer Singularität an einer einspringenden Ecke, auftreten. Der Fehler kann hier sowohl örtlich als auch zeitlich weitab von der Stelle auftreten, wo das Gitter nicht optimal war.

Es ist nun fraglich, ob der duale Schätzer in der Lage ist, die Beeinflussung der Zielgröße durch teilweise Reflexion zu „sehen“ und das Gitter entsprechend zu verändern. Im allgemeinen ist das zweifelhaft und es scheint nicht ausgeschlossen, daß der duale Schätzer diese Art von Fehlern nur schlecht unterdrücken kann; hier sind weitere Untersuchungen notwendig, die im Rahmen dieser Arbeit aber nicht mehr durchgeführt werden konnten.

Anhang A

A priori Fehlerschranken für einige Zeitschrittverfahren

In diesem Abschnitt sollen einige a priori Abschätzungen der Zeitdiskretisierung des verwendeten Systems

$$\begin{aligned} u_t - v &= 0, && \text{(A.1)}\\ v_t - \Delta u &= 0, && \text{(A.2)}\\ u(\mathbf{x}, 0) &= u_0,\\ v(\mathbf{x}, 0) &= v_0 \end{aligned}$$

gegeben werden. Der Einfachheit halber werden die Koeffizienten als konstant angenommen. Darüberhinaus wird nur der Effekt der Zeitdiskretisierung betrachtet, was es erlaubt, scharfe Abschätzungen des L^∞-Fehlers zu erhalten; der durch die Ortsdiskretisierung verursachte Fehler wird nicht einbezogen.

Die Rechnungen sind nur für eine Raum- und eine Zeitdimension, sowie ein örtlich unbeschränktes Gebiet durchgeführt, um die Formeln übersichtlich zu halten. Die Erweiterung auf mehrere Raumdimensionen ist leicht durchführbar, indem man die entsprechenden GREENschen Funktionen verwendet; diese sind auch für zwei und mehr Raumdimensionen bekannt (vgl. beispielsweise [9]). Die Einschränkung auf ein beschränktes Gebiet ist nur für Punkte des Gebiets relevant, deren Abstand zum Rand kleiner als die Zeitschrittweite mal der Ausbreitungsgeschwindigkeit ist; man kann sich leicht überlegen, daß die Mehrzahl der Ergebnisse gültig bleibt, falls reflektierende (d. h. DIRICHLET- oder NEUMANN-) Randbedingungen verwendet werden.

A.1 Vorbemerkungen

Wir betrachten die exakte Lösung der Wellengleichung zum Zeitpunkt $t = k$, d. h. nach dem ersten Zeitschritt, die wir mit der semidiskretisierten Lösung vergleichen werden. Erstere ist gemäß der D'ALAMBERTschen Formel durch

$$\begin{aligned} u(x, t = k) &= \frac{1}{2}\left(u_0(x-k) + u_0(x+k)\right) + \frac{1}{2}\int_{-k}^{k} dy\ v_0(x-y)\\ &= \int_{-\infty}^{\infty} dy\ \frac{1}{2}\left(\delta(y-k) + \delta(y+k)\right) u_0(x-y)\\ &\quad + \int_{-\infty}^{\infty} dy\ \frac{1}{2}\theta(k-|y|)\ v_0(x-y) \qquad \text{(A.3)} \end{aligned}$$

gegeben. Sie läßt sich also angeben als Faltung der Anfangswerte mit Integralkernen. Aus dem Vergleich der exakten Integralkerne mit den bei der Diskretisierung entstehenden werden wir

versuchen, die Güte eines Verfahrens abzuschätzen. Der Vergleich wird jeweils für den ersten Zeitschritt durchgeführt werden, und da wir nur die Semidiskretisierung betrachten werden, sind u_0 und u^0 synonym.

Für die folgende Analyse benötigen wir noch zwei Lemmata:

Lemma 1 *Es gilt*

$$\| (\alpha * \omega)(x)\|^2_{L^2(x)} = 2\pi \ \left(|\tilde{\alpha}(p)|^2, |\tilde{\omega}(p)|^2\right)_{L^2(p)},$$

wobei $*$ das Faltungsprodukt $(\alpha * \omega)(x) = \int \alpha(y)\omega(x-y)\, dy$ bezeichne und $\tilde{\alpha}$ und $\tilde{\omega}$ die FOURIER-*Transformierten von α und ω sind.*

Beweis: Zuerst zeigen wir, daß das Faltungsprodukt im FOURIER-Raum dem normalen Produkt entspricht:

$$\begin{aligned}
\mathcal{F}\left[(\alpha * \omega)(x)\right](p) &= \int dx \ \frac{1}{\sqrt{2\pi}} e^{ipx} \int dy \ \alpha(y)\ \omega(x-y) \\
&= \int dx \ \frac{1}{\sqrt{2\pi}} e^{ipx} \int dy \ \times \\
&\qquad \times \left(\int dq \ \frac{1}{\sqrt{2\pi}} e^{-iqy} \tilde{\alpha}(q)\right) \left(\int dr \ \frac{1}{\sqrt{2\pi}} e^{-ir(x-y)} \tilde{\omega}(r)\right) \\
&= \frac{1}{\sqrt{2\pi}^3} \int dq \int dr \int dx \ e^{i(p-r)x} \int dy \ e^{i(r-q)y} \tilde{\alpha}(q)\tilde{\omega}(r) \\
&= \frac{1}{\sqrt{2\pi}^3} \int dq \int dr \ 2\pi\delta(p-r) \ 2\pi\delta(r-q)\tilde{\alpha}(q)\tilde{\omega}(r) \\
&= \sqrt{2\pi} \ \tilde{\alpha}(p) \ \tilde{\omega}(p).
\end{aligned}$$

$\mathcal{F}[\cdot]$ bezeichne dabei die FOURIER-Transformation. Nun gilt

$$\begin{aligned}
\|\alpha * \omega\|^2_{L^2} &= \int dx \ |\alpha * \omega|^2 \\
&= \int dx \ \left|\mathcal{F}^{-1}\left[\mathcal{F}\left[\alpha * \omega\right]\right]\right|^2 \\
&= \int dx \ \left|\int dp \ \frac{1}{\sqrt{2\pi}} e^{-ipx} \left(\sqrt{2\pi}\tilde{\alpha}(p)\tilde{\omega}(p)\right)\right|^2 \\
&= \int dx \left(\int dp \ e^{-ipx} \tilde{\alpha}(p)\tilde{\omega}(p)\right) \left(\int dq \ e^{-iqx} \tilde{\alpha}(q)\tilde{\omega}(q)\right)^* \\
&= \int dp \int dq \int dx \ e^{-i(p-q)x} \tilde{\alpha}(p)\tilde{\omega}(p)\tilde{\alpha}^*(q)\tilde{\omega}^*(q) \\
&= \int dp \int dq \ 2\pi\delta(p-q) \ \tilde{\alpha}(p)\tilde{\omega}(p)\tilde{\alpha}^*(q)\tilde{\omega}^*(q) \\
&= 2\pi \int dp \ |\tilde{\alpha}(p)|^2 \ |\tilde{\omega}(p)|^2
\end{aligned}$$

Das vollendet den Beweis. □

Lemma 2 *Mit dem obigen Lemma und der* HÖLDER*schen Ungleichung folgt*

$$\|\alpha * \omega\|^2_{L^2} \leq 2\pi \ \|\tilde{\alpha}\|^2_{L^{2m}} \|\tilde{\omega}\|^2_{L^{2n}}$$

mit $\frac{1}{m} + \frac{1}{n} = 1$, sofern die entsprechenden Normen existieren. Falls diese Normen existieren, gelten wegen

$$\|\tilde{\alpha}\|^2_{L^2} = \|\alpha\|^2_{L^2}$$

mit $m = 1, n = \infty$ insbesondere die folgenden Ungleichungen:

$$\|\alpha * \omega\|_{L^2}^2 \leq 2\pi \ \|\alpha\|_{L^2}^2 \|\tilde{\omega}\|_{L^\infty}^2 ,$$
$$\|\alpha * \omega\|_{L^2}^2 \leq 2\pi \ \|\tilde{\alpha}\|_{L^\infty}^2 \|\omega\|_{L^2}^2 .$$

Beweis: Die Aussagen folgen durch Einsetzen, weshalb nur die mittlere Identität bewiesen sei:

$$\begin{aligned}
\|\tilde{\alpha}\|_{L^2}^2 &= \int dp \ \tilde{\alpha}(p)\tilde{\alpha}^*(p) \\
&= \int dp \ \left(\int dx \ \frac{1}{\sqrt{2\pi}} e^{ipx} \alpha(x) \right) \left(\int dy \ \frac{1}{\sqrt{2\pi}} e^{-ipy} \alpha^*(y) \right) \\
&= \frac{1}{2\pi} \int dx \int dy \int dp \ e^{i(x-y)p} \alpha(x)\alpha^*(y) \\
&= \frac{1}{2\pi} \int dx \int dy \ 2\pi\delta(x-y) \ \alpha(x)\alpha^*(y) \\
&= \int dx \ |\alpha(x)|^2 \\
&= \|\alpha\|_{L^2}^2 .
\end{aligned}$$

□

A.2 Das implizite Euler-Verfahren

A.2.1 Verfahren

Verwendet man das implizite EULER-Verfahren, so erhält man aus (A.1) und (A.2) für die Lösung u^1 zum ersten Zeitschritt die Gleichungen

$$\begin{aligned}
u^1 &= u^0 + kv^1, \\
v^1 &= v^0 + k\Delta u^1.
\end{aligned}$$

Dabei ist k die Zeitschrittweite. Für die folgenden Zeitschritte gelten natürlich analoge Formeln.

Die beiden Gleichungen lassen sich in

$$\left(1 - k^2\Delta\right) u^1 = u^0 + kv^0, \tag{A.4}$$
$$v^1 = \frac{1}{k}\left(u^1 - u^0\right) \tag{A.5}$$

umformen. Es ist also nur noch eine echte Gleichung zu lösen, die zweite Gleichung ist eine Art Postprocessing-Schritt.

A.2.2 Analyse

Die GREENsche Funktion zum Operator $(1 - k^2\Delta)$ ist in einer Dimension durch

$$g(y) = \frac{1}{2k} e^{-\frac{|y|}{k}},$$

gegeben, so daß für die Näherung u^1 gilt:

$$\begin{aligned}
u^1 &= \int_{-\infty}^{\infty} dy \ \frac{1}{2k} e^{-\frac{|y|}{k}} \left(u^0(x-y) + kv^0(x-y) \right) \\
&= \int_{-\infty}^{\infty} dy \ \frac{1}{2k} e^{-\frac{|y|}{k}} \ u^0(x-y) + \int_{-\infty}^{\infty} dy \ \frac{1}{2} e^{-\frac{|y|}{k}} \ v^0(x-y).
\end{aligned} \tag{A.6}$$

Während der zweite Term tatsächlich im wesentlichen eine Mittelung über den Bereich $-k \leq y \leq k$ ist und somit den kastenförmigen Integralkern im zweiten Term von (A.3) einigermaßen vernünftig annähert, zeigt der erste Term keinerlei Ähnlichkeit mit der exakten Lösung. Schaubilder der approximativen und der exakten Integralkerne sind in Abbildung A.1 zu finden. Beide Terme verletzen die Endlichkeit der Ausbreitungsgeschwindigkeit, was angesichts des elliptischen Charakters von (A.4) auch nicht weiter überrascht.

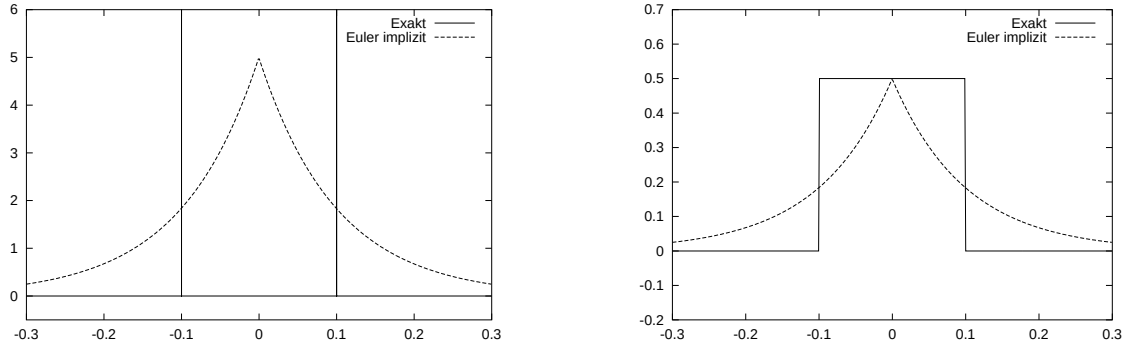

Abbildung A.1: *Erster und zweiter Integralkern für das implizite* EULER*-Verfahren bei* $k = \frac{1}{10}$.

L^∞-Fehler

Es gilt wegen (A.3) und (A.6)

$$\left(u(k) - u^1\right)(x) = (a * u^0)(x) + k(b * v^0)(x) \tag{A.7}$$

mit

$$\begin{aligned} a(y) &= \frac{1}{2}\delta(y-k) + \frac{1}{2}\delta(y+k) - \frac{1}{2k}e^{-\frac{|y|}{k}}, \\ b(y) &= \frac{1}{2k}\theta(k-|y|) - \frac{1}{2k}e^{-\frac{|y|}{k}}. \end{aligned}$$

Aus b wurde eine k-Potenz aus Gründen, die weiter unten klar werden, herausgezogen.

Damit ist

$$\begin{aligned} \left|\left(u(k) - u^1\right)(x)\right| &= \left|\left(a(y), u^0(x-y)\right)_{L^2(y)} + k\left(b(y), v^0(x-y)\right)_{L^2(y)}\right| \\ &\leq \left|\left(a(y), u^0(x-y)\right)_{L^2(y)}\right| + k\left|\left(b(y), v^0(x-y)\right)_{L^2(y)}\right|. \end{aligned}$$

Das wollen wir gerne durch die HÖLDERsche Ungleichung auswerten. Für den ersten Term mit a kommt dabei nur die L^1-Norm in Frage, da die L^∞-Norm nicht existiert und wir sonst nicht wissen, wie wir zum Beispiel für die L^2-Norm das Quadrat einer Delta-Funktion auswerten sollen; für den zweiten Summanden können wir die Normen mit $\frac{1}{n} + \frac{1}{m} = 1$ frei wählen:

$$\leq \|a(y)\|_{L^1(y)} \left\|u^0(x-y)\right\|_{L^\infty(y)} + k\|b(y)\|_{L^m(y)} \left\|v^0(x-y)\right\|_{L^n(y)}.$$

Die rechte Seite hängt jetzt nicht mehr von x ab. Die Normen von a und b können wir ausrechnen und erhalten beispielsweise für $m = n = 2$

$$\left\|u(k) - u^1\right\|_{L^\infty} \leq 2\left\|u^0\right\|_{L^\infty} + \sqrt{\frac{1}{e} - \frac{1}{4}}\,\sqrt{k}\left\|v^0\right\|_{L^2}. \tag{A.8}$$

Daß wir nur die Ordnung $\mathcal{O}(1)$ erhalten, ist ebenso wie der Vorfaktor kein technisches Problem der Abschätzung von a, sondern eine scharfe Abschätzung, wie sich an einem einfachen Beispiel demonstrieren läßt. Seien dafür als Anfangswerte

$$u^0(x) = \begin{cases} 1 & \text{für } x = \pm k, \\ 0 & \text{sonst,} \end{cases}$$
$$v^0(x) = 0$$

oder eine geeignete, gegen diese Funktionen konvergierende Funktionenfolge gegeben. Nach A.3 ist dann $u(0,k) = 1$, jedoch ist $u^1(0) = 0$ nach (A.6). Damit ist schon $\left\|u(k) - u^1\right\|_{L^\infty} = 1 = \left\|u^0\right\|_{L^\infty}$. Der Vorfaktor läßt sich mit einem etwas technischeren Beispiel demonstrieren, wenn man v^0 so setzt, daß die beiden Punktwerte von u^0 nicht in jeweils beide Richtungen laufen, sondern beide ausschließlich zum Ursprung hin. Dort ist dann $u(0,k) = 2$, die L^∞-Norm von u^0 hat sich aber nicht geändert.

L^2-Fehler

Mit (A.7) ist

$$\begin{aligned} \left\|u(k) - u^1\right\|_{L^2}^2 &= \int dx \ \left| \left(a(y), u^0(x-y)\right)_{L^2(y)} + k^2 \left(b(y), v^0(x-y)\right)_{L^2(y)} \right|^2 \\ &\leq 2 \int dx \ \left| \left(a(y), u^0(x-y)\right) \right|^2 + 2k^2 \int dx \ \left| \left(b(y), v^0(x-y)\right) \right|^2 . \end{aligned}$$

Diesmal ist die Aufteilung mit der HÖLDERschen Ungleichung nicht möglich, da dann aus dem ersten Term das Integral $2\int_{-\infty}^{\infty} dx\ 2^2 \left\|u^0\right\|_{L^\infty}^2$ entstehen würde, also ein Integral über eine Konstante, was natürlich divergiert.

Man muß hier versuchen, die Lokalisierung von a besser auszunutzen. Verwendet man jedoch Lemma 1, so erhält man für den ersten Term

$$\left\| \left(a * u^0\right)(x) \right\|_{L^2}^2 = 2\pi \int dp \ |\tilde{a}|^2 \left|\tilde{u}^0\right|^2 ,$$

wobei wir $\tilde{a}$ aus der Berechnung von g und von der Fouriertransformierten der Deltafunktionen kennen:

$$= 2\pi \int dp \ |\tilde{u}^0(p)|^2 \left[\frac{1}{\sqrt{2\pi}} \left(\cos kp - \frac{1}{1+k^2p^2} \right) \right]^2 .$$

Dieser Term ist ohne Abschätzung zustande gekommen und läßt sich wie folgt deuten: für kleine kp ist der Kosinus eine Approximation erster Ordnung für den Bruchterm, das heißt der zweite Term ist bis etwa $kp = 0.2$ klein. Der ganze Term ist also klein, falls $|\tilde{u}^0(p)|^2$ außerhalb von $|p| < \frac{0.2}{k}$ verschwindet. Die Bedeutung dieser Bedingung ist in Abschnitt A.5 genauer erläutert.

Nimmt man für u^0 als Regularität an, daß $\tilde{u}^0(p) \equiv 0$ für $|p| \geq p_0$, p_0 klein genug, dann gilt für diesen Term aufgrund der HÖLDERschen Ungleichung mit $\frac{1}{n} + \frac{1}{m} = 1$:

$$\leq \| \, |\tilde{u}^0|^2 \|_{L^n} \left\{ \int_{-p_0}^{p_0} dp \ \left(\cos kp - \frac{1}{1+k^2p^2} \right)^{2m} \right\}^{\frac{1}{m}} .$$

Für kleine p_0 kann man das Integral aus der Reihenentwicklung heraus berechnen und erhält

$$\left\| \left(a * u^0\right)(x) \right\|_{L^2}^2 \leq 2^{-2+\frac{1}{m}} \ (4m+1)^{-\frac{1}{m}} \ p_0^{4+\frac{1}{m}} \ k^4 \|\tilde{u}^0\|_{L^{2n}}^2 .$$

Allerdings konvergiert die Reihenentwicklung des Integranden sehr schlecht, so daß die Abschätzung mit dem ersten Reihenglied zwar für alle Werte kp richtig, aber nur bis etwa $kp_0 = 0.5$

gut ist. Wie wir gleich sehen werden, sind wir hauptsächlich an der L^∞-Norm von $\tilde{a}$ interessiert. Wir definieren daher

$$a_\infty^2(s) = \max_{-s\le x\le s}\left(\cos x - \frac{1}{1+x^2}\right)^2. \tag{A.9}$$

Diese Funktion ist leicht numerisch auswertbar.

Für den zweiten Term erhält man ganz analog:

$$\left\|\left(b * v^0\right)(x)\right\|_{L^2}^2 = \int dp\, |\tilde{v}^0(p)|^2 \left(\frac{\sin kp}{kp} - \frac{1}{1+k^2p^2}\right)^2.$$

Hätten wir nicht ein k aus b herausgezogen, so würde der Integrand außer von kp auch noch von k abhängen. Integriert man wieder für kleine kp_0 nur das erste Reihenglied, so ist $\left\|\left(b * v^0\right)(x)\right\|_{L^2}^2 \le c_b^2 p_0^{4+\frac{1}{\mu}}\, k^4 \|\tilde{v}^0\|_{L^{2\nu}}^2$. Außerdem definieren wir noch

$$b_\infty(s)^2 = \max_{-s\le x\le s}\left(\frac{\sin x}{x} - \frac{1}{1+x^2}\right)^2. \tag{A.10}$$

Damit haben wir für kleine kp_0 insgesamt

$$\left\|u(k) - u^1\right\|_{L^2}^2 \le c_a^2 p_0^{4+\frac{1}{m}}\, k^4 \|\tilde{u}^0\|_{L^{2n}}^2 + c_b^2 p_0^{4+\frac{1}{\mu}}\, k^6 \|\tilde{v}^0\|_{L^{2\nu}}^2, \tag{A.11}$$

mit dem gewählten p_0 und den wie beschrieben berechenbaren Konstanten $c_{a,b}^2$. Insbesondere erhält man wegen $\|\tilde{u}\|_{L^2} = \|u\|_{L^2}$ mit $n = \nu = 1, m = \mu = \infty$:

$$\left\|u(k) - u^1\right\|_{L^2}^2 \le \frac{1}{4} p_0^4\, k^4 \|u^0\|_{L^2}^2 + \frac{25}{36} p_0^4\, k^6 \|v^0\|_{L^2}^2.$$

Die bessere Abschätzung mit den numerisch auswertbaren Funktionen a_∞ und b_∞ ergibt

$$\left\|u(k) - u^1\right\|_{L^2}^2 \le a_\infty^2(kp_0)\|u^0\|_{L^2}^2 + b_\infty^2(kp_0)k^2\|v^0\|_{L^2}^2.$$

Um eine vorgegebene Genauigkeit zu erreichen, müssen wir k so klein wählen, daß die Vorfaktoren bei gegebenem p_0 klein genug werden. a_∞^2 und b_∞^2 sind in den Abbildungen A.3 für verschiedene Verfahren gegenübergestellt.

A.3 Das θ-Verfahren

A.3.1 Verfahren

Diskretisiert man (A.1) und (A.2) mit dem Fractional-θ-Verfahren, so erhält man die folgenden beiden Gleichungen:

$$\begin{aligned} u^1 &= u^0 + k\theta_1 v^1 + k(1-\theta_1)v^0, \\ v^1 &= v^0 + k\theta_2 \Delta u^1 + k(1-\theta_2)\Delta u^0, \end{aligned}$$

oder nach einer Umformung:

$$\begin{aligned} (\mathbf{1} - k^2\theta_1\theta_2\Delta)u^1 &= u^0 + kv^0 + k^2\theta_1(1-\theta_2)\Delta u^0, \\ v^1 &= \frac{1}{k\theta_1}(u^1 - u^0) - \frac{1-\theta_1}{\theta_1}v^0. \end{aligned} \tag{A.12}$$

Der Allgemeinheit halber haben wir zwei verschiedene Konstanten θ_1, θ_2 gewählt. Ebenso wie beim impliziten EULER-Verfahren ist auch hier nur eine echte Gleichung zu lösen.

A.3.2 Analyse

Für die erste Gleichung ist die GREENsche Funktion durch

$$g(x) = \frac{1}{2k\sqrt{\theta_1\theta_2}} e^{-\frac{|x|}{k\sqrt{\theta_1\theta_2}}}$$

gegeben, und wir erhalten für die semidiskretisierte Lösung nach einem Zeitschritt

$$u^1 = \left(g * \left(u^0 + kv^0 + k^2\theta_1(1-\theta_2)\Delta u^0\right)\right)(x). \tag{A.13}$$

Wir können den Ableitungsoperator durch partielle Integration auf g wirken lassen und erhalten mit der Definitionsgleichung von g:

$$\begin{aligned} u^1 = &\int_{-\infty}^{\infty} dy \left(\frac{1}{2k\theta_2\sqrt{\theta_1\theta_2}} e^{-\frac{|y|}{k\sqrt{\theta_1\theta_2}}} - \frac{1-\theta_2}{\theta_2}\delta(y)\right) u^0(x-y) \\ &+ \int_{-\infty}^{\infty} dy \, \frac{1}{2\sqrt{\theta_1\theta_2}} e^{-\frac{|y|}{k\sqrt{\theta_1\theta_2}}} v^0(x-y). \end{aligned} \tag{A.14}$$

Im ersten Integralkern wird in der Mitte des unerwünschten exponentiellen Peaks (vgl. Abbildung A.1) eine Deltafunktion abgezogen. Im Finite-Elemente-Kontext können wir diese Deltafunktion wohl als Hutfunktion der Breite $2h$ und der Höhe $\frac{1}{h}$ ansehen. Im günstigsten Fall ist $h = k$, dann ergibt sich die in Abbildung A.2 gezeigte Funktion. Für andere Werte von h ist die Approximation schlechter, was oft der Fall sein wird, da wir wegen des impliziten Charakters des Verfahrens die Zeitschrittweite größer als h machen können.

Das θ, bei dem das Verhältnis der abgezogenen Funktion zur Exponentialfunktion maximal wird, ist wie erwartet $\theta = \frac{1}{2}$, entsprechend dem CRANK-NICOLSON-Verfahren.

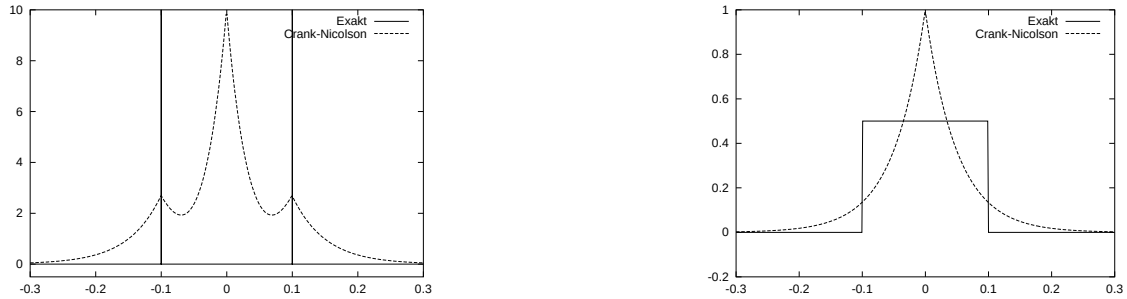

Abbildung A.2: *Erster und zweiter Integralkern für das* CRANK-NICOLSON*-Verfahren bei* $k = \frac{1}{10}$ *und links* $h = k$.

L^∞-Fehler

Den Fehler kann man mit (A.3) und (A.13) wieder als

$$\left(u(k) - u^1\right)(x) = (a * u^0)(x) + k(b * v^0)(x) \tag{A.15}$$

darstellen, mit

$$\begin{aligned} a(y) &= \frac{1}{2}\delta(y-k) + \frac{1}{2}\delta(y+k) - \frac{1}{2k\theta_2\sqrt{\theta_1\theta_2}} e^{-\frac{|y|}{k\sqrt{\theta_1\theta_2}}} + \frac{1-\theta_2}{\theta_2}\delta(y), \\ b(y) &= \frac{1}{2k}\theta(k-|y|) - \frac{1}{2k\sqrt{\theta_1\theta_2}} e^{-\frac{|y|}{k\sqrt{\theta_1\theta_2}}}. \end{aligned}$$

Wir erhalten ganz analog zu oben

$$\|u(k)-u^1\|_{L^\infty} \le \|a\|_1 \|u^0\|_\infty + k\|b\|_{L^m}\|v^0\|_{L^n},$$

mit $\frac{1}{m}+\frac{1}{n}=1$. Wählt man wieder $m=n=2$ und rechnet die Integrale aus, so erhält man

$$\|u(k)-u^1\|_{L^\infty} \le \frac{2}{\theta_2}\|u^0\|_{L^\infty} + \sqrt{e^{-\frac{1}{\sqrt{\theta_1\theta_2}}} - \frac{1}{2} + \frac{1}{4\sqrt{\theta_1\theta_2}}}\ \sqrt{k}\|v^0\|_{L^2}.$$

Auch hier ist der Vorfaktor des ersten Terms wieder scharf, wofür sich Beispiele analog dem oben gezeigten konstruieren lassen. Im allgemeinen ist der L^∞-Fehler bei diesem Verfahren also schlechter als beim impliziten EULER-Verfahren. Für $\theta_1=\theta_2=1$ erhält man natürlich wieder die Formel für das implizite EULER-Verfahren.

L^2-Fehler

Man geht wieder ebenso wie oben vor und erhält

$$\begin{aligned}\left\|u(k)-u^1\right\|_{L^2}^2 &\le \left\|a * u^0\right\|_{L^2}^2 + k^2\left\|b * v^0\right\|_{L^2}^2 \\ &\le 2\pi\|\tilde a\|_{L^{2m}}^2\|\tilde u^0\|_{L^{2n}}^2 + 2\pi k^2\|\tilde b\|_{L^{2\mu}}^2\|\tilde v^0\|_{L^{2\nu}}^2\end{aligned}$$

mit

$$\begin{aligned}\|\tilde a\|_{L^{2m}}^2 &= \frac{1}{2\pi}\left\{\int dp\left(\cos pk - \frac{1}{\theta_2}\ \frac{1}{1+k^2\theta_1\theta_2p^2} + \frac{1-\theta_2}{\theta_2}\right)^{2m}\right\}^{\frac{1}{m}},\\ \|\tilde b\|_{L^{2\mu}}^2 &= \frac{1}{2\pi}\left\{\int dp\left(\frac{\sin(kp)}{kp} - \frac{1}{1+k^2\theta_1\theta_2p^2}\right)^{2\mu}\right\}^{\frac{1}{\mu}}.\end{aligned}$$

Betrachtet man die ersten Reihenglieder der Entwicklung nach p, so ergibt sich

$$\begin{aligned}\|\tilde a\|_{L^{2m}}^{2m} &= \frac{1}{2\pi}\int dp\Big((1-2\theta_1)^2\frac{k^4p^4}{4} + (1-2\theta_1)(24\theta_1^2\theta_2-1)\frac{k^6p^6}{24}\\ &\qquad + (8640\theta_1^4\theta_2^2 - 2880\theta_1^3\theta_2^2 - 240\theta_1^2\theta_2 - 8\theta_1 + 9)\frac{k^8p^8}{2880} + \cdots\Big)^m,\\ \|\tilde b\|_{L^{2\mu}}^{2\mu} &= \frac{1}{2\pi}\int dp\Big((1-6\theta_1\theta_2)\frac{k^4p^4}{36} - (1-6\theta_1\theta_2)(1-120\theta_1^2\theta_2^2)\frac{k^6p^6}{360}\\ &\qquad + (3\theta_1^4\theta_2^4 - \frac{1}{3}\theta_1^3\theta_2^3 - \frac{1}{60}\theta_1^2\theta_2^2 - \frac{1}{2520}\theta_1\theta_2 + \frac{41}{302400})p^8k^8 + \cdots\Big)^\mu.\end{aligned}$$

Wählt man also $\theta_1=\frac{1}{2}$ und $\theta_2=\frac{1}{3}$ und führt wieder die Abschneidefrequenz p_0 ein, so erhält man die maximale Ordnung:

$$\begin{aligned}\|\tilde a\|_{L^{2m}}^2 &\le \frac{1}{2\pi}\frac{1}{576}\left(\frac{2}{8m+1}\right)^{\frac{1}{m}} {p_0}^{8+\frac{1}{m}}k^8,\\ \|\tilde b\|_{L^{2m}}^2 &\le \frac{1}{2\pi}\frac{49}{129600}\left(\frac{2}{8\mu+1}\right)^{\frac{1}{\mu}} {p_0}^{8+\frac{1}{\mu}}k^8.\end{aligned}$$

Zusammen ergäbe sich für $n=\nu=1, m=\mu=\infty$:

$$\left\|u(k)-u^1\right\|_{L^2}^2 \le \frac{1}{576}p_0^8\ k^8\|u^0\|_{L^2}^2 + \frac{49}{129600}p_0^8\ k^8\|v^0\|_{L^2}^2. \tag{A.16}$$

Allerdings ist das Verfahren für diese Wahl der θ_i nicht stabil.

Wählt man dagegen mit $\theta = \frac{1}{2}$

$$a_\infty^2(s) = \max_{-s\le x\le s}\left(\cos x - \frac{1}{\theta}\,\frac{1}{1+\theta^2 x} + \frac{1-\theta}{\theta}\right)^2,$$
$$b_\infty^2(s) = \max_{-s\le x\le s}\left(\frac{\sin(x)}{x} - \frac{1}{1+\theta^2 x^2}\right)^2,$$

so ergibt sich wieder

$$\left\|u(k) - u^1\right\|_{L^2}^2 \le a_\infty^2(kp_0)\|u^0\|_{L^2}^2 + b_\infty^2(kp_0)k^2\|v^0\|_{L^2}^2.$$

a_∞^2 und b_∞^2 sind wieder in den Abbildungen A.3 zu finden.

A.4 Das DG(1)-Verfahren

A.4.1 Verfahren

Da wegen der Linearität der Wellengleichung das implizite EULER-Verfahren identisch mit dem unstetigen GALERKIN-Verfahren nullter Ordnung, DG(0), ist, sind die Fehlergrenzen für dieses Verfahren bereits oben hergeleitet. Wir betrachten daher nur noch das DG(1)-Verfahren.

Mit $U = (u, v)^T$ lassen sich (A.1) und (A.2) als

$$U_t = \begin{pmatrix} 0 & 1 \\ \Delta & 0 \end{pmatrix} U =: BU$$

schreiben, wobei B die Systemmatrix sei. Das DG(1)-Verfahren für Systeme schreibt sich

$$U_-^1 = \int_0^k BU(t)\,dt + U_-^0,$$
$$U_-^1 - U_+^0 = \int_0^k BU(t)\frac{t}{k}\,dt.$$

Setzt man $U(t) = \frac{t}{k}U_-^1 + \frac{k-t}{k}U_+^0$ ein, so ergibt sich daraus

$$U_-^1 = \frac{k}{2}B(U_-^1 + U_+^0) + U_-^0,$$
$$U_-^1 - U_+^0 = \frac{k}{3}BU_-^1 + \frac{k}{6}BU_+^0,$$

oder wieder in „alter“ Schreibweise:

$$u_-^1 = \frac{k}{2}(v_-^1 + v_+^0) + u_-^0, \tag{A.17}$$
$$v_-^1 = \frac{k}{2}\Delta(u_-^1 + u_+^0) + v_-^0, \tag{A.18}$$
$$u_-^1 - u_+^0 = \frac{k}{3}v_-^1 + \frac{k}{6}v_+^0, \tag{A.19}$$
$$v_-^1 - v_+^0 = \frac{k}{3}\Delta u_-^1 + \frac{k}{6}\Delta u_+^0. \tag{A.20}$$

Aus (A.17) und (A.19) ergibt sich der Postprocessing-Schritt, um U_-^1 zu erhalten:

$$U_-^1 = P_+U_+^0 + P_-U_-^0$$

mit

$$P_+ = \begin{pmatrix} 3 & -\frac{k}{2} \\ \frac{6}{k} & -2 \end{pmatrix}, \qquad P_- = \begin{pmatrix} -2 & 0 \\ -\frac{6}{k} & 0 \end{pmatrix}.$$

Die beiden Matrizen $P_\pm$ propagieren sozusagen die Lösung auf dem Intervall $(0, k)$ von einem Zeitschritt zum nächsten.

Eingesetzt in (A.18) und (A.20) ergibt sich die Gleichung, die uns über den Sprung in $t = 0$ hinweghilft:

$$\begin{pmatrix} \frac{6}{k} - 2k\Delta & -2 + \frac{k^2}{4}\Delta \\ \frac{6}{k} - \frac{7}{6}k\Delta & -3 + \frac{k^2}{6}\Delta \end{pmatrix} U_+^0 = \begin{pmatrix} \frac{6}{k} - k\Delta & 1 \\ \frac{6}{k} - \frac{2}{3}k\Delta & 0 \end{pmatrix} U_-^0. \tag{A.21}$$

A.4.2 Analyse

Man kann Gleichung (A.21) diagonalisieren:

$$(144 - 76k^2\Delta + k^4\Delta^2)U_+^0 = \Lambda(\partial)U_-^0,$$

mit dem Matrixoperator

$$\Lambda(\partial) = \begin{pmatrix} 4(36 - 7k^2\Delta) & 4k(18 - k^2\Delta) \\ 4k\Delta(18 - k^2\Delta) & 4(36 - 7k^2\Delta) \end{pmatrix}.$$

Diese Darstellung ist zwar für die Numerik ungeeignet, da sie zuviel Regularität von der Lösung fordert; sie erlaubt aber eine einfachere Analyse, da die GREENsche Funktion $g(y)$ zum Operator auf der linken Seite skalar ist. $g(y)$ ist in diesem Fall nicht mehr berechenbar; das macht aber nichts, da wir g auch gar nicht brauchen, sondern die FOURIER-Transformierte $\tilde{g}$, die direkt durch

$$\tilde{g}(p) = \frac{1}{\sqrt{2\pi}} \frac{1}{144 + 76k^2p^2 + k^4p^4} \tag{A.22}$$

gegeben ist.

Damit gilt:

$$U_+^0 = \left(g * \Lambda(\partial)U_-^0\right)(x),$$

und mit den $P_\pm$-Matrizen

$$\begin{aligned} U_-^1(x) &= \left(g * P_+\Lambda(\partial)U_-^0\right)(x) + P_-U_-^0(x) \\ &= \int dy \left(\mathcal{F}^{-1}[\tilde{g}](x-y)P_+\Lambda(\partial_y) + \delta(x-y)P_-\right) U_-^0(y). \end{aligned}$$

Da g skalar ist, vertauscht es mit den Matrizen und wir können die partielle Integration einfach durchführen:

$$= \int dy \left(P_+\Lambda(\partial_y)\mathcal{F}^{-1}[\tilde{g}](y) + \delta(y)P_-\right) U_-^0(x-y).$$

Eigentlich würde man im Argument zu Λ durch die partielle Integration ein anderes Vorzeichen erhalten, das jedoch durch das Vorzeichen von y im Argument zu g wieder verschwindet. Man kann nun Λ mit unter das Integral der Fouriertransformation ziehen und erhält damit

$$\begin{aligned} U_+^0(x) &= \int dy\ \mathcal{F}^{-1}\left[P_+\Lambda(-ip)\tilde{g} + \frac{1}{\sqrt{2\pi}}P_-\right](y)\ U_-^0(x-y) \\ &= \left(T * U_-^0\right)(x) \end{aligned}$$

mit der Transfermatrix

$$T = \mathcal{F}^{-1}\left[P_+\Lambda(-ip)\tilde{g} + \frac{1}{\sqrt{2\pi}}P_-\right].$$

Daraus erhalten wir wieder unsere Fehleridentität:

$$\left(u(k) - u^1\right)(x) = (a * u^0)(x) + k(b * v^0)(x)$$

mit den Größen

$$a(y) = \frac{1}{2}(\delta(y-k) + \delta(y+k)) - T_{11},$$
$$b(y) = \frac{1}{2k}\theta(k - |y|) - \frac{1}{k}T_{12}.$$

L^∞-Fehler

Der L^∞-Fehler ist diesmal nicht so leicht berechenbar, da wir $a(y)$ und $b(y)$ nicht explizit kennen, sondern nur deren FOURIER-Transformierten. Möglicherweise ist es trotzdem machbar, indem man Eigenschaften der inversen FOURIER-Transformation trickreich nutzt, um die L^1-Norm der Integralkerne zu berechnen; dieser Weg sei hier aber nicht weiter verfolgt.

L^2-Fehler

Die Normen der FOURIER-Transformierten der Integralkerne sind jetzt mit der Abschneidefrequenz p_0 durch

$$\|\tilde{a}\|^2_{L^{2m}} = \frac{1}{2\pi}\left\{\int_{-p_0}^{p_0} dp \left(\cos pk - \frac{2(216 + 60k^2p^2 + k^4p^4)}{144 + 76k^2p^2 + k^4p^4} + 2\right)^{2m}\right\}^{\frac{1}{m}},$$
$$\|\tilde{b}\|^2_{2\mu} = \frac{1}{2\pi}\left\{\int_{-p_0}^{p_0} dp \left(\frac{\sin(kp)}{kp} - \frac{144 - 2k^2p^2}{144 + 76k^2p^2 + k^4p^4}\right)^{2\mu}\right\}^{\frac{1}{\mu}}$$

gegeben. Diese Normen hier explizit zu berechnen ist zwar möglich (und ergibt Proportionalitäten zu $p_0^{4+\frac{1}{m}}k^4$ und $p_0^{4+\frac{1}{\mu}}k^4$), ist aber nicht sinnvoll, da die Reihenentwicklungen sehr schlecht konvergieren; insbesondere haben die ersten Terme das gleiche Vorzeichen. Wir definieren deshalb wieder

$$a^2_\infty(s) = \max_{-s\leq x\leq s}\left(\cos x - \frac{2(216 + 60x^2 + x4)}{144 + 76x^2 + x^4} + 2\right)^2,$$
$$b^2_\infty(s) = \max_{-s\leq x\leq s}\left(\frac{\sin x}{x} - \frac{144 - 2x}{144 + 76x^2 + x^4}\right)^2,$$

und erhalten

$$\left\|u(k) - u^1\right\|^2_{L^2} \leq a^2_\infty(kp_0)\|u^0\|^2_{L^2} + b^2_\infty(kp_0)k^2\|v^0\|^2_{L^2}.$$

A.5 Bedeutung der Bedingung „$\tilde{u}(p) = 0$ für $|p| \geq p_0$“

In der Herleitung der L^2-Fehlerabschätzungen wurde von der Bedingung

$$\tilde{u}(p) = 0 \qquad \text{für} \qquad |p| \geq p_0$$

Gebrauch gemacht, um die Integrale asymptotisch berechnen zu können. Die Forderung dieser Bedingung erscheint auf Anhieb etwas willkürlich, ist aber bei genauerer Betrachtung durchaus

vernünftig: nimmt man an, daß die Bedingung gilt, so gibt es eine größte Frequenz p_0 in der Fourierentwicklung von $u(x,t)$. Die (räumlich) am schnellsten variierenden Anteile von $u(x,t)$ haben also die Frequenz p_0 und damit die Wellenlänge $\lambda = \frac{2\pi}{p_0}$.

Im Rahmen der späteren räumlichen Diskretisierung legen wir durch die Wahl einer minimalen vetretbaren Gitterweite fest, bis zu welcher Größenordnung wir die Eigenschaften der exakten Lösung approximieren wollen. Eine vernünftige Auflösung ist beispielsweise für Variationen der Wellenlänge $\lambda \geq 2h$ mit der Ortsgitterweite h gegeben (eine Periode eines Sinus mit weniger als zwei Geradenstücken anzunähern geht vernünftigerweise nicht). Wenn wir also sagen, daß wir mit Gitterweiten bis hinunter zu (zum Beispiel) $h = 0.01$ rechnen, dann implizieren wir, daß wir keine Variationen der Lösung mehr auflösen können, deren Wellenlänge kleiner als $\lambda_0 = 0.02$ ist. Das entspricht aber eine maximalen Frequenz von $p_0 = \frac{2\pi}{\lambda} = 100\pi$. In der Praxis erhält man eine vernünftige numerische Lösung sogar nur für $\lambda \geq 10h \ldots 20h$.

Daraus folgt für die Festlegung von p_0, daß es als das Minimum der beiden folgenden Größen zu wählen ist:

- die größte tatsächlich in der Lösung vorhandene Frequenz, falls die Lösung glatt ist und wir die kleinsten Eigenschaften von $u(x)$ mit mehr als zwei Ortsgitterweiten auflösen wollen; oder
- die größte Frequenz, die wir auf einem gegebenen Gitter noch auflösen können, das heißt $\lambda \approx 2h$.

In den Fällen wo die Lösung glatt genug ist (Fall 1) ist p_0 eine Konstante und wir können Vorfaktoren wie Cp_0^2kT wie beim EULER-Verfahren durch Wahl von k beliebig klein machen. Wollen wir zum Beispiel eine relative Genauigkeit von 10 Prozent (bezüglich der L^2-Norm) erreichen, so muß $\frac{1}{2}p_0^2kT \leq 0.1$ sein. Nimmt man beispielsweise an, daß wir bis zur Zeit $T = 10$ rechnen wollen und die kleinsten Dimensionen von $u(x)$ bei $\frac{1}{100}$ liegen, wobei $h < \frac{1}{200}$, dann ist $p_0 \approx 630$ und wir erhalten für die Zeitschrittweite $k \leq 8 \cdot 10^{-5}$. Das ist sicher zu klein, wir hätten $k \approx \frac{h}{T} = 5 \cdot 10^{-4}$ erwartet. Die a-priori-Abschätzung liegt aber damit immerhin in einem realistischen Bereich. Für das Crank-Nicolson-Verfahren erhält man aus $\frac{1}{12}p_0^4k^3T \leq 0.1$ die Schrittweite $k \leq 2 \cdot 10^{-4}$.

Mit diesen Zahlen ist ersichtlich, weshalb beim Crank-Nicolson-Verfahren (A.16) der erste Term den zweiten trotz der besseren Ordnung dominiert: setzt man p_0 und k von eben ein, so erhält man

$$\left\|u(k) - u^1\right\|_{L^2}^2 \leq 0.044\|u^0\|_{L^2}^2 + 7 \cdot 10^{-8}\|v^0\|_{L^2}^2$$

Im zweiten Fall nehmen wir an, daß die maximale Frequenz durch unsere kleinste Gitterweite gegeben ist, d. h. $p_0 = \frac{2\pi}{2h}$. Dann ist die Bedingung $\frac{1}{2}p_0^2kT \leq 0.1$ gleichbedeutend mit $k \leq 0.03\frac{h}{T}h$. Beim Crank-Nicolson-Verfahren erhielte man $k \leq 0.25\left(\frac{h}{T}\right)^{\frac{1}{3}}h$.

A.6 Vergleich der Verfahren

Alle Verfahren haben die Fehlerabschätzung

$$\left\|u(k) - u^1\right\|_{L^2}^2 \leq a_\infty^2(kp_0)\|u^0\|_{L^2}^2 + b_\infty^2(kp_0)k^2\|v^0\|_{L^2}^2 \tag{A.23}$$

mit verschieden definierten a_∞ und b_∞, die in den Abbildungen A.3 einander gegenübergestellt sind. Man kann deutlich die verschiedenen Ordnungen der Verfahren ablesen.

Bei der Wellengleichung heben sich einmal gemachte Fehler nicht wieder weg, so daß wir bei einer gewünschten Genauigkeit ε zur Endzeit T fordern müssen, daß

$$\left\|u(k) - u^1\right\|_{L^2} \leq \frac{k}{T}\varepsilon$$

gilt. Nimmt man einmal an, daß in (A.23) nur der erste Term existierte, dann müßte $\frac{a_\infty^2(kp_0)}{k^2} \approx \frac{\varepsilon^2}{T^2}$ gelten. Bei einer geforderten Genauigkeit von 0.1, einer Endzeit von 10 und $p_0 = 300$ (entsprechend

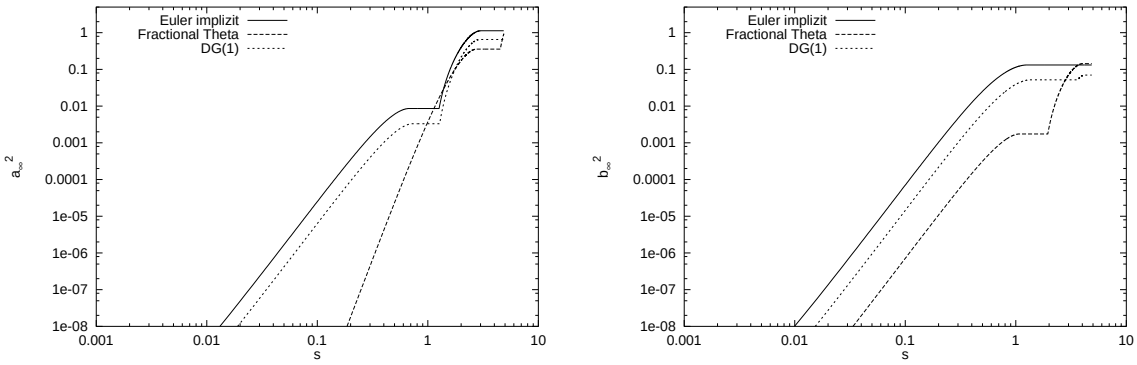

Abbildung A.3: *Vergleich der Funktionen $a_\infty^2(s)$ und $b_\infty^2(s)$ für die verschiedenen untersuchten Verfahren.*

einer kleinsten Wellenlänge von 0.02 und damit $h \leq 0.01$) ergibt das bereits für das Fractional-θ für k die obere Schranke $5 \cdot 10^{-5}$, also ein zweihundertstel der größten vernünftigen Ortsgitterweite. Für die anderen Verfahren ist das maximale k um etliche weitere Größenordnungen kleiner. Der maximale Fehler darf daher nicht als Kriterium für die Wahl der Zeitschrittweite angesehen werden.

Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurden Verfahren zur adaptiven Lösung der Wellengleichung in inhomogenen Medien mittels Finiten Elementen entwickelt. Dabei wurde zuerst eine Diskretisierung in Ort und Zeit vorgestellt; als GALERKIN-Verfahren ist sie in der Lage, als Fundament der adaptiven Methoden mit dualen Fehlerschätzern zu dienen.

Auf dieser Basis wurde ein allgemeiner Ansatz entwickelt, der es erlaubt, Gitter zu erzeugen, die neben der numerischen Lösung auch an die Quantität angepaßt ist, an der man interessiert ist. Für diese Größe kommen nahezu alle Funktionale in Frage, die sich auf die Lösung anwenden lassen. Dazu gehören unter anderem Linien- und Gebietsintegrale, mit und ohne Integration über die Zeit, aber auch andere Auswertungen; der Ansatz ist insbesondere nicht auf lineare Zielfunktionale beschränkt, es wurden auch Wege aufgezeigt, wie nichtlineare Funktionale verwendet werden können. Zur Berücksichtigung der Zielgröße ist es nötig, ein zweites, „duales“ Problem zu lösen, das die Information enthält, welche Orte zu einem gegebenen Zeitpunkt zum Ergebnis beitragen, und mit welchem Gewicht sie dies tun.

Dieser Formalismus wurde in den darauffolgenden Kapiteln auf einige Beispiele angewandt. Dabei wurden eine Reihe von Fällen mit verschiedenen Zielfunktionalen untersucht, bei denen sich die Überlegenheit des dualen Fehlerschätzers über gleichförmige Verfeinerung, aber auch gegenüber „traditionellen“ Fehlerindikatoren auf der Basis der Energienorm zeigen ließ. Es konnte gezeigt werden, daß der duale Schätzer in der Lage ist, erheblich effizientere Gitter zu erzeugen, mit denen sich die gewünschte Genauigkeit mit deutlich geringerer Anzahl an Freiheitsgraden berechnen läßt.

In der Hauptanwendung der Arbeit wurde der Energietransport am Übergang zwischen solarer Chromosphäre und Korona untersucht. Es zeigte sich, daß die Verfeinerung mit dem Energiefehlerindikator gute Ergebnisse zeigte, während der duale Fehlerschätzer im wesentlichen nicht zu befriedigenden Resultaten führte. Als Gründe dafür kommen unzureichende Linearisierungen der Zielfunktionale um die primale Lösung, die starke Inhomogenität des Mediums und Diskretisierungsfehler in Frage. Am Beispiel des Energieflusses durch eine Linie konnte gezeigt werden, daß die Linearisierung in der Tat nur sehr schlecht konvergiert. Es ist allerdings nicht klar, ob dies der alleinige Grund für das Versagen ist; hier sind noch weitere Untersuchungen nötig, um die genauen Gründe zu bestimmen.

Die Ergebnisse dieser Arbeit, vor allem am Beispiel aus der Sonnenphysik, werfen eine Reihe neuer Fragen auf, die zu untersuchen aus zeitlichen Gründen nicht mehr möglich waren. Zu diesen Fragen gehören:

- Ist das Nichtfunktionieren des dualen Schätzers endgültig, d. h. divergieren die Ergebnisse auch bei weiterer VerfeinerungΓ Da im zeitlichen Rahmen dieser Arbeit die Implementation von Mehrgitteralgorithmen nicht möglich war, konnten aus Rechenzeitgründen nur eine sehr beschränkte Zahl von Verfeinerungen mit dem dualen Schätzer durchgeführt werden. Man beobachtet auch bei der Verfeinerung mit dem Energiefehlerindikator, daß starke Veränderungen des Gitters zu zeitweilig gestörter Konvergenz führen kann; da nach Umschalten vom Energiefehlerindikator auf den dualen Schätzer größere Umstrukturierungen des Gitters stattfinden, wäre es möglich, daß nur die ersten Schritte danach ein divergentes Verhalten suggerieren, anschließend jedoch eine beschleunigte Konvergenz stattfindet. Hier könnten bereits wenige weitere Verfeinerungsschritte Indizien geben, was jedoch erheblich verbesserte Lösungsalgorithmen voraussetzt.

- Welchen Effekt hat die schlechte Linearisierung der nichtlinearen ZielfunktionaleΓ Aus den berechneten Verläufen des Energieflusses durch die Auswertungslinie würde man erwarten, daß der Fehlerschätzer nicht dazu führt, daß Verfeinerung an den richtigen Stellen unterlassen wird; man erwartet eher, daß der verfeinerte Bereich zu groß ist. Es ist daher nicht klar, ob die schlechte Linearisierung tatsächlich für das Versagen verantwortlich ist. Hier wäre in einem ersten Schritt zu untersuchen, ob bei dem gegebenen Problem die Verfeinerung mit dem dualen Schätzer zu einem *linearen* Zielfunktional zum Erfolg führt.

- Welche Rolle spielen die stark variierenden KoeffizientenΓ Sie sollten die Konvergenz von primaler und dualer Lösung wesentlich verschlechtern und könnten daher dazu sowohl zur schlechten Linearisierung als auch zu ungenauer Berechnung der dualen Lösung beitragen. Darüberhinaus verschlechtern sie die Kondition der Systemmatrizen erheblich und erhöhen damit den numerischen Aufwand zu deren Invertierung stark, was weitere Konvergenzanalysen erschwert. In vielen Anwendungen aus der Praxis sind die Variationen der Koeffizienten erheblich geringer (in der Geophysik beispielsweise in der Größenordnung von unter zehn, während sie hier beinahe einen Faktor Tausend ausmachte), so daß die gleichen Verfahren dort erfolgreich sein könnten; einige der gerechneten nicht anwendungsbezogenen Beispiele legen diese Vermutung nahe.

Daneben sind noch die folgenden, allgemeineren Fragen unbeantwortet:

- Ist der duale Fehlerschätzer in der Lage, teilweise Reflexion an Gitterunstetigkeiten zu erkennen und gegebenenfalls zu verhindern, sofern das für die Auswertung des Zielfunktionals notwendig istΓ

- Zu welchen Ergebnissen führt die Verwendung anderer Möglichkeiten der Auswertung der FehleridentitätΓ Dazu existieren neben der skizzierten Verwendung des Bramble-Hilbert-Lemmas die Extrapolation einer numerischen dualen Lösung auf einen höheren Ansatzraum auf dem Patch. Letztlich ist die Ersetzung des verwendeten Verfahrens durch ein anderes unabdingbar, da sich der sehr erhebliche zusätzliche Aufwand zur Berechnung der dualen Lösung in der Praxis weder rechtfertigen noch bereitstellen läßt. Erste Experimente in diese Richtung lassen vermuten, daß sich die Qualität der erzeugten Gitter nicht wesentlich ändert; da quantitative Fehlerschätzung in den meisten Fällen ohnehin nicht möglich erscheint, ist dies ausreichend zur Verwendung „billigerer“ Fehlerschätzer.

- Weshalb wurde ein Versagen des Konzepts der dualen Fehlerschätzer bisher nie beobachtetΓ Was unterscheidet die Wellengleichung von den bisher betrachteten SystemenΓ Sowohl bei nichtlinearen Gleichungen wie der Elasto-Plastizität oder den Navier-Stokes-Gleichungen, als auch bei zeitabhängigen Problemen wie der Wärmeleitungsgleichung wurden duale Fehlerschätzer erfolgreich angewandt. Diese Gleichungen wurden allerdings entweder zeitunabhängig gelöst (Elasto-Plastizität, Navier-Stokes), oder hatten stark diffusive Eigenschaften (Wärmeleitung), so daß es denkbar ist, daß erst die völlige Abwesenheit von Dämpfungseigenschaften und Fehlerakkumulation bei zeitabhängigen Problemen in Kombination mit den hohen rechentechnischen Anforderungen dazu führten, daß es nicht möglich war, in den Bereich der Konvergenz der Berechnung von Fehlerschätzern und Gitterverfeinerung zu gelangen. Diese Vermutung ist bisher allerdings in keiner Weise untermauert.

Abschließend läßt sich sagen, daß die Vorteile adaptiver Verfahren gegenüber globaler Verfeinerung deutlich gezeigt werden konnte, ebenso wie die Überlegenheit des dualen Schätzers gegenüber traditionellen Fehlerindikatoren bei einfachen Problemen. Allerdings scheiterte dieses Verfahren in einem komplexen Anwendungsfall aus der Praxis weitgehend; die Gründe des Scheiterns sind zu einem erheblichen Teil unverstanden und bieten Stoff und Anlaß für weiterführende Untersuchungen auf diesem Gebiet.

Notationen

$\mathbf{w} = (u, v)$	Exakte Lösung
$\mathbf{w}_h = (u_h, v_h)$	Numerische Approximation der Lösung
$\mathbf{w}_h^n = (u_h^n, v_h^n)$	Numerische Approximation der Lösung zum Zeitpunkt t_n
$\bar{\mathbf{w}} = (\bar{u}, \bar{v})$	Exakte duale Lösung
$\bar{\mathbf{w}}_h = (\bar{u}_h, \bar{v}_h)$	Numerische Approximation der dualen Lösung
$\mathbf{e} = \mathbf{w} - \mathbf{w}_h$	Fehler zwischen exakter und numerischer Lösung
$\mathbf{t} = (\varphi, \psi)$	Testfunktion des kontinuierlichen Problems
$\mathbf{t}_h = (\varphi_h, \psi_h)$	Testfunktion des diskreten Problems
$\mathcal{W}_h$	Diskreter Ansatzraum
$\mathcal{T}_h$	Diskreter Testraum
k_n	$t_n - t_{n-1}$ Zeitschrittweite im $n$ten Zeitintervall I_n
h_K	Durchmesser der Zelle K
$\mathbb{T}^n$	Triangulierung des Gebiets Ω auf dem Zeitintervall I_n
$\rho(\mathbf{x})$	Dichte
$a(\mathbf{x})$	Steifigkeitskoeffizient
$(f, g)_\Omega$	$\int_\Omega dx\ f^*(\mathbf{x}) g(\mathbf{x})$
$\|u\|$	$(u, u)_\Omega^{1/2}$
Δ	LAPLACE-Operator $\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$
$\mathcal{A}$	$-\nabla \cdot a(\mathbf{x}) \nabla$
$a(u, v)$	$(a(\mathbf{x}) \nabla u, \nabla v)_{\Omega \times [0,T]}$
$\mathcal{B}$	$\begin{pmatrix} 0 & -\rho(\mathbf{x}) \\ \mathcal{A} & 0 \end{pmatrix}$
$b(\mathbf{w}, \mathbf{t})$	$\left(\begin{pmatrix} 0 & -\rho \\ a(\mathbf{x})\nabla & 0 \end{pmatrix} \mathbf{w}, \begin{pmatrix} 1 & 0 \\ 0 & \nabla \end{pmatrix} \mathbf{t} \right)_{\Omega \times [0,T]} - (g_N, t_2)_{\Gamma_N \times [0,T]}$
g_D	DIRICHLET-Randwerte
g_N	NEUMANN-Randwerte
$\mathcal{J}(\cdot)$	Zielfunktional
$J(\cdot)$	Exaktes Fehlerfunktional
$\tilde{J}(\cdot)$	Linearisiertes Fehlerfunktional
$\Delta J(\cdot) = J(\cdot) - \tilde{J}(\cdot)$	Linearisierungsfehler des Fehlerfunktionals
$\mathcal{E}_{K,n}$	Exakter Beitrag von $K \times I_n$ zum Fehler $J(\mathbf{e})$
$\eta_{K,n}$	Fehlerschätzer für $K \times I_n$ zum Fehler $J(\mathbf{e})$

Literaturverzeichnis

[1] Najib A. Abboud and Peter M. Pinksky. Finite element dispersion analysis for the three-dimensional second-order scalar wave equation. *Int. J. Numer. Meth. Eng.*, 35:1183–1218, 1992.

[2] Shin Aoi and Hiroyuki Fujiwara. 3-d finite-difference method using discontinuous grids. *submitted to BSSA*, 1998.

[3] Wolfgang Bangerth. DEAL.II *technical reference*. Universität Heidelberg, `http://gaia.iwr.uni-heidelberg.de/~wolf`, 1998.

[4] Roland Becker. *An adaptive finite element method for the incompressible Navier-Stokes equations on time-dependent domains*. Doktorarbeit, Universität Heidelberg, 1995.

[5] Roland Becker, Guido Kanschat, and Franz-Theo Suttmeier. DEAL – a differential equations analysis library. Technical report, Preprint, IWR, Universität Heidelberg, 1995.

[6] Roland Becker and Rolf Rannacher. A feed-back approach to error control in finite element methods: Basic analysis and examples. *East-West J. Numer. Math.*, 4(4):237–264, 1996.

[7] S. C. Brenner and R. L. Scott. *The Mathematical Theory of Finite Elements*. Springer, Berlin-Heidelberg-New York, 1994.

[8] M. O. Bristeau, R. Glowinski, and J. Periaux. Numerical methods for the Navier-Stokes equations: applications to the simulation of compressible and incompressible viscous flows. *Report UH/MD-4, Univ. of Houston, in Computer Physics Report*, 1987.

[9] Anatoliy G. Butkovski. *Green's Functions and Transfer Functions Handbook*. Series in Mathematics and its Applications. Ellis Horwood Ltd., New York, Brisbane, Chichester, Toronto, 1982.

[10] Mats Carlsson and Robert F. Stein. Radiation shock dynamics in the solar chromosphere – results of numerical simulations. In Mats Carlsson, editor, *Proceedings of a mini-workshop held at Institute of Theoretical Astrophysics, University of Oslo, Norway, June 6-8, 1994*, 1994.

[11] C. Cerjan, D. Kosloff, R. Kosloff, and M. Reshef. A nonreflecting boundary condition for discrete acoustic and elastic wave equations. *Geophysics*, 50:705–708, 1985.

[12] R. Clayton and B. Engquist. Absorbing boundary conditions for acoustic and elastic wave equations. *Bull. Seism. Soc. Am.*, 67:1529–1540, 1977.

[13] Richard Courant, K. O. Friedrichs, and Hans Lewy. Über die partiellen Differentialgleichungen der Mathematischen Physik. *Mathematische Annalen*, 100:32–74, 1928.

[14] Bjorn Engquist and Andrew Majda. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comp.*, 31(139):629–651, July 1977.

[15] J. M. Fontenla, E. H. Avrett, and R. Loeser. Energy balance in the solar transition region. III. Helium emission in hydrostatic, constant-abundance models with diffusion. *Ap. J.*, 406:319–345, March 1993.

[16] Donald A. French and Todd E. Peterson. A continuous space-time finite element method for the wave equation. *Math. Comp.*, 65(214):491–506, April 1996.

[17] M. J. Grote. Nonreflecting boundary conditions for electromagnetic scattering. Research Report 98-09, Seminar für Angewandte Mathematik, Eidgenössische Technische Hochschule, Zürich, September 1998.

[18] M. J. Grote and J. B. Keller. Exact nonreflecting boundary condition for elastic waves. Research Report 98-08, Seminar für Angewandte Mathematik, Eidgenössische Technische Hochschule, Zürich, September 1998.

[19] Ralf Hartmann. A-posteriori Fehlerschätzung und adaptive Schrittweiten- und Ortsgittersteuerung bei Galerkin-Verfahren für die Wärmeleitungsgleichung. Diplomarbeit, Universität Heidelberg, 1998.

[20] R. L. Higdon. Absorbing boundary conditions for acoustic and elastic wave equations in stratified media. *J. Comp. Phys.*, 101:386–418, 1992.

[21] Patrick Joly. Developments in numerical methods for transient scattering problems. In Guy Chavent and Pierre C. Sabatier, editors, *Inverse Problems of Wave Propagation. Proceedings of the conference held in Aix-les-Bains, France, September 23–27, 1996*, Lecture Notes in Physics 486. Springer, 1997.

[22] Guido Kanschat. *Parallel and adaptive Galerkin methods for radiative transfer problems.* Doktorarbeit, Universität Heidelberg, 1996.

[23] D. W. Kelly, J. R. Gago, O. C. Zienkiewicz, and I. Babuška. A posteriori error analysis and adaptive processes in the finite element method. *Internat. J. Numer. Methods Engrg.*, 19:1593–1619, 1983.

[24] P. Kloucek and F. S. Rys. On the stability of fractional step θ-scheme for the nonstationary Navier-Stokes equations. *SIAM J. Numer. Anal.*, 31(5):1312–1335, 1994.

[25] Peter Korevaar. *Time-dependent models of stellar coronae.* Doktorarbeit, Rijksuniversiteit Utrecht, 1989.

[26] Y. Q. Lou, R. Rosner, and P. Ulmschneider. A computational code for two-dimensional unsteady magnetohydrodynamics by the methods of characteristics. *Ap. J.*, 315:349–370, April 1987.

[27] S. Müller, A. Prohl, R. Rannacher, and S. Turek. Implicit time-discretization of the nonstationary incompressible Navier-Stokes equations. In W. Hackbusch and G. Wittum, editors, *Fast Solvers for Flow Problems, Proceedings of the tenth GAMM-Seminar, Kiel January 14–16, 1994*. F. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 1994.

[28] R. Rannacher. Numerical analysis of nonstationary fluid flow. A survey. In V. C. Boffi and H. Neunzert, editors, *Applications of Mathematics in Industry and Technology*, pages 34–53. B. G. Teubner, Stuttgart, 1989.

[29] Endre Süli and Catherine Wilkins. Adaptive finite element methods for the damped wave equation. Report no. 96/23, Oxford University Computing Laboratory, 1996.

[30] Franz-Theo Suttmeier. Private Mitteilung.

[31] J. Theurer, P. Ulmschneider, and W. Kalkofen. Acoustic wave propagation in the solar atmosphere. V. Observation versus simulation. *Astron. Astrophys.*, 324:717–724, 1997.

[32] Joachim Theurer. *Generation and Propagation of Acoustic Wave Spectra in Late-Type Stellar Atmosphere*. Doktorarbeit, Universität Heidelberg, 1998.

[33] Peter Ulmschneider. Acoustic waves in the solar atmosphere. VIII. Extrapolation in time. *Astron. Astrophys.*, 168:308–310, 1986.

[34] J. E. Vernazza, E. H. Avrett, and R. Loeser. Structure of the solar chromosphere. III. Models of the EUV brightness of the quiet sun. *Ap. J. Supp.*, 45:635–725, April 1981.

[35] J. Virieux. P-SV wave propagation in heterogeneous media: velocity-stress finite-difference method. *Geophysics*, 51:889–901, 1986.