Stetige Galerkinverfahren für zeitabhängige Maxwellgleichungen mit Kerr-Nichtlinearität

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Dipl.–Math. techn. Hannes Gerner
aus Wiesloch

Tag der mündlichen Prüfung: 23. Oktober 2013

Referent: Prof. Dr. Willy Dörfler
Korreferent: Prof. Dr. Tobias Jahnke

Diese Arbeit widme ich meinen Großeltern und Peter.

Inhaltsverzeichnis

Danksagungen

Ich danke Prof. Dr. Willy Dörfler dafür, dass er meine Doktorarbeit in so vielen Stunden betreut und unterstützt hat, sowie meinem Zweitbetreuer Prof. Dr. Tobias Jahnke für die Diskussionen und auch mal aufbauenden Worte. Ich danke Prof. Dr. Roland Schnaubelt für die Hinweise, die zum Beweis der eindeutigen Lösbarkeit der in dieser Arbeit behandelten Differentialgleichung geführt haben. Weiterhin danke ich JProf. Dr. Tomáš Dohnal für die Fragestellung dieser Doktorarbeit.

Für die Finanzierung während meiner Promotionszeit im Graduiertenkolleg 1294 „Analysis, Simulation und Design nanotechnologischer Prozesse“ als Stipendiat und Mitarbeiter bin ich der Deutschen Forschungsgesellschaft, sowie Prof. Dr. Dörfler und Prof. Dr. Wieners zu Dank verpflichtet.

Ich danke herzlich Dr. Markus Richter dafür, dass er mir seine FEM-Bibliothek zur Verfügung gestellt und ausführlich erklärt und dass er in den Diskussionen des Graduiertenkollegs moderierend eingegriffen hat. Außerdem danke ich ihm und dem Rest des Instituts für Numerische Mathematik für die immer wieder amüsanten Kaffeestunden.

Für eine sehr angenehme Atmosphäre bei der Arbeit danke ich meinen Kollegen des Graduiertenkollegs. Dabei geht besonderer Dank an meine Zimmerkollegen Kai Sandfort, Piotr Idzik und Hans-Jürgen Freisinger für die vielen mathematischen und nichtmathematischen Diskussionen, aber auch an Bernhard Barth, Stefan Findeisen, Dominik Müller, Philipp Schmalkoke und Anton Verbitsky.

Da fähigen Lehrern viel zu selten gedankt wird, möchte ich an dieser Stelle dreien meiner Mathematiklehrer meiner Gymnasialzeit am E.I. Heidelberg meinen Dank aussprechen. Zuerst ist dabei Frau S. Rühtz zu nennen, deren Gewissenhaftigkeit und teilweise Gnadenlosigkeit bei der Einübung von Rechenfertigkeiten das Fundament geschaffen haben, auf das ich bis heute zäh-

len kann. Dann danke ich Frau C. Eisenlohr, die durch Zusatzaufgaben und den spielerischen Umgang mit der Mathematik erst richtig mein Interesse an mathematischen Problemen geweckt hat und zu guter Letzt Herrn H. Volkert für das Ermöglichen der Teilnahmen an den „Tagen der Mathematik“ in Karlsruhe und Heidelberg, die mich schlussendlich nach Karlsruhe geführt haben.

Schließlich danke ich meinen Eltern, ohne deren Unterstützung das alles nicht möglich gewesen wäre.

Einführung 1

Diese Arbeit habe ich im Rahmen des Graduiertenkollegs 1294 der Deutschen Forschungsgemeinschaft durchgeführt. Das Arbeitsgebiet dieses Graduiertenkollegs ist „Analysis, Simulation und Design nanotechnologischer Prozesse“. Das beinhaltet die mathematische Modellierung und Simulation nichtlinearer optischer Effekte und der Wellenausbreitung in sogenannten photonischen Kristallen. Eine Einführung zu diesem Themenkomplex liefert beispielsweise [DLP+11]. In dieser Arbeit befassen wir uns mit einem nichtlinearen optischen Effekt, dem sogenannten Kerr-Effekt. Dieser Effekt kann in photonischen Kristallen kontrolliert werden. Dabei versteht man unter einem photonischen Kristall einen Kristall aus verschiedenen dielektrischen Stoffen, der analog zu einem elektrischen Halbleiter zur Manipulation elektromagnetischer Wellen verwendet werden kann, siehe beispielsweise [JJWM11].

1.1 Der Kerr-Effekt und Solitonen

Im Jahr 1875 entdeckte der Physiker John Kerr, dass gewisse Substanzen durch das Anlegen eines äußeren elektrischen Feldes ihren Brechungsindex intensitätsabhängig ändern. Dieser Effekt wird heute der *elektrooptische Kerr-Effekt* genannt. Er nutzte diesen Effekt für die Konstruktion einer *Kerr-Zelle*, einer mit einer speziellen Flüssigkeit gefüllten Zelle, durch die er die Polarisation einfallenden Lichts modulieren konnte: Schickt man durch die Zelle eine Lichtwelle mit bekannter Polarisation, beispielsweise durch Verwendung eines Lasers, so kann man ein elektrisches Signal in ein optisches Signal umwandeln, die Änderung der Polarisation der Lichtwelle entspricht der Änderung des elektrischen Feldes. Der Kerr-Effekt ist ein nichtlinearer Effekt, der den Brechungsindex einer Substanz quadratisch vom anliegenden elektrischen Feld abhängen lässt. Zwar tritt dieser Effekt bei allen Substanzen auf, er ist aber meist so klein, dass er von anderen Effekten überlagert wird. In gewissen Flüssigkeiten, wie dem von Kerr verwendeten Nitrobenzol, aber auch in manchen Kristallen ist der Effekt nutzbar.

Der Kerr-Effekt hat eine selbstphasenmodulierende Wirkung auf Wellen. Das heißt, dass ein Puls, der sich in einem Kerr-Medium ausbreitet, eine Phasenverschiebung erfährt. Unter bestimmten Bedingungen wirkt dieser Effekt der Dispersion eines Wellenpulses entgegen, verhindert also ein örtliches Auseinanderlaufen des Pulses und schafft dadurch ein sogenanntes Soliton. Unter einem *Soliton* versteht man im Wesentlichen eine örtlich begrenzte Welle (eine *Einzelwelle*, englisch *solitary wave*) die sich unter Beibehaltung ihrer Form ausbreitet und die mit anderen Solitonen nur in Form einer Phasenänderung interagiert. Das heißt, dass sich zwei aufeinandertreffende Solitonen zwar durchlaufen, danach aber wieder ihre ursprüngliche Form annehmen. Eine genauere Beschreibung sowie Gleichungen, unter deren Lösungen Solitonen zu finden sind, bietet [SCM73]. Mathematisch wird ein Soliton meist durch den Sekans Hyperbolicus beschrieben, siehe beispielsweise auch [EGCB73] und [Tom80]. In [Tom80] wird dabei ein Solitonbeispiel für eine nichtlineare Wellengleichung in einem Medium bestehend aus zwei dielektrischen Stoffen betrachtet und der Brechungsindex einer der Stoffe hängt nichtlinear vom elektrischen Feld ab, so dass der optische Kerr-Effekt auftritt. Im Grunde handelt es sich also um einen photonischen Kristall, der eine Solitonausbreitung begünstigt.

Man findet viele Anwendungen des Kerr-Effektes. In [LMY+13] werden beispielsweise Solitonen in photonischen Kristallfasern frequenzmoduliert. In [JYY+12] wird der Kerr-Effekt in Flüssigkristallen zur schnellen Messung von Elektroden verwendet und in [AAEH13] benutzt man ihn in einer Silizium-Nanokavität (auch ein photonischer Kristall) zur Herstellung elektronischer Bauteile. Weitere Anwendungen finden sich beispielsweise in [AMK05, Kapitel 10].

1.2 Mathematische Beschreibung

Mathematisch wird der Kerr-Effekt durch einen nichtlinearen Brechungsindex n beschrieben, genauer handelt es sich um einen quadratischen Einfluss des elektrischen Feldes $\mathcal{E}$,

$$n(\mathcal{E}) = n_0 + \chi^{(3)}|\mathcal{E}|^2,$$

wobei $n_0 \geq 1$ und $\chi^{(3)}$ die Suszeptibilität dritter Ordnung ist. Diese Beziehung verwendet man selten direkt in den Maxwellgleichungen, die die Ausbreitung elektromagnetischer Wellen in Materie beschreiben. Wir werden darauf in Kapitel 3 zu sprechen kommen. Gewöhnlich findet man die Beschreibung von Wellen unter dem Einfluss des Kerr-Effektes in Form einer nichtlinearen Schrödingergleichung,

$$\mathrm{i}\partial_t u + \Delta u + \chi^{(3)}|u|^2 u = 0.$$

Eine Rechtfertigung dieser Gleichung findet man in [DLP+11, Chapter 5] und [NM92], für eine Diskussion der auftretenden Effekte siehe auch [TN83, Section 3.7].

Dr. T. Dohnal[1] machte mich auf die folgende Fragestellung aufmerksam: Kann man Solitonen nicht auch direkt über die Maxwellgleichungen simulieren, ohne den Umweg über die Schrödingergleichung zu gehen? Wenn man das versucht, dann erhält man die nichtlineare Wellengleichung

$$\partial_t^2(u + f(u)) = \Delta u, \tag{1.1}$$

wobei $f(u) = \chi^{(3)}|u|^2 u$ ist. Zu dieser nichtlinearen Wellengleichung haben wir keine Theorie gefunden, die Existenz und Eindeutigkeit einer Lösung liefern würde, ebensowenig waren uns numerische Verfahren bekannt. Es bietet sich erst einmal an, die linke Seite auszudifferenzieren, dann erhält man

$$\partial_t^2 u + f'(u)\partial_t^2 u + f''(u)(\partial_t u)^2 = \Delta u.$$

Auch diese Art der Gleichung passt in keine der bekannten Klassen, das liegt an dem Term $(\partial_t u)^2$ auf der linken Seite.

Eine numerische Behandlung dieser Gleichung ist schwierig. Anders als beispielsweise bei der nichtlinearen Schrödingergleichung liegt für (1.1) keine einfache Energieerhaltung vor. Dies erschwert es, Kriterien zur Konstruktion eines numerischen Verfahrens aufzustellen, aus gleichem Grunde sind Konvergenzbeweise eines solchen Verfahren nicht direkt ersichtlich.

1.3 Feine Strukturen

Ein zusätzliches Problem bei der Simulation von Wellenausbreitungen entsteht, wenn das Medium, in dem sich die Welle befindet, an lokalen Stellen einer höheren Auflösung bedarf als im restlichen Medium. Das kann beispielsweise der Fall sein, wenn in ein Medium sehr kleine Strukturen eingefügt wurden oder wenn ein Wellenleiter an einer Stelle verjüngt ist. In diesem Fall muss man lokal das Ortsgitter sehr fein wählen, allerdings ist es dann aber auch zu empfehlen, in der Zeit höher aufzulösen, damit der Einfluss der örtlich verfeinerten Stelle auch zur Geltung kommt. Bei expliziten Verfahren kommt hinzu, dass die Stabilität eine kleine, durch die kleinste Ortsgitterweite bestimmte Zeitschrittweite erzwingt. Für eine derartige Flexibilität benötigt man ein Verfahren, das einen ortsabhängigen Zeitschritt zulässt, wie es in [DG09] für die Wellengleichung und in [GM10] für die Maxwellgleichungen gemacht wurde. Die Zeitschrittverfahren sind jeweils explizit, die Frage ist, ob man dieses Ziel nicht auch mit einem stetigen Galerkinverfahren in der Zeit bewerkstelligen kann.

[1] Aktuell ist Tomáš Dohnal Juniorprofessor an der Universität Dortmund.

1.4 Ziele der Arbeit

Wir verfolgen mit dieser Arbeit drei Ziele.

1. Wir möchten das Modellproblem (1.1) besser verstehen.
2. Wir möchten numerische Verfahren mittels der Finite-Elemente-Methode für die nichtlineare Wellengleichung definieren und Konvergenz zeigen.
3. Für die Problematik feiner Strukturen möchten wir ein lokales Zeitschrittverfahren definieren.

Um das erste Ziel zu erreichen, werden wir aus der nichtlinearen Wellengleichung (1.1) mehrere unterschiedliche Problemformulierungen herleiten und Stabilitätseigenschaften besprechen. Das führt dann zu einem Existenz- und Eindeutigkeitsresultat.

Der zweite Punkt ist das Hauptziel dieser Arbeit. Wir möchten die nichtlineare Wellengleichung mit stetigen Finiten Elementen in Zeit und Ort diskretisieren und zumindest für ein Verfahren die Konvergenz zeigen. Wir werden zwar nicht in der Lage sein, die vermutlich optimale Konvergenzergebnisse in der Theorie zu zeigen, gehen aber darauf ein, wo der Verlust der Konvergenzordnung herkommt.

Unser drittes Ziel umfasst die Klärung der Frage, wie man die Idee eines lokalen Zeitschrittverfahrens auf Finite-Elemente-Diskretisierungen anwenden kann und wie sich das dann strukturell in die Implementierung eines der zuvor vorgestellten Verfahren einfügt.

Wir möchten hervorheben, dass diese Arbeit nicht das Ziel verfolgt, möglichst effiziente Verfahren zu konstruieren. Über Effizienz machen mir uns nur ganz am Rande Gedanken, hier geht es eher um die Frage der Machbarkeit.

1.5 Aufbau der Arbeit

Diese Arbeit ist folgendermaßen aufgebaut. In Kapitel 2 stellen wir die in der ganzen Arbeit geltenden Notationen und Konventionen vor und wiederholen wichtige, wohlbekannte Abschätzungen.

In Kapitel 3 leiten wir aus den Maxwellgleichungen das Modellproblem her, die nichtlineare Wellengleichung (1.1). Mit dieser Wellengleichung werden wir uns in dieser Arbeit beschäftigen. In diesem Kapitel beschreiben wir darüber hinaus mehrere verschiedene Formulierungen des Modellproblems und geben ein Resultat an, das Existenz und Eindeutigkeit einer Lösung liefert.

Danach, in Kapitel 4, leiten wir in die Numerik ein und stellen Finite Elemente Methoden (FEM) und ihre Begrifflichkeiten vor. Wir besprechen die

Anwendung auf elliptische Differentialgleichungen, aber auch auf zeitabhängige Probleme und erläutern die Ideen, die hinter Konvergenzbeweisen stehen.

Dann folgt der Hauptteil der Arbeit in Kapitel 5. In diesem Kapitel definieren wir zwei stetige Finite-Elemente-Verfahren zur Approximation der Lösungen der nichtlinearen Wellengleichung. Für das erste Verfahren zeigen wir außerdem den Konvergenzsatz 5.32. Wir besprechen für beide Verfahren die Implementierung und zeigen in numerischen Experimenten, dass das theoretische Resultat in mehreren Beispielproblemen bestätigt wird.

Schließlich untersuchen wir in Kapitel 6 den Aspekt des lokalen Zeitschrittes für stetige Galerkinverfahren und wenden das Verfahren auf die nichtlineare Wellengleichung an. Numerische Experimente zeigen dann, wie gut sich das lokale Zeitschrittverfahren bei linearen und nichtlinearen Problemen verhält.

Vorbereitungen 2

In diesem Kapitel geben wir einen Überblick über die Notationen, Konventionen und wichtigsten allgemeinen Resultate, die in der Arbeit verwendet werden.

In Abschnitt 2.1 führen wir die grundlegenden Notationen ein und treffen Konventionen. Sobolevräume und ihre Normen besprechen wir in Abschnitt 2.2. Zuletzt geben wir in 2.3 einen Überblick über wichtige Abschätzungen.

2.1 Notationen und Konventionen

Die Menge der natürlichen Zahlen $\mathbb{N}$ beinhaltet in dieser Arbeit nicht die 0, wir definieren durch $\mathbb{N}_0$ die Menge $\mathbb{N} \cup \{0\}$. $\mathbb{R}$ sei die Menge der reellen Zahlen und für ein $p \in \mathbb{N}_0$ und eine Menge M sei $\mathbb{P}_p(M)$ die Menge der Polynome vom maximalen Grad p auf M.

Wir bezeichnen gewöhnlich mit Großbuchstaben Vektoren – dies tun wir aber nur wenn es sich um Koeffizientenvektoren in den numerischen Verfahren handelt – und mit fettgeschriebenen Großbuchstaben Matrizen. Für einen Vektor U sei U_i die i-te Komponente von U für $i \in \mathbb{N}$ und für eine Matrix $\mathbf{A}$ sei $\mathbf{A}_{ij}$ die Komponente in der i-ten Zeile und der j-ten Spalte für $i, j \in \mathbb{N}$. Ab und zu verwenden wir für Vektoren U, V die aus MATLAB® [1] bekannte Schreibweise $[U; V]$ für einen Vektor, der durch das Hintereinanderhängen von U und V entsteht. Die Komponenten von $x \in \mathbb{R}^3$ bezeichnen wir mit x_i, $i = 1, 2, 3$.

Wenn eine Aussage für alle Elemente e aus einer Menge E gilt oder gelten soll, dann schreiben wir das meist hinter die Aussage in Klammern in der Form $(e \in E)$.

Es sei ∇ der Gradient, $\nabla\cdot$ die Divergenz und $\nabla\times$ die Rotation, sowie Δ der Laplace-Operator. Wir benutzen für partielle Ableitungen die abkürzen-

[1] MATLAB® ist eine eingetragene Marke von *The MathWorks, Inc.*

den Schreibweisen

$$\partial_t = \frac{\partial}{\partial t}, \qquad \partial_t^2 = \frac{\partial^2}{\partial t^2}, \qquad \partial_{x_i} = \frac{\partial}{\partial x_i}, \qquad \partial_{x_i}^2 = \frac{\partial^2}{\partial x_i^2} \qquad (i = 1, 2, 3).$$

Ferner sei $\Delta_{(x_1,x_2)} := \partial_{x_1}^2 + \partial_{x_2}^2$.

Ω bezeichnet immer ein beschränktes Gebiet in $\mathbb{R}^d$ für $d = 1, 2, 3$ und $\partial\Omega$ den Rand von Ω. $\overline{\Omega}$ sei der Abschluss von Ω.

Für eine Menge M und $N \subset M$ sei $\mathbb{1}_N$ die *charakteristische Funktion von N*, also

$$\mathbb{1}_N(x) = \begin{cases} 0, & x \notin N, \\ 1, & x \in N. \end{cases}$$

2.2 Sobolevräume

Es sei G ein beschränktes Gebiet, in unserem Fall ist $G = \Omega$ oder ein Zeitintervall. Wir bezeichnen mit $L^p(G)$ für $p \in [1, \infty]$ die bekannten Räume der p-fach Lebesgue-integrierbaren Funktionen und mit $H^s(G)$ für $s \in \mathbb{N}_0$ die Sobolevräume bezüglich $L^2(G)$, wobei $H^0(G) := L^2(G)$. Ferner sei $H_0^1(G)$ der Sobolev-Raum $H^1(G)$ mit verschwindender Spur. Genaueres findet man in [AF03]).

Die Normen der L^p-Räume seien gegeben durch

$$\|u\|_{L^p(G)} = \left(\int_G |u|^p \, \mathrm{d}x \right)^{1/p} \qquad \text{für } p \in [1, \infty),$$
$$\|u\|_{L^\infty(G)} = \operatorname{esssup}_{x \in G} |u(x)|$$

und für die Sobolevräume verwenden wir

$$\|u\|_{H^s(G)}^2 = \sum_{i=0}^{s} \left\| \nabla^i u \right\|_{L^2(G)}^2 .$$

Da wir die L^2-Norm und das zugehörige Skalarprodukt häufig benötigen, schreiben wir kurz $\|\cdot\|_2$ für die L^2-Norm auf G und $(\cdot, \cdot)$ für das L^2-Skalarprodukt über G. Im Falle von $G = \Omega$ werden wir in Beweisen die Ω-Abhängigkeit der Normen weglassen, wenn keine Verwechslungsgefahr besteht.

Weiterhin benötigen wir Sobolev-Räume für Funktionen $u \colon [0, T] \to X$ für einen reellen Banachraum X. Wir verwenden die Räume $L^\infty(0, T; X)$ und $L^2(0, T; X)$ sowie die Räume $H^s(0, T; X)$ für $s \in \mathbb{N}$, wobei die letzteren Räume über verallgemeinerte (schwache) Ableitungen in der Zeit definiert seien,

siehe beispielsweise [Zei90, Section 23.5] und [Eva98, Section 5.9.2]. Als Normen verwenden wir

$$\begin{aligned}
\|u\|_{L^\infty(0,T;H^s(\Omega))} &= \operatorname{ess\,sup}_{\tau\in[0,T]} \|u(\tau,\cdot)\|_{H^s(\Omega)}\,,\\
\|u\|^2_{L^2(0,T;H^s(\Omega))} &= \int_0^T \|u(\tau,\cdot)\|^2_{H^s(\Omega)}\ \mathrm{d}\tau,\\
\|u\|^2_{H^m(0,T;H^s(\Omega))} &= \sum_{i=0}^{m} \left\|u^{(i)}\right\|^2_{L^2(0,T;H^s(\Omega))},
\end{aligned}$$

wobei $u^{(i)}$ die i-te verallgemeinerte Ableitung von u bezüglich der Zeitvariablen sei.

Wir werden nur die folgende, einfachste Version der Sobolev-Einbettung in dieser Arbeit verwenden, siehe [AF03, Lemma 5.17].

2.1 Lemma. (Sobolev-Einbettung)
Es sei $\Omega \subset \mathbb{R}$. Dann ist die Einbettung $H^1(\Omega) \hookrightarrow C(\overline{\Omega})$ stetig. Es existiert also eine Konstante $C_S > 0$ mit

$$\|u\|_{L^\infty(\Omega)} \le C_S \|u\|_{H^1(\Omega)} \tag{2.1}$$

für alle $u \in H^1(\Omega)$.

Ferner benötigen wir die Poincaré-Ungleichung, siehe [Eva98, Section 5.6, Theorem 3].

2.2 Lemma. (Poincaré-Ungleichung)
Es sei $\Omega \subset \mathbb{R}^d$ und $u \in H^1_0(\Omega)$, dann existiert eine Konstante $C = C(\Omega) > 0$ mit

$$\|u\|_2 \le C \|\nabla u\|_2\,. \tag{2.2}$$

Dieses Lemma sagt aus, dass die H^1-Norm und die H^1-Halbnorm auf $H^1_0(\Omega)$ äquivalent sind; es existieren also Konstanten $C_1, C_2 > 0$, sodass für alle $u \in H^1_0(\Omega)$ die Ungleichung

$$C_1 \|\nabla u\|_2 \le \|u\|_2 \le C_2 \|\nabla u\|_2\,.$$

Daher machen wir häufiger keinen Unterschied zwischen diesen beiden Normen.

2.3 Wichtige Abschätzungen

Wir wiederholen jetzt noch drei grundlegende Abschätzungen, die wir in Abschnitt 5.2.4 ständig benutzen werden. Allerdings werden wir ihre Anwendung meist nicht explizit erwähnen.

2.3 Proposition. (Cauchy-Schwarz)
Es seien $u, v \in L^2(\Omega)$, dann gilt $uv \in L^1(\Omega)$ und

$$\|uv\|_{L^1(\Omega)} \leq \|u\|_2 \|v\|_2 .$$

Als Youngsche Ungleichung bezeichnen wir folgende skalierte Ungleichung, die sich einfach mit der binomischen Formel zeigen lässt.

2.4 Proposition. (Youngsche Ungleichung)
Es seien $a, b \geq 0$, dann gilt für alle $\varepsilon > 0$

$$ab \leq \frac{\varepsilon}{2} a^2 + \frac{1}{2\varepsilon} b^2.$$

Für die Beweise der letzten beiden Abschätzungen siehe [Eva98, Appendix B].

Das Gronwall-Lemma benötigen wir für die Fehlerabschätzung eines numerischen Verfahrens, das Lemma findet man in [Pla10, Lemma 8.14]

2.5 Lemma. (Diskrete Gronwall-Ungleichung)
Es seien $(a_n)_{n\in\mathbb{N}_0}, (b_n)_{n\in\mathbb{N}_0}$ nicht-negative Folgen und es existiere ein $\xi > 0$ mit $a_0 \leq \xi$ und

$$a_N \leq \xi + \sum_{n=0}^{N-1} b_n a_n \qquad (N \in \mathbb{N}).$$

Dann lässt sich a_N abschätzen durch

$$a_N \leq \xi \exp\left(\sum_{n=0}^{N-1} b_n \right) \qquad (N \in \mathbb{N}).$$

3 Mathematische Modellierung und Analysis

In diesem Kapitel leiten wir aus den Maxwellgleichungen unser Modellproblem, eine nichtlineare Wellengleichung, her. Diese Wellengleichung schreiben wir auf mehrere Weisen in ein System erster Ordnung um und zitieren, wo wir etwas über Existenz und Eindeutigkeit von Lösungen sagen können.

Wir beginnen in Abschnitt 3.1 mit einer kurzen Vorstellung der Maxwellgleichungen mit den für diese Arbeit wichtigen Substitutionsbedingungen und verwenden einen speziellen Lösungsansatz, um in Abschnitt 3.2 die nichtlineare Wellengleichung herzuleiten, auf der diese Arbeit basiert. In Abschnitt 3.3 formen wir die Wellengleichung in eine quasilineare Differentialgleichung erster Ordnung um und stellen ein Resultat vor, welches eine eindeutige Lösung dieses Problems liefert. Schließlich schreiben wir die Wellengleichung in Abschnitt 3.4 noch in verschiedene semilineare Differentialgleichungen erster Ordnung um. Eines dieser semilinearen Probleme bildet die Grundlage sowohl für eines der Galerkin-Verfahren, die wir später definieren, als auch für das lokale Zeitschrittverfahren in Kapitel 6. Weiterhin werden wir erläutern, warum ein Existenz- und Eindeutigkeitsbeweis wie in Abschnitt 3.3.2 hier nicht funktioniert.

3.1 Die Maxwellgleichungen

Die Darstellung dieses Abschnittes basiert auf [Jac75].

In der klassischen Elektrodynamik werden elektromagnetische Wellen beschrieben durch die Vektorfelder $\mathcal{E}, \mathcal{D}, \mathcal{H}, \mathcal{B} \colon \mathbb{R} \times \mathbb{R}^3 \to \mathbb{C}^3$, die Lösungen der *Maxwellgleichungen* sind. Diese Vektorfelder sind Funktionen in den Variablen $(t, x) \in \mathbb{R} \times \mathbb{R}^3$, wobei t einen Zeitpunkt darstellt und x einen Punkt im Raum.

$\mathcal{E}$ und $\mathcal{H}$ sind das *elektrische* und das *magnetische Feld*, $\mathcal{D}$ und $\mathcal{B}$ sind die *elektrische* und die *magnetische Flussdichte*.

Die *makroskopischen Maxwellgleichungen* in SI-Einheiten lauten

$$\nabla \times \mathcal{E} + \frac{\partial \mathcal{B}}{\partial t} = 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}^3, \tag{3.1a}$$

$$\nabla \cdot \mathcal{B} = 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}^3, \tag{3.1b}$$

$$\nabla \times \mathcal{H} - \frac{\partial \mathcal{D}}{\partial t} = \mathcal{J} \qquad \text{in } \mathbb{R} \times \mathbb{R}^3, \tag{3.1c}$$

$$\nabla \cdot \mathcal{D} = \rho \qquad \text{in } \mathbb{R} \times \mathbb{R}^3. \tag{3.1d}$$

Dabei ist $\mathcal{J} \colon \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ die *elektrische Stromdichte* und $\rho \colon \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}$ die *elektrische Ladungsdichte*. Wir betrachten nur Medien, in denen weder freie Ströme noch freie Ladungen existieren, es ist also

$$\mathcal{J} = 0, \qquad \rho = 0.$$

Weiterhin benötigen wir Bedingungen, die jeweils $\mathcal{E}$ und $\mathcal{D}$ sowie $\mathcal{H}$ und $\mathcal{B}$ miteinander in Beziehung setzen. Im Vakuum gilt

$$\mathcal{D} = \varepsilon_0 \mathcal{E}, \quad \mathcal{H} = \mu_0 \mathcal{B}$$

mit der *Permittivität* ε_0 und der *Permeabilität* μ_0, die durch

$$\varepsilon_0 \mu_0 = \frac{1}{c_0^2}$$

zusammenhängen, wobei c_0 die *Lichtgeschwindigkeit im Vakuum* bezeichnet. In einem Medium dagegen führt das elektrische Feld $\mathcal{E}$ zu einer Verschiebung der Ladungen und damit zu einer *Polarisation* $\mathcal{P}$ der Atome. Daher gilt die Beziehung

$$\mathcal{D} = \varepsilon_0 \mathcal{E} + \mathcal{P}(\mathcal{E}).$$

Für lineare Medien lässt sich die Polarisation angeben durch

$$\mathcal{P}(\mathcal{E}) = \varepsilon_0 \chi^{(1)} \mathcal{E},$$

wobei $\chi^{(1)}$ die *elektrische Suszeptibilität* bezeichnet. Wir nehmen an, dass es sich um ein *isotropes* Medium handelt, also $\chi^{(1)}$ eine skalare Funktion ist, die vom Ort abhängt. Weiterhin haben wir implizit angenommen, dass die Suszeptibilität nicht von der Frequenz der Welle abhängt, das Medium also *nicht-dispersiv* ist. Mathematische Aspekte der frequenzabhängigen Maxwellgleichungen werden beispielsweise in [Sch13] behandelt. Wir schreiben jetzt

$$\mathcal{D} = \varepsilon_0 \varepsilon_r \mathcal{E}$$

mit der *relativen Permittivität* $\varepsilon_r = 1 + \chi^{(1)}$. In nichtlinearen Medien treten noch zusätzliche Effekte auf, die man durch Hinzufügen von Termen höherer Ordnung beschreibt (siehe [Fox12, Kapitel 11]). Wir beachten nichtlineare Effekte dritter Ordnung und nehmen an, dass wir ein inversionssymmetrisches Medium haben, weil in diesem Fall keine quadratischen Nichtlinearitäten auftreten. Dann erhalten wir die Beziehung

$$\mathcal{P}(\mathcal{E}) = \varepsilon_0(\chi^{(1)}\mathcal{E} + \chi^{(3)}|\mathcal{E}|^2\mathcal{E}).$$

$\chi^{(3)}$, die *nichtlineare Suszeptibilität dritter Ordnung*, ist ebenso wie $\chi^{(1)}$ eine skalare Funktion, ein Medium mit diesen Eigenschaften nennt man *isotrop*.Es ist also

$$\mathcal{D} = \varepsilon_0 \left(\varepsilon_r + \chi^{(3)}|\mathcal{E}|^2\right) \mathcal{E}.$$

Weiterhin skalieren wir die Felder $\mathcal{E}, \mathcal{B}$, sodass wir $\varepsilon_0 = \mu_0 = 1$ setzen können.

Insgesamt erhalten wir die Gleichungen

$$\nabla \times \mathcal{E} + \frac{\partial \mathcal{B}}{\partial t} = 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}^3, \tag{3.2a}$$

$$\nabla \cdot \mathcal{B} = 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}^3, \tag{3.2b}$$

$$\nabla \times \mathcal{B} - \frac{\partial(\varepsilon_r\mathcal{E} + \chi^{(3)}|\mathcal{E}|^2\mathcal{E})}{\partial t} = 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}^3, \tag{3.2c}$$

$$\nabla \cdot (\varepsilon_r\mathcal{E} + \chi^{(3)}|\mathcal{E}|^2\mathcal{E}) = 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}^3. \tag{3.2d}$$

3.2 Eine nichtlineare Wellengleichung

Um auf unser Modellproblem zu kommen, verwenden wir einen speziellen Ansatz in den Maxwellgleichungen (3.2). Zunächst wenden wir den Operator $\nabla\times$ auf (3.2a) an und leiten (3.2c) nach der Zeit ab. Zusammen folgt dann für $\mathcal{E}$ die Gleichung

$$\nabla \times (\nabla \times \mathcal{E}) + \varepsilon_r \frac{\partial^2 \mathcal{E}}{\partial t^2} + \chi^{(3)} \frac{\partial^2(|\mathcal{E}|^2\mathcal{E})}{\partial t^2} = 0. \tag{3.3}$$

Außerdem muss Gl. (3.2d) erfüllt sein.

Wir betrachten ein Medium, dass sich in x_3-Richtung nicht ändert. Trifft eine elektromagnetische Welle in der x_1-x_2-Ebene auf dieses Medium (das heißt, der Wellenvektor liegt in der x_1-x_2-Ebene), so lassen sich die Maxwellgleichung aufspalten in zwei unabhängige Probleme, die man $\mathcal{E}$-Polarisation und $\mathcal{H}$-Polarisation oder TE-Welle und TM-Welle nennt (eine genaue Erklärung findet man in [JJWM11, Chapter 3] und mit umgekehrter Benennung

der TE- und TM-Welle in [Jac75, Chapter 8]). Das elektrische Feld der $\mathcal{E}$-Polarisation lässt sich durch den Ansatz

$$\mathcal{E}(x_1, x_2) = \begin{pmatrix} 0 \\ 0 \\ u(x_1, x_2) \end{pmatrix} \tag{3.4}$$

beschreiben, u ist dann die x_3-Komponente des elektrischen Feldes. Da sich das Medium nur in x_1- und x_2-Richtung verändert, gilt für die Materialparameter $\chi^{(1)} = \chi^{(1)}(x_1, x_2)$ und $\chi^{(3)} = \chi^{(3)}(x_1, x_2)$. Das führt außerdem dazu, dass die Bedingung (3.2d) erfüllt ist, denn

$$\begin{aligned} &\nabla\cdot(\varepsilon_r\mathcal{E} + \chi^{(3)}|\mathcal{E}|^2\mathcal{E}) \\ &\quad = \partial_{x_3}((1 + \chi^{(1)}(x_1, x_2))u(x_1, x_2) + \chi^{(3)}(x_1, x_2)|u(x_1, x_2)|^2 u(x_1, x_2)) \\ &\quad = 0. \end{aligned}$$

Weiterhin gilt

$$\nabla \times (\nabla \times \mathcal{E}) = (\nabla(\nabla\cdot\mathcal{E}) - \Delta\mathcal{E}),$$

und mit (3.4) ist

$$\nabla\cdot\mathcal{E}(x_1, x_2) = \partial_{x_3} u(x_1, x_2) = 0,$$

also gilt

$$\nabla \times (\nabla \times \mathcal{E}) = -\Delta u = -\Delta_{(x_1,x_2)} u.$$

Daher führt der Ansatz (3.4) zu der nichtlinearen Wellengleichung

$$\partial_t^2(\varepsilon_r u + \chi^{(3)}|u|^2 u) - \Delta u = 0. \tag{3.5}$$

In dieser Arbeit machen wir die folgenden zusätzlichen Annahmen und Änderungen der Notation.

1. $\chi^{(1)}$ sei Null, also ist $\varepsilon_r = 1$.
2. $\chi^{(3)}$ sei eine konstante Funktion, wir nennen sie ab jetzt λ und schreiben für den nichtlinearen Term

   $$f(u) := \lambda|u|^2 u.$$

3. Die obige Herleitung kann man auch mit $u(x_1, x_2) = u(x_1)$ durchführen, physikalisch entspricht das dann einem Medium, dass nur in der x_1-Richtung inhomogen ist. Wir können die Gleichung (3.5) also auch in $\mathbb{R}$ betrachten.

4. Wir fügen auf der rechten Seite noch eine Funktion g hinzu.

Damit gelangen wir zu unserem Modellproblem

$$\partial_t^2(u + f(u)) = \Delta u + g. \tag{3.6}$$

Wir betrachten diese Gleichung für $t \geq 0$ auf einem beschränkten Gebiet $\Omega \subset \mathbb{R}^d$ für $d = 1, 2, 3$. Daher benötigen wir die Anfangsbedingungen

$$\begin{aligned} u(0, \cdot) &= u_0, \\ \partial_t u(0, \cdot) &= v_0, \end{aligned}$$

wobei $u_0, v_0 \colon \Omega \to \mathbb{C}$, und auf $\Gamma = \partial\Omega$ wählen wir die homogene Dirichlet-Randbedingung

$$u(t, \cdot) = 0, \quad t \geq 0.$$

Für komplexwertige Funktionen u ist die Differentialgleichung (3.6) zunächst nicht definiert, weil f nur in 0 komplex differenzierbar ist. Da f aber unendlich oft reell differenzierbar ist, können wir der Differentialgleichung durch die Identifizierung von $\mathbb{C}$ mit $\mathbb{R}^2$ einen Sinn geben. Wir werden diese nichtlineare Wellengleichung nicht in dieser Form verwenden, sondern sie in ein System erster Ordnung umschreiben. Zur Vorbereitung rechnen wir die Zeitableitung auf der linken Seite formal aus.

Für reellwertige Funktionen folgt im ersten Schritt

$$\partial_t((1 + f'(u))\partial_t u) = \Delta u + g \tag{3.7}$$

und schließlich

$$(1 + f'(u))\partial_t^2 u + f''(u)(\partial_t u)^2 = \Delta u + g. \tag{3.8}$$

Für komplexwertige Funktionen setzen wir

$$\begin{aligned} u_{\mathrm{re}} &:= \mathrm{Re}(u), \qquad u_{\mathrm{im}} := \mathrm{Im}(u), \\ g_{\mathrm{re}} &:= \mathrm{Re}(g), \qquad g_{\mathrm{im}} := \mathrm{Im}(g). \end{aligned}$$

Diese Indizierung verwenden wir auch für andere Funktionen. Fasse f als Funktion $f \colon \mathbb{R}^2 \to \mathbb{R}^2$ auf, dann ist

$$f(u_{\mathrm{re}}, u_{\mathrm{im}}) = \lambda(u_{\mathrm{re}}^2 + u_{\mathrm{im}}^2) \begin{pmatrix} u_{\mathrm{re}} \\ u_{\mathrm{im}} \end{pmatrix}.$$

In diesem Fall lautet (3.7)

$$\partial_t(1 + f'(u)[\partial_t u]) = \Delta u + g, \tag{3.9}$$

wobei 1 hier die konstante Funktion sei, die in beiden Komponenten den Wert 1 annimmt. Der Laplace-Operator wird komponentenweise angewendet und es gilt

$$\begin{aligned} f'(u)[v] &= \lambda \begin{pmatrix} (3u_{\mathrm{re}}^2 + u_{\mathrm{im}}^2)v_{\mathrm{re}} + 2u_{\mathrm{re}}u_{\mathrm{im}}v_{\mathrm{im}} \\ 2u_{\mathrm{re}}u_{\mathrm{im}}v_{\mathrm{re}} + (u_{\mathrm{re}}^2 + 3u_{\mathrm{im}}^2)v_{\mathrm{im}} \end{pmatrix} \\ &= \lambda \begin{pmatrix} 3u_{\mathrm{re}}^2 + u_{\mathrm{im}} & 2u_{\mathrm{re}}u_{\mathrm{im}} \\ 2u_{\mathrm{re}}u_{\mathrm{im}} & u_{\mathrm{re}}^2 + 3u_{\mathrm{im}}^2 \end{pmatrix} \begin{pmatrix} v_{\mathrm{re}} \\ v_{\mathrm{im}} \end{pmatrix}, \end{aligned}$$

Die Gleichung (3.8) lautet im komplexwertigen Fall

$$\partial_t^2(u + f(u)) = \partial_t^2 u + f'(u)[\partial_t^2 u] + f''(u)[\partial_t u, \partial_t u] = \Delta u + g, \tag{3.10}$$

wobei

$$f''(u)[v, w] = \lambda \begin{pmatrix} 6u_{\mathrm{re}}v_{\mathrm{re}}w_{\mathrm{re}} + 4u_{\mathrm{im}}v_{\mathrm{re}}w_{\mathrm{im}} + 2u_{\mathrm{re}}v_{\mathrm{im}}w_{\mathrm{im}} \\ 2u_{\mathrm{im}}v_{\mathrm{re}}w_{\mathrm{re}} + 4u_{\mathrm{re}}v_{\mathrm{im}}w_{\mathrm{re}} + 6u_{\mathrm{im}}v_{\mathrm{im}}w_{\mathrm{im}} \end{pmatrix}.$$

In der Umformung von (3.6) in ein System erster Ordnung beschränken wir uns auf den reellwertigen Fall. Dafür reicht es, reellwertige Anfangswerte vorzuschreiben. Den komplexwertigen Fall werden wir für die numerischen Verfahren wieder aufgreifen, siehe Abschnitt 5.2.5.2 und Abschnitt 5.3.2.2. Existenz und Eindeutigkeit einer Lösung von (3.6) lässt sich unter gewissen Voraussetzungen zeigen, wir werden einen Satz mit einem entsprechenden Resultat im nachfolgenden Abschnitt 3.3.2 angeben.

Wir unterscheiden im Nachfolgenden zwei verschiedene Arten der Umformung von (3.6) in ein System erster Ordnung. Zuerst betrachten wir die Formulierungen, in denen die Ortsableitungen einen von u abhängigen Koeffizienten besitzen, das führt auf quasilineare Wellengleichungen. Danach schauen wir uns Formulierungen an, in denen die nichtlinearen Terme von den Differentialoperatoren getrennt sind, das sind dann semilineare Wellengleichungen.

3.3 Quasilineare Wellengleichung

Wir können (3.6) beispielsweise für den Fall $g = 0$ wie folgt in ein System umschreiben,

$$\partial_t(u + f(u)) = \nabla \cdot \mathbf{q}, \tag{3.11a}$$

$$\partial_t \mathbf{q} = \nabla u. \tag{3.11b}$$

Wenn die Gleichung $v = u + f(u)$ invertierbar ist, also $u = \Phi(v)$, und Φ differenzierbar ist, dann erhalten wir daraus die quasilineare Gleichung

$$\begin{aligned} \partial_t v &= \nabla \cdot \mathbf{q}, \\ \partial_t \mathbf{q} &= \Phi'(v)\nabla v. \end{aligned}$$

Das ähnelt dann in der Form den Maxwellgleichungen (3.2) (man denke sich die Divergenz und den Gradienten ersetzt durch die Rotation). Diese Formulierung lässt sich unter Umständen als hyperbolisches System schreiben, diesen Weg wollen wir in dieser Arbeit aber nicht gehen.

Nun wenden wir uns einer der Formulierungen zu, die uns in dieser Arbeit interessieren werden. Eine naheliegende Transformation von Gl. (3.6) auf ein System erster Ordnung erfolgt durch $v := \partial_t u$. Setzen wir v in (3.8) ein, so folgt die Gleichung

$$(1 + f'(u))\partial_t v + f''(u)v^2 = \Delta u + g.$$

Wir werden dieses System in zwei äquivalenten Formulierungen benutzen. Eines der numerischen Verfahren basiert auf

$$\partial_t u = v, \tag{3.12a}$$

$$(1 + f'(u))\partial_t v = \Delta u - f''(u)v^2 + g. \tag{3.12b}$$

Wir werden nun voraussetzen, dass

$$1 + f'(u) > 0$$

auf der ganzen Lösungstrajektorie gilt. Wäre dem nicht so, dann würde sich der Typ der Differentialgleichung ändern; das wollen wir vermeiden. Diese Bedingung führt aber für den kritischen Fall $\lambda < 0$ dazu, dass u klein genug bleiben muss. Wir werden mehrfach auf diese Problematik stoßen. Für die Analysis setzen wir $g = 0$ und teilen die zweite Gleichung durch $1 + f'(u)$. Wir erhalten dann das System

$$\begin{aligned}\partial_t \begin{pmatrix} u \\ v \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ \frac{1}{1+f'(u)}\Delta & -\frac{f''(u)v}{1+f'(u)} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ \frac{1}{1+f'(u)}\Delta & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} 0 \\ \frac{f''(u)}{1+f'(u)}v^2 \end{pmatrix}. \end{aligned} \tag{3.13}$$

3.3.1 Stabilität

In diesem Abschnitt werden wir aus der quasilinearen Formulierung zwei *Stabilitätsgleichungen* herleiten. Dazu schreiben wir die Gleichungen (3.12) in variationeller Form. Wir suchen also $u, v \in L^2(0, T; H_0^1(\Omega)) \cap H^1(0, T; L^2(\Omega))$ mit

$$\int_0^T (\partial_t u, \phi_1)\, \mathrm{d}t = \int_0^T (v, \phi_1)\, \mathrm{d}t, \tag{3.14a}$$

$$\begin{aligned}\int_0^T ((1 + f'(u))\partial_t v, \phi_2)\, \mathrm{d}t = &- \int_0^T (\nabla u, \nabla \phi_2) + (f''(u)v^2, \phi_2)\, \mathrm{d}t \\ &+ \int_0^T (g, \phi_2)\, \mathrm{d}t \end{aligned} \tag{3.14b}$$

für alle $\phi_1 \in L^2(0,T;L^2(\Omega))$, $\phi_2 \in L^2(0,T;H_0^1(\Omega))$. Es sei $\mathcal{A}\colon H_0^1(\Omega) \to L^2(\Omega)$ definiert durch

$$(\mathcal{A}w_1, w_2) := (\nabla w_1, \nabla w_2) \qquad (w_2 \in H_0^1(\Omega)).$$

Wir testen (3.14) mit $\phi_1 := \mathcal{A}u$ und $\phi_2 := v$, dabei ist die Anwendung von $\mathcal{A}$ punktweise in der Zeit zu verstehen, und erhalten

$$\begin{aligned}
\frac{1}{2}\int_0^T \partial_t(\|\nabla u\|_2^2)\,\mathrm{d}t &= \int_0^T (\nabla v, \nabla u)\,\mathrm{d}t,\\
\frac{1}{2}\int_0^T \partial_t(\|v\|_2^2)\,\mathrm{d}t + \int_0^T (f'(u)\partial_t v, v)\,\mathrm{d}t &= -\int_0^T (\nabla u, \nabla v)\,\mathrm{d}t\\
&\quad - \int_0^T (f''(u)v^2, v)\,\mathrm{d}t\\
&\quad + \int_0^T (g, v)\,\mathrm{d}t.
\end{aligned}$$

Addieren wir diese beiden Gleichungen, so folgt

$$\begin{aligned}
\frac{1}{2}\int_0^T \partial_t(\|\nabla u\|_2^2 + \|v\|_2^2)\,\mathrm{d}t + \int_0^T (f'(u)\partial_t v + f''(u)v^2, v)\,\mathrm{d}t&\\
= \int_0^T (g, v)\,\mathrm{d}t.& \quad (3.15)
\end{aligned}$$

Es gilt mittels partieller Integration

$$\begin{aligned}
\int_0^T (f'(u)\partial_t v, v)\,\mathrm{d}t &= \int_0^T (f'(u)v, \partial_t v)\,\mathrm{d}t\\
&= [(f'(u)v, v)]_0^T - \int_0^T (f''(u)v^2 + f'(u)\partial_t v, v)\,\mathrm{d}t\\
&= \frac{1}{2}[(f'(u)v, v)]_0^T - \frac{1}{2}\int_0^T (f''(u)v^2, v)\,\mathrm{d}t.
\end{aligned}$$

Also lässt sich der Term mit den Nichtlinearitäten in (3.15) umschreiben zu

$$\begin{aligned}
\int_0^T (f'(u)\partial_t v + f''(u)v^2, v)\,\mathrm{d}t&\\
= \frac{1}{2}\left([(f'(u)v, v)]_0^T + \int_0^T (f''(u)v^2, v)\,\mathrm{d}t\right).&
\end{aligned}$$

Setzen wir das in (3.15) ein, so folgt

$$\begin{aligned}
\int_0^T \partial_t(\|\nabla u\|_2^2 + \|v\|_2^2)\,\mathrm{d}t + [(f'(u)v, v)]_0^T&\\
= -\int_0^T (f''(u)v^2, v)\,\mathrm{d}t + \int_0^T (g, v)\,\mathrm{d}t.& \quad (3.16)
\end{aligned}$$

Die linke Seite kann man als eine Art Energie auffassen und die rechte Seite lässt sich abschätzen. Wir werden später sehen, wie man diese Gleichung verwenden kann.

In $H_0^1(\Omega) \times L^2(\Omega)$ findet man also keine Energieerhaltung, denn dazu müsste die ganze rechte Seite von (3.16) gleich Null sein. Man kann aber eine Energieerhaltung in $L^2(\Omega) \times H^{-1}(\Omega)$ zeigen. Zu diesem Zweck testet man mit $\phi_1 := u$ und $\phi_2 := \mathcal{A}^{-1}((1 + f'(u))v)$ (siehe auch [BL94]) und erhält zum einen

$$\frac{1}{2} \int_0^T \partial_t (\|u\|_2^2) \, \mathrm{d}t = \int_0^T (v, u) \, \mathrm{d}t \tag{3.17}$$

und zum anderen

$$\begin{aligned} &\int_0^T (\mathcal{A}^{-1/2}((1 + f'(u))\partial_t v + f''(u)v^2), \mathcal{A}^{-1/2} v) \\ &\qquad = -\int_0^T (u, (1 + f'(u))v) \, \mathrm{d}t + \int_0^T (\mathcal{A}^{-1/2} g, \mathcal{A}^{-1/2} v) \, \mathrm{d}t. \end{aligned} \tag{3.18}$$

Dabei haben wir ausgenutzt, dass $\mathcal{A}^{-1}$ selbstadjungiert und positiv ist, daher lässt sich die Wurzel des Operators definieren. Unter Beachtung der Gleichungen

$$(1 + f'(u))\partial_t v + f''(u)v^2 = \partial_t((1 + f'(u))v)$$

und

$$\begin{aligned} ((1 + f'(u))v, u) &= (v, u) + (f'(u)u, v) \\ &= (v, u) + 3\lambda(u^3, v) = (v, u) + \frac{3}{4}\lambda \partial_t \|u\|_{L^4}^4 \end{aligned}$$

erhalten wir nach Addition der Gleichungen (3.17) und (3.18) die Gleichung

$$\begin{aligned} &\frac{1}{2} \int_0^T \partial_t \left(\|u\|_2^2 + \left\| \mathcal{A}^{-1/2}((1 + f'(u))v) \right\|_2^2 + \frac{3}{4}\lambda \|u\|_{L^4}^4 \right) \mathrm{d}t \\ &\qquad = \int_0^T (\mathcal{A}^{-1/2} g, \mathcal{A}^{-1/2} v) \, \mathrm{d}t. \end{aligned} \tag{3.19}$$

Für $g = 0$ erhalten wir dann tatsächlich eine Energie

$$E(t) := \|u(t)\|_2^2 + \left\| \mathcal{A}^{-1/2}((1 + f'(u(t)))v(t)) \right\|_2^2 + \frac{3}{4}\lambda \|u(t)\|_{L^4}^4,$$

die erhalten wird, unter der Voraussetzung, dass $E(t) \geq 0$ ist. Wir bemerken, dass $(1 + f'(u))v = \partial_t(u + f(u))$ gilt. Auf dieser Energie wird diese Arbeit aber nicht aufbauen. Denn die nachfolgende Analysis funktioniert dafür nicht und die Numerik würde so undurchsichtig (womit testet man im diskreten Fall?), dass wir darauf keine Zeit verwendet haben.

3.3.2 Ein Existenz- und Eindeutigkeitsresultat

In diesem Abschnitt gehen wir kurz auf einen Existenz- und Eindeutigkeitssatz ein, den wir zusammen mit Prof. R. Schnaubelt für das Problem (3.13) bewiesen haben (siehe [DGS13]).

Um den Satz formulieren zu können, müssen wir einige Voraussetzungen erklären. Wir schreiben (3.13) für $g = 0$ in der Form

$$\partial_t u = v, \tag{3.20a}$$
$$\partial_t v = K_1(u)\Delta u + K_2(u)v^2, \tag{3.20b}$$

wobei wir wieder Nullrandbedingungen voraussetzen. Wir definieren

$$w(t) := (u(t), v(t))$$

und schreiben das System in der Form

$$\frac{\mathrm{d}}{\mathrm{d}t} w(t) = A(w(t))w(t)$$

mit dem nichtlinearen Operator

$$A(w(t)) := \begin{pmatrix} 0 & I \\ K_1(u(t))\Delta & K_2(u(t))v(t) \end{pmatrix}.$$

Wir setzen voraus, dass es $\rho, \delta > 0$ gibt mit

$$K_1, K_2 \in C^2([-\rho, \rho]) \quad \text{und} \quad K_1 \geq \delta. \tag{3.21}$$

Für $\lambda < 0$ bedeutet das, dass wir $\rho < \sqrt{-3\lambda}$ wählen müssen.

Es sei $\Omega \subset \mathbb{R}^d$ für $d = 1, 2, 3$ ein beschränktes Gebiet mit einem C^3-Rand. Wir beschränken uns auf reellwertige Funktionen. Wir betrachten den Laplace-Operator mit Definitionsbereich

$$\mathcal{D}(-\Delta) = H^2(\Omega) \cap H_0^1(\Omega).$$

Wir definieren die Räume

$$\mathcal{H}_0 := L^2(\Omega), \quad \mathcal{H}_k := \mathcal{D}\big((-\Delta)^{k/2}\big)$$

für $k \in \mathbb{N}$ mit den Normen

$$|\varphi|_{\mathcal{H}_0} := \|\varphi\|_2, \quad |\varphi|_{\mathcal{H}_k} := \left\|(-\Delta)^{k/2}\varphi\right\|_2.$$

Weiterhin sei

$$\mathcal{X}_k := \mathcal{H}_{k+1} \times \mathcal{H}_k$$

für $k \in \mathbb{N}_0$ mit den Normen

$$|(u,v)|^2_{\mathcal{X}_k} := |u|^2_{\mathcal{H}_{k+1}} + |v|^2_{\mathcal{H}_k}.$$

Wir wählen ein $r > 0$ fest mit

$$|u|_{\mathcal{H}_2} \leq r \quad \Rightarrow \quad \|u\|_{L^\infty} \leq \rho.$$

Die Konstante C_1 im folgenden Satz tritt im Beweis dieses Satzes auf und hängt mit der Stabilität einer Familie von Operatoren zusammen.

3.1 Satz. (Existenz und Eindeutigkeit)
Wir nehmen an, dass (3.21) *gilt und es sei* $w_0 = (u_0, v_0) \in \mathcal{X}_2$ *mit* $|w_0|_{\mathcal{X}_1} \leq C_1 \eta r$ *für ein* $\eta \in (0,1)$. *Dann existieren ein Zeitpunkt* $T = T(\eta, r, |w_0|_{\mathcal{X}_2}) > 0$ *und eine Funktion* $w \in C^1(0,T;\mathcal{X}_1) \cap C(0,T;\mathcal{X}_2)$ *mit* $|w(t)|_{\mathcal{X}_1} \leq r$ *für* $t \in [0,T]$, *die*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} w(t) &= A(w(t))w(t), \quad t \in [0,T], \\ w(0) &= w_0 \end{aligned} \tag{3.22}$$

erfüllt. Setzen wir $w(t) = (u(t), v(t))$, *dann ist* $v = \partial_t u$ *und*

$$u \in C(0,T;\mathcal{H}_3) \cap C^1(0,T;\mathcal{H}_2) \cap C^2(0,T;\mathcal{H}_1)$$

erfüllt $|(u(t), \partial_t u(t))|_{\mathcal{X}_1} \leq r$ *sowie die Differentialgleichung*

$$\begin{aligned} \partial_t^2 u &= K_1(u)\Delta u + K_2(u)(\partial_t u)^2, \quad t \in [0,T], \\ u(0) &= u_0, \\ \partial_t u(0) &= v_0. \end{aligned} \tag{3.23}$$

Jede andere Lösung von (3.23) *mit den obigen Voraussetzungen auf einem Zeitintervall* $[0,T']$ *stimmt mit u auf* $[0, \min\{T,T'\})]$ *überein.*

Gilt für f aus (3.6) $f \in C^4(\mathbb{R})$ *und* K_1, K_2 *erfüllen* (3.21), *dann gelten die Ergebnisse auch, wenn wir in* (3.23) *die Differentialgleichung durch*

$$\partial_t^2 (u + f(u)) = \Delta u$$

ersetzen.

3.2 Korollar.
Gilt $f(z) = \lambda z^3$ *mit* $\lambda > 0$, *so gilt Satz 3.1 ohne alle Bedingungen, die r beinhalten.*

Der Beweis basiert auf einer Konstruktion von Lösungen quasilinearer Differentialgleichungen von Kato [Kat70] durch die Verwendung der Halbgruppentheorie. Wir werden im Folgenden ein paar Begriffe dieser Theorie

verwenden, ohne sie hier zu definieren. Daher verweisen wir auf das Standardwerk [Paz83]. Im Beweis von Satz 3.1 geht man in drei Schritten vor.

Vorbereitend spalten wir den nichtlinearen Operator A auf und schreiben (3.22) wie folgt,

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = \underbrace{\begin{pmatrix} 0 & \mathrm{Id} \\ K_1(u)\Delta & 0 \end{pmatrix}}_{=:A_0(w)} w + \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & K_2(u)v^2 \end{pmatrix}}_{=:B(w)}.$$

Im ersten Schritt linearisiert man das quasilineare Problem, indem man eine beliebige aber feste Funktion $\widetilde{w} \in \mathcal{X}_2$ mit gewissen Eigenschaften wählt und das autonome Problem

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = (A_0(\widetilde{w}) + B(\widetilde{w}))w(t)$$

betrachtet. Es lässt sich nun zeigen, dass $(A_0(\widetilde{w}), \mathcal{X}_1)$ eine unitäre Halbgruppe auf $\mathcal{X}_0$ erzeugt und dass $B(\widetilde{w})$ eine stetige Störung ist. Daraus können wir folgern, dass $(A(\widetilde{w}), \mathcal{X}_1)$ eine Kontraktionshalbgruppe auf $\mathcal{X}_0$ erzeugt.

Im zweiten Schritt erweitert man dieses Resultat, indem man eine zeitabhängige fixierte Funktion $\widetilde{w}(t)$ zulässt und daher das Problem

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = A(\widetilde{w}(t))w(t) \tag{3.24}$$

betrachtet. Für die Konstruktion eines Lösungsoperators benötigt man die sogenannte Stabilität der Operatorenfamilie $(A(\widetilde{w}(t)))_{t\in[0,T]}$. Das ist eine Eigenschaft, die aus einer Finite-Differenzen-Approximation herrührt, es muss Konstanten $M, \beta > 0$ geben, sodass

$$\left\| \mathrm{e}^{\tau_n A(\widetilde{w}(t_n))} \dots \mathrm{e}^{\tau_1 A(\widetilde{w}(t_1))} \right\|_{\mathcal{B}(\mathcal{X}_1)} \leq M \mathrm{e}^{\beta(\tau_1+\dots+\tau_n)}$$

für alle $\tau_j \geq 0$, $j = 1, \dots, n$ und $0 \leq t_1 \leq \dots \leq t_n \leq T$ mit $n \in \mathbb{N}$ gilt. Diese Eigenschaft direkt auf $\mathcal{X}_1$ zu zeigen ist schwierig. Sie lässt sich aber mit dem ersten Schritt auf $\mathcal{X}_0$ folgern und über eine Isometrie von $\mathcal{X}_0 \to \mathcal{X}_2$ auch auf $\mathcal{X}_2$, sodass ein Interpolationsargument die Stabilität auf $\mathcal{X}_1$ liefert. Damit lässt sich dann ein Lösungsoperator $U_{\widetilde{w}}(t)w_0$ für dieses nichtautonome lineare Problem konstruieren.

Im letzten Schritt sucht man dann eine Lösung von (3.24) mit $\widetilde{w}(t) = w(t)$, man löst also das Fixpunktproblem

$$U_w(t)w_0 = w(t)$$

auf einer gewissen Menge, die die Wohldefiniertheit der Nichtlinearitäten gewährleistet. Man zeigt, dass auf der linken Seite ein kontraktiver Operator steht und kann dann die Existenz einer eindeutigen Lösung garantieren.

3.3 Bemerkung.
Zwei Bedingungen verhindern den Beweis globaler Existenz von Lösungen. Zunächst ist das die Bedingung, dass $|w|_{\mathcal{X}_1}$ für alle t klein bleiben muss, damit durch $1 + f'(u)$ geteilt werden kann. Das Korollar 3.2 sagt gerade aus, dass diese Bedingung wegfällt, wenn $\lambda > 0$ ist. Die zweite Bedingung ist die lokale Lipschitz-Stetigkeit der „rechten Seite". Diese Bedingung muss für beliebige $\lambda \in \mathbb{R}$ erfüllt sein. Hier besteht ein Zusammenhang zwischen $|w|_{\mathcal{X}_1}$ und dem Zeitpunkt T. Je größer $|w|_{\mathcal{X}_1}$ ist, desto kleiner ist T. Es gibt nun Beispiele, in denen man die globale Existenz einer Lösung nicht zeigen kann, weil für $t \to T^-$ die $\mathcal{X}_1$-Norm von w gegen Unendlich geht. Ein Beispiel dafür werden wir in den numerischen Experimenten in Abschnitt 5.4.1 kennenlernen. Man nennt ein solches Verhalten der Lösung einen blow-up.

3.4 Semilineare Wellengleichung

Bevor wir eine semilineare Formulierung der Wellengleichung herleiten, erhalten wir das System

$$v = u + f(u), \tag{3.25a}$$
$$\partial_t v = w, \tag{3.25b}$$
$$\partial_t w = \Delta u + g. \tag{3.25c}$$

Diese Formulierung eignet sich nicht für eine analytische Untersuchung des Problems, ist aber die einfachste Form, um ein numerisches Verfahren für die nichtlineare Wellengleichung (3.6) zu definieren. Wir werden diese Formulierung in Abschnitt 5.3 als Grundlage verwenden. Fassen wir die ersten zwei Gleichungen zusammen, so ist

$$w = (1 + f'(u))\partial_t u.$$

Dann erhalten wir zur Bestimmung von (u, w) das System

$$\begin{aligned}(1 + f'(u))\partial_t u &= w,\\ \partial_t w &= \Delta u + g.\end{aligned}$$

Dieses System sieht angenehmer aus als (3.12), weil sich der nichtlineare Teil nicht mehr in der Gleichung mit dem Laplace-Operator befindet. Indem wir wieder durch $1 + f'(u)$ teilen, erhalten wir die semilineare Formulierung

$$\begin{aligned}\partial_t u &= w + \frac{f'(u)}{1 + f'(u)} w,\\ \partial_t w &= \Delta u + g.\end{aligned}$$

Wir schreiben dieses System in der Form

$$\partial_t \begin{pmatrix} u \\ w \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ \Delta & 0 \end{pmatrix}}_{=:A} \begin{pmatrix} u \\ w \end{pmatrix} + \begin{pmatrix} \frac{f'(u)}{1+f'(u)} w \\ g \end{pmatrix}. \tag{3.26}$$

Das ist nichts anderes als die Wellengleichung mit einem nichtlinearen Kraftterm.

Der Operator A erzeugt eine kontraktive Halbgruppe auf $H_0^1(\Omega) \times L^2(\Omega)$ mit Definitionsbereich

$$\mathcal{D}(A) = (H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega),$$

dies zeigt man mit dem Lemma von Lumer-Phillips, siehe [Eva98, 7.4 Thm. 5]. Daher drängt sich die Erwartung auf, dass man mit bekannter Halbgruppentheorie zumindest die Existenz einer Lösung der Gleichung (3.26) zeigen könnte. Damit das funktionieren kann, benötigt man aber die (lokale) Lipschitzstetigkeit der rechten Seite (diese Bedingung kennt man schon aus der Theorie gewöhnlicher Differentialgleichung, sie taucht im Satz von Picard-Lindelöf auf). Die Funktion

$$\Phi(u, w) := \frac{f'(u)}{1 + f'(u)} w$$

ist aber keine Lipschitz-stetige Funktion von $H_0^1(\Omega) \times L^2(\Omega)$ nach $H_0^1(\Omega)$, sie ist noch nicht einmal lokal Lipschitz-stetig, da w nur in $L^2(\Omega)$ liegt. Wir haben versucht, dieses Problem für $\Omega \subset \mathbb{R}$ durch die Einführung eines gestörten Operators A zu umgehen, genauer sei

$$A_\varepsilon = \begin{pmatrix} 0 & 1 - \varepsilon \partial_x^2 \\ \partial_x^2 & 0 \end{pmatrix}$$

für $\varepsilon > 0$ auf $X^\varepsilon := H_0^1(\Omega) \times H_0^1(\Omega)$ mit Definitionsbereich

$$\mathcal{D}(A^\varepsilon) = \left\{ y \in H^3(\Omega) \cap H_0^1(\Omega) \colon \partial_x^2 y \lfloor_{\partial\Omega} \equiv 0 \right\}^2.$$

Bei geeigneter Wahl des Innenproduktes erhält man durch A^ε wieder einen Erzeuger einer kontraktiven Halbgruppe. Die Idee ist es, für jedes $\varepsilon > 0$ die Existenz und Eindeutigkeit einer Lösung $(u^\varepsilon, w^\varepsilon)$ des gestörten Problems zu zeigen und dann die Konvergenz vom $(u^\varepsilon, w^\varepsilon)$ für $\varepsilon \to 0$ zu beweisen. In diesem Fall ist die rechte Seite lokal Lipschitz-stetig. Die Lipschitz-Konstante hängt aber reziprok von ε ab, geht für $\varepsilon \to 0$ also gegen Unendlich. Das hat die Folge, dass sich die H^1-Norm von u^ε nicht unabhängig von ε beschränken lässt. Diese Beschränktheit ist aber notwendig, um $1 + f'(u) > 0$ garantieren zu können. Aus diesem Grund führt auch der Störungsansatz zu keinem Ergebnis.

Die Finite-Elemente-Methode

4

In diesem Kapitel werden wir auf die Begriffe und Ideen der Finite-Elemente-Methoden eingehen, die wir im nächsten Kapitel verwenden werden.

Wir beginnen mit den grundlegenden Begriffen der Finite-Elemente-Methoden und stellen die wichtigsten Resultate für elliptische Differentialgleichungen in Abschnitt 4.1 vor. In Abschnitt 4.2 erweitern wir die Anwendung von Finiten Elementen auf zeitabhängige Probleme, wir wenden dabei Finite Elemente auch in der Zeitdiskretisierung an und besprechen das Vorgehen zum Beweis der Konvergenz eines solchen Verfahrens. Schließlich gehen wir in Abschnitt 4.3 noch auf eine Technik der Vereinfachung der diskretisierten Gleichungen für nichtlineare Probleme und auf das Lösen nichtlinearer Gleichungssysteme ein.

4.1 Finite Elemente für elliptische Probleme

Es sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Gebiet mit genügend glattem Rand. Wir betrachten die Poisson-Gleichung

$$-\Delta u = f \qquad \text{in } \Omega, \tag{4.1a}$$

$$u = 0 \qquad \text{auf } \partial\Omega. \tag{4.1b}$$

Ein Finite-Elemente-Verfahren für diese Differentialgleichung basiert auf der zugehörigen Variationsformulierung: Wir suchen ein $u \in H_0^1(\Omega)$ mit

$$\int_\Omega \nabla u \cdot \nabla \varphi \, \mathrm{d}x = \int_\Omega f\varphi \, \mathrm{d}x \qquad (\varphi \in H_0^1(\Omega)). \tag{4.2}$$

Allgemeiner sei $a\colon V \times V \to \mathbb{R}$ eine Bilinearform auf dem Hilbertraum V und b ein Funktional auf V. Gesucht ist dann ein $u \in V$ mit

$$a(u,v) = b(v) \qquad (v \in V). \tag{4.3}$$

Auf dieses Problem kann man das Lemma von Lax-Milgram anwenden.

4.1 Lemma. (Lax-Milgram)
Es sei V ein Banachraum und $a\colon V \times V \to \mathbb{R}$, sowie $b \in V'$, wobei V' der Raum der stetigen Funktionale auf V sei. Wenn a eine stetige, koerzive Bilinearform ist, dann existiert genau eine Lösung $u \in V$ des Variationsproblems

$$a(u, v) = b(v) \quad (v \in V).$$

Beweis. Siehe beispielsweise [GR05, Lemma 3.6]. □

Im Falle des Poisson-Problems (4.2) sind die Bedingungen erfüllt, wenn $f \in L^2(\Omega)$ gilt (siehe beispielsweise [Hac96, Beispiel 7.2.10]).

Galerkin-Verfahren Um die Lösung u von (4.3) zu approximieren, wählen wir endlichdimensionale Vektorräume V_h, W_h und suchen ein $u_h \in V_h$ mit

$$a(u_h, v_h) = b(v_h) \qquad (v_h \in W_h). \tag{4.4}$$

Dabei nennen wir den Raum V_h *Ansatzraum* und den Raum W_h *Testraum*. Wenn $V_h = W_h$ gilt, dann nennt man ein solches Verfahren ein *Galerkin-Verfahren*, ist $V_h \neq W_h$, dann nennt man das Verfahren zur Abgrenzung häufig ein *Petrov-Galerkin-Verfahren*. In diesem Abschnitt beschränken wir uns auf den Fall $V_h = W_h$. Das Verfahren heißt *konform*, wenn $V_h \subset V$ gilt. In diesem Fall kann man auf Gl. (4.4) wieder das Lax-Milgram-Lemma anwenden und erhält Existenz und Eindeutigkeit einer Lösung wie in Gl. (4.3). Die Idee hinter dem Galerkin-Verfahren ist es, eine Folge von Räumen $(V_h)_{h\in\mathbf{H}}$ für eine Indexmenge $\mathbf{H}$ zu wählen, sodass die zugehörige Lösung u_h für $h \to 0$ gegen die analytische Lösung u konvergiert. Der Index h ist hier schon auf die Verwendung mit Finiten Elementen ausgelegt; h ist die maximale Gitterweite.

In dieser Arbeit nennen wir das Problem (4.3) häufig das *kontinuierliche Problem* und Gl. (4.4) das *diskretisierte Problem*.

Finite Elemente Ein Finite-Elemente-Verfahren ist ein Galerkin-Verfahren, das sich durch besondere Wahl des endlichdimensionalen Raumes V_h auszeichnet, obwohl die Begriffe „Galerkin-Verfahren“ und „Finite-Elemente-Verfahren“ häufig synonym verwendet werden.

Wir zerlegen das Gebiet Ω in eine Familie $\mathcal{T}_h$ von Polyedern, die bis auf deren Ränder disjunkt seien. Liegt eine Ecke eines Polyeders $K_1 \in \mathcal{T}_h$ auf dem Rand eines anderen Polyeders $K_2 \in \mathcal{T}_h$, dann soll die Ecke auch eine Ecke von K_2 sein. h bezeichne den maximalen Umkreisdurchmesser aller $K \in \mathcal{T}_h$; verkleinern wir also h, so werden die K entsprechend kleiner. Die Vereinigung aller $K \in \mathcal{T}_h$ sei eine polygonal berandete Menge $\Omega_h \subset \overline{\Omega}$, deren Randknoten auf $\partial\Omega$ liegen.

4.2 Definition.
Man nennt $\mathcal{T}_h$ *Triangulierung* von Ω. $h > 0$ ist der *Diskretisierungsparameter*, den wir *Gitterweite* nennen. Weiterhin ist jedes $K \in \mathcal{T}_h$ ein *Element*, wenn wir von der Triangulierung reden.

In $\mathbb{R}$ besteht eine Triangulierung aus einer Zerlegung von Ω in Intervalle, in $\mathbb{R}^2$ beispielsweise aus einer Zerlegung in Dreiecke und in $\mathbb{R}^3$ beispielsweise aus einer Zerlegung in Tetraeder. Der Einfachheit halber sei Ω polygonal berandet und $\Omega_h = \Omega$. Das Vorgehen für den allgemeinen Fall nicht polygonal berandeter Gebiet mit isoparametrischen Elementen findet man beispielsweise in [Cia02, Chapter 4.3].

Für ein $p \in \mathbb{N}$ definiert man dann den Finite-Elemente-Raum $\mathcal{S}_h^p = \mathcal{S}_h(\mathcal{T}_h)$ auf der Triangulierung $\mathcal{T}_h$. Die Funktionen aus $\mathcal{S}_h^p$ seien stückweise definiert. Gewöhnlich sind sie auf jedem Element $K \in \mathcal{T}_h$ Polynome. Dazu kommt bei konformen Verfahren noch eine globale Stetigkeitsbedingung, die vom verwendeten Raum in der variationellen Formulierung abhängt. Denn damit beispielsweise $\mathcal{S}_h^p \subset H_0^1(\Omega)$ erfüllt ist, müssen die Funktionen $\mathcal{S}_h^p$ auf $\overline{\Omega}$ stetig sein. Diese Stetigkeitsbedingung liefert der folgende Satz [Bra03, Satz 5.2].

4.3 Satz.
Sei $k \geq 1$. Eine stückweise beliebig oft differenzierbare Funktion $v\colon \overline{\Omega} \to \mathbb{R}$ gehört genau dann zu $H^k(\Omega)$, wenn $v \in C^{k-1}(\overline{\Omega})$ gilt.

Dieser Satz liefert uns die notwendige Information, um die Finite-Elemente-Räume für H_0^1-konforme Finite Elemente definieren zu können.

4.4 Definition. (Finite-Elemente-Räume)
Für $p \in \mathbb{N}$ ist der *Finite-Elemente-Raum des Grades p* für $H_0^1(\Omega)$-konforme Finite Elemente gegeben durch

$$\mathcal{S}_h^p = \left\{ \varphi \in C(\overline{\Omega})\colon \varphi\lfloor_{\partial\Omega} = 0, \varphi\lfloor_K \in \mathbb{P}_p(K), K \in \mathcal{T}_h \right\}.$$

Für $p = 1$ spricht man von *linearen Elementen* und für $p = 2$ von *quadratischen Elementen*.

Ein Aspekt, der für die Implementierung eines Finite-Elemente-Verfahrens wichtig ist, ist die Auswahl einer Basis des Finite-Elemente-Raumes $\mathcal{S}_h^p$. Jede Funktion in $\mathcal{S}_h^p$ kann durch die Koeffizienten bezüglich der gewählten Basis dargestellt werden. Wir möchten nur noch diese Koeffizienten berechnen. Die Basis legt man durch die Wahl sogenannter *Freiheitsgrade* fest, einer Familie

$$\mathcal{F}_h = \left\{ \Phi_i\colon \mathcal{S}_h^p \to \mathbb{R}, i = 1, \ldots, M \right\}$$

von M linearen Funktionalen auf $\mathcal{S}_h^p$. Diese Menge soll so gewählt sein, dass die Vorgabe der Werte dieser Funktionale eine Funktion in $\mathcal{S}_h^p$ eindeutig festlegt. Die zugehörige Basis $\{\varphi_i \in \mathcal{S}_h^p : i = 1, \dots, M\}$ erhält man aus der Bedingung

$$\varphi_j \in \mathcal{S}_h^p, \quad \Phi_i(\varphi_j) = \delta_{ij} \qquad (i, j = 1, \dots, M).$$

Ein wichtiges Beispiel wollen wir jetzt betrachten.

4.5 Beispiel. (Lineare Lagrange-Elemente)
Es sei $\mathcal{N}_h = \{1, \dots, M\}$ die Menge der Indizes der im Inneren von Ω liegenden Ecken x_i (auch *Gitterpunkte* genannt) aller $K \in \mathcal{T}_h$. M ist die Dimension von $\mathcal{S}_h^1$. Wir nennen $\mathcal{N}_h$ die *Knotenmenge* und identifizieren die x_i mit ihren Indizes. Die linearen *Lagrange-Elemente* (auch *Knotenelemente* oder *nodale Elemente*) bestehen aus dem Raum $\mathcal{S}_h^p$ und den Freiheitsgraden

$$\mathcal{F}_h = \left\{\Phi_\nu : \Phi_\nu(u) = u(x_\nu), u \in C(\overline{\Omega}), \nu \in \mathcal{N}_h\right\}.$$

Das sind die Punktauswertungen in den Knoten aus $\mathcal{N}_h$. Für $\Omega \subset \mathbb{R}$ sind die entsprechenden Basisfunktionen dann die *Hutfunktionen*, das heißt stückweise lineare Funktionen φ_ν mit

$$\varphi_\nu(x_{\nu'}) = \delta_{\nu\nu'} \qquad (\nu, \nu' \in \mathcal{N}_h).$$

Für $p > 1$ muss man die Knotenmenge entsprechend erweitern, bei quadratischen Elementen nimmt man beispielsweise zusätzlich zu den Ecken aller Elemente noch die Seitenmittelpunkte der $K \in \mathcal{T}_h$ hinzu. Häufig gibt man einfach nur die Basis von $\mathcal{S}_h^p$ an, ohne eine Erwähnung der Freiheitsgrade. In Abschnitt 6.1 werden wir genau dies für eine hierarchische Basis machen. Wir beachten, dass die Gitterpunkte und die Knoten nur im Fall linearer Knotenelemente gleich sind.

Im Folgenden nehmen wir an, dass die Freiheitsgrade der Knotenmenge durch $\nu \mapsto \Phi_\nu$ für $\nu \in \mathcal{N}_h$ zugeordnet werden können. Dann identifizieren wir sprachlich jeden Knoten mit seinem zugehörigen Freiheitsgrad.

4.6 Bemerkung.
Wir beachten, dass der Begriff „Element" kontextabhängig verschiedene Dinge bezeichnen kann.

1. *Es kann sich um ein $K \in \mathcal{T}_h$ handeln,*
2. *um den Funktionenraum mit dem Fokus auf dem zugrundeliegenden Polynomgrad (zum Beispiel „lineare Elemente"),*

3. *um den Funktionenraum mit dem Fokus auf die Freiheitsgrade oder die Basis (zum Beispiel „nodale Elemente“)*

4. *oder um das Finite Element mit Fokus auf die globale Stetigkeitsbedingung (zum Beispiel „stetige Elemente“, „stetig differenzierbare Elemente“).*

Nach dieser Vorarbeit können wir Gl. (4.4) als ein lineares Gleichungssystems schreiben.

4.7 Definition. (Diskretisierungsmatrizen, Koeffizientenvektor)
Die *Massematrix* $\mathbf{M} \in \mathbb{R}^{M\times M}$ und die *Steifigkeitsmatrix* $\mathbf{K} \in \mathbb{R}^{M\times M}$ der Ortsdiskretisierung seien für die Knoten $\nu, \nu' \in \mathcal{N}_h$ gegeben durch

$$\mathbf{M}_{\nu',\nu} := (\varphi_\nu, \varphi_{\nu'}), \quad \mathbf{K}_{\nu',\nu} := a(\varphi_\nu, \varphi_{\nu'}).$$

Man beachte dabei die Reihenfolge der Indizes. Weiterhin sei

$$U := (U_\nu)_{\nu\in\mathcal{N}_h} \in \mathbb{R}^M$$

der *Koeffizientenvektor* von u_h, das heißt auf Ω ist

$$u_h(x) = \sum_{\nu\in\mathcal{N}_h} U_\nu \varphi_\nu(x).$$

Dann ist Gl. (4.4) äquivalent zu

$$\mathbf{K}U = F, \tag{4.5}$$

mit $F = (F_\nu)_{\nu\in\mathcal{N}_h}$ und $F_\nu := b(\varphi_\nu)$. Die Massematrix wird in diesem Problem nicht benötigt, sie taucht im nächsten Abschnitt bei der Zeitdiskretisierung auf.

Wir nennen $u - u_h$ die *Fehlerfunktion* oder kurz den *Fehler*. Eine Fehlerabschätzung für das Poisson-Problem erhält man mithilfe des Lemmas von Céa. Dabei bezeichne $\|\cdot\|_V$ die Norm auf V.

4.8 Lemma. (Céa)
Für den Fehler $u - u_h$ gilt

$$\|u - u_h\|_V \leq C \inf_{v_h\in\mathcal{S}_h^p} \|u - v_h\|_V$$

für eine Konstante $C \geq 1$, die nur von der Koerzivitäts- und von der Stetigkeitskonstante der Bilinearform a abhängt.

Beweis. Siehe beispielsweise [Hac96, Satz 8.2.1]. □

Der Fehler in der Norm, die man häufig auch *Energienorm* nennt, lässt sich also im Wesentlichen durch die Energienorm der Bestapproximation im

Finite-Elemente-Raum abschätzen. Die Bestapproximation lässt sich wiederum dadurch abschätzen, dass wir ein Element $v_h \in \mathcal{S}_h^p$ wählen, das „genügend nahe“ an u liegt. Dafür verwendet man gewöhnlich eine Interpolation von u in $\mathcal{S}_h^p$. Alle Finiten Elemente besitzen eine natürliche Interpolation gegeben durch

$$I_h u := \sum_{\nu \in \mathcal{N}_h} \Phi_\nu(u) \varphi_\nu.$$

Damit diese Interpolation wohldefiniert ist, muss $\Phi_\nu(u)$ definiert sein. Dafür benötigt man eine gewisse Glattheit von u. Für nodale Elemente bedeutet das beispielsweise, dass die Knotenauswertung wohldefiniert sein muss. Also ist die Stetigkeit von u erforderlich. Ist u nicht stetig, so greift man auf eine andere Interpolation zurück, beispielsweise eine Clément-Interpolation, die glättend wirkt und nur $u \in L^2(\Omega)$ benötigt (vergleiche [BS08, Chapter 4.8] oder das Originalpaper [Clé75]). Die Soboleveinbettung (2.1) liefert die Stetigkeit von u für

$$u \in \begin{cases} H^1(\Omega), & d = 1, \\ H^2(\Omega), & d = 2, 3. \end{cases}$$

Dann kann man durch das Bramble-Hilbert-Lemma [Bra03, Lemma 6.3] beweisen, dass für $u \in H^r(\Omega)$ mit $2 \leq r \leq p+1$ unter weiteren Bedingungen an die Triangulierung die folgende Abschätzung gilt, siehe dazu [Cia02, Theorem 3.2.1],

$$\|u - I_h u\|_{L^2(\Omega)} + h \|\nabla(u - I_h u)\|_{L^2(\Omega)} \leq C h^r \|\nabla^r u\|_{L^2(\Omega)}.$$

Zusammen mit dem Lemma von Céa folgt die Fehlerabschätzung

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \leq C \inf_{v_h \in \mathcal{S}_h^p} \|\nabla(u - v_h)\|_{L^2(\Omega)} \leq C h^{r-1} \|\nabla^r u\|_{L^2(\Omega)}. \quad (4.6)$$

Mittels eines Dualitätsarguments und zusätzlicher Bedingungen an das Gebiet Ω kann man für die L^2-Norm des Fehlers eine Ordnung in h gewinnen, vergleiche mit [Cia02, Theorem 3.2.5]. Dieses Argument werden wir aber in dieser Arbeit nicht benötigen.

4.9 Bemerkung. (Elliptische Projektion)
Wir beobachten, dass aus (4.2) *und* (4.4) *für konforme Verfahren die Gleichung*

$$a(u - u_h, v_h) = 0 \qquad (v_h \in V_h)$$

folgt, die sogenannte Galerkin-Orthogonalität. *Für die Poisson-Gleichung bedeutet das*

$$(\nabla(u - u_h), \nabla\varphi) = 0 \qquad (\varphi \in \mathcal{S}_h^p).$$

u_h ist also die orthogonale Projektion von u auf den Raum $\mathcal{S}_h^p$ bezüglich des H_0^1-Skalarproduktes. Diese orthogonale Projektion können wir auf dem Raum $H_0^1(\Omega)$ definieren, wir schreiben sie $\mathcal{P}_E$ und nennen sie die elliptische Projektion *auf $\mathcal{S}_h^p$. Für diese elliptische Projektion können wir jetzt die Abschätzung* (4.6) *verwenden, es gilt also für $w \in H^r(\Omega) \cap H_0^1(\Omega)$*

$$\|w - \mathcal{P}_E w\|_{L^2} + h \|\nabla(w - \mathcal{P}_E w)\|_{L^2} \leq Ch^r \|\nabla^r w\|_{L^2}.$$

Das gilt deswegen, weil w trivialerweise (4.3) *mit $a(u,v) = (\nabla u, \nabla v)$ und $b(v) := (\nabla w, \nabla v)$ löst. b ist dank der Stetigkeit von a ein stetiges Funktional auf $H_0^1(\Omega)$. $\mathcal{P}_E w$ ist dann die Lösung des zugehörigen Problems* (4.4) *mit $V_h = W_h = \mathcal{S}_h^p$.*

4.2 Ein Petrov-Galerkin-Verfahren für zeitabhängige Probleme

In diesem Abschnitt erweitern wir die Diskretisierung elliptischer Differentialgleichungen auf die Diskretisierung zeitabhängiger Differentialgleichungen durch ein so genanntes *Petrov-Galerkin-Verfahren*. Ein solches Verfahren unterscheidet sich vom klassischen Galerkin-Verfahren dadurch, dass der Ansatz- und der Testraum nicht übereinstimmen. Wir beginnen damit, das Standardbeispiel parabolischer Differentialgleichungen, die Wärmeleitungsgleichung, mit einem Petrov-Galerkin-Verfahren in Zeit und Ort zu diskretisieren. Dies dient zur Illustrierung der Anwendung des stetigen Galerkinverfahrens in der Zeitdiskretisierung. Das dort besprochene Vorgehen wenden wir danach auf die Wellengleichung an und zeigen, auf welche Art und Weise man eine Abschätzung des Diskretisierungsfehlers erhalten kann.

4.2.1 Zeitdiskretisierung durch ein Petrov-Galerkin-Verfahren

Es sei $\Omega_T := (0,T] \times \Omega$ und wir betrachten das Rand-Anfangswertproblem

$$\partial_t u + \Delta u = f \qquad \text{in } \Omega_T, \tag{4.7a}$$
$$u = 0 \qquad \text{auf } [0,T] \times \partial\Omega, \tag{4.7b}$$
$$u = u_0 \qquad \text{auf } \{0\} \times \Omega. \tag{4.7c}$$

Wir folgen der Darstellung schwacher Lösungen in [Eva98, Chapter 7]. Wir fassen u als Funktion von $[0,T]$ nach $H_0^1(\Omega)$ auf, genauso sei $f\colon [0,T] \to L^2(\Omega)$. Wir definieren dann eine *schwache Lösung* von Gl. (4.7) durch eine Funktion $u \in L^2(0,T;H_0^1(\Omega))$ mit $u' \in L^2(0,T;H^{-1}(\Omega))$, die

$$\langle u', \varphi \rangle - (\nabla u, \nabla \varphi) = (f, \varphi) \qquad (\varphi \in H_0^1(\Omega)) \tag{4.8}$$

für fast alle $t \in [0,T]$ erfüllt und der Anfangsbedingung $u(0) = u_0$ genügt. Dabei ist $\langle \cdot, \cdot \rangle$ die duale Paarung auf $H^{-1}(\Omega) \times H_0^1(\Omega)$. Die Anfangsbedingung ist wohldefiniert, da unter den obigen Voraussetzungen für die Lösung $u \in C([0,T];L^2(\Omega))$ gilt (vergleiche [Eva98, Section 5.9.2, Theorem 3]).

Im Folgenden sei der Einfachheit halber $f = 0$. Wir setzen wie zuvor

$$a(u, \varphi) := (\nabla u, \nabla \varphi)$$

und schreiben Gl. (4.7) als Variationsproblem in $L^2(0, T; H_0^1(\Omega))$, wir suchen also ein $u \in H^1(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$ mit

$$\int_0^T (\partial_t u, \phi) + a(u, \phi) \, \mathrm{d}t = 0 \tag{4.9}$$

für alle $\phi \in L^2(0, T; H_0^1(\Omega))$ mit der Anfangsbedingung

$$u(0) = u_0 \in H_0^1(\Omega). \tag{4.10}$$

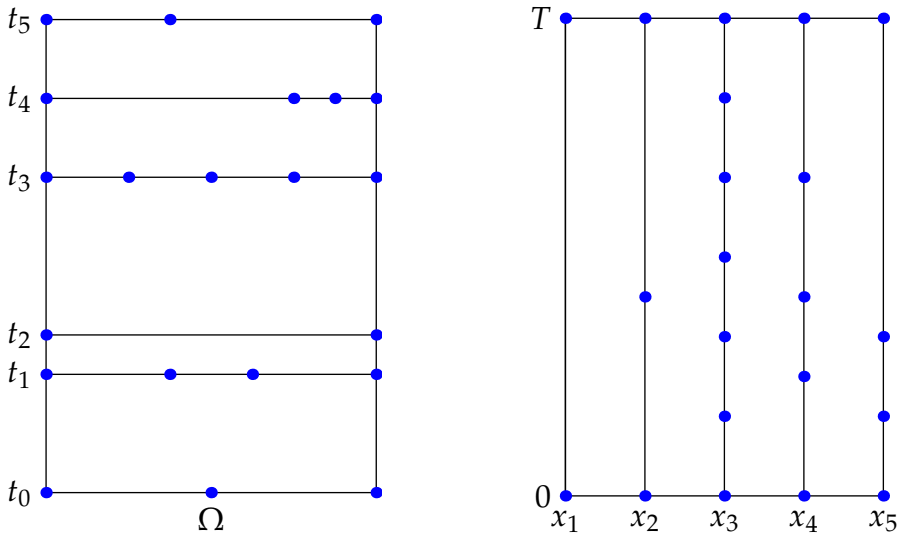
Für die Definition einer Volldiskretisierung von (4.9) gibt es verschiedene Vorgehensweisen.

1. Die *horizontale Linienmethode* oder *Rothe-Methode* besteht aus einer Semidiskretisierung von (4.9) bezüglich der Zeit. Man kann dazu jedes Verfahren zur Diskretisierung bezüglich der Zeitvariablen benutzen, zum Beispiel Finite Differenzen (siehe [GR05, Kapitel 5.2.4]) oder das unstetige Galerkinverfahren (siehe [Tho97, Chapter 12]). Das führt auf ein *Zeitschrittverfahren*, in diesem Fall eine Folge elliptischer Probleme, die man schrittweise löst und in jedem Zeitschritt die Lösung des semidiskretisierten Problems zu einem bestimmten Zeitpunkt in $[0, T]$ erhält. Die elliptischen Probleme werden dann wieder diskretisiert, beispielsweise durch Finite Elemente wie im vorherigen Abschnitt. Die Rothe-Methode ermöglicht es, die Ortsdiskretisierung für jeden Zeitschritt zu variieren, siehe beispielsweise [BJ74] und [VF80].

2. Die *(vertikale) Linienmethode* besteht aus einer Semidiskretisierung bezüglich des Ortes. Man erhält ein System gewöhnlicher Differentialgleichungen, das man mit dem präferierten numerischen Verfahren lösen kann. Diese Methode bietet sich an, wenn man den Zeitschritt variabel (im Ort) wählen möchte, die Ortsdiskretisierung aber nicht zeitabhängig sein soll.

3. Eine dritte Möglichkeit verwendet unstrukturierte Gitter auf $\overline{\Omega_T}$. In diesem Fall macht man keinen Unterschied zwischen der Zeit- und den Ortsvariablen. Daher erhält man auch kein Zeitschrittverfahren. Dieses Vorgehen vergrößert das zu lösende Gleichungssystem, beträchtlich (man muss es allerdings auch nur einmal lösen). Es gibt Probleme, in denen ein unstrukturiertes Gitter notwendig ist, sei es bei der Anwendung adaptiver Strategien simultan in Ort und Zeit, sei es, weil man nur auf diese Weise eine gute Approximation erhält, siehe beispielsweise [HH90].

In Abb. 4.1 sehen wir den Unterschied der Zeit-Raum-Gitter zwischen der Rothe- und der Linienmethode. In der Rothe-Methode ändert sich das Ortsgitter in jedem Zeitschritt, in der Linienmethode ändert sich das Zeitgitter in jedem Ortsknoten. In diesem und dem nächsten Kapitel sehen wir davon



(a) Rothe-Methode **(b)** Linienmethode

Abbildung 4.1: Illustrierung der Unterschiede zwischen Rothe- und Linienmethode.

ab, das Ortsgitter zeitabhängig oder das Zeitgitter ortsabhängig zu definieren. Daher spielt es keine Rolle, ob wir nach der Rothe-Methode oder der Linienmethode vorgehen. In Kapitel 6 aber definieren wir ein Zeitschrittverfahren, welches einen ortsabhängigen Zeitschritt zulässt. Daher werden wir die Linienmethode verwenden.

Wir wählen im Ort $H_0^1(\Omega)$-konforme Elemente, wie wir es im vorherigen Abschnitt besprochen haben. Wir haben also einen endlichdimensionalen Unterraum $\mathcal{S}_h^p$ von $H_0^1(\Omega)$ und definieren die Semidiskretisierung von (4.9) durch die Bestimmung von $u_h \in H^1(0,T;\mathcal{S}_h^p)$ mit

$$\begin{aligned} (\partial_t u_h, \varphi) + a(u_h, \varphi) &= 0 \qquad (\varphi \in \mathcal{S}_h^p), \\ u_h(0) &= u_h^0 \in \mathcal{S}_h^p \end{aligned} \tag{4.11}$$

für fast alle $t \in (0,T)$, wobei u_h^0 eine geeignete Projektion von u_0 auf den Finite-Elemente-Raum sei. Gl. (4.11) lässt sich als lineares Anfangswertproblem schreiben, es ist

$$\mathbf{M}U' + \mathbf{K}U = 0 \tag{4.12}$$

für $t \in (0,T)$ mit der Anfangsbedingung $U(0) = U_0$, wobei U_0 der Koeffizientenvektor von u_h^0 ist. U ist der zeitabhängige Koeffizientenvektor der Finite-Elemente-Lösung u_h, das heißt

$$u_h(t,x) = \sum_{\nu \in \mathcal{N}_h} U_\nu(t)\varphi_\nu(x).$$

Im nächsten Schritt diskretisieren wir Gl. (4.11) auch in der Zeit. Wir behandeln Zeit und Ort ähnlich, benutzen also auch ein Finite-Elemente-Verfahren zur Zeitdiskretisierung. Dazu sei $0 = t_0 < t_1 < \cdots < t_N = T$ und wir definieren damit ein *Zeitgitter*

$$\mathcal{T}_k = \mathcal{T}_k(0,T) := \{t_0, \ldots, t_N\}$$

über $[0,T]$ für $N \in \mathbb{N}$. Wir definieren die *Zeitschrittweite* sowie die *maximale Zeitschrittweite* durch

$$\begin{aligned} k_n &:= t_{n+1} - t_n \qquad (n = 0, \ldots, N-1), \\ k &:= \max_{n=0,\ldots,N-1} k_n. \end{aligned}$$

Weiterhin setzen wir

$$I_n := [t_n, t_{n+1}] \qquad (n = 0, \ldots, N-1).$$

4.10 Definition.
Wir definieren den Finite-Elemente-Raum $H^1(0,T)$-konformer Elemente für $q \in \mathbb{N}$ durch

$$\mathcal{S}_k^q = \mathcal{S}_k^q(0,T) := \left\{ \psi \in C([0,T]) \colon \psi_{\lfloor I_n} \in \mathbb{P}_q(I_n), n = 0, \ldots, N-1 \right\}.$$

Außerdem setzen wir

$$\mathcal{S}_k^0 := \left\{ \psi \colon [0,T] \to \mathbb{R} \mid \psi_{\lfloor I_n} \in \mathbb{P}_0(I_n), n = 0, \ldots, N-1 \right\}.$$

Damit wir eine Knotenbasis dieses Raumes definieren können, benötigen wir für $q > 1$ mehr Knoten als in $\mathcal{T}_k$ liegen und wir setzen für $i = 0, \ldots, q$

$$t_{n,i} := t_n + k_n \frac{i}{q}.$$

Es gilt $t_{n,q} = t_{n+1,0} = t_{n+1}$. Wir definieren dann eine nodale Basis von $\mathcal{S}_k^q$ durch

$$\psi_j^m \in \mathcal{S}_k^q, \quad \psi_j^m(t_{n,i}) = \delta_{nm}\delta_{ij} \quad \text{für } i,j = 0, \ldots, q \text{ und } m,n = 0, \ldots, N.$$

Gl. (4.11) ist jetzt die gewöhnliche Differentialgleichung, die wir mit Finiten Elementen behandeln wollen. Wir machen also den Ansatz

$$u_{kh}(t,x) = \sum_{\nu\in\mathcal{N}_h} \left(U_\nu^{0,0}\psi_0^0(t) + \sum_{n=0}^{N-1}\sum_{i=1}^{q} U_\nu^{n,i}\psi_i^n(t) \right)\varphi_\nu(x).$$

Das ist ein *Tensorproduktansatz*, wir definieren den *Tensorproduktraum* entsprechend wie folgt.

4.11 Definition. (Tensorproduktraum)
Für $q \in \mathbb{N}_0$ definieren wir den *Tensorproduktraum* durch

$$\mathcal{S}_k^q \otimes \mathcal{S}_h^p = \left\{ \phi\colon \Omega_T \to \mathbb{R} \mid \phi(t,x) = \psi(t)\varphi(x),\ \psi \in \mathcal{S}_k^q, \varphi \in \mathcal{S}_h^p \right\}.$$

Wir bestimmen die Funktion u_{kh} schrittweise für jedes Teilintervall I_n, indem wir Gl. (4.11) mit Basisfunktionen von $\mathbb{P}_{q-1}(I_n)$, die durch 0 auf $(0,T)$ fortgesetzt werden, für alle $n = 0,\dots,N-1$ multiplizieren und über t von 0 bis T integrieren. Wir suchen also $u_{kh} \in \mathcal{S}_k^q \otimes \mathcal{S}_h^p$ mit

$$\int_{I_n} (\partial_t u_{kh}, \phi) + a(u_{kh}, \phi)\ \mathrm{d}t = 0 \qquad (\phi \in \mathbb{P}_{q-1}(I_n) \otimes \mathcal{S}_h^p), \tag{4.13}$$

für alle $n = 0,\dots,N-1$ und mit der Anfangsbedingung $u_{kh}(0) = u_h^0$. Die Diskretisierung der Wellengleichung findet man in dieser Form bei [AM89].

4.12 Bemerkung.
In Gl. (4.13) *unterscheiden sich der Testraum und der Ansatzraum. Zum einen sind die Polynome im Testraum bezüglich der Zeit um einen Grad kleiner als die im Ansatzraum und zum anderen besteht der Testraum aus unstetigen Funktionen auf* $[0,T]$.

Diese Wahl des Testraumes führt zu einem wohldefinierten Verfahren, weil von den zu bestimmenden Koeffizienten $U_\nu^{n,i}$, $i = 0,\dots,q$, *der Koeffizient* $U_\nu^{n,0}$ *entweder ein Anfangswert* $U_\nu^{0,0}$ *ist oder schon im vorherigen Zeitschritt berechnet wurde, denn für* $n > 0$ *gilt dank der Stetigkeit der Funktionen in* $\mathcal{S}_k^q \otimes \mathcal{S}_h^p$ *die Gleichung* $U_\nu^{n,0} = U_\nu^{n-1,q}$.

Galerkin-Verfahren, bei denen sich Ansatz- und Testraum unterscheiden, nennt man Petrov-Galerkin-Verfahren.

Das Petrov-Galerkin-Verfahren (4.13) in der Zeit nennt man auch *stetiges Galerkinverfahren*, man findet die Bezeichnung cG(q) für „**c**ontinuous **G**alerkin“, wohingegen das *unstetige Galerkinverfahren* mit dG(q) für „**d**iscontinuous **G**alerkin“ abgekürzt wird (vergleiche [Tho97]). Die Volldiskretisierung, bei der man auch im Ort stetige Finite Elemente verwendet, schreiben wir kurz als

cG(q) cG(p)-Verfahren. Die Diskretisierung im Ort kann mittels des Tensorproduktansatzes aber durch einen beliebigen Finite-Elemente-Raum erfolgen.

4.13 Bemerkungen.

1. *Das stetige Galerkinverfahren in der Zeit findet man beinahe nur in Randbemerkungen in der gängigen Literatur. In [GR05, S. 333] heißt es beispielsweise „Diese stetige Galerkin-Methode [...] ist nicht sonderlich beliebt, weil Raum und Zeit ähnlich behandelt werden und stetige Raum-Zeit-Elemente vermieden werden können...". Es ist aber nicht klar, warum stetige Raum-Zeit-Elemente vermieden werden sollten und warum eine ähnliche Behandlung von Raum und Zeit so unvorteilhaft ist. Einen sehr viel einsichtigeren Kommentar findet man in [Tho97, S. 185], der da lautet „Because [the continuous Galerkin method] has less advantageous smoothing properties than the discontinuous Galerkin method [...], we shall refrain from a detailed analysis here." Der Grund dafür ist, dass sich cG-Methoden im homogenen Fall mit Kollokationsverfahren bezüglich der Gauß-Legendre-Knoten in der Zeit identifizieren lassen, und diese Kollokationsverfahren wiederum sind äquivalent zur Verwendung einer diagonalen Padé-Approximation (vgl. [AM89] und [Hul72]). Diagonale Padé-Approximationen führen zu nichtdispersiven Verfahren, für die Wellengleichung erhält ein cG-Verfahren die Energie, wie es gewünscht ist. dG-Methoden entsprechen im homogenen Fall einer subdiagonalen Padé-Approximation. Diese Klasse von Verfahren ist dissipativ, dies bedeutet eine Glättung der Lösung (das sind die „smoothing properties", von denen im obigen Zitat die Rede ist). Eine solche Eigenschaft ist vorteilhaft für die Diskretisierung parabolischer Differentialgleichung, da man für diese weiß, dass sie dissipativ sind, also dass die Lösungen über die Zeit an Energie verlieren. Für hyperbolische Gleichungen dagegen, bei denen meist eine Energie erhalten wird, ist eine energieerhaltende Diskretisierung geeigneter.*

2. *Die Wahl der Finite-Elemente-Räume als Tensorprodukt von Finite-Elemente-Räumen für die Zeit- und die Ortsdiskretisierung erleichtert sowohl die Fehleranalysis als auch die Implementierung, da wir die Diskretisierungen von Zeit und Raum weitestgehend getrennt behandeln können.*

4.2.2 Diskretisierung der linearen Wellengleichung

In diesem Abschnitt gehen wir kurz darauf ein, wie man die Konvergenz des im vorherigen Abschnitt vorgestellten Petrov-Galerkin-Verfahrens für die Wellengleichung zeigt. Für hyperbolische Systeme erster Ordnung in der Zeit wird die Konvergenz eines cG-Verfahrens in [Dup73] gezeigt. Für die Wellengleichung in der Form (4.14) finden sich Konvergenzbeweise eines cG-Verfahrens beispielsweise in [FP96] und [BL94]. Man beachte, dass [FP96]

von [AM89] (zur Wärmeleitungsgleichung) beeinflusst wurde und wiederum [BL94] beeinflusste (in Form eines „Research Report“, der schon 1991 vorlag). In beiden Arbeiten zur Wellengleichung werden Abschätzungen sowohl in der $H_0^1(\Omega) \times L^2(\Omega)$-Norm als auch in der $L^2(\Omega) \times H^{-1}(\Omega)$-Norm bewiesen. Wir beschränken uns hier auf den ersten Fall, auch wenn der zweite Fall in Hinblick auf die Energieerhaltung (3.19) interessant erscheint.

Wir betrachten die Wellengleichung als System erster Ordnung

$$\partial_t u = v, \tag{4.14a}$$
$$\partial_t v = \Delta u + f \tag{4.14b}$$

auf $\Omega_T := (0, T] \times \Omega$ mit homogenen Dirichlet-Randbedingungen, den Anfangsbedingungen $u(0, x) = u_0(x)$ und $v(0, x) = v_0(x)$ für $x \in \Omega$ und mit $f \in L^2(0, T; L^2(\Omega))$. In variationeller Formulierung suchen wir nach $u \in H^1(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$ und $v \in H^1(0, T; L^2(\Omega))$ mit

$$\int_0^T (\partial_t u, \phi_1)\, \mathrm{d}t = \int_0^T (v, \phi_1)\, \mathrm{d}t, \tag{4.15a}$$
$$\int_0^T (\partial_t v, \phi_2)\, \mathrm{d}t = -\int_0^T (\nabla u, \nabla \phi_2)\, \mathrm{d}t + \int_0^T (f, \phi_2)\, \mathrm{d}t, \tag{4.15b}$$

für alle $\phi_1, \phi_2 \in L^2(0, T; H_0^1(\Omega))$. Wir verwenden wie schon in Abschnitt 3.3.1 den Operator $\mathcal{A}\colon H_0^1(\Omega) \to L^2(\Omega)$, der für $w \in H_0^1(\Omega)$ definiert ist durch

$$(\mathcal{A}w, \phi) = (\nabla w, \nabla \phi) \qquad (\phi \in H_0^1(\Omega)).$$

Testen wir in (4.15) mit $\phi_1 = \mathcal{A}u$ und $\phi_2 = v$, so folgt

$$\frac{1}{2} \int_0^T \partial_t \|\nabla u\|_2^2 \, \mathrm{d}t = \int_0^T (\nabla v, \nabla u)\, \mathrm{d}t,$$
$$\frac{1}{2} \int_0^T \partial_t \|v\|_2^2 \, \mathrm{d}t = -\int_0^T (\nabla u, \nabla v)\, \mathrm{d}t + \int_0^T (f, v)\, \mathrm{d}t.$$

Addieren wir die beiden Gleichungen, so erhalten wir

$$\|\nabla u(T)\|_2^2 + \|v(T)\|_2^2 = \|\nabla u_0\|_2^2 + \|v_0\|_2^2 + 2\int_0^T (f, v)\, \mathrm{d}t. \tag{4.16}$$

Für $f \equiv 0$ ist das eine Energiegleichung bezüglich der Norm auf $H_0^1(\Omega) \times L^2(\Omega)$, die für $w = (w_1, w_2) \in H_0^1(\Omega) \times L^2(\Omega)$ definiert ist durch

$$\|w\|_{H_0^1 \times L^2}^2 := \|\nabla w_1\|_2^2 + \|w_2\|_2^2.$$

Diese Energiegleichung wird eine wichtige Rolle bei der Fehleranalyse des zugehörigen Petrov-Galerkin-Verfahrens spielen.

Wir definieren die Finite-Elemente-Lösung der Wellengleichung durch $u_{kh}, v_{kh} \in \mathcal{S}_k^q \otimes \mathcal{S}_h^p$ mit

$$\int_{I_n} (\partial_t u_{kh}, \phi_1) \, \mathrm{d}t = \int_{I_n} (v_{kh}, \phi_1) \, \mathrm{d}t, \tag{4.17a}$$

$$\int_{I_n} (\partial_t v_{kh}, \phi_2) \, \mathrm{d}t = -\int_{I_n} (\nabla u_{kh}, \nabla \phi_2) \, \mathrm{d}t + \int_{I_n} (f, \phi_2) \, \mathrm{d}t \tag{4.17b}$$

für alle $n = 0, \ldots, N-1$ und $\phi_1, \phi_2 \in \mathbb{P}_{q-1} \otimes \mathcal{S}_h^p$. Die Anfangswerte für $n = 0$ seien geeignete Approximationen u_{kh}^0, v_{kh}^0 von u_0 und v_0 im Finite-Elemente-Raum $\mathcal{S}_h^p$, beispielsweise die elliptischen Projektionen der genannten Funktionen. Die Anfangswerte für $n > 0$ ergeben sich aus der Berechnung von u_{kh}, v_{kh} auf dem Intervall I_{n-1}. Ein Vorteil des cG-Verfahrens in der Zeit ist, dass die Energiegleichung (4.16) des kontinuierlichen Problems automatisch auch für das diskretisierte Problem gilt. Um diese Eigenschaft zu zeigen, können wir nicht direkt so vorgehen wie im kontinuierlichen Problem, da der Testraum sich vom Ansatzraum unterscheidet. Wir benutzen daher die $L^2(I_n; L^2(\Omega))$-orthogonale Projektion $\mathcal{Q}_n^{q-1}$ auf den Testraum (die präzise Definition verschieben wir auf später, sie findet sich in (5.7)). Weiterhin definieren wir den diskreten Laplace-Operator $\mathcal{A}_h \colon \mathcal{S}_h^p \to \mathcal{S}_h^p$ durch

$$(\mathcal{A}_h w, \phi) = (\nabla w, \nabla \phi) \qquad (w, \phi \in \mathcal{S}_h^p). \tag{4.18}$$

Wir testen in (4.17) mit $\phi_1 = \mathcal{A}_h \mathcal{Q}_n^{q-1} u_{kh}$ und $\phi_2 = \mathcal{Q}_n^{q-1} v_{kh}$ und erhalten wie zuvor

$$\|\nabla u_{kh}(T)\|_2^2 + \|v_{kh}(T)\|_2^2 = \|\nabla u_{kh}^0\|_2^2 + \|v_{kh}^0\|_2^2 + \sum_{n=0}^{N-1} \int_{I_n} (f, \mathcal{Q}_n^{q-1} v_{kh}) \, \mathrm{d}t.$$

Aus dieser Gleichung folgen sofort Existenz und Eindeutigkeit, denn für $f \equiv 0$ und verschwindende Anfangswerte folgt

$$\|\nabla u_{kh}(T)\|_2^2 + \|v_{kh}(T)\|_2^2 = 0$$

und damit $v_{kh} = 0$ und $u_{kh} = 0$ aufgrund der Nullrandbedingung. Um Abschätzungen für die Fehlerfunktionen $e_u := u - u_{kh}$ und $e_v := v - v_{kh}$ zu erhalten, möchten wir die Gleichungen (4.15) und (4.17) verwenden. Wir können eine variationelle Formulierung für den Fehler aufstellen. Sie lautet

$$\int_{I_n} (\partial_t e_u, \phi_1) \, \mathrm{d}t = \int_{I_n} (e_v, \phi_1) \, \mathrm{d}t,$$

$$\int_{I_n} (\partial_t e_v, \phi_2) \, \mathrm{d}t = -\int_{I_n} (\nabla e_u, \nabla \phi_2) \, \mathrm{d}t,$$

für alle $\phi_1, \phi_2 \in \mathbb{P}_{q-1}(I_n) \otimes \mathcal{S}_h^p$ mit gegebenen Anfangswerten e_u^n, e_v^n. Es ist jetzt nicht möglich mit $\phi_1 = \mathcal{A}_h \mathcal{Q}_n^{q-1} e_u$ und $\phi_2 = \mathcal{Q}_n^{q-1} e_v$ zu testen, diese

Funktionen liegen nämlich nicht im Testraum, da im Allgemeinen für die Fehler $e_u(t), e_v(t) \notin \mathcal{S}_h^p$ für $t \in [0,T]$ gilt. Daher spaltet man den Fehler beispielsweise wie folgt auf.

$$e_u = (u - \widetilde{u}) + (\widetilde{u} - u_{kh}) =: \rho_u + \theta_u, \tag{4.19a}$$
$$e_v = (v - \widetilde{v}) + (\widetilde{v} - v_{kh}) =: \rho_v + \theta_v, \tag{4.19b}$$

mit $\widetilde{u}, \widetilde{v} \in \mathcal{S}_k^q \otimes \mathcal{S}_h^p$. Die *Zwischenelemente* $\widetilde{u}, \widetilde{v}$ kann man auf verschiedene Weisen definieren. Wir werden darauf in Abschnitt 5.2.4 noch genauer eingehen und führen hier nur einen Weg vor. Es seien $\mathcal{P}\colon L^2(0,T;H_0^1(\Omega)) \to L^2(0,T) \otimes \mathcal{S}_h^p$ und $\mathcal{Q}\colon H^1(0,T;L^2(\Omega)) \to \mathcal{S}_k^q \otimes L^2(\Omega)$ Projektionen. Wir definieren

$$\widetilde{u} := \mathcal{P}\mathcal{Q}u, \quad \widetilde{v} := \mathcal{P}\mathcal{Q}v.$$

Im ersten Schritt treffen wir unter gewissen Glattheitsvoraussetzungen an u und v Aussagen darüber, wie sich (ρ_u, ρ_v) in der Energienorm abhängig von den Diskretisierungsparametern k und h verhält. Diese Aussagen legen dann auch schon das bestmögliche Konvergenzverhalten für den Fehler (e_u, e_v) fest, die wir mit der Aufspaltung Gl. (4.19) erhalten können.

Im zweiten Schritt müssen wir (θ_u, θ_v) mit der gleichen Fehlerordnung abschätzen können. Hier verwendet man die Energiemethode, zeigt also, dass (θ_u, θ_v) einer gestörten Version der Differentialgleichung genügt, also

$$\int_{I_n} (\partial_t \theta_u, \phi_1)\,\mathrm{d}t = \int_{I_n} (\theta_v, \phi_1)\,\mathrm{d}t + \int_{I_n} (R_1, \phi_1)\,\mathrm{d}t,$$
$$\int_{I_n} (\partial_t \theta_v, \phi_2)\,\mathrm{d}t = -\int_{I_n} (\nabla\theta_u, \nabla\phi_2)\,\mathrm{d}t + \int_{I_n} (R_2, \phi_2)\,\mathrm{d}t,$$

und testet dann mit $\phi_1 = \mathcal{A}_h \mathcal{Q}_n^{q-1}\theta_u$ und $\phi_2 = \mathcal{Q}_n^{q-1}\theta_v$. In diesem Fall funktioniert das, weil sowohl θ_u als auch θ_v in $\mathcal{S}_k^q \otimes \mathcal{S}_h^p$ liegen. R_1 und R_2 sind Restterme, in diesem Fall wieder Projektionsfehler. Es folgt also die Abschätzung

$$\begin{aligned}&\|\nabla\theta_u(t_{n+1})\|_2^2 - \|\nabla\theta_u(t_n)\|_2^2 + \|\theta_v(t_{n+1})\|_2^2 - \|\theta_v(t_n)\|_2^2 \\ &\quad\le \|\nabla\theta_u\|_{L^2(I_n;L^2)}^2 + \|\theta_v\|_{L^2(I_n;L^2)}^2 + \|\nabla R_1\|_{L^2(I_n;L^2)}^2 + \|R_2\|_{L^2(I_n;L^2)}^2.\end{aligned}$$

Für die Restterme benötigen wir wieder Abschätzungen der Projektionsfehler. Wenn man diese hat, dann verbleibt nur noch, die Teilresultate in (4.19) einzusetzen und das Gronwall-Lemma zu verwenden.

Dieses grundlegende Vorgehen werden wir im Beweis der Konvergenz eines Petrov-Galerkin-Verfahrens für das nichtlineare Problem verfolgen (siehe Abschnitt 5.2). Da wir für die Formulierung dieses Verfahrens weitere Projektionen benötigen, werden wir die entsprechenden Definitionen und Abschätzungen am Anfang des nächsten Kapitels tätigen.

4.3 Aspekte der Diskretisierung nichtlinearer Probleme

In diesem letzten Einführungsabschnitt zu Finite-Elemente-Methoden gehen wir noch auf zwei Aspekte der Diskretisierung nichtlinearer Probleme ein. Wir streifen dabei kurz die Diskretisierung stationärer, nichtlinearer Probleme und wie man mittels der *Produktapproximation* genannten Technik die Implementierung eines Finite-Elemente-Verfahren für nichtlineare Probleme vereinfachen kann. Schließlich besprechen wir das Lösen der resultierenden nichtlinearen Gleichungssysteme durch das Newton-Verfahren.

Wir betrachten in diesem Abschnitt als Beispiel die Gleichung

$$-\Delta u = f(u) \qquad \text{in } \Omega, \tag{4.20a}$$

$$u = 0 \qquad \text{auf } \partial\Omega. \tag{4.20b}$$

f sei dabei eine Funktion auf $\mathbb{R}$. Wir können dieses Problem genauso diskretisieren wie in Abschnitt 4.1. Wir suchen also $u_h \in \mathcal{S}_h^p$ mit

$$(\nabla u_h, \nabla \varphi) = (f(u_h), \varphi) \qquad (\varphi \in \mathcal{S}_h^p).$$

Dort steht jetzt kein lineares Gleichungssystem mehr, sondern ein nichtlineares Gleichungssystem. In Matrix-Vektor-Form schreibt sich diese Gleichung als

$$\mathbf{K}U = F(U), \tag{4.21}$$

wobei $F(U) = (F_\nu(U))_{\nu \in \mathcal{N}_h}$ mit $F_\nu(U) = (f(u_h), \varphi_\nu)$ ist. Um dieses Gleichungssystem zu lösen, kann man ein iteratives Verfahren zum Lösen nichtlinearer Gleichungen verwenden, beispielsweise das Newton-Verfahren. In jedem Iterationsschritt müssten wir dabei die nichtlineare Funktion neu assemblieren. Um das zu kann man die Produktapproximation anwenden.

4.3.1 Produktapproximation

Die rechte Seite in Gl. (4.21) ist aus recheneffizienter Sicht problematisch. Wir wissen bereits, dass der Fehler des linearen Problems mit $f = 0$ in der H^1-Norm von der Ordnung p ist. Es stellt sich die Frage, ob es notwendig ist, $(f(u_h), \varphi_\nu)$ zu berechnen oder ob es nicht ausreicht, statt $f(u_h)$ eine Projektion dieser Funktion auf $\mathcal{S}_h^p$ zu verwenden. Mit der Darstellung

$$u_h = \sum_{\nu \in \mathcal{N}_h} U_\nu \varphi_\nu$$

erreicht man eine solche Projektion ganz einfach durch

$$f(u_h) \approx \sum_{\nu \in \mathcal{N}_h} f(U_\nu) \varphi_\nu.$$

Diese Idee der Knoteninterpolation von $f(u_h)$ auf $\mathcal{S}_h^p$ stammt aus [CGMSS81] und wurde dort *Produktapproximation* getauft. Es folgt

$$(f(u_h), \varphi_{\nu'}) \approx \sum_{\nu \in \mathcal{N}_h} f(U_\nu)(\varphi_\nu, \varphi_{\nu'}).$$

Setzen wir jetzt $F(U) := (f(U_\nu))_{\nu \in \mathcal{N}_h}$, so lässt sich Gl. (4.21) schreiben als

$$\mathbf{K}U = \mathbf{M}F(U). \tag{4.22}$$

Ob diese Approximation in dieser Formulierung funktioniert, hängt maßgeblich davon ab, welche Basis man für $\mathcal{S}_h^p$ gewählt hat. Unter Umständen muss man noch eine Basistransformation durchführen. Wir werden beispielsweise in Abschnitt 6.2.1 sehen, dass für gewisse hierarchische Basen die Produktapproximation eine Basistransformation auf die Knotenbasis erfordert, um die gewünschte Genauigkeitsordnung zu erhalten. Denn für die Knotenbasis funktioniert die Produktapproximation ohne weitere Modifikation, siehe beispielsweise [CGMSS81] und [TM88].

4.3.2 Newton-Verfahren

Die hier gewählte Darstellung findet sich in [Kel03, Chapter 1.6]. Das *(ungedämpfte) Newton-Verfahren* zur Lösung eines Nullstellenproblems

$$G(x) = 0$$

für $G\colon \mathbb{R}^N \to \mathbb{R}^N$ besteht für $x_1 \in \mathbb{R}^N$ in der Iteration

$$x_{n+1} = x_n - G'(x_n)^{-1} G(x_n) \qquad (n \in \mathbb{N}).$$

Dabei ist vorausgesetzt, dass G' invertierbar ist (zumindest für alle x_n der Newtonfolge).

Im Fall (4.21) lautet das Nullstellenproblem

$$G(U) = \mathbf{K}U - \mathbf{M}F(U) = 0.$$

Wir wählen einen Startwert U_1 und lösen iterativ das Gleichungssystem

$$G'(U_n) U_n^{\mathrm{kor}} = -G(U_n)$$

für die Newton-Korrektur U_n^{kor} und setzen dann

$$U_{n+1} = U_n + U_n^{\mathrm{kor}}.$$

Wir müssen uns noch Gedanken darüber machen, wie man G' bestimmt. Es sei $U \in \mathbb{R}^N$ und $\mathbf{A} \in \mathbb{R}^{N \times N}$. F sei wie im vorhergehenden Abschnitt eine

Vektorfunktion mit $F(U) = (f(U_v))_{v \in \mathcal{N}_h}$. Dann gilt für $i, j = 1, \ldots, N$

$$\begin{aligned}\partial_{U_j}(\mathbf{A}F(U))_i &= \partial_{U_j}\left(\sum_{l=1}^{N} \mathbf{A}_{il} f(U_l)\right) \\ &= \sum_{l=1}^{N} \mathbf{A}_{il} \partial_{U_j} f(U_l) \\ &= \mathbf{A}_{ij} f'(U_j).\end{aligned}$$

Wir setzen $F'(U) := (f(U_v))_{v \in \mathcal{N}_h}$ und es folgt

$$\frac{\mathrm{d}}{\mathrm{d}U}(\mathbf{A}F(U)) = \begin{pmatrix} \mathbf{A}_{11} f'(U_1) & \ldots & \mathbf{A}_{1N} f'(U_N) \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{N1} f'(U_1) & \ldots & \mathbf{A}_{NN} f'(U_N) \end{pmatrix} =: \mathrm{mult}(A, F'(U)). \quad (4.23)$$

Natürlich führt man das Newtonverfahren nur für eine endliche Anzahl von Schritten durch. Wir wählen drei Abbruchkriterien. Es sei $n_{\max} \in \mathbb{N}$ die maximale Anzahl an Iterationen und $\mathrm{TOL}_U, \mathrm{TOL}_F > 0$ zwei Toleranzen. Wir brechen die Iteration ab, wenn gilt

1. $n > n_{\max}$ oder
2. $\left\|U^{\mathrm{kor}}\right\| < \mathrm{TOL}_U$ oder
3. $\|F(U_n)\| < \mathrm{TOL}_F$.

Dabei ist $\|\cdot\|$ eine beliebige Vektornorm, wir wählen immer die euklidische Norm.

cG-Verfahren für die nichtlineare Wellengleichung

5

In diesem Kapitel definieren wir zwei Petrov-Galerkin-Verfahren zur Approximation der Lösung der nichtlinearen Wellengleichung (3.6).

In Kapitel 3 haben wir aus der nichtlinearen Wellengleichung verschiedene Problemformulierungen hergeleitet, die sich in die zwei Typen quasilinearer und semilinearer Differentialgleichungen einteilen lassen. Als Ausgangspunkt der Definition der numerischen Verfahren wählen wir auf der einen Seite die quasilineare Differentialgleichung (3.12), für die wir die Existenz und Eindeutigkeit einer Lösung in Abschnitt 3.3.2 besprochen haben, und auf der anderen Seite die semilineare Differentialgleichung (3.25). Wir sprechen nachfolgend nur noch von der quasilinearen und der semilinearen Formulierung, wenn wir uns auf diese zwei Differentialgleichungen beziehen.

Im ersten Abschnitt 5.1 stellen wir die Projektionen vor, die wir im Abschnitt 5.2 zur Definition und Konvergenzanalysis eines Petrov-Galerkin-Verfahrens für die quasilineare Formulierung (3.12) der nichtlinearen Wellengleichung benötigen werden. Im Abschnitt 5.3 definieren wir ein Verfahren für die semilineare Formulierung (3.25) und besprechen, warum die Konvergenz dieses Verfahrens nicht gezeigt werden konnte. In den letzten beiden Abschnitten 5.4 und 5.5 werden beide Verfahren mit drei Beispielen getestet und die Ergebnisse diskutiert.

5.1 Vorbereitungen

5.1.1 Inverse Abschätzungen

Den Beweis der folgenden inversen Abschätzungen findet man in [Cia02, Theorem 3.2.6].

5.1 Proposition. (Inverse Abschätzungen)
Es sei $m \in \mathbb{N}_0$ und $I \subset \mathbb{R}$ ein beliebiges Intervall. Es existiert eine Konstante $C_m > 0$, die nur von m abhängt, sodass für alle $y \in \mathbb{P}_m(I)$ die inversen Abschätzungen

$$\|y\|_{L^\infty(I)} \leq C_m \frac{1}{\sqrt{|I|}} \|y\|_{L^2(I)}, \tag{5.1}$$

$$\|y_t\|_{L^2(I)} \leq C_m \frac{1}{|I|} \|y\|_{L^2(I)} \tag{5.2}$$

gelten.

5.1.2 Projektionen

In diesem Abschnitt behandeln wir die Projektionen, die wir für die Definition eines Petrov-Galerkin-Verfahrens für die quasilineare Formulierung verwenden werden und deren Approximationseigenschaften im Konvergenzbeweis zur Anwendung kommen. Wir unterscheiden dabei zwischen den Projektionen bezüglich der Zeit- und denen bezüglich der Ortsvariablen.

Im gesamten Abschnitt gelte $r, s \in \mathbb{N}_0$. Wir bemerken schon hier, dass die meisten Resultate, die wir zitieren oder zeigen, auch für allgemeinere Voraussetzungen gelten. Wir beschränken uns aber auf die Ergebnisse, die wir tatsächlich benötigen.

5.1.2.1 Projektionen in der Zeit

Es sei $T > 0$ gegeben. Wir verwenden die Notationen aus Abschnitt 4.2.

5.2 Definition.

1. Für $q \in \mathbb{N}$ bezeichnen wir mit $\mathcal{Q}_t^q \colon H^1(0,T) \to \mathcal{S}_k^q$ die Projektion definiert durch

$$\begin{aligned} \mathcal{Q}_t^q w(0) &:= w(0), \\ \left(\partial_t(\mathcal{Q}_t^q w), \partial_t \phi_k\right)_{L^2(0,T)} &= \left(\partial_t w, \partial_t \phi_k\right)_{L^2(0,T)} \qquad (\phi_k \in \mathcal{S}_k^q), \end{aligned}$$

   für $w \in H^1(0,T)$.

2. Es sei für $q \in \mathbb{N}_0$ die lokale L^2-orthogonale Projektion in der Zeit $\mathcal{Q}_n^q \colon L^2(I_n) \to \mathbb{P}_q(I_n)$ für $w \in L^2(I_n)$ gegeben durch

$$\left(\mathcal{Q}_n^q w, \phi_k\right)_{L^2(I_n)} = \left(w, \phi_k\right)_{L^2(I_n)} \qquad (\phi_k \in \mathbb{P}_q(I_n)).$$

5.3 Proposition.
Für alle $q \in \mathbb{N}$ ist $\mathcal{Q}_t^q$ interpolierend bezüglich des Zeitgitters, das heißt

$$\mathcal{Q}_t^q w(t_n) := w(t_n) \qquad (n = 0, \dots, N). \tag{5.3}$$

Daher ist es möglich, $\mathcal{Q}_t^q$ lokal auf I_n unter Vorgabe des Wertes in t_n zu definieren. Das heißt, dass die globale Definition von $\mathcal{Q}_t^q$ äquivalent ist zur Definition

$$\begin{aligned} \mathcal{Q}_t^q w(t_n) &:= w(t_n), \\ \left(\partial_t(\mathcal{Q}_t^q w), \partial_t \phi_k\right)_{L^2(I_n)} &= \left(\partial_t w, \partial_t \phi_k\right)_{L^2(I_n)} \qquad (\phi_k \in \mathbb{P}_q(I_n)), \end{aligned}$$

für alle $n = 0, \dots, N-1$.

Beweis. Wir setzen für $t \in [0, T]$

$$\phi_k(t) := t \mathbb{1}_{[0,t_n]}(t) + t_n \mathbb{1}_{[t_n,T]}(t).$$

ϕ_k ist stetig auf $[0, T]$, linear auf $[0, t_n]$ und sonst konstant. Also gilt $\phi_k \in \mathcal{S}_k^q$ für beliebiges $q \in \mathbb{N}$. Für $w \in H^1(0, T)$ ist dann

$$\begin{aligned} \mathcal{Q}_t^q w(t_n) - \mathcal{Q}_t^q w(0) &= \int_0^T \partial_t(\mathcal{Q}_t^q w)(t) \partial_t \phi_k(t) \, \mathrm{d}t = \int_0^T \partial_t w(t) \partial_t \phi_k(t) \, \mathrm{d}t \\ &= \int_0^{t_n} \partial_t w(t) \, \mathrm{d}t = w(t_n) - w(0). \end{aligned}$$

Da $\mathcal{Q}_t^q w(0) = w(0)$ nach Voraussetzung gilt, folgt die Interpolationseigenschaft. Unter Verwendung dieser Eigenschaft können wir zeigen, dass man $\mathcal{Q}_t^q$ lokal definieren kann. Wir testen mit

$$\phi_k(t) := t_n \mathbb{1}_{[0,t_n)}(t) + t \mathbb{1}_{[t_n,t_{n+1}]}(t) + t_{n+1} \mathbb{1}_{(t_{n+1},T]}(t) \qquad (t \in [0, T]).$$

ϕ_k ist stetig auf $[0, T]$, linear auf I_n und jeweils konstant auf den Intervallen $[0, t_n]$ und $[t_{n+1}, T]$, also gilt $\phi_k \in \mathcal{S}_k^q$ und die Ableitung verschwindet außerhalb von I_n. Genauer gilt für die Ableitung von ϕ_k im schwachen Sinne, dass $\phi_k'(t) = \mathbb{1}_{I_n}(t)$ ist. Zusammen mit $\mathcal{Q}_t^q w(t_n) = w(t_n)$ (dies ist dank der Interpolationseigenschaft konsistent mit der globalen Definition), können wir die Definition also auf I_n beschränken.

Definiert man andersherum $\mathcal{Q}_t^q$ lokal für alle $n = 0, \dots, N-1$, so folgt die globale Definition aus der Linearität des Integrals, der stückweisen Definition der Funktionen in $\mathcal{S}_k^q$ sowie der erzwungenen Stetigkeitsbedingung in den Gitterpunkten.

□

5.4 Lemma.

1. *Es sei $w \in H^2(I_n)$. Dann existiert ein C unabhängig von w mit*

$$\left\| w - \mathcal{Q}_t^1 w \right\|_{L^2(I_n)} + k_n \left\| w - \mathcal{Q}_t^1 w \right\|_{H^1(I_n)} \leq C k_n^2 \left\| w \right\|_{H^2(I_n)}. \tag{5.4}$$

2. *Es sei $w \in H^1(I_n)$. Dann existiert ein C unabhängig von w, aber abhängig von T mit*

$$\left\|w - \mathcal{Q}_n^0 w\right\|_{L^2(I_n)} \leq Ck_n \left\|w\right\|_{H^1(I_n)}. \tag{5.5}$$

Beweis. In Proposition 5.3 haben wir gezeigt, dass $\mathcal{Q}_t^1 w$ für $w \in H^1(I_n)$ die lineare Interpolation von w auf I_n ist. Diese erfüllt die behauptete Abschätzung, siehe beispielsweise [KA03, Theorem 3.25].

Die Abschätzung für $\mathcal{Q}_n^0$ folgt daraus, dass $\mathcal{Q}_n^0$ die L^2-orthogonale Projektion auf $\mathbb{P}_0$ ist und daher gilt

$$\left\|w - \mathcal{Q}_n^0 w\right\|_{L^2(I_n)} = \inf_{w_0 \in \mathbb{P}_0} \left\|w - w_0\right\|_{L^2(I_n)} \leq \left\|w - \mathcal{I}w\right\|_{L^2(I_n)}$$

für einen beliebigen Interpolationsoperator $\mathcal{I}\colon H^1(I_n) \to \mathbb{P}_0$. Wir können beispielsweise den Interpolationsoperator wählen, der zur Knotenauswertung im Mittelpunkt des Intervalls I_n gehört, da dieser Operator auf H^1 stetig ist. Der oben genannte Satz, [KA03, Theorem 3.25], liefert auch in diesem Fall die Behauptung. □

5.5 Bemerkung.
Gilt nur $w \in H^1(I_n)$, dann erhalten wir entsprechend die schwächere Abschätzung

$$\left\|w - \mathcal{Q}_t^1 w\right\|_{L^2(I_n)} \leq Ck_n \left\|w\right\|_{H^1(I_n)}.$$

Wir erweitern beide Definitionen von $\mathcal{Q}_t^q, \mathcal{Q}_n^q$ im L^2-Sinne auf Funktionen, die von t und x abhängen.

5.6 Definition.

1. Wir definieren $\mathcal{Q}_t^q\colon H^1(0,T;L^2(\Omega)) \to \mathcal{S}_k^q \otimes L^2(\Omega)$ durch

$$\begin{aligned} \left((\mathcal{Q}_t^q w)(0,\cdot), \phi_k\right) &= \left(w(0,\cdot), \phi_k\right) && (\phi_k \in L^2(\Omega)), \\ \int_0^T \left(\partial_t(\mathcal{Q}_t^q w), \partial_t \phi_k\right)\, \mathrm{d}t &= \int_0^T \left(\partial_t w, \partial_t \phi_k\right)\, \mathrm{d}t && (\phi_k \in \mathcal{S}_k^q \otimes L^2(\Omega)), \end{aligned}$$

   für $w \in H^1(0,T;L^2(\Omega))$.

2. Es sei $\mathcal{Q}_n^q\colon L^2(I_n;L^2(\Omega)) \to \mathcal{S}_k^q \otimes L^2(\Omega)$ für $w \in L^2(I_n;L^2(\Omega))$ definiert durch

$$\int_{I_n} \left(\mathcal{Q}_n^q w, \phi_k\right)\, \mathrm{d}t = \int_{I_n} \left(w, \phi_k\right)\, \mathrm{d}t \qquad (\phi_k \in \mathcal{S}_k^q \otimes L^2(\Omega)).$$

Mit den Ergebnissen von Lemma 5.4 können wir die Fehlerabschätzungen für die erweiterten Projektionen zeigen.

5.7 Lemma. (Projektionsfehler bzgl. Zeit)

1. *Es sei $w \in H^2(I_n; L^2(\Omega))$. Dann existiert ein C unabhängig von w mit*

$$\left\| w - \mathcal{Q}_t^1 w \right\|_{L^2(I_n;L^2(\Omega))} + k_n \left\| w - \mathcal{Q}_t^1 w \right\|_{H^1(I_n;L^2(\Omega))} \leq C k_n^2 \left\| w \right\|_{H^2(I_n;L^2(\Omega))}. \tag{5.6}$$

2. *Es sei $w \in H^1(I_n; L^2(\Omega))$. Dann existiert ein C unabhängig von w mit*

$$\left\| w - \mathcal{Q}_n^0 w \right\|_{L^2(I_n;L^2(\Omega))} \leq C k_n \left\| w \right\|_{H^1(I_n;L^2(\Omega))}. \tag{5.7}$$

Beweis. Beide Abschätzungen werden auf die gleiche Art und Weise mithilfe von Lemma 5.4 bewiesen, siehe auch den Beweis von [AM89, Lemma 2.2]. Wir beschränken uns auf die Abschätzung (5.6) und zeigen sie zunächst für Funktionen in $C^\infty(I_n \times \Omega)$. Dabei können wir uns auf ein Intervall I_n beschränken, da $\mathcal{Q}_t^q$ nach Proposition 5.3 lokal auf I_n definiert werden kann. In diesem Fall kann die Abschätzung in der Zeit punktweise im Ort angewendet werden, denn es ist

$$\begin{aligned}
\left\| w - \mathcal{Q}_t^q w \right\|_{L^2(I_n;L^2)}^2 &= \int_\Omega \left\| w(\cdot, x) - \mathcal{Q}_t^q w(\cdot, x) \right\|_{L^2(I_n)}^2 \, \mathrm{d}x \\
&\overset{(5.4)}{\leq} \int_\Omega C k_n^4 \left\| w(\cdot, x) \right\|_{H^2(I_n)}^2 \, \mathrm{d}x \\
&= C k_n^4 \left\| w \right\|_{H^2(I_n;L^2(\Omega))}^2.
\end{aligned}$$

Betrachten wir nun eine Folge $(w_i)_{i \in \mathbb{N}} \subset C^\infty(I_n \times \Omega)$, die gegen w in der $H^2(I_n; L^2(\Omega))$-Norm konvergiert, dann gilt die obige Abschätzung für jedes w_i und daher auch für den Grenzwert w. Genauso zeigt man die Abschätzung für die $H^1(I_n; L^2(\Omega))$-Norm. □

5.1.2.2 Projektion im Ort

Im Ort werden wir nur eine Projektion benötigen, die *elliptische Projektion*. Für $p \in \mathbb{N}$ sei $\mathcal{S}_h^p$ ein $H_0^1(\Omega)$-konformer Finite-Elemente-Raum mit Polynomen vom Grad $p \in \mathbb{N}$, wie wir ihn in Abschnitt 4.1 definiert haben. Dieser Raum erfülle die Bedingungen von [Cia02, Theorem 3.2.1], um hohe Approximationsordnungen zu ermöglichen. Für $d = 1$ fallen dabei keine neuen Bedingungen an, daher werden wir darauf hier nicht näher eingehen.

5.8 Definition.
Die elliptische Projektion $\mathcal{P}_E \colon H_0^1(\Omega) \to \mathcal{S}_h^p$ sei gegeben durch

$$(\nabla \mathcal{P}_E w, \nabla \varphi_h) = (\nabla w, \nabla \varphi_h) \qquad (\varphi_h \in \mathcal{S}_h^p), \tag{5.8}$$

für $w \in H_0^1(\Omega)$.

Für die elliptische Projektion gilt die folgende Approximationsabschätzung, wie wir schon in Bemerkung 4.9 angedeutet haben,

5.9 Lemma.
Es sei $1 \leq r \leq p+1$ *und* $w \in H_0^1(\Omega) \cap H^r(\Omega)$. *Dann gilt unter geeigneten Voraussetzungen an* Ω *und den verwendeten Finite-Elemente-Raum* $\mathcal{S}_h^p$ *die Abschätzung*

$$\|w - \mathcal{P}_E w\|_{H^s(\Omega)} \leq Ch^{r-s} \|w\|_{H^r(\Omega)} \tag{5.9}$$

für $s \in \{0, 1\}$.

Beweis. Siehe [Cia02, Theorem 3.2.1 und Theorem 3.2.5]. □

5.10 Bemerkung.
Die „geeigneten Voraussetzungen" reduzieren sich für $d = 1$ *auf die Forderung eines regulären adjungierten Problems. Im Fall der Poissongleichung ist das adjungierte Problem wieder eine Poissongleichung und es lässt sich zeigen, dass unter den schon getroffenen Voraussetzungen das adjungierte Problem regulär ist. Siehe dazu die im Beweis zitierten Theoreme.*

Auch diese Projektion erweitern wir zu einer Projektion in t und x im L^2-Sinne.

5.11 Definition.
Wir definieren $\mathcal{P}_E \colon H^1(I_n; H_0^1(\Omega)) \to L^2(I_n) \otimes \mathcal{S}_h^p$ für $w \in H^1(I_n; H_0^1(\Omega))$ durch

$$\int_{I_n} (\nabla \mathcal{P}_E w, \nabla \varphi_h) \, \mathrm{d}t = \int_{I_n} (\nabla w, \nabla \varphi_h) \, \mathrm{d}t \qquad (\varphi_h \in L^2(I_n) \otimes \mathcal{S}_h^p). \tag{5.10}$$

Und genau wie zuvor erhalten wir auch für diese erweiterte Projektion die Abschätzung, die wir nach Lemma 5.9 erwarten würden.

5.12 Proposition.
Es sei $1 \leq r \leq p+1$ *und* $w \in L^2(I_n; H_0^1(\Omega) \cap H^r(\Omega))$, *dann gilt*

$$\|u - \mathcal{P}_E w\|_{L^2(I_n; H^s(\Omega))} \leq Ch^{r-s} \|w\|_{L^2(I_n; H^r(\Omega))} \tag{5.11}$$

für $s \in \{0, 1\}$.

Beweis. Der Beweis wird genauso geführt wie der Beweis von Lemma 5.7. □

5.1.2.3 Weitere Eigenschaften der Projektionen

Wir werden immer wieder benötigen, dass zum einen die Projektionen bezüglich der Zeit mit der elliptischen Projektion vertauschen und dass zum anderen die Projektionen bezüglich der Zeit mit den Differentialoperatoren im Ort vertauschen, ebenso die elliptische Projektion mit den Zeitableitungen. Weiterhin brauchen wir die Stetigkeit der Projektionen in verschiedenen Normen.

5.13 Proposition. (Vertauschungseigenschaften)
Es sei $w \in L^2(0,T;L^2(\Omega))$. Die folgenden Gleichungen gelten für $q \in \mathbb{N}$, wenn sowohl die linken als auch die rechten Seiten wohldefiniert sind.

$$\partial_t \mathcal{P}_E w = \mathcal{P}_E \partial_t w, \tag{5.12}$$
$$\nabla \mathcal{Q}_t^q w = \mathcal{Q}_t^q \nabla w, \tag{5.13}$$
$$\nabla \mathcal{Q}_n^{q-1} w = \mathcal{Q}_n^{q-1} \nabla w, \tag{5.14}$$
$$\mathcal{P}_E \mathcal{Q}_n^{q-1} w = \mathcal{Q}_n^{q-1} \mathcal{P}_E w, \tag{5.15}$$
$$\mathcal{P}_E \mathcal{Q}_t^q w = \mathcal{Q}_t^q \mathcal{P}_E w. \tag{5.16}$$

Beweis. Die erste Gleichung wird in [AM89, Lemma 2.1] bewiesen. Wir führen den Beweis hier auf, weil sich die restlichen Gleichungen analog zeigen lassen.

Es sei $\phi \in H_0^1(0,T) \otimes \mathcal{S}_h^p$. Dann gilt

$$\begin{aligned}
&\int_0^T (\nabla(\mathcal{P}_E(\partial_t w)), \nabla\phi)\,\mathrm{d}t \\
&\quad = \int_0^T (\nabla(\partial_t w), \nabla\phi)\,\mathrm{d}t = \int_0^T (\partial_t(\nabla w), \nabla\phi)\,\mathrm{d}t \\
&\quad = -\int_0^T (\nabla w, \partial_t \nabla\phi)\,\mathrm{d}t = -\int_0^T (\nabla(\mathcal{P}_E w), \partial_t\nabla\phi)\,\mathrm{d}t \\
&\quad = \int_0^T (\nabla(\partial_t(\mathcal{P}_E w)), \nabla\phi)\,\mathrm{d}t.
\end{aligned}$$

Also ist

$$\mathcal{P}_E(\partial_t w) = \mathcal{P}_E(\partial_t(\mathcal{P}_E w)) = \partial_t(\mathcal{P}_E w),$$

da $\partial_t(\mathcal{P}_E w) \in L^2(0,T) \otimes \mathcal{S}_h^p$.

Die zweite Gleichung zeigt man ähnlich. Es sei $\phi \in \mathcal{S}_k^q \otimes H_0^1(\Omega)$ und $i \in \{1, \dots, d\}$, dann folgt

$$\begin{aligned}
&\int_{I_n} (\partial_t(\partial_{x_i}(\mathcal{Q}_t^q w)), \partial_t \phi) \, \mathrm{d}t \\
&\quad = \int_{I_n} (\partial_{x_i}(\partial_t(\mathcal{Q}_t^q w)), \partial_t \phi) \, \mathrm{d}t = - \int_{I_n} (\partial_t(\mathcal{Q}_t^q w), \partial_{x_i} \partial_t \phi) \, \mathrm{d}t \\
&\quad = - \int_{I_n} (\partial_t w, \partial_t \partial_{x_i} \phi) \, \mathrm{d}t = \int_{I_n} (\partial_t \partial_{x_i} w, \partial_t \phi) \, \mathrm{d}t \\
&\quad = \int_{I_n} (\partial_t(\mathcal{Q}_t^q(\partial_{x_i} w)), \partial_t \phi) \, \mathrm{d}t.
\end{aligned}$$

Also gilt $\mathcal{Q}_t^q(\partial_{x_i} w) = \mathcal{Q}_t^q(\partial_{x_i}(\mathcal{Q}_t^q w)) = \partial_{x_i}(\mathcal{Q}_t^q w)$.

Die dritte Gleichung zeigt man mit $w \in \mathcal{S}_k^{q-1}$ genauso wie die vorherige Gleichung, man ersetzt nur $(\partial_t \cdot, \partial_t \cdot)$ durch $(\cdot, \cdot)$, da $\mathcal{Q}_n^{q-1}$ die L^2-orthogonale Projektion auf $\mathcal{S}_k^{q-1} \otimes L^2(\Omega)$ ist.

Mit den Vertauschungseigenschaften der Differentialoperatoren können wir nun die Vertauschungseigenschaften zwischen Orts- und Zeitprojektionen zeigen. Für die vierte Gleichung sei $\phi \in \mathcal{S}_k^{q-1} \otimes \mathcal{S}_h^p$, dann ist

$$\begin{aligned}
&\int_{I_n} (\nabla(\mathcal{P}_E(\mathcal{Q}_n^{q-1} w)), \nabla \phi) \, \mathrm{d}t \\
&\quad = \int_{I_n} (\nabla(\mathcal{Q}_n^{q-1} w), \nabla \phi) \, \mathrm{d}t \overset{(5.14)}{=} \int_{I_n} (\mathcal{Q}_n^{q-1} \nabla w, \nabla \phi) \, \mathrm{d}t \\
&\quad = \int_{I_n} (\nabla w, \nabla \phi) \, \mathrm{d}t = \int_{I_n} (\nabla(\mathcal{P}_E w), \nabla \phi) \, \mathrm{d}t \\
&\quad = \int_{I_n} (\mathcal{Q}_n^{q-1} \nabla(\mathcal{P}_E w), \nabla \phi) \, \mathrm{d}t \overset{(5.14)}{=} \int_{I_n} (\nabla(\mathcal{Q}_n^{q-1}(\mathcal{P}_E w)), \nabla \phi) \, \mathrm{d}t.
\end{aligned}$$

Es folgt

$$\mathcal{P}_E(\mathcal{Q}_n^{q-1} w) = \mathcal{P}_E(\mathcal{Q}_n^{q-1}(\mathcal{P}_E w)) = \mathcal{Q}_n^{q-1}(\mathcal{P}_E w),$$

unter Ausnutzung von $\mathcal{Q}_n^{q-1}(\mathcal{P}_E w) \in \mathcal{S}_k^{q-1} \otimes \mathcal{S}_h^p$. Die letzte Gleichung zeigen wir wieder genauso wie zuvor, nur dass wir nun $(\partial_t \cdot, \partial_t \cdot)$ statt $(\cdot, \cdot)$ nehmen, da $\mathcal{Q}_t^q$ die orthogonale Projektion bezüglich dieses Innenproduktes ist. □

5.14 Bemerkung.
Wir benötigen die Vertauschungseigenschaften der Operatoren in den beiden Fällen $w \in H^2(0, T; H^2(\Omega))$ und $w \in \mathcal{S}_k^q \otimes \mathcal{S}_h^p$, in diesen Fällen sind sowohl die linken als auch die rechten Seiten der Gleichungen wohldefiniert und damit die Proposition anwendbar.

5.15 Proposition.
Aus der Definition der Projektionen $\mathcal{Q}_n^{q-1}$, $\mathcal{Q}_t^q$ und $\mathcal{P}_E$ erhalten wir für ausreichend glatte w die folgenden Abschätzungen,

$$\left\| \mathcal{Q}_n^{q-1} w \right\|_{L^2(I_n,L^2)} \leq \|w\|_{L^2(I_n,L^2)}, \tag{5.17}$$

$$\left\| \partial_t (\mathcal{Q}_t^q w) \right\|_{L^2(I_n,L^2)} \leq \|\partial_t w\|_{L^2(I_n,L^2)}, \tag{5.18}$$

$$\|\nabla (\mathcal{P}_E w)\|_{L^2(I_n,L^2)} \leq \|\nabla w\|_{L^2(I_n,L^2)}. \tag{5.19}$$

Ferner gilt

$$\left\| \mathcal{Q}_t^1 w \right\|_{L^2(I_n,L^2)} \leq C \left(\|w\|_{L^2(I_n,L^2)} + k_n \|\partial_t w\|_{L^2(I_n,L^2)} \right), \tag{5.20}$$

und für $\Omega \subset \mathbb{R}$

$$\|\mathcal{P}_E w\|_{L^2(I_n,L^2(\Omega))} \leq C \|w\|_{L^2(I_n,H^1(\Omega))}. \tag{5.21}$$

Beweis. Die ersten drei Ungleichungen folgen direkt aus den jeweiligen Definitionen.

Um (5.20) zu zeigen, verwenden wir die folgende skalierte Sobolevungleichung, siehe [Bur98, Theorem 4.2]. Für alle $w \in H^1(I_n)$ gilt

$$\|w\|_{L^\infty(I_n)} \leq C \left(k_n^{-1/2} \|w\|_{L^2(I_n)} + k_n^{1/2} \|\partial_t w\|_{L^2(I_n)} \right).$$

Mit dieser Ungleichung und der Interpolationseigenschaft von $\mathcal{Q}_t^q$, Proposition 5.3, erhalten wir

$$\begin{aligned} \left\| \mathcal{Q}_t^q w \right\|_{L^2(I_n)} &\leq k_n^{1/2} \left\| \mathcal{Q}_t^q w \right\|_{L^\infty(I_n)} \leq k_n^{1/2} \|w\|_{L^\infty(I_n)} \\ &\leq C k_n^{1/2} \left(k_n^{-1/2} \|w\|_{L^2(I_n)} + k_n^{1/2} \|\partial_t w\|_{L^2(I_n)} \right) \\ &= C \left(\|w\|_{L^2(I_n)} + k_n \|\partial_t w\|_{L^2(I_n)} \right). \end{aligned}$$

Dieses Ergebnis lässt sich wie zuvor auf die erweiterte Projektion übertragen. Die letzte Abschätzung (5.21) folgt direkt aus der Sobolevungleichung (2.1) und der schon behandelten Stetigkeit von $\mathcal{P}_E$ bezüglich der H_0^1-Norm. □

Weiterhin benötigen wir L^∞-Abschätzung für die Projektionen.

5.16 Proposition.
Es sei $\mathcal{H} = L^\infty(\Omega)$ oder $\mathcal{H} = H^k(\Omega)$ für $k \in \mathbb{N}_0$. Dann gilt für $w \in L^\infty(I_n; \mathcal{H})$

$$\left\| \mathcal{Q}_t^1 w \right\|_{L^\infty(I_n,\mathcal{H})} \leq C \|w\|_{L^\infty(I_n,\mathcal{H})}, \tag{5.22}$$

$$\left\| \partial_t (\mathcal{Q}_t^1 w) \right\|_{L^\infty(I_n,\mathcal{H})} \leq C \|\partial_t w\|_{L^\infty(I_n,\mathcal{H})}, \tag{5.23}$$

$$\left\| \mathcal{Q}_n^0 w \right\|_{L^\infty(I_n,\mathcal{H})} \leq C \|w\|_{L^\infty(I_n,\mathcal{H})}. \tag{5.24}$$

Für $d = 1$, $p \in \{2, \infty\}$ und $w \in L^p(I_n; L^\infty(\Omega))$ gilt außerdem

$$\|\mathcal{P}_E w\|_{L^p(I_n,L^\infty)} \leq C \|w\|_{L^p(I_n,H^1)}. \tag{5.25}$$

Beweis. Die erste Ungleichung (5.22) folgt sofort aus der Interpolationseigenschaft (5.3).

Für die nächsten beiden Abschätzungen verwenden wir die inverse Ungleichung (5.1) für Polynome und die Stetigkeit der Projektionen in der L^2-Norm. Wir zeigen nur (5.24), da (5.23) exakt genauso gezeigt wird.

$$\begin{aligned}\left\|\mathcal{Q}_n^0 w\right\|_{L^\infty(I_n,H)} &= k_n^{-1/2} \left\|\mathcal{Q}_n^0 w\right\|_{L^2(I_n,H)} \leq k_n^{-1/2} \|w\|_{L^2(I_n,H)} \\ &\leq k_n^{-1/2} k_n^{1/2} \|w\|_{L^\infty(I_n,H)} = \|w\|_{L^\infty(I_n,H)}.\end{aligned}$$

Die erste Gleichung gilt, weil $\mathcal{Q}_n^0 w$ konstant in der Zeit ist. Also erhalten wir auch die zweite Abschätzung (5.24).

Für (5.25) verwenden wir punktweise die Sobolevungleichung (2.1). □

5.2 Ein cG-Verfahren für die quasilineare Wellengleichung

In diesem Abschnitt definieren wir ein Petrov-Galerkin-Verfahren, genauer ein $\mathrm{cG}(1)\,\mathrm{cG}(p)$-Verfahren für das Modellproblem (3.6) basierend auf der quasilinearen Formulierung (3.12). Wir beschränken uns zunächst auf den Fall reeller Anfangsbedingungen und daher reeller Lösungen, und leiten eine Fehlerabschätzung her. Der Fall komplexer Anfangsbedingungen ist aufwändiger zu formulieren, vergleiche mit Abschnitt 3.3. Das numerische Verfahren lässt sich dann aber analog zum reellen Fall definieren. Wir werden darauf im Abschnitt 5.2.5.2 näher eingehen. In Abschnitt 5.4.3 präsentieren wir dann auch ein numerisches Beispiel mit einer komplexwertigen Lösung.

5.2.1 Verfahrensformulierung qlw-$\mathrm{cG}(1)\,\mathrm{cG}(p)$

Es sei $\sum_{n=0}^{N-1} k_n = T$ und es seien die Anfangswerte $u_{kh}^0, v_{kh}^0 \in \mathcal{S}_h^p$ gegeben.

Wir definieren ein $\mathrm{cG}(1)\,\mathrm{cG}(p)$-Verfahren für (3.12) wie folgt. Wir suchen Funktionen $u_{kh}, v_{kh} \in \mathcal{S}_k^1 \otimes \mathcal{S}_h^p$ mit

$$\int_0^T (\partial_t u_{kh}, \phi_1)\,\mathrm{d}t = \int_0^T (v_{kh}, \phi_1)\,\mathrm{d}t, \tag{5.26a}$$

$$\begin{aligned}\int_0^T (\mathcal{Q}_t^1(1 + f'(u_{kh}))\partial_t v_{kh}, \phi_2)\,\mathrm{d}t &= -\int_0^T (\nabla u_{kh}, \nabla\phi_2)\,\mathrm{d}t \\ &\quad - \int_0^T (\mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}, \phi_2)\,\mathrm{d}t \\ &\quad + \int_0^T (g, \phi_2)\,\mathrm{d}t,\end{aligned} \tag{5.26b}$$

für alle $\phi_1, \phi_2 \in \mathcal{S}_k^0 \otimes \mathcal{S}_h^p$ und den Anfangsbedingungen

$$u_{kh}(0,\cdot) = u_{kh}^0,$$
$$v_{kh}(0,\cdot) = v_{kh}^0.$$

Wir nennen dieses Verfahren das **qlw-cG(1) cG(p)-Verfahren**. Im Folgenden schreiben wir kurz für alle $n = 0, \dots, N-1$

$$u_{kh}^n := u_{kh}(t_n,\cdot),\ v_{kh}^n := v_k(t_n,\cdot),$$

wobei wir beachten, dass sowohl u_{kh} als auch v_{kh} stetig bezüglich der Zeitvariablen sind.

Aufgrund der fehlenden Stetigkeit der Testfunktionen in $\mathcal{S}_k^0$ können wir statt des Problems (5.26) auch eine Folge von Problemen auf den Intervallen I_n lösen. Das Lösen jedes einzelnen Teilproblems nennt man auch das Durchführen eines *Zeitschrittes*. Wir suchen in diesem Sinne $u_{kh}, v_{kh} \in \mathcal{S}_k^1 \otimes \mathcal{S}_h^p$, die für alle $n = 0, \dots, N-1$ die Gleichungen

$$\int_{I_n} (\partial_t u_{kh}, \phi_1)\,\mathrm{d}t = \int_{I_n} (v_{kh}, \phi_1)\,\mathrm{d}t, \tag{5.27a}$$

$$\begin{aligned}\int_{I_n} (\mathcal{Q}_t^1(1 + f'(u_{kh}))\partial_t v_{kh}, \phi_2)\,\mathrm{d}t = &- \int_{I_n} (\nabla u_{kh}, \nabla\phi_2)\,\mathrm{d}t \\ &- \int_{I_n} (\mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}, \phi_2)\,\mathrm{d}t \\ &+ \int_{I_n} (g, \phi_2)\,\mathrm{d}t,\end{aligned} \tag{5.27b}$$

für alle $\phi_1, \phi_2 \in \mathbb{P}_0(I_n) \otimes \mathcal{S}_h^p$ erfüllen, wobei die Werte u_{kh}^n, v_{kh}^n entweder durch den vorherigen Zeitschritt berechnet wurden ($n \geq 1$) oder durch die Anfangswerte gegeben sind ($n = 0$). Welche Werte N annehmen kann, werden wir im Lemma 5.19 über die Existenz einer Lösung dieses Gleichungssystems bestimmen, sodass T hier nicht notwendigerweise das gleiche T ist wie im Satz 3.1.

5.17 Bemerkung.
Aus (5.27) *folgt die Beziehung*

$$u_{kh}^{n+1} - u_{kh}^n = \frac{k_n}{2}(v_{kh}^{n+1} + v_{kh}^n). \tag{5.28}$$

5.2.2 Existenz und Eindeutigkeit

Zunächst möchten wir klären, was wir über Existenz und Eindeutigkeit einer Lösung von (5.27) aussagen können. Die notwendigen Bedingungen dafür werden den Bedingungen für die eindeutige Lösbarkeit ähneln, wir werden

an der entsprechenden Stelle näher darauf eingehen. Wir setzen der Einfachheit halber $g = 0$. g spielt nur bei der Existenz eine Rolle und liefert einen zusätzlichen Term, der klein genug sein muss. Um die Abschätzungen nicht noch weiter zu erschweren, vernachlässigen wir also g.

Wir erhalten durch (5.27) ein Zeitschrittverfahren

$$\left(u_{kh}^{n+1} - u_{kh}^{n}, \varphi_1\right) = \frac{k_n}{2}\left(v_{kh}^{n+1} + v_{kh}^{n}, \varphi_1\right), \tag{5.29a}$$

$$\begin{aligned}\left(v_{kh}^{n+1} - v_{kh}^{n}, \varphi_2\right) = &-\frac{k_n}{2}\left(\nabla(u_{kh}^{n+1} + u_{kh}^{n}), \nabla\varphi_2\right)\\ &-\frac{k_n}{8}\Big((f''(u_{kh}^{n+1}) + f''(u_{kh}^{n}))(v_{kh}^{n+1} + v_{kh}^{n})^2, \varphi_2\Big)\\ &-\frac{1}{2}\left((f'(u_{kh}^{n+1}) + f'(u_{kh}^{n}))(v_{kh}^{n+1} - v_{kh}^{n}), \varphi_2\right),\end{aligned} \tag{5.29b}$$

für alle $\varphi_1, \varphi_2 \in \mathcal{S}_h^p$. Dabei seien $u_{kh}^n, v_{kh}^n \in \mathcal{S}_h^p$ gegeben. Wir ersetzen ein v_{kh} in (5.27b) gemäß (5.28) durch u_{kh}, da uns der Term $(v_{kh}^{n+1} + v_{kh}^{n})^2$ vor Probleme stellen würde, die wir nicht auflösen können. Wir suchen dann die Funktionen $u_{kh}^{n+1}, v_{kh}^{n+1} \in \mathcal{S}_h^p$ als Lösung von

$$\left(u_{kh}^{n+1} - u_{kh}^{n}, \varphi_1\right) = \frac{k_n}{2}\left(v_{kh}^{n+1} + v_{kh}^{n}, \varphi_1\right), \tag{5.30a}$$

$$\begin{aligned}\left(v_{kh}^{n+1} - v_{kh}^{n}, \varphi_2\right) = &-\frac{k_n}{2}\left(\nabla(u_{kh}^{n+1} + u_{kh}^{n}), \nabla\varphi_2\right)\\ &-\frac{1}{4}\Big((f''(u_{kh}^{n+1}) + f''(u_{kh}^{n}))\\ &\qquad\qquad\times (u_{kh}^{n+1} - u_{kh}^{n})(v_{kh}^{n+1} + v_{kh}^{n}), \varphi_2\Big)\\ &-\frac{1}{2}\left((f'(u_{kh}^{n+1}) + f'(u_{kh}^{n}))(v_{kh}^{n+1} - v_{kh}^{n}), \varphi_2\right),\end{aligned} \tag{5.30b}$$

für alle $\varphi_1, \varphi_2 \in \mathcal{S}_h^p$.

Weiterhin führen wir eine Substitution durch, um die Beweise einfacher zu halten. Wir nutzen aus, dass das stetige Galerkinverfahren in der Zeit bei linearen Elementen der impliziten Mittelpunktregel entspricht, auch wenn unsere Verfahrensdefinition in den nichtlinearen Termen der Anwendung der Trapezregel entspricht. Daher setzen wir

$$u_{kh}^{n+1/2} := \frac{1}{2}(u_{kh}^{n+1} + u_{kh}^{n}), \quad v_{kh}^{n+1/2} := \frac{1}{2}(v_{kh}^{n+1} + v_{kh}^{n}).$$

Dann erhalten wir aus (5.30) das Gleichungssystem

$$\begin{aligned}
&\left(u_{kh}^{n+1/2} - u_{kh}^{n}, \varphi_1\right) \\
&\quad = \frac{k_n}{2}\left(v_{kh}^{n+1/2}, \varphi_1\right), \qquad (5.31a)\\
&\left(v_{kh}^{n+1/2} - v_{kh}^{n}, \varphi_2\right) \\
&\quad = -\frac{k_n}{2}\left(\nabla u_{kh}^{n+1/2}, \nabla\varphi_2\right) \\
&\qquad - \frac{1}{2}\left((f''(2u_{kh}^{n+1/2} - u_{kh}^{n}) + f''(u_{kh}^{n}))(u_{kh}^{n+1/2} - u_{kh}^{n})v_{kh}^{n+1/2}, \varphi_2\right) \\
&\qquad - \frac{1}{2}\left((f'(2u_{kh}^{n+1/2} - u_{kh}^{n}) + f'(u_{kh}^{n}))(v_{kh}^{n+1/2} - v_{kh}^{n}), \varphi_2\right), \qquad (5.31b)
\end{aligned}$$

für alle $\varphi_1, \varphi_2 \in \mathcal{S}_h^p$.

5.2.2.1 Existenz

Für die Existenz einer Lösung des Systems (5.31) benutzen wir den folgenden Nullstellensatz, der aus dem Brouwerschen Fixpunktsatz folgt (siehe [ADK91, Lemma 3.1]).

5.18 Satz. (Nullstellensatz)
Es sei $(\mathcal{H}, (\cdot,\cdot)_{\mathcal{H}})$ ein endlichdimensionaler Innenproduktraum mit Norm $\|\cdot\|_{\mathcal{H}} = (\cdot,\cdot)_{\mathcal{H}}^{1/2}$. Ferner sei $F\colon \mathcal{H} \to \mathcal{H}$ eine Abbildung mit den folgenden Eigenschaften.

1. *F ist stetig,*
2. *es existiert ein $\alpha > 0$, sodass für alle $z \in \mathcal{H}$ mit $\|z\|_{\mathcal{H}} = \alpha$ die Abschätzung*

$$\mathrm{Re}(F(z), z)_{\mathcal{H}} \geq 0$$

gilt.

Dann existiert ein $z^* \in \mathcal{H}$ mit $F(z^*) = 0$ und es gilt $\|z^*\|_{\mathcal{H}} \leq \alpha$.

Wir setzen $\mathcal{H} := (\mathcal{S}_h^p)^2$ und für $(u, v), (\tilde{u}, \tilde{v}) \in \mathcal{H}$ sei

$$\left(\begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix}\right)_{\mathcal{H}} := (\nabla u, \nabla\tilde{u}) + (v, \tilde{v}).$$

Im nächsten Schritt definieren wir für gegebene u_{kh}^n, v_{kh}^n und k_n die Funktion $F = (F_1, F_2)\colon \mathcal{H} \to \mathcal{H}$. Es sei für $(u, v) \in \mathcal{H}$

$$F_1(u, v) := u - u_{kh}^n - \frac{k_n}{2}v.$$

Wir definieren $\Phi\colon \mathcal{H} \to \mathcal{S}_h^p$ durch

$$\begin{aligned}(\Phi(u,v),\varphi) = \frac{k_n}{2}(\nabla u, \nabla\varphi) \\ + \frac{1}{2}\left((f''(2u-u_{kh}^n)+f''(u_{kh}^n))(u-u_{kh}^n)v,\varphi\right) \\ + \frac{1}{2}\left((f'(2u-u_{kh}^n)+f'(u_{kh}^n))(v-v_{kh}^n),\varphi\right)\end{aligned}$$

für alle $\varphi \in \mathcal{S}_h^p$ und setzen

$$F_2(u,v) := v - v_{kh}^n + \Phi(u,v).$$

Wir geben $\widetilde{\delta} \in (0,1)$ beliebig, aber fest, vor und starten mit Anfangswerten innerhalb der Kugel

$$B_{\mathcal{H}}(\widetilde{\delta}) := \{z \in \mathcal{H}\colon\ \|z\|_{\mathcal{H}} \le \widetilde{\delta}\}.$$

Wie wir in der Bemerkung 3.3 erwähnt haben, benötigen wir für die Lösbarkeit unseres Problem die Bedingung $1 + f'(u) > 0$ für alle $t \in [0,T)$ und $x \in \Omega$. Dafür wiederum müssen wir $\|u\|_{L^\infty}$ beschränkt halten, zumindest wenn $f'(u) \le 0$ gilt. Der Zeitraum, in dem das möglich ist, ist genau der Zeitraum, in dem eine Lösung existiert. Diese Bedingung werden wir in leicht abgeänderter Form auch im Beweis der Existenz benötigen.

Da wir mit H^1-konformen Elementen arbeiten, erhalten wir Kontrolle von $\|u_{kh}\|_{L^\infty}$ mittels Sobolevschem Einbettungssatz 2.1 nur in einer Raumdimension durch $\|\partial_x u_{kh}\|_2$. Das ist der Grund, warum wir uns jetzt auf eine Raumdimension beschränken. Wir verwenden also, dass wir für $z \in B_{\mathcal{H}}(\widetilde{\delta})$ mit $z = (z_1, z_2)$ die Abschätzung

$$\|z_1\|_{L^\infty} \le C_S \|\partial_x z_1\|_2 \le C_S\widetilde{\delta} =: \delta$$

zur Verfügung haben. Wir setzen

$$\begin{aligned}\left|f'\right|_{\infty,\delta} &:= \sup_{x\in B(\delta)} \left|f'(x)\right|, \\ \left|f''\right|_{\infty,\delta} &:= \sup_{x\in B(\delta)} \left|f''(x)\right|.\end{aligned}$$

f' ist auf $\mathbb{R}$ lokal Lipschitz-stetig mit Lipschitz-Konstante $L_{f',\delta}$ auf $B(\delta)$ und f'' ist als Funktion auf $\mathbb{R}$ Lipschitz-stetig mit Lipschitz-Konstante $L_{f''}$.

5.19 Lemma. (Existenz)
Es existiert ein $N \in \mathbb{N}$ so, dass das volldiskretisierte Problem (5.31) *eine Lösung (u_{kh}^n, v_{kh}^n) in $(\mathcal{S}_h^p)^2$ für alle $n = 1,\dots,N$ besitzt. Ferner existiert ein $\gamma > 0$, sodass*

$$\|\partial_x u_{kh}^n\|_2 + \|v_{kh}^n\|_2 \le \gamma \qquad (n = 0,\dots,N).$$

Beweis. Die Stetigkeit von F ist klar aufgrund der endlichdimensionalen Räume, auf denen F definiert ist.

Es sei $\delta := \left\|\partial_x u_{kh}^n\right\|_2 + \left\|v_{kh}^n\right\|_2$. Wir zeigen die zweite benötigte Eigenschaft von F des Nullstellensatzes 5.18. Es gilt mit $z = (u, v) \in \mathcal{H}$

$$\begin{aligned}
(F(z), z)_{\mathcal{H}} &= \left(\partial_x \left(u - u_{kh}^n - \frac{k_n}{2} v\right), \partial_x u\right) + (v - v_{kh}^n, v) + \frac{k_n}{2}(\partial_x u, \partial_x v) \\
&\quad + \frac{1}{2}\left((f''(2u - u_{kh}^n) + f''(u_{kh}^n))(u - u_{kh}^n)v, v\right) \\
&\quad + \frac{1}{2}\left((f'(2u - u_{kh}^n) + f'(u_{kh}^n))(v - v_{kh}^n), v\right) \\
&= \|\partial_x u\|_2^2 - (\partial_x u_{kh}^n, \partial_x u) + \|v\|_2^2 - (v_{kh}^n, v) \\
&\quad + \frac{1}{2}\left((f''(2u - u_{kh}^n) + f''(u_{kh}^n))(u - u_{kh}^n)v, v\right) \\
&\quad + \frac{1}{2}\left((f'(2u - u_{kh}^n) + f'(u_{kh}^n))(v - v_{kh}^n), v\right) \\
&\geq \|\partial_x u\|_2^2 - \|\partial_x u_{kh}^n\|_2 \|\partial_x u\|_2 + \|v\|_2^2 - \|v_{kh}^n\|_2 \|v\|_2 \\
&\quad - \frac{1}{2}\left\|f''(2u - u_{kh}^n) + f''(u_{kh}^n)\right\|_{L^\infty} \|u - u_{kh}^n\|_{L^\infty} \|v\|_2^2 \\
&\quad - \frac{1}{2}\left\|f'(2u - u_{kh}^n) + f'(u_{kh}^n)\right\|_{L^\infty} \left(\|v\|_2^2 + \|v_{kh}^n\|_2 \|v\|_2\right) \\
&= \left(\|\partial_x u\|_2 - \|\partial_x u_{kh}^n\|_2\right) \|\partial_x u\|_2 \\
&\quad + \Big(1 - \frac{1}{2}\left\|f''(2u - u_{kh}^n) + f''(u_{kh}^n)\right\|_{L^\infty} \|u - u_{kh}^n\|_{L^\infty} \\
&\qquad - \left\|f'(2u - u_{kh}^n) + f'(u_{kh}^n)\right\|_{L^\infty}\Big) \|v\|_2^2 \\
&\quad - \left(1 + \frac{1}{2}\left\|f'(2u - u_{kh}^n) + f'(u_{kh}^n)\right\|_{L^\infty}\right) \|v_{kh}^n\|_2 \|v\|_2 .
\end{aligned}$$

Wir wählen u so, dass $\|\partial_x u\|_2 = \left\|\partial_x u_{kh}^n\right\|_2$, dadurch fällt der erste Term weg. Desweiteren definieren wir

$$\begin{aligned}
A &:= 1 - \frac{1}{2}\left\|f''(2u - u_{kh}^n) + f''(u_{kh}^n)\right\|_{L^\infty} \|u - u_{kh}^n\|_{L^\infty} \\
&\quad - \left\|f'(2u - u_{kh}^n) + f'(u_{kh}^n)\right\|_{L^\infty}, \\
B &:= 1 + \frac{1}{2}\left\|f'(2u - u_{kh}^n) + f'(u_{kh}^n)\right\|_{L^\infty}.
\end{aligned}$$

Mit diesen Definitionen lautet die zu erfüllende Bedingung für die Existenz einer Nullstelle

$$A \|v\|_2 - B \left\|v_{kh}^n\right\|_2 \geq 0. \tag{5.32}$$

Das wiederum erfordert zunächst, dass A größer als Null ist. Es gilt

$$A \geq 1 - \frac{1}{2}\left|f''\right|_{\infty,2\delta} 2\delta - 2\left|f'\right|_{\infty,3\delta}. \tag{5.33}$$

Hier kommt die Lösbarkeitsbedingung ins Spiel. Wir setzen voraus, dass δ klein genug ist, damit $A \geq \frac{1}{2}$ gilt. Dies ist mindestens für $n = 0$ möglich, indem wir die Anfangswerte klein genug wählen. Es existiert also mindestens ein $N \in \mathbb{N}$, sodass $A \geq \frac{1}{2}$ für alle $n = 1, \ldots, N$ erfüllt werden kann. B ist immer positiv. Es verbleibt also, Bedingung (5.32) zu erfüllen. Es ist

$$B \leq 1 + \left|f'\right|_{\infty,3\delta}.$$

Wir wählen v so, dass gilt

$$\|v\|_2 = 2\left(1 + \left|f'\right|_{\infty,3\delta}\right)\|v_{kh}^n\|_2.$$

Dann folgt

$$\begin{aligned} A\|v\|_2 - B\|v_{kh}^n\|_2 &\geq \frac{1}{2}2\left(1 + \left|f'\right|_{\infty,3\delta}\right)\|v_{kh}^n\|_2 - \left(1 + \left|f'\right|_{\infty,3\delta}\right)\|v_{kh}^n\|_2 \\ &= 0. \end{aligned}$$

Der Nullstellensatz 5.18 liefert die Existenz einer Lösung $z^* \in \mathcal{H}$ mit der Abschätzung

$$\|z^*\|_{\mathcal{H}} \leq \|u_{kh}^n\|_{H^1} + 2\left(1 + \left|f'\right|_{\infty,3\delta}\right)\|v_{kh}^n\|_2 \leq 2\left(1 + \left|f'\right|_{\infty,3\delta}\right)\delta. \quad (5.34)$$

□

5.20 Bemerkung.
(5.34) *liefert uns eine grobe Schranke für die Norm der Anfangswerte des jeweils nächsten Zeitschrittes. Setzen wir die notwendige Lösbarkeitsbedingung des kontinuierlichen Problems an, also* $1 + f'(u) > 0$*, so muss* $|f'|_{\infty,3\delta}$ *kleiner als* 1 *bleiben. Im schlimmsten Fall vervierfacht sich also die Schranke in jedem Schritt.*

Weiterhin bemerken wir, dass in diesem Beweis keinerlei Stabilitätsbedingung in Form von Zeitschrittbeschränkungen auftauchen. Dies liegt daran, dass wir uns der Substitutionsgleichung (5.28) *bedient haben, wodurch die* k_n*-Abhängigkeit verschwunden ist. Die Anwendung der Substitution führt dazu, dass wir statt der Bedingung* $1 + f'(u) > 0$ *im Kontinuierlichen hier im Diskreten die Entsprechung zu* $1 + f'(u) + f''(u)u > 0$ *fordern. Würden wir versuchen ohne die Substitution zu arbeiten, dann würde* v_{kh} *in der dritten Potenz auftauchen. Es ist nicht klar, wie man mit diesem Term umgehen muss, da wir keine Kontrolle über* v_{kh} *in* L^∞ *haben. Für* $f' \geq 0$ *könnte man den Beweis noch geringfügig verbessern, die Ungleichung* (5.33) *lässt sich in diesem Fall zu*

$$A \geq 1 - |f''|_{\infty,2\delta}\delta$$

ändern. Das ändert aber nichts Wesentliches an der Aussage des Lemmas. Zur Bedeutung der Existenzaussage siehe Bemerkung 5.22.

5.2.2.2 Eindeutigkeit

5.21 Lemma. (Eindeutigkeit)
Das Problem (5.27) *besitze für ein n eine Lösung* (u_{kh}^n, v_{kh}^n) *und es sei* $f'(0) = 0$. *Ist für ein genügend kleines* $\gamma > 0$ *die Bedingung*

$$\|u_{kh}^n\|_{H^1(\Omega)} + \|v_{kh}^n\|_{L^2(\Omega)} \leq \gamma$$

erfüllt, so ist die Lösung eindeutig.

Beweis. Es seien $(U, V), (\widetilde{U}, \widetilde{V}) \in \mathcal{H}$ Lösungen von (5.31) mit

$$\|U\|_{H^1} + \|V\|_2 \leq \frac{\gamma}{C_S},$$
$$\left\|\widetilde{U}\right\|_{H^1} + \left\|\widetilde{V}\right\|_2 \leq \frac{\gamma}{C_S}$$

für ein $\gamma > 0$. Wir haben hier schon die Konstante C_S der Sobolevungleichung eingefügt, da dies später die Notation vereinfacht.

Wir ziehen die $(U, V), (\widetilde{U}, \widetilde{V})$ definierenden Gleichungen (5.31) voneinander ab und erhalten für die Differenzen $U - \widetilde{U}$ und $V - \widetilde{V}$ die Gleichungen

$$(U - \widetilde{U}, \varphi_1) = \frac{k_n}{2}(V - \widetilde{V}, \varphi_1), \tag{5.35a}$$

$$\begin{aligned}(V - \widetilde{V}, \varphi_2) = &-\frac{k_n}{2}(\partial_x(U - \widetilde{U}), \partial_x\varphi_2)\\ &-\frac{1}{2}\big((f''(2U - u_{kh}^n) + f''(u_{kh}^n))(U - u_{kh}^n)V\\ &\qquad - (f''(2\widetilde{U} - u_{kh}^n) + f''(u_{kh}^n))(\widetilde{U} - u_{kh}^n)\widetilde{V}, \varphi_2\big)\\ &-\frac{1}{2}\big((f'(2U - u_{kh}^n) + f'(u_{kh}^n))(V - v_{kh}^n)\\ &\qquad - (f'(2\widetilde{U} - u_{kh}^n) + f'(u_{kh}^n))(\widetilde{V} - v_{kh}^n), \varphi_2\big),\end{aligned} \tag{5.35b}$$

für alle $\varphi_1, \varphi_2 \in \mathcal{S}_h^p$. Wir testen (5.35) mit $\varphi_1 = \mathcal{A}_h(U - \widetilde{U})$, $\varphi_2 = V - \widetilde{V}$ (zur Definition des diskreten Laplace-Operators $\mathcal{A}_h$ siehe (4.18)) und addieren die beiden Gleichungen. Es folgt

$$\begin{aligned}&\left\|\partial_x(U - \widetilde{U})\right\|_2^2 + \left\|V - \widetilde{V}\right\|_2^2\\ &\qquad = -\frac{1}{2}\big((f''(2U - u_{kh}^n) + f''(u_{kh}^n))(U - u_{kh}^n)V\\ &\qquad\qquad - (f''(2\widetilde{U} - u_{kh}^n) + f''(u_{kh}^n))(\widetilde{U} - u_{kh}^n)\widetilde{V}, V - \widetilde{V}\big)\\ &\qquad\quad -\frac{1}{2}\big((f'(2U - u_{kh}^n) + f'(u_{kh}^n))(V - v_{kh}^n)\\ &\qquad\qquad - (f'(2\widetilde{U} - u_{kh}^n) + f'(u_{kh}^n))(\widetilde{V} - v_{kh}^n), V - \widetilde{V}\big)\\ &\qquad = -\frac{1}{2}(I_1 + I_2 + I_3 + I_4, V - \widetilde{V}),\end{aligned} \tag{5.36}$$

mit

$$
\begin{aligned}
I_1 &:= f''(2U - u_{kh}^n)(U - u_{kh}^n)V - f''(2\widetilde{U} - u_{kh}^n)(\widetilde{U} - u_{kh}^n)\widetilde{V},\\
I_2 &:= f''(u_{kh}^n)((U - u_{kh}^n)V - (\widetilde{U} - u_{kh}^n)\widetilde{V}),\\
I_3 &:= f'(2U - u_{kh}^n)(V - v_{kh}^n) - f'(2\widetilde{U} - u_{kh}^n)(\widetilde{V} - v_{kh}^n),\\
I_4 &:= f'(u_{kh}^n)((V - v_{kh}^n) - (\widetilde{V} - v_{kh}^n)) = f'(u_{kh}^n)(V - \widetilde{V}).
\end{aligned}
$$

Es ist

$$
\begin{aligned}
I_1 = (&f''(2U - u_{kh}^n) - f''(2\widetilde{U} - u_{kh}^n))(U - u_{kh}^n)V\\
&+ f''(2\widetilde{U} - u_{kh}^n)(U - \widetilde{U})V + f''(2\widetilde{U} - u_{kh}^n)(\widetilde{U} - u_{kh}^n)(V - \widetilde{V}).
\end{aligned}
$$

Für die einzelnen Summanden von I_1 folgt

$$
\begin{aligned}
&\Big|\Big((f''(2U - u_{kh}^n) - f''(2\widetilde{U} - u_{kh}^n))(U - u_{kh}^n)V, V - \widetilde{V}\Big)\Big|\\
&\qquad\le \left\|f''(2U - u_{kh}^n) - f''(2\widetilde{U} - u_{kh}^n)\right\|_{L^\infty} \|U - u_{kh}^n\|_{L^\infty}\\
&\qquad\quad \times \|V\|_2 \left\|V - \widetilde{V}\right\|_2\\
&\qquad\le L_{f''} \left\|U - \widetilde{U}\right\|_{L^\infty} \|U - u_{kh}^n\|_{L^\infty} \|V\|_2 \left\|V - \widetilde{V}\right\|_2\\
&\qquad\le L_{f''} \left\|U - \widetilde{U}\right\|_{H^1} 2\gamma^2 \left\|V - \widetilde{V}\right\|_2\\
&\qquad\le \gamma^2 L_{f''}\Big(\left\|U - \widetilde{U}\right\|_{H^1}^2 + \left\|V - \widetilde{V}\right\|_2^2\Big),\\
&\Big|\Big(f''(2\widetilde{U} - u_{kh}^n)(U - \widetilde{U})V, V - \widetilde{V}\Big)\Big|\\
&\qquad\le \left\|f''(2\widetilde{U} - u_{kh}^n)\right\|_{L^\infty} \left\|U - \widetilde{U}\right\|_{L^\infty} \|V\|_2 \left\|V - \widetilde{V}\right\|_2\\
&\qquad\le |f''|_{\infty,3\gamma} C_S \left\|U - \widetilde{U}\right\|_{H^1} \gamma \left\|V - \widetilde{V}\right\|_2\\
&\qquad\le \frac{1}{2} |f''|_{\infty,3\gamma} C_S \gamma\Big(\left\|U - \widetilde{U}\right\|_{H^1}^2 + \left\|V - \widetilde{V}\right\|_2^2\Big),\\
&\Big|\Big(f''(2\widetilde{U} - u_{kh}^n)(\widetilde{U} - u_{kh}^n)(V - \widetilde{V}), V - \widetilde{V}\Big)\Big|\\
&\qquad\le \left\|f''(2\widetilde{U} - u_{kh}^n)\right\|_{L^\infty} \left\|\widetilde{U} - u_{kh}^n\right\|_{L^\infty} \left\|V - \widetilde{V}\right\|_2^2\\
&\qquad\le |f''|_{\infty,3\gamma} 2\gamma \left\|V - \widetilde{V}\right\|_2^2.
\end{aligned}
$$

Insgesamt können wir $(I_1, V - \widetilde{V})$ abschätzen durch

$$
\begin{aligned}
\Big|(I_1, V - \widetilde{V})\Big| \le &\left(\gamma^2 L_{f''} + \frac{1}{2}\gamma |f''|_{\infty,3\gamma}\right) \left\|U - \widetilde{U}\right\|_{H^1}^2\\
&+ \left(\gamma^2 L_{f''} + \frac{5}{2}\gamma |f''|_{\infty,3\gamma}\right) \left\|V - \widetilde{V}\right\|_2^2. \quad (5.37)
\end{aligned}
$$

Im nächsten Schritt ist

$$I_2 = f''(u_{kh}^n)\left((U-\widetilde{U})V + (\widetilde{U}-u_{kh}^n)(V-\widetilde{V})\right),$$

und damit

$$\begin{aligned}
&\left|\left(f''(u_{kh}^n)(U-\widetilde{U})V, V-\widetilde{V}\right)\right| \\
&\quad\le |f''|_{\infty,\gamma}\left\|U-\widetilde{U}\right\|_{L^\infty}\|V\|_2\left\|V-\widetilde{V}\right\|_2 \\
&\quad\le |f''|_{\infty,\gamma} C_S\left\|U-\widetilde{U}\right\|_{H^1}\gamma\left\|V-\widetilde{V}\right\|_2 \\
&\quad\le \frac{1}{2}|f''|_{\infty,\gamma}\gamma\left(\left\|U-\widetilde{U}\right\|_{H^1}^2 + \left\|V-\widetilde{V}\right\|_2^2\right), \\
&\left|\left(f''(u_{kh}^n)(\widetilde{U}-u_{kh}^n)(V-\widetilde{V}), V-\widetilde{V}\right)\right| \\
&\quad\le |f''|_{\infty,\gamma}\left\|\widetilde{U}-u_{kh}^n\right\|_{L^\infty}\left\|V-\widetilde{V}\right\|_2^2 \\
&\quad\le 2|f''|_{\infty,\gamma}\gamma\left\|V-\widetilde{V}\right\|_2^2.
\end{aligned}$$

Daher gilt

$$\left|(I_2, V-\widetilde{V})\right| \le \frac{1}{2}|f''|_{\infty,\gamma}\gamma\left\|U-\widetilde{U}\right\|_{H^1}^2 + \frac{5}{2}|f''|_{\infty,\gamma}\gamma\left\|V-\widetilde{V}\right\|_2^2. \quad (5.38)$$

Den dritten Summanden I_3 schreiben wir als

$$I_3 = (f'(2U-u_{kh}^n) - f'(2\widetilde{U}-u_{kh}^n))(V-v_{kh}^n) + f'(2\widetilde{U}-u_{kh}^n)(V-\widetilde{V}).$$

Dann ist

$$\begin{aligned}
&\left|((f'(2U-u_{kh}^n) - f'(2\widetilde{U}-u_{kh}^n))(V-v_{kh}^n), V-\widetilde{V})\right| \\
&\qquad\le \left\|f'(2U-u_{kh}^n) - f'(2\widetilde{U}-u_{kh}^n)\right\|_{L^\infty} \\
&\qquad\quad\times \|V-v_{kh}\|_2\left\|V-\widetilde{V}\right\|_2 \\
&\qquad\le 2L_{f',3\gamma}\left\|U-\widetilde{U}\right\|_{H^1}2\gamma\left\|V-\widetilde{V}\right\|_2 \\
&\qquad\le 2L_{f',3\gamma}\gamma\left(\left\|U-\widetilde{U}\right\|_{H^1}^2 + \left\|V-\widetilde{V}\right\|_2^2\right),
\end{aligned}$$

und

$$\begin{aligned}
\left|\left(f'(2\widetilde{U}-u_{kh}^n)(V-\widetilde{V}), \widetilde{V}\right)\right| &\le \left\|f'(2\widetilde{U}-u_{kh}^n)\right\|_{L^\infty}\left\|V-\widetilde{V}\right\|_2^2 \\
&\le |f'|_{\infty,3\gamma}\left\|V-\widetilde{V}\right\|_2^2.
\end{aligned}$$

Zusammen ist

$$\left|(I_3, V-\widetilde{V})\right| \leq 2L_{f',3\gamma}\gamma \left\|U-\widetilde{U}\right\|_{H^1}^2 + (2L_{f',3\gamma}\gamma + |f'|_{\infty,3\gamma}) \left\|V-\widetilde{V}\right\|_2^2. \quad (5.39)$$

Zu guter Letzt erhalten wir für I_4

$$\left|(I_4, V-\widetilde{V})\right| \leq |f'|_{\infty,\gamma} \left\|V-\widetilde{V}\right\|_2^2. \quad (5.40)$$

Wir schätzen nun (5.36) durch (5.37) bis (5.40) ab und erhalten

$$\left\|\partial_x(U-\widetilde{U})\right\|_2^2 + \left\|V-\widetilde{V}\right\|_2^2 \leq C_U \left\|\partial_x(U-\widetilde{U})\right\|_2^2 + C_V \left\|V-\widetilde{V}\right\|_2^2,$$

mit

$$C_U = \gamma(\gamma L_{f''} + 2L_{f',3\gamma}) + \gamma |f''|_{\infty,3\gamma},$$
$$C_V = \gamma(\gamma L_{f''} + 2L_{f',3\gamma}) + 5\gamma |f''|_{\infty,3\gamma} + 2|f'|_{\infty,3\gamma}.$$

Solange γ klein genug bleibt, sind C_U und C_V kleiner als 1, dabei müssen wir $f'(0) = 0$ fordern. In diesem Fall können wir die Eindeutigkeit der Lösung folgern. □

5.22 Bemerkung.
Die Ergebnisse zu Existenz und Eindeutigkeit sind unbefriedigend, da sie nur einen Index $N \in \mathbb{N}$ liefern, bis zu dem eine eindeutige numerische Lösung existiert. Das bedeutet aber, dass t_N gegen Null gehen kann, wenn wir die Zeitschrittweite gegen Null gehen lassen. Mithilfe der verwendeten Techniken ist ein besseres Ergebnis allerdings auch nicht zu erwarten. Das hängt damit zusammen, dass einerseits u_{kh} in der H^1-Norm klein bleiben muss, andererseits die Schranke für diese Norm aber exponentiell abhängig von n wächst. Um ein Existenzintervall $[0,T]$ unabhängig von k zu erhalten, ist das gewählte Vorgehen nicht geeignet. Mithilfe des Konvergenzresultates Satz 5.32 werden wir aber in der Lage sein, ein solches Intervall zu garantieren, siehe Korollar 5.34.

5.23 Bemerkung.
Im Nachfolgenden ist der Zeitschrittindex n immer aus $\{0,\dots,N-1\}$ und wir setzen $T := t_N$.

5.2.3 Stabilität des Verfahrens

Wir zeigen, dass die Volldiskretisierung einer Stabilitätsgleichung analog zur kontinuierlichen Stabilitätsgleichung (3.16) genügt. Das Vorgehen im Beweis dieser Stabilitätsabschätzung werden wir später in der Konvergenzbetrachtung wieder benötigen.

5.24 Proposition. (Stabilität Volldiskretisierung)
Die Lösung von (5.31) *genügt für* $f(z) = \lambda z^3$ *der Gleichung*

$$\begin{aligned}\left\|\partial_x u_{kh}^{n+1}\right\|_2^2 + \left\|v_{kh}^{n+1}\right\|_2^2 + 3\lambda\left(|u_{kh}^{n+1}|^2, |v_{kh}^{n+1}|^2\right) \\ = \|\partial_x u_{kh}^n\|_2^2 + \|v_{kh}^n\|_2^2 + 3\lambda\left(|u_{kh}^n|^2, |v_{kh}^n|^2\right) \\ - \frac{3}{2}\lambda k_n\left(u_{kh}^{n+1} + u_{kh}^n, (v_{kh}^{n+1} + v_{kh}^n)v_{kh}^{n+1}v_{kh}^n\right) \\ + \int_{I_n}(g, \mathcal{Q}_n^0 v_{kh})\,\mathrm{d}t. \end{aligned} \tag{5.41}$$

Beweis. Zunächst beachten wir wieder, dass wir nicht direkt wie im Kontinuierlichen mit $\phi_1 = \mathcal{A}_h u_{kh}$, $\phi_2 = v_{kh}$ testen können, da die Testfunktionen in der Zeit eine polynomielle Ordnung niedriger besitzen müssen als die Ansatzfunktionen. Daher testen wir stattdessen mit $\phi_1 = \mathcal{A}_h \mathcal{Q}_n^0 u_{kh}$ und $\phi_2 = \mathcal{Q}_n^0 v_{kh}$. $\mathcal{A}_h$ und $\mathcal{Q}_n^0$ kommutieren nach Proposition 5.13. Es folgt

$$\int_{I_n}\frac{1}{2}\partial_t\left(\|\partial_x u_{kh}\|_2^2\right)\,\mathrm{d}t = \int_{I_n}\left(\partial_x v_{kh}, \partial_x \mathcal{Q}_n^0 u_{kh}\right)\,\mathrm{d}t \tag{5.42}$$

und

$$\begin{aligned}\int_{I_n}\frac{1}{2}\partial_t\left(\|v_{kh}\|_2^2\right)\,\mathrm{d}t = -\int_{I_n}&(\mathcal{Q}_t^1(f'(u_{kh}))\partial_t v_{kh}, \mathcal{Q}_n^0 v_{kh}) \\ &+ (\partial_x u_{kh}, \partial_x \mathcal{Q}_n^0 v_{kh}) \\ &+ \left(\mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}, \mathcal{Q}_n^0 v_{kh}\right)\,\mathrm{d}t \\ + \int_{I_n}(g, \mathcal{Q}_n^0 v_{kh})\,\mathrm{d}t.\end{aligned} \tag{5.43}$$

Da $\mathcal{Q}_n^0$ symmetrisch bezüglich des L^2-Skalarproduktes auf I_n ist, fallen die Gradiententerme bei Addition der beiden Gleichungen (5.42) und (5.43) weg. Außerdem können wir die nichtlinearen Terme explizit in Abhängigkeit der Funktionswerte in den Gitterpunkten t_n und t_{n+1} angeben,

$$\begin{aligned}&\int_{I_n}(\mathcal{Q}_t^1(f'(u_{kh}))\partial_t v_{kh} + \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}, \mathcal{Q}_n^0 v_{kh})\,\mathrm{d}t \\ &\overset{(5.28)}{=} \frac{3}{4}\lambda\Big((|u_{kh}^{n+1}|^2 + |u_{kh}^n|^2)(v_{kh}^{n+1} - v_{kh}^n) \\ &\qquad + \frac{1}{2}(u_{kh}^{n+1} + u_{kh}^n)k_n(v_{kh}^{n+1} + v_{kh}^n)^2, v_{kh}^{n+1} + v_{kh}^n\Big) \\ &= \frac{3}{4}\lambda\Big[\left(|u_{kh}^{n+1}|^2 + |u_{kh}^n|^2, |v_{kh}^{n+1}|^2 - |v_{kh}^n|^2\right) \\ &\qquad + \left(|u_{kh}^{n+1}|^2 - |u_{kh}^n|^2, (v_{kh}^{n+1} + v_{kh}^n)^2\right)\Big] \\ &= \frac{3}{2}\lambda\left[\left(|u_{kh}^{n+1}|^2, |v_{kh}^{n+1}|^2\right) - \left(|u_{kh}^n|^2, |v_{kh}^n|^2\right)\right] \\ &\qquad + \frac{3}{2}\lambda\left((u_{kh}^{n+1} + u_{kh}^n)(u_{kh}^{n+1} - u_{kh}^n), v_{kh}^{n+1}v_{kh}^n\right)\end{aligned}$$

$$\overset{(5.28)}{=} \frac{3}{2}\lambda \left[\left(|u_{kh}^{n+1}|^2, |v_{kh}^{n+1}|^2 \right) - \left(|u_{kh}^{n}|^2, |v_{kh}^{n}|^2 \right) \right] + \frac{3}{4}\lambda k_n \left(u_{kh}^{n+1} + u_{kh}^{n}, (v_{kh}^{n+1} + v_{kh}^{n}) v_{kh}^{n+1} v_{kh}^{n} \right).$$

Insgesamt folgt also

$$\begin{aligned} \left\| \partial_x u_{kh}^{n+1} \right\|_2^2 + \left\| v_{kh}^{n+1} \right\|_2^2 + 3\lambda \left(|u_{kh}^{n+1}|^2, |v_{kh}^{n+1}|^2 \right) &= \|\partial_x u_{kh}^{n}\|_2^2 + \|v_{kh}^{n}\|_2^2 + 3\lambda (|u_{kh}^{n}|^2, |v_{kh}^{n}|^2) \\ &\quad - \frac{3}{2}\lambda k_n \left(u_{kh}^{n+1} + u_{kh}^{n}, (v_{kh}^{n+1} + v_{kh}^{n}) v_{kh}^{n+1} v_{kh}^{n} \right) \\ &\quad + \int_{I_n} (g, \mathcal{Q}_n^0 v_{kh}) \, \mathrm{d}t. \end{aligned} \tag{5.44}$$

□

5.2.4 Konvergenz des Verfahrens

Bevor wir die Konvergenz des qlw-cG(1) cG(p)-Verfahrens zeigen, führen wir die in Abschnitt 4.2.2 schon angedeutete Beweisidee genauer aus. Konvergenz mit Hilfe der Energiemethode für Galerkinverfahren zeigt man für gewöhnlich wie folgt. Man wählt eine geeignete Funktion $\widetilde{u}$ im Finite-Elemente-Raum und spaltet die Differenz der Lösung u des kontinuierlichen Problems und der Lösung des Galerkinverfahrens u_h auf.

$$u - u_h = (u - \widetilde{u}) + (\widetilde{u} - u_h) =: \rho + \theta.$$

Meist wählt man für $\widetilde{u}$ eine Projektion von u auf den Finite-Elemente-Raum, für das Poisson-Problem beispielsweise die elliptische Projektion. Dann kann man Ergebnisse der Approximationstheorie benutzen, um ρ abzuschätzen. Für θ lassen sich wieder Energieargumente verwenden: Zunächst zeigt man, dass θ auch eine Differentialgleichung in Variationsform löst. Dann kann man ausnutzen, dass θ ein Element des Finite-Elemente-Raumes ist, denn jetzt kann man mit θ testen. Dies führt schlussendlich auf eine Abschätzung für θ in der Energienorm und insgesamt auf eine Fehlerabschätzung.

Für zeitabhängige Probleme mit der numerischen Lösung u_{kh} definiert man häufig zwei *Zwischenelemente*, $\widetilde{u}$ und $\widetilde{u}_{kh}$, wobei $\widetilde{u}$ eine Funktion ist, die nur in der Ortsvariablen oder nur in der Zeitvariablen diskret ist, und $\widetilde{u}_{kh}$ liegt im Finite-Elemente-Raum des volldiskretisierten Problems. Dann schreibt man den Fehler

$$u - u_{kh} = (u - \widetilde{u}) + (\widetilde{u} - \widetilde{u}_{kh}) + (\widetilde{u}_{kh} - u_{kh}) =: \rho + \eta + \theta.$$

Man findet im Wesentlichen zwei verschiedene Vorgehensweisen zur Definition von $\widetilde{u}$ und $\widetilde{u}_{kh}$.

1. Man verwendet eine Semidiskretisierung (im Ort oder in der Zeit) des Problems und setzt $\widetilde{u}$ gleich der Lösung dieses Problems. Man erhält $\widetilde{u}_{kh}$, indem man $\widetilde{u}$ auf den Finite-Elemente-Raum der Volldiskretisierung projiziert. Das beschriebene Vorgehen finden wir beispielsweise für ein nichtlineares parabolisches Problem in [Tho97, Kapitel 13] oder für die lineare Wellengleichung in [FP96].

   In diesem Fall beweist man zunächst eine Fehlerabschätzung für ρ und erhält dann mittels Approximationstheorie eine Abschätzung von η, falls $\widetilde{u}$ regulär genug ist. θ wird genauso behandelt wie im stationären Fall.

2. Man arbeitet ausschließlich mit Projektionen. $\widetilde{u}$ definiert man als Projektion im Ort oder in der Zeit von u auf den nur im Ort oder nur in der Zeit diskretisierten Raum. Und man definiert $\widetilde{u}_{kh}$ als Projektion in Ort und Zeit von u auf den Finite-Elemente-Raum. Dieses Vorgehen finden wir für eine nichtlineare Wellengleichung in [KM05] und für eine nichtlineare Schrödingergleichung in [ADK91].

   In diesem Fall argumentiert man für ρ und η ausschließlich mit einer mit Abschnitt 5.1.2 vergleichbaren Approximationstheorie. Alle Regularitätsvoraussetzungen beziehen sich dann direkt auf die Lösung u.

Für unser Problem hat sich der zweite Ansatz als zielführend erwiesen. Der erste Ansatz scheiterte daran, dass Regularitätsaussagen über die Lösung des semidiskretisierten Problems nicht einfach aus der Regularität der Lösung des kontinuierlichen Problems abgeleitet werden konnten. Das liegt daran, dass das im Ort semidiskretisierte Problem ein singulär gestörtes Problem ist (die Zeitableitung führt zu einem Faktor k, vergleiche auch mit [GR05, Kapitel 6]).

5.2.4.1 Annahmen und Notationen

Wir definieren ein Zwischenelement $(\widetilde{u}, \widetilde{v}) \in (\mathcal{S}_k^1 \otimes \mathcal{S}_h^p)^2$ durch

$$(\widetilde{u}, \widetilde{v}) := (\mathcal{P}_E \mathcal{Q}_t^1 u, \mathcal{P}_E \mathcal{Q}_t^1 v).$$

Dann teilen wir die Fehlerfunktionen gemäß

$$\begin{aligned} u - u_{kh} &= (u - \mathcal{P}_E u) + \mathcal{P}_E(u - \mathcal{Q}_t^1 u) + (\mathcal{P}_E \mathcal{Q}_t^1 u - u_{kh}) \\ &=: \rho_u + \mathcal{P}_E \eta_u + \theta_u, \qquad (5.45a) \\ v - v_{kh} &= (v - \mathcal{P}_E v) + \mathcal{P}_E(v - \mathcal{Q}_t^1 v) + (\mathcal{P}_E \mathcal{Q}_t^1 v - v_{kh}) \\ &=: \rho_v + \mathcal{P}_E \eta_v + \theta_v \qquad (5.45b) \end{aligned}$$

auf. Die ρ- und η-Terme sind Projektionsfehler und können daher mithilfe der entsprechenden Lemmata aus Abschnitt 5.1.2 abgeschätzt werden. Wir konzentrieren uns daher erst auf die θ-Terme und zeigen, dass sie ein gestörtes System partieller Differentialgleichung in Variationsformulierung lösen.

Annahme Wir treffen die folgende Annahme. Es existiere ein $C_A > 0$ und ein $C_B > 0$ so, dass für alle Diskretisierungsparameter $k, h > 0$ gilt, dass

$$\|u_{kh}\|_{L^\infty(0,T;L^\infty)} \leq C_A, \tag{5.46a}$$
$$\|v_{kh}\|_{L^\infty(0,T;L^\infty)} \leq C_B. \tag{5.46b}$$

Wir setzen also im Prinzip voraus, dass unsere Diskretisierung in L^∞ stabil ist. Da wir aber problembedingt sowieso benötigen, dass u_{kh} beschränkt bleibt, um Existenz und Eindeutigkeit der numerischen Lösung gewährleisten zu können (siehe Lemma 5.19 und Lemma 5.21), ist zumindest die erste Voraussetzung nicht weiter einschränkend. Die Bedingung an v_{kh} ist dagegen eine recht starke Bedingung, sie wird aber für die Behandlung der kubischen Nichtlinearität in v benötigt.

Notationen und technische Details Im nachfolgenden Konvergenzbeweis wiederholen sich eine Handvoll Abschätzungen mehrfach. Um die Beweisschritte nicht zu überladen, sammeln wir hier im Vorhinein diese Abschätzungen und verweisen später nur noch auf sie.

Wir verwenden für $\gamma > 0$ wie zuvor die Schreibweisen

$$\left|f'\right|_{\infty,\gamma} := \sup_{x \in B(\gamma)} \left|f'(x)\right|,$$
$$\left|f''\right|_{\infty,\gamma} := \sup_{x \in B(\gamma)} \left|f''(x)\right|.$$

Wir setzen nun

$$\gamma := \max\left\{\|u\|_{L^\infty(0,T;L^\infty)}, \|u\|_{L^\infty(0,T;H^1)}\right\}.$$

Für $\widetilde{u}$ gilt die folgende Abschätzung.

$$\begin{aligned}\|\widetilde{u}\|_{L^\infty(0,T;L^\infty)} &\overset{(2.1)}{\leq} C_S \|\partial_x \widetilde{u}\|_{L^\infty(0,T;L^2)} \overset{(5.19)}{\leq} C_S \left\|\partial_x(\mathcal{Q}_t^1 u)\right\|_{L^\infty(0,T;L^2)} \\ &\overset{(5.13),(5.22)}{\leq} C \|u\|_{L^\infty(0,T;H^1)} \leq C\gamma. \end{aligned} \tag{5.47}$$

Ferner ist für $w \in \{u, v\}$ mit der Aufspaltung (5.45)

$$\begin{aligned}\|w - \widetilde{w}\|^2_{L^2(I_n;L^2)} &\leq 2\left(\|\rho_w\|^2_{L^2(I_n;L^2)} + \|\mathcal{P}_E \eta_w\|^2_{L^2(I_n;L^2)}\right) \\ &\overset{(2.1),(5.19)}{\leq} 2\left(\|\rho_w\|^2_{L^2(I_n;L^2)} + C_S^2 \|\eta_w\|^2_{L^2(I_n;H^1)}\right) \\ &\overset{(5.11),(5.6)}{\leq} C\left(h^{2p} \|w\|^2_{L^2(I_n;H^p)} + k_n^2 \|w\|^2_{H^1(I_n;H^1)}\right) \end{aligned} \tag{5.48}$$

und

$$\begin{aligned}\|\partial_t(w-\widetilde{w})\|^2_{L^2(I_n;L^2)} &\leq 2\left(\|\partial_t\rho_w\|^2_{L^2(I_n;L^2)} + \|\mathcal{P}_E\partial_t\eta_w\|^2_{L^2(I_n;L^2)}\right)\\ &\overset{(2.1),(5.19)}{\leq} 2\left(\|\partial_t\rho_w\|^2_{L^2(I_n;L^2)} + C_S^2\,\|\partial_t\eta_w\|^2_{L^2(I_n;H^1)}\right)\\ &\overset{(5.11),(5.6)}{\leq} C\left(h^{2p}\,\|w\|^2_{H^1(I_n;H^p)} + k_n^2\,\|w\|^2_{H^2(I_n;H^1)}\right).\end{aligned} \tag{5.49}$$

5.25 Proposition. (Lokale Lipschitzstetigkeit von f)
Es seien $w_1, w_2 \in H^1(I_n;H^1)$ mit $\|w_i\|_{H^1(I_n;H^1)} \leq \gamma$ für $i = 1,2$. Dann gibt es für $f(z) = \lambda z^3$ mit $\lambda \in \mathbb{R}$ Konstanten $L_{f',\gamma}, L_{f''} > 0$ unabhängig von k_n mit

$$\left\|f'(w_1) - f'(w_2)\right\|_{L^2(I_n;L^2)} \leq L_{f',\gamma}\,\|w_1 - w_2\|_{L^2(I_n;L^2)}, \tag{5.50}$$

$$\left\|f'(w_1) - f'(w_2)\right\|^2_{L^2(I_n;H^1)} \leq C\left(L^2_{f''}\gamma^2 + L^2_{f',\gamma}\right)\|w_1 - w_2\|^2_{L^2(I_n;H^1)}, \tag{5.51}$$

und

$$\begin{aligned}&\left\|\partial_t(f'(w_1) - f'(w_2))\right\|_{L^2(I_n;L^2)}\\ &\quad\leq C_S L_{f''}\,\|\partial_t w_1\|_{L^2(I_n;L^2)}\,\|w_1 - w_2\|_{L^\infty(I_n;H^1)}\\ &\qquad + \left|f''\right|_{\infty,\gamma}\,\|\partial_t(w_1 - w_2)\|_{L^2(I_n;L^2)}.\end{aligned} \tag{5.52}$$

Beweis. Wir spalten die Differenzen geeignet auf und schätzen die einzelnen Ausdrücke dann mittels der Approximationssätze für die Projektionen ab.
Abschätzung (5.50): Es ist

$$f'(w_1) - f'(w_2) = 3\lambda(w_1^2 - w_2^2) = 3\lambda(w_1 + w_2)(w_1 - w_2).$$

Dies liefert

$$\left\|f'(w_1) - f'(w_2)\right\|_{L^2(I_n;L^2)} \leq \underbrace{\left|f''\right|_{\infty,\gamma}}_{=L_{f',\gamma}}\|w_1 - w_2\|_{L^2(I_n;L^2)}.$$

Abschätzung (5.51): Wir schreiben

$$\begin{aligned}\partial_x(f'(w_1) - f'(w_2)) &= f''(w_1)\partial_x w_1 - f''(w_2)\partial_x w_2\\ &= (f''(w_1) - f''(w_2))\partial_x w_1 + f''(w_2)\partial_x(w_1 - w_2).\end{aligned}$$

Also ist

$$\begin{aligned}&\left\|\partial_x(f'(w_1) - f'(w_2))\right\|^2_{L^2(I_n;L^2)}\\ &\qquad\leq \frac{3}{2}\Big(\left\|(f''(w_1) - f''(w_2))\partial_x w_1\right\|^2_{L^2(I_n;L^2)}\\ &\qquad\quad + \left\|f''(w_2)\partial_x(w_1 - w_2)\right\|^2_{L^2(I_n;L^2)}\Big).\end{aligned}$$

Den ersten Summanden schätzen wir wie folgt ab. Wir beachten dabei, dass f'' linear ist.

$$\begin{aligned}
&\left\|(f''(w_1)-f''(w_2))\partial_x w_1\right\|^2_{L^2(I_n;L^2)} \\
&\qquad = \int_{I_n} \left\|(f''(w_1)-f''(w_2))\partial_x w_1\right\|^2_{L^2}\,\mathrm{d}t \\
&\qquad \le \int_{I_n} \left\|f''(w_1)-f''(w_2)\right\|^2_{L^\infty} \left\|\partial_x w_1\right\|^2_{L^2}\,\mathrm{d}t \\
&\qquad \le \left\|f''(w_1)-f''(w_2)\right\|^2_{L^2(I_n;L^\infty)} \left\|\partial_x w_1\right\|^2_{L^\infty(I_n;L^2)} \\
&\qquad \le L^2_{f''} \left\|w_1-w_2\right\|^2_{L^2(I_n;L^\infty)} \left\|\partial_x w_1\right\|^2_{L^\infty(I_n;L^2)} \\
&\qquad \le L^2_{f''} C C_S^2 \left\|\partial_x(w_1-w_2)\right\|^2_{L^2(I_n;L^2)} \left\|w_1\right\|^2_{L^\infty(I_n,H^1)} \\
&\qquad \le C L^2_{f''} \gamma^2 \left\|w_1-w_2\right\|^2_{L^2(I_n;H^1)}.
\end{aligned}$$

Der zweite Summand lässt sich abschätzen durch

$$\left\|f''(w_2)\partial_x(w_1-w_2)\right\|^2_{L^2(I_n;L^2)} \le \left|f''\right|^2_{\infty,\gamma} \left\|\partial_x(w_1-w_2)\right\|^2_{L^2(I_n;L^2)}.$$

Zusammen ergibt sich

$$\left\|f'(w_1)-f'(w_2)\right\|^2_{L^2(I_n;H^1)} \le C(L^2_{f''}\gamma^2 + L^2_{f',\gamma}) \left\|w_1-w_2\right\|^2_{L^2(I_n;H^1)}.$$

Abschätzung (5.52): In diesem Fall ist

$$\begin{aligned}
\partial_t(f'(w_1)-f'(w_2)) &= f''(w_1)\partial_t w_1 - f''(w_2)\partial_t w_2 \\
&= (f''(w_1)-f''(w_2))\partial_t w_1 + f''(w_2)\partial_t(w_1-w_2).
\end{aligned}$$

Folglich ist

$$\begin{aligned}
&\left\|\partial_t(f'(w_1)-f'(w_2))\right\|_{L^2(I_n;L^2)} \\
&\qquad \le \left\|(f''(w_1)-f''(w_2))\partial_t w_1\right\|_{L^2(I_n;L^2)} \\
&\qquad\quad + \left\|f''(w_2)\partial_t(w_1-w_2)\right\|_{L^2(I_n;L^2)} \\
&\qquad \le \left\|f''(w_1)-f''(w_2)\right\|_{L^\infty(I_n;L^\infty)} \left\|\partial_t w_1\right\|_{L^2(I_n;L^2)} \\
&\qquad\quad + \left\|f''(w_2)\right\|_{L^\infty(I_n;L^\infty)} \left\|\partial_t(w_1-w_2)\right\|_{L^2(I_n;L^2)} \\
&\qquad \le C_S L_{f''} \left\|w_1-w_2\right\|_{L^\infty(I_n;H^1)} \left\|\partial_t w_1\right\|_{L^2(I_n;L^2)} \\
&\qquad\quad + \left|f''\right|_{\infty,\gamma} \left\|\partial_t(w_1-w_2)\right\|_{L^2(I_n;L^2)}.
\end{aligned}$$

□

5.2.4.2 Fehlergleichung

Das Ziel ist die Abschätzung des Fehlers (θ_u, θ_v). Zu diesem Zweck beweisen wir zunächst, dass diese Fehlerfunktionen eine gestörte Differentialgleichung in Variationsformulierung lösen. Diese Gleichungen nennen wir *Fehlergleichungen*.

5.26 Proposition. (Fehlergleichungen)
(θ_u, θ_v) lösen die Fehlergleichungen

$$\int_{I_n} (\partial_t \theta_u, \phi_1)\, \mathrm{d}t = \int_{I_n} (\theta_v, \phi_1)\, \mathrm{d}t + \int_{I_n} (\rho_v - \partial_t \rho_u + \mathcal{P}_E \eta_v, \phi_1)\, \mathrm{d}t. \tag{5.53}$$

$$\begin{aligned}\int_{I_n} (\partial_t \theta_v, \phi_2)\, \mathrm{d}t = &- \int_{I_n} (\partial_x \theta_u, \partial_x \phi_2)\, \mathrm{d}t \\ &- \int_{I_n} (f'(u)\partial_t v - \mathcal{Q}_t^1(f'(u_{kh}))\partial_t v_{kh}, \phi_2) \\ &\quad + (f''(u)v^2 - \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}, \phi_2)\, \mathrm{d}t \\ &+ \int_{I_n} (\partial_x^2 \eta_u - \partial_t \rho_v, \phi_2)\, \mathrm{d}t. \end{aligned} \tag{5.54}$$

für alle $\phi_1, \phi_2 \in \mathbb{P}_0(I_n) \otimes \mathcal{S}_h^p$.

Beweis. Wir verwenden die Definitionen der Funktionen u, v durch (3.14) sowie die Definitionen der Finite-Elemente-Funktionen u_{kh}, v_{kh} durch (5.27) und beachten, dass $\mathcal{P}_E$ und $\mathcal{Q}_t^1$ vertauschen (siehe (5.16)). Dann gilt

$$\begin{aligned}\int_{I_n} (\partial_t \theta_u, \phi_1)\, \mathrm{d}t &= \int_{I_n} (\partial_t (\mathcal{P}_E \mathcal{Q}_t^1 u - u_{kh}), \underbrace{\phi_1}_{\in \mathbb{P}_0(I_n)})\, \mathrm{d}t \\ &= \int_{I_n} (\partial_t (\mathcal{P}_E u - u_{kh}), \phi_1)\, \mathrm{d}t \\ &= \int_{I_n} (\partial_t (\underbrace{\mathcal{P}_E u - u}_{=-\rho_u}), \phi_1)\, \mathrm{d}t + \int_{I_n} (\partial_t (u - u_{kh}), \phi_1)\, \mathrm{d}t \\ &= -\int_{I_n} (\partial_t \rho_u, \phi_1)\, \mathrm{d}t + \int_{I_n} (v - v_{kh}, \phi_1)\, \mathrm{d}t \\ &= -\int_{I_n} (\partial_t \rho_u, \phi_1)\, \mathrm{d}t + \int_{I_n} (\rho_v + \mathcal{P}_E \eta_v + \theta_v, \phi_1)\, \mathrm{d}t \\ &= \int_{I_n} (\theta_v, \phi_1)\, \mathrm{d}t + \int_{I_n} (\rho_v - \partial_t \rho_u + \mathcal{P}_E \eta_v, \phi_1)\, \mathrm{d}t. \end{aligned}$$

Damit ist die erste Behauptung gezeigt. Weiter ist

$$\begin{aligned}\int_{I_n} (\partial_t \theta_v, \phi_2)\,\mathrm{d}t &= \int_{I_n} (\partial_t(\mathcal{P}_E \mathcal{Q}_t^1 v - v_{kh}), \phi_2)\,\mathrm{d}t \\ &= \int_{I_n} (\partial_t(\mathcal{P}_E v - v_{kh}), \phi_2)\,\mathrm{d}t \\ &= -\int_{I_n} (\partial_t \rho_v, \phi_2)\,\mathrm{d}t + \int_{I_n} (\partial_t(v - v_{kh}), \phi_2)\,\mathrm{d}t\end{aligned}$$

und der zweite Summand lässt sich schreiben als

$$\begin{aligned}&\int_{I_n} (\partial_t(v - v_{kh}), \phi_2)\,\mathrm{d}t \\ &\quad= -\int_{I_n} (\partial_x u, \partial_x \phi_2) + (f'(u)\partial_t v + f''(u)v^2, \phi_2)\,\mathrm{d}t \\ &\qquad+ \int_{I_n} (\partial_x u_{kh}, \partial_x \phi_2)\,\mathrm{d}t \\ &\qquad+ \int_{I_n} (\mathcal{Q}_t^1(f'(u_{kh}))\partial_t v_{kh} + \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}, \phi_2)\,\mathrm{d}t \\ &\quad= -\int_{I_n} (\partial_x(\rho_u + \eta_u + \theta_u), \partial_x \phi_2)\,\mathrm{d}t \\ &\qquad- \int_{I_n} (f'(u)\partial_t v - \mathcal{Q}_t^1(f'(u_{kh}))\partial_t v_{kh}, \phi_2)\,\mathrm{d}t \\ &\qquad- \int_{I_n} (f''(u)v^2 - \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}, \phi_2)\,\mathrm{d}t \\ &\quad= -\int_{I_n} (\partial_x \theta_u, \partial_x \phi_2)\,\mathrm{d}t - \int_{I_n} \underbrace{(\partial_x \rho_u, \partial_x \phi_2)}_{=0}\,\mathrm{d}t + \int_{I_n} (\partial_x^2 \eta_u, \phi_2)\,\mathrm{d}t \\ &\qquad- \int_{I_n} (f'(u)\partial_t v - \mathcal{Q}_t^1(f'(u_{kh}))\partial_t v_{kh}, \phi_2)\,\mathrm{d}t \\ &\qquad- \int_{I_n} (f''(u)v^2 - \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}, \phi_2)\,\mathrm{d}t.\end{aligned}$$

Zusammen lautet die zweite Fehlergleichung

$$\begin{aligned}\int_{I_n} (\partial_t \theta_v, \phi_2)\,\mathrm{d}t = &-\int_{I_n} (\partial_x \theta_u, \partial_x \phi_2)\,\mathrm{d}t + \int_{I_n} (\partial_x^2 \eta_u - \partial_t \rho_v, \phi_2)\,\mathrm{d}t \\ &- \int_{I_n} (f'(u)\partial_t v - \mathcal{Q}_t^1(f'(u_{kh}))\partial_t v_{kh}, \phi_2)\,\mathrm{d}t \\ &- \int_{I_n} (f''(u)v^2 - \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}, \phi_2)\,\mathrm{d}t.\end{aligned}$$

□

5.2.4.3 Ein Konvergenzsatz

Jetzt können wir die Fehlergleichungen benutzen, um Abschätzungen für θ_u und θ_v zu gewinnen. Wir testen die erste Fehlergleichung (5.53) mit $\phi_1 =$

$\mathcal{Q}_n^0 \mathcal{A}_h \theta_u$ und die zweite Fehlergleichung (5.54) mit $\phi_2 = \mathcal{Q}_n^0 \theta_v$ und addieren die zwei resultierenden Gleichungen. Es folgt unter Verwendung der Vertauschungseigenschaften in Proposition 5.13 und der Definition der ρ-Terme

$$\begin{aligned}
&\int_{I_n} \frac{1}{2} \partial_t \Big(\|\partial_x \theta_u\|_2^2 + \|\theta_v\|_2^2 \Big) \, \mathrm{d}t \\
&\quad = \underbrace{\int_{I_n} (\partial_x \theta_v, \partial_x \mathcal{Q}_n^0 \theta_u) - (\partial_x \theta_u, \partial_x \mathcal{Q}_n^0 \theta_v) \, \mathrm{d}t}_{=0} \\
&\qquad - \int_{I_n} (\underbrace{f'(u)\partial_t v - \mathcal{Q}_t^1(f'(u_{kh}))\partial_t v_{kh}}_{=:J_1}, \mathcal{Q}_n^0 \theta_v) \, \mathrm{d}t \\
&\qquad - \int_{I_n} (\underbrace{f''(u)v^2 - \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh}}_{=:J_2}, \mathcal{Q}_n^0 \theta_v) \, \mathrm{d}t \\
&\qquad + \int_{I_n} \underbrace{(\partial_x(\rho_v - \partial_t \rho_u), \partial_x \mathcal{Q}_n^0 \theta_u)}_{=0} \, \mathrm{d}t \\
&\qquad + \int_{I_n} (\partial_x^2 \eta_u - \partial_t \rho_v, \mathcal{Q}_n^0 \theta_v) - (\partial_x \eta_v, \partial_x \mathcal{Q}_n^0 \theta_u) \, \mathrm{d}t \\
&\quad = - \int_{I_n} (J_1 + J_2, \mathcal{Q}_n^0 \theta_v) \, \mathrm{d}t \\
&\qquad + \int_{I_n} (\partial_x^2 \eta_u - \partial_t \rho_v, \mathcal{Q}_n^0 \theta_v) - (\partial_x \eta_v, \partial_x \mathcal{Q}_n^0 \theta_u) \, \mathrm{d}t. \qquad (5.55)
\end{aligned}$$

Die nichtlinearen Terme spalten wir weiter auf, damit wir die Differenzen in J_1 und J_2 abschätzen können. Wir schreiben

$$\begin{aligned}
J_1 &= f'(u)\partial_t v - \mathcal{Q}_t^1(f'(u_{kh}))\partial_t v_{kh} \\
&= (f'(u) - \mathcal{Q}_t^1(f'(u)))\partial_t v + \mathcal{Q}_t^1(f'(u) - f'(\widetilde{u}))\partial_t v \\
&\quad + \mathcal{Q}_t^1(f'(\widetilde{u}))\partial_t(v - \widetilde{v}) + \mathcal{Q}_t^1(f'(\widetilde{u}) - f'(u_{kh}))\partial_t \widetilde{v} \\
&\quad + \mathcal{Q}_t^1(f'(u_{kh}))\partial_t \theta_v \qquad (5.56)
\end{aligned}$$

und

$$\begin{aligned}
J_2 &= f''(u)v^2 - \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0 v_{kh} \\
&= (\mathrm{Id} - \mathcal{Q}_n^0)(f''(u))v^2 + \mathcal{Q}_n^0(f''(u) - f''(\widetilde{u}))v^2 + \mathcal{Q}_n^0(f''(\widetilde{u}))(v - \widetilde{v})v \\
&\quad + \mathcal{Q}_n^0(f''(\widetilde{u}))\widetilde{v}(v - \widetilde{v}) + \mathcal{Q}_n^0(f''(\widetilde{u}))\widetilde{v}(\mathrm{Id} - \mathcal{Q}_n^0)\widetilde{v} \\
&\quad + \mathcal{Q}_n^0(f''(\widetilde{u}) - f''(u_{kh}))\widetilde{v}\mathcal{Q}_n^0\widetilde{v} + \mathcal{Q}_n^0(f''(u_{kh}))\theta_v \mathcal{Q}_n^0 \widetilde{v} \\
&\quad + \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0\theta_v. \qquad (5.57)
\end{aligned}$$

Auf den ersten Blick stellt nur der letzte Term in (5.56) noch ein Problem dar. Denn um diesen Term abschätzen zu können, müssten wir voraussetzen, dass wir $\partial_t v_{kh}$ gleichmäßig in $L^\infty(0,T;L^\infty(\Omega))$ abschätzen können. Da

diese Bedingung sehr restriktiv ist und schwierig zu begründen ist, möchten wir den Beweis ohne sie führen. Daher fehlt uns eine Abschätzung von $\partial_t \theta_v$. Allerdings erinnern wir an den Beweis der Stabilität der Volldiskretisierung, Proposition 5.24, wo wir gesehen haben, wie wir die Zeitableitung von v_{kh} mit dem zu f'' gehörigem Term verrechnen können. Der zu diesem Vorgehen passende Term ist der letzte in (5.57), wie die folgende Proposition 5.27 zeigt. Die Gleichung, die wir in dieser Proposition erhalten, ist in gewisser Weise das Kernstück des Konvergenzbeweises, da sich die nichtlinearen Terme erst unter Verwendung dieser Gleichung abschätzen lassen.

5.27 Proposition.
Es gilt

$$\begin{aligned}\int_{I_n} (\mathcal{Q}_t^1(f'(u_{kh}))\partial_t\theta_v + \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0\theta_v, \mathcal{Q}_n^0\theta_v)\,\mathrm{d}t \\ = \frac{3}{2}\lambda\left[\left(\left|u_{kh}^{n+1}\right|^2, \left|\theta_v^{n+1}\right|^2\right) - \left(|u_{kh}^n|^2, |\theta_v^n|^2\right)\right] \\ + \frac{3}{4}k_n\lambda\left((u_{kh}^{n+1}+u_{kh}^n)(v_{kh}^{n+1}+v_{kh}^n), \theta_v^{n+1}\theta_v^n\right). \quad (5.58)\end{aligned}$$

Beweis. Da die auftretenden Funktionen im Argument des Integrals allesamt im Finite-Elemente-Raum liegen, ist es uns möglich, sie explizit in Abhängigkeit der Gitterwerte anzugeben. Es gilt dank der Interpolationseigenschaft (5.3) auf dem Intervall I_n

$$\mathcal{Q}_t^1(f'(u_{kh}))(t) = f'(u_{kh}^{n+1})\psi^{n+1}(t) + f'(u_{kh}^n)\psi^n(t).$$

Dabei sind ψ^{n+1}, ψ^n die linearen Lagrange-Basisfunktionen auf I_n. Weiterhin gilt auf I_n

$$\begin{aligned}\partial_t\theta_v &= \frac{1}{k_n}(\theta_v^{n+1} - \theta_v^n),\\ \mathcal{Q}_n^0\theta_v &= \frac{1}{2}(\theta_v^{n+1} + \theta_v^n).\end{aligned}$$

Daher ist

$$\begin{aligned}\int_{I_n} (\mathcal{Q}_t^1(f'(u_{kh}))\partial_t\theta_v, \mathcal{Q}_n^0\theta_v)\,\mathrm{d}t \\ = \frac{3}{4}\lambda\left(\left|u_{kh}^{n+1}\right|^2 + |u_{kh}^n|^2, \left|\theta_v^{n+1}\right|^2 - |\theta_v^n|^2\right) \quad (5.59)\end{aligned}$$

und mit der zusätzlichen Hilfe von (5.28) zur Ersetzung von v_{kh} folgt

$$\begin{aligned}\int_{I_n} (\mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0\theta_v, \mathcal{Q}_n^0\theta_v)\,\mathrm{d}t \\ = \frac{3}{4}\lambda\left(\left|u_{kh}^{n+1}\right|^2 - |u_{kh}^n|^2, \left|\theta_v^{n+1}\right|^2 + 2\theta_v^{n+1}\theta_v^n + |\theta_v^n|^2\right). \quad (5.60)\end{aligned}$$

Summieren wir die beiden Gleichungen (5.59) und (5.60) auf, so folgt

$$\begin{aligned}
&\int_{I_n} (\mathcal{Q}_t^1(f'(u_{kh}))\partial_t\theta_v + \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0\theta_v, \mathcal{Q}_n^0\theta_v)\,\mathrm{d}t \\
&\quad= \frac{3}{2}\lambda\left[\left(\left|u_{kh}^{n+1}\right|^2, \left|\theta_v^{n+1}\right|^2\right) - \left(\left|u_{kh}^n\right|^2, \left|\theta_v^n\right|^2\right)\right] \\
&\qquad+ \frac{3}{2}\lambda\left(\left|u_{kh}^{n+1}\right|^2 - \left|u_{kh}^n\right|^2, \theta_v^{n+1}\theta_v^n\right) \\
&\quad\overset{(5.28)}{=} \frac{3}{2}\lambda\left[\left(\left|u_{kh}^{n+1}\right|^2, \left|\theta_v^{n+1}\right|^2\right) - \left(\left|u_{kh}^n\right|^2, \left|\theta_v^n\right|^2\right)\right] \\
&\qquad+ \frac{3}{4}k_n\lambda\left((u_{kh}^{n+1} + u_{kh}^n)(v_{kh}^{n+1} + v_{kh}^n), \theta_v^{n+1}\theta_v^n\right). \qquad (5.61)
\end{aligned}$$

□

5.28 Bemerkung.
Der Grund der Beschränkung auf lineare Elemente in der Zeitdiskretisierung hängt mit der gerade bewiesenen Proposition zusammen. Für ein beliebige Ordnung $q \in \mathbb{N}$ in der Zeit müsste man das Verfahren so formulieren, dass man eine zu (5.61) *analoge Gleichung erhält. Hat man eine solche Gleichung nicht zur Verfügung, so steht man, wie zuvor erwähnt, vor dem Problem $\partial_t\theta_v$ abschätzen zu müssen. Dies ist aber nicht ohne Verlust einer k_n-Ordnung möglich.*

Da es mir nicht gelungen ist, eine Verfahrensformulierung für $q \in \mathbb{N}$ zu finden, die eine solche Gleichung liefert, konnte ich nur für das Verfahren mit linearen Elementen in der Zeit Konvergenz zeigen.

5.29 Lemma.
Es seien $u, v \in H^2(0,T;H^2(\Omega)) \cap H^1(0,T;H^p(\Omega))$ und $k \in [0,K]$ für ein beliebiges $K > 0$. Dann lässt sich der Approximationsfehler (θ_u, θ_v) abschätzen durch

$$\begin{aligned}
&\|\partial_x\theta_u(T)\|_2^2 + \|\theta_v(T)\|_2^2 - \|\partial_x\theta_u(0)\|_2^2 - \|\theta_v(0)\|_2^2 \\
&\quad\le C_{u,v}\Big(h^{2p} + k^2 + \|\partial_x\theta_u\|_{L^2(0,T,L^2)}^2 + \|\theta_v\|_{L^2(0,T,L^2)}^2\Big) \\
&\qquad- 3\lambda\left[\left(\left|u_{kh}(T)\right|^2, \left|\theta_v(T)\right|^2\right) - \left(\left|u_{kh}(0)\right|^2, \left|\theta_v(0)\right|^2\right)\right] \\
&\qquad- \frac{3}{2}\lambda\sum_{n=0}^{N-1} k_n\left((u_{kh}^{n+1} + u_{kh}^n)(v_{kh}^{n+1} + v_{kh}^n), \theta_v^{n+1}\theta_v^n\right). \qquad (5.62)
\end{aligned}$$

wobei $k = \max_{n=1,\dots,N} k_n$ die maximale Zeitschrittweite bezeichnet und $C_{u,v}$ von u, v sowie K und der Nichtlinearität f abhängt.

Beweis. Unser Ausgangspunkt ist die Fehlergleichung (5.55),

$$\begin{aligned}\int_{I_n} \frac{1}{2}\partial_t \Big(\|\partial_x \theta_u\|_2^2 + \|\theta_v\|_2^2 \Big) \,\mathrm{d}t \\ = -\int_{I_n} (J_1 + J_2, \mathcal{Q}_n^0 \theta_v) \,\mathrm{d}t \\ + \int_{I_n} (\partial_x^2 \eta_u - \partial_t \rho_v, \mathcal{Q}_n^0 \theta_v) - (\partial_x \eta_v, \partial_x \mathcal{Q}_n^0 \theta_u) \,\mathrm{d}t,\end{aligned}$$

wobei J_1, J_2 durch (5.56) und (5.57) gegeben sind. Wir wiederholen deren Definitionen sobald sie gebraucht werden. Wir haben

$$\gamma := \max \left\{ \|u\|_{L^\infty(0,T;L^\infty)}, \|u\|_{L^\infty(0,T;H^1)} \right\}$$

gesetzt. Wir schätzen zuerst alle Terme auf den Teilintervallen I_n ab und summieren am Ende über alle n auf. $C_{u,v}$ bezeichne eine Konstante, die nur von u und v auf $(0,T)$ sowie K und der Nichtlinearität f abhängt, sie kann aber von Abschätzung zu Abschätzung unterschiedlich sein. Wir beginnen mit J_1 und erinnern daran, dass J_1 gegeben ist durch

$$\begin{aligned}J_1 = (f'(u) - \mathcal{Q}_t^1(f'(u)))\partial_t v + \mathcal{Q}_t^1(f'(u) - f'(\widetilde{u}))\partial_t v \\ + \mathcal{Q}_t^1(f'(\widetilde{u}))\partial_t(v - \widetilde{v}) + \mathcal{Q}_t^1(f'(\widetilde{u}) - f'(u_{kh}))\partial_t \widetilde{v} \\ + \mathcal{Q}_t^1(f'(u_{kh}))\partial_t \theta_v.\end{aligned}$$

Wir schätzen die einzelnen Terme in J_1 getrennt ab. Für den ersten Term erhalten wir mit der Cauchy-Schwarzschen und der Youngschen Ungleichung (siehe Abschnitt 2.3)

$$\begin{aligned}\Big| \int_{I_n} ((f'(u) - \mathcal{Q}_t^1(f'(u)))\partial_t v, \mathcal{Q}_n^0 \theta_v) \,\mathrm{d}t \Big| \\ \leq \|\partial_t v\|_{L^\infty(I_n;L^\infty)} \left\| (\mathrm{Id} - \mathcal{Q}_t^1)(f'(u)) \mathcal{Q}_n^0 \theta_v \right\|_{L^1(I_n;L^1)} \\ \leq \frac{1}{2} \|\partial_t v\|_{L^\infty(I_n;L^\infty)} \left(\left\| (\mathrm{Id} - \mathcal{Q}_t^1)(f'(u)) \right\|_{L^2(I_n;L^2)}^2 + \left\| \mathcal{Q}_n^0 \theta_v \right\|_{L^2(I_n;L^2)}^2 \right) \\ \overset{(5.6)}{\leq} \frac{1}{2} \|\partial_t v\|_{L^\infty(I_n;L^\infty)} \left(Ck^2 \left\| f'(u) \right\|_{H^1(I_n;L^2)}^2 + \left\| \mathcal{Q}_n^0 \theta_v \right\|_{L^2(I_n;L^2)}^2 \right) \\ \overset{(5.17)}{\leq} C_{u,v} \Big(C \left\| f'(u) \right\|_{H^1(I_n;L^2)}^2 k^2 + \|\theta_v\|_{L^2(I_n;L^2)}^2 \Big). \end{aligned} \tag{5.63}$$

Für den zweiten Summanden in J_1 müssen wir beachten, dass $\mathcal{Q}_t^1$ nicht stetig in L^2 ist. Stattdessen haben wir die Abschätzung (5.20). Es gilt

$$\begin{aligned}\left\| \mathcal{Q}_t^1(f'(u) - f'(\widetilde{u})) \right\|_{L^2(I_n;L^2)}^2 \\ \overset{(5.20)}{\leq} C \left(\left\| f'(u) - f'(\widetilde{u}) \right\|_{L^2(I_n;L^2)}^2 + k_n^2 \left\| \partial_t (f'(u) - f'(\widetilde{u})) \right\|_{L^2(I_n;L^2)}^2 \right)\end{aligned}$$

$$
\begin{aligned}
&\overset{(5.50)}{\underset{(5.52)}{\leq}} C\Big(L^2_{f',\gamma}\left\|u-\widetilde{u}\right\|^2_{L^2(I_n;L^2)} + Ck_n^2\Big(L_{f''}\left\|v\right\|^2_{L^\infty(I_n;L^\infty)}\left\|u-\widetilde{u}\right\|^2_{L^2(I_n;L^2)} \\
&\qquad + \left|f''\right|^2_{\infty,\gamma}\left\|\partial_t(u-\widetilde{u})\right\|^2_{L^2(I_n;L^2)}\Big)\Big) \\
&\overset{(5.48)}{\underset{(5.49)}{\leq}} C\Bigg(\Big(L^2_{f',\gamma} + k_n^2 L^2_{f''}\left\|v\right\|^2_{L^\infty(I_n;L^\infty)}\Big)\Big(h^{2p}\left\|u\right\|^2_{L^2(I_n;H^p)} \\
&\qquad\qquad + k_n^2\left\|u\right\|^2_{H^1(I_n;H^1)}\Big) \\
&\qquad + Ck_n^2\left|f''\right|^2_{\infty,\gamma}\Big(h^{2p}\left\|u\right\|^2_{H^1(I_n;H^p)} + k_n^2\left\|u\right\|^2_{H^2(I_n;H^1)}\Big)\Bigg) \\
&\overset{k\leq K}{\leq} C_{u,v}\Big(h^{2p}\left\|u\right\|^2_{L^2(I_n;H^p)} + k_n^2\left\|u\right\|^2_{H^2(I_n;H^1)}\Big). \qquad (5.64)
\end{aligned}
$$

Mit diesem Ergebnis können wir den zweiten Term in J_1 abschätzen.

$$
\begin{aligned}
&\Big|\int_{I_n}\left((\mathcal{Q}_t^1(f'(u)-f'(\widetilde{u})))\partial_t v, \mathcal{Q}_n^0\theta_v\right)\mathrm{d}t\Big| \\
&\leq \left\|\partial_t v\right\|_{L^\infty(I_n;L^\infty)}\left\|\mathcal{Q}_t^1(f'(u)-f'(\widetilde{u}))\mathcal{Q}_n^0\theta_v\right\|_{L^1(I_n;L^1)} \\
&\leq \frac{1}{2}\left\|\partial_t v\right\|_{L^\infty(I_n;L^\infty)}\left(\left\|\mathcal{Q}_t^1(f'(u)-f'(\widetilde{u}))\right\|^2_{L^2(I_n;L^2)} + \left\|\mathcal{Q}_n^0\theta_v\right\|^2_{L^2(I_n;L^2)}\right) \\
&\overset{(5.64)}{\leq} \frac{1}{2}\left\|\partial_t v\right\|_{L^\infty(I_n;L^\infty)}\Big(C_{u,v}\Big(h^{2p}\left\|u\right\|^2_{L^2(I_n;H^p)} + k_n^2\left\|u\right\|^2_{H^1(I_n;H^1)}\Big) \\
&\qquad + \left\|\mathcal{Q}_n^0\theta_v\right\|^2_{L^2(I_n;L^2)}\Big) \\
&\overset{(5.17)}{\leq} \frac{1}{2}\left\|\partial_t v\right\|_{L^\infty(I_n;L^\infty)}\Big(C_{u,v}\Big(h^{2p}\left\|u\right\|^2_{L^2(I_n;H^p)} + k_n^2\left\|u\right\|^2_{H^1(I_n;H^1)}\Big) \\
&\qquad + \left\|\theta_v\right\|^2_{L^2(I_n;L^2)}\Big) \\
&\leq C_{u,v}\left(h^{2p}\left\|u\right\|^2_{L^2(I_n;H^p)} + k_n^2\left\|u\right\|^2_{H^1(I_n;H^1)} + \left\|\theta_v\right\|^2_{L^2(I_n;L^2)}\right). \qquad (5.65)
\end{aligned}
$$

Der nächste abzuschätzende Term ist $\mathcal{Q}_t^1(f'(\widetilde{u}))\partial_t(v-\widetilde{v})$. Es gilt

$$
\begin{aligned}
&\Big|\int_{I_n}\left(\mathcal{Q}_t^1(f'(\widetilde{u}))\partial_t(v-\widetilde{v}), \mathcal{Q}_n^0\theta_v\right)\mathrm{d}t\Big| \\
&\leq \frac{1}{2}\left\|\mathcal{Q}_t^1(f'(\widetilde{u}))\right\|_{L^\infty(I_n;L^\infty)}\Big(\left\|\partial_t(v-\widetilde{v})\right\|^2_{L^2(I_n;L^2)} + \left\|\mathcal{Q}_n^0\theta_v\right\|^2_{L^2(I_n;L^2)}\Big) \\
&\overset{(5.22)}{\underset{(5.47)}{\leq}} \frac{1}{2}\left|f'\right|_{\infty,\gamma}\Big(\left\|\partial_t(v-\widetilde{v})\right\|^2_{L^2(I_n;L^2)} + \left\|\mathcal{Q}_n^0\theta_v\right\|^2_{L^2(I_n;L^2)}\Big) \\
&\overset{(5.49)}{\underset{(5.17)}{\leq}} \frac{1}{2}\left|f'\right|_{\infty,\gamma}\Big(Ch^{2p}\left\|v\right\|^2_{H^1(I_n;H^p)} + Ck_n^2\left\|v\right\|^2_{H^2(I_n;H^1)} + \left\|\theta_v\right\|^2_{L^2(I_n;L^2)}\Big) \\
&\leq C_{u,v}\Big(h^{2p}\left\|v\right\|^2_{H^1(I_n;H^p)} + k_n^2\left\|v\right\|^2_{H^2(I_n;H^1)} + \left\|\theta_v\right\|^2_{L^2(I_n;L^2)}\Big). \qquad (5.66)
\end{aligned}
$$

In der Abschätzung von $(\mathcal{Q}_t^1(f'(\widetilde{u}) - f'(u_{kh})))\partial_t\widetilde{v}$ nutzen wir aus, dass sowohl $\widetilde{u}$ als auch u_{kh} im Finite-Elemente-Raum liegen, also bezüglich eines Intervalls I_n Polynome sind. Diese Eigenschaft erlaubt die Anwendung einer inversen Abschätzung.

$$
\begin{aligned}
&\Big| \int_{I_n} \left((\mathcal{Q}_t^1(f'(\widetilde{u}) - f'(u_{kh})))\partial_t\widetilde{v}, \mathcal{Q}_n^0\theta_v \right) \mathrm{d}t \Big| \\
&\quad \leq \frac{1}{2} \left\| \partial_t\widetilde{v} \right\|_{L^\infty(I_n;L^2)} \Bigg(\left\| \mathcal{Q}_t^1(f'(\widetilde{u}) - f'(u_{kh})) \right\|^2_{L^2(I_n;H^1)} \\
&\qquad\qquad + \left\| \mathcal{Q}_n^0\theta_v \right\|^2_{L^2(I_n;L^2)} \Bigg) \\
&\quad \overset{(5.23)}{\underset{(5.21)}{\leq}} \frac{1}{2} \left\| \partial_t v \right\|_{L^\infty(I_n;H^1)} \Bigg(\left\| \mathcal{Q}_t^1(f'(\widetilde{u}) - f'(u_{kh})) \right\|^2_{L^2(I_n;H^1)} \\
&\qquad\qquad + \left\| \mathcal{Q}_n^0\theta_v \right\|^2_{L^2(I_n;L^2)} \Bigg) \\
&\quad \overset{(5.20)}{\leq} \frac{1}{2} \left\| \partial_t v \right\|_{L^\infty(I_n;L^2)} \Bigg(C \Big(\left\| f'(\widetilde{u}) - f'(u_{kh}) \right\|^2_{L^2(I_n;H^1)} \\
&\qquad\qquad + k_n^2 \left\| \partial_t(f'(\widetilde{u}) - f'(u_{kh})) \right\|^2_{L^2(I_n;H^1)} \Big) + \left\| \mathcal{Q}_n^0\theta_v \right\|^2_{L^2(I_n;L^2)} \Bigg) \\
&\quad \overset{(5.2)}{\leq} C \left\| \partial_t v \right\|_{L^\infty(I_n;L^2)} \Bigg(\left\| f'(\widetilde{u}) - f'(u_{kh}) \right\|^2_{L^2(I_n;H^1)} + \left\| \mathcal{Q}_n^0\theta_v \right\|^2_{L^2(I_n;L^2)} \Bigg) \\
&\quad \overset{(5.52)}{\leq} C \left\| \partial_t v \right\|_{L^\infty(I_n;L^2)} \Bigg(C(L_{f''}^2 + L_{f',\gamma}^2) \left\| \partial_x\theta_u \right\|^2_{L^2(I_n;L^2)} \\
&\qquad\qquad + \left\| \mathcal{Q}_n^0\theta_v \right\|^2_{L^2(I_n;L^2)} \Bigg) \\
&\quad \overset{(5.17)}{\leq} C_{u,v} \left(\left\| \partial_x\theta_u \right\|^2_{L^2(I_n;L^2)} + \left\| \theta_v \right\|^2_{L^2(I_n;L^2)} \right). \qquad (5.67)
\end{aligned}
$$

Den letzten Term $\mathcal{Q}_t^1(f'(u_{kh}))\partial_t\theta_v$ in J_1 verrechnen wir mithilfe von Proposition 5.27 mit dem entsprechenden Term in J_2, also mit $\mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0\theta_v$. Wir fahren mit den Termen in J_2 fort und schätzen die zugehörigen Integrale wie zuvor für J_1 ab. Es ist

$$
\begin{aligned}
J_2 = {} & (\mathrm{Id} - \mathcal{Q}_n^0)(f''(u))v^2 + \mathcal{Q}_n^0(f''(u) - f''(\widetilde{u}))v^2 + \mathcal{Q}_n^0(f''(\widetilde{u}))(v - \widetilde{v})v \\
& + \mathcal{Q}_n^0(f''(\widetilde{u}))\widetilde{v}(v - \widetilde{v}) + \mathcal{Q}_n^0(f''(\widetilde{u}))\widetilde{v}(\mathrm{Id} - \mathcal{Q}_n^0)\widetilde{v} \\
& + \mathcal{Q}_n^0(f''(\widetilde{u}) - f''(u_{kh}))\widetilde{v}\mathcal{Q}_n^0\widetilde{v} + \mathcal{Q}_n^0(f''(u_{kh}))\theta_v\mathcal{Q}_n^0\widetilde{v} \\
& + \mathcal{Q}_n^0(f''(u_{kh}))v_{kh}\mathcal{Q}_n^0\theta_v.
\end{aligned}
$$

Wir beginnen mit

$$\begin{aligned}
&\Big| \int_{I_n} \left((\mathrm{Id} - \mathcal{Q}_n^0)(f''(u))v^2, \mathcal{Q}_n^0 \theta_v \right) \mathrm{d}t \Big| \\
&\quad \leq \frac{1}{2} \left\| v^2 \right\|_{L^\infty(I_n;L^\infty)} \left(\left\| (\mathrm{Id} - \mathcal{Q}_n^0)(f''(u)) \right\|_{L^2(I_n;L^2)}^2 + \left\| \mathcal{Q}_n^0 \theta_v \right\|_{L^2(I_n;L^2)}^2 \right) \\
&\quad \overset{(5.7)}{\leq} \frac{1}{2} \left\| v \right\|_{L^\infty(I_n;L^\infty)}^2 \left(k_n^2 \left\| \partial_t f''(u) \right\|_{L^2(I_n;L^2)}^2 + \left\| \theta_v \right\|_{L^2(I_n;L^2)}^2 \right) \\
&\quad = C_{u,v} \left(k_n^2 \left\| \partial_t f''(u) \right\|_{L^2(I_n;L^2)}^2 + \left\| \theta_v \right\|_{L^2(I_n;L^2)}^2 \right). \qquad (5.68)
\end{aligned}$$

Für den zweiten Summanden in J_2 erhalten wir

$$\begin{aligned}
&\Big| \int_{I_n} \left(\mathcal{Q}_n^0(f''(u) - f''(\widetilde{u}))v^2, \mathcal{Q}_n^0 \theta_v \right) \mathrm{d}t \Big| \\
&\quad \leq \frac{1}{2} \left\| v^2 \right\|_{L^\infty(I_n;L^\infty)} \Big(\left\| \mathcal{Q}_n^0(f''(u) - f''(\widetilde{u})) \right\|_{L^2(I_n;L^2)}^2 \\
&\qquad\qquad + \left\| \mathcal{Q}_n^0 \theta_v \right\|_{L^2(I_n;L^2)}^2 \Big) \\
&\quad \overset{(5.17)}{\leq} \frac{1}{2} \left\| v \right\|_{L^\infty(I_n;L^\infty)}^2 \left(L_{f''} \left\| u - \widetilde{u} \right\|_{L^2(I_n;L^2)}^2 + \left\| \theta_v \right\|_{L^2(I_n;L^2)}^2 \right) \\
&\quad \overset{(5.48)}{\leq} \frac{1}{2} \left\| v \right\|_{L^\infty(I_n;L^\infty)}^2 \Big(C L_{f''} \left(h^2 \left\| u \right\|_{L^2(I_n;H^1)}^2 + k_n^2 \left\| u \right\|_{H^1(I_n;H^1)}^2 \right) \\
&\qquad\qquad + \left\| \theta_v \right\|_{L^2(I_n;L^2)}^2 \Big) \\
&\quad \leq C_{u,v} \left(h^2 \left\| u \right\|_{L^2(I_n;H^1)}^2 + k_n^2 \left\| u \right\|_{H^1(I_n;H^1)}^2 + \left\| \theta_v \right\|_{L^2(I_n;L^2)}^2 \right). \qquad (5.69)
\end{aligned}$$

Desweiteren ist

$$\begin{aligned}
&\Big| \int_{I_n} \left(\mathcal{Q}_n^0(f''(\widetilde{u}))(v - \widetilde{v})v, \mathcal{Q}_n^0 \theta_v \right) \mathrm{d}t \Big| \\
&\quad \leq \frac{1}{2} \left\| \mathcal{Q}_n^0(f''(\widetilde{u})) \right\|_{L^\infty(I_n;L^\infty)} \left\| v \right\|_{L^\infty(I_n;L^\infty)} \\
&\qquad\qquad \times \left(\left\| v - \widetilde{v} \right\|_{L^2(I_n;L^2)}^2 + \left\| \mathcal{Q}_n^0 \theta_v \right\|_{L^2(I_n;L^2)}^2 \right) \\
&\quad \overset{(5.24)}{\leq} \left| f'' \right|_{\infty,\gamma} \left\| v \right\|_{L^\infty(I_n;L^\infty)} \left(\left\| v - \widetilde{v} \right\|_{L^2(I_n;L^2)}^2 + \left\| \mathcal{Q}_n^0 \theta_v \right\|_{L^2(I_n;L^2)}^2 \right) \\
&\quad \underset{(5.17)}{\overset{(5.48)}{\leq}} C_{u,v} \left(h^{2p} \left\| v \right\|_{L^2(I_n;H^p)}^2 + k_n^2 \left\| v \right\|_{H^1(I_n;H^1)}^2 + \left\| \theta_v \right\|_{L^2(I_n;L^2)}^2 \right). \qquad (5.70)
\end{aligned}$$

Der nächste Term unterscheidet sich nur durch die Funktion $\widetilde{v}$ statt dem letzten v vom vorherigen Ausdruck, und es gilt

$$\left\| \widetilde{v} \right\|_{L^\infty(I_n;L^\infty)} \leq C \left\| v \right\|_{L^\infty(I_n;L^\infty)}$$

aufgrund der L^∞-Stabilität der Projektionen $\mathcal{Q}_t^1, \mathcal{P}_E$ (siehe (5.22) und (5.25)), daher ändert sich die Abschätzung nicht und es gilt

$$\begin{aligned}
&\Big| \int_{I_n} \left(\mathcal{Q}_n^0(f''(\widetilde{u}))(v-\widetilde{v})\widetilde{v}, \mathcal{Q}_n^0\theta_v\right) \mathrm{d}t\Big) \\
&\qquad \leq C_{u,v}\left(h^{2p}\left\|v\right\|^2_{L^2(I_n;H^p)} + k_n^2\left\|v\right\|^2_{H^1(I_n;H^1)} + \left\|\theta_v\right\|^2_{L^2(I_n;L^2)}\right). \quad (5.71)
\end{aligned}$$

Der fünfte Summand in J_2 führt uns zu

$$\begin{aligned}
&\Big| \int_{I_n} \left(\mathcal{Q}_n^0(f''(\widetilde{u}))\widetilde{v}(\mathrm{Id}-\mathcal{Q}_n^0)\widetilde{v}, \mathcal{Q}_n^0\theta_v\right) \mathrm{d}t\Big| \\
&\quad \leq \frac{1}{2}\left\|\mathcal{Q}_n^0(f''(\widetilde{u}))\right\|_{L^\infty(I_n;L^\infty)}\left\|\widetilde{v}\right\|_{L^\infty(I_n;L^\infty)} \\
&\qquad\qquad \times\left(\left\|(\mathrm{Id}-\mathcal{Q}_n^0)\widetilde{v}\right\|^2_{L^2(I_n;L^2)} + \left\|\mathcal{Q}_n^0\theta_v\right\|^2_{L^2(I_n;L^2)}\right) \\
&\quad \overset{(5.7)}{\leq} C_{u,v}\left(Ck_n^2\left\|\partial_t\widetilde{v}\right\|^2_{L^2(I_n;L^2)} + \left\|\theta_v\right\|^2_{L^2(I_n;L^2)}\right) \\
&\quad \underset{(5.21)}{\overset{(5.18)}{\leq}} C_{u,v}\left(Ck_n^2\left\|v\right\|^2_{H^1(I_n;H^1)} + \left\|\theta_v\right\|^2_{L^2(I_n;L^2)}\right). \quad (5.72)
\end{aligned}$$

Für den Term $\mathcal{Q}_n^0(f''(u_{kh}))\theta_v\mathcal{Q}_n^0\widetilde{v}$ erhalten wir wieder unter Verwendung der L^2-Stabilität von $\mathcal{Q}_n^0$ (siehe (5.17)) und der L^∞-Stabilität der Projektionen $\mathcal{Q}_t^1$ und $\mathcal{P}_E$ (siehe (5.22) und (5.25))

$$\begin{aligned}
&\Big| \int_{I_n} \left(\mathcal{Q}_n^0(f''(\widetilde{u}) - f''(u_{kh}))\widetilde{v}\mathcal{Q}_n^0\widetilde{v}, \mathcal{Q}_n^0\theta_v\right) \mathrm{d}t\Big| \\
&\qquad \leq \frac{1}{2}\left\|\widetilde{v}\right\|_{L^\infty(I_n;L^\infty)}\left\|\mathcal{Q}_n^0\widetilde{v}\right\|_{L^\infty(I_n;L^\infty)} \\
&\qquad\qquad \times\left(\left\|\mathcal{Q}_n^0(f''(\widetilde{u}) - f''(u_{kh}))\right\|^2_{L^2(I_n;L^2)} + \left\|\mathcal{Q}_n^0\theta_v\right\|^2_{L^2(I_n;L^2)}\right) \\
&\qquad \leq \frac{1}{2}C\left\|v\right\|^2_{L^\infty(I_n;H^1)}\left(C\left\|f''(\widetilde{u}) - f''(u_{kh})\right\|^2_{L^2(I_n;L^2)} + \left\|\mathcal{Q}_n^0\theta_v\right\|^2_{L^2(I_n;L^2)}\right) \\
&\qquad \leq \frac{1}{2}C\left\|v\right\|^2_{L^\infty(I_n;H^1)}\left(L_{f''}\left\|\theta_u\right\|^2_{L^2(I_n;L^2)} + \left\|\theta_v\right\|^2_{L^2(I_n;L^2)}\right) \\
&\qquad \overset{(2.2)}{\leq} C_{u,v}\left(\left\|\partial_x\theta_u\right\|^2_{L^2(I_n;L^2)} + \left\|\theta_v\right\|^2_{L^2(I_n;L^2)}\right). \quad (5.73)
\end{aligned}$$

Für den vorletzten Summanden in J_2 benötigen wir die gleichen Stabilitätseigenschaften der Projektionen wie bei der vorherigen Abschätzung.

$$\begin{aligned}
&\Big| \int_{I_n} \left(\mathcal{Q}_n^0(f''(u_{kh}))\theta_v \mathcal{Q}_n^0\widetilde{v}, \mathcal{Q}_n^0\theta_v\right) \mathrm{d}t \Big| \\
&\quad \leq \frac{1}{2} \left\| \mathcal{Q}_n^0(f''(u_{kh})) \right\|_{L^\infty(I_n;L^\infty)} \left\| \mathcal{Q}_n^0\widetilde{v} \right\|_{L^\infty(I_n;L^\infty)} \\
&\qquad\qquad \times \left(\|\theta_v\|^2_{L^2(I_n;L^2)} + \left\| \mathcal{Q}_n^0\theta_v \right\|^2_{L^2(I_n;L^2)} \right) \\
&\quad \leq \frac{1}{2} L_{f''} \underbrace{\|u_{kh}\|_{L^\infty(I_n;L^\infty)}}_{\leq C_A} \|v\|_{L^\infty(I_n;L^\infty)} \left(\|\theta_v\|^2_{L^2(I_n;L^2)} + \|\theta_v\|^2_{L^2(I_n;L^2)} \right) \\
&\quad \leq C_{u,v} \|\theta_v\|^2_{L^2(I_n;L^2)} .
\end{aligned} \tag{5.74}$$

Der letzte Summand in J_2 wird, wie zuvor erwähnt, mit dem letzten Summand in J_1 gemäß Proposition 5.27 verrechnet. Es verbleibt das letzte Integral in (5.55), dafür erhalten wir

$$\begin{aligned}
&\Big| \int_{I_n} (\partial_x^2\eta_u - \partial_t\rho_v, \mathcal{Q}_n^0\theta_v) - (\partial_x\eta_v, \partial_x\mathcal{Q}_n^0\theta_u) \, \mathrm{d}t \Big| \\
&\quad \leq \frac{1}{2} \Big(\left\|\partial_x^2\eta_u\right\|^2_{L^2(I_n,L^2)} + \|\partial_t\rho_v\|^2_{L^2(I_n,L^2)} + 2 \left\| \mathcal{Q}_n^0\theta_v \right\|^2_{L^2(I_n,L^2)} \\
&\qquad + \|\partial_x\eta_v\|^2_{L^2(I_n,L^2)} + \left\| \mathcal{Q}_n^0\partial_x\theta_u \right\|^2_{L^2(I_n,L^2)} \Big) \\
&\quad \overset{\substack{(5.6)\\(5.11)}}{\leq} \frac{1}{2} C \Big(k_n^2 \left\|\partial_x^2 u\right\|^2_{H^1(I_n,L^2)} + h^{2p} \|\partial_t v\|^2_{L^2(I_n,H^p)} + 2\|\theta_v\|^2_{L^2(I_n,L^2)} \\
&\qquad + k_n^2 \|\partial_x v\|^2_{L^2(I_n,L^2)} + \|\partial_x\theta_u\|^2_{L^2(I_n,L^2)} \Big) \\
&\quad \leq C_{u,v} \Big(k_n^2 \Big(\|u\|^2_{H^1(I_n;H^2)} + \|v\|^2_{L^2(I_n;H^1)} \Big) + h^{2p} \|v\|^2_{H^1(I_n;H^p)} \\
&\qquad + \|\partial_x\theta_u\|^2_{L^2(I_n;L^2)} + \|\theta_v\|^2_{L^2(I_n;L^2)} \Big).
\end{aligned} \tag{5.75}$$

Wir schätzen nun die linke Seite von (5.55) durch die Summe der rechten Seiten der Ungleichungen (5.63) und (5.65) bis (5.75) ab und wenden die Pro-

position 5.27 an.

$$\begin{aligned}
&\int_{I_n} \frac{1}{2}\partial_t \Big(\|\partial_x \theta_u\|_2^2 + \|\theta_v\|_2^2 \Big) \,\mathrm{d}t \\
&\quad \leq C_{u,v} \Bigg((h^{2p} + k_n^2) \Big(\|u\|^2_{H^2(I_n;H^2)} + \|v\|^2_{H^2(I_n;H^2)} + \|u\|^2_{H^1(I_n;H^p)} \\
&\qquad\qquad + \|v\|^2_{H^1(I_n;H^p)} \Big) + \|\partial_x \theta_u\|^2_{L^2(I_n,L^2)} + \|\theta_v\|^2_{L^2(I_n,L^2)} \Bigg) \\
&\quad - \frac{3}{2}\lambda \left[\Big(|u_{kh}^{n+1}|^2, |\theta_v^{n+1}|^2 \Big) - \Big(|u_{kh}^n|^2, |\theta_v^n|^2 \Big) \right] \\
&\quad - \frac{3}{4} k_n \lambda \Big((u_{kh}^{n+1} + u_{kh}^n)(v_{kh}^{n+1} + v_{kh}^n), \theta_v^{n+1}\theta_v^n \Big) . \qquad (5.76)
\end{aligned}$$

Wir beachten die negativen Vorzeichen von λ. Schließlich summieren wir (5.76) über $n = 0, \ldots, N-1$ auf und erhalten

$$\begin{aligned}
&\|\partial_x \theta_u(T)\|_2^2 + \|\theta_v(T)\|_2^2 - \|\partial_x \theta_u(0)\|_2^2 - \|\theta_v(0)\|_2^2 \\
&\quad \leq C_{u,v} \Bigg((h^{2p} + k^2) \Big(\|u\|^2_{H^2(0,T;H^2)} + \|v\|^2_{H^2(0,T;H^2)} + \|u\|^2_{H^1(0,T;H^p)} \Big) \\
&\qquad\qquad + \|v\|^2_{H^1(0,T;H^p)} + \|\partial_x \theta_u\|^2_{L^2(0,T,L^2)} + \|\theta_v\|^2_{L^2(0,T,L^2)} \Bigg) \\
&\quad - 3\lambda \left[\Big(|u_{kh}(T)|^2, |\theta_v(T)|^2 \Big) - \Big(|u_{kh}(0)|^2, |\theta_v(0)|^2 \Big) \right] \\
&\quad - \frac{3}{2}\lambda \sum_{n=0}^{N-1} k_n \Big((u_{kh}^{n+1} + u_{kh}^n)(v_{kh}^{n+1} + v_{kh}^n), \theta_v^{n+1}\theta_v^n \Big) . \qquad (5.77)
\end{aligned}$$

Das war zu zeigen. □

5.30 Korollar.
Mit der Konstante C_A aus der Annahme (5.46) *gelte für $\lambda < 0$*

$$1 + 3C_A^2 \lambda > 0. \qquad (5.78)$$

Dann können wir aus der vorhergehenden Lemma 5.29 die Fehlerabschätzung

$$\|\partial_x \theta_u(T)\|_2^2 + \|\theta_v(T)\|_2^2 \leq C_{u,v} \Big(\|\partial_x \theta_u(0)\|_2^2 + \|\theta_v(0)\|_2^2 + k^2 + h^{2p} \Big) \qquad (5.79)$$

folgern. Die Konstante $C_{u,v}$ hängt von T, $\|u\|_{H^2(0,T;H^2(\Omega))}$ und $\|v\|_{H^2(0,T;H^2(\Omega))}$, sowie von $\|u\|_{H^1(0,T;H^p(\Omega))}$ und $\|v\|_{H^1(0,T;H^p(\Omega))}$ ab.

Beweis. Wir möchten das diskrete Gronwall-Lemma 2.5 auf das Ergebnis von Lemma 5.29 anwenden. Dazu müssen wir zuerst noch die Punktauswertungen in $t = 0$ und $t = T$ miteinander verrechnen.

Wir verwenden für $\lambda < 0$ unter Zuhilfenahme der Annahme (5.46) die Abschätzungen

$$\begin{aligned}
3\lambda \left(|u_{kh}(T)|^2, |\theta_v(T)|^2 \right) &\geq -3|\lambda| \, \|u_{kh}(T)\|_{L^\infty}^2 \, \|\theta_v(T)\|_{L^2}^2 \\
&\geq -3|\lambda| C_A^2 \, \|\theta_v(T)\|_{L^2}^2 , \\
3\lambda \left(|u_{kh}(0)|^2, |\theta_v(0)|^2 \right) &\leq 3|\lambda| \, \|u_{kh}(0)\|_{L^\infty}^2 \, \|\theta_v(0)\|_{L^2}^2 \\
&\leq 3|\lambda| C_A^2 \, \|\theta_v(0)\|_{L^2}^2 ,
\end{aligned}$$

Daraus erhalten wir

$$\begin{aligned}
&\|\partial_x \theta_u(T)\|_2^2 + \|\theta_v(T)\|_2^2 + 3\lambda \left(|u_{kh}(T)|^2, |\theta_v(T)|^2 \right) \\
&\quad \geq C_1 \left(\|\partial_x \theta_u(T)\|_2^2 + \|\theta_v(T)\|_2^2 \right) , \qquad (5.80a) \\
&\|\partial_x \theta_u(0)\|_2^2 + \|\theta_v(0)\|_2^2 + 3\lambda \left(|u_{kh}(0)|^2, |\theta_v(0)|^2 \right) \\
&\quad \leq C_2 \left(\|\partial_x \theta_u(0)\|_2^2 + \|\theta_v(0)\|_2^2 \right) . \qquad (5.80b)
\end{aligned}$$

Für $\lambda > 0$ ist $C_1 = 1$, für $\lambda < 0$ dagegen

$$C_1 = 1 + 3\lambda C_A^2 .$$

Wir benötigen $C_1 > 0$, daher resultiert die Bedingung (5.78) dieses Korollars.

Wir müssen noch die letzte Summe in Gl. (5.62) geeignet abschätzen und die $L^2(0,T;L^2)$-Norm von $\partial_x \theta_u$ und θ_v durch eine Summe der diskreten Werte abschätzen, um das diskrete Gronwall-Lemma anwenden zu können. Es gilt

$$\begin{aligned}
&\sum_{n=0}^{N-1} k_n \left((u_{kh}^{n+1} + u_{kh}^n)(v_{kh}^{n+1} + v_{kh}^n), \theta_v^{n+1} \theta_v^n \right) \\
&\quad \leq \sum_{n=0}^{N-1} k_n 4 \, \|u_{kh}\|_{L^\infty(I_n;L^\infty)} \, \|v_{kh}\|_{L^\infty(I_n;L^\infty)} \left| \left(\theta_v^{n+1}, \theta_v^n \right) \right| \\
&\quad \leq 2 \, \|u_{kh}\|_{L^\infty(0,T;L^\infty)} \, \|v_{kh}\|_{L^\infty(0,T;L^\infty)} \sum_{n=0}^{N-1} k_n \left(\left\| \theta_v^{n+1} \right\|_2^2 + \|\theta_v^n\|_2^2 \right) \\
&\quad \leq 2 C_A C_B \sum_{n=0}^{N-1} k_n \left(\left\| \theta_v^{n+1} \right\|_2^2 + \|\theta_v^n\|_2^2 \right) , \qquad (5.81)
\end{aligned}$$

außerdem ist

$$\begin{aligned}
&\|\theta_v\|^2_{L^2(0,T;L^2)} \\
&\quad= \sum_{n=0}^{N-1} \int_{I_n} \|\theta_v\|_2^2 \,\mathrm{d}t \\
&\quad= \sum_{n=0}^{N-1} \int_{I_n} \Big(\theta_v^{n+1}\psi^{n+1}(t) + \theta_v^n\psi^n(t), \theta_v^{n+1}\psi^{n+1}(t) + \theta_v^n\psi^n(t)\Big) \,\mathrm{d}t \\
&\quad= \sum_{n=0}^{N-1} \int_{I_n} \left\|\theta_v^{n+1}\right\|_2^2 (\psi^{n+1}(t))^2 + 2(\theta_v^{n+1}, \theta_v^n)\psi^{n+1}(t)\psi^n(t) \\
&\qquad\qquad + \|\theta_v^n\|_2^2 (\psi^n(t))^2 \,\mathrm{d}t \\
&\quad= \sum_{n=0}^{N-1} \frac{1}{3}k_n \left\|\theta_v^{n+1}\right\|_2^2 + \frac{1}{3}k_n(\theta_v^{n+1}, \theta_v^n) + \frac{1}{3}k_n \|\theta_v^n\|_2^2 \\
&\quad= \frac{1}{3}\sum_{n=0}^{N-1} k_n \Big(\left\|\theta_v^{n+1}\right\|_2^2 + (\theta_v^{n+1}, \theta_v^n) + \|\theta_v^n\|_2^2 \Big) \\
&\quad\le \frac{1}{3}\sum_{n=0}^{N-1} k_n \Big(\left\|\theta_v^{n+1}\right\|_2^2 + \frac{1}{2}\left(\left\|\theta_v^{n+1}\right\|_2^2 + \|\theta_v^n\|_2^2 \right) + \|\theta_v^n\|_2^2 \Big) \\
&\quad= \frac{1}{2}\sum_{n=0}^{N-1} k_n \Big(\left\|\theta_v^{n+1}\right\|_2^2 + \|\theta_v^n\|_2^2 \Big). \qquad (5.82)
\end{aligned}$$

Wir verwenden nun (5.80), (5.81) und (5.82) in Gl. (5.62) und erhalten für k_{N-1} klein genug

$$\begin{aligned}
&\left\|\partial_x\theta_u^N\right\|_2^2 + \left\|\theta_v^N\right\|_2^2 \\
&\quad\le C_{u,v}\Big(h^{2p} + k^2 + \|\partial_x\theta_u(0)\|_2^2 + \|\theta_v(0)\|_2^2 \\
&\qquad\qquad + \sum_{n=0}^{N-1} k_n \Big(\|\partial_x\theta_u^n\|_2^2 + \|\theta_v^n\|_2^2 \Big) \Big).
\end{aligned}$$

Darauf wenden wir das diskrete Gronwall-Lemma 2.5 an und erhalten die Behauptung. □

5.31 Bemerkung.
Die Bedingung (5.78) *ist genau die gleiche Bedingung, die wir im kontinuierlichen Problem für die Existenz einer Lösung benötigt haben, siehe Bemerkung 3.3. Der Fall $\lambda < 0$ ist dabei der interessantere Fall. Die Konvergenzanalysis ist also näher am kontinuierlichen Problem als die Existenz- und Eindeutigkeitsbeweise für das diskretisierte Problem, in denen die notwendigen Bedingungen teilweise deutlich von den in der Analysis benötigten Bedingungen abwichen.*

5.32 Satz. (Fehler der Volldiskretisierung)
Es sei $\Omega \subset \mathbb{R}$ ein offenes Intervall. Für die Lösung (u, v) von (3.14) *gelte*

$$u \in H^2(0,T;H^2(\Omega)) \cap H^1(0,T;H^{p+1}(\Omega)),$$
$$v \in H^2(0,T;H^2(\Omega)) \cap H^1(0,T;H^{p}(\Omega)).$$

Es sei (u_{kh}, v_{kh}) die Lösung der Volldiskretisierung (5.27) *mit*

$$\|u_{kh}\|_{L^\infty(0,T;L^\infty)} \le C_A, \qquad \|v_{kh}\|_{L^\infty(0,T;L^\infty)} \le C_B,$$

für von k und h unabhängige Konstanten $C_A, C_B > 0$ und es gelte für $\lambda < 0$

$$1 + 3C_A^2\lambda > 0.$$

Dann gilt

$$\begin{aligned}\|(u-u_{kh})(T)\|^2_{H^1(\Omega)} &+ \|(v-v_{kh})(T)\|_2^2 \\ &\le C\Big(\|(u-u_{kh})(0)\|^2_{H^1(\Omega)} + \|(v-v_{kh})(0)\|_2^2 + k^2 + h^{2p}\Big).\end{aligned}$$

Beweis. Wir verwenden die Fehleraufspaltung (5.45), also ist

$$\begin{aligned}&\|(u-u_{kh})(T)\|^2_{H^1(\Omega)} \\ &\quad\le C\left(\|\rho_u(T)\|^2_{H^1(\Omega)} + \|\mathcal{P}_E\eta_u(T)\|^2_{H^1(\Omega)} + \|\theta_u(T)\|^2_{H^1(\Omega)}\right), \\ &\|(v-v_{kh})(T)\|^2_{L^2(\Omega)} \\ &\quad\le C\left(\|\rho_v(T)\|^2_{L^2(\Omega)} + \|\mathcal{P}_E\eta_v(T)\|^2_{L^2(\Omega)} + \|\theta_v(T)\|^2_{L^2(\Omega)}\right).\end{aligned}$$

$\eta_u(T)$ und $\eta_v(T)$ verschwinden, da $\mathcal{Q}_t^1$ auf dem Zeitgitter interpolierend ist, siehe Gl. (5.3). Weiterhin folgt aus Lemma 5.9, dass

$$\begin{aligned}\|\rho_u(T)\|^2_{H^1(\Omega)} &\le Ch^{2p}\,\|u(T)\|^2_{H^{p+1}(\Omega)}, \\ \|\rho_v(T)\|^2_{L^2(\Omega)} &\le Ch^{2p}\,\|v(T)\|^2_{H^{p}(\Omega)}.\end{aligned}$$

Zusammen mit den Abschätzungen für (θ_u, θ_v) aus Korollar 5.30 folgt die Behauptung. □

5.33 Bemerkung.
Die Anfangsfehler haben für die Wahl

$$u_{kh}(0) := \mathcal{P}_E u(0),\ v_{kh}(0) := \mathcal{P}v(0)$$

die gleiche Fehlerordnung wie der Rest, wobei $\mathcal{P}$ die L^2-orthogonale Projektion auf $\mathcal{S}_h^p$ sei.

Die Konvergenzordnung in der Zeit ist vermutlich nicht optimal. [FP96] und [BL94] liefern für die lineare Wellengleichung eine quadratische Konvergenzordnung bezüglich k und nicht nur eine lineare Konvergenzordnung. Die meisten Abschätzungen des Zeitfehlers in unserem Beweis nutzen nicht die volle zur Verfügung stehende Regularität von u und v aus. Die Abschätzungen von $\mathrm{Id} - \mathcal{Q}_t^1$ *in der* L^2*-Norm bezüglich der Zeit kann man meist noch verbessern, um den Faktor* k_n^2 *gewinnen. Es sind nur zwei Terme, in denen das nicht möglich ist. Diese treten in den Abschätzungen* (5.66) *und* (5.72)*, und zwar sind das*

$$\|\partial_t(v - \widetilde{v})\|_{L^2(I_n;L^2)} \quad \textit{und} \quad \left\|(\mathrm{Id} - \mathcal{Q}_n^0)\widetilde{v}\right\|_{L^2(I_n;L^2)},$$

aus denen man nur eine k_n*-Ordnung erhalten kann.*

Wir halten fest, dass bisher T für $k \to 0$ gegen 0 gehen kann, da wir T definiert haben als den maximalen Zeitpunkt, zu dem eine eindeutige Lösung der diskretisierten Gleichungen existiert. Das Konvergenzresultat gibt uns jetzt aber eine Möglichkeit an die Hand, mit der wir das vorherige Existenzresultat Lemma 5.19 soweit verbessern können, dass T nicht mehr von k abhängig ist.

5.34 Korollar. (Existenzintervall)
Es sei T_{end} *der Zeitpunkt, bis zu dem die Lösung* (u, v) *des kontinuierlichen Problems mit* $\Omega \subset \mathbb{R}$ *existiert, siehe Satz 3.1. Dann gibt es ein* $\widetilde{T} \in (0, T_{\mathrm{end}}]$ *unabhängig von k, sodass für k, h klein genug eine eindeutige Lösung* (u_{kh}, v_{kh}) *von* (5.26) *existiert.*

Beweis. Wir setzen für $t \in [0, T_{\mathrm{end}}]$ und $n = 0, \ldots, N$

$$\begin{aligned} \delta(t) &:= \|u(t)\|_{H^1} + \|v(t)\|_2, \\ \delta_n &:= \|u_{kh}^n\|_{H^1} + \|v_{kh}^n\|_2. \end{aligned}$$

Die kritische Stelle im Beweis des Existenzresultates Lemma 5.19 steckt in (5.33) und ist die Forderung, dass die Lösbarkeitsbedingung

$$1 - |f''|_{\infty,\delta_{N-1}}\delta_{N-1} - 2|f'|_{\infty,2\delta_{N-1}} \geq \frac{1}{2}$$

gelten soll. Diese Bedingung erfordert es, dass δ_{N-1} klein genug ist.

Wir gehen nun wie folgt vor: Wir erfüllen diese Bedingung mit der kontinuierlichen Lösung für ein $\widetilde{T} < T_{\mathrm{end}}$ und wählen dann k und h so klein, dass auch die numerische Lösung für alle $t_n \leq \widetilde{T}$ noch die Lösbarkeitsbedingung erfüllt. Dann können wir folgern, dass das diskretisierte Problem eine Lösung bis zu $\widetilde{T}$ besitzt. Es besteht eine gewisse Wechselbeziehung zwischen $\widetilde{T}$ und k: Je näher $\widetilde{T}$ am maximalen Zeitpunkt ist, für den die Lösbarkeitsbedingung erfüllt ist, desto kleiner müssen k und h gewählt werden.

Zunächst sei $0 < \widetilde{T} < T_{\text{end}}$ so gewählt, dass

$$1 - |f''|_{\infty,\delta(\widetilde{T})}\delta(\widetilde{T}) - 2|f'|_{\infty,2\delta(\widetilde{T})} > \frac{1}{2}.$$

Die Existenz eines solchen $\widetilde{T}$ ist für genügend kleine Anfangswerte gesichert. Wir beachten, dass der Ausdruck $|f''|_{\infty,\delta}\delta + 2|f'|_{\infty,2\delta}$ monoton steigend in δ ist. Wir gehen davon aus, dass der Anfangsfehler verschwindet oder die Genauigkeit $k + h^p$ besitzt. Dann liefert Satz 5.32 für den Fehler mit $T = t_N$ die Aussage

$$\|(u - u_{kh})(T)\|^2_{H^1(\Omega)} + \|(v - v_{kh})(T)\|^2_2 \leq C(k^2 + h^{2p}).$$

Dabei ist $C = C(t)$ monoton steigend in t. Also liegt $(u_{kh}(t_n), v_{kh}(t_n))$ für alle n in einer Kugel mit Radius $C(\widetilde{T})(k + h^p)$ um (u, v). Das bedeutet, dass wir k, h klein genug wählen können, sodass

$$1 - |f''|_{\infty,\delta_n}\delta_n - 2|f'|_{\infty,2\delta_n} \geq \frac{1}{2}$$

für alle $n = 1, \ldots, N$ gilt. Das liefert dank Lemma 5.19 schon die Existenz der numerischen Lösung zur Zeit t_{N+1} und wir können weitere Zeitschritte durchführen, bis $t_{N+m} = \widetilde{T}$ erreicht ist. Das liefert die Behauptung. □

5.35 Bemerkung.
Mit einer analogen Argumentation wie im Beweis zu Korollar 5.34 lässt sich zeigen, dass die Bedingung

$$1 + 3C_A > 0$$

aus dem Satz 5.32 für ein $T > 0$ unabhängig von k und für k, h klein genug erfüllt ist. Die Beschränkung auf den eindimensionalen Fall $\Omega \subset \mathbb{R}$ ist auch hier notwendig, um die Soboleveinbettung anwenden zu können.

5.2.5 Implementierung

Für die Implementierung des Zeitschrittverfahrens geben wir in diesem Abschnitt das nichtlineare Gleichungssystem (5.27) in Matrix-Vektor-Form an. Da wir zur Lösung dieses Gleichungssystems das Newtonverfahren anwenden möchten, bestimmen wir außerdem die Ableitung der zugehörigen Nullstellenfunktion. Zusätzlich zu den reellwertigen Anfangswerten, die wir bisher vorausgesetzt haben, um das Problem übersichtlicher zu halten, bestimmen wir jetzt auch das nichtlineare Gleichungssystem im Fall komplexwertiger Anfangswerte. Dieses gewinnt man einfach daraus, dass man Real- und Imaginärteil getrennt betrachtet.

5.2.5.1 Reellwertiger Fall

Die Gleichung (5.27a) lässt sich sofort in Matrix-Vektor-Form schreiben und lautet

$$\mathbf{M}(U^{n+1} - U^n) = \frac{k_n}{2}\mathbf{M}(V^{n+1} + V^n). \tag{5.83}$$

Diese Gleichung lässt sich nach V^{n+1} umformen und lautet dann ($\mathbf{M}$ ist invertierbar)

$$V^{n+1} = \frac{2}{k_n}(U^{n+1} - U^n) - V^n. \tag{5.84}$$

5.36 Bemerkung.
Alle Produkte von Vektoren, das Quadrieren und die Funktionen f', f'', die auf Vektoren angewendet werden, verstehen wir komponentenweise.

Für die zweite Gleichung (5.27b) beachten wir Gl. (5.3) und erhalten mithilfe der Produktapproximation für die nichtlinearen Terme (vgl. Abschnitt 4.3.1) die Gleichung

$$\begin{aligned}
&\mathbf{M}\left[\left(1 + \frac{1}{2}\left(f'(U^{n+1}) + f'(U^n)\right)\right)(V^{n+1} - V^n)\right] \\
&\quad = -\frac{k_n}{2}\mathbf{K}(U^{n+1} + U^n) \\
&\qquad - \frac{k_n}{4}\left[\mathbf{M}(f''(U^{n+1}) + f''(U^n))(V^{n+1} + V^n)^2\right] + G^n
\end{aligned} \tag{5.85}$$

mit $G^n = (\int_{I_n}(g, \varphi_v)\,\mathrm{d}t)_{v\in\mathcal{N}_h}$. Es gibt nun zwei Möglichkeiten. Entweder man löst das Gleichungssystem bestehend aus Gl. (5.83) und Gl. (5.85) zusammen oder man setzt Gl. (5.84) in Gl. (5.85) ein, um so zwei entkoppelte Gleichungssysteme zu erhalten, von denen eines noch dazu nicht das Invertieren einer Matrix erfordert. Es folgt nach Multiplikation mit $2k_n$

$$\begin{aligned}
&4\mathbf{M}\left[\left(1 + \frac{1}{2}\left(f'(U^{n+1}) + f'(U^n)\right)\right)(U^{n+1} - U^n - k_n V^n)\right] \\
&\quad = -k_n^2\mathbf{K}(U^{n+1} + U^n) \\
&\qquad - \mathbf{M}\left[(f''(U^{n+1}) + f''(U^n))(U^{n+1} - U^n)^2\right] + 2k_n G^n
\end{aligned} \tag{5.86}$$

Diese Gleichung können wir nach U^{n+1} umsortieren,

$$\begin{aligned}
&(4\mathbf{M} + k_n^2\mathbf{K})U^{n+1} + \mathbf{M}[2(f'(U^{n+1}) + f'(U^n))(U^{n+1} - U^n) \\
&\qquad + (f''(U^{n+1}) + f''(U^n))(U^{n+1} - U^n)^2] \\
&\qquad - 2k_n\mathbf{M}[f'(U^{n+1})V^n] \\
&\quad = (4\mathbf{M} - k_n^2\mathbf{K})U^n + 4k_n\mathbf{M}V^n + 2k_n\mathbf{M}[f'(U^n)V^n] + 2k_n G^n.
\end{aligned} \tag{5.87}$$

Wir lösen (5.87) durch die Anwendung des Newton-Verfahrens, siehe Abschnitt 4.3.2. Dafür benötigen wir zuerst eine Funktion, deren Nullstelle wir suchen. U^n und V^n sind gegeben. Wir definieren

$$\begin{aligned}\mathbf{A} &:= 4\mathbf{M} + k_n^2\mathbf{K},\\ F(U) &:= 2(f'(U) + f'(U^n))(U - U^n)\\ &\quad + (f''(U) + f''(U^n))(U - U^n)^2 - 2k_n f'(U)V^n,\\ B &= (4\mathbf{M} - k_n^2\mathbf{K})U^n + 4k_n\mathbf{M}V^n + 2k_n\mathbf{M}[f'(U^n)V^n] + 2k_nG^n.\end{aligned}$$

Mit diesen Definitionen können wir das nichtlineare Gleichungssystem (5.87) umschreiben in das Nullstellenproblem

$$\Phi(U) := \mathbf{A}U + \mathbf{M}F(U) - B = 0. \tag{5.88}$$

Dann gilt, wie wir im Abschnitt 4.3.2 besprochen haben mit der Funktion mult aus (4.23)

$$\Phi'(U) = \mathbf{A} + \mathrm{mult}(\mathbf{M}, F'(U)),$$

wobei

$$\begin{aligned}F'(U) = {} & 2(f''(U))(U - U^n) + 2(f'(U) + f'(U^n)) + f'''(U)(U - U^n)^2\\ & + 2(f''(U) + f''(U^n))(U - U^n) - 2k_n f''(U)V^n.\end{aligned}$$

Damit können wir die Lösung von (5.88) approximativ berechnen. Das Zeitschrittverfahren hat den Nebeneffekt, dass wir keinen guten Startwert suchen müssen, U^n ist gewöhnlich schon ein sehr guter Startwert.

5.37 Bemerkung.
Die Verwendung der inexakten Newtoniteration mit der approximativen Ableitung $\widetilde{\Phi}'(U) = \mathbf{A}$ ist auch möglich. In allen reellwertigen Beispielproblemen wird sich zeigen, dass die inexakte Newtoniteration nennenswert schneller ist als die exakte Newtoniteration. Denn obwohl man bei der inexakten Newtoniteration deutlich mehr Newtonschritte benötigt, lohnt sich das im Vergleich zu den sehr zeitintensiven Funktionsauswertungen der exakten Ableitung.

5.2.5.2 Komplexwertiger Fall

Im Fall komplexer Anfangswerte lässt sich das nichtlineare System mittels der Definitionen aus dem Abschnitt 3.2 ohne größere Änderungen zum reellwertigen Fall angeben. Wie in Abschnitt 3.2 erwähnt wurde, ist f nur im Nullpunkt komplex differenzierbar, wir müssen also von $\mathbb{C}$ auf $\mathbb{R}^2$ übergehen.

Wir können also (5.27) wieder verwenden, angewendet auf das Problem

$$\begin{aligned} \partial_t u &= v, \\ \partial_t v + f'(u)[\partial_t v] &= \Delta u - f''(u)[v,v] + g. \end{aligned}$$

Die Koeffizientenvektoren U^n, V^n bestehen aufgrund der Identifizierung von $\mathbb{C}$ mit $\mathbb{R}^2$ aus je zwei Teilvektoren, dem Real- und dem Imaginärteil, also $U^n = [U^n_{\mathrm{re}}; U^n_{\mathrm{im}}]$ und $V^n = [V^n_{\mathrm{re}}; V^n_{\mathrm{im}}]$.

5.38 Definition. (Vektorielle Funktionsauswertungen)
Auch hier müssen wir wieder definieren, wie wir Funktionsauswertungen mit Vektorargumenten verstehen. In der Nullstellenfunktion treten die Funktionen f' und f'' mit zwei bzw. drei Argumenten auf, die jeweils aus Real- und aus Imaginärteil bestehen. Wir haben also für $m = 2, 3$ Funktionen

$$F\colon \bigtimes_{i=1}^{m} \mathbb{R}^2 \to \mathbb{R}^2,$$

wobei $F = (F_{\mathrm{re}}, F_{\mathrm{im}})$.

Vektoriell aufgefasst ist dann F für $M := |\mathcal{N}_h|$ gegeben durch

$$F\colon \bigtimes_{i=1}^{m} \mathbb{R}^{2M} \to \mathbb{R}^{2M}.$$

Wir schreiben für die Argumente $X^i = [X^i_{\mathrm{re}}; X^i_{\mathrm{im}}]$ und als Funktionswert $Y = [Y_{\mathrm{re}}; Y_{\mathrm{im}}]$, also

$$F(X^1, \dots, X^m) = Y.$$

Wir definieren die Auswertung komponentenweise für $\nu \in \mathcal{N}_h$ durch

$$\begin{aligned} (Y_{\mathrm{re}})_\nu &:= F_{\mathrm{re}}([(X^1_{\mathrm{re}})_\nu; (X^1_{\mathrm{im}})_\nu], \dots, [(X^m_{\mathrm{re}})_\nu; (X^m_{\mathrm{im}})_\nu]), \\ (Y_{\mathrm{im}})_\nu &:= F_{\mathrm{im}}([(X^1_{\mathrm{re}})_\nu; (X^1_{\mathrm{im}})_\nu], \dots, [(X^m_{\mathrm{re}})_\nu; (X^m_{\mathrm{im}})_\nu]). \end{aligned}$$

Noch eine Stufe weiter müssen wir bei der Berechnung der Ableitung der Nullstellenfunktion gehen. Dort treten für $m = 2, 3$ matrixwertige Funktionen der Form $G\colon \times_{i=1}^m \mathbb{R}^2 \to \mathbb{R}^{2\times 2}$ auf. Die Komponenten der Argumente indizieren wir wie oben mit re und im, die vier Komponenten der Funktionswerte indizieren wir, wie es bei Matrizen üblich ist, durch G_{11}, G_{12}, G_{21} und G_{22}. Entsprechend definieren wir $G\colon \times_{i=1}^m \mathbb{R}^{2M} \to \mathbb{R}^{2M\times 2}$ für $\nu \in \mathcal{N}_h$ durch

$$G(X^1, \dots, X^m) := Y = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix} \in \mathbb{R}^{2M\times 2}$$

mit $Y_{ij} \in \mathbb{R}^{M\times 1}$, $i, j = 1, 2$ und

$$(Y_{ij})_\nu := G_{ij}([(X^1_{\mathrm{re}})_\nu; (X^1_{\mathrm{im}})_\nu], \dots, [(X^m_{\mathrm{re}})_\nu; (X^m_{\mathrm{im}})_\nu]) \qquad (i, j = 1, 2).$$

Jetzt können wir genauso vorgehen wie im reellwertigen Fall und erhalten die zwei Gleichungssysteme

$$V^{n+1} = \frac{2}{k_n}(U^{n+1} - U^n) - V^n \tag{5.89}$$

und

$$\begin{aligned}(4\widetilde{\mathbf{M}} + k_n^2\widetilde{\mathbf{K}})U^{n+1} + \widetilde{\mathbf{M}}\Big[&2\big(f'(U^{n+1})[U^{n+1} - U^n] + f'(U^n)[U^{n+1} - U^n]\big)\\ &+ \big(f''(U^{n+1})[U^{n+1} - U^n, U^{n+1} - U^n]\\ &+ f''(U^n)[U^{n+1} - U^n, U^{n+1} - U^n]\big)\Big] - 2k_n\widetilde{\mathbf{M}}f'(U^{n+1})[V^n]\\ = (4\widetilde{\mathbf{M}} - k_n^2\widetilde{\mathbf{K}})U^n + 4k_n\widetilde{\mathbf{M}}V^n + 2k_n\widetilde{\mathbf{M}}f'(U^n)[V^n] + 2k_nG^n.\end{aligned}$$

Masse- und Steifigkeitsmatrix wurden erweitert zu

$$\widetilde{\mathbf{M}} := \begin{pmatrix}\mathbf{M} & \\ & \mathbf{M}\end{pmatrix}, \qquad \widetilde{\mathbf{K}} := \begin{pmatrix}\mathbf{K} & \\ & \mathbf{K}\end{pmatrix}. \tag{5.90}$$

Analog zum reellwertigen Fall definieren wir

$$\begin{aligned}\widetilde{\mathbf{A}} &:= 4\widetilde{\mathbf{M}} + k_n^2\widetilde{\mathbf{K}},\\ F(U) &:= 2\big(f'(U^{n+1})[U^{n+1} - U^n] + f'(U^n)[U^{n+1} - U^n]\big),\\ &\quad + \big(f''(U^{n+1})[U^{n+1} - U^n, U^{n+1} - U^n]\\ &\quad + f''(U^n)[U^{n+1} - U^n, U^{n+1} - U^n]\big)\\ \widetilde{B} &= (4\widetilde{\mathbf{M}} - k_n^2\widetilde{\mathbf{K}})U^n + 4k_n\widetilde{\mathbf{M}}V^n + 2k_n\widetilde{\mathbf{M}}f'(U^n)[V^n] + 2k_nG^n.\end{aligned}$$

Wir suchen also nach der Nullstelle von

$$\Phi(U) = \widetilde{\mathbf{A}}U + \widetilde{\mathbf{M}}F(U) - \widetilde{B}.$$

Um jetzt die Ableitung $\Phi'(U)$ für die Anwendung des Newton-Verfahrens bestimmen zu können, benötigen wir die partiellen Ableitungen von $f'(u)[v]$ und $f''(u)[v,w]$. Für $f(z) = \lambda|z|^2 z$ gilt

$$\begin{aligned}\partial_u f'(u)[v] &= \lambda\begin{pmatrix}6u_{\mathrm{re}}v_{\mathrm{re}} + 2u_{\mathrm{im}}v_{\mathrm{im}} & 2u_{\mathrm{im}}v_{\mathrm{re}} + 2u_{\mathrm{re}}v_{\mathrm{im}}\\ 2u_{\mathrm{im}}v_{\mathrm{re}} + 2u_{\mathrm{re}}v_{\mathrm{im}} & 2u_{\mathrm{re}}v_{\mathrm{re}} + 6u_{\mathrm{im}}v_{\mathrm{im}}\end{pmatrix},\\ \partial_v f'(u)[v] &= \lambda\begin{pmatrix}3u_{\mathrm{re}}^2 + u_{\mathrm{im}}^2 & 2u_{\mathrm{re}}u_{\mathrm{im}}\\ 2u_{\mathrm{re}}u_{\mathrm{im}} & u_{\mathrm{re}}^2 + 3u_{\mathrm{im}}^2\end{pmatrix}\end{aligned}$$

und

$$\begin{aligned}\partial_u f''(u)[v,w] &= \lambda\begin{pmatrix}6v_{\mathrm{re}}w_{\mathrm{re}} + 2v_{\mathrm{im}}w_{\mathrm{im}} & 4v_{\mathrm{re}}w_{\mathrm{im}}\\ 4v_{\mathrm{im}}w_{\mathrm{re}} & 2v_{\mathrm{re}}w_{\mathrm{re}} + 6v_{\mathrm{im}}w_{\mathrm{im}}\end{pmatrix},\\ \partial_v f''(u)[v,w] &= \lambda\begin{pmatrix}6u_{\mathrm{re}}w_{\mathrm{re}} + 4u_{\mathrm{im}}w_{\mathrm{im}} & 2u_{\mathrm{re}}w_{\mathrm{im}}\\ 2u_{\mathrm{im}}w_{\mathrm{re}} & 4u_{\mathrm{re}}w_{\mathrm{re}} + 6u_{\mathrm{im}}w_{\mathrm{im}}\end{pmatrix},\\ \partial_w f''(u)[v,w] &= \lambda\begin{pmatrix}6u_{\mathrm{re}}v_{\mathrm{re}} & 4u_{\mathrm{im}}v_{\mathrm{re}} + 2u_{\mathrm{re}}v_{\mathrm{im}}\\ 2u_{\mathrm{im}}v_{\mathrm{re}} + 4u_{\mathrm{re}}v_{\mathrm{im}} & 6u_{\mathrm{im}}v_{\mathrm{im}}\end{pmatrix}.\end{aligned}$$

Damit erhalten wir dann

$$\begin{aligned} F'(U) = 2\Big(&\partial_u f'(U)[U-U^n] + \partial_v f'(U)[U-U^n] + \partial_v f'(U^n)[U-U^n]\Big) \\ &+ \partial_u f''(U+U^n)[U-U^n, U-U^n] \\ &+ \partial_v f''(U+U^n)[U-U^n, U-U^n] \\ &+ \partial_w f''(U+U^n)[U-U^n, U-U^n]. \end{aligned}$$

Das erlaubt uns dann, die Ableitung der Funktion Φ anzugeben, es ist

$$\Phi'(U) = \widetilde{\mathbf{A}} + \begin{pmatrix} \mathrm{mult}(\mathbf{M}, (F'(U))_{11}) & \mathrm{mult}(\mathbf{M}, (F'(U))_{12}) \\ \mathrm{mult}(\mathbf{M}, (F'(U))_{21}) & \mathrm{mult}(\mathbf{M}, (F'(U))_{22}) \end{pmatrix}.$$

5.39 Bemerkung.
Wir haben im reellwertigen Fall gesehen, dass eine inexakte Newton-Iteration sehr viel Zeit sparen kann. Es ist wichtig zu bemerken, dass im komplexwertigen Fall eine inexakte Newton-Iteration mit $\widetilde{\Phi}'(U) = \widetilde{\mathbf{A}}$ nicht *verwendet werden kann. In diesem Fall ergeben unsere Experimente nämlich, dass die inexakte Newton-Iteration im allgemeinen nicht konvergiert.*

5.3 Ein cG-Verfahren für die semilineare Wellengleichung

In diesem Abschnitt stellen wir ein cG-Verfahren basierend auf der semilinearen Formulierung (3.25) vor.

Wir haben in Abschnitt 3.4 gesehen, dass wir die Existenz einer eindeutigen Lösung der semilinearen Formulierung nicht zeigen können. Aus ähnlichen Gründen scheitert auch ein Konvergenzbeweis des nachfolgend definierten Verfahrens für diese Formulierung. Dennoch zeigen die numerischen Beispiele des nächsten Abschnittes, dass dieses Verfahren vergleichbare Konvergenzergebnisse liefert wie das zuvor analysierte qlw-cG(1) cG(p)-Verfahren. Außerdem werden wir das Verfahren des aktuellen Abschnittes aufgrund seiner einfachen Struktur als Grundlage für das lokale Zeitschrittverfahren verwenden, das wir in Kapitel 6 definieren werden. Wir nehmen also die Problemformulierung

$$\begin{aligned} u + f(u) &= v, \\ \partial_t v &= w, \\ \partial_t w &= \Delta u + g \end{aligned}$$

als Ausgangspunkt.

5.3.1 Verfahrensformulierung slw-cG(1) cG(p)

Die variationelle Formulierung ist

$$\begin{aligned}\int_{I_n} (u + f(u), \phi_1)\,\mathrm{d}t &= \int_{I_n} (v, \phi_1)\,\mathrm{d}t,\\ \int_{I_n} (\partial_t v, \phi_2)\,\mathrm{d}t &= \int_{I_n} (w, \phi_2)\,\mathrm{d}t\\ \int_{I_n} (\partial_t w, \phi_3)\,\mathrm{d}t &= -\int_{I_n} (\nabla u, \nabla\phi_2)\,\mathrm{d}t + \int_{I_n} (g, \phi_3)\,\mathrm{d}t\end{aligned}$$

für $\phi_1, \phi_2 \in L^2(0,T;L^2(\Omega))$ und $\phi_3 \in L^2(0,T;H_0^1(\Omega))$. Die Diskretisierung führen wir genauso durch wie beim qlw-cG(1) cG(p)-Verfahren. Wir suchen also Funktionen u_{kh}, v_{kh} und w_{kh} aus $\mathcal{S}_k^1 \otimes \mathcal{S}_h^p$ mit

$$\begin{aligned}\int_{I_n} (u_{kh} + \mathcal{I}_1(f(u_{kh})), \phi_1)\,\mathrm{d}t &= \int_{I_n} (v_{kh}, \phi_1)\,\mathrm{d}t,\\ \int_{I_n} (\partial_t v_{kh}, \phi_2)\,\mathrm{d}t &= \int_{I_n} (w_{kh}, \phi_2)\,\mathrm{d}t\\ \int_{I_n} (\partial_t w_{kh}, \phi_3)\,\mathrm{d}t &= -\int_{I_n} (\nabla u_{kh}, \nabla\phi_2)\,\mathrm{d}t + \int_{I_n} (g, \phi_3)\,\mathrm{d}t\end{aligned}$$

für alle $\phi_1, \phi_2, \phi_3 \in \mathbb{P}_0(I_n) \otimes \mathcal{S}_h^p$. Dabei sei $\mathcal{I}_1$ ein Interpolationsoperator bezüglich der Zeitvariablen. Dann entspricht $\int_{I_n} \mathcal{I}_1 w\ \mathrm{d}t$ einer Quadraturformel. Um die erwünschte Fehlerordnung zu erreichen, kann man $\mathcal{I}_1$ so wählen, dass man die Trapez- oder Mittelpunktregel erhält. Beide führen zu vergleichbaren Fehlern in den numerischen Experimenten, wir werden uns im Folgenden auf die Mittelpunktregel beschränken, also ist

$$\mathcal{I}_1 w\big\lfloor_{I_n} := w\left(\frac{t_n + t_{n+1}}{2}\right).$$

Wir nennen dieses Verfahren **slw-cG(1) cG(p)**.

Die Probleme, die bei dem Versuch eines Konvergenzbeweises auftreten, können wir kurz skizzieren. Wir haben in Abschnitt 3.4 gesehen, dass der Existenzbeweis für die semilineare Formulierung daran scheitert, dass die nichtlinearen Terme in gewisser Weise in der falschen Komponente auftauchen und dadurch keine lokale Lipschitzstetigkeit der rechten Seite in der betreffenden Norm vorhanden ist. Natürlicherweise trifft man auf dieses Problem auch in dem Versuch eines Konvergenzbeweises (unter der nicht bewiesenen Annahme, dass eine genügend glatte Lösung der Differentialgleichung existiert). Definieren wir die Fehlerfunktionen $e_u := u - u_{kh}$, $e_v := v - v_{kh}$

und $e_w := w - w_{kh}$, so folgt

$$\int_{I_n} (e_u, \phi_1)\,\mathrm{d}t + \int_{I_n} (f(u) - \mathcal{I}_1(f(u_{kh})), \phi_1)\,\mathrm{d}t = \int_{I_n} (e_v, \phi_1)\,\mathrm{d}t, \tag{5.91a}$$

$$\int_{I_n} (\partial_t e_v - e_w, \phi_2)\,\mathrm{d}t = 0, \tag{5.91b}$$

$$\int_{I_n} (\partial_t e_w, \phi_3)\,\mathrm{d}t + \int_{I_n} (\nabla e_u, \nabla \phi_3)\,\mathrm{d}t = 0, \tag{5.91c}$$

für alle $\phi_1,\ \phi_2,\ \phi_3 \in \mathbb{P}_0(I_n) \otimes \mathcal{S}_h^p$. Wenn wir davon absehen, dass wir den Fehler wie zuvor noch weiter aufteilen müssen, so können wir aus diesen Fehlergleichungen schon sehen, dass nicht klar ist, wie eine Strategie zum Testen dieser Gleichungen aussehen könnte. Wir kennen nur eine „Energie“ unseres Problems und diese haben wir aus der quasilinearen Formulierung gewonnen. Diese Energie für die obigen Fehlergleichungen auszunutzen scheint aber aussichtslos. Für Beweise, die mit der Energiemethode durchgeführt werden, ist das Vorhandensein einer Energie aber eine notwendige Voraussetzung. Das können wir auch sehen, wenn wir versuchen ohne Beachtung dieses Wissens Gl. (5.91) geeignet zu testen. Der Einfachheit halber ignorieren wir in dieser kurzen Darstellung die Tatsache, dass die Fehlerfunktionen nicht im Testraum liegen und erhalten mit der natürlichen Wahl $\phi_3 := e_w$

$$\int_{I_n} (\partial_t e_w, e_w)\,\mathrm{d}t = -\int_{I_n} (\nabla e_u, \nabla e_w)\,\mathrm{d}t. \tag{5.92}$$

Die rechte Seite müssen wir loswerden, da e_w auf der linken Seite nur in der $L^2(\Omega)$-Norm vorliegt. Also testen wir weiter mit $\phi_2 := \mathcal{A} e_u$ und erhalten

$$\int_{I_n} (\partial_t \nabla e_v, \nabla e_u)\,\mathrm{d}t = \int_{I_n} (\nabla e_w, \nabla e_u)\,\mathrm{d}t. \tag{5.93}$$

Durch Aufaddieren von (5.92) und (5.93) fällt also der Gradiententerm von e_w weg. Um die linke Seite von (5.93) zu verrechnen, müsste man (5.91a) mit $\phi_1 = \partial_t \mathcal{A} e_u$ testen (wir ignorieren die Randwerte, die bei der partiellen Integration in der Zeit auftreten) und erhielte

$$\int_{I_n} (\nabla e_u, \partial_t \nabla e_u)\,\mathrm{d}t + \int_{I_n} (\nabla(f(u) - \mathcal{I}_1(f(u_{kh}))), \partial_t \nabla e_u)\,\mathrm{d}t = \int_{I_n} (\nabla e_v, \partial_t \nabla e_u)\,\mathrm{d}t. \tag{5.94}$$

Hier sehen wir nun das Problem, das auch im Kontinuierlichen aufgetreten ist: Der nichtlineare Term in (5.94) passt nicht zu den restlichen Termen, die Zeitableitung in $\partial_t \nabla e_u$ ist „zuviel“. Im diskreten Fall kann man diese Zeitableitung durch eine inverse Abschätzung loswerden, verliert dadurch aber eine Ordnung in der Zeit. Bei genauer Behandlung dieses Problems sieht

man, dass es sich dabei nicht nur um den Verlust einer Fehlerordnung handelt, sondern es fehlt ein k_n, das für die Anwendung des diskreten Gronwall-Lemmas notwendig ist. Ähnliche Probleme treten unabhängig davon auf, wie man testet. Dennoch werden wir sehen, dass die numerischen Experimente die Probleme der Theorie nicht bestätigen. Das Konvergenzverhalten dieses Verfahrens entspricht meist recht exakt dem des qlw-cG(1) cG(p)-Verfahrens.

5.3.2 Implementierung

Wir bestimmen die nichtlinearen Gleichungssystem im Fall reellwertiger sowie komplexwertiger Anfangswerte, wie wir es in Abschnitt 5.2.5 getan haben.

5.3.2.1 Reellwertiger Fall

Das Gleichungssystem (5.27) lässt sich mithilfe der Produktapproximation 4.3.1 in die Matrix-Vektor-Form bringen,

$$\frac{1}{2}\mathbf{M}(U^{n+1}+U^n)+\mathbf{M}f((U^{n+1}+U^n)/2)=\frac{1}{2}\mathbf{M}(V^{n+1}+V^n), \tag{5.95a}$$

$$\mathbf{M}(V^{n+1}-V^n)=\frac{k_n}{2}\mathbf{M}(W^{n+1}+W^n), \tag{5.95b}$$

$$\mathbf{M}(W^{n+1}-W^n)=-\frac{k_n}{2}\mathbf{K}(U^{n+1}+U^n)+G^n, \tag{5.95c}$$

wobei $G^n = (\int_{I_n}(g,\varphi_\nu)\ \mathrm{d}t)_{\nu\in\mathcal{N}_h}$. Wir entkoppeln diese Gleichungen, indem wir W^{n+1} in (5.95b) durch (5.95c) eliminieren und dann V^{n+1} in (5.95a) mit (5.95b) ersetzen. Wir erhalten die nichtlineare Gleichung

$$\begin{aligned}&\left(\mathbf{M}+\frac{k_n^2}{4}\mathbf{K}\right)U^{n+1}+2\mathbf{M}f((U^{n+1}+U^n)/2)\\&\quad=-\left(\mathbf{M}+\frac{k_n^2}{4}\mathbf{K}\right)U^n+\mathbf{M}\left(2V^n+\frac{k_n}{2}W^n\right)+\frac{k_n}{2}G^n.\end{aligned} \tag{5.96}$$

Diese Gleichung lösen wir zuerst, mit U^{n+1} können wir dann über (5.95c) W^{n+1} bestimmen und schließlich V^{n+1} mit (5.95b). Mit

$$\begin{aligned}\mathbf{A}&:=\mathbf{M}+\frac{k_n^2}{4}\mathbf{K},\\B&:=-\mathbf{A}U^n+\mathbf{M}\left(2V^n+\frac{k_n}{2}W^n\right)+\frac{k_n}{2}G^n\end{aligned}$$

suchen wir die Nullstelle der Funktion

$$\Phi(U):=\mathbf{A}U+2\mathbf{M}f((U+U^n)/2)-B.$$

Wie in 5.2.5 erhalten wir

$$\Phi'(U) = \mathbf{A} + 2\,\mathrm{mult}(\mathbf{M}, f'((U+U^n)/2))$$

und können damit wieder das Newton-Verfahren durchführen. Wir bemerken, dass wir in der semilinearen Formulierung von der nichtlinearen Funktion f nicht mehr als eine Ableitung benötigen. Bei der quasilinearen Formulierung haben wir im Gegensatz dazu sogar noch die dritte Ableitung für die Newton-Iteration gebraucht.

5.3.2.2 Komplexwertiger Fall

War die Definition der Nullstellenfunktion und ihrer Ableitung für die quasilineare Formulierung in Abschnitt 5.2.5.2 noch umfangreich, so ist das für die semilineare Formulierung sehr einfach.

Es seien $\widetilde{\mathbf{M}}$ und $\widetilde{\mathbf{K}}$ die $\mathbb{R}^2$-Entsprechungen der Masse- und der Steifigkeitsmatrix wie in Gl. (5.90). Dann definieren wir

$$\begin{aligned}
\widetilde{\mathbf{A}} &:= \widetilde{\mathbf{M}} + \frac{k_n^2}{4}\widetilde{\mathbf{K}},\\
\widetilde{B} &:= -\widetilde{\mathbf{A}}U^n + \widetilde{\mathbf{M}}\left(2V^n + \frac{k_n}{2}W^n\right) + \frac{k_n}{2}G^n.
\end{aligned}$$

Wir suchen die Nullstelle der Funktion

$$\Phi(U) := \widetilde{\mathbf{A}}U + 2\widetilde{\mathbf{M}}f((U+U^n)/2) - \widetilde{B}.$$

Wie üblich ist f hier komponentenweise zu verstehen. Es gilt für $f(z) = \lambda|z|^2 z$

$$f'(u) = \lambda \begin{pmatrix} 3u_{\mathrm{re}}^2 + u_{\mathrm{im}}^2 & 2u_{\mathrm{re}}u_{\mathrm{im}} \\ 2u_{\mathrm{re}}u_{\mathrm{im}} & u_{\mathrm{re}}^2 + 3u_{\mathrm{im}}^2 \end{pmatrix}.$$

Also folgt unter Beachtung der Definition 5.38 (mit nur einem Argument bei der Funktionsauswertung)

$$\Phi'(U) = \widetilde{\mathbf{A}} + 2\begin{pmatrix} \mathbf{F}_{11}(U) & \mathbf{F}_{12}(U) \\ \mathbf{F}_{21}(U) & \mathbf{F}_{22}(U) \end{pmatrix}.$$

mit

$$\mathbf{F}_{ij}(U) := \mathrm{mult}(\mathbf{M}, (f'((U+U^n)/2))_{ij}) \qquad (i,j=1,2).$$

5.4 Numerische Ergebnisse

Die in den vorherigen zwei Abschnitten vorgestellten Verfahren wurden in MATLAB® [1] (Version 7.4.0.287 (R2007a) und 7.14.0.739 (R2012a)) implementiert. Bei der zugrundeliegenden Finite-Elemente-Bibliothek für elliptische Probleme handelt es sich um die `femtoolbox` von M. Richter (vgl. die Dissertation [Ric10]), das Newtonverfahren ist von W. Dörfler. Die Implementierungen der Zeitschrittverfahren stammen von mir.

Alle Berechnungen, die wir im Folgenden durchführen, verwenden die gleichen Parameter für das Newton-Verfahren. Zur Bedeutung der Parameter vergleiche mit Abschnitt 4.3.2. Wir setzen

$$\begin{aligned} n_{\max} &:= 20, \\ \mathrm{TOL}_U &:= 10^{-8}, \\ \mathrm{TOL}_F &:= 10^{-10}. \end{aligned}$$

$n_{\max}$ ist zur Sicherheit groß angesetzt, wir werden auf die tatsächlich notwendige Iterationszahl in den jeweiligen Abschnitten eingehen.

Wir werden bei den nachfolgenden drei Testproblemen verschiedene Problemklassen abdecken. Bei dem ersten Problem handelt es sich um ein homogenes Problem in $\mathbb{R}$ mit zu vernachlässigenden Randwerten und $\lambda < 0$. Die Lösung existiert hier nur in einem beschränkten Zeitintervall. Das zweite Problem behandelt den inhomogenen Fall für $\lambda < 0$ in $\mathbb{R}$ mit einer global (in der Zeit) existenten Lösung. Beide Probleme besitzen reelle Anfangswerte. Zu guter Letzt betrachten wir ein homogenes Problem in $\mathbb{R}^2$ mit nicht verschwindenden Randwerten, $\lambda > 0$ und komplexen Anfangswerten. Auch die Lösung dieses Problems existiert global in der Zeit.

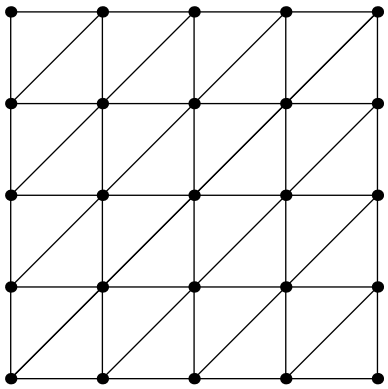


Abbildung 5.1: Einfaches criss-Gitter.

Wir verwenden in allen Beispielen gleichmäßige Diskretisierungen mit den Parametern k und h, die sich aus den jeweiligen Konvergenztabellen ergeben. Für das Solitonproblem in Abschnitt 5.4.3, welches in $\mathbb{R}^2$ formuliert wird, verwenden wir ein sogenanntes *criss-Gitter* (zu weiteren Gittertypen siehe [Bra03, Kapitel 8]). Ein solches Gitter besteht aus Dreiecken die wie in Abb. 5.1 angeordnet sind, wobei nach rechts die x-Richtung und nach oben die y-Richtung aufgetragen sei.

[1] MATLAB® ist eine eingetragene Marke von *The MathWorks, Inc.*

In allen Fällen geben wir den Diskretisierungsfehler $u - u_{kh}$ in der H^1-Halbnorm zum jeweiligen Endzeitpunkt $t = T$ an. Nur für das Solitonproblem müssen wir diese Aussage relativieren: Wenn $t = T$ in N Zeitschritten erreicht sein sollte, dann ist dies der Zeitpunkt, an dem wir den Fehler berechnen. Aufgrund der inexakten Zahldarstellung eines Computers ist dies gewöhnlich *nicht* $t = T$, sondern bei unserer Zeitschrittwahl ein Zeitpunkt, der geringfügig vor T liegt.

Die experimentelle Konvergenzordnung (wir schreiben in den Tabellen kurz *eoc* für „**e**xperimental **o**rder of **c**onvergence“) bestimmen wir für lineare Elemente $p = 1$ in Abhängigkeit von $k + h$, da der Konvergenzsatz 5.32 in diesem Fall die Ordnung 1 in k und h erwarten lässt. Für quadratische Elemente $p = 2$ betrachten wir Konvergenz in Zeit und Ort getrennt. Dazu wählen wir jeweils eine der beiden Parameter k und h so klein, dass der entsprechende Zeit- bzw. Ortsdiskretisierungsfehler vernachlässigbar ist und betrachten dann die Konvergenz im anderen Parameter. In den Abbildungen der Fehler befindet sich immer ein Steigungsdreieck, dass die Ordnung 1 oder 2 exemplarisch anzeigt, damit ein Vergleich möglich ist.

5.4.1 Pulspropagation mit blow-up, $\lambda < 0$

Das erste Referenzproblem ist das Modellproblem (3.6) mit $g = 0$, also ein homogenes Problem. Eine Lösung wurde durch einen speziellen Ansatz in der Arbeit von Busch et al. [PNTB09] hergeleitet, diesen Ansatz werden wir unten vorstellen. Man erhält eine Lösung, die nur für einen endlichen Zeitraum $[0, T_{\text{end}})$ existiert und dann bezüglich der H^1-Norm einen sogenannten *blow-up* erfährt. Das heißt, dass die H^1-Norm für $t \to T_{\text{end}}$ gegen Unendlich geht.

5.4.1.1 Herleitung von Referenzlösungen

Wir beginnen mit der Situation in (3.11) mit $\Omega \subset \mathbb{R}$. Es folgt also das System

$$\partial_t\big((1 + \lambda|u(t,x)|^2)u(t,x)\big) = \partial_x v(t,x), \tag{5.97a}$$

$$\partial_t v(t,x) = \partial_x u(t,x). \tag{5.97b}$$

Wir nehmen weiter an, dass die Anfangsbedingungen an u und v reell sind, so dass die Lösungen auch, solange sie existieren, reell bleiben. In diesem Fall können wir die Zeitableitung auf der linken Seite von (5.97a) auflösen und erhalten

$$\partial_t u(t,x) = \frac{1}{1 + 3\lambda|u(t,x)|^2}\partial_x v(t,x). \tag{5.98}$$

Jetzt machen wir den Ansatz

$$v(t,x) = b(u(t,x))u(t,x)$$

für eine stetig differenzierbare Funktion b auf $\mathbb{R}$. Setzen wir diesen Ansatz in (5.98) ein, so folgt

$$\partial_t u = \frac{1}{1+3\lambda|u|^2}(b'(u)u + b(u))\partial_x u \tag{5.99}$$

und selbiges Vorgehen in (5.97b) liefert

$$(b'(u)u + b(u))\partial_t u = \partial_x u\,.$$

Zusammen erhalten wir

$$\frac{1}{1+3\lambda|u|^2}(b'(u)u + b(u))^2\partial_x u = \partial_x u.$$

Eine nichttriviale Lösung u erfordert daher

$$b'(u)u + b(u) = \pm\sqrt{1+3\lambda|u|^2}\,.$$

Einsetzen dieser Bedingung in (5.99) ergibt

$$\partial_t u = \pm\frac{1}{\sqrt{1+3\lambda|u|^2}}\partial_x u. \tag{5.100}$$

Wir wählen das Minuszeichen, um Ausbreitung nach rechts zu gewährleisten. Dies Lösungen dieser verallgemeinerten Transportgleichung sind gegeben durch die implizite Gleichung

$$u(t,x) = G(u(t,x)) := F\left(x - \frac{1}{\sqrt{1+3\lambda|u(t,x)|^2}}t\right) \tag{5.101}$$

für beliebiges $F \in C^1(\mathbb{R})$ und alle $(t,x) \in (0,T)\times\Omega$. Bei gegebener Anfangsbedingung $u_0(x) = u(0,x)$ folgt $F \equiv u_0$, falls u_0 stetig differenzierbar ist.

5.40 Bemerkung.
Auf den ersten Blick scheint es, dass man damit für gegebene Anfangsbedingungen $u(0), \partial_t u(0)$ mit obigem Vorgehen immer eine Lösung der Differentialgleichung erhält. Dem ist nicht so, die Anfangsbedingung $\partial_t u(0)$ muss einer aus dem Ansatz folgenden Verträglichkeitsbedingung genügen, damit das obige Vorgehen funktioniert.

Schon hier kann man das allgemeine Verhalten einer Lösung (5.101) beurteilen: Die Charakteristiken sind offensichtlich Geraden, die aber nicht parallel zueinander sind (vergleiche auch [Eva98, Chapter 3.2]). Das sorgt dafür, dass glatte Anfangsdaten früher oder später in einen Schockzustand geraten.

Dieses Verhalten kann man auch im Beweis der Existenz einer Lösung beobachten, siehe Bemerkung 3.3. Im Fall (5.101) müssen wir nur punktweise die Fixpunktgleichung lösen. Die Lösbarkeit dieser Gleichung ist garantiert, wenn

$$F'(x) \leq C < 1$$

ist. Darüber hinaus muss

$$1 + 3\lambda|u(t,x)|^2 > 0$$

für alle $(t,x) \in (0,T) \times \Omega$ gelten (oder kleiner Null, nur ein Nulldurchgang wäre problematisch, da das zu differential-algebraischen Gleichungen führen würde.) Diese Bedingung ist im defokussierenden Fall mit $\lambda > 0$ immer erfüllt und muss daher nicht beachtet werden. Im fokussierenden Fall, also mit negativem λ, ist diese Bedingung nicht automatisch erfüllt.

5.4.1.2 Gauß-Puls

Es sei $\Omega = (0,100)$ und $\lambda = -0.08$. Als Anfangsbedingung benutzen wir einen reskalierten Gauß-Puls,

$$u_0(x) := \mathrm{e}^{-\frac{1}{25}(x-25)^2}.$$

Dies legt die Funktion F in (5.101) fest. Für unsere Verfahren benötigen wir weiterhin eine Anfangsbedingung für $v = \partial_t u$. Diese lässt sich aus (5.100) bestimmen, es ist

$$\begin{aligned} v_0(x) &= \frac{1}{\sqrt{1+3\lambda|u_0(x)|^2}} \partial_x u_0(x) \\ &= -\frac{2}{25}(x-25)\frac{1}{\sqrt{1+3\lambda|u_0(x)|^2}} u_0(x). \end{aligned}$$

Wir bestimmen u durch eine einfache Fixpunktiteration mit dem Anfangswert $u^0(t,x) := 0$ für alle $(t,x) \in [0,T] \times \Omega$,

$$u^{k+1}(t,x) := G(u^k(t,x)) \qquad ((t,x) \in [0,T] \times \Omega,\ k \in \mathbb{N}_0).$$

Wir brechen dabei ab, wenn $|u^{k+1}(t,x) - G(u^k(t,x))| < 10^{-12}$ ist. Die Ableitung im Ort von u approximieren wir durch den Differenzenquotienten

$$\partial_x u(t,x) \approx \frac{u(t,x+10^{-8}) - u(t,x)}{10^{-8}}.$$

Die Lösung existiert bis $T = 15$, bis zu diesem Zeitpunkt führen wir unsere Simulationen durch. In Abb. 5.2 ist die Lösung u zum Startzeitpunkt $t = 0$ sowie zum Endzeitpunkt $t = T$ abgebildet. Man sieht, dass sich die Pulsfront

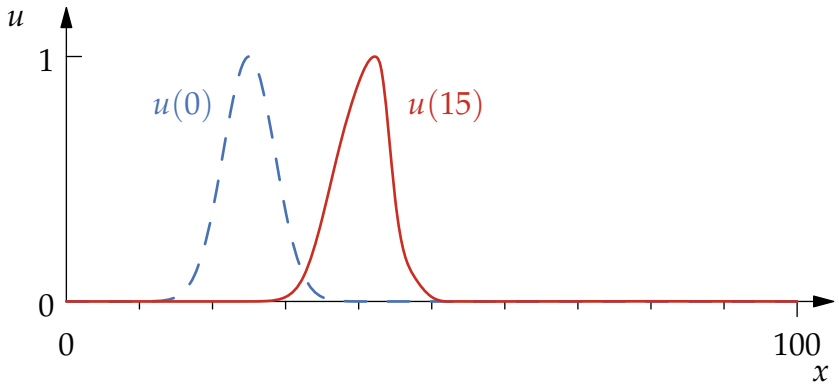


Abbildung 5.2: Anfangsbedingung $u(0)$ des blow-up-Problems (blau) sowie die Lösung zum Zeitpunkt $T = 15$ (rot).

aufgestellt hat, der blow-up ist zu erahnen. Der Diskretisierungsfehler ist an dieser Stelle zwar am größten, es sind aber keine Schwingungen der numerischen Lösungen an dieser Stelle zu beobachten. Abb. 5.3 und Tab. 5.3 zeigen, dass für $p = 1$ die erwartete lineare Konvergenz für beide Verfahren bestätigt werden kann. Dabei ist der Fehler für beide Verfahren im Wesentlichen gleich.

Die Konvergenzordnung in der Zeit für $p = 2$ scheint besser als linear zu sein. Abb. 5.4a und Tab. 5.4 zeigen für beide Verfahren eher eine quadratische Konvergenzordnung in der Zeit an. Im Ort bestätigt sich dagegen die quadratische Konvergenzordnung, siehe Abb. 5.4b und Tab. 5.5, wenngleich die experimentelle Konvergenzordnung immer leicht unter 2 liegt. Das liegt vermutlich am größeren Fehler der Zeitdiskretisierung.

Wie lässt sich die quadratische Ordnung in der Zeit erklären? Wie wir schon in Bemerkung 5.33 erwähnt haben, tauchen im Konvergenzbeweis nur zwei von der Nichtlinearität abhängige Terme auf, die für die lineare Konvergenzordnung in der Zeit verantwortlich sind. Es ist nachvollziehbar, dass für betragsmäßig kleines λ dieser Fehleranteil entsprechend klein und daher nicht zu beobachten ist.

Abschließend möchten wir kurz auf die Performance des Newton-Verfahrens eingehen. Wir haben schon in Abschnitt 5.2.5.1 angemerkt, dass ein inexaktes Newtonverfahren für reellwertige Probleme schneller ist. Um das zu zeigen, betrachten wir das qlw-cG(1)cG(1)- und das slw-cG(1) cG(1)-Verfahren für $k = h = 0.0625$. Mit i bezeichnen wir den Iterationsschritt.

Wir sehen in Tab. 5.1, dass die exakte Newton-Iteration bei Verwendung des qlw-cG(1) cG(1)-Verfahrens nur knapp ein Drittel der Iterationsschritte der inexakten Newton-Iteration benötigt. Dennoch ist die Rechendauer ungefähr fünf Mal so lang. Das liegt an den zusätzlichen Funktionsauswertungen.

Tab. 5.2 liefert ein ähnliches Ergebnis für das slw-cG(1) cG(1)-Verfahren.

	Inexakt		Exakt	
i	$\|U_i^{\text{kor}}\|$	$\|F(U_i)\|$	$\|U_i^{\text{kor}}\|$	$\|F(U_i)\|$
1	1.30e-01	3.47e-03	1.82e-04	1.43e-01
2	1.39e-02	5.71e-04	4.16e-07	8.50e-04
3	2.28e-03	1.06e-04	1.01e-09	1.97e-06
4	4.25e-04	2.11e-05	2.50e-12	4.78e-09
5	8.42e-05	4.33e-06		
6	1.73e-05	9.12e-07		
7	3.64e-06	1.96e-07		
8	7.81e-07	4.25e-08		
9	1.70e-07	9.33e-09		
10	3.73e-08	2.07e-09		
11	8.26e-09	4.61e-10		

Tabelle 5.1: Inexakte und exakte Newton-Iteration des letzten Zeitschritts des qlw-cG(1)cG(1)-Verfahrens.

	Inexakt		Exakt	
i	$\|U_i^{\text{kor}}\|$	$\|F(U_i)\|$	$\|U_i^{\text{kor}}\|$	$\|F(U_i)\|$
1	1.30e-01	8.64e-04	1.44e-01	1.15e-05
2	1.38e-02	1.39e-04	2.14e-04	5.91e-11
3	2.23e-03	2.55e-05		
4	4.07e-04	4.97e-06		
5	7.95e-05	1.01e-06		
6	1.62e-05	2.11e-07		
7	3.38e-06	4.50e-08		
8	7.20e-07	9.73e-09		
9	1.56e-07	2.13e-09		
10	3.40e-08	4.70e-10		
11	7.51e-09	1.04e-10		

Tabelle 5.2: Inexakte und exakte Newton-Iteration des letzten Zeitschritts des slw-cG(1)cG(1)-Verfahrens.

Für dieses Verfahren benötigt die exakte Newtoniteration sogar nur zwei Schritte mit Konvergenz vierter Ordnung. Auch wenn die inexakte Iteration hier sogar über fünf Mal soviele Iterationen braucht, ist sie vier Mal schneller, das ist wieder auf die Funktionsauswertungen zurückzuführen.

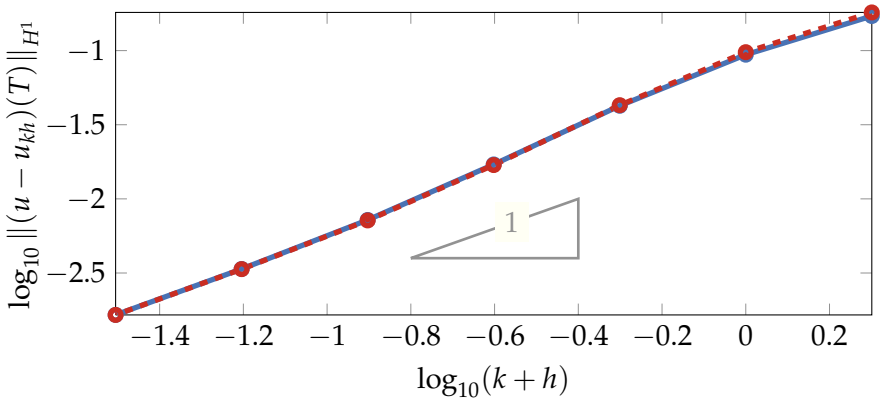
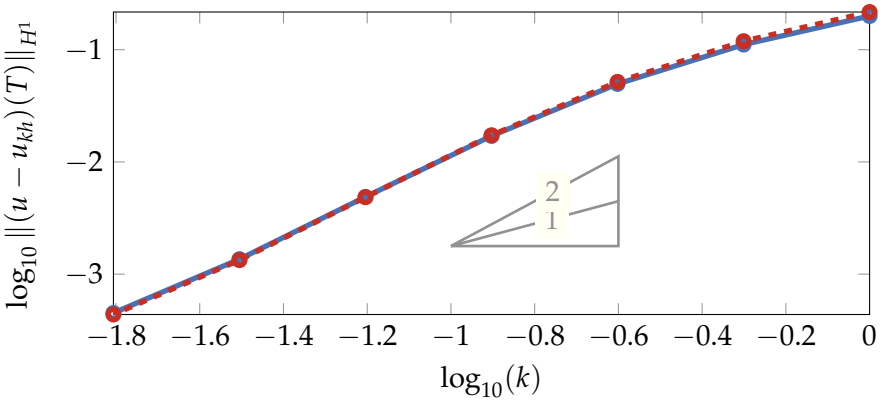
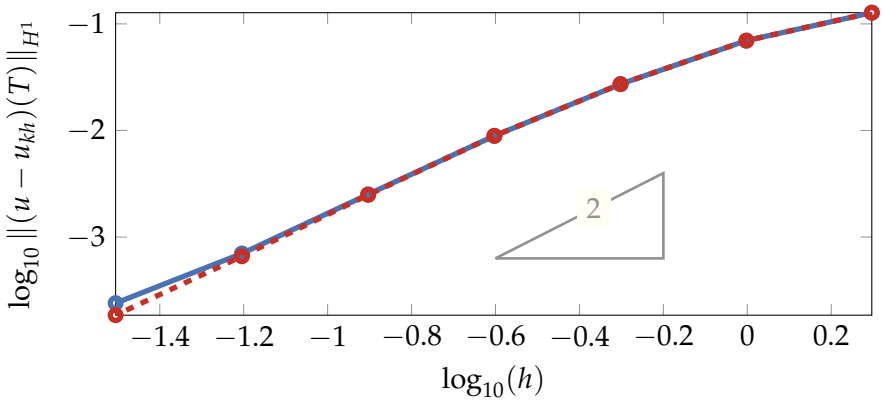


Abbildung 5.3: blow-up-Problem: Konvergenz für $p = 1$ in Abhängigkeit von $k + h$. In blau das qlw-cG(1)cG(1)-Verfahren, in gestrichelt rot das slw-cG(1)cG(1)-Verfahren.



(a) Konvergenz in der Zeit in Abhängigkeit von k, mit $h = 0.03125$ fest.



(b) Konvergenz im Ort in Abhängigkeit von h, mit $k = 0.005$ fest.

Abbildung 5.4: blow-up-Problem: Konvergenz für $p = 2$. In blau das qlw-cG(1)cG(2)-Verfahren, in gestrichelt rot das slw-cG(1)cG(2)-Verfahren.

		qlw-cG(1)cG(1)		slw-cG(1)cG(1)	
k	h	$\|(u-u_{kh})(T)\|_{H^1}$	eoc	$\|(u-u_{kh})(T)\|_{H^1}$	eoc
1.0	1.0	1.71e-01		1.81e-01	
0.5	0.5	9.38e-02	8.66e-01	9.76e-02	8.90e-01
0.25	0.25	4.25e-02	1.14e+00	4.28e-02	1.19e+00
0.125	0.125	1.71e-02	1.31e+00	1.69e-02	1.34e+00
0.0625	0.0625	7.21e-03	1.24e+00	7.15e-03	1.24e+00
0.03125	0.03125	3.37e-03	1.10e+00	3.36e-03	1.09e+00
0.015625	0.015625	1.65e-03	1.03e+00	1.65e-03	1.03e+00

Tabelle 5.3: blow-up-Problem: Konvergenz des qlw- und des slw-cG(1)cG(1)-Verfahrens für lineare Elemente ($p = 1$).

		qlw-cG(1)cG(2)		slw-cG(1)cG(2)	
k	h	$\|(u-u_{kh})(T)\|_{H^1}$	eoc	$\|(u-u_{kh})(T)\|_{H^1}$	eoc
1.0	0.03125	1.20e-01		2.17e-01	
0.5	0.03125	1.11e-01	8.46e-01	1.19e-01	8.64e-01
0.25	0.03125	4.95e-02	1.17e+00	5.17e-02	1.20e+00
0.125	0.03125	1.71e-02	1.54e+00	1.73e-02	1.58e+00
0.0625	0.03125	4.88e-03	1.81e+00	4.85e-03	1.83e+00
0.03125	0.03125	1.36e-03	1.84e+00	1.33e-03	1.86e+00
0.015625	0.03125	4.51e-04	1.59e+00	4.37e-04	1.61e+00

Tabelle 5.4: blow-up-Problem: Konvergenz in der Zeit des qlw- und des slw-cG(1)cG(2)-Verfahrens für quadratische Elemente ($p = 2$).

		qlw-cG(1)cG(2)		slw-cG(1)cG(2)	
k	h	$\|(u-u_{kh})(T)\|_{H^1}$	eoc	$\|(u-u_{kh})(T)\|_{H^1}$	eoc
0.005	2.0	1.27e-01		1.27e-01	
0.005	1.0	6.96e-02	8.78e-01	6.96e-02	8.78e-01
0.005	0.5	2.72e-02	1.36e+00	2.72e-02	1.36e+00
0.005	0.25	8.91e-03	1.61e+00	8.90e-03	1.62e+00
0.005	0.125	2.52e-03	1.82e+00	2.49e-03	1.84e+00
0.005	0.0625	7.01e-04	1.84e+00	6.60e-04	1.92e+00
0.005	0.03125	2.40e-04	1.55e+00	1.85e-04	1.84e+00

Tabelle 5.5: blow-up-Problem: Konvergenz im Ort des qlw- und des slw-cG(1)cG(2)-Verfahrens für quadratische Elemente ($p = 2$).

5.4.2 Global existente Pulspropagation, $\lambda < 0$

Im zweiten Testproblem überprüfen wir die Konvergenz für ein inhomogenes Problem, das heißt ein Problem mit einem nicht verschwindenden Kraftterm g. Dabei entsteht ein zusätzlicher Fehler, da wir die rechte Seite g nicht exakt behandeln, wie wir es in den Verfahrensdefinitionen getan haben. Stattdessen verwenden wir eine Quadraturformel in der Zeit mit genügend hoher Genauigkeit und im Ort die Knoteninterpolation in $\mathcal{S}_h^p$. In der Zeit ist eine Quadratur der Ordnung 2 ausreichend, wir verwenden die Trapezregel.

Wir definieren auf $\Omega = (0, 10)$ und für $t \in (0, T)$ mit $T = 5$ den Kraftterm

$$g(t, x) = 6\lambda(24(x - t - 2)^2 - 2)\mathrm{e}^{-6(x-t-2)^2}.$$

Dann ist die Lösung der zugehörigen Differentialgleichung (3.6) durch den reskalierten Gauß-Puls

$$u(t, x) := \mathrm{e}^{-2(x-t-2)^2}$$

gegeben. Weiterhin setzen wir $\lambda = -0.1$. u können wir offenbar auch für alle $t \geq 0$ definieren, da wir den nichtlinearen Einfluss durch den Kraftterm g eliminieren. Daher tritt anders als im vorherigen Beispiel in diesem Fall auch kein blow-up auf.

Die Ergebnisse liefern die gleichen Beobachtungen wie vorhergehenden im blow-up-Problem. Für lineare Elemente $p = 1$ zeigen Abb. 5.5 und Tab. 5.6 ein Konvergenzverhalten, das leicht besser ist als die erwartete lineare Konvergenz. Nun haben wir schon beim blow-up-Problem gesehen, dass die Konvergenz in der Zeit besser als linear zu sein scheint, aber die Fehlerkonstanten bezüglich der Zeitdiskretisierung höher sind als die der Ortsdiskretisierung. Es kann also auch hier der Fall sein, dass für die betrachteten Diskretisierungsparameter der Zeitfehler dominiert und ein etwas besseres Konvergenzverhalten besitzt als die Ortsdiskretisierung.

Für quadratische Elemente $p = 2$ bestätigt sich bei Betrachtung der Zeitkonvergenz in Abb. 5.6a und Tab. 5.7 die Vermutung, dass zumindest die experimentelle Konvergenzordnung in der Zeit höher ist als theoretisch hergeleitet. Beide Verfahren zeigen ein quadratisches Konvergenzverhalten in der Zeit. Gleiches gilt für die Konvergenz im Ort, siehe Abb. 5.6b und Tab. 5.8, allerdings entspricht das dem Konvergenzresultat Satz 5.32.

Auch für dieses Testproblem sind die Fehler beider Verfahren im Wesentlichen die gleichen.

Das Newtonverfahren verhält sich qualitativ ähnlich wie im vorherigen Problem. Die inexakten Newtoniterationen sind auch hier wieder schneller als die exakten Newtoniterationen, die quadratische Konvergenz des exakten Newtonverfahrens sieht man nur beim slw-cG(1) cG(1)-Verfahren.

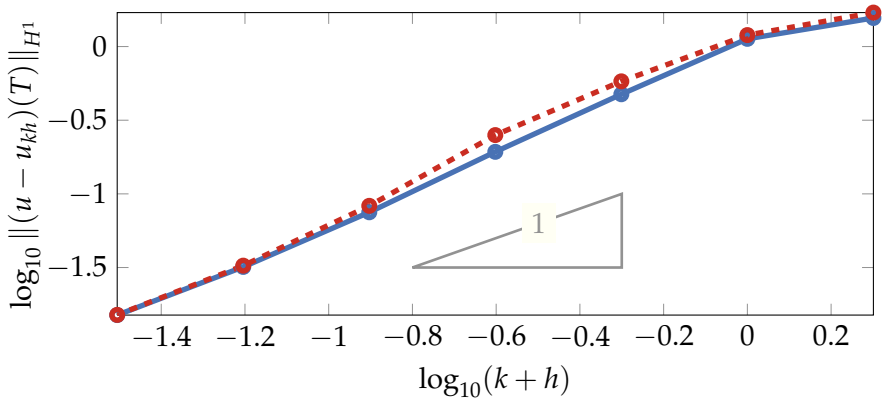
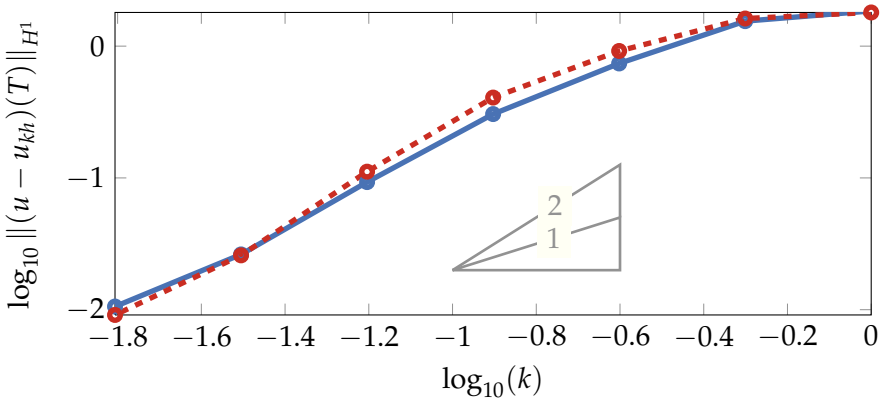
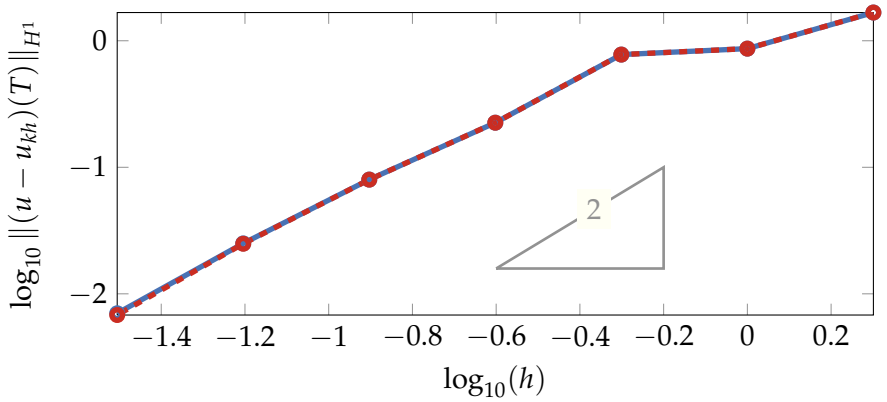


Abbildung 5.5: Pulsausbreitungsproblem: Konvergenz für $p = 1$ in Abhängigkeit von $k + h$. In blau das qlw-cG(1)cG(1)-Verfahren, in gestrichelt rot das slw-cG(1)cG(1)-Verfahren.



(a) Konvergenz in der Zeit in Abhängigkeit von k, mit $h = 0.03125$ fest.



(b) Konvergenz im Ort in Abhängigkeit von h, mit $k = 0.005$ fest.

Abbildung 5.6: Pulsausbreitungsproblem: Konvergenz für $p = 2$. In blau das qlw-cG(1)cG(2)-, in gestrichelt rot das slw-cG(1)cG(2)-Verfahren.

		qlw-cG(1)cG(1)		slw-cG(1)cG(1)	
k	h	$\|(u-u_{kh})(T)\|_{H^1}$	eoc	$\|(u-u_{kh})(T)\|_{H^1}$	eoc
1.0	1.0	1.56e+00		1.70e+00	
0.5	0.5	1.13e+00	4.70e-01	1.20e+00	5.07e-01
0.25	0.25	4.74e-01	1.25e+00	5.81e-01	1.04e+00
0.125	0.125	1.93e-01	1.30e+00	2.51e-01	1.21e+00
0.0625	0.0625	7.50e-02	1.36e+00	8.29e-02	1.60e+00
0.03125	0.03125	3.19e-02	1.23e+00	3.26e-02	1.35e+00
0.015625	0.015625	1.50e-02	1.09e+00	1.51e-02	1.11e+00

Tabelle 5.6: Pulsausbreitungsproblem: Konvergenztabellen des qlw- und des slw-cG(1)cG(1)-Verfahrens für lineare Elemente ($p = 1$).

		qlw-cG(1)cG(2)		slw-cG(1)cG(2)	
k	h	$\|(u-u_{kh})(T)\|_{H^1}$	eoc	$\|(u-u_{kh})(T)\|_{H^1}$	eoc
1.0	0.03125	1.85e+00		1.80e+00	
0.5	0.03125	1.55e+00	2.58e-01	1.62e+00	1.53e-01
0.25	0.03125	7.39e-01	1.07e+00	9.19e-01	8.18e-01
0.125	0.03125	3.05e-01	1.28e+00	4.07e-01	1.17e+00
0.0625	0.03125	9.30e-02	1.71e+00	1.12e-01	1.87e+00
0.03125	0.03125	2.63e-02	1.82e+00	2.58e-02	2.11e+00
0.015625	0.03125	1.06e-02	1.31e+00	9.12e-03	1.50e+00

Tabelle 5.7: Pulsausbreitungsproblem: Konvergenz in der Zeit des qlw- und des slw-cG(1)cG(2)-Verfahrens für quadratische Elemente ($p = 2$).

		qlw-cG(1)cG(2)		slw-cG(1)cG(2)	
k	h	$\|(u-u_{kh})(T)\|_{H^1}$	eoc	$\|(u-u_{kh})(T)\|_{H^1}$	eoc
0.005	2.0	1.67e+00		1.67e+00	
0.005	1.0	8.67e-01	9.48e-01	8.67e-01	9.48e-01
0.005	0.5	7.77e-01	1.56e-01	7.77e-01	1.56e-01
0.005	0.25	2.25e-01	1.79e+00	2.25e-01	1.79e+00
0.005	0.125	8.00e-02	1.49e+00	7.99e-02	1.49e+00
0.005	0.0625	2.50e-02	1.68e+00	2.49e-02	1.68e+00
0.005	0.03125	7.02e-03	1.83e+00	6.80e-03	1.87e+00

Tabelle 5.8: Pulsausbreitungsproblem: Konvergenz im Ort des qlw- und des slw-cG(1)cG(2)-Verfahrens für quadratische Elemente ($p = 2$).

5.4.3 Soliton, $\lambda > 0$

Aus der Herleitung der Schrödingergleichung aus den Maxwellgleichungen können wir uns den Ansatz

$$u(t, x_1, x_2) = \mathrm{e}^{\mathrm{i}(\mathrm{k}x_1 - \omega t)} b(x_2)$$

abschauen. Dabei bezeichnen $\mathrm{k}, \omega \in \mathbb{R}$ die Wellenzahl und die Frequenz der Welle. Einsetzen dieses Ansatzes in das Modellproblem Gl. (3.6) mit $g = 0$ ergibt

$$-\omega^2(b + \lambda|b|^2 b) = -\mathrm{k}^2 b + \partial_{x_2}^2 b \,.$$

Dies lässt sich umschreiben zu

$$(\mathrm{k}^2 - \omega^2)b - \lambda\omega^2|b|^2 b = \partial_{x_2}^2 b \,.$$

Wir setzen $\lambda = 1$, dies entspricht dem defokussierenden Fall bei der nichtlinearen Schrödingergleichung. Damit der Koeffizient des Reaktionsteils positiv ist, setzen wir also für ein $a > 0$

$$\begin{aligned} \omega &:= a \,, \\ \mathrm{k} &:= \sqrt{2}a \,, \end{aligned}$$

woraus wir erhalten:

$$a^2(b + |b|^2 b) = \partial_{x_2}^2 b \,.$$

Diese Differentialgleichung wird gelöst durch

$$b(x_2) = \sqrt{2}\,\mathrm{sech}(ax_2) = \sqrt{2}\frac{1}{\cosh(ax_2)} \,.$$

Es folgt

$$u(t, x_1, x_2) = \sqrt{2}\mathrm{e}^{\mathrm{i}a(\sqrt{2}x_1 - t)}\frac{1}{\cosh(ax_2)} \,.$$

Für die Experimente setzen wir $\Omega = (0, 2) \times (-1, 1)$ und $T = 1.0$. Für $a = 1$ betrachten wir also die Funktion

$$u(t, x_1, x_2) = \sqrt{2}\mathrm{e}^{\mathrm{i}(\sqrt{2}x_1 - t)}\,\mathrm{sech}(x_2).$$

Wie wir anfangs dieses Abschnittes erwähnt haben, verwenden wir zur Triangulierung von Ω ein criss-Gitter. h ist dann die Länge der Hypotenuse eines Elementes. In den Konvergenztabellen geben wir die Gitterweite in Form von $h/\sqrt{2}$ an, das ist gerade die Länge der kürzeren Seite eines Elementes. Die Werte für k und h in den Tabellen 5.10 und 5.12 basieren darauf, dass

wir Ω in Tab. 5.10 in $10, 20, 40, 60, 70$ und 80 Teilintervalle und in Tab. 5.12 in $5, 10, 20, 30, 35$ und 40 Teilintervalle jeweils in x_1- und x_2-Richtung unterteilt haben. Mit einem entsprechenden Verhältnis wird auch k jeweils gewählt, die angegebenen Werte sind nicht exakt. Der Grund für die ungleichmäßige Verfeinerung ist der hohe Speicherverbrauch. Die Randwerte von u werden in jedem Zeitschritt erzwungen, wie man es bei Finite-Elemente-Methoden gewohnt ist.

Bei diesem zweidimensionalen Problem sehen wir jetzt zum einzigen Mal Unterschiede zwischen den beiden Verfahren. Für lineare Elemente lässt sich in Abb. 5.7 zwar beobachten, dass das qlw-cG(1) cG(1)-Verfahren leicht besser ist, aber beide Verfahren besitzen eine lineare experimentelle Konvergenzordnung, das kann man auch genauer in Tab. 5.10 ablesen. Wir werden in der genaueren Aufspaltung für quadratische Elemente sehen, dass es der Fehler in der Zeitdiskretisierung ist, der für die größere Fehlerkonstante von slw-cG(1) cG(p) verantwortlich ist.

Für quadratische Elemente betrachten wir wieder Zeit- und Ortskonvergenz getrennt. In der Abb. 5.8a und der Tab. 5.11 können wir beobachten, dass das qlw-cG(1) cG(2)-Verfahren in der Zeit eine quadratische experimentelle Konvergenzordnung besitzt und für den kleinsten Zeitschritt konvergiert der Fehler offenbar schon gegen den Fehler der Ortsdiskretisierung. Das slw-cG(1) cG(2)-Verfahren zeigt dagegen für die ersten Zeitschritte ein unerklärliches Verhalten, konvergiert dann aber auch quadratisch. Für die Überprüfung der Ortskonvergenz wurde nun $k = 0.005$ ausgesprochen klein gewählt, in Abb. 5.8b und Tab. 5.12 sieht man daher keinen Unterschied zwischen den Verfahren, beide haben eine sehr klare quadratische experimentelle Konvergenzordnung im Ort.

Warum der deutlich unterschiedliche Zeitfehler auftritt, ist nicht klar. Es steht zu vermuten, dass sich in der Implementierung der quadratischen Elemente doch irgendwo ein kleiner Fehler befindet. Ausgeschlossen werden kann nur ein systematischer Fehler, der durch die Produktapproximation eingeführt wurde. Die Berechnungen wurden zur Überprüfung auch ohne Produktapproximation durchgeführt, ohne dass sich ein nennenswerter Unterschied ergab.

Auch hier ist wieder die quadratische Konvergenzordnung in der Zeit zu beobachten. In Abschnitt 5.4.1 haben wir dieses Verhalten noch damit erklärt, dass λ klein war. In diesem Soliton-Problem ist aber $\lambda = 1$, die Nichtlinearität hat also die gleiche Größenordnung wie die linearen Terme. Entweder ist also der Konvergenzbeweis nicht optimal geführt worden oder der Fehleranteil der nichtlinearen Terme ist dennoch deutlich kleiner als der Fehleranteil der linearen Fehleranteile.

Eine Anwendung des inexakten Newtonverfahrens ist in diesem Problem nicht möglich, da es nicht konvergiert. Diese Aussage trifft für beide cG-Verfahren zu. Das exakte Newtonverfahren konvergiert dafür quadratisch in beiden Fällen, wie wir in Tab. 5.9 sehen können. Dabei haben wir $h = 0.05\sqrt{2}$

	qlw-cG(1) cG(1)		slw-cG(1) cG(1)	
i	$\|U_i^{\text{kor}}\|$	$\|F(U_i)\|$	$\|U_i^{\text{kor}}\|$	$\|F(U_i)\|$
1	1.20e+01	9.93e-02	1.18e+01	5.88e-03
2	1.48e+00	4.79e-03	3.85e-01	1.97e-05
3	8.30e-02	1.51e-05	1.32e-03	2.46e-10
4	2.85e-04	1.82e-10	1.69e-08	3.46e-16
5	3.69e-09	1.37e-15		

Tabelle 5.9: Die exakte Newton-Iteration des letzten Zeitschritts für das qlw- und das slw-cG(1)cG(1)-Verfahren.

und $k = 0.25$ verwendet. Beide Verfahren benötigen im Wesentlichen die gleiche Rechendauer.

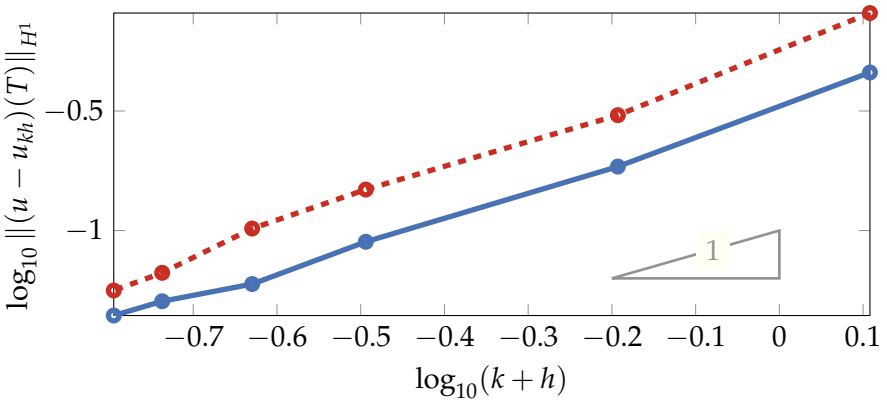
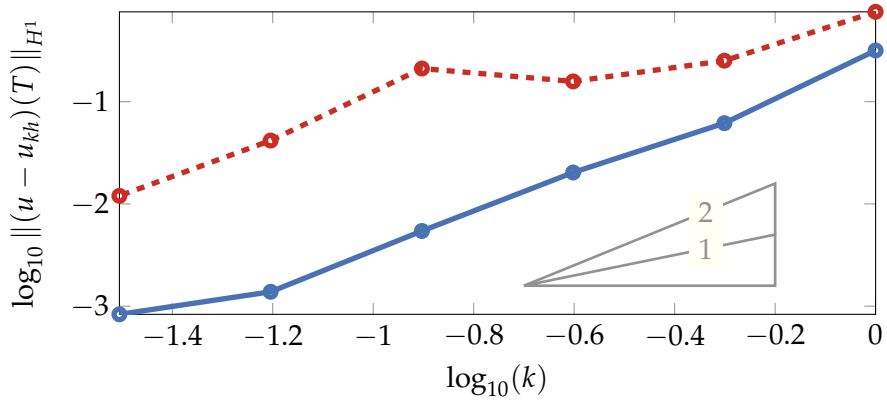
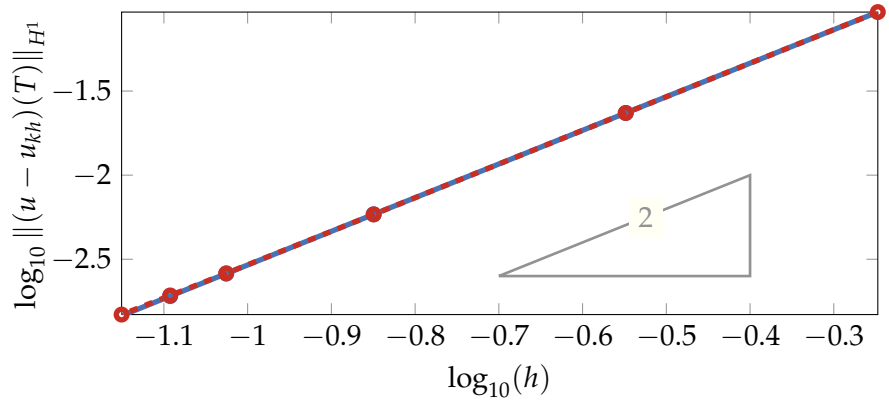


Abbildung 5.7: Solitonproblem: Konvergenz für lineare Elemente, $p = 1$, in Abhängigkeit von $k + h$. In blau das qlw-cG(1)cG(1)-Verfahren, in gestrichelt rot das slw-cG(1)cG(1)-Verfahren.



(a) Konvergenz in der Zeit in Abhängigkeit von k, mit $h = 0.04\sqrt{2}$ fest.



(b) Konvergenz im Ort in Abhängigkeit von h, mit $k = 0.005$ fest.

Abbildung 5.8: Solitonproblem: Konvergenz für $p = 2$. In blau das qlw-cG(1)cG(2)-Verfahren, in gestrichelt rot das slw-cG(1)cG(2)-Verfahren.

		qlw-cG(1)cG(1)		slw-cG(1)cG(1)	
k	$h/\sqrt{2}$	$\|(u-u_{kh})(T)\|_{H^1}$	eoc	$\|(u-u_{kh})(T)\|_{H^1}$	eoc
1.0	0.2	4.58e-01		8.12e-01	
0.5	0.1	1.85e-01	1.31e+00	3.04e-01	1.42e+00
0.25	0.05	8.97e-02	1.04e+00	1.48e-01	1.03e+00
0.1875	0.033	5.97e-02	1.30e+00	1.02e-01	1.20e+00
0.1428	0.02857	5.06e-02	6.70e-01	6.66e-02	1.72e+00
0.125	0.025	4.41e-02	1.03e+00	5.61e-02	1.28e+00

Tabelle 5.10: Solitonproblem: Konvergenz des qlw- und des slw-cG(1)cG(p)-Verfahrens für lineare Elemente ($p = 1$) in Abhängigkeit von $k + h$.

		qlw-cG(1)cG(2)		slw-cG(1)cG(2)	
k	$h/\sqrt{2}$	$\|(u-u_{kh})(T)\|_{H^1}$	eoc	$\|(u-u_{kh})(T)\|_{H^1}$	eoc
1.0	0.04	3.17e-01		7.57e-01	
0.5	0.04	6.17e-02	2.36e+00	2.52e-01	1.59e+00
0.25	0.04	2.03e-02	1.60e+00	1.58e-01	6.74e-01
0.125	0.04	5.44e-03	1.90e+00	2.11e-01	-4.19e-01
0.0625	0.04	1.38e-03	1.98e+00	4.17e-02	2.34e+00
0.03125	0.04	8.32e-04	7.32e-01	1.20e-02	1.80e+00

Tabelle 5.11: Solitonproblem: Konvergenz in der Zeit des qlw- und des slw-cG(1)cG(2)-Verfahrens.

		qlw-cG(1)cG(2)		slw-cG(1)cG(2)	
k	$h/\sqrt{2}$	$\|(u-u_{kh})(T)\|_{H^1}$	eoc	$\|(u-u_{kh})(T)\|_{H^1}$	eoc
0.005	0.4	9.32e-02		9.32e-02	
0.005	0.2	2.34e-02	1.99e+00	2.34e-02	1.99e+00
0.005	0.1	5.85e-03	2.00e+00	5.84e-03	2.00e+00
0.005	0.0667	2.60e-03	2.00e+00	2.61e-03	1.99e+00
0.005	0.057	1.91e-03	2.00e+00	1.93e-03	1.96e+00
0.005	0.05	1.46e-03	2.01e+00	1.48e-03	1.96e+00

Tabelle 5.12: Solitonproblem: Konvergenz im Ort des qlw- und des slw-cG(1)cG(2)-Verfahrens.

5.5 Zusammenfassung und Fazit

In diesem Kapitel haben wir zwei Verfahren zur Diskretisierung der nichtlinearen Wellengleichung (3.6) vorgestellt.

Für das qlw-cG(1) cG(p)-Verfahren haben wir in Satz 5.32 lineare Konvergenz in der Zeit und Konvergenz der Ordnung p im Ort für die Energienorm der Wellengleichung gezeigt. Die numerischen Ergebnisse liefern sogar die optimale Konvergenzordnung, quadratisch in Zeit und Ort, wie wir es schon in Bemerkung 5.33 angesprochen haben. Wir konnten nicht bestimmen, woran das genau liegt, auch wenn wir in Abschnitt 5.4.1 die Vermutung ausgesprochen haben, dass die Fehlerterme, die nur linear konvergieren, möglicherweise deutlich kleiner sind als die quadratisch konvergierenden und daher nicht beobachtet wurden. Weiterhin sind die Existenz- und Eindeutigkeitsbedingungen dieses numerischen Verfahren für das Solitonproblem nicht erfüllt, was darauf hinweist, dass die Existenz- und Eindeutigkeitsresultate in Abschnitt 5.2.2 nicht optimal sind.

Die Konvergenz des zweiten Verfahrens, slw-cG(1) cG(p), konnten wir nicht zeigen, die numerischen Beispiele zeigen aber kaum einen Unterschied zwischen den beiden Verfahren.

Es bleibt zu bemerken, dass das qlw-cG(1) cG(p)-Verfahren aufgrund der in jedem Newtonschritt durchzuführenden Funktionsauswertungen meist etwas mehr Rechenzeit benötigt als das slw-cG(1) cG(p)-Verfahren.

Mit geeigneter Definition der Verfahren müsste es eigentlich auch möglich sein, Verfahren höherer Ordnung in der Zeit zu konstruieren, das ist mir aber nicht gelungen.

Galerkin-Verfahren mit lokalem Zeitschritt

6

In diesem Kapitel beschäftigen wir uns mit der Anpassung des im vorherigen Kapitel eingeführten Zeitschrittverfahrens für die Möglichkeit lokaler Zeitschritte. Es ist bei diesem Verfahren kein Problem, ein beliebiges Zeitgitter zu definieren; allerdings gilt die in jedem Zeitschritt festgelegte Zeitschrittweite k_n für alle Knoten im Ort. In [DG09] und [GM10] wurde aber folgendes Problem vorgestellt: Wenn die Geometrie des Problems eine lokal deutlich höhere Auflösung erfordert als global notwendig wäre, dann bedarf es aus Stabilitätsgründen bei expliziten Zeitschrittverfahren eines der kleinsten Gitterweite im Ort entsprechenden Zeitschrittes. Das lieferte die Idee, ein Verfahren zu entwerfen, welches nur eine lokal (im Ort) verkleinerten Zeitschrittweite benötigt.

Die Klasse der cG-Verfahren in der Zeit hat die gerade beschriebenen Stabilitätsprobleme nicht, da es sich um implizite Verfahren handelt. Die Umsetzung eines lokalen Zeitschrittes hat einen anderen Sinn, der mit der Differentialgleichung und zunächst nicht mit der Geometrie des Problems zusammenhängt. Wie wir zuvor gesehen haben, besitzt unsere nichtlineare Wellengleichung nur für eine endliche Zeit eine Lösung. Anhand der Beweise, aber auch anhand der numerischen Simulationen kann man sehen, dass diese endliche Existenzzeit häufig durch einen blow-up der H^1-Norm der Funktion charakterisiert wird, das heißt die örtliche Ableitung strebt gegen Unendlich. Der numerische Fehler steigt gegen Ende des Existenzintervalls daher deutlich an, wenn man nicht die Zeitschrittweite entsprechend verkleinert. Eine solche Anpassung wäre global zu kostspielig, wir möchten dies auf ein Intervall im Ort beschränken, in dem dieser blow-up auftritt. Ein anderer Punkt ist der Fehler, den ein Zeitschrittverfahren bei einer lokal verfeinerten Geometrie macht. Löst man nicht gleichzeitig höher in der Zeit auf, wenn sich eine Welle durch ein lokal verfeinertes Gebiet bewegt, dann ist der Fehler der Zeitdiskretisierung in diesen Zeitschritten unter Umständen im Ver-

gleich zum Ortsdiskretisierungsfehler zu groß und dominiert. Es bietet sich daher an, den Zeitschritt mit der Ortsdiskretisierung zu koppeln.

Im Gegensatz zur Motivation der zitierten Arbeiten geht es uns also nicht um Stabilität, sondern um Genauigkeit. Zusätzlich zu einem Verfahren, welches lokale Zeitschritte erlaubt, wäre eine Strategie vonnöten, die uns sagt, wo wir wie oft lokal verfeinern müssen. Ersteres möchten wir in den nächsten Abschnitten vorstellen, letzteres kann diese Arbeit nicht leisten.

Die Idee des lokalen Zeitschrittverfahrens in [DG09] und [GM10] besteht im Wesentlichen aus drei Punkten:

1. Diskretisiere das ganze Problem zunächst mit einem expliziten Zeitschrittverfahren in der Zeit und beliebig im Ort.

2. Unterscheide im Lösungsvektor den Anteil $z^{[\text{coarse}]}$, in dem ein großer Zeitschritt durchgeführt werden soll und den Anteil $z^{[\text{fine}]}$, in dem ein kleiner Zeitschritt durchgeführt werden soll.

3. Ersetze $z^{[\text{fine}]}$ durch eine Lösung einer modifizierten Differentialgleichung, um den feineren Zeitschritt definieren zu können.

Im Finite-Elemente-Kontext kann ein analoges Vorgehen systematischer behandelt werden. Eigentlich muss man nur in der Linienmethode jedem Ortsgitterpunkt einen eigenen Finite-Elemente-Raum in der Zeit zuordnen, um ein lokales (da für jeden Ortspunkt mit unterschiedlichem Zeitschritt funktionierendes) Zeitschrittverfahren zu erhalten.

In Abschnitt 6.1 definieren wir das lokale Zeitschrittverfahren für die (lineare) Wellengleichung. Im zweiten Abschnitt 6.2 erweitern wir dieses Konzept auf unsere nichtlineare Wellengleichung in der Form von (3.25). In Abschnitt 6.3 gehen wir dann darauf ein, wie die Matrizen aufgestellt werden, die im Zeitschrittverfahren auftauchen und gehen kurz auf die Besetzungsstruktur dieser Matrizen ein. Die numerischen Beispiele in Abschnitt 6.4 zeigen dann, wie gut sich die Methode bei linearen und nichtlinearen Problemen verhält. Zu guter Letzt diskutieren wir das Verfahren und die erhaltenen Ergebnisse in Abschnitt 6.5.

6.1 Lokaler Zeitschritt für lineare Probleme

Wir stellen zunächst die Diskretisierung für ein lineares Problem vor, damit wir uns auf die Struktur des entstehenden Gleichungssystems konzentrieren können. Wir unterscheiden dabei drei Stufen in der Diskretisierung:

1. Wir beginnen mit einem Verfahren ohne Verfeinerung.

2. Danach betrachten wir die globale Verfeinerung in der Zeit.

3. Und dann leiten wir daraus das Verfahren mit lokaler Verfeinerung her.

6.1.1 Allgemeine Formulierung des Galerkin-Verfahrens

Unser Ausgangsproblem ist die lineare Wellengleichung. Es sei $T > 0$ und wir suchen eine Lösung $u\colon (0,T) \times \Omega \to \mathbb{R}$ mit

$$\begin{aligned} \partial_t^2 u - \Delta u &= g && \text{in } (0,T) \times \Omega, \\ u &= 0 && \text{auf } (0,T) \times \partial\Omega, \end{aligned}$$

mit den Anfangsbedingungen $u(0,\cdot) = u_0$ und $\partial_t u(0,\cdot) = v_0$. Der Einfachheit halber setzen wir $g = 0$. Wir betrachten die schwache Formulierung auf $(0,T) \times \Omega$ und suchen $u \in H^1(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$ und $v \in H^1(0,T;L^2(\Omega))$ mit obigen Anfangsbedingungen $u(0) = u_0 \in H_0^1(\Omega)$ und $v(0) = v_0 \in L^2(\Omega)$, und

$$\int_0^T (\partial_t u, \phi_1)\,\mathrm{d}t = \int_0^T (v, \phi_1)\,\mathrm{d}t, \tag{6.1a}$$

$$\int_0^T (\partial_t v, \phi_2)\,\mathrm{d}t = -\int_0^T a(u, \phi_2)\,\mathrm{d}t \tag{6.1b}$$

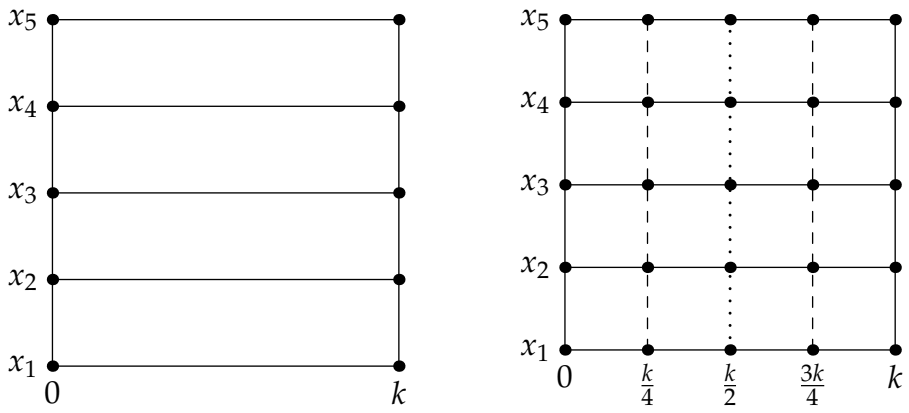
für alle $\phi_1, \phi_2 \in L^2((0,T), H_0^1(\Omega))$, wobei $a(u,\phi_2) := \int_\Omega \nabla u \cdot \nabla \phi_2 \,\mathrm{d}x$. Da u, v als stetig in der Zeit aufgefasst werden können, sind die Anfangsbedingungen wohldefiniert.

Formulierung mit Verfeinerung

Wir diskretisieren dieses Problem wieder mit Hilfe eines Tensorproduktansatzes, wie wir es schon im vorherigen Kapitel 5 getan haben. Das erlaubt es uns, die Diskretisierung in Ort und Zeit unabhängig voneinander zu betrachten. Im Ort verwenden wir einen beliebigen H_0^1-konformen Ansatz mit dem Ansatzraum $\mathcal{S}_h$ und einer Basis $\{\varphi_\nu\colon \nu \in \mathcal{N}_h\}$ bezüglich der Menge der inneren Knoten $\mathcal{N}_h$. Die angesprochene Lokalität des Zeitschrittes soll sich auf $\mathcal{N}_h$ beziehen: Wir wählen problembedingt Knoten aus $\mathcal{N}_h$ aus, in denen wir einen kleineren Zeitschritt verwenden möchten.

Zum besseren Verständnis der Struktur des linearen Gleichungssystems, das wir später erhalten werden, behandeln wir zunächst alle Ortsknoten $\nu \in \mathcal{N}_h$ gleich, beginnen mit einem festen $k > 0$ und definieren das *Grundintervall* $I = [0,k]$. Wir schreiben $\Omega_k := I \times \Omega$. Im vorherigen Kapitel haben wir im Prinzip ein Galerkin-Verfahren auf diesem Intervall definiert und in jedem Zeitschritt nur die Anfangswerte geändert. Wir verfeinern dieses Grundintervall durch wiederholte Bisektion und erhalten zu einem (noch ν-unabhängigem) *Verfeinerungslevel* $l \in \mathbb{N}_0$ das *(gleichmäßige) Zeitgitter zum Level l*

$$\mathcal{T}^l := \left\{ t_{l,i} := \frac{i}{2^l} k\colon i = 0, \dots, 2^l \right\}$$



(a) Ohne Verfeinerung. **(b)** Globale Verfeinerung, $l = 2$.

Abbildung 6.1: Das Zeit-Ort-Gitter bezüglich des Grundintervalls und das in der Zeit global zweimal verfeinerte Zeit-Ort-Gitter.

und bezeichnen mit

$$I^{l,i} := [t_{l,i}, t_{l,i+1}] \qquad (i = 0, \dots, 2^l - 1),$$

die Teilintervalle dieses Gitters. In Abb. 6.1 haben wir das Zeit-Ort-Gitter einmal ohne Verfeinerung und einmal mit globaler Verfeinerung des Levels $l = 2$ dargestellt. Es sei $\mathcal{S}_k = \mathcal{S}_k(I)$ der Finite-Elemente-Raum der linearen Elemente auf I bezüglich des Gitters $\mathcal{T}^l$ und $\mathcal{B}^l = \{\psi^{l,i} \colon i = 0, \dots, 2^l\}$ eine beliebige Basis von $\mathcal{S}_k$. Den Tensorproduktansatzraum über $I \times \Omega$ bezeichnen wir mit

$$\mathcal{V}_1^l := \mathcal{S}_k \otimes \mathcal{S}_h = \operatorname{span}\Big\{\phi \colon \Omega_k \to \mathbb{R} \mid \phi(t,x) = \psi(t)\varphi_\nu(x),\ \nu \in \mathcal{N}_h, \psi \in \mathcal{B}^l\Big\}. \quad (6.2)$$

Die indizierte 1 des Raumes kennzeichnet die Wahl linearer Elemente in der Zeit, darin unterscheidet er sich vom Testraum

$$\mathcal{V}_0^l := \operatorname{span}\Big\{\phi \colon \Omega_k \to \mathbb{R} \mid \phi(t,x) = \mathbb{1}_{I_{l,i}}(t)\varphi_\nu(x) \colon i = 0, \dots, 2^l - 1, \nu \in \mathcal{N}_h\Big\}. \quad (6.3)$$

Mit diesen Vorbereitungen können wir unser vorläufiges Problem stellen. Wir suchen $u_{kh}, v_{kh} \in \mathcal{V}_1^l$ mit

$$\int_I (\partial_t u_{kh}, \phi_1)\, \mathrm{d}t = \int_I (v_{kh}, \phi_1)\, \mathrm{d}t, \quad (6.4a)$$

$$\int_I (\partial_t v_{kh}, \phi_2)\, \mathrm{d}t = -\int_I a(u_{kh}, \phi_2)\, \mathrm{d}t, \quad (6.4b)$$

für alle $\phi_1, \phi_2 \in \mathcal{V}_0^l$. Die Anfangswerte $u_{kh}(0,\cdot), v_{kh}(0,\cdot)$ seien geeignet gewählte Approximationen von u_0, v_0 im Raum $\mathcal{S}_h$, beispielsweise die elliptischen oder L^2-Projektionen von u_0, v_0 auf $\mathcal{S}_h$.

Matrix-Vektor-Form

Wir schreiben u_{kh}, v_{kh} entsprechend des gewählten Ansatzes als

$$u_{kh}(t,x) = \sum_{\nu\in\mathcal{N}_h} \left(\sum_{i=0}^{2^l} U_\nu^{l,i} \psi^{l,i}(t) \right) \varphi_\nu(x), \tag{6.5}$$

$$v_{kh}(t,x) = \sum_{\nu\in\mathcal{N}_h} \left(\sum_{i=0}^{2^l} V_\nu^{l,i} \psi^{l,i}(t) \right) \varphi_\nu(x) \tag{6.6}$$

mit den *Koeffizientenvektoren* $U = (U_\nu^{l,i})_{\nu\in\mathcal{N}_h,\ i=0,\dots,2^l}$, $V = (V_\nu^{l,i})_{\nu\in\mathcal{N}_h,\ i=0,\dots,2^l}$. Wählen wir als Basis für den Testraum $\mathcal{V}_0^l$ die Funktionen

$$\{\phi\colon \Omega_k \to \mathbb{R} \mid \phi(t,x) = \mathbb{1}_{I^{l,j}}(t)\varphi_\nu(x),\ j = 0,\dots,2^l,\ \nu \in \mathcal{N}_h\}, \tag{6.7}$$

so können wir (6.4) auch wie folgt schreiben.

$$\begin{aligned}
&\sum_{\nu\in\mathcal{N}_h} \sum_{i=0}^{2^l} U_\nu^{l,i} \left(\int_{I^{l,j}} \partial_t \psi^{l,i}(t)\ \mathrm{d}t \right) (\varphi_\nu, \varphi_{\nu'}) \\
&\qquad = \sum_{\nu\in\mathcal{N}_h} \sum_{i=0}^{2^l} V_\nu^{l,i} \left(\int_{I^{l,j}} \psi^{l,i}(t)\ \mathrm{d}t \right) (\varphi_\nu, \varphi_{\nu'}), \\
&\sum_{\nu\in\mathcal{N}_h} \sum_{i=0}^{2^l} V_\nu^{l,i} \left(\int_{I^{l,j}} \partial_t \psi^{l,i}(t)\ \mathrm{d}t \right) (\varphi_\nu, \varphi_{\nu'}) \\
&\qquad = -\sum_{\nu\in\mathcal{N}_h} \sum_{i=0}^{2^l} U_\nu^{l,i} \left(\int_{I^{l,j}} \psi^{l,i}(t)\ \mathrm{d}t \right) a(\varphi_\nu, \varphi_{\nu'}),
\end{aligned}$$

für alle $j = 0,\dots,2^l - 1$ und $\nu' \in \mathcal{N}_h$. Wir sortieren dies ein wenig um und erhalten

$$\begin{aligned}
&\sum_{i=0}^{2^l} \left(\int_{I^{l,j}} \partial_t \psi^{l,i}(t)\ \mathrm{d}t \right) \sum_{\nu\in\mathcal{N}_h} U_\nu^{l,i} (\varphi_\nu, \varphi_{\nu'}) \\
&\qquad = \sum_{i=0}^{2^l} \left(\int_{I^{l,j}} \psi^{l,i}(t)\ \mathrm{d}t \right) \sum_{\nu\in\mathcal{N}_h} V_\nu^{l,i} (\varphi_\nu, \varphi_{\nu'}),
\end{aligned} \tag{6.8a}$$

$$\sum_{i=0}^{2^l} \left(\int_{I^{l,j}} \partial_t \psi^{l,i}(t)\, \mathrm{d}t \right) \sum_{\nu \in \mathcal{N}_h} V_\nu^{l,i} (\varphi_\nu, \varphi_{\nu'}) = -\sum_{i=0}^{2^l} \left(\int_{I^{l,j}} \psi^{l,i}(t)\, \mathrm{d}t \right) \sum_{\nu \in \mathcal{N}_h} U_\nu^{l,i} a(\varphi_\nu, \varphi_{\nu'}), \tag{6.8b}$$

Die Zeitintegration und die Ortsintegration sind also wie erwartet nicht gekoppelt. Dieses Gleichungssystem möchten wir in Matrix-Vektor-Form bringen. Dazu beachten wir, dass $U^0 = (U_\nu^{l,0})_{\nu \in \mathcal{N}_h}$, $V^0 = (V_\nu^{l,0})_{\nu \in \mathcal{N}_h}$ bekannt sind, wir also in der Summe $i = 0$ herausnehmen können, um uns auf die unbekannten Koeffizienten zu beschränken. Die zu bestimmenden Vektoren seien durch

$$U^1 := (U_\nu^{l,i})_{\nu \in \mathcal{N}_h,\, i=1,\dots,2^l},$$
$$V^1 := (V_\nu^{l,i})_{\nu \in \mathcal{N}_h,\, i=1,\dots,2^l}$$

definiert. Es ist also $U^1, V^1 \in \mathbb{R}^{|\mathcal{N}_h| 2^l}$, im Gegensatz zu $U^0, V^0 \in \mathbb{R}^{|\mathcal{N}_h|}$.

6.1 Definition. (Diskretisierungsmatrizen)
Die *Massematrix* $\mathbf{M}$ und die *Steifigkeitsmatrix* $\mathbf{K}$ der Ortsdiskretisierung seien für die Knoten $\nu, \nu' \in \mathcal{N}_h$ wieder gegeben durch

$$\mathbf{M}_{\nu',\nu} := (\varphi_\nu, \varphi_{\nu'}), \quad \mathbf{K}_{\nu',\nu} := -a(\varphi_\nu, \varphi_{\nu'}).$$

Für die Zeitdiskretisierung definieren wir die *Zeitschrittmatrizen* $\mathbf{Z^t}, \mathbf{Z} \in \mathbb{R}^{2^l, 2^l}$ sowie $\mathbf{Z^{t,0}}, \mathbf{Z^0} \in \mathbb{R}^{2^l}$ durch

$$\mathbf{Z^t}_{j+1,i} := \int_{I^{l,j}} \partial_t \psi^{l,i}\, \mathrm{d}t, \quad \mathbf{Z^{t,0}}_{j+1} := \int_{I^{l,j}} \partial_t \psi^{l,0}\, \mathrm{d}t$$

und

$$\mathbf{Z}_{j+1,i} := \int_{I^{l,j}} \psi^{l,i}\, \mathrm{d}t, \quad \mathbf{Z^0}_{j+1} := \int_{I^{l,j}} \psi^{l,0}\, \mathrm{d}t$$

für $i = 1, \dots, 2^l$ und $j = 0, \dots, 2^l - 1$.

Die in (6.8) auftretende Verknüpfung der Orts- und der Zeitdiskretisierungsmatrizen ist das *Kroneckerprodukt* zweier Matrizen.

6.2 Definition.
Es seien $\mathbf{A} \in \mathbb{R}^{n_1 \times m_1}$ und $\mathbf{B} \in \mathbb{R}^{n_2 \times m_2}$. Dann ist das *Kroneckerprodukt* dieser Matrizen gegeben durch

$$\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{(n_1 n_2) \times (m_1 m_2)},$$
$$(\mathbf{A} \otimes \mathbf{B})_{(i_1-1)n_1+i_2, (j_1-1)n_2+j_2} := \mathbf{A}_{i_1,j_1} \mathbf{B}_{i_2,j_2},$$

wobei $i_1 = 1, \dots, n_1$, $i_2 = 1, \dots, n_2$ und $j_1 = 1, \dots, m_1$, $j_2 = 1, \dots, m_2$.

Da diese Definition unanschaulich ist und man daher die Struktur der so definierten Matrix nicht sofort sehen kann, schreiben wir diese Matrix in der Blockform

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} \mathbf{A}_{1,1}\mathbf{B} & \dots & \mathbf{A}_{1,n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{n,1}\mathbf{B} & \dots & \mathbf{A}_{n,n}\mathbf{B} \end{pmatrix}.$$

Mithilfe des Kroneckerproduktes können wir das Gleichungssystem (6.8) wie folgt schreiben.

$$(\mathbf{Z}^{\mathbf{t}} \otimes \mathbf{M})U^1 + (\mathbf{Z}^{\mathbf{t},\mathbf{0}} \otimes \mathbf{M})U^0 = (\mathbf{Z} \otimes \mathbf{M})V^1 + (\mathbf{Z}^{\mathbf{0}} \otimes \mathbf{M})V^0, \tag{6.9a}$$

$$(\mathbf{Z}^{\mathbf{t}} \otimes \mathbf{M})V^1 + (\mathbf{Z}^{\mathbf{t},\mathbf{0}} \otimes \mathbf{M})V^0 = (\mathbf{Z} \otimes \mathbf{K})U^1 + (\mathbf{Z}^{\mathbf{0}} \otimes \mathbf{K})U^0. \tag{6.9b}$$

Entkopplung der Gleichungen

Unter der Annahme, dass $\mathbf{Z}^{\mathbf{t}}$ invertierbar ist, lässt sich mit den folgenden Rechenregeln für das Kroneckerprodukt das vorhergehende System entkoppeln. Die Annahme ist gerechtfertigt, siehe Proposition 6.9.

6.3 Proposition. (Rechenregeln für das Kroneckerprodukt)
Es seien $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times n_2}, \mathbf{A}_2 \in \mathbb{R}^{n_2 \times n_3}$ *und* $\mathbf{B}_1 \in \mathbb{R}^{m_1 \times m_2}, \mathbf{B}_2 \in \mathbb{R}^{m_2 \times m_3}$, *dann gilt*

$$(\mathbf{A}_1 \otimes \mathbf{B}_1)(\mathbf{A}_2 \otimes \mathbf{B}_2) = (\mathbf{A}_1\mathbf{A}_2) \otimes (\mathbf{B}_1\mathbf{B}_2). \tag{6.10}$$

Wenn auch $\mathbf{A}_2 \in \mathbb{R}^{n_1 \times n_2}$ *ist, dann gilt*

$$(\mathbf{A}_1 \otimes \mathbf{B}_1) + (\mathbf{A}_2 \otimes \mathbf{B}_1) = (\mathbf{A}_1 + \mathbf{A}_2) \otimes \mathbf{B}_1. \tag{6.11}$$

Beweis. Siehe [Gra81, Chapter 2]. □

Wir können nun V^1 aus (6.9a) eliminieren, indem wir $\widetilde{\mathbf{Z}} := \mathbf{Z}(\mathbf{Z}^{\mathbf{t}})^{-1}$ setzen und (6.9a) von links mit $\widetilde{\mathbf{Z}} \otimes \mathrm{Id}$ multiplizieren, denn dann erhalten wir

$$\begin{aligned}(\mathbf{Z} \otimes \mathbf{M})V^1 \\ = -((\widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{t},\mathbf{0}}) \otimes \mathbf{M})V^0 + ((\widetilde{\mathbf{Z}}\mathbf{Z}) \otimes \mathbf{K})U^1 + ((\widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{0}}) \otimes \mathbf{K})U^0.\end{aligned} \tag{6.12}$$

Wir setzen (6.12) in (6.9a) ein und erhalten für die Bestimmung von U^1 die Gleichung

$$\begin{aligned}(\mathbf{Z}^{\mathbf{t}} \otimes \mathbf{M} - (\widetilde{\mathbf{Z}}\mathbf{Z}) \otimes \mathbf{K})U^1 \\ = ((\widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{0}} - \mathbf{Z}^{\mathbf{t},\mathbf{0}}) \otimes \mathbf{M})U^0 + ((\mathbf{Z}^{\mathbf{0}} - \widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{t},\mathbf{0}}) \otimes \mathbf{M})V^0.\end{aligned} \tag{6.13}$$

Statt mit (6.9) ein gekoppeltes System zu lösen, können wir also auch in zwei Schritten vorgehen und jeweils ein System der halben Dimension lösen. Zuerst bestimmen wir dann U^1 über (6.13) und aktualisieren V^1 danach über (6.12) oder (6.9b).

6.4 Bemerkung.

1. *Es stellt sich sofort die Frage, wie teuer die Operationen in der Berechnung der Produkte von $\widetilde{\mathbf{Z}}$ mit $\mathbf{Z}^{\mathbf{t},\mathbf{0}}$ und $\mathbf{Z}^{\mathbf{0}}$ sind. Problematisch sieht insbesondere die Anwendung der Inversen von $\mathbf{Z}^{\mathbf{t}}$ aus. Wir werden allerdings in Abschnitt 6.3.5 sehen, dass wir durch eine geeignete Basistransformation erreichen können, dass $\mathbf{Z}^{\mathbf{t}}$ eine Diagonalmatrix ist.*

2. *Die Besetzungsstruktur der durch das Kroneckerprodukt enstandenen Matrizen hängt maßgeblich von den Zeitschrittmatrizen ab. Zwar sind die Massematrix $\mathbf{M}$ und die Steifigkeitsmatrix $\mathbf{K}$ bei der Wahl Finiter Elemente schwach besetzt, sollten aber die Zeitschrittmatrizen vollbesetzt sein, dann haben die Matrizen, die aus den Kroneckerprodukten entstehen, 4^l mal soviele Einträge wie die Ortsdiskretisierungsmatrizen. Wir werden in Abschnitt 6.3.5 sehen, dass die Zeitschrittmatrizen auch schwach besetzt sind.*

3. *Das Vorgehen in diesem Abschnitt scheint zu einem unnötigen (Speicher-) Mehraufwand in der Lösung des Problems zu führen. Statt 2^l Zeitschritte mit je zwei Gleichungssystemen der Dimension N durchzuführen, müssen wir nun zwei Gleichungssysteme der Dimension $2^l N$ lösen. Dies ist allerdings nicht das Ziel. Wir benötigen die Erkenntnisse über die Struktur des Problems bei l Bisektionen, um in den nächsten Abschnitten zu erklären, wie man daraus ein Verfahren mit lokalem Zeitschritt konstruieren kann.*

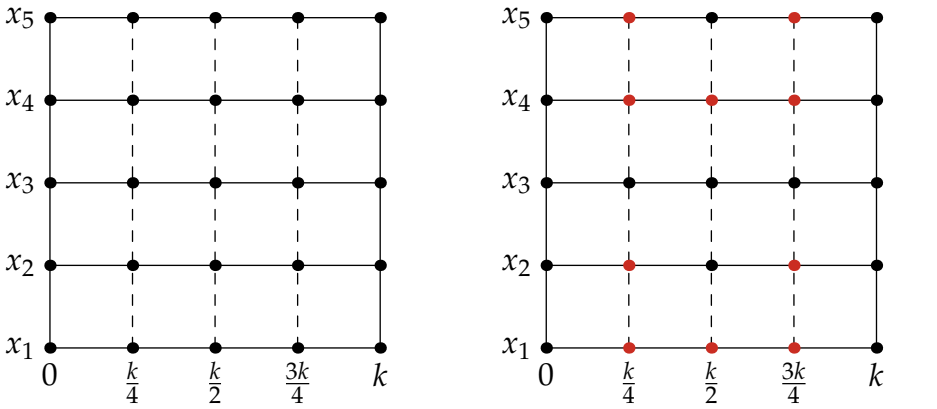
6.1.2 Lokale Verfeinerung

In diesem Abschnitt entfernen wir die Einschränkung, dass das Verfeinerungslevel l für alle $\nu \in \mathcal{N}_h$ gleich gewählt wird. Dies führt dann dazu, dass sich die Zeitschrittweite für verschiedene $\nu \in \mathcal{N}_h$ unterscheiden kann. Wir werden sehen, dass die Wahl der Basis des Ansatzraumes wichtig ist. Die kanonische Wahl der Lagrange-Basis führt nicht dazu, dass wir die Ergebnisse des vorhergenden Abschnittes verwenden können, erlaubt aber eine systematische Berechnung der auftauchenden Matrizen bezüglich der hierarchischen Basis. Wir betrachten daher zuerst die Lagrange-Basis und gehen danach über zu einer hierarchischen Basis, dank der wir lokale Zeitschritte recht einfach mit Hilfe von (6.9) erhalten können.

Wir ordnen also jedem Knoten $\nu \in \mathcal{N}_h$ ein *Verfeinerungslevel* $l_\nu \in \mathbb{N}_0$ zu und definieren den *Verfeinerungsvektor* $\ell := (l_\nu)_{\nu \in \mathcal{N}_h}$. Entsprechend ändert sich der Ansatzraum aus (6.2).

$$\mathcal{V}_1^\ell := \operatorname{span}\left\{\phi \colon \Omega_k \to \mathbb{R} \,|\, \phi(t,x) = \psi(t)\varphi_\nu(x),\ \psi \in \mathcal{B}^{l_\nu}, \nu \in \mathcal{N}_h\right\}.$$

Die Basis für die Zeitdiskretisierung ist im Gegensatz zur globalen Verfeinerung jetzt knotenabhängig. In (6.2) haben wir ein global verfeinertes Gitter



(a) Globale Verfeinerung, $l = 2$. **(b)** Lokale Verfeinerung mit variablem Level l_ν.

Abbildung 6.2: Das Zeit-Ort-Gitter bezüglich des Grundintervalls, einmal mit globaler Verfeinerung des Levels $l = 2$ und einmal reduziert auf eine lokale Verfeinerung. Die schwarzen Punkte geben die Knoten dieses Gitter an, die roten Knoten sind die hängenden Knoten des Zeitgitters.

auf ein lokal verfeinertes Gitter mit $l_1 = l_4 = 0$, $l_2 = l_5 = 1$, $l_3 = 2$ reduziert, wobei die roten Knoten sogenannte *hängende Knoten* sind. Das heißt, dass die Funktionswerte in diesen Knoten durch Interpolation der Werte in den benachbarten (nicht hängenden) Knoten bestimmt werden.

Weiterhin ändert sich der Testraum (6.3),

$$\mathcal{V}_0^\ell := \operatorname{span}\Big\{\phi\colon \Omega_k \to \mathbb{R} \mid \phi(t,x) = \mathbb{1}_{I_{l_\nu,i}}(t)\varphi_\nu(x),\ \nu \in \mathcal{N}_h,\ i = 0,\dots,2^{l_\nu}-1\Big\}.$$

Das Variationsproblem (6.4) passen wir entsprechend an: Wir suchen jetzt u_{kh} und v_{kh} aus $\mathcal{V}_1^\ell$ mit

$$\int_I (\partial_t u_{kh}, \phi_1)\,\mathrm{d}t = \int_I (v_{kh}, \phi_1)\,\mathrm{d}t, \tag{6.14a}$$

$$\int_I (\partial_t v_{kh}, \phi_2)\,\mathrm{d}t = -\int_I a(u_{kh}, \phi_2)\,\mathrm{d}t \tag{6.14b}$$

für alle $\phi_1, \phi_2 \in \mathcal{V}_0^\ell$.

Existenz und Eindeutigkeit

Wir wissen, dass das Problem (6.4) eindeutig lösbar ist. Es ist nicht sofort klar, ob sich diese Eigenschaft auch auf das Problem (6.14) übertragen lässt. Tatsächlich können wir eine positive Antwort geben, es existiert eine eindeutige Lösung für jeden beliebigen Verfeinerungsvektor. Bevor wir das beweisen,

wiederholen wir die Beweisideen zweier Spezialfälle, dann lässt sich die Beweisidee für den allgemeinen Fall einfacher nachvollziehen.

Da das Problem (6.14) linear ist, reicht es aus $u_{kh}(0) = v_{kh}(0) = 0$ zu folgern, dass u_{kh} und v_{kh} auf ganz I verschwinden.

Im Fall $\max_{\nu \in \mathcal{N}_h} l_\nu = 0$, also ohne Verfeinerung, haben wir das Problem (6.4) für $l = 0$. Um dann aus homogenen Anfangswerten herzuleiten, dass $u_{kh}(k) = v_{kh}(k) = 0$ gilt, testet man mit $\phi_1 = \mathcal{A}_h \mathcal{Q}^0 u_{kh}$ und $\phi_2 = \mathcal{Q}^0 v_{kh}$ mit der $L^2(I; L^2(\Omega))$-orthogonalen Projektion $\mathcal{Q}^0$ auf $\mathbb{P}_0(I) \otimes \mathcal{S}_h$ (siehe (5.7)). Dabei ist $\mathcal{A}_h$ wieder der diskrete Laplace-Operator. Damit folgt

$$\begin{aligned}\int_I a(\partial_t u_{kh}, u_{kh}) \,\mathrm{d}t &= \int_I a(v_{kh}, \mathcal{Q}^0 u_{kh}) \,\mathrm{d}t,\\ \int_I (\partial_t v_{kh}, v_{kh}) \,\mathrm{d}t &= -\int_I a(u_{kh}, \mathcal{Q}^0 v_{kh}) \,\mathrm{d}t.\end{aligned}$$

Nach Addition dieser zwei Gleichungen folgt

$$\|u_{kh}(k)\|_{H^1}^2 + \|v_{kh}(k)\|_2^2 = \|u_{kh}(0)\|_{H^1}^2 + \|v_{kh}(0)\|_2^2 = 0.$$

Dass die Funktionswerte in $t = k$ im global verfeinerten Problem (6.4) für $l > 0$ verschwinden, beweist man ähnlich. Zuerst zeigt man $u_{kh}(t_{l,1}) = v_{kh}(t_{l,1}) = 0$, indem man vorheriges Vorgehen auf $I^{l,0}$ anwendet. Dies geht aus dem Grund, dass die Funktionen

$$\begin{aligned}\phi_1(t,x) &= (\mathcal{A}_h(\mathcal{Q}_0^0 u_{kh}))(t,x)\mathbb{1}_{I^{l,0}}(t),\\ \phi_2(t,x) &= (\mathcal{Q}_0^0 v_{kh})(t,x)\mathbb{1}_{I^{l,0}}(t)\end{aligned}$$

für $(t,x) \in \Omega_k$ im Testraum liegen, wobei $\mathcal{Q}_0^0$ die $L^2(I^{l,0}; L^2(\Omega))$-orthogonale Projektion auf $\mathbb{P}_0(I^{l,0}) \times \mathcal{S}_h$ ist.

Im lokal verfeinerten Fall wäre eine erste Idee, genauso vorzugehen wie zuvor diskutiert. Die Idee des nichtverfeinerten Falles kann beispielsweise mit einer geeigneten Projektion auf I dazu verwendet werden zu zeigen, dass die Funktionswerte von u_{kh} und v_{kh} im Endpunkt k verschwinden. Allerdings ist dann zunächst nicht offensichtlich, wie man zeigt, dass die Werte in den anderen Knoten verschwinden. Die Wahl der Testfunktionen analog zum global verfeinerte Problem mit $l > 0$ funktioniert hier nicht, da die in diesem Fall verwendeten Testfunktionen nicht im Testraum $\mathcal{V}_0^\ell$ liegen, wenn die lokalen Verfeinerungslevel sich voneinander unterscheiden. Wir werden eine Mischung beider Ideen verwenden.

Für den Beweis benötigen wir L^2-Projektionen auf die Testräume bezüglich des lokal verfeinerten Gitters. Wir definieren einen Vektorraum, der eine Obermenge unserer Ansatz- und Testräume ist, damit wir die Projektion definieren können und verwenden den Ansatzraum ohne die globale Stetigkeitsbedingung in der Zeit (daher der zusätzliche Index mit der Bedeutung

„**dis**continuous“, unstetig),

$$\mathcal{V}_{1,\mathrm{dc}}^{\ell} := \operatorname{span}\Big\{\phi\colon \Omega_k \to \mathbb{R} \,|\, \phi(t,x) = \psi(t)\varphi_\nu(x), \\ \psi\big\lfloor_{I^{l_\nu,i}} \text{ affin linear}, \nu \in \mathcal{N}_h,\ i = 0,\dots,2^{l_\nu}-1\Big\}.$$

Dann gilt $\mathcal{V}_1^\ell \subset \mathcal{V}_{1,\mathrm{dc}}^\ell$ und auch $\mathcal{V}_0^\ell \subset \mathcal{V}_{1,\mathrm{dc}}^\ell$ und wir können die $L^2(I;L^2(\Omega))$-orthogonale Projektion $\mathcal{L}_0\colon \mathcal{V}_{1,\mathrm{dc}}^\ell \to \mathcal{V}_0^\ell$ definieren durch

$$\int_I (\mathcal{L}_0 w_1, w_0)\,\mathrm{d}t = \int_I (w_1, w_0)\,\mathrm{d}t$$

für alle $w_1 \in \mathcal{V}_{1,\mathrm{dc}}^\ell$, $w_0 \in \mathcal{V}_0^\ell$. Auch $\mathcal{L}_0$ lässt sich im $L^2(I;L^2(\Omega))$-Skalarprodukt wieder wie die Projektion $\mathcal{Q}_n^{q-1}$ aus Abschnitt 5.1.2.1 vertauschen, es gilt also mit $w_1, w_2 \in \mathcal{V}_{1,\mathrm{dc}}^\ell$

$$\int_I (\mathcal{L}_0 w_1, w_2)\,\mathrm{d}t = \int_I (\mathcal{L}_0 w_1, \mathcal{L}_0 w_2)\,\mathrm{d}t = \int_I (w_1, \mathcal{L}_0 w_2)\,\mathrm{d}t. \tag{6.15}$$

Da wir auch eine lokale Version dieser Projektion benötigen, definieren wir für eine Teilmenge $\widetilde{\mathcal{N}}_h$ der Knotenmenge $\mathcal{N}_h$ und ein Intervall $I^{l,j}$ für ein $j = 0,\dots,2^l-1$ den Vektorraum $\mathcal{V}_{1,\mathrm{dc}}^\ell(I^{l,j})$ *bezüglich* $\widetilde{\mathcal{N}}_h$ durch

$$\mathcal{V}_{1,\mathrm{dc}}^{\ell}(I^{l,j}) := \operatorname{span}\{\phi\colon I^{l,j}\times\Omega \to \mathbb{R} \,|\, \phi(t,x) = \psi(t)\varphi_\nu(x), \\ \psi\big\lfloor_{I^{l_\nu,i}} \text{ affin linear}, \nu \in \widetilde{\mathcal{N}}_h,\ i = 0,\dots,2^{l_\nu} \text{ mit } I^{l_\nu,i} \subset I^{l,j}\}.$$

Dies ist einfach nur der Raum $\mathcal{V}_{1,\mathrm{dc}}^\ell$ eingeschränkt auf die Knoten $\widetilde{\mathcal{N}}_h$ und auf das Intervall $I^{l,j}$. Genauso bezeichnen wir mit $\mathcal{V}_0^\ell(I^{l,j})$ den Raum $\mathcal{V}_0^\ell$ eingeschränkt auf die Knoten $\widetilde{\mathcal{N}}_h$ und auf das Intervall $I^{l,j}$ und analog $\mathcal{V}_1^\ell(I^{l,j})$. Dann können wir analog zu $\mathcal{L}_0$ die Projektion $\mathcal{L}_0^{l,j}\colon \mathcal{V}_{1,\mathrm{dc}}^\ell(I^{l,j}) \to \mathcal{V}_0^\ell(I^{l,j})$ definieren durch

$$\int_{I^{l,j}} (\mathcal{L}_0^{l,j} w_1, w_0)\,\mathrm{d}t = \int_{I^{l,j}} (w_1, w_0)\,\mathrm{d}t$$

für alle $w_1 \in \mathcal{V}_{1,\mathrm{dc}}^\ell(I^{l,j})$, $w_0 \in \mathcal{V}_0^\ell(I^{l,j})$.

6.5 Lemma. (Lösbarkeit des lokal verfeinerten Problems)
Das Problem (6.14) *besitzt für jeden beliebigen Verfeinerungsvektor ℓ eine eindeutige Lösung.*

Beweis. Wir setzen voraus, dass $u_{kh}(0) = v_{kh}(0) = 0$ gilt und zeigen, dass $u_{kh}(t) = v_{kh}(t) = 0$ für alle $t \in I$ folgt. Weiterhin benutzen wir in der Zeitdiskretisierung die Lagrange- oder die hierarchische Basis, so dass die Lösungsfunktionen u_{kh} und v_{kh} durch Koeffizienten $U_\nu^{l_\nu,i}$ in den Knoten aus $\mathcal{T}^{l_\nu}$ für alle $\nu \in \mathcal{N}_h$ gegeben sind.

Im ersten Schritt betrachten wir das volle Problem (6.14) und wir möchten $u_{kh}(k)$ und $v_{kh}(k)$ bestimmen. Die Situation ist in Abb. 6.3a veranschaulicht. Die Anfangswerte sind Null, gekennzeichnet durch rote Knoten, gesucht sind die Koeffizienten bezüglich der blauen Knoten. Wie in der dem Lemma vorangestellten Diskussion besprochen, testen wir in (6.14) mit $\phi_1 := \mathcal{A}_h \mathcal{L}_0 u_{kh}$ und $\phi_2 := \mathcal{L}_0 v_{kh}$. Nach Definition der Projektion folgt unter Beachtung von $\partial_t u_{kh}, \partial_t v_{kh} \in \mathcal{V}_0^\ell$

$$\begin{aligned}\int_I a(\partial_t u_{kh}, u_{kh}) \,\mathrm{d}t &= \int_I a(v_{kh}, \mathcal{L}_0 u_{kh}) \,\mathrm{d}t,\\ \int_I (\partial_t v_{kh}, v_{kh}) \,\mathrm{d}t &= -\int_I a(u_{kh}, \mathcal{L}_0 v_{kh}) \,\mathrm{d}t.\end{aligned}$$

Wir addieren die beiden Gleichungen unter Ausnutzung der Vertauschungseigenschaft (6.15) und es folgt

$$\|u_{kh}(k)\|_{H^1}^2 + \|v_{kh}(k)\|_2^2 = \|u_{kh}(0)\|_{H^1}^2 + \|v_{kh}(0)\|_2^2 = 0. \tag{6.16}$$

Bis zu diesem Punkt ist die Argumentation bekannt. Für alle Knoten $\nu \in \mathcal{N}_h$ mit $l_\nu = 0$ haben wir damit alle möglichen Koeffizienten bestimmt, sie verschwinden. Zur Bestimmung der restlichen Koeffizienten können wir also die Level-0-Knoten rausnehmen und den Ansatzraum entsprechend verkleinern. Daher brauchen wir auch nicht mehr den vollen Testraum, auch hier entfernen wir alle Komponenten, die zu Level-0-Knoten gehören. Sei also

$$\mathcal{N}_h^1 := \{\nu \in \mathcal{N}_h \colon l_\nu \geq 1\}$$

und verwende im Folgenden die Vektorräume $\mathcal{V}_1^\ell(I^{1,0})$, $\mathcal{V}_{1,\mathrm{dc}}^\ell(I^{1,0})$, $\mathcal{V}_0^\ell(I^{1,0})$ bezüglich $\mathcal{N}_h^1$. Vom Ausgangsproblem (6.14) bleibt dann die folgende Formulierung übrig. Wir suchen jetzt $u_{kh}, v_{kh} \in \mathcal{V}_1^\ell(I^{1,0})$ mit

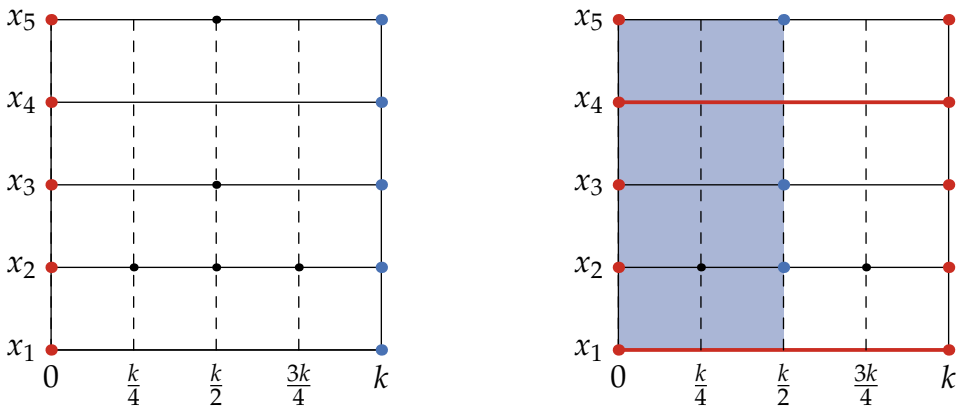
$$\int_{I^{1,0}} (\partial_t u_{kh}, \phi_1) \,\mathrm{d}t = \int_{I^{1,0}} (v_{kh}, \phi_1) \,\mathrm{d}t, \tag{6.17a}$$

$$\int_{I^{1,0}} (\partial_t v_{kh}, \phi_2) \,\mathrm{d}t = -\int_{I^{1,0}} a(u_{kh}, \phi_2) \,\mathrm{d}t \tag{6.17b}$$

für alle $\phi_1, \phi_2 \in \mathcal{V}_0^\ell(I^{1,0})$. Diese Situation ist in Abb. 6.3b dargestellt. Die durch rote Linien gekennzeichneten Knoten wurden aus der Problemformulierung entfernt und wir lösen das Problem auf dem blau eingefärbten Teilgebiet von $I \times \Omega$, um die Koeffizienten bezüglich des Zeitknotens $t_{1,1} = \frac{k}{2}$ zu bestimmen. Wir testen (6.17) mit $\phi_1 := \mathcal{A}_h \mathcal{L}_0^{1,0} u_{kh}$ und $\phi_2 := \mathcal{L}_0^{1,0} v_{kh}$, beide Funktionen liegen in $\mathcal{V}_0^\ell(I^{1,0})$. Wie im ersten Schritt des Beweises folgt also $u_{kh}(t_{1,1}) = v_{kh}(t_{1,1}) = 0$.

Im dritten Schritt setzen wir

$$\mathcal{N}_h^2 := \{\nu \in \mathcal{N}_h \colon l_\nu \geq 2\}.$$



(a) Schritt 1 **(b)** Schritt 2

(c) Schritt 3

Abbildung 6.3: Illustrierung Beweis 6.5. Rote Knoten haben Koeffizienten 0, Koeffizienten in blauen Knoten werden bestimmt, rote Linien bezeichnen x_ν, die aus dem Problem gestrichen wurden, auf dem blau gefärbten Teilgebiet wird die Lösung bestimmt.

Wir möchten die Koeffizienten auf den Knoten des zweiten Levels bestimmen. Dazu müssen wir zwei Probleme lösen, eines auf dem Intervall $I^{2,0}$ mit den Räumen $\mathcal{V}_1^\ell(I^{2,0})$, $\mathcal{V}_{1,\mathrm{dc}}^\ell(I^{2,0})$, $\mathcal{V}_0^\ell(I^{2,0})$ bezüglich $\mathcal{N}_h^2$ und eines auf dem Intervall $I^{2,2}$ mit den Räumen $\mathcal{V}_1^\ell(I^{2,2})$, $\mathcal{V}_{1,\mathrm{dc}}^\ell(I^{2,2})$, $\mathcal{V}_0^\ell(I^{2,2})$ bezüglich $\mathcal{N}_h^2$, siehe auch Abb. 6.3c. Das funktioniert aber genauso wie im zweiten Beweisschritt und es folgt $u_{kh}(t_{2,1}) = v_{kh}(t_{2,1}) = 0$ sowie $u_{kh}(t_{2,3}) = v_{kh}(t_{2,3}) = 0$. Indem man dieses Vorgehen iterativ bis $\max_{\nu\in\mathcal{N}_h} l_\nu$ fortführt, erhält man die Behauptung. □

6.6 Bemerkung. (Energieerhaltung)
Aus dem Beweis des vorhergehenden Lemmas können wir auch direkt die Energieer-

haltung und damit die unbedingte Stabilität des lokalen Zeitschrittverfahrens ablesen. Denn die Gleichung (6.16) *lautet*

$$\|u_{kh}(k)\|_{H^1}^2 + \|v_{kh}(k)\|_2^2 = \|u_{kh}(0)\|_{H^1}^2 + \|v_{kh}(0)\|_2^2,$$

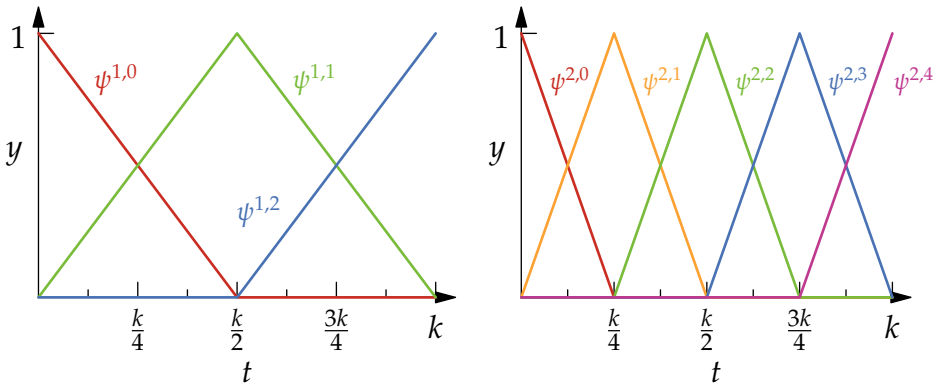
dies liefert die Behauptung.

Als Nächstes stellen wir die Frage, ob die strukturellen Eigenschaften des Variationsproblems (6.14) mit denen von (6.4) übereinstimmen, ob wir also die Zeit- und die Ortsdiskretisierung in der Matrix-Vektor-Formulierung in Form eines Kroneckerproduktes trennen können. Die Antwort auf diese Frage fällt je nach verwendeter Basis unterschiedlich aus. Wir wollen daher im Folgenden untersuchen, welche Basis sich am besten zur Verwendung in diesem Problem eignet.

Lagrange-Basis

Die *Lagrange-Basis* (oder *Knotenbasis*, *nodale Basis*) linearer Elemente bezüglich des Gitters $\mathcal{T}^l$ ist definiert durch

$$\begin{aligned}\mathcal{B}^l := \{\psi^{l,i} \in C(I)\colon\ & \psi^{l,i}\big|_{I^{l,j}} \text{ affin linear } (j = 0,\dots,2^l-1),\\ & \psi^{l,i}(t_{l,j}) = \delta_{ij}\ (j = 0,\dots,2^l),\ \text{für alle } i = 0,\dots,2^l\}.\end{aligned}$$



(a) Die Basis $\mathcal{B}^1$. **(b)** Die Basis $\mathcal{B}^2$.

Abbildung 6.4: Die nodalen Basen für zwei verschiedene Verfeinerungslevel.

Die Lagrange-Basis für $l = 2$ zeigt Abb. 6.4. Das zentrale Problem dieser Basen in Zusammenhang mit dem Zeitschrittverfahren können wir jetzt beschreiben. Gehen wir vor wie im vorherigen Abschnitt, so erhalten wir für

$u_{kh} \in \mathcal{V}_1^\ell$ die Darstellung

$$u_{kh}(t,x) = \sum_{\nu \in \mathcal{N}_h} \left(\sum_{i=0}^{2^{l_\nu}} U_\nu^{l_\nu,i} \psi^{l_\nu,i}(t) \right) \varphi_\nu(x). \tag{6.18}$$

Für v_{kh} müssen wir nur $U_\nu^{l_\nu,i}$ durch $V_\nu^{l_\nu,i}$ ersetzen. Setzen wir diesen Ansatz in (6.14) ein, so erhalten wir

$$\begin{aligned}
&\sum_{\nu \in \mathcal{N}_h} \sum_{i=0}^{2^{l_\nu}} U_\nu^{l_\nu,i} \left(\int_{I^{l_{\nu'},j}} \partial_t \psi^{l_\nu,i}(t)\, \mathrm{d}t \right) (\varphi_\nu, \varphi_{\nu'}) \\
&\qquad = \sum_{\nu \in \mathcal{N}_h} \sum_{i=0}^{2^{l_\nu}} V_\nu^{l_\nu,i} \left(\int_{I^{l_{\nu'},j}} \psi^{l_\nu,i}(t)\, \mathrm{d}t \right) (\varphi_\nu, \varphi_{\nu'}), \\
&\sum_{\nu \in \mathcal{N}_h} \sum_{i=0}^{2^{l_\nu}} V_\nu^{l_\nu,i} \left(\int_{I^{l_{\nu'},j}} \partial_t \psi^{l_\nu,i}(t)\, \mathrm{d}t \right) (\varphi_\nu, \varphi_{\nu'}) \\
&\qquad = - \sum_{\nu \in \mathcal{N}_h} \sum_{i=0}^{2^{l_\nu}} U_\nu^{l_\nu,i} \left(\int_{I^{l_{\nu'},j}} \psi^{l_\nu,i}(t)\, \mathrm{d}t \right) a(\varphi_\nu, \varphi_{\nu'}),
\end{aligned}$$

für alle $\nu' \in \mathcal{N}_h$ und $j = 0, \ldots, 2^{l_{\nu'}} - 1$. Der Unterschied zu (6.8) ist die Kopplung aller Faktoren über die Abhängigkeit von ν und dass in den Zeitintegralen die Kopplung zwischen allen auftauchenden Verfeinerungsleveln berechnet wird. Die Struktur ähnelt zwar der des global verfeinerten Falles, man kann aber die Orts- und Zeitdiskretisierungsmatrizen nicht mehr entkoppeln. Stattdessen kann man beispielsweise

$$\sum_{\nu \in \mathcal{N}_h} \sum_{i=1}^{2^{l_\nu}} U_\nu^{l_\nu,i} \left(\int_{I^{l_{\nu'},j}} \partial_t \psi^{l_\nu,i}(t)\, \mathrm{d}t \right) (\varphi_\nu, \varphi_{\nu'})$$

für alle $\nu' \in \mathcal{N}_h$ und $j = 0, \ldots, 2^{l_{\nu'}} - 1$ in die Form

$$\begin{pmatrix} \mathbf{M}_{1,1}\mathbf{Z^t}_{1,1} & \ldots & \mathbf{M}_{1,N}\mathbf{Z^t}_{1,N} \\ \vdots & \ddots & \vdots \\ \mathbf{M}_{N,1}\mathbf{Z^t}_{N,1} & \ldots & \mathbf{M}_{N,N}\mathbf{Z^t}_{N,N} \end{pmatrix} \begin{pmatrix} U_1^{l_1} \\ \vdots \\ U_N^{l_N} \end{pmatrix}$$

bringen. Dabei sind die $\mathbf{M}_{\nu,\nu'}$ die Einträge der Massematrix $\mathbf{M}$ nach Nummerierung der Freiheitsgrade von 1 bis $N := |\mathcal{N}_h|$. Die Zeitschrittmatrizen $\mathbf{Z^t}_{\nu,\nu'}$ sind aber knotenabhängig definiert durch

$$(\mathbf{Z^t}_{\nu,\nu'})_{i,j} := \int_{I^{l_{\nu'},j}} \partial_t \psi^{l_\nu,i}(t)\, \mathrm{d}t,$$

für $i = 1, \ldots, 2^{l_\nu}$ und $j = 0, \ldots, 2^{l_{\nu'}} - 1$. Ferner ist $U_\nu^{l_\nu} := (U_\nu^{l_\nu,i})_{i=1,\ldots,2^{l_\nu}}$. Um diese Matrix aufstellen zu können, müsste man alle $\mathbf{Z^t}_{\nu,\nu'}$ berechnen oder

zumindest eine Matrix für jedes auftretende Paar von Verfeinerungsleveln (l, l'). Das ist zwar möglich, wir wünschen uns aber mehr Flexibilität. Beispielsweise möchten wir ohne viel Aufwand den Vektor ℓ in jedem Zeitschritt ändern können. Für die Lagrange-Elemente würde das bedeuten, dass wir die Matrizen komplett neu aufstellen müssten. Abhilfe bietet eine andere Basis des Ansatzraumes.

Hierarchische Elemente

Wir definieren jetzt eine hierarchische Basis, das heisst bei jedem Übergang von einem Verfeinerungslevel auf das nächsthöhere Level ergänzen wir die vorherige Basis nur, bestimmen sie aber nicht wie bei der Knotenbasis ganz neu. In diesem Fall setzt man a priori ein maximales Verfeinerungslevel fest, bestimmt dann bezüglich dieses Levels das Gleichungssystem gemäß Abschnitt 6.1.1 und fügt eine Zwangsbedingung hinzu: Alle Koeffizienten, deren zugehöriges Level höher ist als das gewünschte, werden Null gesetzt. Ein analoges Vorgehen scheitert für Lagrange-Elemente, ein Übergang auf ein niedrigeres Verfeinerungslevel erfordert eine ganz andere Basis.

Wir definieren die *hierarchischen Basen* induktiv. Die Zeitgitter $\mathcal{T}^l$ sind nach Definition schon hierarchisch aufgebaut, es gilt also

$$\mathcal{T}^0 \subset \mathcal{T}^1 \subset \mathcal{T}^2 \subset \cdots$$

und wir können die Menge der Knoten, die bei jeder Erhöhung des Verfeinerungslevels hinzukommt, bestimmen. Diese *Menge der Knoten des l-ten Levels* bezeichnen wir mit

$$\mathcal{R}^l := \mathcal{T}^l \setminus \mathcal{T}^{l-1} \qquad (l \in \mathbb{N})$$

und wir setzen $\mathcal{R}^0 := \mathcal{T}^0$. Dann gilt die Beziehung

$$\mathcal{T}^l = \mathcal{T}^{l-1} \cup \mathcal{R}^l \qquad (l \in \mathbb{N}).$$

Die Menge der *Indizes der Knoten des l-ten Levels* bezeichnen wir mit

$$R^l := \{i \in \mathbb{N}_0 \colon t_{l,i} \in \mathcal{R}^l\} \qquad (l \in \mathbb{N}_0).$$

Weiterhin benötigen wir diese Indizes auch bezüglich eines höheren Levels, wir definieren daher die Menge der *Indizes der Knoten des l-ten Levels bezüglich des $\hat{l}$-ten Levels* (mit $\hat{l} \geq l$) durch

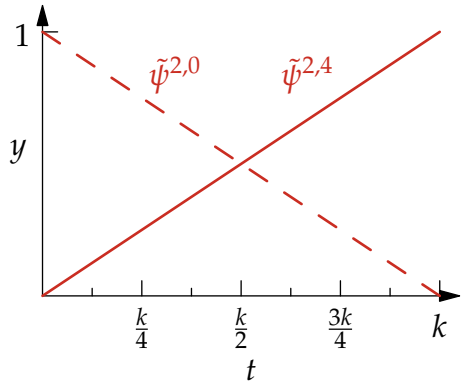
$$R^{\hat{l},l} := \{i \in \mathbb{N}_0 \colon t_{\hat{l},i} \in \mathcal{R}^l\} \qquad (\hat{l} \in \mathbb{N}_0,\ l \leq \hat{l}).$$

In Abb. 6.5 sind diese Mengen beispielhaft für das Referenzlevel 2 angegeben.

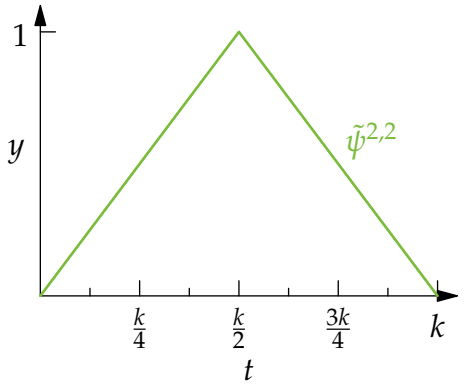
Eine hierarchische Basis bezüglich des Levels l kann dann ausgehend von den Knotenbasen der Level $j \leq l$ aufgebaut werden durch

$$\mathcal{B}^l_{\text{hier}} = \left\{ \tilde{\psi}^{l,i} : i = 0, \ldots, 2^l \right\} := \bigcup_{j=0}^{l} \left\{ \psi^{j,i} : i \in R^{j,j} \right\},$$

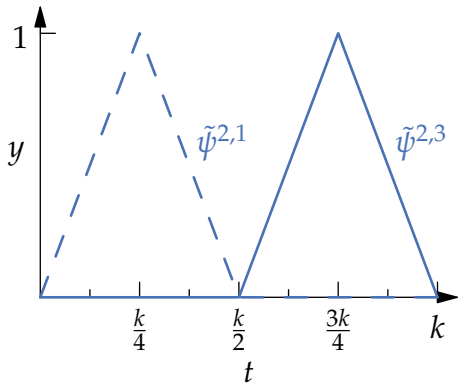
wobei $\tilde{\psi}^{l,i}$ dasjenige $\psi \in \mathcal{B}^l_{\text{hier}}$ sei, für das $\psi(t_{l,i}) = 1$ gilt. Das bedeutet, dass wir ausgehend von einer Basis $\mathcal{B}^{l-1}_{\text{hier}}$ zur Basis $\mathcal{B}^l_{\text{hier}}$ gelangen, indem wir die Lagrange-Basisfunktionen aus $\mathcal{B}^l$ hinzufügen, die den Knoten des l-ten Levels $\mathcal{R}^l$ zugeordnet sind. Diese Konstruktion hat die Eigenschaft, die wir weiter oben erwähnt hatten: Setzen wir die Freiheitsgrade in den Knoten $\mathcal{R}^l$ gleich Null, so bleibt eine Darstellung der Funktion bezüglich der Basis $\mathcal{B}^{l-1}_{\text{hier}}$ übrig.



(a) Level 0 mit $\mathcal{R}^0 = \{0, k\}, R^{2,0} = \{0, 4\}$



(b) Level 1 mit $\mathcal{R}^1 = \{k/2\}, R^{2,1} = \{2\}$



(c) Level 2 mit $\mathcal{R}^2 = \{k/4, 3k/4\}, R^{2,2} = \{1, 3\}$

Abbildung 6.5: Basiselemente der hierarchischen Basis zu Level 2 aufgeteilt nach den zugehörigen Leveln

Die Darstellung von u_{kh} bezüglich der hierarchischen Basis auf dem Grund-

intervall können wir sortiert nach dem jeweilig aktiven Level angeben.

$$u_{kh}(t,x) = \sum_{\nu\in\mathcal{N}_h}\left(\sum_{i=0}^{2^{l_\nu}} \widetilde{U}_\nu^{l_\nu,i}\widetilde{\psi}^{l_\nu,i}(t)\right)\varphi_\nu(x)$$
$$= \sum_{\nu\in\mathcal{N}_h}\left(\sum_{l=0}^{l_\nu}\sum_{i\in R^{l_\nu,l}} \widetilde{U}_\nu^{l_\nu,i}\widetilde{\psi}^{l_\nu,i}(t)\right)\varphi_\nu(x). \tag{6.19}$$

Wir beginnen zuerst mit der globalen Verfeinerung aus Abschnitt 6.1.1, also $l_\nu = l$ für alle $\nu \in \mathcal{N}_h$. Verwenden wir den Ansatz (6.19) in Gl. (6.4a) und testen mit den aufspannenden Funktionen von $\mathcal{V}_0^\ell$, $\phi_1(t,x) = \mathbb{1}_{I^{l',j}}(t)\varphi_{\nu'}(x)$ für $\nu' \in \mathcal{N}_h$, $l' = 0,\dots,\hat{l}$, $j \in R^{\hat{l},l'}$, so erhalten wir

$$\sum_{\nu\in\mathcal{N}_h}\left(\sum_{l=0}^{\hat{l}}\sum_{i\in R^{\hat{l},l}} \widetilde{U}_\nu^{\hat{l},i}\left(\int_{I^{l',j}}\partial_t\widetilde{\psi}^{\hat{l},i}(t)\,\mathrm{d}t\right)\right)(\varphi_\nu,\varphi_{\nu'})$$
$$= \sum_{\nu\in\mathcal{N}_h}\left(\sum_{l=0}^{\hat{l}}\sum_{i\in R^{\hat{l},l}} \widetilde{V}_\nu^{\hat{l},i}\left(\int_{I^{l',j}}\widetilde{\psi}^{\hat{l},i}(t)\,\mathrm{d}t\right)\right)(\varphi_\nu,\varphi_{\nu'}). \tag{6.20a}$$

Analog verfahren wir mit Gl. (6.4b) und erhalten

$$\sum_{\nu\in\mathcal{N}_h}\left(\sum_{l=0}^{\hat{l}}\sum_{i\in R^{\hat{l},l}} \widetilde{V}_\nu^{\hat{l},i}\left(\int_{I^{l',j}}\partial_t\widetilde{\psi}^{\hat{l},i}(t)\,\mathrm{d}t\right)\right)(\varphi_\nu,\varphi_{\nu'})$$
$$= \sum_{\nu\in\mathcal{N}_h}\left(\sum_{l=0}^{\hat{l}}\sum_{i\in R^{\hat{l},l}} \widetilde{U}_\nu^{\hat{l},i}\left(\int_{I^{l',j}}\widetilde{\psi}^{\hat{l},i}(t)\,\mathrm{d}t\right)\right)a(\varphi_\nu,\varphi_{\nu'}). \tag{6.20b}$$

Die Dimension des Ansatzraums unterscheidet sich von der Dimension des Testraums um 1. Dieser Dimensionsunterschied ist auch hier wieder notwendig, da die Werte $\widetilde{U}_\nu^{\hat{l},0}, \widetilde{V}_\nu^{\hat{l},0}$ bekannt sind, also nicht der volle Ansatzraum verwendet werden kann. Wir müssen daher nur die Werte $\widetilde{U}_\nu^{\hat{l},i}, \widetilde{V}_\nu^{\hat{l},i}$ für $i > 0$ bestimmen.

Reduktion der Freiheitsgrade bei variabler Verfeinerung Wir betrachten nun die Reduktion für eine nicht gleichmäßige Verfeinerung, falls also $l_\nu < \hat{l}$ für mindestens ein $\nu \in \mathcal{N}_h$ gilt. Im Falle von $l_\nu < \hat{l}$ ergeben sich zwei Änderungen in (6.20).

1. Die Levelsumme für $\nu \in \mathcal{N}_h$ läuft nur noch bis l_ν, dies entspricht der Reduktion des Ansatzraumes.
2. Aufgrund der Reduktion des Ansatzraumes kann auch der Testraum reduziert werden. Wir testen nur noch mit den Funktionen $\phi_1(t,x) = \mathbb{1}_{I^{l',j}}(t)\varphi_{\nu'}(x)$ für $\nu' \in \mathcal{N}_h$, $l' = 0,\dots,l_\nu$, $j \in R^{l_{\nu'},l'}$. Diese Funktionen bilden eine Basis von $\mathcal{V}_0^\ell$.

Der Unterschied zur Verwendung der Lagrange-Basis ist, dass die Struktur erhalten bleibt, es werden nur gewisse Teile weggelassen. Mathematisch entspricht das der Anwendung einer geeigneten Projektion.

Wir möchten also das lineare Gleichungssystem Gl. (6.9) hernehmen und so reduzieren, dass wir eine Version mit lokal verfeinertem Zeitschritt erhalten. Wir beachten dazu die zwei oben beschriebenen Änderungen. Zunächst entfernen wir aus den Koeffizientenvektoren U^1, V^1 alle Komponenten, die durch die Levelbeschränkung Null sein sollen. Das gewährleisten wir durch Multiplikation mit einer geeigneten Projektion, die wir im Folgenden definieren. Es bezeichne wieder $\ell = (l_\nu)_{\nu \in \mathcal{N}_h}$ den Vektor, der die lokalen Verfeinerungslevel beinhaltet, und weiterhin sei

$$\mathcal{N}_h^l := \{\nu \in \mathcal{N}_h \colon l \leq l_\nu\}$$

die Menge der Knoten, deren Zeitverfeinerungslevel (mindestens) l ist. Mithilfe dieser Menge können wir den Koeffizientenvektor reduzieren. Wir ordnen dem Zeitindex $i \in R^{\hat{l},l}$ für $N_l := |\mathcal{N}_h^l|$ den Vektor $U^{\ell,i} \in \mathbb{R}^{N_l}$ definiert durch

$$U^{\ell,i} := (\widetilde{U}_\nu^{l,i})_{\nu \in \mathcal{N}_h^l}$$

zu. Aus diesen Teilvektoren setzen wir dann den vollständigen Koeffizientenvektor $U^{1,\ell} \in \mathbb{R}^{N_\ell}$ zusammen, wobei $N_\ell := \sum_{l=1}^{\hat{l}} |R^l| N_l$. Es sei also

$$U^{1,\ell} := (U^{\ell,i})_{i=1,\dots,2^{\hat{l}}}.$$

Der Vektor U^0, der die Anfangswerte enthält, ändert sich nicht. Es sei $\mathbf{P}_\ell \in \{0,1\}^{N_\ell, N 2^{\hat{l}}}$ die Projektionsmatrix, die durch

$$\mathbf{P}_\ell U^1 = U^{1,\ell}$$

definiert ist. Dieser Vorgang entspricht der Reduktion des Ansatzraumes. Wir erhalten U^1 wieder, wenn wir $U^1 = \mathbf{P}_\ell^T U^{1,\ell}$ berechnen. In Gl. (6.9) setzen wir also U^1, V^1 ein und reduzieren den Testraum durch Anwendung der Projektionsmatrix $\mathbf{P}_\ell$ von links.

$$\begin{aligned}
&\mathbf{P}_\ell(\mathbf{Z^t} \otimes \mathbf{M})\mathbf{P}_\ell^T U^{1,\ell} + \mathbf{P}_\ell(\mathbf{Z^{t,0}} \otimes \mathbf{M})U^0 \\
&\qquad = \mathbf{P}_\ell(\mathbf{Z} \otimes \mathbf{M})\mathbf{P}_\ell^T V^{1,\ell} + \mathbf{P}_\ell(\mathbf{Z^0} \otimes \mathbf{M})V^0, \quad (6.21a) \\
&\mathbf{P}_\ell(\mathbf{Z^t} \otimes \mathbf{M})\mathbf{P}_\ell^T V^{1,\ell} + \mathbf{P}_\ell(\mathbf{Z^{t,0}} \otimes \mathbf{M})V^0 \\
&\qquad = \mathbf{P}_\ell(\mathbf{Z} \otimes \mathbf{K})\mathbf{P}_\ell^T U^{1,\ell} + \mathbf{P}_\ell(\mathbf{Z^0} \otimes \mathbf{K})U^0. \quad (6.21b)
\end{aligned}$$

Die Zeitschrittmatrizen wurden hier gemäß der hierarchischen Basis aufgestellt. Verwenden wir also eine hierarchische Basis, so können wir die Struktur von Gl. (6.9) im Wesentlichen bewahren. Das bietet den Vorteil der einfacheren Implementierung und lässt es zu, dass wir in jedem Zeitschritt die Knoten, in denen lokal verfeinert werden soll, bequem durch die Änderung der Projektionsmatrix austauschen können.

6.7 Bemerkung.
Wir können die Projektion auch auf das entkoppelte System (6.13), (6.9b) *anwenden.*

6.2 Lokaler Zeitschritt für ein nichtlineares Problem

In diesem Schritt wenden wir uns dem nichtlinearen Problem in semilinearer Formulierung (3.25) zu. Zwar konnten wir die Konvergenz des zugehörigen Galerkinverfahrens in Abschnitt 5.3 nicht beweisen, die numerischen Experimente zeigten aber zumindest im Eindimensionalen das gewünschte Konvergenzverhalten (siehe Abschnitt 5.4). Wir verwenden die semilineare Formulierung, weil die Anwendung der Idee des lokalen Zeitschrittverfahrens mit nur wenigen Anpassungen möglich ist. Prinzipiell könnte man natürlich auch die quasilineare Formulierung mit dem qlw-cG(1) cG(p)-Verfahren (5.27) nehmen, strukturelle Vorteile, wie wir sie in den vorherigen Abschnitten erarbeitet haben, finden sich dann aber nicht mehr. Wir betrachten also wieder

$$v = u + f(u)\,, \tag{6.22a}$$
$$\partial_t v = w\,, \tag{6.22b}$$
$$\partial_t w = \Delta u\,. \tag{6.22c}$$

Wir verwenden zur Zeitdiskretisierung die hierarchische Basis, dann lassen sich die letzteren beiden Gleichungen mit Hilfe der vorhergehenden Abschnitte sofort diskretisieren. Wir erhalten also vor Anwendung der Projektionen

$$\begin{aligned}(\mathbf{Z}\otimes\mathbf{M})V^1 + (\mathbf{Z^0}\otimes\mathbf{M})V^0 &= (\mathbf{Z}\otimes\mathbf{M})U^1 + (\mathbf{Z^0}\otimes\mathbf{M})U^0 \\ &\quad + F(u_{kh}),\end{aligned} \tag{6.23a}$$
$$(\mathbf{Z^t}\otimes\mathbf{M})V^1 + (\mathbf{Z^{t,0}}\otimes\mathbf{M})V^0 = (\mathbf{Z}\otimes\mathbf{M})W^1 + (\mathbf{Z^0}\otimes\mathbf{M})W^0, \tag{6.23b}$$
$$(\mathbf{Z^t}\otimes\mathbf{M})W^1 + (\mathbf{Z^{t,0}}\otimes\mathbf{M})W^0 = (\mathbf{Z}\otimes\mathbf{K})U^1 + (\mathbf{Z^0}\otimes\mathbf{K})U^0. \tag{6.23c}$$

Dabei ist für alle $v' \in \mathcal{N}_h$

$$F_{v'}(u_{kh}) = \left(\int_{\hat{I},j} \int_\Omega f(u_{kh}(t,x))\varphi_{v'}(x)\,\mathrm{d}x\,\mathrm{d}t \right)_{j=0,\dots,2^{\hat{l}}-1}. \tag{6.24}$$

Es bleibt die Frage, wie man (6.24) im Newtonverfahren behandelt. Es wäre zu aufwändig, diesen Vektor in jedem Newtonschritt neu zu assemblieren. Würden wir die Produktapproximation (siehe Abschnitt 4.3.1) zur Hilfe nehmen, dann könnten wir wieder mit den Projektionsmatrizen aus vorherigem Abschnitt das Gleichungssystem reduzieren und die Idee des lokalen Zeitschrittverfahrens verwenden. Wie gut die Produktapproximation funktioniert hängt aber von der Basis der Zeitdiskretisierung ab, wie wir jetzt sehen werden.

6.2.1 Produktapproximation und hierarchische Basis

Die Anwendung der Produktapproximation in Zeit und Ort unter Verwendung der hierarchischen Basisdarstellung in der Zeit ist gegeben durch

$$(f(u_{kh}))(t,x) \approx \sum_{v \in \mathcal{N}_h} \left(\sum_{l=0}^{l_v} \sum_{i \in R^{l_v,l}} f\left(\tilde{U}_v^{l_v,i}\right) \tilde{\psi}^{l_v,i}(t) \right) \varphi_v(x).$$

Eine Implementierung mit Lagrange-Elementen $\mathcal{S}_h$ und Verwendung der Produktapproximation liefert folgendes Ergebnis.

1. Für feste Zeitdiskretisierung und $h \to 0$ konvergiert das Verfahren wie erwartet, bis der Fehler durch die Zeitdiskretisierung erreicht wird.
2. Für feste Ortsdiskretisierung und $l_v \to \infty$ konvergiert das Verfahren nicht, der Fehler ist sogar im Wesentlichen für alle $l_v \in \mathbb{N}_0$ gleich.

Dieses Verhalten legt nahe, dass die Produktapproximation bezüglich der Ortsdiskretisierung mindestens die gleiche Fehlerordnung besitzt wie das Verfahren selber, der Fehler der Zeitdiskretisierung aber durch Erhöhung des Verfeinerungslevels nicht gegen 0 konvergiert.

Wir möchten daher kurz auf die Ursache dieses Phänomens eingehen. Dazu betrachten wir ein einfaches Beispiel, welches die zugrundeliegende Problematik aufzeigt. Wir definieren das Grundintervall $I := [0,1]$ und darauf eine affin lineare Funktion $u\colon I \to \mathbb{R}$, $u(t) := a + (b-a)t$ für $a, b \in \mathbb{R}$. Also ist $u(0) = a, u(1) = b$. Diese Funktion können wir exakt durch lineare Elemente auf jedem beliebigen Gitter über I darstellen. Für das Verfeinerungslevel l mit entsprechendem Gitter $\mathcal{T}^l$ lässt sich u in Lagrange-Darstellung schreiben als

$$u(t) = \sum_{i=0}^{2^l} u(t_{l,i}) \psi^{l,i}(t).$$

Insbesondere enthält diese Darstellung Informationen über u an allen Stützstellen $t_{l,i}$. Dies ist bei der hierarchischen Darstellung nicht so, denn es ist

$$u(t) = a\tilde{\psi}^{0,0}(t) + b\tilde{\psi}^{0,1}(t),$$

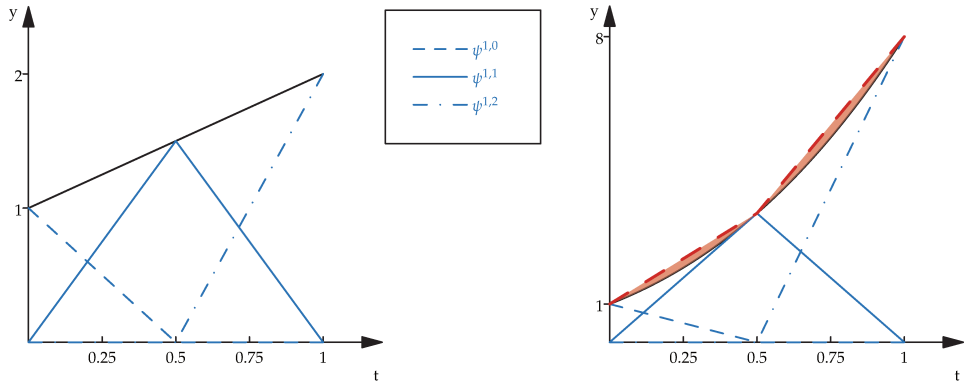
unabhängig vom gewählten Verfeinerungslevel. Zunächst sieht die hierarchische Darstellung geschickter aus, da sie mit der vorliegenden Information optimal umgeht, man benötigt nur zwei Parameter zur exakten Beschreibung der Funktion. Diese Eigenschaft stellt sich bei Anwendung der Produktapproximation als problematisch heraus. Denn wir approximieren in der Lagrange-Darstellung $f(u(t))$ durch

$$f(u(t)) \approx \sum_{i=0}^{2^l} f(u(t_{l,i})) \psi^{l,i}(t),$$

also durch die lineare Interpolation in den Gitterpunkten des (feinen) Gitters $\mathcal{T}^l$. In der hierarchischen Darstellung approximieren wir $f(u(t))$ dagegen durch

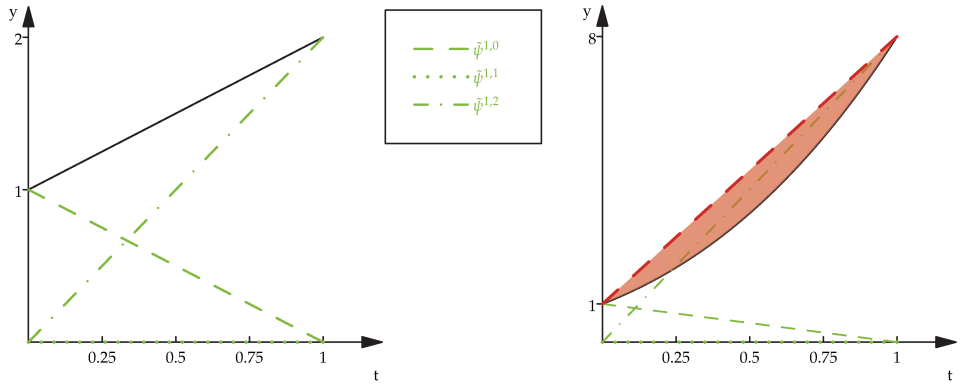
$$f(u(t)) \approx f(a)\tilde{\psi}^{0,0}(t) + f(b)\tilde{\psi}^{0,1}(t),$$

also die lineare Interpolation im gröbsten Gitter $\mathcal{T}^0$.



(a) u in Lagrange-Darstellung.

(b) $f(u)$ und Produktapproximation in Lagrange-Darstellung.



(c) u in hierarchischer Darstellung.

(d) $f(u)$ und Produktapproximation in hierarchischer Darstellung

Abbildung 6.6: Darstellung einer affin linearen Funktion u auf I durch Lagrange- und hierarchische Basis. Auf der rechten Seite sind für $f(x) = x^3$ die Funktion $f(u)$ (schwarz) sowie die Produktapproximationen (rot gestrichelt) dargestellt. Die Fläche zwischen $f(u)$ und den Approximationen ist rot eingefärbt.

In Abb. 6.6 ist diese Situation für das Verfeinerungslevel $l = 1$ mit $f(x) = x^3$ dargestellt. Auf der rechten Seite sehen ist der Unterschied zwischen $f(u)$ und den jeweiligen Approximationen rot eingefärbt. Diese Fläche wird für Lagrange-Elemente mit jeder Erhöhung des Verfeinerungslevels kleiner, für hierarchische Elemente bleibt sie immer gleich groß.

Dieses Beispiel ist ein Extremfall, in dem überhaupt keine Verbesserung durch die Wahl eines höheren Levels erreicht werden kann. Doch auch für Funktionen u, die nicht affin linear sind, treten vergleichbare Probleme auf. In Abschnitt 6.3.3 werden wir sehen, dass die Koeffizienten der hierarchischen Basisdarstellung im Wesentlichen um einen Faktor $1/2^l$ kleiner sind als die nodalen Werte. Dies kann ein wünschenswerter Nebeneffekt sein, wie man in threshold-Verfahren für Wavelets sehen kann (siehe [Li11]), führt hier aber dazu, dass durch höhere Verfeinerungslevel nicht die erwünschte Verbesserung der Konvergenzordnung erreicht wird.

Eine Möglichkeit, das vorgestellte Verhalten zu umgehen, stellen wir jetzt vor. Wir definieren durch $\Pi_{\text{knot}}^{\text{hier}}$ die Transformation eines Vektors in Knotenbasisdarstellung in den zugehörigen Vektor in hierarchischer Basisdarstellung, umgekehrt sei $\Pi_{\text{hier}}^{\text{knot}}$ die Transformation eines Vektors in hierarchischer Basisdarstellung in den Koeffizientenvektor in Knotenbasisdarstellung. Wie man diese Transformationen berechnet, beschreiben wir in Abschnitt 6.3.3.

1. Bestimme den Koeffizientenvektor U zu u_{kh} in hierarchischer Darstellung.
2. Transformiere U durch $\Pi_{\text{hier}}^{\text{knot}} U$ auf Knotenbasisdarstellung.
3. Jetzt können wir die nichtlineare Funktion auf diesen Vektor anwenden, erhalten also $f(\Pi_{\text{hier}}^{\text{knot}} U)$.
4. Diesen Vektor können wir dann wieder ohne Verlust von Genauigkeit in die hierarchische Darstellung transformieren und erhalten den Vektor $\Pi_{\text{knot}}^{\text{hier}} f(\Pi_{\text{hier}}^{\text{knot}} U)$.

Durch dieses Vorgehen bleibt die Kroneckerprodukt-Struktur der Galerkin-Formulierung erhalten. Ohne die Reduktion aus dem vorherigen Abschnitt führt das zu dem Gleichungssystem

$$\begin{aligned}(\mathbf{Z}\otimes\mathbf{M})V^1 + (\mathbf{Z^0}\otimes\mathbf{M})V^0 &= (\mathbf{Z}\otimes\mathbf{M})(U^1 + \Pi_{\text{knot}}^{\text{hier}} f(\Pi_{\text{hier}}^{\text{knot}} U^1)) \\ &\quad + (\mathbf{Z^0}\otimes\mathbf{M})(U^0 + f(U^0)), \qquad (6.25a)\\ (\mathbf{Z^t}\otimes\mathbf{M})V^1 + (\mathbf{Z^{t,0}}\otimes\mathbf{M})V^0 &= (\mathbf{Z}\otimes\mathbf{M})W^1 + (\mathbf{Z^0}\otimes\mathbf{M})W^0, \qquad (6.25b)\\ (\mathbf{Z^t}\otimes\mathbf{M})W^1 + (\mathbf{Z^{t,0}}\otimes\mathbf{M})W^0 &= (\mathbf{Z}\otimes\mathbf{K})U^1 + (\mathbf{Z^0}\otimes\mathbf{K})U^0. \qquad (6.25c)\end{aligned}$$

Die Reduktion auf lokale Verfeinerungslevel funktioniert wie in (6.38) durch Anwendung der Projektion $\mathbf{P}_\ell$.

Entkopplung der Gleichungen

Auch das System (6.25) lässt sich unter der Annahme, dass $\mathbf{Z^t}$ invertierbar ist, entkoppeln. Die Invertierbarkeit zeigen wir in Proposition 6.9.

Wir gehen nun wie folgt vor. Wir lösen (6.25c) nach W^1 auf, setzen dies in (6.25b) ein, lösen die resultierende Gleichung nach V^1 auf, und setzen dies

dann in (6.25a) ein. Dies ergibt dann ein Gleichungssystem für U^1. Haben wir U^1 bestimmt, können wir V^1 berechnen und schließlich W^1. Multiplizieren wir (6.25c) von links mit $(\mathbf{Z}(\mathbf{Z}^t)^{-1}) \otimes \mathrm{Id}$, so erhalten wir dank (6.10)

$$(\mathbf{Z} \otimes \mathbf{M})W^1 + ((\widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{0}}) \otimes \mathbf{M})W^0 = ((\widetilde{\mathbf{Z}}\mathbf{Z}) \otimes \mathbf{K})U^1 + ((\widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{0}}) \otimes \mathbf{K})U^0, \quad (6.26)$$

wobei wir $\widetilde{\mathbf{Z}} := \mathbf{Z}(\mathbf{Z}^{\mathbf{t}})^{-1}$ gesetzt haben. Damit ersetzen wir in (6.25b) den Term $(\mathbf{Z} \otimes \mathbf{M})W^1$ und erhalten mit (6.11)

$$\begin{aligned}(\mathbf{Z}^{\mathbf{t}} \otimes \mathbf{M})V^1 + (\mathbf{Z}^{\mathbf{t,0}} \otimes \mathbf{M})V^0 \\ = ((\mathbf{Z}^{\mathbf{0}} - \widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{t,0}}) \otimes \mathbf{M})W^0 + ((\widetilde{\mathbf{Z}}\mathbf{Z}) \otimes \mathbf{K})U^1 + ((\widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{0}}) \otimes \mathbf{K})U^0.\end{aligned} \quad (6.27)$$

Wie zuvor multiplizieren wir (6.27) von links mit $(\widetilde{\mathbf{Z}} \otimes \mathrm{Id})$, so folgt

$$\begin{aligned}(\mathbf{Z} \otimes \mathbf{M})V^1 = -((\widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{t,0}}) \otimes \mathbf{M})V^0 + ((\widetilde{\mathbf{Z}}(\mathbf{Z}^{\mathbf{0}} - \widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{t,0}})) \otimes \mathbf{M})W^0 \\ + ((\widetilde{\mathbf{Z}}^2\mathbf{Z}) \otimes \mathbf{K})U^1 + ((\widetilde{\mathbf{Z}}^2\mathbf{Z}^{\mathbf{0}}) \otimes \mathbf{K})U^0.\end{aligned} \quad (6.28)$$

Dies setzen wir in (6.25a) ein und erhalten die Bestimmungsgleichung für U^1.

$$\begin{aligned}(\mathbf{Z} \otimes \mathbf{M})(U^1 + f(U^1)) = -(\mathbf{Z}^{\mathbf{0}} \otimes \mathbf{M})(U^0 + f(U^0)) \\ + ((\mathbf{Z}^{\mathbf{0}} - \widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{t,0}}) \otimes \mathbf{M})V^0 \\ + ((\widetilde{\mathbf{Z}}(\mathbf{Z}^{\mathbf{0}} - \widetilde{\mathbf{Z}}\mathbf{Z}^{\mathbf{t,0}})) \otimes \mathbf{M})W^0 \\ + ((\widetilde{\mathbf{Z}}^2\mathbf{Z}^{\mathbf{0}}) \otimes \mathbf{K})U^0 + ((\widetilde{\mathbf{Z}}^2\mathbf{Z}) \otimes \mathbf{K})U^1.\end{aligned} \quad (6.29)$$

Um einen Zeitschritt zu berechnen, gehen wir also wie folgt vor.

1. Berechne U^1 aus (6.29).
2. Setze U^1 in (6.27) ein und berechne V^1.
3. Setze U^1 in Gl. (6.25c) ein und berechne W^1.

6.3 Details des Verfahrens

In diesem Abschnitt erläutern wir einige Details des Verfahrens mit lokalem Zeitschritt. Zuerst beschreiben wir, wie man die Zeitschrittmatrizen $\mathbf{Z}^{\mathbf{t}}, \mathbf{Z}^{\mathbf{t,0}}$ und $\mathbf{Z}, \mathbf{Z}^{\mathbf{0}}$ am einfachsten berechnet. Zwar benötigen wir diese Matrizen bezüglich der hierarchischen Basis, es erweist sich aber als einfacher, sie zunächst bezüglich der Lagrange-Basis aufzustellen und dann eine Transformation auf die hierarchische Basis durchzuführen. Desweiteren lässt sich dank dieser Diskussion über den Zusammenhang der Zeitschrittmatrizen in den verschiedenen Darstellungen die Invertierbarkeit von $\mathbf{Z}^{\mathbf{t}}$, die für die Entkopplung der System erforderlich ist, einfach nachweisen.

Zwei weitere Aspekte, die den Basiswechsel zwischen nodaler und hierarchischer Basis betreffen, besprechen wir danach. Dort geben wir an, wie man für den im nichtlinearen Problem auftretenden Vektor U einen Basiswechsel von der hierarchischen zur nodalen Basis und umgekehrt durchführt. Außerdem wollen wir noch einen kleinen Anstoß dafür geben, dass die Wahl der Basis des Testraumes numerisch auch eine Rolle spielt.

6.3.1 Zeitschrittmatrizen für die Lagrange-Basis

Wir beginnen damit, die Zeitschrittmatrizen für die Lagrange-Basis zu berechnen. Die Definition der Matrizen, Definition 6.1, liefert in diesem Fall sofort die folgenden Gleichungen. Es sei $l \in \mathbb{N}_0$, dann sind $\mathbf{Z^t}, \mathbf{Z} \in \mathbb{R}^{2^l \times 2^l}$ und $\mathbf{Z^{t,0}}, \mathbf{Z^0} \in \mathbb{R}^{2^l}$ für das Grundintervall $[0, k]$ gegeben durch

$$\mathbf{Z^t} = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & -1 & 1 & \\ & & & -1 & 1 \end{pmatrix}, \quad \mathbf{Z^{t,0}} = \begin{pmatrix} -1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix} \tag{6.30}$$

und

$$\mathbf{Z} = \frac{k}{2^{l+1}} \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ & 1 & \ddots & & \\ & & \ddots & 1 & \\ & & & 1 & 1 \end{pmatrix}, \quad \mathbf{Z^0} = \frac{k}{2^{l+1}} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}. \tag{6.31}$$

6.3.2 Zeitschrittmatrizen für die hierarchische Basis

Um die Zeitschrittmatrizen auf die hierarchische Darstellung zu transformieren, müssen wir die Basis der Ansatzfunktionen entsprechend transformieren.

Anstatt die hierarchische Basis iterativ ausgehend von der Knotenbasis auf $\mathcal{T}^0$ zu definieren, kann man auch für ein $l \in \mathbb{N}$ umgekehrt vorgehen und die hierarchische Basis zum Level l aus der Knotenbasis desselben Levels konstruieren. Die Idee der Iteration ist wie folgt.

1. Beginne mit der nodalen Basis auf $\mathcal{T}^l$.
2. Lasse alle Basisfunktionen, die zu den Knoten $\mathcal{R}^l$ gehören, fest. Diese Funktionen gehören schon der gesuchten hierarchischen Basis an.
3. Transformiere alle verbleibenden Funktionen in die Knotenbasis bezüglich $\mathcal{T}^{l-1}$.

4. Behalte die Funktionen, die zu den Knoten $\mathcal{R}^{l-1}$ gehören.

5. Führe das obige Verfahren iterativ fort.

Wir müssen nun nur noch Schritt 3 genauer ausführen, also beschreiben, wie die Transformation aussieht. Daher machen wir die folgende Beobachtung. Wir betrachten die nodale Basis $\mathcal{B}^l$ und wollen daraus die Basisfunktionen aus $\mathcal{B}^{l-1}$ konstruieren. Es sei $i \in R^{l,l-1}$, also ist $t_{i,l} \in \mathcal{T}^{l-1}$. Dann gilt

$$\psi^{l-1,i}(t) = \psi^{l,i}(t) + \begin{cases} \frac{1}{2}\psi^{l,i+1}(t), & i = 0, \\ \frac{1}{2}\psi^{l,i-1}(t), & i = 2^l, \\ \frac{1}{2}(\psi^{l,i-1}(t) + \psi^{l,i+1}(t)), & \text{sonst.} \end{cases}$$

Wir benötigen also den linken und rechten Nachbar *bezüglich des Ausgangslevels* um eine Basisfunktion des nächstniedrigeren Levels zu bekommen. Für den allgemeinen Fall setzen wir

$$K^{l,j} := \{m2^{l-j} \colon m = 0, \dots, 2^j\} = \bigcup_{n=0}^{j} R^{l,n}$$

für $j = 0, \dots, l$. Dies sind gerade die Indizes der Knoten in $\mathcal{T}^j$ bezüglich des Levels l. Dann definieren wir den *linken* und den *rechten Nachbar zu* $i = m2^{l-j} \in K^{l,j}$ *bezüglich* $j = 1, \dots, l$

$$\begin{aligned} \text{left}_j(i) &:= (m-1)2^{l-j}, \\ \text{right}_j(i) &:= (m+1)2^{l-j}. \end{aligned}$$

Damit haben wir alles beisammen, was wir für die Definition der Transformation benötigen. Wir definieren iterativ eine Folge $(\widetilde{\psi}_j^{l,i})_{j=0,\dots,l}$ für alle $i = 0, \dots, 2^l$. Dabei sei $\widetilde{\psi}_j^{l,i} = 0$, falls $i < 0$ oder $i \geq 2^l$. Das erspart uns die Fallunterscheidung, ob $t_{i,l}$ ein Randknoten ist oder nicht. Wir setzen zunächst $\widetilde{\psi}_l^{l,i} := \psi^{l,i}$ für alle $i = 0, \dots, 2^l$ und durchlaufen dann rückwärts für $j = l-1$ bis $j = 0$ die Vorschrift

$$\widetilde{\psi}_j^{l,i} := \begin{cases} \widetilde{\psi}_{j+1}^{l,i} + \frac{1}{2}(\widetilde{\psi}_{j+1}^{l,\text{left}_{j+1}(i)} + \widetilde{\psi}_{j+1}^{l,\text{right}_{j+1}(i)}), & i \in K^{l,j}, \\ \widetilde{\psi}_{j+1}^{l,i}, & \text{sonst.} \end{cases} \tag{6.32}$$

Am Ende setzen wir $\widetilde{\psi}^{l,i} := \widetilde{\psi}_0^{l,i}$ für $i = 0, \dots, 2^l$. Mittels dieser Vorschrift können wir Transformationsmatrizen $\mathbf{T}_{j+1}$ definieren, die unsere Zeitschrittmatrizen bezüglich der nodalen Basis des Ansatzraumes in die Zeitschrittmatrizen bezüglich der hierarchischen Basis des Ansatzraumes überführen. Es sei $\mathbf{T}_{j+1}$ die Einheitsmatrix mit den zusätzlichen Einträgen

$$(\mathbf{T}_{j+1})_{i,\text{left}_{j+1}(i)} = \frac{1}{2}, \ (\mathbf{T}_{j+1})_{i,\text{right}_{j+1}(i)} = \frac{1}{2}, \quad \text{für } i \in K^{l,j}. \tag{6.33}$$

Das ist gerade die Beziehung (6.32) in Matrixschreibweise. Zur Unterscheidbarkeit der Zeitschrittmatrizen in den verschiedenen Darstellungen bezeichnen wir hier mit $\widetilde{\mathbf{Z}}^{\mathbf{t,0}}, \widetilde{\mathbf{Z}}^{\mathbf{t}}$ die Zeitschrittmatrizen in hierarchischer und mit $\mathbf{Z}^{\mathbf{t,0}}, \mathbf{Z}^{\mathbf{t}}$ die Zeitschrittmatrizen in Knotenbasisdarstellung und wir definieren durch $[\widetilde{\mathbf{Z}}^{\mathbf{t,0}}, \widetilde{\mathbf{Z}}^{\mathbf{t}}]$ die Matrix, die durch Zusammensetzung der zwei Matrizen entsteht. Dann gilt also

$$[\widetilde{\mathbf{Z}}^{\mathbf{t,0}}, \widetilde{\mathbf{Z}}^{\mathbf{t}}] = [\mathbf{Z}^{\mathbf{t,0}}, \mathbf{Z}^{\mathbf{t}}]\mathbf{T}_l^T \cdots \mathbf{T}_1^T. \tag{6.34}$$

Analog gilt diese Beziehung für die anderen Zeitschrittmatrizen $\mathbf{Z}$ und $\mathbf{Z^0}$.

6.3.3 Basiswechsel für Vektoren

Für das nichtlineare Problem benötigen wir die Transformationen eines Vektors von der hierarchischen Basisdarstellung in die nodale und umgekehrt. Das funktioniert so ähnlich wie im vorherigen Abschnitt und wird bei Multi-Level-Methoden schon lange verwendet, siehe beispielsweise [Yse86, Section 4].

Der Einfachheit halber betrachten wir in diesem Abschnitt nur einen Polygonzug in der Zeit ohne zusätzliche Abhängigkeit vom Ort. Es sei U der Koeffizientenvektor in nodaler Basisdarstellung von

$$u(t) = \sum_{i=0}^{2^l} U^{l,i}\psi^{l,i}(t).$$

Wir suchen also die Transformation, um aus U den Koeffizientenvektor der hierarchischen Basisdarstellung $\widetilde{U}$ mit

$$u(t) = \sum_{i=0}^{2^l} \widetilde{U}^{l,i}\widetilde{\psi}^{l,i}(t)$$

zu bestimmen. Da es einfacher ist, die Koeffizienten der nodalen Darstellung aus den Koeffizienten der hierarchischen Darstellung zu berechnen als umgekehrt, werden wir diesen Weg gehen. Wieder definieren wir eine Folge aus Vektoren $(U_j^{l,i})_{j=0,\dots,l}$ für alle $i = 0, \dots, 2^l$ und setzen

$$U_0^{l,i} := \widetilde{U}^{l,i}.$$

Wir gehen levelweise vor und berechnen für $j = 1, \dots, l$

$$U_j^{l,i} := \begin{cases} U_{j-1}^{l,i} + \frac{1}{2}(\widetilde{U}_{j-1}^{l,\mathrm{left}_{j-1}(i)} + \widetilde{U}_{j-1}^{l,\mathrm{right}_{j-1}(i)}), & i \in R^{j,l}, \\ U_{j-1}^{l,i}, & \text{sonst.} \end{cases}$$

Dann ist $U^{l,i} = U_l^{l,i}$. Wie im vorherigen Abschnitt können wir diese Transformation als Matrix darstellen. Ist $\mathbf{S}_j$ die Einheitsmatrix mit den zusätzlichen Einträgen

$$(\mathbf{S}_j)_{i,\text{left}_{j-1}(i)} = \frac{1}{2},\ (\mathbf{S}_j)_{i,\text{right}_{\text{j-1}}(i)} = \frac{1}{2} \quad \text{für } i \in R^{j,l},$$

so ist

$$U = \mathbf{S}_l \cdots \mathbf{S}_1 \widetilde{U}. \tag{6.35}$$

Die Transformation $\Pi_{\text{hier}}^{\text{knot}}$, die wir für das nichtlineare Problem benötigen, ist dann durch $\mathbf{S}_l \cdots \mathbf{S}_1$ gegeben. Dabei müssen wir beachten, dass der Eintrag der Zeile i und der Spalte j in dem Fall dann auf den Vektor $(\widetilde{U}_\nu^{l,j})_{\nu \in \mathcal{N}_h}$ angewendet wird. Die $\mathbf{S}_j$ sind invertierbar, denn es ist $\mathbf{S}_j^{-1}$ die Einheitsmatrix mit den zusätzlichen Einträgen

$$(\mathbf{S}_j^{-1})_{i,\text{left}_{j-1}(i)} = -\frac{1}{2},\ (\mathbf{S}_j^{-1})_{i,\text{right}_{j-1}(i)} = -\frac{1}{2}, \quad \text{für } i \in R^{j,l}. \tag{6.36}$$

Daher ist $\Pi_{\text{knot}}^{\text{hier}}$ gegeben durch $(\mathbf{S}_l \cdots \mathbf{S}_1)^{-1} = \mathbf{S}_1^{-1} \cdots \mathbf{S}_l^{-1}$.

6.3.4 Invertierbarkeit von $Z^{\mathbf{t}}$

In diesem Abschnitt zeigen wir die Invertierbarkeit von $\mathbf{Z}^{\mathbf{t}}$ in der hierarchischen Basisdarstellung über die einfach zu zeigende Invertierbarkeit dieser Matrix in Knotenbasisdarstellung. Dazu benötigen wir aber zusätzliche Informationen über die Transformation der einen in die andere Darstellung.

6.8 Proposition.
Die Matrizen $\mathbf{T}_{j+1}$ sind für alle $j = 0, \ldots, l-1$ invertierbar, wobei die Inversen wie folgt gegeben sind. $\mathbf{T}_{j+1}^{-1}$ ist die Einheitsmatrix mit den zusätzlichen Einträgen

$$(\mathbf{T}_{j+1})_{i,\text{left}_{j+1}(i)} = -\frac{1}{2},\ (\mathbf{T}_{j+1})_{i,\text{right}_{j+1}(i)} = -\frac{1}{2}, \quad \textit{für } i \in K^{l,j}.$$

Beweis. Die Aussage ist einfach, wenn man sich die Beziehung (6.32) anschaut, durch welche $\mathbf{T}_{j+1}$ definiert ist. Um von

$$\widetilde{\psi}_{j+1}^{l,i} + \frac{1}{2}(\widetilde{\psi}_{j+1}^{l,\text{left}_{j+1}(i)} + \widetilde{\psi}_{j+1}^{l,\text{right}_{j+1}(i)})$$

für $i \in K^{l,j}$ die Funktion $\widetilde{\psi}_{j+1}^{l,i}$ zurückzuerhalten, muss man wieder

$$\frac{1}{2}(\widetilde{\psi}_{j+1}^{l,\text{left}_{j+1}(i)} + \widetilde{\psi}_{j+1}^{l,\text{right}_{j+1}(i)})$$

abziehen. Dabei muss man nur beachten, dass

$$\widetilde{\psi}_{j+1}^{l,\mathrm{left}_{j+1}(i)} = \widetilde{\psi}_{j}^{l,\mathrm{left}_{j+1}(i)} \quad \text{und} \quad \psi_{j+1}^{l,\mathrm{right}_{j+1}(i)} = \widetilde{\psi}_{j}^{l,\mathrm{right}_{j+1}(i)}$$

gilt, da $\mathrm{left}_{j+1}(i)$ und $\mathrm{right}_{j+1}(i)$ nach Definition in $R^{j+1,l}$ und somit nicht in $K^{l,j}$ liegen. □

6.9 Proposition.
Die Zeitschrittmatrix $\mathbf{Z^t}$ ist für ein beliebiges Verfeinerungslevel $l \in \mathbb{N}_0$ invertierbar.

Beweis. Wir verwenden die Gleichung (6.34). Wir schreiben die Transformationsmatrizen in der Form

$$\mathbf{T}_{j+1} = \begin{pmatrix} 1 & e_{j+1} \\ 0_{2^l} & \widetilde{\mathbf{T}}_{j+1} \end{pmatrix},$$

wobei $0_{2^l} \in \mathbb{R}^{2^l \times 1}$ der Nullvektor ist und $e_{j+1} \in \mathbb{R}^{1\times 2^l}$ der Nullvektor ist bis auf den Eintrag, der dem rechten Nachbar des Anfangsknoten zugeordnet ist, dort ist e_{j+1} gleich $\frac{1}{2}$. Die erste Spalte verschwindet bis auf den ersten Eintrag, weil $\mathrm{left}_{j+1}(i) \neq 0$ für alle $i \in K^{l,j}$ ist. Also ist

$$[\widetilde{\mathbf{Z}}^{\mathbf{t},\mathbf{0}}, \widetilde{\mathbf{Z}}^{\mathbf{t}}] = [\mathbf{Z}^{\mathbf{t},\mathbf{0}}, \mathbf{Z}^{\mathbf{t}}]\mathbf{T}_l^T = [\mathbf{Z}^{\mathbf{t},\mathbf{0}} + \mathbf{Z}^{\mathbf{t}} e_l^T, \mathbf{Z}^{\mathbf{t}}\widetilde{\mathbf{T}}_l^T].$$

Führen wir das iterativ fort, so folgt $\widetilde{\mathbf{Z}}^{\mathbf{t}} = \mathbf{Z}^{\mathbf{t}}\widetilde{\mathbf{T}}_l^T \cdots \widetilde{\mathbf{T}}_1^T$. Nun ist offensichtlich $\det(\mathbf{Z}^{\mathbf{t}}) = 1$ (siehe (6.30)). Es verbleibt also nur noch zu zeigen, dass die $\widetilde{\mathbf{T}}_{j+1}$ invertierbar sind. Aus Proposition 6.8 wissen wir, dass $\mathbf{T}_{j+1}$ invertierbar ist und da $\det(\mathbf{T}_{j+1}) = \det(1)\det(\widetilde{\mathbf{T}}_{j+1})$ gilt, folgt die Behauptung. □

6.3.5 Basiswechsel des Testraumes

Bisher haben wir nur die Basis des Ansatzraumes dem Problem angepasst. Die Besetzungsstruktur lässt sich durch die zusätzliche Anpassung des Testraumes weiter ausdünnen. Ohne weitere Anpassungen haben die Zeitschrittmatrizen beispielsweise für $l = 2$ die Form

$$\mathbf{Z^t} = \begin{pmatrix} 1 & \frac{1}{2} & 0 & \frac{1}{4} \\ -1 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & -\frac{1}{2} & 1 & \frac{1}{4} \\ 0 & -\frac{1}{2} & -1 & \frac{1}{4} \end{pmatrix}, \quad \mathbf{Z} = \frac{k}{32}\begin{pmatrix} 4 & 2 & 0 & 1 \\ 4 & 6 & 0 & 3 \\ 0 & 6 & 4 & 5 \\ 0 & 2 & 4 & 7 \end{pmatrix}.$$

Die Anzahl der verschwindenden Einträge pro Spalte hängt davon ab, zu welchem Index in $R^{l,m}$ sich die zur Spalte gehörige Basisfunktion des Ansatzraumes befindet. Mit $1 \leq m \leq l$ sind für einen Index in $R^{l,m}$ genau 2^{l-m+1}

Einträge ungleich 0 und für die Funktion $\widetilde{\psi}^{l,2^l}$ verschwindet keiner der Einträge. Beide Behauptungen ergeben sich sofort aus Definition 6.1. Wählen wir aber statt (6.7) eine andere Basis des Testraumes, so können wir Verbesserungen erreichen. So führt die Wahl

$$\left\{\phi\colon \Omega_k \to \mathbb{R} \,|\, \phi(t,x) = \partial_t \widetilde{\psi}^{l,j}(t)\varphi_\nu(x),\ j = 1,\dots,2^l, \nu \in \mathcal{N}_h\right\}$$

auf

$$\overline{\mathbf{Z}}^{\mathbf{t}} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \overline{\mathbf{Z}} = \frac{k}{16}\begin{pmatrix} 0 & -2 & 0 & -1 \\ 4 & 0 & -4 & 4 \\ 0 & 2 & 0 & 1 \\ 4 & 8 & 4 & 2 \end{pmatrix}.$$

Diese Matrizen erhält man aus $\mathbf{Z}^{\mathbf{t}}, \mathbf{Z}$ durch Linksmultiplikation mit einer geeigneten Transformationsmatrix, für das Level 2 wäre das beispielsweise

$$\mathbf{P}_2 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Die Einträge von $\overline{\mathbf{Z}}^{\mathbf{t}}$ sind für diese Basis durch

$$(\overline{\mathbf{Z}}^{\mathbf{t}})_{i,j} = \int_I \partial_t \widetilde{\psi}^{l,j} \partial_t \widetilde{\psi}^{l,i}\,\mathrm{d}t = \left\|\partial_t \widetilde{\psi}^{l,i}\right\|^2_{L^2(I)} \delta_{ij}$$

für $i,j = 1,\dots,2^l$ gegeben. Die Orthogonalität lässt sich einfach einsehen. Es gibt im Wesentlichen zwei Fälle.

1. Gilt $i,j \in R^{l,m}$ für ein $1 \le m \le l$, gehören also beide Indizes zum gleichen Level, dann interessiert uns nur der Fall $i \neq j$. In diesem Fall sind aber die Träger von $\partial_t \widetilde{\psi}^{l,i}$ und $\partial_t \widetilde{\psi}^{l,j}$ disjunkt oder schneiden sich nur in einem Punkt. Daher verschwindet das Integral des Produktes dieser beiden Funktionen.

2. Gilt $i \in R^{l,m}$ und $j \in R^{l,m+n}$ für $0 \le m \le l$ und $1 \le n \le m-l$, so lässt sich feststellen, dass $\partial_t \widetilde{\psi}^{l,i}$ auf dem Träger von $\partial_t \widetilde{\psi}^{l,j}$ konstant 1 oder konstant -1 ist. Das Integral auf I über $\partial_t \widetilde{\psi}^{l,j}$ ist aber Null.

In der Matrix $\overline{\mathbf{Z}}^{\mathbf{t}}$ lassen sich also die einzelnen Verfeinerungslevel komplett entkoppeln. Es ist nicht möglich, dies simultan auch in $\overline{\mathbf{Z}}$ zu schaffen. Die Einträge für $\mathbf{Z}$ lauten

$$(\overline{\mathbf{Z}})_{i,j} = \int_I \widetilde{\psi}^{l,j} \partial_t \widetilde{\psi}^{l,i}\,\mathrm{d}t$$

für $i,j = 1,\dots,2^l$. Wieviele nichtverschwindende Einträge stehen dann in der Spalte j? Es sei $j \in R^{l,m}$ und $i \in R^{l,n}$. Wir unterscheiden die Fälle $m < n$, $m = n$ und $m > n$.

1. Sei $m < n$. Der Träger von $\widetilde{\psi}^{l,j}$ ist also größer als der von $\partial_t \widetilde{\psi}^{l,i}$, sogar genau 2^{n-m} mal so groß. Entsprechend verschwinden genau 2^{n-m} Einträge nicht.

2. Sei $m = n$. In diesem Fall sind die Träger der zu multiplizierenden Funktionen entweder bis auf einen Punkt disjunkt oder $i = j$. In beiden Fällen ist das Ergebnis der Integration des Produktes Null.

3. Sei $m > n$. Der Träger von $\widetilde{\psi}^{l,j}$ ist kleiner als der von $\partial_t \widetilde{\psi}^{l,i}$, es existiert also genau ein $i \in R^{l,n}$ für welches die Träger von $\widetilde{\psi}^{l,j}$ und $\partial_t \widetilde{\psi}^{l,i}$ nicht bis auf einen Punkt disjunkt sind.

Wir zählen für $j \in R^{l,m}$, $m \geq 1$ zusammen, wieviele Einträge es gibt. Es sind

$$\underbrace{\left(\sum_{n=m+1}^{l} 2^{n-m} \right)}_{=\sum_{n=1}^{l-m} 2^n} + 0 + \left(\sum_{i=0}^{m-1} 1 \right) = m + 2^{l-m+1} - 2.$$

Also haben wir pro Spalte $m-2$ nichtverschwindende Einträge mehr in $\overline{\mathbf{Z}}$ als in $\mathbf{Z}$, andererseits ist die Diagonalmatrix $\overline{\mathbf{Z}}^{\mathbf{t}}$ eine deutliche Verbesserung gegenüber $\mathbf{Z}^{\mathbf{t}}$. In den implementierten Verfahren wurde nicht explizit versucht, Vorteile aus der speziellen Struktur des Problems zu ziehen. Da für $k \to 0$ die Matrizen mit $\mathbf{Z}^{\mathbf{t}}$ oder $\overline{\mathbf{Z}}^{\mathbf{t}}$ das Problem dominieren, könnte man entsprechende Vorkonditionierer zur Lösung der Gleichungssysteme im Newton-Verfahren verwenden.

6.4 Numerische Ergebnisse

Wir werden das Galerkin-Verfahren mit lokalem Zeitschritt an zwei Problemen testen. Als erstes diskretisieren wir die Wellengleichung, also ein lineares Problem, und zeigen damit, dass lineare Probleme mit der gewünschten Ordnung $k + h^p$ approximiert werden. Das zweite Problem ist das blow-up-Problem aus Abschnitt 5.4.1.

Wir diskretisieren beide Probleme auf $(0, T) \times \Omega$ jeweils mit linearen und quadratischen Elementen im Ort. Wir werden den Fehler auf verschiedenen Teilgebieten von Ω berechnen. Zum einen geben wir den Fehler auf Ω und zum anderen auf Ω_{loc} an, dies ist das Teilgebiet, auf dem wir einen lokalen Teilschritt durchführen. Alle Konvergenzuntersuchungen werden bezüglich des lokalen Zeitschrittes $k_{\text{loc}} = 2^{-l}k$ für ein vorgegebenes k und variable Zeitlevel l durchgeführt. Der Fehler, der außerhalb des verfeinerten Gebietes gemacht wird, ist in beiden Problemen um Größenordnungen kleiner als der globale Fehler, wie wir in den Fehlerdiagrammen sehen werden. Wir untersuchen also die Probleme auf Konvergenz für $h \to 0$ und $l \to \infty$.

Im Fall linearer Elemente im Ort geben wir den H^1-Fehler für die Funktion u zur Zeit T in Abhängigkeit von $k_{\text{loc}} + h$ an, im Fall quadratischer Elemente gehen wir vor wie in Abschnitt 5.4. Wir betrachten in diesem Fall die Konvergenz in der Zeit und im Ort getrennt, wobei wir die jeweils andere Diskretisierung so genau wählen, dass sie keinen großen Einfluss hat. Auch hier sind in allen Abbildungen wieder Steigungsdreiecke eingezeichnet, die einen Vergleich zur Bestimmung der experimentellen Konvergenzordnung liefern.

6.4.1 Lineare Wellengleichung

Wir betrachten die Wellengleichung in Form eines Systems erster Ordnung.

$$\partial_t u = v, \tag{6.37a}$$
$$\partial_t v = \Delta u, \tag{6.37b}$$

auf $\Omega = (-50, 50)$ und für $t \in (0, T)$, $T = 10$, mit den Anfangsbedingungen

$$u_0(x) := \mathrm{e}^{-x^2},$$
$$v_0(x) := 0,$$

für $x \in \Omega$. Die exakte Lösung ist

$$u(t,x) = \frac{1}{2}\left(\mathrm{e}^{-(x-t)^2} + \mathrm{e}^{-(x+t)^2}\right),$$

bei passenden Dirichlet-Randbedingungen. Wir verwenden eine dynamische lokale Verfeinerung, das heißt, wir verfeinern nur in Knoten im Bereich

$$\Omega_{\text{loc}}(t) := \{x \in \Omega \colon |x \pm t| < 5\}.$$

In der Verfahrensdurchführung bedeutet das, dass wir in jedem Zeitschritt den Vektor ℓ anpassen, im n-ten Zeitschritt sei

$$\ell^n := \begin{cases} l & , \nu \in \mathcal{N}_h,\ x_\nu \in \Omega_{\text{loc}}(t^n), \\ 0 & , \text{sonst.} \end{cases}$$

Das in Gl. (6.38) definierte Verfahren berechnet dann eine Folge von Vektoren $U^{n,\ell^n}, V^{n,\ell^n}$ für $n = 1, \dots, N$ mit $Nk = T$. Wir setzen $k = 0.1$. U^0, V^0 bestimmen wir als Interpolation von u_0, v_0 in $\mathcal{S}_h^p$. Wir führen Berechnungen für $p = 1, 2$ durch, also diskretisieren im Ort mit linearen und quadratischen Elementen. Wir lösen also für $n = 1, \dots, N$ das Gleichungssystem

$$\begin{aligned} &\mathbf{P}_{\ell^n}(\mathbf{Z}^{\mathbf{t}} \otimes \mathbf{M})\mathbf{P}_{\ell^n}^T U^{n,\ell^n} + \mathbf{P}_{\ell^n}(\mathbf{Z}^{\mathbf{t,0}} \otimes \mathbf{M})U^{n-1} \\ &\qquad = \mathbf{P}_{\ell^n}(\mathbf{Z} \otimes \mathbf{M})\mathbf{P}_{\ell^n}^T V^{n,\ell^n} + \mathbf{P}_{\ell^n}(\mathbf{Z}^{\mathbf{0}} \otimes \mathbf{M})V^{n-1}, \end{aligned} \tag{6.38a}$$
$$\begin{aligned} &\mathbf{P}_{\ell^n}(\mathbf{Z}^{\mathbf{t}} \otimes \mathbf{M})\mathbf{P}_{\ell^n}^T V^{n,\ell^n} + \mathbf{P}_{\ell^n}(\mathbf{Z}^{\mathbf{t,0}} \otimes \mathbf{M})V^{n-1} \\ &\qquad = \mathbf{P}_{\ell^n}(\mathbf{Z} \otimes \mathbf{K})\mathbf{P}_{\ell^n}^T U^{n,\ell^n} + \mathbf{P}_{\ell^n}(\mathbf{Z}^{\mathbf{0}} \otimes \mathbf{K})U^{n-1}. \end{aligned} \tag{6.38b}$$

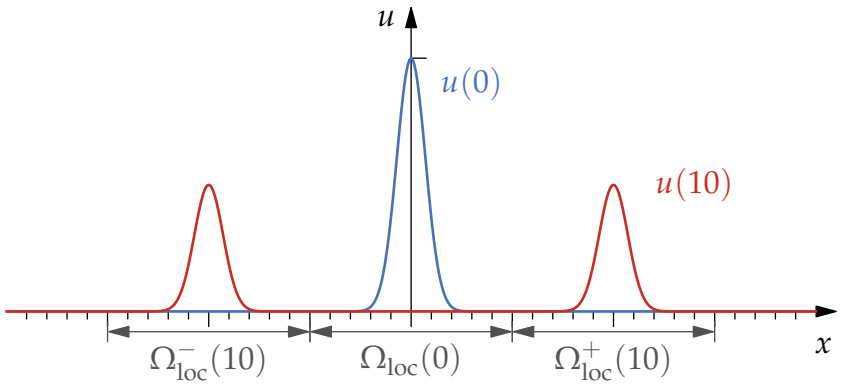


Abbildung 6.7: Lösungen der Wellengleichung zum Zeitpunkt $t = 0$ (blau) und $t = 10$ (blau). Die lokal verfeinerten Teilgebiete $\Omega_{\text{loc}}(t)$ sind unter der x-Achse angegeben, wobei $\Omega_{\text{loc}}(10) = \Omega^-_{\text{loc}}(10) \cup \Omega^+_{\text{loc}}(10)$

U^{n-1}, V^{n-1} tragen im Exponenten kein ℓ^{n-1}, da das nur die Werte von u_{kh}, v_{kh} im Punkt t^{n-1} sind. Nachdem wir $U^{n,\ell^n}, V^{n,\ell^n}$ berechnet haben, entfernen wir daher alle Einträge, die zu einem Level $l \geq 1$ gehören und behalten nur die Koeffizientenvektoren des Zeitpunktes t^n.

Für lineare Elemente sehen wir in Abb. 6.8 und Tab. 6.1, dass sich der Fehler auf Ω und Ω_{loc} im Wesentlichen gleich verhält, die Wahl von Ω_{loc} ist also offensichtlich gut. Weiterhin liegt die experimentelle Fehlerordnung teilweise deutlich über der erwarteten linearen Ordnung.

Für quadratische Elemente betrachten wir die Konvergenz in der Zeit und im Ort getrennt. Abb. 6.9a und Tab. 6.2 zeigen ein Verhalten, dass wir schon in Abschnitt 5.4 beobachten konnten: Die experimentelle Konvergenzordnung in der Zeit ist 2, da es sich aber um ein lineares Problem handelt, war das zu erwarten (vergleiche Bemerkung 5.33). Die experimentelle Konvergenzordnung im Ort, zu sehen in Abb. 6.9b und Tab. 6.3, zeigt eine quadratische Konvergenzordnung und entspricht damit der Theorie.

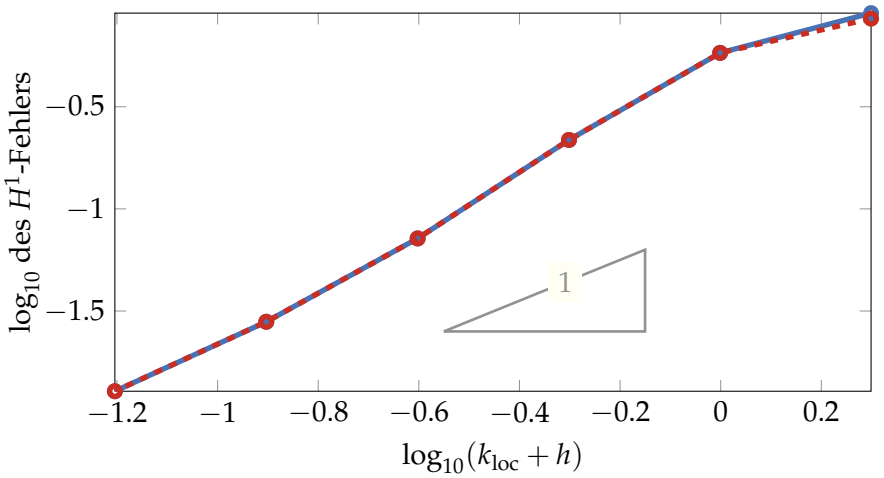


Abbildung 6.8: Lineare Wellengleichung: Konvergenz für $p = 1$ in Abhängigkeit von $k_{\text{loc}} + h$. Blau der Fehler in Ω, rot gestrichelt der Fehler in Ω_{loc}.

l	k_{loc}	h	$\|(u - u_{kh})(T)\|_{H^1(\Omega)}$	eoc
0	1.0	1.0	9.09e-01	
1	0.5	0.5	5.81e-01	6.49e-01
2	0.25	0.25	2.18e-01	1.42e+00
3	0.125	0.125	7.17e-02	1.60e+00
4	0.0625	0.0625	2.80e-02	1.36e+00
5	0.03125	0.03125	1.28e-02	1.13e+00

(a) Konvergenz in Ω.

l	k_{loc}	h	$\|(u - u_{kh})(T)\|_{H^1(\Omega_{\text{loc}}(T))}$	eoc
0	1.0	1.0	8.54e-01	
1	0.5	0.5	5.81e-01	5.59e-01
2	0.25	0.25	2.18e-01	1.42e+00
3	0.125	0.125	7.17e-02	1.60e+00
4	0.0625	0.0625	2.80e-02	1.36e+00
5	0.03125	0.03125	1.28e-02	1.13e+00

(b) Konvergenz in $\Omega_{\text{loc}}(T)$.

Tabelle 6.1: Lineare Wellengleichung: Konvergenztabelle für lineare Elemente in Abhängigkeit von $k_{\text{loc}} + h$.

l	k_{loc}	h	$\|(u-u_{kh})(T)\|_{H^1(\Omega)}$	eoc
0	1.0	0.0625	1.07e+00	
1	0.5	0.0625	8.10e-01	4.07e-01
2	0.25	0.0625	3.57e-01	1.18e+00
3	0.125	0.0625	1.03e-01	1.80e+00
4	0.0625	0.0625	2.62e-02	1.97e+00
5	0.03125	0.0625	6.60e-03	1.99e+00

(a) Konvergenz in Ω.

l	k_{loc}	h	$\|(u-u_{kh})(T)\|_{H^1(\Omega_{\text{loc}}(T))}$	eoc
0	1.0	0.0625	9.96e-01	
1	0.5	0.0625	8.07e-01	3.03e-01
2	0.25	0.0625	3.57e-01	1.18e+00
3	0.125	0.0625	1.03e-01	1.80e+00
4	0.0625	0.0625	2.62e-02	1.97e+00
5	0.03125	0.0625	6.60e-03	1.99e+00

(b) Konvergenz in $\Omega_{\text{loc}}(T)$.

Tabelle 6.2: Lineare Wellengleichung: Konvergenz in der Zeit für $p = 2$.

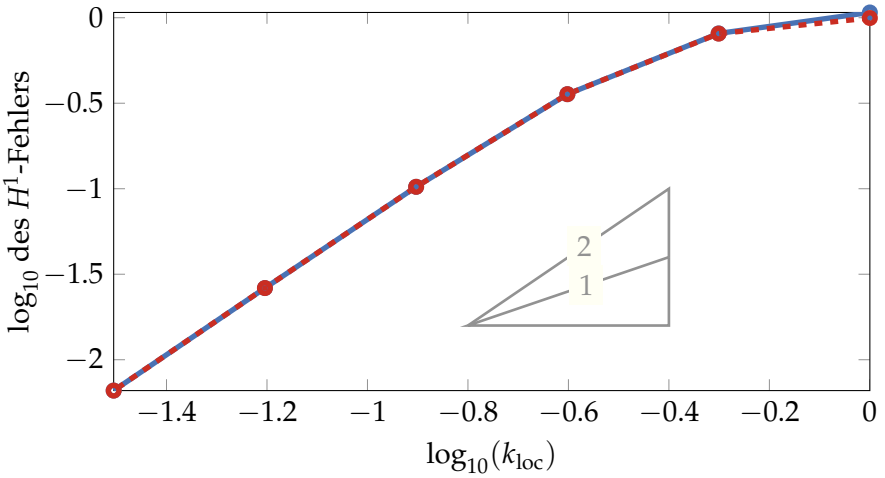
l	k_{loc}	h	$\|(u-u_{kh})(T)\|_{H^1(\Omega)}$	eoc
5	0.003125	2.0	9.10e-01	
5	0.003125	1.0	3.45e-01	1.40e+00
5	0.003125	0.5	4.08e-02	3.08e+00
5	0.003125	0.25	7.39e-03	2.47e+00
5	0.003125	0.125	1.78e-03	2.05e+00
5	0.003125	0.0625	4.50e-04	1.99e+00

(a) Konvergenz in Ω.

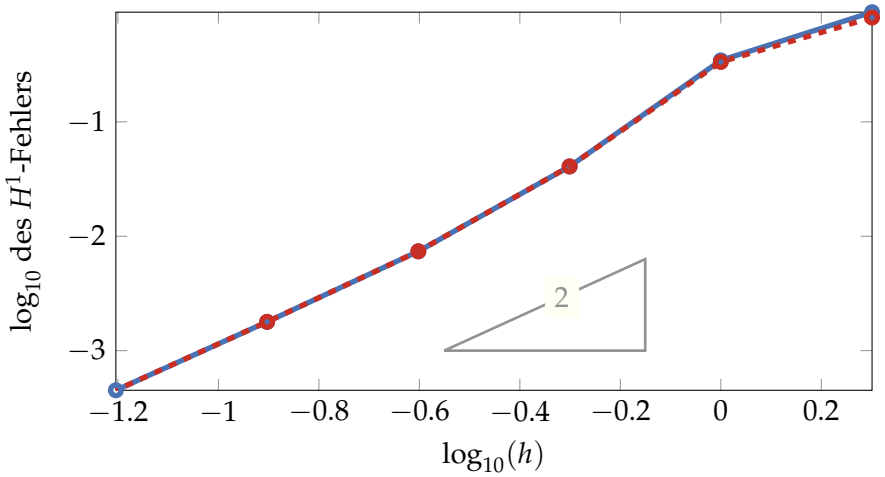
l	k_{loc}	h	$\|(u-u_{kh})(T)\|_{H^1(\Omega_{\text{loc}}(T))}$	eoc
5	0.003125	2.0	8.17e-01	
5	0.003125	1.0	3.34e-01	1.29e+00
5	0.003125	0.5	4.08e-02	3.03e+00
5	0.003125	0.25	7.39e-03	2.47e+00
5	0.003125	0.125	1.78e-03	2.05e+00
5	0.003125	0.0625	4.50e-04	1.99e+00

(b) Konvergenz in $\Omega_{\text{loc}}(T)$.

Tabelle 6.3: Lineare Wellengleichung: Konvergenz im Ort für $p = 2$.



(a) Konvergenz in der Zeit in Abhängigkeit von k_{loc}, mit $h = 0.0625$ fest.



(b) Konvergenz im Ort in Abhängigkeit von h, wobei $k_{\text{loc}} = 0.003125$ fest.

Abbildung 6.9: Lineare Wellengleichung: Konvergenz für $p = 2$. Blau der Fehler in Ω, rot gestrichelt der Fehler in Ω_{loc}.

6.4.2 Das blow-up-Problem

Um die nichtlineare Version des lokalen Zeitschrittverfahrens zu überprüfen, verwenden wir das gleiche nichtlineare blow-up-Problem wie in Abschnitt 5.4.2. Wir berechnen die Lösung bis $T = 10$ und benutzen lokal verfeinerte Zeitschritte in $\Omega_{\text{loc}} := (10, 60)$, so dass der Puls im Wesentlichen in Ω_{loc} verläuft. Die exakte Lösung u ist in Abb. 6.10 zum Startzeitpunkt $t = 0$ sowie zum Zeitpunkt $T = 10$ dargestellt, Ω_{loc} ist unter der x-Achse eingezeichnet. Die numerische Lösung wird durch Gl. (6.25) mit Projektion berechnet, wir

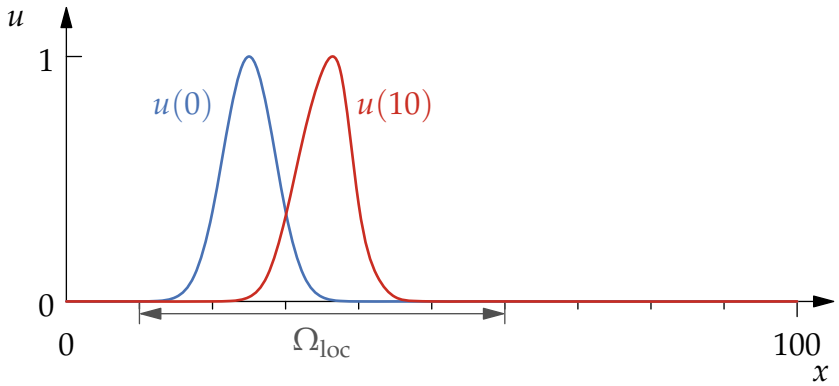


Abbildung 6.10: Lösung des blow-up-Problems zum Zeitpunkt $t = 0$ (blau) und zum Zeitpunkt $t = 10$ (rot). Unter der x-Achse ist das lokal verfeinerte Gebiet Ω_{loc} gekennzeichnet.

lösen also das nichtlineare Gleichungssystem

$$\begin{aligned}
\mathbf{P}_\ell(\mathbf{Z}\otimes\mathbf{M})\mathbf{P}_\ell^T V^{n,\ell} + \mathbf{P}_\ell(\mathbf{Z^0}\otimes\mathbf{M})V^{n-1} & \\
= \mathbf{P}_\ell(\mathbf{Z}\otimes\mathbf{M})\mathbf{P}_\ell^T(U^{n,\ell} + \Pi^{\text{hier}}_{\text{knot}} f(\Pi^{\text{knot}}_{\text{hier}} U^{n,\ell})) & \\
+ \mathbf{P}_\ell(\mathbf{Z^0}\otimes\mathbf{M})(U^{n-1} + f(U^{n-1})), & \\
\mathbf{P}_\ell(\mathbf{Z^t}\otimes\mathbf{M})\mathbf{P}_\ell^T V^{n,\ell} + \mathbf{P}_\ell(\mathbf{Z^{t,0}}\otimes\mathbf{M})V^{n-1} & \\
= \mathbf{P}_\ell(\mathbf{Z}\otimes\mathbf{M})\mathbf{P}_\ell^T W^{n,\ell} + \mathbf{P}_\ell(\mathbf{Z^0}\otimes\mathbf{M})W^{n-1}, & \\
\mathbf{P}_\ell(\mathbf{Z^t}\otimes\mathbf{M})\mathbf{P}_\ell^T W^{n,\ell} + \mathbf{P}_\ell(\mathbf{Z^{t,0}}\otimes\mathbf{M})W^{n-1} & \\
= \mathbf{P}_\ell(\mathbf{Z}\otimes\mathbf{K})\mathbf{P}_\ell^T U^{n,\ell} + \mathbf{P}_\ell(\mathbf{Z^0}\otimes\mathbf{K})U^n &
\end{aligned}$$

für $n = 1, \ldots, N$ mit $Nk = 10$ und $k = 0.1$. Dieses nichtlineare Gleichungssystem lösen wir durch das ungedämpfte Newtonverfahren (genauere Diskussion siehe Abschnitt 4.3.2) mit den folgenden Parametern

$$\begin{aligned}
n_{\max} &= 20, \\
\text{TOL}_U &= 10^{-8}, \\
\text{TOL}_F &= 10^{-10}.
\end{aligned}$$

Die Ergebnisse der Experimente decken sich gut mit denen der linearen Wellengleichung in Abschnitt 6.4.1.

Für lineare Elemente zeigen Abb. 6.11 und Tab. 6.4 wieder eine lineare Abhängigkeit des Fehlers von $k_{\text{loc}} + h$, wobei hier die experimentelle Fehlerordnung immer ganz leicht unter 1 liegt. Der Fehler in $\Omega \setminus \Omega_{\text{loc}}$ ist auch wieder um Größenordnungen kleiner als der in Ω_{loc}, weshalb der globale Fehler und der Fehler im lokal verfeinerten Teilgebiet beinahe identisch sind.

Auch bei diesem nichtlinearen Problem ist die experimentelle Konvergenzordnung in der Zeit für quadratische Elemente in Abb. 6.12a und Tab. 6.5

wieder deutlich besser als linear. Die experimentelle Konvergenzordnung im Ort nähert sich für kleine h der 2 an, wie man in Abb. 6.12b und Tab. 6.6 sehen kann.

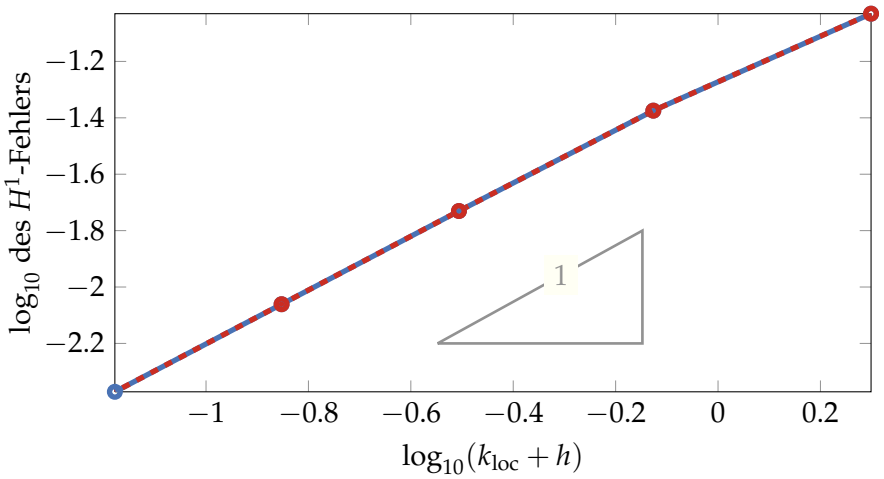


Abbildung 6.11: blow-up-Problem: Konvergenz für lineare Elemente in Abhängigkeit von $k_{\text{loc}} + h$.

l	k_{loc}	h	$\|(u - u_{kh})(T)\|_{H^1(\Omega)}$	eoc
0	1.0	1.0	9.32e-02	
1	0.5	0.5	4.22e-02	8.09e-01
2	0.25	0.25	1.86e-02	9.37e-01
3	0.125	0.125	8.69e-03	9.54e-01
4	0.0625	0.0625	4.25e-03	9.54e-01

(a) Konvergenz in Ω.

l	k_{loc}	h	$\|(u - u_{kh})(T)\|_{H^1(\Omega_{\text{loc}})}$	eoc
0	1.0	1.0	9.32e-02	
1	0.5	0.5	4.22e-02	8.09e-01
2	0.25	0.25	1.86e-02	9.37e-01
3	0.125	0.125	8.69e-03	9.54e-01
4	0.0625	0.0625	4.25e-03	9.54e-01

(b) Konvergenz in Ω_{loc}.

Tabelle 6.4: blow-up-Problem: Konvergenztabelle für lineare Elemente in Abhängigkeit von $k_{\text{loc}} + h$.

l	k_{loc}	h	$\|(u-u_{kh})(T)\|_{H^1(\Omega)}$	eoc
0	1.0	0.125	9.72e-02	
1	0.5	0.125	3.88e-02	1.33e+00
2	0.25	0.125	1.26e-02	1.63e+00
3	0.125	0.125	3.77e-03	1.74e+00
4	0.0625	0.125	1.39e-03	1.44e+00

(a) Konvergenz in Ω.

l	k_{loc}	h	$\|(u-u_{kh})(T)\|_{H^1(\Omega_{\text{loc}})}$	eoc
0	1.0	0.125	9.72e-02	
1	0.5	0.125	3.88e-02	1.33e+00
2	0.25	0.125	1.26e-02	1.63e+00
3	0.125	0.125	3.77e-03	1.74e+00
4	0.0625	0.125	1.39e-03	1.44e+00

(b) Konvergenz in Ω_{loc}.

Tabelle 6.5: blow-up-Problem: Konvergenz in der Zeit für $p = 2$.

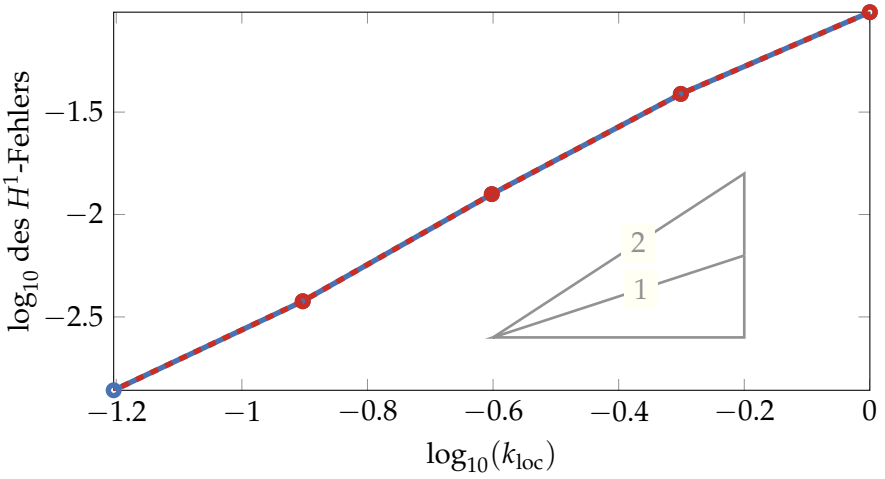
l	k_{loc}	h	$\|(u-u_{kh})(T)\|_{H^1(\Omega)}$	eoc
4	0.00625	2.0	6.32e-02	
4	0.00625	1.0	2.57e-02	1.30e+00
4	0.00625	0.5	8.16e-03	1.65e+00
4	0.00625	0.25	2.25e-03	1.86e+00
4	0.00625	0.125	5.83e-04	1.95e+00

(a) Konvergenz in Ω.

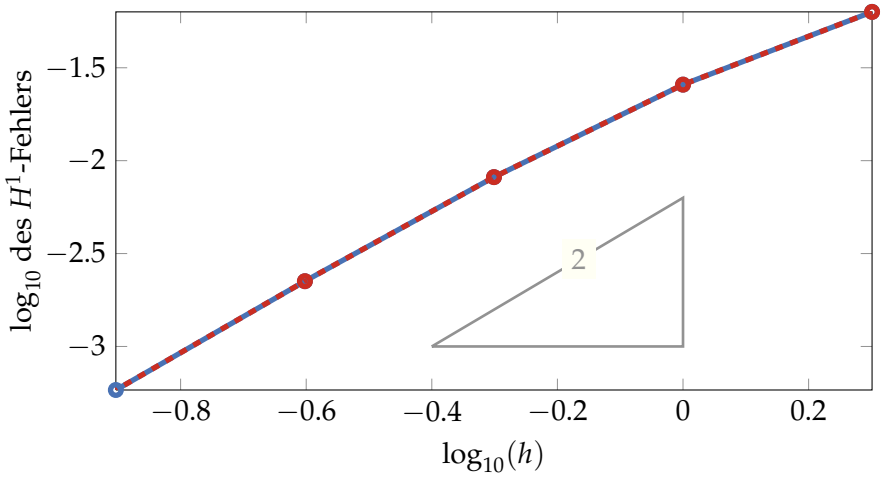
l	k_{loc}	h	$\|(u-u_{kh})(T)\|_{H^1(\Omega_{\text{loc}})}$	eoc
4	0.00625	2.0	6.32e-02	
4	0.00625	1.0	2.57e-02	1.30e+00
4	0.00625	0.5	8.16e-03	1.65e+00
4	0.00625	0.25	2.25e-03	1.86e+00
4	0.00625	0.125	5.83e-04	1.95e+00

(b) Konvergenz in Ω_{loc}.

Tabelle 6.6: blow-up-Problem: Konvergenz im Ort für $p = 2$.



(a) Konvergenz in der Zeit in Abhängigkeit von k_{loc}, wobei $h = 0.125$ fest.



(b) Konvergenz im Ort in Abhängigkeit von h, wobei $k_{\text{loc}} = 0.00625$ fest.

Abbildung 6.12: blow-up-Problem: Konvergenz für $p = 2$. Blau der Fehler in Ω, rot gestrichelt der Fehler in Ω_{loc}.

6.5 Zusammenfassung und Fazit

Die Verwendung eines stetigen Galerkinverfahrens erlaubt die Definition eines lokalen Zeitschrittverfahrens auf eine für Galerkinverfahren natürliche Art und Weise. Man koppelt die Orts- und Zeitdiskretisierung in den Ansatz- und Testräumen aneinander und erreicht dadurch, dass die Wahl der Räume der Zeitdiskretisierung von den Freiheitsgraden der Ortsdiskretisierung abhängig sein dürfen. Durch die Wahl einer hierarchischen Basis für den Ansatzraum in der Zeit wird sogar die Struktur erhalten, die man von Tensorelementansätzen kennt.

Das vorgestellte Verfahren bietet sich an, wenn man weiß, dass sich ein örtlich begrenzter Puls ausbreitet oder wenn Ω lokal stark aufgelöst werden muss. Um durch einen uniformen Zeitschritt nicht an Genauigkeit zu verlieren oder zu viel Rechenzeit für einen global feineren Zeitschritt zu brauchen, kann man in den entsprechend ausgezeichneten Teilgebieten einen lokal feineren Zeitschritt durchführen.

Für beide numerischen Beispiele gilt $\Omega \subset \mathbb{R}$, das bedeutet aber nicht, dass das Verfahren nur im Eindimensionalen funktioniert. In der Herleitung haben wir generell keine Voraussetzungen daran gestellt, in welcher Raumdimension wir uns befinden oder was für einen Finite-Elemente-Raum wir für die Ortsdiskretisierung benutzen. Prinzipiell müsste es sogar möglich sein, ein lokales Zeitschrittverfahren für höhere Ordnungen der Elemente der Zeitdiskretisierung zu definieren. Für zweidimensionale oder dreidimensionale Probleme ist die verwendete MATLAB-Implementierung aber nicht speichereffizient genug.

Es verbleibt die Frage, ob das lokale Zeitschrittverfahren weniger Rechendauer benötigt, als wenn man global mehrere aber kleinere Zeitschritte macht. Dazu lässt sich sagen, dass die bisherige Implementierung des Verfahrens noch nicht darauf angelegt war, besonders effizient zu arbeiten, insbesondere wird der Standardlöser für Gleichungssysteme von MATLAB verwendet. Dennoch ist für das lineare Wellenproblem in den numerischen Ergebnissen das lokale Zeitschrittverfahren mit Level l ungefähr doppelt so schnell wie eine Durchführung ohne lokalen Zeitschritt, aber mit Zeitschritt $2^{-l}k$. Die Rechenzeit hängt entscheidend davon ab, wie groß das Teilgebiet ist, in dem man einen lokalen Zeitschritt verwendet und davon, wie groß l ist. Für große Verfeinerungslevel l werden die auftretenden Gleichungssysteme zu groß. Für das nichtlineare Problem sind beide Vorgehensweisen ungefähr gleich schnell. In diesem Fall fällt die Anwendung des Newton-Verfahrens negativ ins Gewicht. Um dort das lokale Zeitschrittverfahren effizienter zu machen, müsste man darüber nachdenken, wie man die Struktur der entstehenden Gleichungssysteme ausnutzen kann, damit man einen schnelleren Löser erhält.

Literaturverzeichnis

[AAEH13] Shirin Afzal, Vahid Ahmadi, and Majid Ebnali-Heidari: *All-optical tunable photonic crystal nor gate based on the nonlinear kerr effect in a silicon nanocavity*. J. Opt. Soc. Am. B, 30(9):2535–2539, 2013. http://josab.osa.org/abstract.cfm?URI=josab-30-9-2535. (Zitiert auf Seite 2)

[ADK91] Georgios D. Akrivis, Vassilios A. Dougalis, and Ohannes A. Karakashian: *On fully discrete Galerkin methods of second-order temporal accuracy for the nonlinear Schrödinger equation*. Numer. Math., 59(1):31–53, 1991. http://dx.doi.org/10.1007/BF01385769. (Zitiert auf den Seiten 55 und 65)

[AF03] Robert A. Adams and John J. F. Fournier: *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003. (Zitiert auf den Seiten 8 und 9)

[AM89] A. K. Aziz and Peter Monk: *Continuous finite elements in space and time for the heat equation*. Math. Comp., 52(186):255–274, 1989. http://dx.doi.org/10.2307/2008467. (Zitiert auf den Seiten 35, 36, 37, 47 und 49)

[AMK05] Manfred von Ardenne, Gerhard Musiol und Uwe Klemradt (Herausgeber): *Effekte der Physik und ihre Anwendungen*. Harri Deutsch Verlag, Frankfurt am Main, 3. Auflage, 2005. (Zitiert auf Seite 2)

[BJ74] R. Bonnerot and P. Jamet: *A second order finite element method for the one-dimensional Stefan problem*. Int. J. Numer. Meth. Engng., 8(4):811–820, 1974. http://dx.doi.org/10.1002/nme.1620080410. (Zitiert auf Seite 32)

[BL94] L. Bales and I. Lasiecka: *Continuous finite elements in space and time for the nonhomogeneous wave equation*. Comput. Math.

Appl., 27(3):91–102, 1994. http://dx.doi.org/10.1016/0898-1221(94)90048-5. (Zitiert auf den Seiten 19, 36, 37 und 84)

[Bra03] Dietrich Braess: *Finite Elemente*. Springer, Berlin, 3. Auflage, 2003. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie. (Zitiert auf den Seiten 27, 30 und 95)

[BS08] Susanne C. Brenner and L. Ridgway Scott: *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008. http://dx.doi.org/10.1007/978-0-387-75934-0. (Zitiert auf Seite 30)

[Bur98] Victor I. Burenkov: *Sobolev spaces on domains*, volume 137 of *Teubner Texts in Mathematics*. Teubner, Stuttgart, 1998. (Zitiert auf Seite 51)

[CGMSS81] I. Christie, D. F. Griffiths, A. R. Mitchell, and J. M. Sanz-Serna: *Product approximation for nonlinear problems in the finite element method*. IMA J. Numer. Anal., 1(3):253–266, 1981. http://dx.doi.org/10.1093/imanum/1.3.253. (Zitiert auf Seite 41)

[Cia02] Philippe G. Ciarlet: *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. http://dx.doi.org/10.1137/1.9780898719208, Reprint of the 1978 original. (Zitiert auf den Seiten 27, 30, 44, 47 und 48)

[Clé75] Ph. Clément: *Approximation by finite element functions using local regularization*. Rev. Française Automat. Informat. Recherche Opérationnelle Sér. (RAIRO) Analyse Numérique, 9(R-2):77–84, 1975. (Zitiert auf Seite 30)

[DG09] Julien Diaz and Marcus J. Grote: *Energy conserving explicit local time stepping for second-order wave equations*. SIAM J. Sci. Comput., 31(3):1985–2014, 2009. http://dx.doi.org/10.1137/070709414. (Zitiert auf den Seiten 3, 113 und 114)

[DGS13] Willy Dörfler, Hannes Gerner, and Roland Schnaubelt: *Local wellposedness of a quasilinear wave equation*. Eingereicht, 2013. (Zitiert auf Seite 20)

[DLP+11] Willy Dörfler, Armin Lechleiter, Michael Plum, Guido Schneider, and Christian Wieners: *Photonic Crystals: Mathematical Analysis and Numerical Approximation*, volume 42 of *Oberwolfach Seminars*. Birkhäuser/Springer, Basel, 2011. (Zitiert auf den Seiten 1 und 3)

[Dup73] Todd Dupont: *Galerkin methods for first order hyperbolics: an example*. SIAM J. Numer. Anal., 10:890–899, 1973. (Zitiert auf Seite 36)

[EGCB73] J. C. Eilbeck, J. D. Gibbon, P. J. Caudrey, and R. K. Bullough: *Solitons in nonlinear optics. i. a more accurate description of the 2π pulse in self-induced transparency*. J. Phys. A: Math., Nucl. Gen., 6(9):1337–1347, 1973. http://dx.doi.org/10.1088/0305-4470/6/9/009. (Zitiert auf Seite 2)

[Eva98] Lawrence C. Evans: *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998. (Zitiert auf den Seiten 9, 10, 24, 31 und 97)

[Fox12] Mark Fox: *Optische Eigenschaften von Festkörpern*. Oldenbourg, München, 2012. (Zitiert auf Seite 13)

[FP96] Donald A. French and Todd E. Peterson: *A continuous space-time finite element method for the wave equation*. Math. Comp., 65(214):491–506, 1996. http://dx.doi.org/10.1090/S0025-5718-96-00685-0. (Zitiert auf den Seiten 36, 65 und 84)

[GM10] Marcus J. Grote and Teodora Mitkova: *Explicit local time-stepping methods for Maxwell's equations*. J. Comput. Appl. Math., 234(12):3283–3302, 2010. http://dx.doi.org/10.1016/j.cam.2010.04.028. (Zitiert auf den Seiten 3, 113 und 114)

[GR05] Christian Großmann und Hans Görg Roos: *Numerische Behandlung partieller Differentialgleichungen*. Teubner Studienbücher Mathematik. Teubner, 3. Auflage, 2005. (Zitiert auf den Seiten 26, 32, 36 und 65)

[Gra81] Alexander Graham: *Kronecker products and matrix calculus: with applications*. Ellis Horwood Ltd., Chichester, 1981. Ellis Horwood Series in Mathematics and its Applications. (Zitiert auf Seite 119)

[Hac96] Wolfgang Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner Studienbücher Mathematik. Teubner, Stuttgart, 1996. Mit Beispielen und Übungsaufgaben. (Zitiert auf den Seiten 26 und 29)

[HH90] Gregory M. Hulbert and Thomas J. R. Hughes: *Space-time finite element methods for second-order hyperbolic equations*. Comput. Methods Appl. Mech. Engrg., 84(3):327–348, 1990. http://dx.doi.org/10.1016/0045-7825(90)90082-W. (Zitiert auf Seite 32)

[Hul72] Bernie L. Hulme: *One-step piecewise polynomial Galerkin methods for initial value problems*. Math. Comp., 26:415–426, 1972. (Zitiert auf Seite 36)

[Jac75] John David Jackson: *Classical electrodynamics*. John Wiley & Sons Inc., New York, second edition, 1975. (Zitiert auf den Seiten 11 und 14)

[JJWM11] John D. Joannopoulos, Steven G. Johnson, Joshua N. Winn, and Robert D. Meade: *Photonic crystals: molding the flow of light*. Princeton University Press, Princeton, 2011. (Zitiert auf den Seiten 1 und 13)

[JYY+12] Ru Long Jin, Yan Hao Yu, Han Yang, Feng Zhu, Qi Dai Chen, Mao Bin Yi, and Hong Bo Sun: *Electro-optical detection based on large kerr effect in polymer-stabilized liquid crystals*. Opt. Lett., 37(5):842–844, 2012. http://dx.doi.org/10.1364/OL.37.000842. (Zitiert auf Seite 2)

[KA03] Peter Knabner and Lutz Angermann: *Numerical methods for elliptic and parabolic partial differential equations*, volume 44 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2003. (Zitiert auf Seite 46)

[Kat70] Tosio Kato: *Linear evolution equations of "hyperbolic" type*. J. Fac. Sci. Univ. Tokyo Sect. I, 17:241–258, 1970. (Zitiert auf Seite 21)

[Kel03] C. T. Kelley: *Solving nonlinear equations with Newton's method*, volume 1 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003. http://dx.doi.org/10.1137/1.9780898718898. (Zitiert auf Seite 41)

[KM05] Ohannes Karakashian and Charalambos Makridakis: *Convergence of a continuous Galerkin method with mesh modification for nonlinear wave equations*. Math. Comp., 74(249):85–102, 2005. http://dx.doi.org/10.1090/S0025-5718-04-01654-0. (Zitiert auf Seite 65)

[Li11] Haojun Li: *Numerical simulation of a micro-ring resonator with adaptive wavelet collocation method*. Dissertation, Karlsruher Institut für Technologie (KIT), Karlsruhe, 2011. http://digbib.ubka.uni-karlsruhe.de/volltexte/1000024186. (Zitiert auf Seite 135)

[LMY+13] Lai Liu, Xiangwei Meng, Feixiang Yin, Meisong Liao, Dan Zhao, Guanshi Qin, Yasutake Ohishi, and Weiping Qin: *Soliton self-frequency shift controlled by a weak seed laser in tellurite photonic*

crystal fibers. Opt. Lett., 38(15):2851–2854, 2013. http://ol.osa.org/abstract.cfm?URI=ol-38-15-2851. (Zitiert auf Seite 2)

[NM92] Alan C. Newell and Jerome V. Moloney: *Nonlinear optics*. Advanced Topics in the Interdisciplinary Mathematical Sciences. Addison-Wesley Publishing Company Advanced Book Program, Redwood City, CA, 1992. (Zitiert auf Seite 3)

[Paz83] A. Pazy: *Semigroups of linear operators and applications to partial differential equations*, volume 44 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1983. http://dx.doi.org/10.1007/978-1-4612-5561-1. (Zitiert auf Seite 22)

[Pla10] Robert Plato: *Numerische Mathematik kompakt*. Vieweg+Teubner Verlag, Wiesbaden, 4. Auflage, 2010. http://dx.doi.org/10.1007/978-3-8348-9644-5, Grundlagenwissen für Studium und Praxis. (Zitiert auf Seite 10)

[PNTB09] Martin Pototschnig, Jens Niegemann, Lasha Tkeshelashvili, and Kurt Busch: *Time-domain simulations of the nonlinear Maxwell equations using operator-exponential methods*. IEEE Trans. Antennas and Propagation, 57(2):475–483, 2009. http://dx.doi.org/10.1109/TAP.2008.2011181. (Zitiert auf Seite 96)

[Ric10] Markus Richter: *Optimization of Photonic Band Structures*. Dissertation, Karlsruher Institut für Technologie (KIT), Karlsruhe, 2010. http://digbib.ubka.uni-karlsruhe.de/volltexte/1000021317. (Zitiert auf Seite 95)

[Sch13] Philipp Schmalkoke: *On the spectral Properties of Dispersive Photonic Crystals*. Dissertation, Karlsruher Institut für Technologie (KIT), Karlsruhe, 2013. http://digbib.ubka.uni-karlsruhe.de/volltexte/1000034730. (Zitiert auf Seite 12)

[SCM73] Alwyn C. Scott, F. Y. F. Chu, and David W. McLaughlin: *The soliton: a new concept in applied science*. Proc. IEEE, 61:1443–1483, 1973. (Zitiert auf Seite 2)

[Tho97] Vidar Thomée: *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1997. (Zitiert auf den Seiten 32, 35, 36 und 65)

[TM88] Y. Tourigny and J. Ll. Morris: *An investigation into the effect of product approximation in the numerical solution of the cubic*

nonlinear Schrödinger equation. J. Comput. Phys., 76(1):103–130, 1988. http://dx.doi.org/10.1016/0021-9991(88)90133-7. (Zitiert auf Seite 41)

[TN83] Tosiya Taniuti and Katsunobu Nishihara: *Nonlinear waves*, volume 15 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1983. Translated from the Japanese by Taniuti and Alan Jeffrey. (Zitiert auf Seite 3)

[Tom80] W. J. Tomlinson: *Surface wave at a nonlinear interface*. Opt. Lett., 5(7):323–325, 1980. http://ol.osa.org/abstract.cfm?URI=ol-5-7-323. (Zitiert auf Seite 2)

[VF80] Erol Varoğlu and W. D. Liam Finn: *Space-time finite elements incorporating characteristics for the Burgers equation*. Internat. J. Numer. Methods Engrg., 16(Special Issue):171–184, 1980. http://dx.doi.org/10.1002/nme.1620160112. (Zitiert auf Seite 32)

[Yse86] Harry Yserentant: *On the multilevel splitting of finite element spaces*. Numer. Math., 49(4):379–412, 1986. http://dx.doi.org/10.1007/BF01389538. (Zitiert auf Seite 139)

[Zei90] Eberhard Zeidler: *Nonlinear functional analysis and its applications. II/A*. Springer-Verlag, New York, 1990. http://dx.doi.org/10.1007/978-1-4612-0985-0, Linear monotone operators, Translated from the German by the author and Leo F. Boron. (Zitiert auf Seite 9)

Index