Adaptive Finite-Element-Ausgleichsformulierungen für parabolische Anfangs-Randwertaufgaben

Vom Fachbereich Mathematik der Universität Hannover
zur Erlangung des Grades

Doktor der Naturwissenschaften

Dr. rer. nat.

genehmigte Dissertation
von

Dipl.-Math. Mohammad Majidi

geboren am 16. August 1975 in Rasht (Iran)

2002

Referent: Prof. Dr. Gerhard Starke

Korreferent: Prof. Dr. Alexander Ostermann

Tag der Promotion: 03.05.2002

Meiner Frau
und
meinem Sohn

Danksagungen

An dieser Stelle möchte ich mich insbesondere bei Herrn Prof. Dr. Gerhard Starke dafür bedanken, daß er mir durch seine Hinweise und Vorschläge sehr geholfen hat. Herr Prof. Dr. Starke hat in meiner mathematischen Entwicklung eine wichtige Rolle gespielt. Einerseits war er ein großartiger Lehrer, andererseits war er für mich auch ein großer Motivator, als ich an manchen Stellen nicht weiter kam. Bei Entstehung dieser Arbeit hat er mit mir sehr viel Geduld gehabt. Für diese einzigartige Leistung sei ihm gedankt.

Ferner möchte ich Herrn Prof. Dr. Ostermann für seine Ratschläge, die zur Entstehung dieser Arbeit sehr viel beigetragen haben, danken.

Als nächtes möchte ich Frau Dr. Haschke, Herrn Dipl. Math. Geilenkothen und Herrn Dr. Korsawe danken, die diese Arbeit Korrektur gelesen haben und mir sehr sinnvolle Ratschläge zur Ergänzung gaben.

Mein spezieller Dank geht an all die Menschen, die mich auf meinem mathematischen Weg unterstützt haben.

Weiterhin möchte ich mich bei meiner Familie, insbesondere bei meiner Mutter bedanken, die mir immer mit Rat und Tat zu Seite stand.

Meiner Frau möchte ich danken, die beim Entstehen dieser Arbeit mit mir sehr viel Geduld bewies.

Zum Schluß geht mein Dank an den Allmächtigen, der mir die Kraft und das Selbstvertrauen gegeben hat, diese Arbeit zu schreiben. Ich möchte ihm dafür danken, daß er mir in jeder Lage meines Lebens geholfen hat und helfen wird.

Mai 2002 Mohammad Majidi

Zusammenfassung

Diese Arbeit ist eine Erweiterung der Finite-Element-Ausgleichsformulierungsmethode für lineare parabolische Anfangs-Randwertprobleme aus den Arbeiten [33, 34] auf semilineare parabolische Probleme. Die Grundidee dieser Methode ist die Minimierung eines Least-Squares-Funktionals (LSF) für ein System erster Ordnung in diskreten Räumen. Es sind also Ausgleichsprobleme, die mit Finite-Element-Methoden gelöst werden. In dieser Dissertation werden wir ein Einschrittverfahren zur Zeit-Diskretisierung von semilinearen parabolischen Problemen herleiten und analysieren. Einer der wichtigsten Vorteile dieses Ansatzes ist, daß das LSF als ein a-posteriori Fehlerschätzer benutzt werden kann. Wir werden dann in dieser Arbeit (bzgl. der Zeitvariablen) stückweise lineare (nicht notwendig stetige) Funktionen für die Approximation des Flusses mit (bzgl. der Zeitvariablen) stückweise linearen stetigen Funktionen für die Approximation der skalaren Unbekannte kombinieren.

Stichworte: a-posteriori Fehlerschätzer, adaptive Zeitschritt-Kontrolle, semilineare parabolische Probleme

Abstract

This thesis generalizes the least-squares Galerkin methods for linear parabolic initial-boundary value problems of the papers [33, 34] to semilinear parabolic problems. These methods are based on the minimization of a least-squares functional for an equivalent first-order system over space and time with respect to suitable discrete spaces. This thesis will present the derivation and analysis of one-step methods for semi-discretization in time from least-squares principles for semilinear parabolic problems. One of the most important features of the least-squares methodology is the built-in a posteriori estimate for the approximation error. For the presentation here, we focus our attention on the specific combination of piecewise linear, not necessarily continuous, functions with continuous piecewise linears for the flux and scalar variable, respectively.

Keywords: a posteriori error estimators, adaptive timestep control, semilinear parabolic problems

Inhaltsverzeichnis

Kapitel 0

Einleitung

Viele Vorgänge in der Natur werden mittels Differentialgleichungen modelliert. Darum sind viele Modelle im Bereich der Biologie, Chemie, Mechanik, Medizin und Physik durch parabolische Gleichungen beschrieben. Einige dieser Modelle werden durch lineare Differentialgleichungen repräsentiert. Die meisten der Modelle werden jedoch mittels Gleichungen nichtlinearer Natur beschrieben. In der Mathematik ist das Interesse an der Analyse der Eigenschaften dieser Modelle stark gestiegen, die Existenz, Eindeutigkeit und Regularität ihrer Lösung beinhaltet (vgl. z.B. [3, 4, 45, 50]). Auch wenn die Theorie der Differentialgleichungen eine eindeutige Lösung eines Modells garantiert, so läßt sich diese im Allgemeinen nicht geschlossen angeben. In der numerischen Mathematik entwickelt man deshalb Algorithmen, welche die Lösungen solcher Differentialgleichungen approximieren. Das Hauptproblem dabei ist, daß sehr wenig über die exakte Lösung bekannt ist. Dabei ist Adaptivität wegen der Beschränktheit der Computer-Ressourcen sehr gefragt. Um adaptive Algorithmen entwickeln zu können und um die Genauigkeit zu erhöhen, benötigt man a-posteriori Fehlerschätzer, welche die Qualität berechneter numerischer Approximationen beurteilen.
Es erweist sich als notwendig bei der Behandlung parabolischer Differentialgleichungen, sowohl in der Theorie, als auch in der Praxis, die Zeitkomponente unterschiedlich von der Ortskomponente zu behandeln. In der numerischen Analysis der parabolischen Differentialgleichungen unterscheidet man zwischen zwei verschiedenen Methoden zur Approximation der Lösung:

- Einerseits hat man die Möglichkeit zuerst im Ort zu diskretisieren. Damit erhält man ein System von gewöhnlichen Differentialgleichungen, das sehr effizient gelöst werden kann. Diese Vorgehensweise ist als *Linienmethode* bekannt (vgl. z.B. [46]).
- Andererseits kann man zuerst bzgl. der Zeit diskretisieren. Man wählt ein *Gitter* bestehend aus endlich vielen Zeitpunkten $\Delta = \{0, t_1, \ldots, t_M\} \subset [0, T]$. An jedem Zeitpunkt ist ausgehend von den Approximationen der letzten Zeitschritte eine bestimmte elliptische Differentialgleichung zu lösen. Dieses Verfahren ist als *Rothe-Methode* bekannt.

In dieser Arbeit betrachten wir nur die zweite Methode, die sich für unsere Zwecke als praktisch erwiesen hat (vgl. [32]). Die Rothe-Methode hat unter anderem die folgende Eigenschaft: Man formuliert die parabolische Differentialgleichung als ein abstraktes Cauchy-Problem in einem Hilbertraum H, also etwa

$$\partial_t p + F(t, p) = 0, \;\; p(0) = P_0, \;\; 0 < t \leq T, \tag{1}$$

wobei $p(t) \in H$, für $0 \leq t \leq T$. Die Gleichung (1) hat formal die Form einer gewöhnlichen Differentialgleichung. Es stellt sich tatsächlich heraus, daß sich viele Ergebnisse aus dem Bereich der Systeme gewöhnlicher Differentialgleichungen auch auf (1) verallgemeinern lassen. Der Grund dafür ist, daß diese Ergebnisse aus der Vollständigkeit von $\mathbb{R}^n$ folgen. Diese sehr praktische Vorgehensweise wird in vielen Werken benutzt (vgl. für den linearen Fall z.B. [9, 10] und für den nichtlinearen Fall [37]). Es stellt sich heraus, daß es wenig Sinn macht, alle Zeitpunkte am Anfang festzulegen, weil man a priori nicht weiß, wie groß diese sinnvoll gewählt werden sollten. Damit man einerseits nicht unnötigen Aufwand betreiben muß und andererseits möglichst gute Approximationen berechnen kann, ist der Einsatz von a-posteriori Fehlerschätzern in der Zeit unentbehrlich. Die Effizienz eines Fehlerschätzers soll den betriebenen Aufwand optimieren, während seine Zuverlässigkeit die Berechnung einer akzeptablen Approximation der Lösung garantiert (vgl. [9, 31]). Mit Hilfe eines a-posteriori Fehlerschätzers ist es also möglich, optimale Zeitschrittweiten zur Approximation von parabolischen Differentialgleichungen zu wählen. Auf der anderen Seite ist es klar, daß die optimalen Zeitschrittweiten wertlos sind, solange wir keinen Algorithmus haben, der in jedem Zeitschritt möglichst gut die anstehende elliptische Differentialgleichung löst. Bei der Entwicklung dieses Algorithmus kann die numerische Methode praktisch *vergessen*, woher das elliptische Problem kommt, und in jedem Zeitschritt nur die zugrundeliegende elliptische Differentialgleichung numerisch lösen. In diesem Zusammenhang hat die Finite-Elemente-Methode (FEM) in jüngster Zeit sehr viel an Popularität gewonnen (vgl. [11, 12]).
Die FEM basiert auf der Idee, abstrakte (unendlichdimensionale) Räume durch (endlichdimensionale) FE-Räume zu ersetzen, um eine numerische Approximation der Lösung der elliptischen Differentialgleichung zu berechnen.
Bei der FEM ist Adaptivität in einfacher Weise zu erreichen: Es genügt nämlich einen a-posteriori Fehlerschätzer im Ort zu konstruieren, der auf einzelnen Elementen der Zerlegung auswertbar ist. Hiermit kann man entscheiden, wie man eine Triangulierung verfeinern soll, um eine bessere Approximation der elliptischen Differentialgleichungen in jedem Zeitschritt zu bekommen (für Standard-Elemente vgl. [48]). Wir verdeutlichen dies anhand der linearen elliptischen Differentialgleichung:

$$p - \Delta p + f = 0, \tag{2}$$

Dies ist eine typische lineare elliptische Differentialgleichung, die (in etwa) in jedem Zeitschritt einer linearen *Wärmeleitungsgleichung* (vgl. Kapitel 2) gelöst werden muß. Ist man dabei nur an einer Approximation von p interessiert, so geht man wie in [9, 18] vor und damit läßt sich p sehr effizient approximieren. Es sind für diese Art von Problemen sehr viele Arbeiten erschienen, die sogar in einigen nichtlinearen Fällen sehr effiziente Fehlerschätzer in der Zeit und im Ort anbieten (vgl. z.B. [18, 19, 20, 21, 22]). Für viele Modellierungen hat jedoch die vektorwertige Funktion $u = -\nabla p$ eine physikalische Interpretation und ihre Approximation ist genauso wichtig wie die Approximation an p. In der Wärmeleitungsgleichung bezeichnet u den Wärmefluß.
Das Ziel dieser Arbeit ist die Entwicklung einer adaptiven Methode zur Approximation der parabolischen Differentialgleichung - bekannt als *Richardsgleichung* - aus dem Gebiet der Bodenmechanik

$$\partial_t \theta(p) - \ \text{div}\ (K(\theta(p))\nabla p) = 0, \ \ p(0) = P_0. \tag{3}$$

Ferner ist $u = -K(\theta(p))\nabla p$ zu approximieren. Es empfiehlt sich dann, statt Gleichung (3) das parabolische System

$$\begin{aligned} \partial_t \theta(p) + \ \text{div}\ u &= 0, \\ u + K(\theta(p))\nabla p &= 0, \\ p(0) &= P_0 \end{aligned} \tag{4}$$

numerisch zu lösen (für die genaue Herleitung der Gleichungen und die Bedeutung der Variablen vgl. Abschnitt 4.1). Es gibt viele Arbeiten, die sich sowohl mit der Existenz, Eindeutigkeit und Regularität der Lösung des Systems (4), als auch mit der Approximation dieser Lösung beschäftigen (vgl. z.B. [3, 52] und [44, 43]). Dabei steht man immer vor dem Problem, die Zeitschritte optimal zu wählen.
Es gibt zwei verschiedene Wege das System (4) numerisch zu approximieren:

- Man könnte einerseits (4) durch die gemischte Methode von Raviart und Thomas (vgl. [40, 24]) approximieren.
- Auf der anderen Seite kommt die Least-Squares-Methode (LSM) zur Approximation von (4) in Frage, die auf der Minimierung eines Ausgleichsfunktionals in FE-Räumen beruht (vgl. [8]).

Wir gehen in dieser Arbeit den zweiten Weg, da die erste Methode einige Nachteile besitzt. So benötigt man für die Variationsfomulierung die sogenannte inf-sup-Bedingung (vgl. z.B. [11, Section III.4]), welche die Wahl der Ansatzräume zur Approximation der Lösung einschränkt. Weiter werden die zu lösenden LGS indefinit. Für diese Probleme kommen also einfache iterative Verfahren - wie das CG-Verfahren - nicht in Frage (vgl. [11, Section IV.5]). Im Gegensatz dazu benötigt die LSM keine inf-sup-Bedingung und die anstehenden LGS sind positiv definit, so daß der Einsatz von Multilevel-Verfahren den optimalen Aufwand von $\mathcal{O}(N_h)$ liefert, wobei N_h die Anzahl der Unbekannten aus dem entsprechenden LGS bezeichnet. In [29, 44, 43] wird das System (4) durch die LSM approximiert, wobei das Least-Squares-Funktional (LSF) als ein Fehlerschätzer (im Ort) benutzt wird. Diese Vorgehensweise hat die Vorteile,

- daß man keine Zusatzbedingungen wie die *Saturierungsbedingung* (wie z.B. beim hierarchischen Fehlerschätzer) benötigt und
- keinen Aufwand zur Berechnung des Fehlerschätzers braucht, da man das LSF sowieso berechnen muß.

Aus einer ganz anderen Perspektive wird die LSM in [33, 34] behandelt. Dort wurde eine Methode zur Approximation von (2) vorgestellt, die den LS-Ansatz dazu benutzt, ein LSF zu definieren, das als Fehlerschätzer in der Zeit benutzt wird. Diese Methode wird LS-Formulierung in der Zeit genannt.
Die wesentlichen Ergebnisse der vorliegenden Arbeit lassen sich wie folgt darstellen:
Die in [33, 34] beschriebene LS-Formulierung in der Zeit wird auf semilineare parabolische Gleichungen der Form

$$\begin{aligned} \partial_t p + \ \text{div}\ u + f(p) &= 0 \\ u + \nabla p &= 0 \end{aligned}$$

verallgemeinert. Die wesentliche Modifikation besteht hierbei in der Abhängigkeit der rechten Seite f von p.
Ersetzt man (u, p) durch die bzgl. der Zeit diskrete Funktion (u_τ, p_τ) (z.B. stückweise linear bzgl. einer Zerlegung Δ des Intervalls $[0, T]$), so ist das zugehörige Least-Squares-Funktional $\mathcal{F}(u_\tau, p_\tau)$ immer noch nichtlinear von p_τ abhängig. Diese Tatsache verhindert, daß man $\mathcal{F}(u_\tau, p_\tau)$ durch eine geeignete Quadraturformel wie in [33] darstellen kann. Damit ist es im semilinearen Fall sehr viel aufwendiger, Konvergenz der LSM zu zeigen.
Für den allgemeinen semilinearen Fall (f nicht autonom) wird ein Konvergenzresultat bewiesen, das auf einigen Annahmen, wie der Äquivalenz von $\mathcal{F}(u_\tau, p_\tau)$ mit dem Fehler beruht (vgl. Satz 3.5). Ist andererseits f autonom, so läßt sich das Konvergenzresultat ohne diese Annahmen beweisen (vgl. Satz 3.7). Die Grundidee der Beweise von Satz 3.5 bzw. Satz 3.7 läßt sich auch in [37] finden.
Die genannten Beweise werden in drei Schritten durchgeführt (hier für die Konvergenzordnung 2 und gleichbleibende Schrittweite τ):

1. Man schätzt den Defekt durch $C\tau^3$ ab.
2. Man schätzt den gesamten Fehler $\|e_n\|$ im n-ten Schritt durch $C(\tau^2 + \tau \sum_{j=1}^{n-1} \|e_n\|)$ ab. Dazu ist die lokale Lipschitzstetigkeit der rechten Seite f notwendig.
3. Das Lemma von Gronwall (Lemma 3.4) liefert das endgültige Resultat.

Im linearen Fall kann die Konvergenzordnung 3 gezeigt werden (vgl. Bemerkung 2.9.(a)), im semilinearen Fall ist dagegen nur die Konvergenzordnung 2 zu erreichen. Es ist also eine Ordnungsreduktion in Vergleich zum linearen Fall zu beobachten. Numerische Experimente bestätigen das Konvergenzresultat.
Danach wird die genannte LS-Formulierung auf das System (4) erweitert und einige numerische Experimente durchgeführt.
Es ist zu erwähnen, daß die neue Methode, angewandt auf (4), eine Konvergenzordnung (welche analytisch nicht bewiesen wird) in der Zeit besitzt. Die genannten Resultate werden in der Arbeit wie folgt dargestellt:
Im Kapitel 1 werden einige Grundlagen aus der Lösung- und Approximationstheorie der partiellen Differentialgleichungen wiederholt. Dabei beschäftigen wir uns mit der Variationsformulierung und Diskretisierung. Danach werden wir als Beispiele das diskontinuierliche Galerkin-Verfahren und die Runge-Kutta-Methode vorstellen.
In Kapitel 2 werden die LS-Formulierung in der Zeit für den linearen Fall wiederholt, sowie einige numerische Simulationen durchgeführt. In Kapitel 3 wird dieser Ansatz auf den semilinearen Fall erweitert und die B-Stabilität sowie eine Konvergenzaussage gezeigt. An dieser Konvergenzaussage ist bereits zu erkennen, daß eine Ordnungsreduktion zum linearen Fall stattfindet. Anschließend bestätigen numerische Berechnungen die analytischen Beweise. Im vierten und letzten Kapitel werden zunächst die Gleichungen aus der Bodenmechanik hergeleitet. Danach wird die LS-Formulierung in der Zeit auch auf diese Gleichungen verallgemeinert. Numerische Experimente runden das Kapitel ab.

Kapitel 1

Approximationstheorie parabolischer Differentialgleichungen

Die numerische Lösung der parabolischen Differentialgleichungen läßt sich in mehrere Teilprobleme untergliedern: Die Variationsformulierung und Diskretisierung in der Zeit (kurz Zeitdiskretisierung genannt) sowie die Variationsformulierung und Diskretisierung der elliptischen Teilprobleme (kurz Ortsdiskretisierung genannt).
In diesem Kapitel werden wir die Variationsformulierung und Diskretisierung der parabolischen Differentialgleichungen herleiten. Wir werden zunächst die Diskretisierung abstrakt einführen, sie aber anhand eines Beispiels (Diskontinuierliches Galerkin-Verfahren) verdeutlichen. Anschließend folgt die Übertragung der Runge-Kutta-Methode auf parabolische Differentialgleichungen durch die Formulierung des Problems als *abstraktes Cauchy-Problem*.
Es sei $\Omega \subset \mathbb{R}^2$ ein beschränktes, mit Polygonzügen berandetes Gebiet, $T > 0$ und $f, P_0 \in L^2(\Omega)$. Außerdem sei für $x \in \Omega$

$$A(x, \partial) = a_0(x) - \sum_{i=1}^{2} \partial_i (a_i(x) \partial_i) \tag{1.1}$$

ein Differentialoperator der Ordnung 2, wobei $a_i(x) \geq \alpha > 0$, $i = 1, 2$ und $a_0(x) \geq 0$ in Ω.

1.1 Elliptische Differentialgleichungen

1.1.1 Variationsformulierung

Gegeben sei die elliptische Differentialgleichung

$$A(x, \partial) p = f, \tag{1.2}$$

wobei $p\big|_{\partial\Omega} = 0$ und $A(x, \partial)$ der Operator aus (1.1).
Ist p eine starke Lösung des Problems (1.2), d.h. gilt $p \in H_0^1(\Omega) \cap H^2(\Omega)$ und p erfüllt

(1.2), so gilt für alle $\varphi \in C_0^\infty(\Omega)$ nach dem Gaußschen Satz:

$$\begin{aligned}(Ap, \varphi)_{0,\Omega} &= (a_0 p, \varphi)_{0,\Omega} - \sum_{i=1}^{2} (\partial_i (a_i \partial_i) p, \varphi)_{0,\Omega} \\ &= (a_0 p, \varphi)_{0,\Omega} + \sum_{i=1}^{2} (a_i \partial_i p, \partial_i \varphi)_{0,\Omega} .\end{aligned}$$

Satz 1.1 Es sei für $p, q \in H_0^1(\Omega)$

$$a(p,q) = (a_0 p, q)_{0,\Omega} + \sum_{i=1}^{2} (a_i \partial_i p, \partial_i q)_{0,\Omega} ,$$

wobei a die aus dem Operator $A(x, \partial)$ (vgl. (1.1)) durch den Gaußschen Satz gewonnene Bilinearform ist. Gilt $a_i \in L^\infty(\Omega)$, $i = 0, 1, 2$, so ist a eine stetige, $H_0^1(\Omega)$-elliptische, symmetrische Bilinearform.

Beweis:

Vgl. z.B. [11, Section II.2].

■

Wir schwächen die Forderung nach einer starken Lösung mit der folgenden Variationsformulierung ab: Gesucht ist $p \in H_0^1(\Omega)$ mit

$$a(p,q) = (f,q)_{0,\Omega} \qquad \forall \ q \in H_0^1(\Omega). \tag{1.3}$$

Diese Lösung nennen wir *schwache Lösung* von (1.2). Für $p \in H^2(\Omega)$ sind die beiden Aussagen: „p ist eine starke Lösung“ und „p ist eine schwache Lösung“ natürlich äquivalent. Der Satz von Lax-Milgram (vgl. [12, (2.7.7) Theorem]) garantiert die Existenz und Eindeutigkeit schwacher Lösungen. Es gibt schwache Lösungen $p \in H_0^1(\Omega)$, die keine starken Lösungen sind.

1.1.2 Diskretisierung

Der Satz von Lax-Milgram garantiert nicht nur die Existenz und Eindeutigkeit der schwachen Lösung, er liefert auch die Grundidee für die Diskretisierung. Setzt man nämlich $H = Q_h$, wobei $Q_h \subset H_0^1(\Omega)$ ein endlichdimensionaler Unterraum von $H_0^1(\Omega)$ ist, so ist Q_h wieder ein Hilbertraum und $a_{\big|_{Q_h \times Q_h}}$ ist stetig, symmetrisch und Q_h-elliptisch, so daß die Variationsformulierung:
Finde $p_h \in Q_h$ mit

$$a(p_h, q_h) = (f, q_h)_{0,\Omega} \qquad \forall \ q_h \in Q_h \tag{1.4}$$

eine eindeutige Lösung besitzt, welche *Galerkin-Approximation* heißt.
Der folgende Satz macht über die Qualität der berechneten Galerkin-Approximation eine wichtige Aussage.

Satz 1.2 (Céa - Lemma) Die symmetrische Bilinearform $a(\cdot,\cdot)$ sei stetig und H-elliptisch (mit den Konstanten C_s und C_e). Dann gilt für die Lösungen p und p_h des Variationsproblems (1.3) in H bzw. (1.4) in $Q_h \subset H$:

$$\|p - p_h\|_H \leq \left(\frac{C_s}{C_e}\right)^{1/2} \inf_{q_h \in Q_h} \|p - q_h\|_H \,.$$

Beweis:
Vgl. z.B. [12, (2.8.1) Theorem].

■

Das Céa - Lemma zeigt, daß p_h *quasi-optimal* bzgl. der Minimierung des Fehlers $\|p - p_h\|_H$ ist, d.h. dieser Ausdruck wird bis auf eine Konstante minimiert. Die Genauigkeit der Galerkin-Näherung hängt also nach dem Céa - Lemma davon ab, wie gut die Lösung $p \in H$ im Funktionenraum Q_h approximiert werden kann.

1.2 Parabolische Differentialgleichungen

Gegeben sei die folgende parabolische Differentialgleichung:

$$\begin{aligned} \partial_t p(t,x) + A(x,\partial)p(t,x) &= f(t,x), \qquad (t,x) \in (0,T] \times \Omega, \\ p(t,\cdot)\big|_{\partial\Omega} &= 0, \qquad \forall\ t \in (0,T], \\ p(0,\cdot) &= P_0. \end{aligned} \tag{1.5}$$

wobei $A(x,\partial)$ der Differentialoperator aus (1.1) ist. Ferner sei $f \in L^2((0,T); L^2(\Omega))$ und $P_0 \in L^2(\Omega)$. Im Folgenden soll nun eine Variationsformulierung für dieses Problem gefunden werden.

1.2.1 Variationsformulierung

Wir bilden gemäß Abschnitt 1.1.1 die zu $A(x,\partial)$ gehörige Bilinearform a, die laut Satz 1.1 $H_0^1(\Omega)$-elliptisch und stetig in $H_0^1(\Omega)$ ist.
Sei zunächst $H = H_0^1(\Omega)$ und p eine starke Lösung der Differentialgleichung, d.h. $p \in C^1((0,T]; H_0^1(\Omega) \cap H^2(\Omega)) \cap C([0,T]; L^2(\Omega)) \subset L^2((0,T); H)$ und p erfülle (1.5). Für alle $\theta \in L^2((0,T); H)$ gilt dann:

$$(\partial_t p, \theta)_{L^2((0,T);H)} + \int_0^T a((p(t), \theta(t))dt = (f, \theta)_{L^2((0,T);H)}$$

Bedenkt man die Tatsache, daß

$$C_0^\infty([0,T)) \otimes H \subset L^2((0,T); H),$$

so haben wir

$$\begin{aligned} (\int_0^T \partial_t p(t)\ \varphi(t)\ dt, q)_{0,\Omega} + a(\int_0^T p(t)\ \varphi(t)dt, q) \\ = (\int_0^T f(t)\ \varphi(t)\ dt, q)_{0,\Omega} \quad \forall\ q \in H, \varphi \in C_0^\infty([0,T)). \end{aligned}$$

Partielle Integration, $\varphi(T) = 0$ und $p(0) = P_0$ liefern:

$$\begin{aligned}(-\int_0^T p(t)\,\partial_t\,\varphi(t)\,dt, q)_{0,\Omega} + a(\int_0^T p(t)\,\varphi(t)dt, q)\\ = (P_0\,\varphi(0), q)_{0,\Omega} + (\int_0^T f(t)\,\varphi(t)\,dt, q)_{0,\Omega} \quad \forall\ q \in H, \varphi \in C_0^\infty([0,T)).\end{aligned} \tag{1.6}$$

Sei umgekehrt $p \in C^1((0,T]; H_0^1(\Omega) \cap H^2(\Omega)) \cap C\left([0,T]\,; L^2\left(\Omega\right)\right)$ und p erfülle die Bedingung (1.6). Dann erhalten wir durch partielle Integration:

$$-\int_0^T p(t)\,\partial_t\,\varphi(t)\,dt = \int_0^T \partial_t\,p(t)\,\varphi(t)\,dt + p(0)\,\varphi(0) - p(T)\,\underbrace{\varphi(T)}_{=0}. \tag{1.7}$$

Setzen wir (1.7) in (1.6) ein, so erhalten wir

$$\begin{aligned}(\int_0^T \partial_t p(t)\varphi(t)\,dt, q)_{0,\Omega} + a(\int_0^T p(t)\,\varphi(t)dt, q) + (p(0)\varphi(0), q)_{0,\Omega}\\ = (P_0\varphi(0), q)_{0,\Omega} + (\int_0^T f(t)\,\varphi(t)\,dt, q)_{0,\Omega}, \quad \forall\ q \in H, \varphi \in C_0^\infty([0,T)).\end{aligned} \tag{1.8}$$

Da $C_0^\infty([0,T))$ dicht in $L^2(0,T)$ enthalten ist, können wir in (1.8) $\varphi^{(n)} = \left(1 - \frac{t}{T}\right)^n$ für $n \in \mathbb{N}$ setzen. Aus $\lim\limits_{n\to\infty} \int\limits_0^T (\varphi^{(n)}(t))^2\,dt = 0$ folgern wir mit der Cauchy-Schwarzschen Ungleichung für die drei Bochner-Integrale (vgl. Abschnitt A.3)

$$\lim_{n\to\infty} \int_0^T \partial_t p(t)\,\varphi^{(n)}(t)\,dt = 0,$$

$$\lim_{n\to\infty} \int_0^T p(t)\,\varphi^{(n)}(t)\,dt = 0$$

und

$$\lim_{n\to\infty} \int_0^T f(t)\,\varphi^{(n)}(t)\,dt = 0.$$

Schließlich folgt aus $\lim\limits_{n\to\infty} \varphi^{(n)}(0) = 1$:

$$(p(0), q)_{0,\Omega} = (P_0, q)_{0,\Omega}, \qquad \forall\ q \in H.$$

Wegen der Dichtheit von H in $L^2\left(\Omega\right)$ gilt $p(0) = P_0$.
Aus der Tatsache, daß $C_0^\infty((0,T)) \otimes H \underset{\text{dicht}}{\subset} L^2((0,T); H)$ gilt, erhalten wir aus (1.6) schließlich:

$$\partial_t p + A(x,\partial)p = f$$

in $L^2((0,T); H)$. Damit ist p auch eine starke Lösung von (1.5). Mit der folgenden Variationsformulierung wollen wir den Begriff der Lösung abschwächen: Bestimme zu vorgegebenem $f \in L^2\left((0,T); L^2\left(\Omega\right)\right)$ und $P_0 \in L^2\left(\Omega\right)$ ein $p \in L^2\left((0,T); H\right) \cap C\left([0,T]\,; L^2\left(\Omega\right)\right)$, so daß (1.6) erfüllt ist. Die Lösung dieser Variationsformulierung heißt schwache Lösung von (1.5). Die Existenz und Eindeutigkeit der schwachen Lösung wird in [39, Theorem 11.1.1] bewiesen.

1.2.2 Diskretisierung

Bei der Behandlung von parabolischen Differentialgleichungen sollte in der Theorie sowie auch bei der numerischen Approximation die Zeitkomponente gesondert behandelt werden. Dies führt dazu, die Zeitvariable separat von der Ortsvariablen zu diskretisieren.
Dazu entwickelt man die Theorie für den „halbdiskreten“ Fall. Dann vollzieht man die zweite Diskretisierung (innere Diskretisierung), und diesen „volldiskreten“ Fall betrachtet man als eine Störung des halbdiskreten Falls. Solange man feste Gitter für den Ort verwendet und von vornherein feste Zeitschrittweiten benutzt, spielt die Reihenfolge der Diskretisierungen keine Rolle. Aber für addaptiv-verfeinerte Gitter oder mögliche Zeitschrittvariationen und adaptive Zeitschrittweiten ist die Reihenfolge in der Tat von Belang.
Da wir in dieser Arbeit Adaptivität in der Zeit und im Ort haben wollen, benutzen wir die Rothe-Methode. Das heißt, es wird zuerst bzgl. der Zeit diskretisiert.
Die vorkommenden elliptischen Subprobleme könnten durch Multilevel-Methoden (vgl. [11]) gelöst werden. Man beachte, daß die hier entwickelte Theorie nicht nur für lineare parabolische Differentialgleichungen vom Typ (1.5) anwendbar ist, sondern allgemein für alle Typen von parabolischen Differentialgleichungen benutzt werden kann. Obwohl wir im folgenden übersichtshalber die Bezeichnungen von (1.5) benutzen, ist eine Verallgemeinerung auf andere (nicht notwendig lineare) parabolische Differentialgleichungen einfach.
Noch ausführlicher gehen wir in den beiden Abschnitten 2.1 bzw. 2.2 im Falle der Wärmeleitungsgleichung (2.1) auf die angesprochenen Themen ein.

Der semidiskrete Fall

Wir diskretisieren zunächst die parabolische Differentialgleichung in der Zeit.

Man beachte, daß man für eine beliebige parabolische Differentialgleichung, für die eine eindeutige Lösung existiert, den *Evolutionsoperator* E für $t \in [0,T)$ und $\tau > 0$ (mit $t+\tau \leq T$) als

$$E(t+\tau;t)p(t) = p(t+\tau) \tag{1.9}$$

definieren kann. Der Evolutionsoperator ist also eine Abbildung von dem Raum, in dem $p(t)$ enthalten ist, in den Raum, der $p(t+\tau)$ enthält. In dieser Arbeit sind die beiden Räume, die $p(t)$ bzw. $p(t+\tau)$ enthalten, stets Teilräume von $H^1(\Omega)$. Für spezielle parabolische Probleme kann man den Evolutionsoperator genauer angeben (vgl. Bemerkung B.4.c,d) und (2.8)).
Wir unterteilen das kontinuierliche Intervall $[0,T]$ durch $M+1$ ausgewählte *diskrete* Zeitpunkte

$$0 = t_0 < t_1 < \ldots < t_M = T.$$

Diese diskreten Zeitpunkte bilden ein *Gitter* $\Delta = \{0 = t_0 < t_1 < \ldots < t_M = T\}$ auf $[0,T]$ und heißen daher *Gitterpunkte*. Die Anzahl der Gitterpunkte hängt offenbar von der Wahl des Gitters ab, weshalb wir in Zukunft M_Δ statt M schreiben wollen, sofern wir diese Abhängigkeit betonen wollen. Ferner bezeichnen wir mit

$$\tau_j = t_j - t_{j-1}, \qquad j = 1, \ldots, M_\Delta$$

die *Zeitschrittweiten* von einem Gitterpunkt zum nächsten, die unter ihnen maximale Zeitschrittweite bezeichnen wir mit

$$\tau_\Delta = \max_{1\le j\le M_\Delta} \tau_j \,.$$

Als nächstes wollen wir für die numerische Approximation der Lösung p einer Anfangs-Randwertaufgabe eine *Gitterfunktion*

$$p_\Delta : \Delta \to H^1(\Omega)$$

angeben, die p an den Gitterpunkten (möglichst gut) approximiert,

$$p_\Delta(t) \approx p(t) \qquad \forall\ t \in \Delta.$$

Solche Verfahren zur Approximation an diskreten Stellen heißen auch *Diskretisierungsverfahren* oder kurz *Diskretisierungen*.
Außerdem soll die Funktion p_Δ rekursiv ermittelbar sein,

$$p_\Delta(0) = P_0 \mapsto p_\Delta(t_1) \mapsto \ldots \mapsto p_\Delta(t_\Delta) = p_\Delta(T),$$

wobei bei *Einschrittverfahren* der hier enthaltene Rechenvorgang für alle Gitter Δ einheitlich durch eine Zweiterm-Rekursion beschrieben wird:

$$\begin{aligned} &\text{(i) } p_\Delta(0) = P_0, \\ &\text{(ii) } p_\Delta(t_{j+1}) = E_d(t_{j+1}; t_j) p_\Delta(t_j) \quad \text{für} \quad j = 0, 1, \ldots, M_\Delta - 1, \end{aligned} \tag{1.10}$$

mit einer von Δ unabhängigen Funktion E_d. Mit Blick auf den Evolutionsoperator E in (1.9) nennen wir E_d einen *diskreten Evolutionsoperator*. Ein Einschrittverfahren ordnet jeder Differentialgleichung, repräsentiert durch ihre rechte Seite f, eine diskrete Evolution zu:

$$f \mapsto E_d = E_d[f].$$

In dieser Arbeit wird in Kapitel 2 eine Methode vorgestellt, mit deren Hilfe es möglich ist, die Funktion p_Δ zu einer Funktion

$$p_\tau : [0, T] \to H^1(\Omega)$$

zu erweitern, so daß der Fehler $p_\tau - p$, gemessen in einer passenden Norm auf dem Lösungsraum der Anfangs-Randwertaufgabe, für entsprechenden Aufwand beliebig klein gehalten wird.
Nun kommen wir zur Diskretisierung von (1.6):

Ersetzen wir in der Variationsformulierung (1.6) den Raum $C_0^\infty([0, T))$ durch einen endlichdimensionalen Unterraum $Q_\Delta \subset L^2((0, T))$ der Dimension N_Δ und der Basis $\{\Phi_j\}_{j=1}^{N_\Delta}$, so muß die Variationsgleichung (1.6) wegen der Linearität nur für diese Basisfunktionen gelten. Man hat nun ein p_Δ zu suchen, welches alle hierdurch entstandene Gleichungen löst. Damit haben wir das Problem in der Zeit diskretisiert und haben somit zeitdiskrete, aber noch keine ortsdiskreten Probleme.

Der volldiskrete Fall

Bis jetzt haben wir angenommen, daß die exakte Lösung der Subprobleme im halbdiskreten Fall bekannt ist. Da es in der Praxis im Allgemeinen nicht möglich ist, in jedem Schritt eine elliptische Gleichung zu lösen, müssen wir sie durch numerische Verfahren approximieren. In dieser Arbeit beschäftigen wir uns ausschließlich mit der FEM basierend auf einer Zerlegung des Gebietes in Dreieckselemente (Triangulierung). Als Ansatzfunktionen wählen wir stetige bzgl. der Triangulierung stückweise lineare Funktionen. Näheres dazu erfährt man in [11]. Sei Q_h ein FE-Raum. Für jeden Zeitschritt im semidiskreten Fall ist ein elliptisches Problem in $H = H_0^1(\Omega)$ zu lösen. Setzen wir in der Variationsformulierung (1.6), wie im Abschnitt 1.1.2, $H = Q_h$, so haben wir ein volldiskretes Problem, das auf ein LGS führt.

1.2.3 Diskontinuierliches Galerkin-Verfahren für parabolische Probleme

Nun haben wir die Diskretisierungsmethoden abstrakt eingeführt. Hier werden wir als Beispiel das diskontinuierliche Galerkin-Verfahren für parabolische Differentialgleichungen herleiten.
Dieses Verfahren führt (im halbdiskreten Fall) zu einer Funktion

$$p_\tau : [0,T] \to H_0^1(\Omega).$$

Analog zu (1.6) fordern wir, gemäß obiger Überlegungen, daß die folgende Variationsformulierung gelte:
Für $l \in \mathbb{N}$ finde man

$$p_\tau \in Q_\tau^l = \left\{ q_\tau : [0,T] \to H_0^1(\Omega) \;\middle|\; q_\tau\big|_{(t_{j-1},t_j]} \text{ ein Polynom vom Grad } l-1 \right\}$$

mit

$$\begin{aligned} &\int_0^T \Big[-\left(p_\tau(t)\varphi_\tau'(t), q\right)_{0,\Omega} + a\left(p_\tau(t)\varphi_\tau(t), q\right) \Big]\, dt \\ &\quad = \left(P_0\varphi_\tau(0), q\right)_{0,\Omega} + \int_0^T \left(f(t)\varphi_\tau(t), q\right)_{0,\Omega} dt\,, \end{aligned} \tag{1.11}$$

für alle $q \in H_0^1(\Omega)$ und alle

$$\varphi_\tau \in \mathcal{P}_\tau^l = \left\{ \varphi : [0,T] \to \mathbb{R} \;\middle|\; \varphi\big|_{(t_{j-1},t_j]} \text{ ein Polynom vom Grad } l-1 \right\}. \tag{1.12}$$

Da

$$\mathcal{P}_\tau^l \otimes H_0^1(\Omega) \underset{\text{dicht}}{\subset} Q_\tau^l\,,$$

ist die Variationsformulierung (1.12) äquivalent zu:
Finde $p_\tau \in Q_\tau^l$ mit

$$\begin{aligned} &\int_0^T \Big[-\left(p_\tau(t), \partial_t q_\tau(t)\right)_{0,\Omega} + a\left(p_\tau(t), q_\tau(t)\right) \Big]\, dt \\ &\quad = \left(P_0, q_\tau(0)\right)_{0,\Omega} + \int_0^T \left(f(t), q_\tau(t)\right)_{0,\Omega} dt\,, \end{aligned} \tag{1.13}$$

für alle $q_\tau \in Q_\tau^l$.
Man beachte, daß jeweils am Zeitpunkt t_j für $j = 1, \ldots, M$ nur die linksseitige Stetigkeit verlangt wird.
Für $q_\tau \in Q_\tau^l$ setzen wir $q_j = q_\tau(t_j)$ und $q_j^+ = \lim\limits_{t \to t_j^+} q_\tau(t)$.

Dann liefert (1.13)

$$\underbrace{\int_0^T -\left(p_\tau(t), \partial_t q_\tau(t)\right)_{0,\Omega} dt}_{=:\mathrm{Int}} + \int_0^T a(p_\tau(t), q_\tau(t))dt = \left(P_0, q_\tau(0)\right)_{0,\Omega} + \int_0^T \left(f(t), q_\tau(t)\right)_{0,\Omega} dt,$$

wobei Int wie folgt umgeformt wird (man beachte, daß $q_\tau(T) = 0$):

$$\begin{aligned}
\mathrm{Int} &= -\sum_{j=1}^{M} \int_{t_{j-1}}^{t_j} \left(p_\tau(t), \partial_t q_\tau(t)\right)_{0,\Omega} dt \\
&= -\sum_{j=1}^{M} \Big[\left(p_\tau(t), q_\tau(t)\right)_{0,\Omega}\Big]_{t_{j-1}}^{t_j} - \int_{t_{j-1}}^{t_j} \left(\partial_t p_\tau(t), q_\tau(t)\right)_{0,\Omega} dt \\
&= \left(p_0^+, q_\tau(0)\right)_{0,\Omega} + \sum_{j=1}^{M-1} \left(p_j^+ - p_j, q(t_j)\right)_{0,\Omega} + \sum_{j=1}^{M} \int_{t_{j-1}}^{t_j} \left(\partial_t p_\tau(t), q_\tau(t)\right)_{0,\Omega} dt\,.
\end{aligned}$$

Da $q_\tau \in Q_\tau^l$, erhalten wir:

$$\begin{aligned}
&\left(p_0^+, q_0^+\right)_{0,\Omega} + \sum_{j=1}^{M-1} \left(p_j^+ - p_j, q_j^+\right)_{0,\Omega} \\
&+ \sum_{j=1}^{M} \int_{t_{j-1}}^{t_j} \Big[\left(\partial_t p_\tau(t), q_t(t)\right)_{0,\Omega} + a\left(p_\tau(t), q_\tau(t)\right)\Big] dt \qquad (1.14) \\
&= \left(P_0, q_0^+\right)_{0,\Omega} + \sum_{j=1}^{M} \int_{t_{j-1}}^{t_j} \left(f(t), q_\tau(t)\right)_{0,\Omega} dt\,, \qquad \forall\ q_\tau \in Q_\tau^l.
\end{aligned}$$

Man beachte, daß $q_\tau(t_j)$ durch q_j^+ ersetzt wurde.
Durch geschickte Wahl der Basis von Q_τ^l zerfällt das Variationsproblem (1.14) in Teilprobleme auf $(t_{j-1}, t_j]$, $j = 1, \ldots, M$, d.h. es entsteht ein Einschrittverfahren.

Diskontinuierliches Galerkin-Verfahren

Ausgehend von $p_0 = P_0$ löse man für $j = 1, \ldots, M$ nacheinander die Variationsprobleme:

$$\begin{aligned}
&\left(p_{j-1}^+, q_{j-1}^+\right) + \int_{t_{j-1}}^{t_j} \Big[\left(\partial_t p_\tau(t), q_\tau(t)\right)_{0,\Omega} + a\left(p_\tau(t), q_\tau(t)\right)\Big] dt \\
&= \left(p_{j-1}, q_{j-1}^+\right)_{0,\Omega} + \int_{t_{j-1}}^{t_j} \left(f(t), q_\tau(t)\right)_{0,\Omega} dt\,, \ \forall\ q_\tau \in Q_\tau^l.
\end{aligned} \qquad (1.15)$$

Satz 1.3 Für jedes $l \in \mathbf{N}$ besitzt die obige Folge von Variationsproblemen (1.15) eine eindeutige Lösung.

Beweis:
Vgl. [46, Chapter 12, pp. 183].

■

Das Verfahren für $l = 1$ (**implizites Euler-Verfahren**):
Jedes $q_\tau \in Q_\tau^1$ läßt sich wie folgt darstellen:

$$q_\tau(t) = q_{j-1}^+ = q_j \qquad \text{in} \qquad (t_{j-1}, t_j] \qquad \text{(stückweise konstant)},$$

für $j = 1, \ldots, M$. Dann vereinfacht sich (1.15) zu

$$(p_j, q)_{0,\Omega} + \tau_j a\,(p_j, q) = (p_{j-1}, q)_{0,\Omega} + \left(\int_{t_{j-1}}^{t_j} f(t)dt, q \right)_{0,\Omega}, \qquad \forall\ q \in H_0^1(\Omega). \quad (1.16)$$

Nun betrachten wir den Fall $l = 2$ (stückweise linear): Jedes $q_\tau \in Q_\tau^2$ läßt sich wie folgt darstellen:

$$q_\tau(t) = q_{j-1}^+ + \frac{1}{\tau_j}\,\left(q_j - q_{j-1}^+\right)(t - t_{j-1}) \qquad \text{in} \qquad (t_{j-1}, t_j]\,.$$

Damit vereinfacht sich (1.15) zu

$$\begin{aligned} &\left(\frac{p_j - p_{j-1}^+}{\tau_j}\,, \int_{t_{j-1}}^{t_j} q_\tau(t)dt \right)_{0,\Omega} + a\left(p_{j-1}^+, \int_{t_{j-1}}^{t_j} q_\tau(t)dt \right) \\ &+ \int_{t_{j-1}}^{t_j} \frac{t - t_{j-1}}{\tau_j}\ a\left(p_j - p_{j-1}^+, q_\tau(t)\right) dt + \left(p_{j-1}^+, q_{j-1}^+\right)_{0,\Omega} \\ &= \left(p_{j-1}, q_{j-1}^+\right)_{0,\Omega} + \int_{t_{j-1}}^{t_j} \left(f(t), q_\tau(t)\right)_{0,\Omega} dt\,, j = 1, \ldots, n. \end{aligned} \quad (1.17)$$

Sei $w_j(t) = \frac{t - t_{j-1}}{\tau_j}$ für $j = 1, \ldots, M$. Ist $q_\tau \in Q_\tau^2$, so ist

$$q_\tau\big|_{(t_{j-1}, t_j]} \in H_0^1(\Omega) \oplus w_j H_0^1(\Omega),\ j = 1, \ldots, M,$$

d.h.

$$q_\tau\big|_{(t_{j-1}, t_j]}(t) = q_1 + w_j(t) q_2, \quad \text{für geeignete } q_1, q_2 \in H_0^1(\Omega),$$

womit (1.17) sich zu

$$\begin{aligned} &\left(p_j - p_{j-1}^+, q_1\right)_{0,\Omega} + \tau_j\ a\left(p_{j-1}^+, q_1\right) + \frac{\tau_j}{2}\ a(p_j - p_{j-1}^+, q_1) + (p_{j-1}^+, q_1)_{0,\Omega} \\ &= (p_{j-1}, q_1)_{0,\Omega} + \left(\int_{t_{j-1}}^{t_j} f(t)dt, q_1 \right)_{0,\Omega} \quad \forall\ q_1 \in H_0^1(\Omega),\ j = 1, \ldots, M \end{aligned} \quad (1.18)$$

und

$$\begin{aligned}&\frac{1}{2}\left(p_j - p_{j-1}^+, q_2\right)_{0,\Omega} + \frac{\tau_j}{2}\, a(p_{j-1}^+, q_2) + \frac{\tau_j}{3}\, a(p_j - p_{j-1}^+, q_2)\\ &= \frac{1}{\tau_j}\left(\int_{t_{j-1}}^{t_j} (t - t_{j-1}) f(t) dt, q_2\right)_{0,\Omega} \quad \forall\ q_2 \in H_0^1(\Omega),\ j = 1, \ldots, M\end{aligned} \tag{1.19}$$

vereinfacht. Oder anders zusammengefaßt:

$$\begin{aligned}&(p_j, q_1)_{0,\Omega} + \frac{\tau_j}{2}\, a(p_{j-1}^+ + p_j, q_1)\\ &= (p_{j-1}, q_1)_{0,\Omega} + \left(\int_{t_{j-1}}^{t_j} f(t) dt, q_1\right)_{0,\Omega} \quad \forall\ q_1 \in H_0^1(\Omega),\ j = 1, \ldots, M\end{aligned} \tag{1.20}$$

und

$$\begin{aligned}&\frac{1}{2}(p_j - p_{j-1}^+, q_2)_{0,\Omega} + \frac{\tau_j}{6}\, a(3p_j - p_{j-1}^+, q_2)\\ &= \frac{1}{\tau_j}\left(\int_{t_{j-1}}^{t_j} (t - t_{j-1})\, f(t) dt, q_2\right)_{0,\Omega} \quad \forall\ q_2 \in H_0^1(\Omega),\ j = 1, \ldots, M\,.\end{aligned} \tag{1.21}$$

Wie man leicht sieht, bilden (1.20) und (1.21) ein System aus zwei elliptischen Variationsproblemen für $p_{j-1}^+, p_j \in H_0^1(\Omega)$. Für zwei Größen $A_{\tau,h}$ und $B_{\tau,h}$ definieren wir:

Definition 1.4 $A_{\tau,h} \gtrsim B_{\tau,h} :\Leftrightarrow$ Es existiert eine Konstante $C > 0$, die nicht von Diskretisierungsparametern wie τ und h (Zeit- und Ort-Diskretisierungsparameter) abhängt mit $A_{\tau,h} \geq C B_{\tau,h}$.
$A_{\tau,h} \lesssim B_{\tau,h} :\Leftrightarrow$ Es gilt $B_{\tau,h} \gtrsim A_{\tau,h}$
$A_{\tau,h} \eqsim B_{\tau,h} :\Leftrightarrow$ Es gilt $A_{\tau,h} \lesssim B_{\tau,h}$ und $A_{\tau,h} \gtrsim B_{\tau,h}$.

Mit der letzten Definition kann man ein Konvergenzresultat im halbdiskreten Fall wie folgt angeben:

Satz 1.5 Sei $l \geq 1$ und $1 \leq j \leq M$. Dann gilt für die mit der diskontinuierlichen Galerkin-Methode gemäß (1.15) berechnete Näherung:

$$\|p_j - p(t_j)\|_{0,\Omega} \lesssim \left(\sum_{i=1}^{j} \tau_i^{2l} \int_{t_{i-1}}^{t_i} \left|\partial_t^l p(t)\right|_{1,\Omega}^2 dt\right)^{1/2}$$

für $j = 1, \ldots, M$.

Beweis:
Vgl. [46, Theorem 12.1].

■

Satz 1.5 macht nur eine Aussage über die Konvergenz an den Gitterpunkten. Der nächste Satz macht eine Konvergenzaussage auf dem jeweiligen Intervall.

Satz 1.6 Sei $l \geq 1$ und $1 \leq j \leq M$. Es sei für $q \in C^0((t_{j-1}, t_j]; L^2(\Omega))$ die Norm

$$|||q|||_j = \sup_{t \in (t_{j-1}, t_j]} \|q(t)\|_{0,\Omega}$$

definiert. Dann gilt für die mit der diskontinuierlichen Galerkin-Methode gemäß (1.15) berechnete Näherung:

$$|||p_\tau - p|||_j \lesssim \|p_j - p(t_j)\|_{0,\Omega} + \|p_{j-1} - p(t_{j-1})\|_{0,\Omega} + \tau_j^l |||\partial_t^l p|||_j$$

für $j = 1, \ldots, M$.

Beweis:
Vgl. [46, Theorem 12.2].

■

Wir diskretisieren nun auch bzgl. des Ortes und ersetzen $H_0^1(\Omega)$ durch den Teilraum Q_h. Analog dazu definieren wir den volldiskreten Raum

$$Q_{\tau,h}^l = \left\{ q_{\tau,h} : [0,T] \to Q_h \;\middle|\; q_{\tau,h}\big|_{(t_{j-1},t_j]} \text{ Polynom von Grad } l-1 \text{ bzgl. } t \right\}.$$

Für $q_{\tau,h} \in Q_{\tau,h}^l$ setzen wir $q_{j,h} = q_{\tau,h}(t_j)$ und $q_{j,h}^+ = \lim_{t \to t_j^+} q_{\tau,h}(t)$.
Das diskontinuierliche Galerkin-Verfahren lautet nun:
Ausgehend von $p_{0,h} = P_{0,h}$, wobei $P_{0,h} \in Q_h$ die Näherung für $P_0 \in L^2(\Omega)$ ist, berechne man für $j = 1, \ldots, M$ nacheinander $p_{\tau,h}(t)\big|_{(t_{n-1},t_n]}$ für $p_{\tau,h} \in Q_{\tau,h}^l$, so daß

$$\begin{aligned} &\int_{t_{j-1}}^{t_j} \left[(\partial_t p_{\tau,h}(t), q_{\tau,h}(t))_{0,\Omega} + a(p_{\tau,h}(t), q_{\tau,h}(t)) \right] dt + (p_{j-1,h}^+, q_{j-1,h}^+)_{0,\Omega} \\ &= (p_{j-1,h}, q_{j-1,h}^+)_{0,\Omega} + \int_{t_{j-1}}^{t_j} (f(t), q_{\tau,h}(t))_{0,\Omega} dt \quad \forall\ q_{\tau,h} \in Q_{\tau,h}^l. \end{aligned} \tag{1.22}$$

Wir betrachten den Fall $l = 1$: Für $1 \leq j \leq M$ bestimme man $p_{j,h} \in Q_h$, so daß

$$\begin{aligned} &\tau_j a(p_{j,h}, q_h) + (p_{j,h}, q_h)_{0,\Omega} \\ &\quad = (p_{j-1,h}, q_h)_{0,\Omega} + \left(\int_{t_{j-1}}^{t_j} f(t)\, dt, q_h \right)_{0,\Omega} \qquad \forall\ q_h \in Q_h. \end{aligned} \tag{1.23}$$

Bilden $\Phi_{h,1}, \ldots, \Phi_{h,N_h}$ eine Basis für Q_h, so kann man auch die Matrix-Schreibweise benutzen:
Es muß also das LGS

$$(\tau_j \mathbf{A}_{\tau,h} + \mathbf{M}_{\tau,h})\, \mathbf{p}_{h,j} = \mathbf{M}_{\tau,h} \mathbf{p}_{h,j-1} + \mathbf{f}_{\tau,h,j}$$

gelöst werden mit

$$\begin{aligned}
\mathbf{A}_{\tau,h} &= \left[a\left(\phi_{h,\mu},\phi_{h,\nu}\right)\right]_{1\le\nu,\mu\le N_h},\\
\mathbf{M}_{\tau,h} &= \left[\left(\phi_{h,\mu},\phi_{h,\nu}\right)_{0,\Omega}\right]_{1\le\nu,\mu\le N_h},\\
\mathbf{f}_{\tau,h,j} &= \left[\left(\int_{t_{j-1}}^{t_j} f(t)\,dt,\phi_{h,\nu}\right)_{0,\Omega}\right]_{1\le\nu\le N_h} \qquad \text{und}\\
\mathbf{p}_{j-1,h} &= \left[p_{j-1,h}(z_\nu)\right]_{1\le\nu\le N_h},
\end{aligned}$$

wobei z_ν für $1\le\nu\le N_h$ die zu der nodalen Basis gehörigen Punkte sind mit

$$\Phi_\mu(z_\nu)=\begin{cases}1 & \mu=\nu\\ 0 & \mu\neq\nu\end{cases}.$$

Die zugehörige diskrete Evolution lautet dann

$$\begin{aligned}
\mathbf{p}_{j,h} &= E_d(t_j;t_{j-1})\mathbf{p}_{j-1,h}, \qquad \text{mit}\\
E_d(t+\tau;t) &= \left(\tau\mathbf{A}_{\tau,h}+\mathbf{M}_{\tau,h}\right)^{-1}\mathbf{M}_{\tau,h}\\
&= \left(\mathbf{I}_{\tau,h}+\tau\mathbf{M}_{\tau,h}^{-1}\mathbf{A}_{\tau,h}\right)^{-1}\\
&= r_{0,1}\left(\tau\mathbf{M}_{\tau,h}^{-1}\mathbf{A}_{\tau,h}\right), \qquad \text{wobei} \qquad r_{0,1}(z)=\frac{1}{1+z}.
\end{aligned}$$

Die Ordnung der Approximation wird durch genauere Betrachtung des *Konsistenzfehlers*

$$\begin{aligned}
\eta\left(t,\mathbf{p};\tau\right) &= \mathbf{p}\left(t+\tau\right)-E_d(t+\tau;t)\mathbf{p}(t)\\
&= \left[\exp\left(-\tau\mathbf{M}_{\tau,h}^{-1}\mathbf{A}_{\tau,h}\right)-r_{0,1}\left(\tau\mathbf{M}_{\tau,h}^{-1}\mathbf{A}_{\tau,h}\right)\right]\mathbf{p}(t),
\end{aligned}$$

durch

$$\begin{aligned}
\exp\left(-z\right)-r_{0,1}\left(z\right) &= 1-z+\frac{z^2}{2}\ -\frac{z^3}{6}\ +\cdots-\left(1-z+z^2+\cdots\right)\\
&= -\frac{z^2}{2}\ +\cdots=\mathcal{O}(z^2)
\end{aligned}$$

bestimmt. Wie bekannt, hat das implizite Euler-Verfahren die Konsistenzordnung 1.

Betrachten wir nun den Fall $l=2$: Bestimme $p_{j,h},\,p_{j-1,h}^{+}\in Q_h$, für $1\le j\le M$, so daß

$$\begin{aligned}
&\left(p_{j,h}\,,q_h\right)_{0,\Omega}+\frac{\tau_j}{2}\ a\left(p_{j-1,h}^{+}+p_{j,h}\,,q_h\right)\\
&\qquad = \left(p_{j-1,h}\,,q_h\right)_{0,\Omega}+\left(\int_{t_{j-1}}^{t_j} f(t)\,dt\,,q_h\right)_{0,\Omega}
\end{aligned}$$

und

$$\begin{aligned}
&\frac{1}{2}\ \left(p_{j,h}-p_{j-1,h}^{+}\,,s_h\right)_{0,\Omega}+\frac{\tau_j}{6}\ a\left(p_{j-1,h}^{+}+2\,p_{j,h}\,,s_h\right)_{0,\Omega}\\
&\quad = \frac{1}{\tau_j}\left(\int_{t_{j-1}}^{t_j}\left(t-t_{j-1}\right)f(t)\,dt\,,s_h\right)_{0,\Omega}, \qquad \forall\ q_h\,,\ s_h\in Q_h.
\end{aligned}$$

Mit den zugehörigen Matrizen gilt

$$\begin{aligned}
&\mathbf{M}_{\tau,h}\mathbf{p}_{j,h} + \frac{\tau_j}{2}\ \mathbf{A}_{\tau,h}\left(\mathbf{p}^+_{j-1,h} + \mathbf{p}_{j,h}\right) = \mathbf{M}_{\tau,h}\mathbf{p}_{j-1,h} + \mathbf{f}_{\tau,h,j}\\
&\frac{1}{2}\,\mathbf{M}_{\tau,h}\left(\mathbf{p}_{j,h} - \mathbf{p}^+_{j-1,h}\right) + \frac{\tau_j}{6}\ \mathbf{A}_{\tau,h}\left(\mathbf{p}^+_{j-1,h} + 2\,\mathbf{p}_{j,h}\right) = \mathbf{g}_{\tau,h,j},
\end{aligned}$$

wobei

$$\mathbf{g}_{\tau,h,j} = \left[\left(\int_{t_{j-1}}^{t_j} \frac{(t-t_{j-1})}{\tau_j}\, f(t)\, dt, \phi_{h,\nu}\right)_{0,\Omega}\right]_{1\le\nu\le N_h}.$$

Das Verfahren in Blockmatrizen zusammengefaßt lautet:

$$\begin{bmatrix} \frac{\tau_j}{2}\,\mathbf{A}_{\tau,h} & \mathbf{M}_{\tau,h} + \frac{\tau_j}{2}\,\mathbf{A}_{\tau,h} \\ \frac{\tau_j}{6}\,\mathbf{A}_{\tau,h} - \frac{1}{2}\,\mathbf{M}_{\tau,h} & \frac{1}{2}\,\mathbf{M}_{\tau,h} + \frac{\tau_j}{3}\,\mathbf{A}_{\tau,h} \end{bmatrix} \begin{bmatrix} \mathbf{p}^+_{j-1,h} \\ \mathbf{p}_{j,h} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{\tau,h}\mathbf{p}_{j-1,h} + \mathbf{f}_{\tau,h,j} \\ \mathbf{g}_{\tau,h,j} \end{bmatrix}.$$

Die diskrete Evolution ergibt sich wie folgt:

$$\begin{aligned}
\mathbf{p}_{j,h} &= E_d(t_j, t_{j-1})\mathbf{p}_{j-1,h} \qquad \text{mit}\\
E_d(t+\tau;t) &= \left[\mathbf{I}_{\tau,h} + \frac{2}{3}\tau\mathbf{M}^{-1}_{\tau,h}\mathbf{A}_{\tau,h} + \frac{\tau^2}{6}\left(\mathbf{M}^{-1}_{\tau,h}\mathbf{A}_{\tau,h}\right)^2\right]^{-1}\left(\mathbf{I}_{\tau,h} - \frac{\tau}{3}\,\mathbf{M}^{-1}_{\tau,h}\mathbf{A}_{\tau,h}\right)\\
&= r_{1,2}\left(\tau\mathbf{M}^{-1}_{\tau,h}\mathbf{A}_{\tau,h}\right),\\
&\text{wobei} \qquad r_{1,2}(z) = \frac{1-\frac{1}{3}\,z}{1+\frac{2}{3}\,z+\frac{1}{6}\,z^2}.
\end{aligned}$$

Für den Konsistenzfehler

$$\begin{aligned}
\eta\,(t,\mathbf{p};\tau) &= \mathbf{p}\,(t+\tau) - E_d(t+\tau;t)\mathbf{p}(t)\\
&= \left[\exp\left(-\tau\mathbf{M}^{-1}_{\tau,h}\mathbf{A}_{\tau,h}\right) - r_{1,2}\left(\tau\mathbf{M}^{-1}_h\mathbf{A}_{\tau,h}\right)\right]\mathbf{p}(t)
\end{aligned}$$

folgern wir mittels Taylorentwicklung an der Stelle $z=0$:

$$\exp\,(z) - r_{1,2}\,(z) = \mathcal{O}(\tau^4)$$

Dieses Verfahren besitzt offenbar Konsistenzordnung 3.

Die A-Stabilität dieses Verfahrens ergibt sich aus:

$$|r_{1,2}(z)| \le 1$$

für alle $z \in \mathbb{C}$ mit $\mathrm{Re}(z) > 0$.

Fehlerabschätzung für den volldiskreten Fall:

Sei $l = 1, 2$. Den gesamten Fehler im volldiskreten Fall können wir auch wie folgt darstellen.

$$p_{j,h} - p(t_j) = p_{j,h} - E_d(t_j;t_{j-1})p_{j-1,h} + E_d(t_j;t_{j-1})p_{j-1,h} - \underbrace{p_j}_{=E_d(t_j;t_{j-1})p_{j-1}} + p_j - p(t_j).$$

Mittels Dreiecksungleichung folgt dann

$$\begin{aligned}\|p_{j,h} - p(t_j)\|_{0,\Omega} \leq \|p_{j,h} - E_d(t_j;t_{j-1})p_{j-1,h}\|_{0,\Omega} \\ + \|E_d(t_j,t_{j-1})\,(p_{j-1,h} - p_{j-1})\,\|_{0,\Omega} + \|p_j - p(t_j)\|_{0,\Omega}.\end{aligned} \tag{1.24}$$

Betrachten wir zunächst in (1.24) den ersten Term auf der rechten Seite, wobei wir als FE-Raum den Raum der Dreieckselemente mit linearen Ansatzfunktionen wählen. Dieser Term beschreibt den Fehler, der durch die Galerkin-Approximation bzgl. des Ortes entsteht, für den nach [11, Corollary II.7.7] gilt:

$$\|p_{j,h} - E_d(t_j;t_{j-1})p_{j-1,h}\|_{0,\Omega} \lesssim h^2,$$

falls H^2-Regularität der vorkommenden elliptischen Subprobleme vorausgesetzt wird. Für den zweiten Term der rechten Seite in (1.24) erhalten wir wegen der A-Stabilität

$$\|E_d(t_j,t_{j-1})\,(p_{j-1,h} - p_{j-1})\,\|_{0,\Omega} \leq \|p_{j-1,h} - p_{j-1}\|_{0,\Omega}.$$

Schließlich liefert Satz 1.5 für den dritten Term auf der rechten Seite

$$\|p_j - p(t_j)\|_{0,\Omega} \lesssim \tau^l.$$

Für die Wärmeleitungsgleichung (2.1) existieren für die Fälle $l = 1, 2$ Konvergenzsätze für das diskontinuerliche Galerkin-Verfahren (vgl. [46, Theorem 12.6 , 12.7]).

Bemerkung 1.7 Da für $1 \leq j \leq M$

$$f(t_j) \approx \frac{1}{\tau_j} \int_{t_{j-1}}^{t_j} f(t)\,dt$$

gilt, vereinfacht sich das implizite Euler-Verfahren (1.16) zu:
Bestimme $p_{j,h} \in Q_h$, so daß

$$\begin{aligned}&\tau_j a\,(p_{j,h}\,,q_h) + (p_{j,h}\,,q_h)_{0,\Omega} \\ &\quad = (p_{j-1,h}\,,q_h)_{0,\Omega} + \tau_j\,(f(t_j)\,,q_h)_{0,\Omega} \qquad \text{für alle } q_h \in Q_h\,.\end{aligned}$$

Dies ist das bekannte *implizite Euler-Verfahren.*

1.2.4 Runge-Kutta-Verfahren

Wir wollen hier die bekannten Runge-Kutta-Verfahren, die auf Runge (vgl. [42]) und Kutta (vgl. [30]) zurückgehen, einführen. Zunächst verallgemeinern wir den Begriff der parabolischen Differentialgleichungen. Sei $D \subset H^1(\Omega)$ ein Unterraum. Sei $\mathcal{S} : [0,T] \times \Omega \times \mathbb{R} \to \mathbb{R}$. Wir führen das abstrakte Cauchy-Problem

$$\begin{aligned}&(\,\mathrm{i}\,)\ \partial_t p(t,x) + \mathcal{S}(t,x,p(t,x)) = 0 \qquad\qquad \forall (t,x) \in (0,T) \times \Omega \\ &(\mathrm{ii})\ p(0,\cdot) = P_0 \in D\end{aligned} \tag{1.25}$$

ein. Gesucht ist nun ein $p \in L^2((0,T);D)$ (die schwache Lösung von (1.25)). Es ist klar, daß (1.25) im Allgemeinen nicht einmal eine Lösung haben muß. Aber unter

bestimmten Voraussetzungen an $\mathcal{S}$ existiert eine Lösung p und hängt stetig von $\mathcal{S}$ und P_0 ab. Diese Voraussetzungen wollen wir nicht an dieser Stelle diskutieren. Dies wird jedoch in Kapitel 3 nachgeholt. Dort werden wir nämlich eine bestimmte Klasse dieser Probleme als semilineare parabolische Differentialgleichungen näher untersuchen.

Definition 1.8 (Runge-Kutta-Verfahren)
Sei $s \in \mathbb{N}$ beliebig und seien $b_i,\ a_{ij} \in \mathbb{R}, i,j = 1,\dots,s$. Seien $c_i = \sum_{j=1}^{s} a_{ij}$. Es sei $b = (b_1,\dots,b_s)^T, c = (c_1,\dots,c_s)^T$ und $\mathcal{A} = (a_{ij})_{i,j=1}^{s}$. Es ist die Näherung $p_\tau(t)$ der Lösung zum Zeitpunkt t bekannt und es soll eine Näherung $p_\tau(t+\tau)$ zum Zeitpunkt $t+\tau$ berechnet werden. Die Methode

$$\begin{aligned} k_i + &\mathcal{S}(c_i\tau, \cdot, p_\tau(t) + \tau \sum_{j=1}^{s} a_{ij} k_j) = 0, \quad i = 1,\dots,s \\ p_\tau(t+\tau) = &E_d(t+\tau;t)p_\tau(t) = p_\tau(t) + \tau \sum_{i=1}^{s} b_i k_i \end{aligned} \tag{1.26}$$

nennt man eine *s-stufige Runge-Kutta-Methode*. Gilt $a_{ij} = 0$ für $i \geq j$, so haben wir eine *explizite* Runge-Kutta-Methode. Sonst haben wir eine *implizite* Runge-Kutta-Methode.

In [16, Lemma 4.14] wird gezeigt, daß die obige Runge-Kutta-Methode genau dann konsistent ist (d.h. $\frac{d}{d\tau} E_d(t+\tau;t)q\big|_{\tau=0} = \mathcal{S}(t,\cdot,q)$ für jeden Startwert q und $t \in (0,T)$), wenn

$$\sum_{i=1}^{s} b_i = 1. \tag{1.27}$$

Wir wollen deshalb (1.27) für alle in dieser Arbeit besprochenen Runge-Kutta-Methoden voraussetzen. Um die Stabilität nichtlinearer Verfahren untersuchen zu können, braucht man neue Begriffe.

Dissipative Differentialgleichungen und B-Stabilität

Zunächst wollen wir eine Vereinfachung vornehmen: Wir setzen von nun an voraus, daß $\mathcal{S}$ bzgl. $(\cdot,\cdot)_{0,\Omega}$ *dissipativ* ist. Das heißt: Für alle $p, q \in D$ und beliebiges $t \in (0,T)$ gilt

$$(\mathcal{S}(t,\cdot,p) - \mathcal{S}(t,\cdot,q), p-q)_{0,\Omega} \geq 0. \tag{1.28}$$

Laut [16, Lemma 6.49] gilt (1.28) genau dann, wenn die exakte Evolution *nichtexpansiv* ist, d.h. für $p, q \in H^1(\Omega)$

$$\|E(t,0)p - E(t,0)q\|_{0,\Omega} \leq \|p-q\|_{0,\Omega}. \tag{1.29}$$

Runge-Kutta-Verfahren, welche die Nichtexpansivität der exakten Evolution ins Diskrete vererben wurden 1975 von Butcher [13] untersucht. Er nannte sie *B-stabil*. B-stabile Runge-Kutta-Verfahren erzeugen für dissipative, hinreichend glatte Funktion $\mathcal{S}$ eine nichtexpansive diskrete Evolution, d.h.

$$\|E_d(t+\tau,t)p - E_d(t+\tau,t)q\|_{0,\Omega} \leq \|p-q\|_{0,\Omega} \tag{1.30}$$

für alle $p, q \in D$ und alle zulässigen $\tau > 0$. Dieses Konzept umfaßt die A-Stabilität für lineare Probleme, denn nach [16, Lemma 6.50] sind B-stabile Runge-Kutta-Methoden A-stabil.
Für lineare autonome Systeme $p' + Ap = 0$ gilt $E_d(t+\tau, t) = r(\tau A)$, wobei $r(z)$ eine rationale Funktion ist.
Es gilt $r(z) = 1 + zb^T(I - z\mathcal{A})^{-1}\mathbf{e}$, wobei $\mathbf{e} = (1, \ldots, 1)^T \in \mathbb{R}^s$ (vgl. [16, Lemma 6.30]).

Beispiel 1.9 Wir wollen als Beispiel eine zweistufige Runge-Kutta-Methode angeben. Betrachten wir die durch $b = (\frac{12}{13}, \frac{1}{13})^T$ und $\mathcal{A} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{12} \\ 1 & \frac{1}{2} \end{bmatrix}$ definierte Runge-Kutta-Methode. Die zugehörige rationale Funktion lautet

$$r_{2,2}(z) = \frac{1 - \frac{1}{6}z^2}{1 + z + \frac{1}{3}z^2}.$$

Diese rationale Funktion wird uns später wiederbegegnen.

Wir wollen an dieser Stelle ein nützliches Lemma über dissipative Funktionen beweisen, das später noch benötigt wird.

Lemma 1.10 Sei $\mathcal{S}$ bzgl. der 3. Komponente stetig Fréchet-differenzierbar. Setzen wir $DS(\hat{p}) = \partial_3 \mathcal{S}(t, \cdot, \hat{p})$, so ist $\mathcal{S}$ dissipativ genau dann, wenn für alle $\hat{p} \in D$ der Operator $DS(\hat{p})$ positiv semidefinit ist, d.h.

$$(DS(\hat{p})\hat{q}, \hat{q})_{0,\Omega} \geq 0 \quad \forall \hat{q} \in D.$$

Beweis:
Sei $DS(\hat{p})$ für alle $\hat{p} \in D$ positiv semidefinit. Zu zeigen ist:

$$(\mathcal{S}(t, \cdot, p) - \mathcal{S}(t, \cdot, q), p - q)_{0,\Omega} \geq 0 \tag{1.31}$$

für alle $p, q \in D$. Seien nun $p, q \in D$ beliebig. Da $D \subset H^1(\Omega)$ ein Unterraum ist, gilt $q + \mu(p - q) \in D$ für alle $\mu \in \mathbb{R}$. Wir definieren $\varphi : [0,1] \to \mathbb{R}$, wobei $\varphi(\mu) = (\mathcal{S}(t, \cdot, q + \mu(p - q)) - \mathcal{S}(t, \cdot, q), p - q)_{0,\Omega}$. Offenbar ist $\varphi \in C^1([0,1])$ mit $\varphi'(\mu) = (DS(q + \mu(p - q))(p - q), p - q)_{0,\Omega} \geq 0$ und $\varphi(0) = 0$. Damit ist $\varphi(\mu) \geq 0$ für alle $\mu \in [0,1]$. Insbesondere gilt $\varphi(1) \geq 0$, woraus (1.31) folgt.
Ist umgekehrt $\mathcal{S}$ dissipativ, so folgern wir für beliebiges $h \in \mathbb{R} \setminus \{0\}$ und beliebiges $\hat{p}, \hat{q} \in D$ sowie $t \in (0, T)$:

$$(\mathcal{S}(t, \cdot, \hat{p} + h\hat{q}) - \mathcal{S}(t, \cdot, \hat{p}), h\hat{q})_{0,\Omega} \geq 0. \tag{1.32}$$

Multiplikation von (1.32) mit $\frac{1}{h^2} > 0$ liefert:

$$(\frac{\mathcal{S}(t, \cdot, \hat{p} + h\hat{q}) - \mathcal{S}(t, \cdot, \hat{p})}{h}, \hat{q})_{0,\Omega} \geq 0.$$

Damit ist

$$(DS(\hat{p})\hat{q}, \hat{q})_{0,\Omega} = \lim_{h \to 0} (\frac{\mathcal{S}(t, \cdot, \hat{p} + h\hat{q}) - \mathcal{S}(t, \cdot, \hat{p})}{h}, \hat{q})_{0,\Omega} \geq 0.$$

■

Kapitel 2

Lineare Probleme

In diesem Kapitel besprechen wir eine auf dem LSF basierende Methode zur Approximation der Wärmeleitungsgleichung und entwickeln einen a-posteriori Fehlerschätzer, der optimale Zeitschrittweiten garantiert. Diese Methode wird in späteren Kapiteln auf allgemeinere Fälle übertragen. Die meisten Ideen dieses Kapitels sind direkt aus [33, 34] übernommen und werden der Vollständigkeit halber hier ausführlich besprochen. Wir betrachten das folgende parabolische System erster Ordnung, das auch als Wärmeleitungsgleichung bekannt ist,

$$\begin{aligned} c\,\partial_t p + \ \mathrm{div}\ u &= \ f, \\ u + a\,\nabla p &= \ 0, \end{aligned} \tag{2.1}$$

in einem beschränkten, mit Polygonzügen berandetem Gebiet $\Omega \subset \mathbb{R}^2$. Es sei $t \in (0,T)$ mit $T > 0$. $\partial\Omega$ sei der Rand von Ω, der in zwei disjunkte Teilmengen $\partial\Omega = \ \Gamma_D \cup \Gamma_N$ aufgeteilt wird, wobei Γ_D eine positive Länge hat. Auf Γ_D bzw. Γ_N schreiben wir Dirichlet- bzw. Neumannrandbedingungen vor, d.h. wir geben p auf Γ_D und $< n, u >$ auf Γ_N vor, wobei $< \cdot, \cdot >$ das Standardskalarprodukt in $\mathbb{R}^2$ ist und n die äußere Einheitsnormale an Γ_N beschreibt. Wir geben darüberhinaus $p(0) = P_0 \in D_A \subset H^1_{\Gamma_D}(\Omega)$ vor, wobei die Definition von D_A später nachgeholt wird. Ferner sei $f \in L^2(\Omega)$ und $a, c \in L^\infty(\Omega)$ unabhängig von t. Für a und c sei vorausgesetzt, daß beide *gleichmäßig von* 0 *weg beschränkt* sind, d.h. es gibt eine Konstante $k > 0$ mit $a(x), c(x) \geq k$ für alle $x \in \Omega$. Wir wollen ab jetzt nur homogene Dirichlet- bzw. Neumannrandwerte betrachten. Die Verallgemeinerung der Methoden auf beliebige Randwerte ist einfach (sobald man die vorgegebenen Dirichlet- bzw. Neumannrandwerte zu einer Funktion in $H^1(\Omega)$ bzw. in $H(\mathrm{div}, \Omega)$ fortgesetzt hat). Die Anfangs- und Randbedingungen, die hier eingeführt worden sind, werden wir während der ganzen Arbeit benutzen.

Wir führen zunächst die folgenden Räume ein:

$$\begin{aligned} H^1_{\Gamma_D}(\Omega) &= \left\{ q \in H^1(\Omega) \mid q = \ 0 \text{ auf } \Gamma_D \right\}, \\ H_{\Gamma_N}(\mathrm{div}\ , \Omega) &= \left\{ v \in H(\mathrm{div}, \Omega) \mid < n, v >= \ 0 \text{ auf } \Gamma_N \right\}. \end{aligned}$$

Es ist klar, daß jede Lösung $(u,p) \in L^2\left((0,T); H_{\Gamma_N}(\mathrm{div}\ , \Omega)\right) \times H^1\left((0,T); H^1_{\Gamma_D}(\Omega)\right)$ dieses Anfangs-Randwertproblems auch das folgende LSF für beliebiges $\alpha > 0$ mini-

miert:

$$\tilde{\mathcal{F}}(u,p;f) = \int_0^T \left(\alpha \parallel c\,\partial_t p + \text{ div } u - f\parallel^2_{0,\Omega} + \parallel u + a\,\nabla p\parallel^2_{0,\Omega}\right) ds\,.$$

2.1 Der halbdiskrete Fall

Nun diskretisieren wir zunächst in der Zeit. Wir wählen also ein Gitter $\Delta = \{t_0, t_1, \ldots, t_M \mid 0 = t_0 < t_1 < \ldots < t_M \leq T\} \subset [0,T]$, die Zeitpunkte, an denen wir eine Approximation der Lösung von (2.1) berechnen wollen. Seien $\tau_j = t_j - t_{j-1}$, $j = 1,\ldots,M$. Es ist klar, daß die Zeitpunkte t_i, $i = 0,\ldots,M$ und damit auch die Zeitschrittweiten τ_j, $j = 1,\ldots,M$ und die Anzahl der Zeitpunkte M von der gewählten Zerlegung Δ abhängen. Deshalb Schreiben wir für ein vorgegebenen Gitter Δ auch $M_\Delta = M$ und analog $\tau_j^\Delta = \tau_j$, $j = 1,\ldots,M_\Delta$ und $t_i^\Delta = t_i$, $i = 0,\ldots,M_\Delta$ immer wenn wir die Abhängigkeit dieser Größen von der Zerlegung Δ betonen wollen.
Wir werden dann das LSF in den Teilräumen

$$V_\tau\left((0,T); H_{\Gamma_N}(\text{div }, \Omega)\right)) \subset L^2\left((0,T); H_{\Gamma_N}(\text{div }, \Omega))\right),$$
$$Q_\tau\left((0,T); H^1_{\Gamma_D}(\Omega))\right) \subset H^1\left((0,T); H^1_{\Gamma_D}(\Omega))\right)$$

minimieren. In diesem Kapitel betrachten wir bzgl. Δ lineare (nicht notwendig stetige) Ansatzfunktionen für V_τ und stückweise lineare stetige Ansatzfunktionen für Q_τ (also jeweils linear auf den Intervallen $(t_j, t_{j+1}), j = 0,\ldots,M-1$).
Wir werden noch zeigen, daß dieses Verfahren für p die Konvergenzordnung 3 in der H^1-Norm hat. Wir werden außerdem Konvergenz in einer skalierten $H^1 \times H(\text{ div })$-Norm zeigen.
Ein Vorteil dieses Verfahrens ist, daß man mit entsprechendem Aufwand auch beliebig hohe Konvergenzordnung bekommt. Da es sich hierbei um ein Einschrittverfahren handelt, nehmen wir an, daß $p(t)$ bekannt ist und wollen mit unserem Verfahren eine Approximation p_τ^+ an $p(t+\tau)$ berechnen.
Bei der numerischen Durchführung des Verfahrens setzt man nacheinander $t = 0, t_1, \ldots, t_{M-1}$ bzw. $\tau = \tau_1, \ldots, \tau_M$.

Mit $\alpha = \tau$ erhalten wir

$$\hat{\mathcal{F}}(u,p;f) = \int_0^\tau \left(\tau\,\|c\partial_t p(t+\sigma) + \text{ div } u(t+\sigma) - f\|^2_{0,\Omega} + \|u(t+\sigma) + a\nabla p(t+\sigma)\|^2_{0,\Omega}\right) d\sigma.$$

Wir ändern dieses Funktional aus Symmetriegründen etwas ab und betrachten im folgenden

$$\begin{aligned}\mathcal{F}(u,p;f) = \int_0^\tau \Big(\tau \,\big\|\, c^{1/2}\partial_t p(t+\sigma) + c^{-1/2}\,(\text{ div } u(t+\sigma) - f)\big\|^2_{0,\Omega} \\ + \big\|\, a^{-1/2}u(t+\sigma) + a^{1/2}\nabla p(t+\sigma)\big\|^2_{0,\Omega}\Big)\, d\sigma.\end{aligned} \tag{2.2}$$

2.1.1 Variationsformulierung für das lineare LSF

Im folgendem benutzen wir für $q_\tau \in Q_\tau$ und $v_\tau \in V_\tau$ die Abkürzungen:

$$\begin{aligned} q_\tau^+ &= q_\tau(t+\tau), \\ v_\tau^- &= v_\tau^+(t) = \lim_{\substack{\sigma\to 0 \\ \sigma>0}} v_\tau(t+\sigma), \\ v_\tau^+ &= v_\tau^-(t+\tau) = \lim_{\substack{\sigma\to 0 \\ \sigma>0}} v_\tau(t+\tau-\sigma). \end{aligned}$$

Da u_τ und p_τ Polynome sind (bzgl. t), kann man sie mittels Lagrange-Polynomen darstellen. Es ist klar, daß man für p_τ und für u_τ jeweils zwei Lagrange-Polynome braucht. Diese sind:

$$\begin{aligned} P_1(\sigma) &= \frac{\tau-\sigma}{\tau}, \\ P_2(\sigma) &= \frac{\sigma}{\tau} \end{aligned}$$

und

$$\begin{aligned} U_1(\sigma) &= \frac{\tau-\sigma}{\tau}, \\ U_2(\sigma) &= \frac{\sigma}{\tau}\ . \end{aligned}$$

Somit ist $p_\tau(t+\sigma) = P_1(\sigma)p_\tau^{\text{alt}} + P_2(\sigma)p_\tau^+$ und $u_\tau(t+\sigma) = U_1(\sigma)u_\tau^- + U_2(\sigma)u_\tau^+$, wobei p_τ^{alt} bekannt ist und die anderen drei Unbekannten durch die Minimierung des LSF (2.2) bestimmt werden.
Unsere Wahl der Ansatzfunktionen führt zu

$$\begin{aligned} &\mathcal{F}(u_\tau, p_\tau; f) \\ &\quad= \int_0^\tau \tau \| c^{1/2} \frac{p_\tau^+ - p_\tau(t)}{\tau} + c^{-1/2} \left(\frac{\tau-\sigma}{\tau}\ \text{div}\ u_\tau^- + \frac{\sigma}{\tau}\ \text{div}\ u_\tau^+ - f \right) \|_{0,\Omega}^2 \, d\sigma \\ &\quad+ \int_0^\tau \| \frac{\tau-\sigma}{\tau}(a^{-1/2}u_\tau^- + a^{1/2}\nabla p_\tau(t)) + \frac{\sigma}{\tau}(a^{-1/2}u_\tau^+ + a^{1/2}\nabla p_\tau^+) \|_{0,\Omega}^2 \, d\sigma \\ &\quad=: \mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), f)\ . \end{aligned} \tag{2.3}$$

Die Minimierung des LSF $\mathcal{F}$ ist in jedem Zeitschritt in den folgenden Räumen durchzuführen:

$$\begin{aligned} \tilde{V}_\tau &= \left\{ \left(1-\frac{\sigma}{\tau}\right) v_1 + \frac{\sigma}{\tau} v_2 \;\middle|\; v_1, v_2 \in H_{\Gamma_N}(\text{div}\ , \Omega) \right\}, \\ \tilde{Q}_\tau &= \left\{ \frac{\sigma}{\tau} q \;\middle|\; q \in H^1_{\Gamma_D}(\Omega) \right\}\ . \end{aligned}$$

Wir suchen also $(u_\tau, \hat{p}_\tau) \in \tilde{V}_\tau \times \tilde{Q}_\tau$, so daß

$$\mathcal{F}\,(u_\tau, p_\tau; f) = \min_{(v_\tau, \hat{q}_\tau) \in \tilde{V}_\tau \times \tilde{Q}_\tau} \mathcal{F}\,(v_\tau, q_\tau; f) \tag{2.4}$$

mit $p_\tau(t+\sigma) = (1-\frac{\sigma}{\tau})p_\tau(t) + \hat{p}_\tau(\sigma)$ und $q_\tau(\sigma) = (1-\frac{\sigma}{\tau})p_\tau(t) + \hat{q}_\tau(\sigma)$. Definieren wir die Bilinearform

$$\begin{aligned}&\mathcal{B}(u_\tau, p_\tau; v_\tau, \hat{q}_\tau)\\ &= \int_0^\tau \big(\tau(c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}\ \mathrm{div}\ u_\tau(t+\sigma),\ c^{1/2}\partial_t q_\tau(\sigma) + c^{-1/2}\ \mathrm{div}\ v_\tau(\sigma))_{0,\Omega} \\ &\quad + (a^{-1/2}u_\tau(t+\sigma) + a^{1/2}\nabla p_\tau(t+\sigma), a^{-1/2}v_\tau(\sigma) + a^{1/2}\nabla\hat{q}_\tau(\sigma))_{0,\Omega}\big)\ d\sigma\end{aligned} \tag{2.5}$$

(mit $p_\tau(t+\sigma) = (1-\frac{\sigma}{\tau})p_\tau(t)+\hat{p}_\tau(\sigma)$) zu $\mathcal{F}(u_\tau, p_\tau; p_\tau(t), f)$, so genügt $(u_\tau, \hat{p}_\tau) \in \tilde{V}_\tau \times \tilde{Q}_\tau$ dem Variationsproblem

$$\mathcal{B}(u_\tau, p_\tau; v_\tau, \hat{q}_\tau) = \tau \int_0^\tau (c^{-1/2}f, c^{1/2}\partial_t\hat{q}_\tau(\sigma) + c^{-1/2}\ \mathrm{div}\ v_\tau(\sigma))_{0,\Omega} d\sigma \tag{2.6}$$

für alle $v_\tau \in \tilde{V}_\tau$ und $\hat{q}_\tau \in \tilde{Q}_\tau$.

2.1.2 Die Analyse des LSF

Um zu zeigen, daß $\mathcal{F}\left(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau^{\mathrm{alt}}, f\right)$ zu einer Norm des Fehlers äquivalent ist, benötigen wir einige Hilfsmittel, die wir in diesem Abschnitt zusammenstellen. Dieses Ergebnis liefert die Grundlage für die adaptive Wahl der Zeitschrittweiten. Gesucht ist die Lösung des modifizierten Problems aus (2.1):

$$\begin{aligned} c^{1/2}\ \partial_t p + c^{-1/2}\ \mathrm{div}\ u &= c^{-1/2}\ f,\\ a^{-1/2}u + a^{1/2}\ \nabla p &= 0. \end{aligned} \tag{2.7}$$

Nun wollen wir den Evolutionsoperator definieren. Es sei $E(t+\sigma;t) : H^1_{\Gamma_D}(\Omega) \to H^1_{\Gamma_D}(\Omega)$, wobei $E(t+\sigma;t)p(t) = p(t+\sigma)$ die Lösung von (2.7) für $\sigma \in (0,\tau)$ bezeichnet. Dann ist (vgl. Bemerkung B.4.(c),(d) und [46, pp. 9 (1.30)])

$$\begin{aligned} p(t+\sigma) &= E(t+\sigma;t)\ p(t)\\ &= c^{-1/2}\ \exp(-\sigma A)\ c^{1/2}\ p(t) + c^{-1/2}\ (I - \exp(-\sigma A))\ A^{-1}\ c^{-1/2}\ f\ , \end{aligned} \tag{2.8}$$

wobei A definiert ist als

$$\begin{aligned} &A : D_A \subset H^1_{\Gamma_D}(\Omega) \to L^2(\Omega)\\ &\quad (c^{1/2}\ A\ c^{1/2}p, q)_{0,\Omega} = (a\ \nabla p, \nabla q)_{0,\Omega} \text{ für alle } q \in H^1_{\Gamma_D}(\Omega)\ , \end{aligned} \tag{2.9}$$

mit

$$D_A = \{q \in H^1_{\Gamma_D}(\Omega) : a\ \nabla q \in H_{\Gamma_N}(\ \mathrm{div}\ , \Omega)\}.$$

Formal ist $A = -c^{-1/2}\ \mathrm{div}\ (a\ \nabla)\ c^{-1/2}$. Man beachte, daß A als Abbildung von D_A nach $L^2(\Omega)$ bijektiv ist. Es ist auch bekannt, daß $E(t+\sigma;t)$ ein Operator ist, der D_A nach D_A abbildet (vgl. [23, Section 7.4]).

Bemerkung 2.1

(i) Die genaue Definition der Operatoren $r_{2,2}(\tau A)$ und $\exp(-\tau A)$ ist in Bemerkung B.2.(a),(b) zu finden.

(ii) Ist $\tilde{p}$ die gesuchte Lösung von (2.1) mit dem Anfangswert $\tilde{p}_0$, wobei f beliebig ist, dann kann man mit den folgenden Überlegungen $\tilde{p}$ aus der Lösung des homogenen Problems und der Lösung eines elliptischen Problems zusammensetzen:

(a) Sei p_f die Lösung des elliptischen Problems

$$Ac^{1/2}p_f = c^{-1/2}f\,, \quad \text{d.h.} \quad p_f = c^{-1/2}A^{-1}c^{-1/2}f.$$

(b) Sei p die Lösung der homogenen Wärmeleitungsgleichung, d.h. die Lösung von

$$\begin{aligned} c^{1/2}\,\partial_t p + c^{-1/2}\,\text{div}\,u &= 0, \\ u + a\,\nabla p &= 0, \end{aligned} \tag{2.10}$$

mit dem Anfangswert $p_0 = \tilde{p}_0 - p_f$, dann ist $\tilde{p} = p + p_f$ die Lösung des Problems (2.1), denn

$$\begin{aligned} \tilde{p}(t) &= p(t) + p_f \\ &= c^{-1/2}\,\exp(-tA)\,c^{1/2}\,p_0 + p_f \\ &= c^{-1/2}\,\exp(-tA)\,c^{1/2}\,(\tilde{p}_0 - p_f) + p_f \\ &= c^{-1/2}\,\exp(-tA)\,c^{1/2}\,\tilde{p}_0 + c^{-1/2}\,(I - \exp(-tA))\,A^{-1}\,c^{-1/2}\,f\,, \end{aligned} \tag{2.11}$$

was offenbar nach (2.8) richtig ist.

(c) Das heißt wir können p durch die zeitdiskrete Funktion p_τ approximieren und $\tilde{p}_\tau = p_\tau + p_f$ approximiert $\tilde{p}$. Wir nehmen ab diesem Zeitpunkt ohne Beschränkung der Allgemeinheit $f = 0$ an, um die Notation zu vereinfachen.

Lemma 2.2 Die Minimierungsaufgabe (2.4) führt auf eine Einschrittmethode. Die diskrete Evolution dieser Einschrittmethode hat die Form

$$\begin{aligned} p_\tau^+ &= E_d(t+\tau;t)\,p_\tau(t) \\ &= c^{-1/2}\,r_{2,2}(\tau A)\,c^{1/2}\,p_\tau(t) + c^{-1/2}\,(I - r_{2,2}(\tau A))\,A^{-1}\,c^{-1/2}\,f\,, \end{aligned}$$

wobei die rationale Funktion $r_{2,2}$ gegeben ist durch

$$r_{2,2}(z) = \frac{1 - \frac{1}{6}z^2}{1 + z + \frac{1}{3}z^2}\,. \tag{2.12}$$

Beweis:

(i) Die Simpson-Regel, die für Polynome vom Grad 2 exakt ist, liefert:

$$\begin{aligned}\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), f) = \; & \frac{1}{6}\|c^{1/2}(p_\tau^+ - p_\tau(t)) + \tau\, c^{-1/2}(\operatorname{div} u_\tau^- - f)\|_{0,\Omega}^2 \\ & + \frac{2}{3}\|c^{1/2}(p_\tau^+ - p_\tau(t)) + \tau\, c^{-1/2}(\frac{1}{2}(\operatorname{div} u_\tau^- + \operatorname{div} u_\tau^+) - f)\|_{0,\Omega}^2 \\ & + \frac{1}{6}\|c^{1/2}(p_\tau^+ - p_\tau(t)) + \tau\, c^{-1/2}(\operatorname{div} u_\tau^+ - f)\|_{0,\Omega}^2 \\ & + \frac{\tau}{6}\|a^{-1/2}u_\tau^- + a^{1/2}\nabla p_\tau(t)\|_{0,\Omega}^2 + \frac{\tau}{6}\|a^{-1/2}u_\tau^+ + a^{1/2}\nabla p_\tau^+\|_{0,\Omega}^2 \\ & + \frac{\tau}{6}\|a^{-1/2}(u_\tau^- + u_\tau^+) + a^{1/2}(\nabla p_\tau(t) + \nabla p_\tau^+)\|_{0,\Omega}^2 \,.\end{aligned} \tag{2.13}$$

Die zu (2.13) gehörige Bilinearform lautet entsprechend:

$$\begin{aligned}&\mathcal{B}(u_\tau^-, u_\tau^+, p_\tau; v_\tau^-, v_\tau^+, q_\tau) \\ &\quad = \frac{1}{3}(c^{1/2}p_\tau^+ + \tau\, c^{-1/2}\ \operatorname{div}\ u_\tau^-, c^{1/2}q_\tau + \tau\, c^{-1/2}\ \operatorname{div}\ v_\tau^-)_{0,\Omega} \\ &\quad + \frac{1}{3}(c^{1/2}p_\tau^+ + \tau\, c^{-1/2}\ \operatorname{div}\ u_\tau^+, c^{1/2}q_\tau + \tau\, c^{-1/2}\ \operatorname{div}\ v_\tau^+)_{0,\Omega} \\ &\quad + \frac{1}{6}(c^{1/2}p_\tau^+ + \tau\, c^{-1/2}\ \operatorname{div}\ u_\tau^-, c^{1/2}q_\tau + \tau\, c^{-1/2}\ \operatorname{div}\ v_\tau^+)_{0,\Omega} \\ &\quad + \frac{1}{6}(c^{1/2}p_\tau^+ + \tau\, c^{-1/2}\ \operatorname{div}\ u_\tau^+, c^{1/2}q_\tau + \tau\, c^{-1/2}\ \operatorname{div}\ v_\tau^-)_{0,\Omega} \\ &\quad + \frac{\tau}{6}(a^{-1/2}u_\tau^-, a^{-1/2}v_\tau^-)_{0,\Omega} \\ &\quad + \frac{\tau}{6}(a^{-1/2}u_\tau^+ + a^{1/2}\nabla p_\tau^+, a^{-1/2}v_\tau^+ + a^{1/2}\,\nabla q_\tau)_{0,\Omega} \\ &\quad + \frac{\tau}{6}(a^{-1/2}(u_\tau^- + u_\tau^+) + a^{1/2}\nabla p_\tau^+, a^{-1/2}(v_\tau^- + v_\tau^+) + a^{1/2}\nabla q_\tau)_{0,\Omega} \,.\end{aligned} \tag{2.14}$$

Damit genügt die Lösung $(u_\tau^-, u_\tau^+, p_\tau^+)$ dem Variationsproblem

$$\begin{aligned}&\mathcal{B}(u_\tau^-, u_\tau^+, p_\tau^+; v_\tau^-, v_\tau^+, q_\tau) \\ &\quad = (c^{1/2}\, p_\tau(t) + \tau\, c^{-1/2}\, f, c^{1/2}\, q_\tau + \frac{\tau}{2}\, c^{-1/2}\,(\ \operatorname{div}\ v_\tau^- + \ \operatorname{div}\ v_\tau^+))_{0,\Omega} \\ &\quad - \frac{\tau}{6}\,(a^{1/2}\nabla p_\tau(t), a^{-1/2}(2v_\tau^- + v_\tau^+))_{0,\Omega} - \frac{\tau}{6}\,(a^{1/2}\nabla p_\tau(t), a^{1/2}\nabla q_\tau)_{0,\Omega}\end{aligned} \tag{2.15}$$

für alle $v_\tau^-, v_\tau^+ \in H_{\Gamma_N}(\operatorname{div}, \Omega)$ und $q_\tau \in H^1_{\Gamma_D}(\Omega)$.
(ii) Nach Bemerkung 2.1 bleibt nur zu zeigen:

$$p_\tau^+ = \; c^{-1/2}\, r_{2,2}(\tau A)\, c^{1/2}\, p_\tau(t)\,.$$

Das Variationsproblem (2.15) ist äquivalent zu:
Finde $(u_\tau^-, u_\tau^+, p_\tau^+)$ so, daß die folgenden drei Gleichungen erfüllt sind:

$$\begin{aligned}&(c^{1/2}p_\tau^+, c^{1/2}q_\tau)_{0,\Omega} + \frac{\tau}{3}(a^{1/2}\nabla p_\tau^+, a^{1/2}\nabla q_\tau)_{0,\Omega} \\ &+ \frac{\tau}{2}(c^{-1/2}\ \operatorname{div}\ u_\tau^-, c^{1/2}q_\tau)_{0,\Omega} + \frac{\tau}{6}(a^{-1/2}u_\tau^-, a^{1/2}\nabla q_\tau)_{0,\Omega} \\ &+ \frac{\tau}{2}(c^{-1/2}\ \operatorname{div}\ u_\tau^+, c^{1/2}q_\tau)_{0,\Omega} + \frac{\tau}{3}(a^{-1/2}u_\tau^+, a^{1/2}\nabla q_\tau)_{0,\Omega} \\ &= (c^{1/2}p_\tau(t), c^{1/2}q_\tau)_{0,\Omega} - \frac{\tau}{6}(a^{1/2}\nabla p_\tau(t), a^{1/2}\nabla q_\tau)_{0,\Omega}\end{aligned} \tag{2.16}$$

für alle $q_\tau \in H^1_{\Gamma_D}(\Omega)$,

$$\begin{aligned}
&\frac{\tau}{2}(c^{1/2}p_\tau^+, c^{-1/2}\ \mathrm{div}\ v_\tau^-)_{0,\Omega} + \frac{\tau}{6}(a^{1/2}\nabla p_\tau^+, a^{-1/2}v_\tau^-)_{0,\Omega} \\
&+ \frac{\tau^2}{3}(c^{-1/2}\ \mathrm{div}\ u_\tau^-, c^{-1/2}\ \mathrm{div}\ v_\tau^-)_{0,\Omega} + \frac{\tau}{3}(a^{-1/2}u_\tau^-, a^{-1/2}v_\tau^-)_{0,\Omega} \\
&+ \frac{\tau^2}{6}(c^{-1/2}\ \mathrm{div}\ u_\tau^+, c^{-1/2}\ \mathrm{div}\ v_\tau^-)_{0,\Omega} + \frac{\tau}{6}(a^{-1/2}u_\tau^+, a^{-1/2}v_\tau^-)_{0,\Omega} \\
&= \frac{\tau}{2}(c^{1/2}p_\tau(t), c^{-1/2}\ \mathrm{div}\ v_\tau^-)_{0,\Omega} - \frac{\tau}{3}(a^{1/2}\nabla p_\tau(t), a^{-1/2}v_\tau^-)_{0,\Omega}
\end{aligned} \tag{2.17}$$

für alle $v_\tau^- \in H_{\Gamma_N}(\ \mathrm{div}\ , \Omega)$, und

$$\begin{aligned}
&\frac{\tau}{2}(c^{1/2}p_\tau^+, c^{-1/2}\ \mathrm{div}\ v_\tau^+)_{0,\Omega} + \frac{\tau}{3}(a^{1/2}\nabla p_\tau^+, a^{-1/2}v_\tau^+)_{0,\Omega} \\
&+ \frac{\tau^2}{6}(c^{-1/2}\ \mathrm{div}\ u_\tau^-, c^{-1/2}\ \mathrm{div}\ v_\tau^+)_{0,\Omega} + \frac{\tau}{6}(a^{-1/2}u_\tau^-, a^{-1/2}v_\tau^+)_{0,\Omega} \\
&+ \frac{\tau^2}{3}(c^{-1/2}\ \mathrm{div}\ u_\tau^+, c^{-1/2}\ \mathrm{div}\ v_\tau^+)_{0,\Omega} + \frac{\tau}{3}(a^{-1/2}\ u_\tau^+, a^{-1/2}v_\tau^+)_{0,\Omega} \\
&= \frac{\tau}{2}(c^{1/2}p_\tau(t), c^{-1/2}\ \mathrm{div}\ v_\tau^+)_{0,\Omega} - \frac{\tau}{6}(a^{1/2}\nabla p_\tau(t), a^{-1/2}v_\tau^+)_{0,\Omega}
\end{aligned} \tag{2.18}$$

für alle $v_\tau^+ \in H_{\Gamma_N}(\ \mathrm{div}\ , \Omega)$. Für den Rest des Beweises benötigen wir nur (2.16) und (2.17). (2.18) verwenden wir später für den Beweis von Lemma 2.3.
Nach dem Integralsatz von Gauß können wir (2.16) schreiben als

$$\begin{aligned}
&(c^{1/2}p_\tau^+, c^{1/2}q_\tau)_{0,\Omega} + \frac{\tau}{3}(a^{1/2}\nabla p_\tau^+, a^{1/2}\nabla q_\tau)_{0,\Omega} \\
&+ \frac{\tau}{3}(c^{-1/2}\ \mathrm{div}\ u_\tau^-, c^{1/2}q_\tau)_{0,\Omega} + \frac{\tau}{6}(c^{-1/2}\ \mathrm{div}\ u_\tau^+, c^{1/2}q_\tau)_{0,\Omega} \\
&= (c^{1/2}p_\tau(t), c^{1/2}q_\tau)_{0,\Omega} - \frac{\tau}{6}(a^{1/2}\nabla p_\tau(t), a^{1/2}\nabla q_\tau)_{0,\Omega}
\end{aligned} \tag{2.19}$$

für alle $q_\tau \in H^1_{\Gamma_D}(\Omega)$. Nochmaliges Anwenden des Gaußschen Satzes führt zu

$$\begin{aligned}
&\Big((I - \frac{\tau}{3}c^{-1/2}\ \mathrm{div}\ (a\nabla)c^{-1/2})c^{1/2}p_\tau^+ + \frac{\tau}{3}c^{-1/2}\ \mathrm{div}\ u_\tau^- + \frac{\tau}{6}c^{-1/2}\ \mathrm{div}\ u_\tau^+, c^{1/2}q_\tau\Big)_{0,\Omega} \\
&= \Big((I + \frac{\tau}{6}c^{-1/2}\ \mathrm{div}\ (a\nabla)\, c^{-1/2})c^{1/2}p_\tau(t), c^{1/2}q_\tau\Big)_{0,\Omega}
\end{aligned}$$

für alle $q_\tau \in H^1_0(\Omega)$. Da $H^1_0(\Omega)$ in $L^2(\Omega)$ dicht liegt, folgern wir

$$\begin{aligned}
\frac{\tau}{3}c^{-1/2}\ \mathrm{div}\ u_\tau^- + \frac{\tau}{6}c^{-1/2}\ \mathrm{div}\ u_\tau^+ = &\ (I + \frac{\tau}{6}c^{-1/2}\ \mathrm{div}\ (a\nabla)\, c^{-1/2})c^{1/2}p_\tau(t) \\
&- (I - \frac{\tau}{3}c^{-1/2}\ \mathrm{div}\ (a\nabla)\, c^{-1/2})c^{1/2}p_\tau^+
\end{aligned} \tag{2.20}$$

in $L^2(\Omega)$. Andererseits folgt aus (2.19)

$$\begin{aligned}
&\Big(\frac{\tau}{3}c^{-1/2}\ \mathrm{div}\ u_\tau^- + \frac{\tau}{6}c^{-1/2}\ \mathrm{div}\ u_\tau^+, c^{1/2}q_\tau\Big)_{0,\Omega} \\
&= \Big((I + \frac{\tau}{6}c^{-1/2}\ \mathrm{div}\ (a\nabla)\, c^{-1/2})c^{1/2}p_\tau(t), c^{1/2}q_\tau\Big)_{0,\Omega} \\
&- \Big((I - \frac{\tau}{3}c^{-1/2}\ \mathrm{div}\ (a\nabla)\, c^{-1/2})c^{1/2}p_\tau^+, c^{1/2}q_\tau\Big)_{0,\Omega} \\
&- \frac{\tau}{6}(< n, (a\,\nabla p_\tau(t)) >, q_\tau)_{0,\Gamma_N} - \frac{\tau}{3}(< n, (a\,\nabla p_\tau^+) >, q_\tau)_{0,\Gamma_N}
\end{aligned}$$

für alle $q_\tau \in H^1_{\Gamma_D}(\Omega)$. Zusammen mit (2.20) haben wir

$$\langle n, (a\,\nabla p_\tau(t))\rangle + 2\left\langle n, (a\,\nabla p_\tau^+)\right\rangle = 0 \text{ auf } \Gamma_N\,.$$

Zusammen mit unserer generellen Annahme, daß $\langle n, (a\,\nabla p_\tau(0))\rangle = 0$ auf Γ_N, folgt

$$\langle n, (a\,\nabla p_\tau(t))\rangle = \left\langle n, (a\,\nabla p_\tau^+)\right\rangle = 0 \text{ auf } \Gamma_N$$

und deshalb $p_\tau(t), p_\tau^+ \in D_A$. Aus der Definition von A bei (2.9) ergibt sich

$$\begin{aligned}\Big((I+\frac{\tau}{3}A)c^{1/2}p_\tau^+ + \frac{\tau}{3}c^{-1/2}\ \text{div}\ u_\tau^- + \frac{\tau}{6}c^{-1/2}\ \text{div}\ u_\tau^+, c^{1/2}q_\tau\Big)_{0,\Omega}\\ = \Big((I-\frac{\tau}{6}A)c^{1/2}p_\tau(t), c^{1/2}q_\tau\Big)_{0,\Omega}\end{aligned}$$

für alle $q_\tau \in H^1_{\Gamma_D}(\Omega)$ und deshalb ist

$$\frac{\tau}{3}c^{-1/2}\ \text{div}\ u_\tau^- + \frac{\tau}{6}c^{-1/2}\ \text{div}\ u_\tau^+ = (I-\frac{\tau}{6}A)c^{1/2}p_\tau(t) - (I+\frac{\tau}{3}A)c^{1/2}p_\tau^+ \qquad (2.21)$$

in $L^2(\Omega)$ (da $H^1_{\Gamma_D}(\Omega)$ dicht in $L^2(\Omega)$).
Beschränken wir uns auf $q_\tau \in D_A$ und setzen $v_\tau^- = a\,\nabla q_\tau$ so, führt dies mit Hilfe von (2.9) zu

$$c^{-1/2}\ \text{div}\ v_\tau^- = -A\,c^{1/2}q_\tau\,.$$

Aus (2.17) folgt somit

$$\begin{aligned}&-\frac{\tau}{2}(c^{1/2}p_\tau^+, A\,c^{1/2}q_\tau)_{0,\Omega} + \frac{\tau}{6}(a^{1/2}\nabla p_\tau^+, a^{1/2}\nabla q_\tau)_{0,\Omega}\\ &-\frac{\tau^2}{3}(c^{-1/2}\ \text{div}\ u_\tau^-, A\,c^{1/2}q_\tau)_{0,\Omega} - \frac{\tau}{3}(c^{-1/2}\ \text{div}\ u_\tau^-, c^{1/2}q_\tau)_{0,\Omega}\\ &-\frac{\tau^2}{6}(c^{-1/2}\ \text{div}\ u_\tau^+, A\,c^{1/2}q_\tau)_{0,\Omega} - \frac{\tau}{6}(c^{-1/2}\ \text{div}\ u_\tau^+, c^{1/2}q_\tau)_{0,\Omega}\\ &= -\frac{\tau}{2}(A\,c^{1/2}p_\tau(t), c^{1/2}q_\tau)_{0,\Omega} - \frac{\tau}{3}(a^{1/2}\nabla p_\tau(t), a^{1/2}\nabla q_\tau)_{0,\Omega}\,,\end{aligned}$$

für alle $q_\tau \in D_A$. Daraus erhalten wir

$$\begin{aligned}\Big(\frac{\tau}{3}c^{-1/2}\ \text{div}\ u_\tau^- + \frac{\tau}{6}c^{-1/2}\ \text{div}\ u_\tau^+\,, (I+\tau A)c^{1/2}q_\tau\Big)_{0,\Omega}\\ = -\frac{\tau}{3}\left(Ac^{1/2}p_\tau^+, c^{1/2}q_\tau\right)_{0,\Omega} + \left(\frac{5}{6}\tau Ac^{1/2}p_\tau(t), c^{1/2}q_\tau\right)_{0,\Omega}\\ = \left((I+\tau A)\,(I+\tau A)^{-1}\tau Ac^{1/2}\left(-\frac{1}{3}p_\tau^+ + \frac{5}{6}p_\tau(t)\right), c^{1/2}q_\tau\right)_{0,\Omega}\end{aligned}$$

für alle $q_\tau \in D_A$. Da $I+\tau A$ auf D_A selbstadjungiert ist, können wir aus der obigen Gleichung folgern:

$$\begin{aligned}\Big(\frac{\tau}{3}c^{-1/2}\ \text{div}\ u_\tau^- + \frac{\tau}{6}c^{-1/2}\ \text{div}\ u_\tau^+, (I+\tau A)\,c^{1/2}q_\tau\Big)_{0,\Omega}\\ = \left((I+\tau A)^{-1}\tau A\,\left(\frac{5}{6}c^{1/2}p_\tau(t) - \frac{1}{3}c^{1/2}p_\tau^+\right), (I+\tau A)\,c^{1/2}q_\tau\right)_{0,\Omega}\end{aligned}$$

für alle $q_\tau \in D_A$. $I + \tau A$ ist als Operator zwischen D_A und $L^2(\Omega)$ bijektiv und damit haben wir

$$\frac{\tau}{3}c^{-1/2}\ \mathrm{div}\ u_\tau^- + \frac{\tau}{6}c^{-1/2}\ \mathrm{div}\ u_\tau^+ = \tau A\,(I+\tau A)^{-1}\left(\frac{5}{6}c^{1/2}p_\tau(t) - \frac{1}{3}c^{1/2}p_\tau^+\right) \qquad (2.22)$$

in $L^2(\Omega)$. Insgesamt erhalten wir aus (2.21) und (2.22)

$$\left(\frac{1}{3}\tau A\,(I+\tau A)^{-1} - I - \frac{\tau}{3}A\right)c^{1/2}p_\tau^+ = \left(\frac{5}{6}\tau A\,(I+\tau A)^{-1} - I + \frac{\tau}{6}A\right)c^{1/2}p_\tau(t),$$

was Lemma 2.2 beweist.

■

Es ist auch möglich, die Größen u_τ^+ und u_τ^- wie in Lemma 2.2 auszudrücken.

Lemma 2.3 Für die Variationsprobleme (2.14) und (2.15) haben wir

$$u_\tau^- = -a\,\nabla\, c^{-1/2} r_{2,3}^-(\tau A)\, c^{1/2} p_\tau(t),$$
$$u_\tau^+ = -a\,\nabla\, c^{-1/2} r_{0,1}^+(\tau A)\, c^{1/2} p_\tau(t),$$

wobei die rationalen Funktionen $r_{2,3}^-$ und $r_{0,1}^+$ wie folgt gegeben sind:

$$r_{2,3}^-(z) = \frac{1+2z+\frac{5}{6}z^2}{(1+z)(1+z+\frac{1}{3}z^2)}\,,\ r_{0,1}^+(z) = \frac{1}{1+z}\,. \qquad (2.23)$$

Beweis:
Analog zu (2.22) in Beweis von Lemma 2.2 folgern wir aus (2.18):

$$\frac{\tau}{6}c^{-1/2}\ \mathrm{div}\ u_\tau^- + \frac{\tau}{3}c^{-1/2}\ \mathrm{div}\ u_\tau^+ = \tau A(I+\tau A)^{-1}\left(\frac{2}{3}c^{1/2}p_\tau(t) - \frac{1}{6}c^{1/2}p_\tau^+\right)\,.$$

Zusammen mit (2.22) ergibt sich:

$$\tau\, c^{-1/2}\ \mathrm{div}\ u_\tau^- = \tau A\,(I+\tau A)^{-1}(2c^{1/2}p_\tau(t) - c^{1/2}p_\tau^+) = \tau A\, r_{2,3}^-(\tau A)\, c^{1/2}p_\tau(t)\,,$$
$$\tau\, c^{-1/2}\ \mathrm{div}\ u_\tau^+ = \tau A\,(I+\tau A)^{-1}c^{1/2}p_\tau(t) = \tau A\, r_{0,1}^+(\tau A)\, c^{1/2}p_\tau(t)\,.$$

Setzen wir dies und das Ergebnis aus Lemma 2.2 in (2.17) und (2.18) ein, so folgt

$$\frac{1}{3}u_\tau^- + \frac{1}{6}u_\tau^+ = a\,\nabla\, c^{-1/2}\left(\frac{1}{3}r_{2,2}(\tau A) - \frac{5}{6}I + \frac{1}{3}\tau A r_{2,3}^-(\tau A) + \frac{1}{6}\tau A r_{0,1}^+(\tau A)\right)c^{1/2}p_\tau(t),$$
$$\frac{1}{6}u_\tau^- + \frac{1}{3}u_\tau^+ = a\,\nabla\, c^{-1/2}\left(\frac{1}{6}r_{2,2}(\tau A) - \frac{2}{3}I + \frac{1}{6}\tau A r_{2,3}^-(\tau A) + \frac{1}{3}\tau A r_{0,1}^+(\tau A)\right)c^{1/2}p_\tau(t).$$

Hieraus erhält man

$$u_\tau^- = a\,\nabla\, c^{-1/2}\left(r_{2,2}(\tau A) - 2I + \tau A\, r_{2,3}^-(\tau A)\right)\, c^{1/2}p_\tau(t)\,,$$
$$u_\tau^+ = a\,\nabla\, c^{-1/2}\left(-I + \tau A\, r_{0,1}^+(\tau A)\right)\, c^{1/2}p_\tau(t)\,,$$

was äquivalent zu (2.23) ist.

■

2.1.3 Das LSF als Fehlerschätzer

Unser Ziel in diesem Abschnitt ist es, die Äquivalenz des LSF zum Konsistenzfehler zu zeigen.
Dies ermöglicht es uns, das LSF als einen Fehlerschätzer zu benutzen, um die Zeitschrittweiten adaptiv zu wählen. Wir fahren zunächst mit der folgenden Tatsache über das LSF fort.

Lemma 2.4 Für das in (2.3) definierte LSF gilt

$$\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), 0) = (c^{1/2} p_\tau(t), r_{4,3}^{\mathcal{F}}(\tau A) c^{1/2} p_\tau(t))_{0,\Omega} , \tag{2.24}$$

wobei

$$r_{4,3}^{\mathcal{F}}(z) = \frac{z^4}{12(1+z)(1+z+\frac{1}{3}z^2)} .$$

Beweis:
Aus Lemma 2.2 und Lemma 2.3 folgern wir

$$\begin{aligned}
& c^{1/2}\ (p_\tau^+ - p_\tau(t)) + \tau\, c^{-1/2}\ \mathrm{div}\ \ u_\tau(t+\sigma) \\
&\quad = \left((r_{2,2}(\tau A) - I) + \left(1 - \frac{\sigma}{\tau}\right)\ \tau A\, r_{2,3}^-(\tau A) + \frac{\sigma}{\tau}\,\tau A\, r_{0,1}^+(\tau A)\right)\ c^{1/2} p_\tau(t) \qquad \text{und} \\
& a^{-1/2}\ u_\tau(t+\sigma) + a^{1/2}\ \nabla p_\tau(t+\sigma) \\
&\quad = \ a^{1/2} \nabla c^{-1/2} \left(I + \frac{\sigma}{\tau}(r_{2,2}(\tau A) - I) \right. \\
&\qquad\qquad \left. -(\tau A)^{-1} \left(\left(1 - \frac{\sigma}{\tau}\right)\ \tau A\, r_{2,3}^-(\tau A) + \frac{\sigma}{\tau}\,\tau A\, r_{0,1}^+(\tau A) \right) \right)\ c^{1/2} p_\tau(t) .
\end{aligned} \tag{2.25}$$

Setzen wir (2.25) in das LSF ein, so erhalten wir mit (2.9):

$$\begin{aligned}
& \mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), 0) \\
&= \int_0^\tau \left(\frac{1}{\tau} \| \left(r_{2,2}(\tau A) - I + \tau A\, r_{2,3}^-(\tau A) + \frac{\sigma}{\tau}\,\tau A\, (r_{0,1}^+(\tau A) - r_{2,3}^-(\tau A)) \right)\ c^{1/2} p_\tau(t) \|_{0,\Omega}^2 \right. \\
&\quad \left. + \| A^{1/2} \left(I - r_{2,3}^-(\tau A) + \frac{\sigma}{\tau}(r_{2,2}(\tau A) - I - r_{0,1}^+(\tau A) + r_{2,3}^-(\tau A)) \right) c^{1/2} p_\tau(t) \|_{0,\Omega}^2 \right) d\sigma \\
&= \| (r_{2,2}(\tau A) - I + \tau A\, r_{2,3}^-(\tau A))\ c^{1/2} p_\tau(t) \|_{0,\Omega}^2 \\
&\quad + \| (\tau A)^{1/2} (I - r_{2,3}^-(\tau A))\ c^{1/2} p_\tau(t) \|_{0,\Omega}^2 \\
&\quad + \frac{1}{3}\, \| \tau A\, (r_{0,1}^+(\tau A) - r_{2,3}^-(\tau A))\ c^{1/2} p_\tau(t) \|_{0,\Omega}^2 \\
&\quad + \frac{1}{3}\, \| ((\tau A)^{1/2} (r_{2,2}(\tau A) - I) - (\tau A)^{1/2} (r_{0,1}^+(\tau A) - r_{2,3}^-(\tau A)))\ c^{1/2} p_\tau(t) \|_{0,\Omega}^2 \\
&\quad + ((r_{2,2}(\tau A) - I + \tau A\, r_{2,3}^-(\tau A))\ c^{1/2} p_\tau(t), \tau A\, (r_{0,1}^+(\tau A) - r_{2,3}^-(\tau A))\ c^{1/2} p_\tau(t))_{0,\Omega} \\
&\quad + (\tau A (I - r_{2,3}^-(\tau A))\ c^{1/2} p_\tau(t), (r_{2,2}(\tau A) - I - r_{0,1}^+(\tau A) + r_{2,3}^-(\tau A))\ c^{1/2} p_\tau(t))_{0,\Omega} .
\end{aligned}$$

In der obigen Formel tauchen jeweils nur rationale Funktionen von τA auf. Da A selbstadjungiert ist, kann man die obigen Terme zusammenfassen und erhält durch algebraische Umformungen die Behauptung.

■

Das folgende Resultat wird für spätere Zwecke benötigt.

Lemma 2.5 Für $w \in L^2((t,t+\tau);L^2(\Omega))$ gilt:

$$\frac{1}{\tau}\int_0^\tau \left\| \int_0^\sigma w(t+\rho)\,d\rho \right\|_{0,\Omega}^2 d\sigma \leq \tau \int_0^\tau \|w(t+\sigma)\|_{0,\Omega}^2\,d\sigma\,. \tag{2.26}$$

Beweis:
Es ist

$$\begin{aligned}
&\frac{1}{\tau}\int_0^\tau \left\| \int_0^\sigma w(t+\rho)\,d\rho \right\|_{0,\Omega}^2 d\sigma = \frac{1}{\tau}\int_0^\tau \int_\Omega \left(\int_0^\sigma w(t+\rho)d\rho \right)^2 dx\,d\sigma \\
&\underbrace{\leq}_{(*)} \frac{1}{\tau}\int_0^\tau \int_\Omega \sigma \int_0^\sigma (w(t+\rho))^2 d\rho\,dx\,d\sigma \leq \frac{1}{\tau}\int_0^\tau \sigma\,d\sigma \int_\Omega \int_0^\tau (w(t+\rho)(x))^2 d\rho\,dx \\
&\leq \tau \int_\Omega \int_0^\tau (w(t+\rho)(x))^2 d\rho\,dx \underbrace{=}_{(**)} \tau \int_0^\tau \|w(t+\sigma)\|_{0,\Omega}^2\,d\sigma,
\end{aligned} \tag{2.27}$$

wobei in $(*)$ die Cauchy-Schwarzsche Ungleichung und in $(**)$ der Satz von Fubini für nicht negative Funktionen benutzt wurde.

■

Mit diesen Mitteln ist es nun möglich zu zeigen, daß das LSF äquivalent ist zu einer Norm auf $L^2((0,T);H_{\Gamma_N}(\text{div}\,,\Omega)) \times H^1((0,T);H^1_{\Gamma_D}(\Omega))$.

Satz 2.6 Auf $L^2((t,t+\tau);H_{\Gamma_N}(\text{ div }\,,\Omega)) \times H^1((t,t+\tau);H^1_{\Gamma_D}(\Omega))$ sei die folgende Norm definiert:

$$\begin{aligned}
|||(v,q)|||_\tau = \Bigg(\int_0^\tau & (\|v(t+\sigma)\|_{0,\Omega}^2 + \tau \,\|\text{ div }\, v(t+\sigma)\|_{0,\Omega}^2 \\
& + \frac{1}{\tau}\,\|q(t+\sigma)\|_{0,\Omega}^2 + \|\nabla\, q(t+\sigma)\|_{0,\Omega}^2) d\sigma \Bigg)^{1/2}\,.
\end{aligned} \tag{2.28}$$

Den Konsistenzfehler von u bzw. p bezeichnen wir mit

$$\begin{aligned}
\eta_p(t+\sigma) &= p_\tau(t+\sigma) - E(t+\sigma;t)p_\tau(t)\,, \\
\eta_u(t+\sigma) &= u_\tau(t+\sigma) + a\,\nabla E(t+\sigma;t)p_\tau(t)\,.
\end{aligned}$$

Dann gilt:

$$\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), f) \eqsim |||(\eta_u,\eta_p)|||_\tau^2\,. \tag{2.29}$$

Beweis:
(i) Unser erstes Ziel ist es, die folgende Äquivalenz zu zeigen

$$\begin{aligned}
&\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), f) \\
&\eqsim \frac{1}{\tau}\int_0^\tau \|c^{1/2}\,(p_\tau(t+\sigma) - p_\tau(t)) + \int_0^\sigma c^{-1/2}\ \text{div}\ u_\tau(t+\rho)\,d\rho\|_{0,\Omega}^2\,d\sigma \\
&+ \int_0^\tau \|a^{-1/2}\,u_\tau(t+\sigma) + a^{1/2}\,\nabla p_\tau(t+\sigma)\|_{0,\Omega}^2\,d\sigma\,.
\end{aligned}$$

Wir betrachten die Gleichung (2.3)

$$\begin{aligned}
&\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), 0) \\
&= \int_0^\tau \tau\|c^{1/2}\,\frac{p_\tau^+ - p_\tau(t)}{\tau} + c^{-1/2}\ \text{div}\ u_\tau(t+\sigma)\|_{0,\Omega}^2\,d\sigma \\
&\quad + \int_0^\tau \|a^{-1/2}\,u_\tau(t+\sigma) + a^{1/2}\,\nabla p_\tau(t+\sigma)\|_{0,\Omega}^2\,d\sigma\ =: I_1 + I_2.
\end{aligned} \tag{2.30}$$

Da p_τ und u_τ bzgl. t linear sind, kann der erste Term in (2.30) wie folgt umgeformt werden zu

$$\begin{aligned}
I_1 &= \tau\int_0^\tau \|c^{1/2}\,\frac{p_\tau^+ - p_\tau(t)}{\tau} + (1-\frac{\sigma}{\tau})c^{-1/2}\ \text{div}\ u_\tau^- \\
&\qquad + \frac{\sigma}{\tau}c^{-1/2}\ \text{div}\ u_\tau^+\|_{0,\Omega}^2\,d\sigma \\
&= \left(\begin{pmatrix} c^{1/2}(p_\tau^+ - p_\tau(t)) \\ c^{-1/2}\tau\ \text{div}\ u_\tau^- \\ c^{-1/2}\tau\ \text{div}\ u_\tau^+ \end{pmatrix}, \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} c^{1/2}(p_\tau^+ - p_\tau(t)) \\ c^{-1/2}\tau\ \text{div}\ u_\tau^- \\ c^{-1/2}\tau\ \text{div}\ u_\tau^+ \end{pmatrix}\right)_{0,\Omega}.
\end{aligned}$$

Andererseits ist

$$\begin{aligned}
&\frac{1}{\tau}\int_0^\tau \|c^{1/2}(p_\tau(t+\sigma) - p_\tau(t)) + c^{-1/2}\int_0^\sigma \text{div}\ u_\tau(t+\rho)\,d\rho\|_{0,\Omega}^2\,d\sigma \\
&= \frac{1}{\tau}\int_0^\tau \|\frac{\sigma}{\tau}\,c^{1/2}\,(p_\tau^+ - p_\tau(t)) + \frac{\sigma}{\tau}(\tau - \frac{1}{2}\sigma)\,c^{-1/2}\ \text{div}\ u_\tau^- \\
&\qquad + \frac{\sigma^2}{2\tau}\ \text{div}\ u_\tau^+\|_{0,\Omega}^2\,d\sigma \\
&= \left(\begin{pmatrix} c^{1/2}(p_\tau^+ - p_\tau(t)) \\ c^{-1/2}\tau\ \text{div}\ u_\tau^- \\ c^{-1/2}\tau\ \text{div}\ u_\tau^+ \end{pmatrix}, \begin{pmatrix} \frac{1}{3} & \frac{5}{24} & \frac{1}{8} \\ \frac{5}{24} & \frac{2}{15} & \frac{3}{40} \\ \frac{1}{8} & \frac{3}{40} & \frac{1}{20} \end{pmatrix} \begin{pmatrix} c^{1/2}(p_\tau^+ - p_\tau(t)) \\ c^{-1/2}\tau\ \text{div}\ u_\tau^- \\ c^{-1/2}\tau\ \text{div}\ u_\tau^+ \end{pmatrix}\right)_{0,\Omega}.
\end{aligned}$$

Die obigen Matrizen haben den gleichen eindimensionalen Nullraum $\{s(-1,1,1)^T \mid s \in \mathbb{R}\}$, und bei beiden sind alle von Null verschiedenen Eigenwerte positiv. Somit gilt

$$\begin{aligned}
&\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), 0) \\
&\eqsim \frac{1}{\tau}\int_0^\tau \|c^{1/2}\,(p_\tau(t+\sigma) - p_\tau(t)) + \int_0^\sigma c^{-1/2}\ \text{div}\ u_\tau(t+\rho)\,d\rho\|_{0,\Omega}^2\,d\sigma \\
&+ \int_0^\tau \|a^{-1/2}\,u_\tau(t+\sigma) + a^{1/2}\,\nabla p_\tau(t+\sigma)\|_{0,\Omega}^2\,d\sigma\,.
\end{aligned}$$

(ii) Die Definition des Operators $E(t+\sigma;t)$ liefert

$$c^{1/2}\,(E(t+\sigma;t)\,p_\tau(t)-p_\tau(t))-c^{-1/2}\int_0^\sigma \text{ div }\ (a\,\nabla E(t+\rho;t)p_\tau(t))\,d\rho\ =\ 0.$$

Einsetzen der obigen Gleichung in (2.30) ergibt

$$\begin{aligned}\mathcal{F}(u_\tau^-,u_\tau^+,p_\tau^+;p_\tau(t),0)\approx\frac{1}{\tau}\int_0^\tau &\|c^{1/2}\,(p_\tau(t+\sigma)-E(t+\sigma;t)\,p_\tau(t))\\ &+c^{-1/2}\int_0^\sigma \text{ div }\ (u_\tau(t+\rho)+a\,\nabla E(t+\rho;t)p_\tau(t))\,d\rho\|_{0,\Omega}^2 d\sigma\\ &\quad+\int_0^\tau \|a^{-1/2}\,u_\tau(t+\sigma)+a^{1/2}\,\nabla E(t+\sigma;t)p_\tau(t)\\ &\qquad+a^{1/2}\,\nabla(p_\tau(t+\sigma)-E(t+\sigma;t)p_\tau(t))\|_{0,\Omega}^2\,d\sigma\,.\end{aligned}$$

Da a und c beschränkt und gleichmäßig von 0 weg beschränkt sind, gilt

$$\begin{aligned}\mathcal{F}(u_\tau^-,u_\tau^+,p_\tau^+;p_\tau(t),0)\approx\frac{1}{\tau}\int_0^\tau &\big(\|p_\tau(t+\sigma)-E(t+\sigma;t)\,p_\tau(t)\|_{0,\Omega}^2\\ &+\frac{1}{\tau}\|\int_0^\sigma \text{ div }\ (u_\tau(t+\rho)+a\,\nabla E(t+\rho;t)p_\tau(t))\,d\rho\|_{0,\Omega}^2\\ &+\|u_\tau(t+\sigma)+a\,\nabla E(t+\sigma;t)p_\tau(t)\|_{0,\Omega}^2\\ &+\|\nabla(p_\tau(t+\sigma)-E(t+\sigma;t)p_\tau(t))\|_{0,\Omega}^2\\ +\frac{2}{\tau}(p_\tau(t+\sigma)-E(t+\sigma;t)\,p_\tau(t),&\int_0^\sigma \text{ div }\ (u_\tau(t+\rho)+a\,\nabla E(t+\rho;t)p_\tau(t))\,d\rho)_{0,\Omega}\\ +2\,(u_\tau(t+\sigma)+a\,\nabla E(t+\sigma;t)p_\tau(t),&\nabla(p_\tau(t+\sigma)-E(t+\sigma;t)p_\tau(t)))_{0,\Omega})\ d\sigma\,.\end{aligned}$$

Die obige Äquivalenz kann durch den Konsistenzfehler wie folgt ausgedrückt werden

$$\begin{aligned}&\mathcal{F}(u_\tau^-,u_\tau^+,p_\tau^+;p_\tau(t),0)\\ &\approx\int_0^\tau\Big(\frac{1}{\tau}\|\eta_p(t+\sigma)\|_{0,\Omega}^2+\frac{1}{\tau}\|\int_0^\sigma \text{ div }\ \eta_u(t+\rho)\,d\rho\|_{0,\Omega}^2\\ &\qquad\qquad+\|\eta_u(t+\sigma)\|_{0,\Omega}^2+\|\nabla\,\eta_p(t+\sigma)\|_{0,\Omega}^2\\ &\quad+\frac{2}{\tau}\Big(\eta_p(t+\sigma),\int_0^\sigma \text{ div }\ (\eta_u(t+\rho)\,d\rho\Big)_{0,\Omega}+2(\eta_u(t+\sigma),\nabla\,\eta_p(t+\sigma))_{0,\Omega}\Big)\ d\sigma\,.\end{aligned}$$

Aus der Cauchy-Schwarzschen Ungleichung folgt

$$\begin{aligned}\frac{2}{\tau}\Big(\eta_p(t+\sigma),\int_0^\sigma \text{ div }\ (\eta_u(t+\rho)\,d\rho\Big)_{0,\Omega}&\\ \le\ \frac{1}{\tau}\Big(\|\eta_p(t+\sigma)\|_{0,\Omega}^2+\|\int_0^\sigma \text{ div }\ \eta_u(t+\rho)\,d\rho\|_{0,\Omega}^2\Big)&\end{aligned}$$

und

$$2(\eta_u(t+\sigma),\nabla\,\eta_p(t+\sigma))_{0,\Omega}\le\ \|\eta_u(t+\sigma)\|_{0,\Omega}^2+\|\nabla\,\eta_p(t+\sigma)\|_{0,\Omega}^2\,.$$

Somit ist

$$\begin{aligned}
&\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), 0) \\
&\lesssim \int_0^\tau \left(\|\eta_u(t+\sigma)\|_{0,\Omega}^2 + \frac{1}{\tau}\| \int_0^\sigma \ \text{div}\ \ \eta_u(t+\sigma)\, d\rho\|_{0,\Omega}^2 \right. \\
&\qquad\qquad \left. + \frac{1}{\tau}\|\eta_p(t+\sigma)\|_{0,\Omega}^2 + \|\nabla\, \eta_p(t+\sigma)\|_{0,\Omega}^2 \right)\, d\sigma\,.
\end{aligned}$$

Nach Lemma 2.5 (mit $w = \ \text{div}\ \ \eta_u(t+\ \cdot\ )$) haben wir die Abschätzung nach oben in (2.29) gezeigt.
(iii) Um die Abschätzung nach unten zu beweisen, müssen wir zeigen, daß ein $\beta > 0$ existiert, so daß

$$\frac{\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t))}{|||(\eta_u, \eta_p)|||_\tau^2} \geq \beta \tag{2.31}$$

gleichmäßig bzgl. τ für alle $p_\tau(t) \in L^2(\Omega)$. Benutzen wir wieder, wie in Teil (ii), die Beschränktheit von c und a und die Tatsache, daß beide gleichmäßig von 0 weg beschränkt sind, so gilt $a \equiv c \equiv 1$.
Nach Lemma 2.2, Bemerkung 2.1 und Lemma 2.3 haben wir

$$\begin{aligned}
p_\tau(t+\sigma) &= \ \left(I + \frac{\sigma}{\tau}(r_{2,2}(\tau A) - I)\right)\ p_\tau(t)\,, \\
u_\tau(t+\sigma) &= \ -\nabla\ \left(r_{2,3}^-(\tau A) + \frac{\sigma}{\tau}(r_{0,1}^+(\tau A) - r_{2,3}^-(\tau A))\right)\ p_\tau(t)\,.
\end{aligned}$$

Der Konsistenzfehler kann deshalb wie folgt ausgedrückt werden:

$$\begin{aligned}
\eta_p(t+\sigma) &= \ \left(I + \frac{\sigma}{\tau}(r_{2,2}(\tau A) - I) - \exp(-\sigma A)\right) p_\tau(t)\,, \\
\eta_u(t+\sigma) &= \ \nabla \left(\exp(-\sigma A) - \left(\left(1 - \frac{\sigma}{\tau}\right)\ r_{2,3}^-(\tau A) + \frac{\sigma}{\tau}\ r_{0,1}^+(\tau A)\right)\right)\ p_\tau(t)\,.
\end{aligned} \tag{2.32}$$

Mit Hilfe von (2.32) ergibt dies

$$\begin{aligned}
&|||(\eta_u, \eta_p)|||_\tau^2 \\
&= \int_0^\tau \left(\left(A \left(\exp(-\sigma A) - \left(1 - \frac{\sigma}{\tau}\right)\ r_{2,3}^-(\tau A) - \frac{\sigma}{\tau}\ r_{0,1}^+(\tau A)\right)^2\ p_\tau(t), p_\tau(t)\right)_{0,\Omega} \right. \\
&\quad + \tau\ \left(A \left(\exp(-\sigma A) - \left(1 - \frac{\sigma}{\tau}\right)\ r_{2,3}^-(\tau A) - \frac{\sigma}{\tau}\ r_{0,1}^+(\tau A)\right)^2\ p_\tau(t), p_\tau(t)\right)_{0,\Omega} \\
&\quad + \frac{1}{\tau}\ \left(\left(I + \frac{\sigma}{\tau}(r_{2,2}(\tau A) - I) - \exp(-\sigma A)\right)^2 p_\tau(t), p_\tau(t)\right)_{0,\Omega} \\
&\quad \left. + \left(A \left(I + \frac{\sigma}{\tau}(r_{2,2}(\tau A) - I) - \exp(-\sigma A)\right)^2 p_\tau(t), p_\tau(t)\right)_{0,\Omega} \right) d\sigma\,.
\end{aligned}$$

Nach Berechnung dieser vier Integrale, erhalten wir

$$\begin{aligned}
|||(\eta_u, \eta_p)|||_\tau^2 &= \ (g(\tau A)\, p_\tau(t), p_\tau(t))_{0,\Omega}\,, \ \text{wobei} \\
g(z) &= \ -\frac{1}{36}\frac{g_0(z) + g_1(z)e^{-z} + g_2(z)e^{-2z}}{z\,(1+z)\ \left(1 + z + \frac{1}{3}z^2\right)^2}\ \text{mit}
\end{aligned}$$

$$
\begin{aligned}
g_0(z) &= -3\,z^7 - 11\,z^6 - z^5 + 36\,z^4 + 6\,z^3 - 96\,z^2 - 90\,z - 18\,,\\
g_1(z) &= -4\,z^6 - 12\,z^5 + 4\,z^4 + 60\,z^3 + 84\,z^2 + 36\,z\,,\\
g_2(z) &= 2\,z^7 + 18\,z^6 + 72\,z^5 + 164\,z^4 + 228\,z^3 + 192\,z^2 + 90\,z + 18\,.
\end{aligned}
$$

Da A^{-1} ein kompakter Operator ist, existiert ein Orthonormalsystem $\{\Phi_k\}_{k\in\mathbb{N}}$ von $L^2(\Omega)$ aus Eigenfunktionen von A (vgl. Bemerkung B.2.a,b). Wir stellen nun $p_\tau(t)$ mit Hilfe von $\{\Phi_k\}_{k\in\mathbb{N}}$ dar. Dann ist (2.31) erfüllt, falls ein $\tilde{\beta} > 0$ existiert mit

$$
\frac{r_{4,3}^{\mathcal{F}}(\tau\lambda_k)}{g(\tau\lambda_k)} \geq \tilde{\beta} \text{ für } k = 1, 2, \ldots
$$

gleichmäßig für alle $\tau > 0$. Deshalb betrachten wir die Funktion

$$
\frac{r_{4,3}^{\mathcal{F}}(x)}{g(x)} = -3\frac{x^5(1 + x + \frac{1}{3}x^2)}{g_0(x) + g_1(x)e^{-x} + g_2(x)e^{-2x}}
$$

für $x > 0$. Aus Teil (i) und Teil (ii) des Beweises wissen wir, daß diese Funktion beschränkt und nichtnegativ ist. Ferner gibt es keine positiven Nullstellen. Bildet man nun noch den Grenzwert für $x \to 0$ bzw. $x \to \infty$, so ergibt sich

$$
\lim_{x\to 0} \frac{r_{4,3}^{\mathcal{F}}(x)}{g(x)} = 10 \text{ und } \lim_{x\to\infty} \frac{r_{4,3}^{\mathcal{F}}(x)}{g(x)} = \frac{1}{3}\,,
$$

womit die Behauptung bewiesen ist.

■

Das wichtigste Resultat aus Satz 2.6 ist, daß das Auswerten des LSF einen a-posteriori Fehlerschätzer für den Konsistenzfehler bietet.

2.1.4 Konvergenztheorie

In diesem Abschnitt wollen wir eine Konvergenztheorie für die Einschrittmethode entwickeln, die sich aus der Variationsformulierung (2.15) ergibt. Zunächst zeigen wir einige wichtige Eigenschaften der rationalen Funktion, die den Evolutionsoperator $E_d(t+\tau;t)$ definiert.

Lemma 2.7 Die rationale Funktion $r_{2,2}$ erfüllt

$$
r_{2,2}(z) = 1 - z + \frac{1}{2}z^2 - \frac{1}{6}z^3 + \mathcal{O}(z^5) \quad \text{und} \tag{2.33}
$$

$$
|r_{2,2}(z)| \leq 1 \text{ für alle } z \text{ mit } \mathrm{Re}z \geq 0\,. \tag{2.34}
$$

Beweis:
Die Taylorentwicklung von $r_{2,2}(z)$ an der Stelle $z = 0$ führt direkt auf (2.33). Die Stabilität (2.34) folgt aus der Tatsache, daß $r_{2,2}$ eine holomorphe Funktion in der rechten Halbebene ist, wobei

$$
\lim_{z\to\infty} |r_{2,2}(z)| = \frac{1}{2} \text{ und}
$$

$$|r_{2,2}(iy)| = \left(\frac{\left(1+\frac{1}{6}y^2\right)^2}{\left(1-\frac{1}{3}y^2\right)^2+y^2} \right)^{1/2} = \left(\frac{1+\frac{1}{3}y^2+\frac{1}{36}y^4}{1+\frac{1}{3}y^2+\frac{1}{9}y^4} \right)^{1/2} \leq 1$$

gilt. Das Maximumprinzip liefert das gewünschte Resultat.

■

Als erstes zeigen wir die Konvergenz an den Punkten t_n für $n = 1, \ldots, M$.

Satz 2.8 Sei $p \in C^\infty((0,T]; H^1_{\Gamma_D}(\Omega)) \cap C([0,T]; L^2(\Omega))$ die exakte Lösung der Anfangs-Randwertaufgabe (2.1). Sei $(u_\tau, p_\tau) \in V_\tau((0,T); H_{\Gamma_N}(\text{ div },\Omega)) \times Q_\tau((0,T); H^1_{\Gamma_D}(\Omega))$ die Lösung des Variationsproblems (2.15). mit den Zeitschrittweiten $\tau_1, \ldots, \tau_M$ und Zeitschritten $t_n = t_{n-1} + \tau_n$ für $n = 1, \ldots, M$. Dann gelten:

$$\|p(t_n) - p_\tau(t_n)\|_{0,\Omega} \lesssim \left(\sum_{j=1}^{n} \tau_j^4 \int_{t_{j-1}}^{t_j} \|\partial_t^3 p\|_{0,\Omega}^2 \, d\sigma \right)^{1/2}, \tag{2.35}$$

$$\|\nabla(p(t_n) - p_\tau(t_n))\|_{0,\Omega} \lesssim \left(\sum_{j=1}^{n} \tau_j^3 \int_{t_{j-1}}^{t_j} \|\partial_t^3 p\|_{0,\Omega}^2 \, d\sigma \right)^{1/2}. \tag{2.36}$$

Beweis:
Wir betrachten einen beliebigen Zeitschritt von t auf $t+\tau$. Den Approximationsfehler teilen wir auf in den Konsistenzfehler und die Propagation des alten Fehlers (zum Zeitpunkt t) durch die diskrete Evolution. Wir führen den Beweis mit den folgenden Normen durch:

$$\| \cdot \|_{0,c,\Omega} = \|c^{1/2}(\cdot)\|_{0,\Omega},$$

$$\| \cdot \|_{0,a,\Omega} = \|a^{1/2}(\cdot)\|_{0,\Omega},$$

die beide zu $\|\cdot\|_{0,\Omega}$ äquivalent sind.
(i) Zunächst zeigen wir (2.35). Es gilt

$$\begin{aligned} \|p(t+\tau) - p_\tau^+\|_{0,c,\Omega} &= \|E(t+\tau;t)p(t) - E_d(t+\tau;t)p_\tau(t)\|_{0,c,\Omega} \\ &\leq \|(E(t+\tau;t) - E_d(t+\tau;t))p(t)\|_{0,c,\Omega} + \|E_d(t+\tau;t)(p(t) - p_\tau(t))\|_{0,c,\Omega}. \end{aligned} \tag{2.37}$$

Um die Terme auf der rechten Seite abzuschätzen, benutzen wir Eigenfunktionen $\{\Phi_k\}_{k\in\mathbb{N}}$ von A, die ein Orthonormalsystem von $L^2(\Omega)$ bilden (vgl. das Ende des Beweises von Satz 2.6). Wir stellen $c^{1/2}p(t)$ bzgl. dieses Systems wie folgt dar:

$$c^{1/2}p(t) = \sum_{k=1}^{\infty} \beta_k \, \Phi_k \, . \tag{2.38}$$

Damit erhalten wir für den ersten Term in (2.37)

$$\begin{aligned}
\|(E(t+\tau;t) - E_d(t+\tau;t))p(t)\|_{0,c,\Omega} &= \|(e^{-\tau A} - r_{2,2}(\tau A))\, c^{1/2} p(t)\|_{0,\Omega} \\
&= \left(\sum_{k=1}^{\infty} \beta_k^2 \, (e^{-\tau\lambda_k} - r_{2,2}(\tau\lambda_k))^2\right)^{1/2} \\
&\underbrace{\lesssim}_{(*)} \left(\sum_{k=1}^{\infty} \beta_k^2 \, \frac{1}{2}\, \tau^5 \, \lambda_k^5 \, (1 - e^{-2\tau\lambda_k})\right)^{1/2} \\
&= \left(\sum_{k=1}^{\infty} \beta_k^2 \, \tau^5 \, \lambda_k^6 \int_0^{\tau} e^{-2\sigma\lambda_k} \, d\sigma\right)^{1/2} \\
&= \tau^{5/2} \left(\int_0^{\tau} \|A^3 \exp(-\sigma A)\, c^{1/2} p(t)\|_{0,\Omega}^2 d\sigma\right)^{1/2} \\
&\lesssim \tau^{5/2} \left(\int_0^{\tau} \left\| \frac{d^3}{d\sigma^3} \, E(t+\sigma;t)\, p(t)\right\|_{0,\Omega}^2 d\sigma\right)^{1/2} .
\end{aligned}$$

Wir haben in $(*)$ ausgenutzt, daß die Funktion

$$\bar{G}(x) = 2\frac{(e^{-x} - r_{2,2}(x))^2}{x^5(1 - e^{-2x})} \tag{2.39}$$

auf $(0,\infty)$ beschränkt ist. Dies folgt aus der Stetigkeit von $\bar{G}$ auf $(0,\infty)$ und der Tatsache, daß

$$\lim_{x\to 0} \bar{G}(x) = 0\,, \ \lim_{x\to\infty} \bar{G}(x) = 0\,.$$

Außerdem ist

$$\frac{d^3}{d\sigma^3} \, E(t+\sigma;t)\, p(t) = -c^{-1/2} \, A^3 \, \exp(-\sigma A)\, c^{1/2}\, p(t)\,,$$

was direkt aus (2.8) folgt. Um den zweiten Term in (2.37) abzuschätzen, stellen wir $c^{1/2}\,(p(t) - p_\tau(t))$ wie folgt dar:

$$c^{1/2}\,(p(t) - p_\tau(t)) = \sum_{k=1}^{\infty} \gamma_k \, \Phi_k\,.$$

Wir folgern aus (2.34), daß

$$\begin{aligned}
\|E_d(t+\tau;t)\,(p(t) - p_\tau(t))\|_{0,c,\Omega} &= \|r_{2,2}(\tau A)\, c^{1/2}\,(p(t) - p_\tau(t))\|_{0,\Omega} \\
&= \left(\sum_{k=1}^{\infty} \gamma_k^2 \, (r_{2,2}(\tau\lambda_k))^2\right)^{1/2} \le \left(\sum_{k=1}^{\infty} \gamma_k^2\right)^{1/2} = \|p(t) - p_\tau(t)\|_{0,c,\Omega}\,.
\end{aligned}$$

Summation über alle Zeitschritte und Anwendung der Cauchy-Schwarzschen Ungleichung liefert

$$\|p(t_n) - p_\tau(t_n)\|_{0,c,\Omega} \lesssim \sum_{j=1}^{n} \tau_j^{5/2} \left(\int_{t_{j-1}}^{t_j} \left\|\partial_t^3 p\right\|_{0,\Omega}^2 \, d\sigma\right)^{1/2} \lesssim \left(\sum_{j=1}^{n} \tau_j^4 \int_{t_{j-1}}^{t_j} \left\|\partial_t^3 p\right\|_{0,\Omega}^2 \, d\sigma\right)^{1/2},$$

wobei $p(0) = p_\tau(0)$ und $t_n \leq T$. Damit ist (2.35) bewiesen.
(ii) Um (2.36) zu zeigen, teilen wir den Fehler wie in (i) in zwei Teile auf:

$$\begin{aligned}\|\nabla(p(t+\tau)-p_\tau^+)\|_{0,a,\Omega} \\ &= \|\nabla(E(t+\tau;t)p(t) - E_d(t+\tau;t)p_\tau(t))\|_{0,a,\Omega} \\ &\leq \|\nabla(E(t+\tau;t) - E_d(t+\tau;t))p(t)\|_{0,a,\Omega} \\ &+ \|\nabla E_d(t+\tau;t)(p(t) - p_\tau(t))\|_{0,a,\Omega}\,.\end{aligned} \tag{2.40}$$

Für den ersten Term in (2.40) haben wir

$$\begin{aligned}\|\nabla(E(t+\tau;t) - E_d(t+\tau;t))p(t)\|_{0,a,\Omega} \\ &= \|a^{1/2}\,\nabla\,c^{-1/2}\,(e^{-\tau A} - r_{2,2}(\tau A))\,c^{1/2}p(t)\|_{0,\Omega} \\ &= \|A^{1/2}\,(e^{-\tau A} - r_{2,2}(\tau A))\,c^{1/2}p(t)\|_{0,\Omega} \\ &= \left(\sum_{k=1}^{\infty}\beta_k^2\,\lambda_k\,(e^{-\tau\lambda_k} - r_{2,2}(\tau\lambda_k))^2\right)^{1/2} \\ &\underbrace{\lesssim}_{(**)}\left(\sum_{k=1}^{\infty}\beta_k^2\,\frac{1}{2}\,\tau^4\,\lambda_k^5\,(1-e^{-2\tau\lambda_k})\right)^{1/2} \\ &= \left(\sum_{k=1}^{\infty}\beta_k^2\,\tau^4\,\lambda_k^6\int_0^\tau e^{-2\sigma\lambda_k}\,d\sigma\right)^{1/2} \\ &= \tau^2\left(\int_0^\tau \|A^3\,\exp(-\sigma A)\,c^{1/2}p(t)\|_{0,\Omega}^2 d\sigma\right)^{1/2} \\ &\lesssim \tau^2\left(\int_0^\tau \left\|\frac{d^3}{d\sigma^3}\,E(t+\sigma;t)\,p(t)\right\|_{0,\Omega}^2 d\sigma\right)^{1/2}.\end{aligned}$$

Hier ergibt sich $(**)$ daraus, daß die Funktion

$$2\frac{(e^{-x} - r_{2,2}(x))^2}{x^4(1-e^{-2x})} \tag{2.41}$$

auf $(0,\infty)$ beschränkt ist, was man analog zu (2.39) zeigen kann. Für den zweiten Term in (2.40) erhalten wir

$$\begin{aligned}\|\nabla\,E_d(t+\tau;t)\,(p(t)-p_\tau(t))\|_{0,a,\Omega} &= \|a^{1/2}\,\nabla\,r_{2,2}(\tau A)\,c^{1/2}\,(p(t)-p_\tau(t))\|_{0,\Omega} \\ = \left(\sum_{k=1}^{\infty}\alpha_k^2\,\lambda_k\,(r_{2,2}(\tau\lambda_k))^2\right)^{1/2} &\leq \left(\sum_{k=1}^{\infty}\alpha_k^2\,\lambda_k\right)^{1/2} = \|\nabla\,(p(t)-p_\tau(t))\|_{0,a,\Omega}\,,\end{aligned}$$

wobei

$$a^{1/2}\nabla\,(p(t)-p_\tau(t)) = \sum_{k=1}^{\infty}\alpha_k\Phi_k$$

gilt.

Summation über alle Zeitschritte führt zu

$$\|\nabla(p(t_n)-p_\tau(t_n))\|_{0,a,\Omega} \lesssim \sum_{j=1}^{n} \tau_j^2 \left(\int_{t_{j-1}}^{t_j} \left\|\partial_t^3 p\right\|_{0,\Omega}^2 d\sigma \right)^{1/2} \lesssim \left(\sum_{j=1}^{n} \tau_j^3 \int_{t_{j-1}}^{t_j} \left\|\partial_t^3 p\right\|_{0,\Omega}^2 d\sigma \right)^{1/2} .$$

■

Bemerkung 2.9

(a) Man beachte, daß Satz 2.8 nicht die bestmögliche Approximationsordnung für die diskrete Evolution (2.12) angibt. Wie man in (2.33) sieht, erhält man sogar die Konvergenzordnung drei. (2.35) und (2.36) reichen für unsere spätere Betrachtung allerdings aus.

(b) Man beachte, daß die Forderung $p \in C^\infty((0,T]; H^1_{\Gamma_D}(\Omega))$ für alle Lösungen von Typ (2.1) gilt (vgl. Bemerkung B.4.d).

Wir interessieren uns für die Konvergenz in der in Satz 2.6 definierten Norm, d.h. in der zum LSF äquivalenten Norm.

Lemma 2.10 Für die in (2.12) bzw. (2.23) definierten rationale Funktionen $r_{2,2}$ bzw. $r_{2,3}^-$ und $r_{0,1}^+$ existieren Konstanten C_p und C_u, so daß für die Funktionen

$$G_p(x) = \int_0^1 \left((1-s) + s\, r_{2,2}(x)\right)^2 ds,$$
$$G_u(x) = \int_0^1 x \, \left((1-s)\, r_{2,3}^-(x) + s r_{0,1}^+(x)\right)^2 ds$$

gilt:

$$0 \le G_p(x) \le C_p \,,\; 0 \le G_u(x) \le C_u \text{ für alle } x \in (0,\infty) \,. \tag{2.42}$$

Beweis:
Die rationalen Funktionen $r_{2,2}, r_{2,3}^-$ und $r_{0,1}^+$ haben keine Polstellen auf $(0,\infty)$, also sind G_p und G_u stetig auf $(0,\infty)$. Die Existenz der gesuchten Konstanten C_p und C_u, mit denen (2.42) gilt, folgt aus

$$\lim_{x\to 0} G_p(x) = 1 \,,\; \lim_{x\to\infty} G_u(x) = \int_0^1 \left(1 - \frac{3}{2}s\right)^2 ds = \frac{1}{4} \,,$$
$$\lim_{x\to 0} G_u(x) = 0 \,,\; \lim_{x\to\infty} G_u(x) = \int_0^1 \left(1 - \frac{3}{2}s\right)^2 ds = \frac{1}{4} \,.$$

■

Schließlich erhalten wir den wichtigen Konvergenzsatz:

Satz 2.11 Auf $L^2((0,T);H_{\Gamma_N}(\text{ div },\Omega)) \times H^1((0,T);H^1_{\Gamma_D}(\Omega))$ sei für jedes Gitter $\Delta \subset [0,T]$ die Norm $|||(\cdot,\cdot)|||$ wie folgt definiert:

$$|||(v,q)||| = \left(\sum_{j=1}^{M_\Delta} \int_{t_{j-1}^\Delta}^{t_j^\Delta} (\|v(t)\|_{0,\Omega}^2 + \tau_j^\Delta \,\| \text{ div } v(t)\|_{0,\Omega}^2 + \frac{1}{\tau_j^\Delta} \|q(t)\|_{0,\Omega}^2 + \|\nabla q(t)\|_{0,\Omega}^2) dt \right)^{1/2} . \tag{2.43}$$

Sei $(u_\tau^\Delta, p_\tau^\Delta) \in V_\tau^\Delta((0,T);H_{\Gamma_N}(\text{ div },\Omega)) \times Q_\tau^\Delta((0,T);H^1_{\Gamma_D}(\Omega))$ die zu Δ gehörige Lösung des Variationsproblems (2.15).
Falls $\max(M_\Delta - j)\tau_j^\Delta$ gleichmäßig bzgl. Δ beschränkt ist, d.h. $\max(M_\Delta - j)\tau_j^\Delta \le C$, wobei C unabhängig von Δ ist, dann gilt:

$$|||(u-u_\tau^\Delta, p-p_\tau^\Delta)||| \lesssim \left(\sum_{j=1}^{M_\Delta} (\tau_j^\Delta)^3 \int_{t_{j-1}^\Delta}^{t_j^\Delta} \left(\|\partial_t^2 p\|_{0,\Omega}^2 + \|\partial_t^3 p\|_{0,\Omega}^2 \right) d\sigma \right)^{1/2} . \tag{2.44}$$

Beweis:
(i) Wir schätzen zunächst $|||(u-u_\tau, p-p_\tau)|||_\tau$ auf einem beliebigen, aber fest gewählten Intervall $(t, t+\tau)$ ab. Da (u,p) die exakte Lösung der Anfangs-Randwertaufgabe ist, gilt:

$$\begin{aligned} p(t+\sigma) &= E(t+\sigma;t)\, p(t) = c^{-1/2} \exp(-\sigma A)\, c^{1/2} p(t) , \\ u(t+\sigma) &= -a \nabla c^{-1/2} \exp(-\sigma A)\, c^{1/2} p(t) \end{aligned} \tag{2.45}$$

(vgl. (2.8)). Für die Approximation (u_τ, p_τ) gilt andererseits:

$$\begin{aligned} p_\tau(t+\sigma) &= E_d(t+\sigma;t)\, p_\tau(t) = c^{-1/2} \left(\left(1 - \frac{\sigma}{\tau}\right) I + \frac{\sigma}{\tau} r_{2,2}(\tau A) \right) c^{1/2} p_\tau(t) , \\ u_\tau(t+\sigma) &= -a \nabla c^{-1/2} \left(\left(1 - \frac{\sigma}{\tau}\right) r_{2,3}^-(\tau A) + \frac{\sigma}{\tau} r_{0,1}^+(\tau A) \right) c^{1/2} p_\tau(t) \end{aligned} \tag{2.46}$$

(vgl. Lemma 2.2 und Lemma 2.3). Wiederum teilen wir den Approximationsfehler in zwei Teile auf. Seien

$$\begin{aligned} \tilde{p}_\tau(t+\sigma) &= E_d(t+\sigma;t)\, p(t) = c^{-1/2} \left(\left(1 - \frac{\sigma}{\tau}\right) I + \frac{\sigma}{\tau} r_{2,2}(\tau A) \right) c^{1/2} p(t) , \\ \tilde{u}_\tau(t+\sigma) &= -a \nabla c^{-1/2} \left(\left(1 - \frac{\sigma}{\tau}\right) r_{2,3}^-(\tau A) + \frac{\sigma}{\tau} r_{0,1}^+(\tau A) \right) c^{1/2} p(t) \end{aligned}$$

die diskreten Evolutionen der exakten Lösungen p und u vom Zeitpunkt t zum Zeitpunkt $t+\sigma$. Nach der Dreiecksungleichung gilt somit

$$|||(u-u_\tau, p-p_\tau)|||_\tau \le |||(u-\tilde{u}_\tau, p-\tilde{p}_\tau)|||_\tau + |||(\tilde{u}_\tau - u_\tau, \tilde{p}_\tau - p_\tau)|||_\tau . \tag{2.47}$$

Diese Terme werden im folgenden abgeschätzt.
(ii) Der erste Term in (2.47) ist der Konsistenzfehler, der mit dem Anfangswert $p(t)$ erzeugt wird. Satz 2.6 liefert

$$|||(u-\tilde{u}_\tau, p-\tilde{p}_\tau)|||_\tau \lesssim \mathcal{F}(\tilde{u}_\tau(t), \tilde{u}_\tau(t+\tau), \tilde{p}_\tau(t+\tau); p(t), 0)^{1/2} . \tag{2.48}$$

Stellen wir nun $c^{1/2}p(t)$ bzgl. des Orthonormalsystems (siehe Teil (iii) vom Beweis des Satz 2.6)) $\{\Phi_k\}_{k\in\mathbf{N}}$ der Eigenfunktionen von A wie in (2.38) dar, so erhalten wir aus Lemma 2.4

$$\begin{aligned}
&\mathcal{F}(\tilde{u}_\tau(t), \tilde{u}_\tau(t+\tau), \tilde{p}_\tau(t+\tau); p(t), 0) \\
&\quad = \ (c^{1/2}p(t), r^{\mathcal{F}}_{4,3}(\tau A)\, c^{1/2}p(t))_{0,\Omega} = \ \sum_{k=1}^{\infty} \beta_k^2\, r^{\mathcal{F}}_{4,3}(\tau\lambda_k) \\
&\quad \lesssim \sum_{k=1}^{\infty} (\tau\lambda_k)^3 \left(1 - e^{-2\tau\lambda_k}\right) \qquad (*) \\
&\quad = \ \tau^3 \left(c^{1/2}p(t), A^3(I - e^{-2\tau A})\, c^{1/2}p(t)\right)_{0,\Omega} \\
&\quad = \ \tau^3 \left(c^{1/2}p(t), A^4 \int_0^\tau e^{-2\sigma A}\, d\sigma\, c^{1/2}p(t)\right)_{0,\Omega} \\
&\quad = \ \tau^3 \int_0^\tau \|A^2 \exp(-\sigma A)\, c^{1/2}p(t)\|^2_{0,\Omega} = \ \tau^3 \int_0^\tau \left\| \frac{d^2}{d\sigma^2}\, c^{1/2}p(t+\sigma)\right\|^2_{0,\Omega} d\sigma\,.
\end{aligned}$$

In $(*)$ haben wir die Tatsache ausgenutzt, daß

$$G^{\mathcal{F}}(x) = \ \frac{r^{\mathcal{F}}_{4,3}(x)}{x^3(1 - e^{-2x})}$$

für $x \in (0, \infty)$ beschränkt ist. Dies folgt aus der Stetigkeit von $G^{\mathcal{F}}$ für $x \in (0, \infty)$ und aus

$$\lim_{x\to 0} G^{\mathcal{F}}(x) = \ \frac{1}{24}\,, \ \lim_{x\to\infty} G^{\mathcal{F}}(x) = \ 0\,.$$

(iii) Um den zweiten Term in (2.47) abzuschätzen, halten wir zunächst folgendes fest:

$$\begin{aligned}
\tilde{p}_\tau - p_\tau &= \ c^{-1/2} \left(\left(1 - \frac{\sigma}{\tau}\right)\, I + \frac{\sigma}{\tau}\, r_{2,2}(\tau A)\right)\, c^{1/2}(p(t) - p_\tau(t))\,, \\
\tilde{u}_\tau - u_\tau &= \ -a\,\nabla\, c^{-1/2} \left(\left(1 - \frac{\sigma}{\tau}\right)\, r^-_{2,3}(\tau A) + \frac{\sigma}{\tau}\, r^+_{0,1}(\tau A)\right)\, c^{1/2}(p(t) - p_\tau(t))\,.
\end{aligned}$$

Zusammen mit der Definition der Norm in (2.28) führt dies zu

$$\begin{aligned}
&|||(\tilde{u}_\tau - u_\tau, \tilde{p}_\tau - p_\tau)|||_\tau \\
&\quad \lesssim \left(\int_0^\tau \left\|A^{1/2}\left(\left(1 - \frac{\sigma}{\tau}\right)\, r^-_{2,3}(\tau A) + \frac{\sigma}{\tau}\, r^+_{0,1}(\tau A)\right)\, c^{1/2}(p(t) - p_\tau(t))\right\|^2_{0,\Omega} d\sigma\right. \\
&\quad + \int_0^\tau \tau\, \left\|A\, \left(\left(1 - \frac{\sigma}{\tau}\right)\, r^-_{2,3}(\tau A) + \frac{\sigma}{\tau}\, r^+_{0,1}(\tau A)\right)\, c^{1/2}(p(t) - p_\tau(t))\right\|^2_{0,\Omega} d\sigma \\
&\quad + \int_0^\tau \frac{1}{\tau}\, \left\|\left(\left(1 - \frac{\sigma}{\tau}\right)\, I + \frac{\sigma}{\tau}\, r_{2,2}(\tau A)\right)\, c^{1/2}(p(t) - p_\tau(t))\right\|^2_{0,\Omega} d\sigma \\
&\quad \left. + \int_0^\tau \left\|A^{1/2}\left(\left(1 - \frac{\sigma}{\tau}\right)\, I + \frac{\sigma}{\tau}\, r_{2,2}(\tau A)\right)\, c^{1/2}(p(t) - p_\tau(t))\right\|^2_{0,\Omega} d\sigma\right)^{1/2}.
\end{aligned}$$

Stellen wir wieder $c^{1/2}(p(t) - p_\tau(t))$ bzgl. des Orthonormalsystems aus Eigenfunktionen von A dar, so liefert uns Lemma 2.10:

$$\begin{aligned}
&|||(\tilde{u}_\tau - u_\tau, \tilde{p}_\tau - p_\tau)|||_\tau \\
&\quad \lesssim \left(\|p(t) - p_\tau(t)\|^2_{0,\Omega}\,, \ + \tau\, \|\nabla(p(t) - p_\tau(t))\|^2_{0,\Omega}\right)^{1/2}.
\end{aligned}$$

Damit haben wir gezeigt, daß

$$|||(u-u_\tau, p-p_\tau)|||_\tau^2 \lesssim \tau^3 \int_0^\tau \left\| \frac{d^2}{d\sigma^2}\, p(t+\sigma) \right\|_{0,\Omega}^2 d\sigma + \|p(t)-p_\tau(t)\|_{0,\Omega}^2 + \tau \|\nabla(p(t)-p_\tau(t))\|_{0,\Omega}^2 . \tag{2.49}$$

(iv) Aus der Definition der Norm $|||(\cdot,\cdot)|||$ wissen wir

$$|||(u-u_\tau^\Delta, p-p_\tau^\Delta)||| = \left(\sum_{j=1}^{M^\Delta} |||(u-u_\tau^\Delta, p-p_\tau^\Delta)|||_{(t_{j-1}^\Delta, t_j^\Delta)}^2 \right)^{1/2} , \tag{2.50}$$

wobei $|||(\cdot,\cdot)|||_{(t_{j-1}^\Delta, t_j^\Delta)}$ die Norm $|||(\cdot,\cdot)|||_{\tau_j^\Delta}$ aus (2.28) auf dem Interval $(t_{j-1}^\Delta, t_j^\Delta)$ bezeichnet. Aus (2.49) und Satz 2.8 können wir somit folgern, daß

$$|||(u-u_\tau^\Delta, p-p_\tau^\Delta)|||_{(t_{j-1}^\Delta, t_j^\Delta)}^2 \lesssim (\tau_j^\Delta)^3 \int_{t_{j-1}^\Delta}^{t_j^\Delta} \|\partial_t^2 p\|_{0,\Omega}^2 \, d\sigma + \sum_{i=1}^{j-1} \left((\tau_i^\Delta)^4 + \tau_j^\Delta (\tau_i^\Delta)^3 \right) \int_{t_{i-1}^\Delta}^{t_i^\Delta} \|\partial_t^3 p\|_{0,\Omega}^2 \, d\sigma .$$

Summation über alle Zeitschritte ergibt

$$\begin{aligned}
&|||(u-u_\tau^\Delta, p-p_\tau^\Delta)|||^2 \\
&\lesssim \sum_{j=1}^{M_\Delta} (\tau_j^\Delta)^3 \left(\int_{t_{j-1}^\Delta}^{t_j^\Delta} \|\partial_t^2 p\|_{0,\Omega}^2 \, d\sigma + \sum_{i=j+1}^{M_\Delta} (\tau_i^\Delta + \tau_j^\Delta) \int_{t_{j-1}^\Delta}^{t_j^\Delta} \|\partial_t^3 p\|_{0,\Omega}^2 \, d\sigma \right) \\
&= \sum_{j=1}^{M_\Delta} (\tau_j^\Delta)^3 \left(\int_{t_{j-1}}^{t_j} \|\partial_t^2 p\|_{0,\Omega}^2 \, d\sigma + \underbrace{(\tau_j^\Delta (M^\Delta - j) + t_M^\Delta - t_j^\Delta)}_{\le C+T} \int_{t_{j-1}^\Delta}^{t_j^\Delta} \|\partial_t^3 p\|_{0,\Omega}^2 \, d\sigma \right) \\
&\lesssim \sum_{j=1}^{M_\Delta} (\tau_j^\Delta)^3 \left(\int_{t_{j-1}^\Delta}^{t_j^\Delta} \|\partial_t^2 p\|_{0,\Omega}^2 + \|\partial_t^3 p\|_{0,\Omega}^2 \, d\sigma \right) .
\end{aligned}$$

■

Korollar 2.12 Ist $\max(M_\Delta - j)\tau_j^\Delta$ derart beschränkt wie in Satz 2.11 gefordert, dann gilt mit $\tau_\Delta = \max \tau_j^\Delta$:

$$|||(u-u_\tau^\Delta, p-p_\tau^\Delta)||| \lesssim (\tau^\Delta)^{3/2} \left(\int_0^T \left(\|\partial_t^2 p\|_{0,\Omega}^2 + \|\partial_t^3 p\|_{0,\Omega}^2 \right) d\sigma \right)^{1/2} . \tag{2.51}$$

Bemerkung 2.13
$\max(M_\Delta - j)\tau_j^\Delta$ ist sicherlich beschränkt wie in Satz 2.11 gefordert, wenn wir annehmen, daß

$$\underline{\tau} \le \min_{1 \le M_\Delta} \tau_j^\Delta$$

für die Folge von Zerlegungen gilt (für ein a priori gewähltes $\underline{\tau} > 0$). Diese Forderung ist aber viel zu streng, wenn wir bedenken, daß die Zeitschrittweiten adaptiv gewählt werden sollen. Mit einfachen Worten bedeutet diese Art von Beschränktheit, daß nur die Zeitschritte in der Nähe von T relativ zu den anderen groß sein dürfen. Diese Annahme ist vernünftig, da die Lösungen der parabolischen Probleme in der Regel nach einiger Zeit glatter werden.

2.1.5 Lösungen der elliptischen Probleme in jedem Zeitschritt

Stetigkeit und Koerzivität (bzgl. der Standard-Normen) der Bilinearform aus (2.14) kann man aus [14] folgern. Wir interessieren uns aber vielmehr für τ-unabhängige Äquivalenz bzgl. der mit τ gewichteten Normen.

Satz 2.14 Bzgl. der Norm

$$\begin{aligned}|||(v^-,v^+,q)||| = \Big(&\frac{\tau}{3}\|a^{-1/2}v^-\|^2_{0,\Omega} + \frac{\tau^2}{3}\|c^{-1/2}\mathrm{div}\, v^-\|^2_{0,\Omega} + \frac{\tau}{3}\|a^{-1/2}v^+\|^2_{0,\Omega}\\ &+\frac{\tau^2}{3}\|c^{-1/2}\mathrm{div}\, v^+\|^2_{0,\Omega} + \|c^{1/2}q\|^2_{0,\Omega} + \frac{\tau}{3}\|a^{1/2}\nabla q\|^2_{0,\Omega}\Big)^{1/2}\end{aligned} \tag{2.52}$$

genügt die Bilinearform (2.14) den folgenden Ungleichungen:

$$\begin{aligned}\mathcal{B}(v^-,v^+,q;v^-,v^+,q) &\geq \frac{1}{4}|||(v^-,v^+,q)|||^2\,,\\ \mathcal{B}(u^-,u^+,p;v^-,v^+,q) &\leq 2|||(u^-,u^+,p)||| \cdot |||(v^-,v^+,q)|||\,,\end{aligned} \tag{2.53}$$

für alle $(u^-,u^+,p),(v^-,v^+,q) \in H_{\Gamma_N}(\ \mathrm{div}\ ,\Omega)^2 \times H^1_{\Gamma_D}(\Omega)$.

Beweis:
Aus dem Gaußschen Integralsatz folgt

$$(v^-,\nabla q)_{0,\Omega} + (\ \mathrm{div}\ v^-,q)_{0,\Omega} = (v^+,\nabla q)_{0,\Omega} + (\ \mathrm{div}\ v^+,q)_{0,\Omega} = 0\,,$$

für alle $v^-,v^+ \in H_{\Gamma_N}(\ \mathrm{div}\ ,\Omega)$ und $q \in H^1_{\Gamma_D}(\Omega)$. Zusammen mit der Definition der Bilinearform in (2.14) liefert dies

$$\begin{aligned}\mathcal{B}(v^-,v^+,q;v^-,v^+,q) &= \mathcal{B}(v^-,v^+,q;v^-,v^+,q)\\ &\quad -\frac{\tau}{3}\left((v^-,\nabla q)_{0,\Omega} + (\ \mathrm{div}\ v^-,q)_{0,\Omega}\right) - \frac{\tau}{3}\left((v^+,\nabla q)_{0,\Omega} + (\ \mathrm{div}\ v^+,q)_{0,\Omega}\right)\\ &= (\ \mathbf{z}\ ,\begin{pmatrix} \frac{\tau}{3} & 0 & \frac{\tau}{6} & 0 & 0 & -\frac{\tau}{6}\\ 0 & \frac{\tau^2}{3} & 0 & \frac{\tau^2}{6} & \frac{\tau}{6} & 0\\ \frac{\tau}{6} & 0 & \frac{\tau}{3} & 0 & 0 & 0\\ 0 & \frac{\tau^2}{6} & 0 & \frac{\tau^2}{3} & \frac{\tau}{6} & 0\\ 0 & \frac{\tau}{6} & 0 & \frac{\tau}{6} & 1 & 0\\ -\frac{\tau}{6} & 0 & 0 & 0 & 0 & \frac{\tau}{3}\end{pmatrix}\mathbf{z}\ )_{0,\Omega},\end{aligned}$$

mit $\mathbf{z} = (a^{-1/2}v^-\ ,\ c^{-1/2}\ \mathrm{div}\ v^-\ ,\ a^{-1/2}v^+\ ,\ c^{-1/2}\ \mathrm{div}\ v^+\ ,\ c^{1/2}q\ ,\ a^{1/2}\nabla q)^T$. Die obere bzw. untere Schranke in (2.53) hängen vom kleinsten bzw. größten Eigenwert des

verallgemeinerten Eigenwertproblems

$$\begin{pmatrix} \frac{\tau}{3} & 0 & \frac{\tau}{6} & 0 & 0 & -\frac{\tau}{6} \\ 0 & \frac{\tau^2}{3} & 0 & \frac{\tau^2}{6} & \frac{\tau}{6} & 0 \\ \frac{\tau}{6} & 0 & \frac{\tau}{3} & 0 & 0 & 0 \\ 0 & \frac{\tau^2}{6} & 0 & \frac{\tau^2}{3} & \frac{\tau}{6} & 0 \\ 0 & \frac{\tau}{6} & 0 & \frac{\tau}{6} & 1 & 0 \\ -\frac{\tau}{6} & 0 & 0 & 0 & 0 & \frac{\tau}{3} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \end{pmatrix} = \lambda \begin{pmatrix} \frac{\tau}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\tau^2}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\tau}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\tau^2}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\tau}{3} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \end{pmatrix}$$

ab. Diese verallgemeinerten Eigenwerte sind durch die Eigenwerte der Matrix

$$\begin{pmatrix} 1 & 0 & \frac{1}{2} & 0 & 0 & -\frac{1}{2} \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2\sqrt{3}} & 0 \\ \frac{1}{2} & 0 & 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 1 & \frac{1}{2\sqrt{3}} & 0 \\ 0 & \frac{1}{2\sqrt{3}} & 0 & \frac{1}{2\sqrt{3}} & 1 & 0 \\ -\frac{1}{2} & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

gegeben, die in dem Intervall $[1/4, 2]$ enthalten sind.

■

Satz 2.14 liefert, daß die Lösung der Variationsformulierung (2.56) quasi-optimal ist: Denn nach dem Céa-Lemma (Satz 1.2) folgt

$$\begin{aligned} &|||(u_\tau^- - u_{\tau,h}^-, u_\tau^+ - u_{\tau,h}^+, p_\tau^+ - p_{\tau,h}^+)||| \\ &\qquad \leq \sqrt{8} \inf_{v_h^- \in V_h, v_h^+ \in V_h, q_h \in Q_h} |||(u_\tau^- - v_h^-, u_\tau^+ - v_h^+, p_\tau^+ - q_h)||| \,. \end{aligned} \tag{2.54}$$

Eine andere Interpretation ist, daß die in (2.14) definierte Bilinearform die Norm

$$\mathcal{B}(v^-, v^+, q; v^-, v^+, q)^{1/2} =: |||(v^-, v^+, q)|||_{\mathcal{B}} \tag{2.55}$$

definiert, die zu der Norm $|||(\,\cdot\,,\,\cdot\,,\,\cdot\,)|||$ auf $H_{\Gamma_N}(\,\mathrm{div}\,,\Omega)^2 \times H^1_{\Gamma_D}(\Omega)$ äquivalent ist.

2.2 Der volldiskrete Fall

In diesem Abschnitt wollen wir den volldiskreten Fall betrachten und einige Beispiele berechnen. Nach Satz 2.14 liegt in jedem Zeitschritt eine stetige und koerzive Bilinearform vor. Der Satz von Lax-Milgram liefert die Existenz von Lösungen in jedem Zeitschritt.

2.2.1 LS-Galerkin-Formulierung bzgl. der Zeit und des Ortes

Um eine Approximation in jedem Zeitschritt zu berechnen, müssen wir die Minimierung des LSF (bzw. die dazu äquivalente Variationsformulierung (2.15)) in endlichdimensionalen Unterräumen $V_h \subseteq H_{\Gamma_N}(\,\mathrm{div}\,,\Omega)$ und $Q_h \subseteq H^1_{\Gamma_D}(\Omega)$ durchführen, die auf einer Triangulierung $\mathcal{T}_h$ von Ω basieren. Wir entscheiden uns in dieser Arbeit für die Räume

$$Q_h = \left\{ q_h \in H^1_{\Gamma_D}(\Omega) \;\middle|\; \forall\, T \in \mathcal{T}_h \; q_h\big|_T \text{ linear} \right\} \text{ (Standard-FE Raum)}$$

und

$$V_h = \left\{ v_h \in H_{\Gamma_N}(\text{div }, \Omega) \;\middle|\; \forall\, T \in \mathcal{T}_h \; q_h\big|_T = \begin{pmatrix} \alpha_T \\ \beta_T \end{pmatrix} + \gamma_T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right\}$$

(Raum der Raviart-Thomas-Elemente niedrigster Ordnung).
Da wir uns als erstes für adaptiv verfeinerte Triangulierungen interessieren, sind quasi-uniforme Triangulierungen nicht geeignet. Im Folgenden sollen die betrachteten Triangulierungen nicht-entartet (englisch: *shape regular* vgl. [11, Definition 5.1.(3)]) sein.
Somit ist die Minimierung in den endlichdimensionalen Unterräumen äquivalent zu der folgenden Variationsformulierung: Finde $(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}) \in V_h^2 \times Q_h$, so daß

$$\begin{aligned} \mathcal{B}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; v^-_h, v^+_h, q_h) = \; & (c^{1/2}\, p_\tau(t) + \tau\, c^{-1/2}\, f, c^{1/2}\, q_h)_{0,\Omega} \\ & + \frac{\tau}{2}(c^{1/2}\, p_\tau(t) + \tau\, c^{-1/2}\, f, c^{-1/2}\;\text{ div }\; v^-_h)_{0,\Omega} \\ & + \frac{\tau}{2}(c^{1/2}\, p_\tau(t) + \tau\, c^{-1/2}\, f, c^{-1/2}\;\text{ div }\; v^+_h)_{0,\Omega} \\ & - \frac{\tau}{6}\,(a^{1/2}\nabla p_\tau(t), a^{-1/2}(2v^-_h + v^+_h))_{0,\Omega} - \frac{\tau}{6}\,(a^{1/2}\nabla p_\tau(t), a^{1/2}\nabla q_h)_{0,\Omega} \end{aligned} \tag{2.56}$$

für alle $v^-_h, v^+_h \in V_h$ und $q_h \in Q_h$ gilt.
Die Existenz und Eindeutigkeit der Lösung von (2.56) folgt aus Satz 2.14.

Satz 2.15 Der Fehler der Zeitdiskretisierung und der Fehler der Ortsdiskretisierung sind orthogonal. Noch genauer:
Es gilt

$$\begin{aligned} \mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_\tau(t), f) = \; & \mathcal{F}(u^-_\tau, u^+_\tau, p^+_\tau; p_\tau(t), f) \\ & + |||(u^-_\tau - u^-_{\tau,h}, u^+_\tau - u^+_{\tau,h}, p^+_\tau - p^+_{\tau,h})|||^2_{\mathcal{B}}\,. \end{aligned} \tag{2.57}$$

Beweis:
Sei $p = \; E(t+\tau;t)p_\tau(t) \in H^1_{\Gamma_D}(\Omega)$ die exakte Evolution des parabolischen Problems und $u = \; -a\,\nabla p$ der dazugehörige Fluß. Wir definieren die folgende Bilinearform

$$\begin{aligned} \mathcal{B}_0(u,p;v,q) = \int_t^{t+\tau} & \big(\tau(c^{1/2}\partial_t p(s) + c^{-1/2}\;\text{ div }\; u(s)), c^{1/2}\partial_t q(s) + c^{-1/2}\;\text{ div }\; v(s))_{0,\Omega} \\ & + (a^{-1/2}u(s) + a^{1/2}\nabla p(s), a^{-1/2}v(s) + a^{1/2}\nabla q(s))_{0,\Omega}\big)\; ds, \end{aligned}$$

für alle $u, v \in L^2((0,T); H_{\Gamma_N}(\text{ div }, \Omega))$ und $p, q \in H^1((0,T); H^1_{\Gamma_D}(\Omega))$. Der Vergleich mit (2.5), (2.6) und (2.14) zeigt, daß

$$\begin{aligned} \mathcal{B}_0(u_\tau, p_\tau; v_\tau, q_\tau) = \; & \mathcal{B}(u^-_\tau, u^+_\tau, p^+_\tau; v^-_\tau, v^+_\tau, q^+_\tau) \\ & - (c^{1/2}p_\tau(t), c^{1/2}q^+_\tau + \frac{\tau}{2}\, c^{-1/2}(\;\text{ div }\; v^-_\tau + \;\text{ div }\; v^+_\tau))_{0,\Omega} \\ & + \frac{\tau}{6}(a^{1/2}\nabla p_\tau(t), a^{1/2}\nabla q^+_\tau + a^{-1/2}(2v^-_\tau + v^+_\tau))_{0,\Omega} \end{aligned} \tag{2.58}$$

mit

$$u_\tau(t+\sigma) = \; \Big(1 - \frac{\sigma}{\tau}\Big)\, u^-_\tau + \frac{\sigma}{\tau}\, u^+_\tau \quad , \quad v_\tau(t+\sigma) = \; \Big(1 - \frac{\sigma}{\tau}\Big)\, v^-_\tau + \frac{\sigma}{\tau}\, v^+_\tau,$$

für $u_\tau^-, u_\tau^+, v_\tau^-, v_\tau^+ \in H_{\Gamma_N}(\text{ div },\Omega)$ und

$$p_\tau(t+\sigma) = \left(1-\frac{\sigma}{\tau}\right) p_\tau(t) + \frac{\sigma}{\tau}\, p_\tau^+ \quad , \quad q_\tau(t+\sigma) = \frac{\sigma}{\tau}\, q_\tau^+,$$

für $p_\tau^+, q_\tau^+ \in H^1_{\Gamma_D}(\Omega)$. Ferner gilt

$$\mathcal{F}(v_\tau^-, v_\tau^+, q_\tau^+; p_\tau(t), f) = \mathcal{B}_0(u - v_\tau, p - q_\tau; u - v_\tau, p - q_\tau),$$

für alle $(v_\tau^-, v_\tau^+, q_\tau^+) \in H_{\Gamma_N}(\text{ div },\Omega)^2 \times H^1_{\Gamma_D}(\Omega)$. Aus Variationsformulierung (2.6) bzw. Variationsformulierung (2.15) und Gleichung (2.58) folgt, daß die Lösung (u_τ, p_τ) der Minimierungsaufgabe die Gleichung

$$\mathcal{B}_0(u - u_\tau, p - p_\tau; v_\tau, q_\tau) = 0$$

für alle $(v_\tau, q_\tau) \in V_\tau((0,T); H_{\Gamma_N}(\text{ div },\Omega)) \times Q_\tau((0,T); H^1_{\Gamma_D}(\Omega))$ erfüllt. Die zu beweisende Gleichung folgt dann aus

$$\begin{aligned}
&\mathcal{F}(u_{\tau,h}^-, u_{\tau,h}^+, p_{\tau,h}^+; p_\tau(t), f) \\
&\quad = \mathcal{B}_0(u - u_{\tau,h}, p - p_{\tau,h}; u - u_{\tau,h}, p - p_{\tau,h}) \\
&\quad = \mathcal{B}_0(u - u_\tau, p - p_\tau; u - u_\tau, p - p_\tau) \\
&\qquad + \mathcal{B}(u_\tau^- - u_{\tau,h}^-, u_\tau^+ - u_{\tau,h}^+, p_\tau^+ - p_{\tau,h}^+; u_\tau^- - u_{\tau,h}^-, u_\tau^+ - u_{\tau,h}^+, p_\tau^+ - p_{\tau,h}^+) \\
&\quad = \mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), f) + |||(u_\tau^- - u_{\tau,h}^-, u_\tau^+ - u_{\tau,h}^+, p_\tau^+ - p_{\tau,h}^+)|||_{\mathcal{B}}^2 ,
\end{aligned}$$

wobei wieder $u_{\tau,h}^- = u_{\tau,h}(t)$, $u_{\tau,h}^+ = u_{\tau,h}(t+\tau)$ und $p_{\tau,h}^+ = p_{\tau,h}(t+\tau)$. ∎

Aus Satz 2.15 folgt, daß das Funktional $\mathcal{F}(u_{\tau,h}^-, u_{\tau,h}^+, p_{\tau,h}^+; p_\tau(t), f)$ kleiner wird, falls wir eine genauere Lösung bzgl. des Ortes berechnen, d.h. falls wir die Räume V_h und Q_h erweitern (z.B. durch Verfeinerung der Triangulierung). Unsere Aufgabe besteht jetzt darin, einen a-posteriori Fehlerschätzer für den Ausdruck $|||(u_\tau^- - u_{\tau,h}^-, u_\tau^+ - u_{\tau,h}^+, p_\tau^+ - p_{\tau,h}^+)|||_{\mathcal{B}}^2$ zu finden.

2.2.2 Hierarchische Basis und a-posteriori Fehlerschätzer im Ort

Im Folgenden soll ein auf der hierarchischen Basis basierender Fehlerschätzer konstruiert werden. Die Grundidee besteht darin, das Funktional im Raum des hierarchischen Überschusses auf einer einmal uniform verfeinerten Triangulierung $\mathcal{T}_{h/2}$ zu minimieren. Dabei ist:

$$V_{h/2} = V_h \oplus Z_h \ , \ Q_{h/2} = Q_h \oplus Y_h.$$

Als Alternative kann man Polynome höherer Ordnung wie in [7] benutzen. Das Funktional wird dann wie folgt in den Räumen des hierarchischen Überschusses minimiert:

$$\min_{(z_h^-, z_h^+, y_h) \in Z_h^2 \times Y_h} \mathcal{F}(u_{\tau,h}^- + z_h^-, u_{\tau,h}^+ + z_h^+, p_{\tau,h}^+ + y_h; p_\tau(t), f) \, . \tag{2.59}$$

Die Variationsformulierung (2.59) ist äquivalent zu: Man finde $(d_h^-, d_h^+, e_h) \in Z_h^2 \times Y_h$ derart, daß

$$\begin{aligned}
&\mathcal{B}(d_h^-, d_h^+, e_h; z_h^-, z_h^+, y_h) \\
&\quad = (c^{1/2}\, p_\tau(t) + \tau\, c^{-1/2}\, f, c^{1/2}\, y_h + \frac{\tau}{2}\, c^{-1/2}\, (\text{ div } z_h^- + \text{ div } z_h^+))_{0,\Omega} \\
&\quad - \frac{\tau}{6}\, (a^{1/2} \nabla p_\tau(t), a^{-1/2}(2 z_h^- + z_h^+))_{0,\Omega} - \frac{\tau}{6}\, (a^{1/2} \nabla p_\tau(t), a^{1/2} \nabla y_h)_{0,\Omega} \\
&\quad - \mathcal{B}(u_{\tau,h}^-, u_{\tau,h}^+, p_{\tau,h}; z_h^-, z_h^+, y_h),
\end{aligned} \tag{2.60}$$

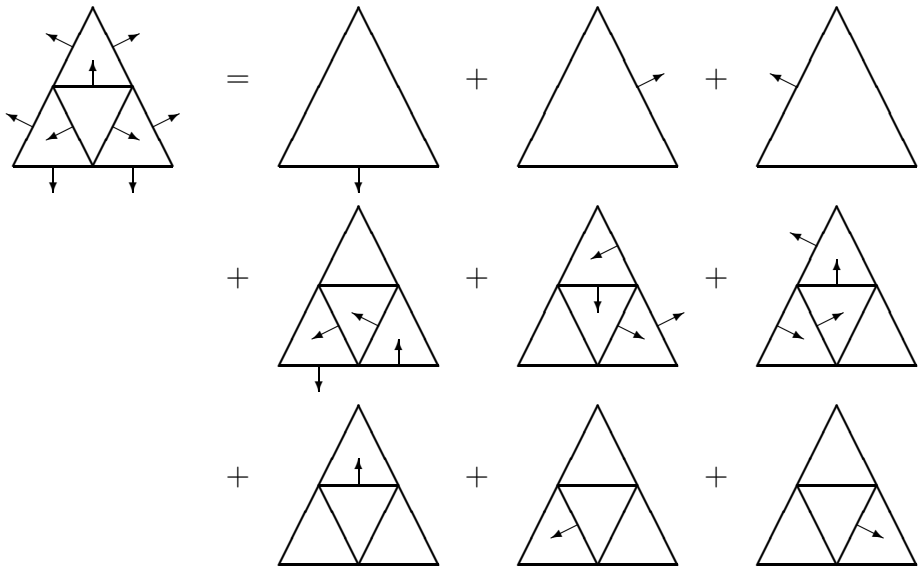

Abbildung 2.1: Hierarchische Basis für Raviart-Thomas Elemente niedrigster Ordnung

für alle $z_h^-, z_h^+ \in Z_h$ und $y_h \in Y_h$ gilt.
Später zeigt sich, daß eine geeignete Wahl der Basis für Z_h und Y_h in der Variationsformulierung (2.60) unabhängig von h und τ gut konditionierte lineare Gleichungssysteme hervorbringt. Ähnliches ist in [51] zu finden, wobei dort der Fehlerschätzer auf der Erweiterung der Räume durch Gebrauch von Polynomen höherer Ordnung basiert. Eine Skizze der hier verwendeten hierarchischen Basis liefert Abbildung 2.1. Die drei Basisfunktionen in der ersten Zeile spannen die Basis für den Raviart-Thomas-Raum (V_h) auf dem gröberen Dreieck. Die mittlere Zeile in Abbildung 2.1 spannt den divergenzfreien Anteil von Z_h auf, wobei die letzte Zeile die zweite zu einer Basis von Z_h ergänzt.
Für die Analyse des hierarchischen Fehlerschätzers werden noch einige Grundlagen benötigt (vgl. [48, Section 1.4]). Als erstes ist die verschärfte Cauchy-Schwarzsche Ungleichung, unabhängig von τ und h, zu zeigen.

Lemma 2.16 Es existiert ein $\gamma \in [0,1)$, so daß

$$\mathcal{B}(v_h^-, v_h^+, q_h; z_h^-, z_h^+, y_h) \leq \gamma |||(v_h^-, v_h^+, q_h)|||_{\mathcal{B}} |||(z_h^-, z_h^+, y_h)|||_{\mathcal{B}} \tag{2.61}$$

für alle $(v_h^-, v_h^+, q_h) \in V_h^2 \times Q_h$ und $(z_h^-, z_h^+, y_h) \in Z_h^2 \times Y_h$ gleichmäßig bzgl. $\tau, h > 0$ ist.

Beweis:
(i) Sei

$$\begin{aligned}\mathcal{A}(u^-, u^+, p; v^-, v^+, q) = {} & \tau(u^-, v^-)_{0,\Omega} + \tau^2(\text{ div } u^-, \text{ div } v^-)_{0,\Omega} \\ & + \tau(u^+, v^+)_{0,\Omega} + \tau^2(\text{ div } u^+, \text{ div } v^+)_{0,\Omega} + (p,q)_{0,\Omega} + \tau(\nabla p, \nabla q)_{0,\Omega}.\end{aligned}$$

Offenbar folgt aus Satz 2.14 und aus den Voraussetzungen an a und c

$$\underline{\alpha}\mathcal{A}(v^-, v^+, q; v^-, v^+, q) \leq \mathcal{B}(v^-, v^+, q; v^-, v^+, q) \leq \bar{\alpha}\mathcal{A}(v^-, v^+, q; v^-, v^+, q)$$

für alle $(v^-, v^+, q) \in H_{\Gamma_N}(\text{ div },\Omega)^2 \times H^1_{\Gamma_D}(\Omega)$ mit Konstanten $\bar{\alpha} \geq \underline{\alpha} > 0$, die nicht von τ abhängen.
Lemma 2.16 kann also formuliert werden als: Es existiert ein $\gamma \in [0,1)$, so daß

$$\mathcal{B}(v_h^-, v_h^+, q_h; z_h^-, z_h^+, y_h) \leq \gamma$$

für alle $(v_h^-, v_h^+, q_h) \in V_h^2 \times Q_h$ mit $|||(v_h^-, v_h^+, q_h)|||_{\mathcal{B}} = 1$ und $(z_h^-, z_h^+, y_h) \in Z_h^2 \times Y_h$ mit $|||(z_h^-, z_h^+, y_h)|||_{\mathcal{B}} = 1$. Mit diesen Annahmen folgt

$$\begin{aligned}|||(v_h^- + z_h^-, v_h^+ + z_h^+, q_h + y_h)|||_{\mathcal{B}}^2 |||(v_h^- - z_h^-, v_h^+ - z_h^+, q_h - y_h)|||_{\mathcal{B}}^2 \\ = 4(1 - \mathcal{B}(v_h^-, v_h^+, q_h; z_h^-, z_h^+, y_h)^2).\end{aligned}$$

Wir haben deshalb nur zu zeigen: Es gibt ein β_0 mit

$$|||(v_h^- + z_h^-, v_h^+ + z_h^+, q_h + y_h)|||_{\mathcal{B}} |||(v_h^- - z_h^-, v_h^+ - z_h^+, q_h - y_h)|||_{\mathcal{B}} \geq \beta_0$$

für alle $(v_h^-, v_h^+, q_h) \in V_h^2 \times Q_h$ mit $|||(v_h^-, v_h^+, q_h)|||_{\mathcal{B}} = 1$ und $(z_h^-, z_h^+, y_h) \in Z_h^2 \times Y_h$ mit $|||(z_h^-, z_h^+, y_h)|||_{\mathcal{B}} = 1$. Anders formuliert: Es gibt ein $\alpha_0 > 0$ mit

$$|||(v_h^- + z_h^-, v_h^+ + z_h^+, q_h + y_h)|||_{\mathcal{A}} |||(v_h^- - z_h^-, v_h^+ - z_h^+, q_h - y_h)|||_{\mathcal{A}} \geq \alpha_0$$

für alle $(v_h^-, v_h^+, q_h) \in V_h^2 \times Q_h$ mit $|||(v_h^-, v_h^+, q_h)|||_{\mathcal{A}} = 1$ und $(z_h^-, z_h^+, y_h) \in Z_h^2 \times Y_h$ mit $|||(z_h^-, z_h^+, y_h)|||_{\mathcal{A}} = 1$. Mit der gleichen Argumentation brauchen wir nur zu zeigen, daß es ein $\tilde{\gamma} \in [0,1)$ gibt mit

$$\mathcal{A}(v_h^-, v_h^+, q_h; z_h^-, z_h^+, y_h) \leq \tilde{\gamma} |||(v_h^-, v_h^+, q_h)|||_{\mathcal{A}} |||(z_h^-, z_h^+, y_h)|||_{\mathcal{A}}$$

für alle $(v_h^-, v_h^+, q_h) \in V_h^2 \times Q_h$ und $(z_h^-, z_h^+, y_h) \in Z_h^2 \times Y_h$.
(ii) Wir zeigen nun: Es gibt ein $\tilde{\gamma} < 1$, so daß

$$\begin{aligned}&\mathcal{A}_K(v_h^-, v_h^+, q_h; z_h^-, z_h^+, y_h) \\ &\quad \leq \tilde{\gamma} \mathcal{A}_K(v_h^-, v_h^+, q_h; v_h^-, v_h^+, q_h)^{1/2} \mathcal{A}_K(z_h^-, z_h^+, y_h; z_h^-, z_h^+, y_h)^{1/2},\end{aligned} \tag{2.62}$$

wobei $\mathcal{A}_K(\cdot,\cdot,\cdot;\cdot,\cdot,\cdot)$ die Bilinearform eingeschränkt auf einzelne Elemente ist, d.h.

$$\begin{aligned}\mathcal{A}_K(u^-, u^+, p; v^-, v^+, q) = \ &\tau(u^-, v^-)_{0,K} + \tau^2(\text{ div } u^-, \text{ div } v^-)_{0,K} \\ &+ \tau(u^+, v^+)_{0,K} + \tau^2(\text{ div } u^+, \text{ div } v^+)_{0,K} + (p,q)_{0,K} + \tau(\nabla p, \nabla q)_{0,K},\end{aligned}$$

für alle $K \in \mathcal{T}_h$. Es ist klar, daß

$$\begin{aligned}(q_h, y_h)_{0,K} &\leq \tilde{\gamma}_0 \|q_h\|_{0,K} \|y_h\|_{0,K}, \\ (\nabla q_h, \nabla y_h)_{0,K} &\leq \tilde{\gamma}_1 \|\nabla q_h\|_{0,K} \|\nabla y_h\|_{0,K}\end{aligned} \tag{2.63}$$

für alle $q_h \in Q_h$ und $y_h \in Y_h$ mit $\tilde{\gamma}_0, \tilde{\gamma}_1 < 1$ (vgl. [7]) gilt.
Benutzen wir die affine Transformation zu einem Referenzelement K_{ref} so, führt dies zu

$$\begin{aligned}(v_h, z_h)_{0,K} &= \frac{|K|}{|K_{\text{ref}}|} (v^{\text{ref}}, z^{\text{ref}})_{0,K_{\text{ref}}} \\ &\leq \tilde{\gamma}_2 \frac{|K|}{|K_{\text{ref}}|} \|v^{\text{ref}}\|_{0,K_{\text{ref}}} \|z^{\text{ref}}\|_{0,K_{\text{ref}}} = \tilde{\gamma}_2 \|v_h\|_{0,K} \|z_h\|_{0,K}\end{aligned} \tag{2.64}$$

für alle $v_h \in V_h$ und $z_h \in Z_h$. Die Ungleichung

$$(v^{\mathrm{ref}}, z^{\mathrm{ref}})_{0,K_{\mathrm{ref}}} \leq \tilde{\gamma_2} \|v^{\mathrm{ref}}\|_{0,K_{\mathrm{ref}}} \|z^{\mathrm{ref}}\|_{0,K_{\mathrm{ref}}}$$

für alle $v^{\mathrm{ref}} \in V^{\mathrm{ref}}$ und $z^{\mathrm{ref}} \in Z^{\mathrm{ref}}$, mit einer Konstante $\tilde{\gamma}_2 < 1$, folgt, weil die Räume V^{ref} und Z^{ref} endlichdimensionale Vektorräume sind (von Dimension drei bzw. sechs, vgl. Abbildung 2.1) und $V^{\mathrm{ref}} \cap Z^{\mathrm{ref}} = \{0\}$.
Schließlich ist $\operatorname{div} v_h$ konstant auf K für $v_h \in V_h$ (da v_h linear ist) und

$$\int_K \operatorname{div} z_h \, dx = \int_{\partial K} < n, z_h > \, ds = 0 \text{ für alle } z_h \in Z_h.$$

Für die Basisfunktionen gilt entweder $\operatorname{div} z_h = 0$ in K (mittlere Zeile in Abbildung 2.1) oder $< n, z_h >= 0$ auf ∂K (untere Zeile in Abbildung 2.1). Damit folgt

$$(\operatorname{div} v_h, \operatorname{div} z_h)_{0,K} = 0 \text{ für alle } v_h \in V_h \text{ und } z_h \in Z_h \,. \tag{2.65}$$

Zusammen folgt aus (2.63), (2.64), (2.65) und der Cauchy-Schwarzschen Ungleichung (für Summen) (2.62) mit $\tilde{\gamma} = \max\{\tilde{\gamma}_0, \tilde{\gamma}_1, \tilde{\gamma}_2\}$.
(iii) Nochmaliges Anwenden der Cauchy-Schwarzschen Ungleichung liefert

$$\begin{aligned}
\mathcal{A}(v_h^-, v_h^+, q_h; z_h^-, z_h^+, y_h) &= \sum_{K \in \mathcal{T}_h} \mathcal{A}_K(v_h^-, v_h^+, q_h; z_h^-, z_h^+, y_h) \\
&\leq \tilde{\gamma} \sum_{K \in \mathcal{T}_h} \mathcal{A}_K(v_h^-, v_h^+, q_h; v_h^-, v_h^+, q_h)^{1/2} \mathcal{A}_K(z_h^-, z_h^+, y_h; z_h^-, z_h^+, y_h)^{1/2} \\
&\leq \tilde{\gamma} \left(\sum_{K \in \mathcal{T}_h} \mathcal{A}_K(v_h^-, v_h^+, q_h; v_h^-, v_h^+, q_h) \right)^{1/2} \left(\sum_{K \in \mathcal{T}_h} \mathcal{A}_K(z_h^-, z_h^+, y_h; z_h^-, z_h^+, y_h) \right)^{1/2} \\
&= \tilde{\gamma} |||(v_h^-, v_h^+, q_h)|||_{\mathcal{A}} |||(z_h^-, z_h^+, y_h)|||_{\mathcal{A}}
\end{aligned}$$

für alle $(v_h^-, v_h^+, q_h) \in V_h^2 \times Q_h$ und $(z_h^-, z_h^+, y_h) \in Z_h^2 \times Y_h$.

■

Der hierarchische Fehlerschätzer ist definiert als(vgl. auch [48, Section 1.4]):

$$\eta_K = \mathcal{B}_K(d_h^-, d_h^+, e_h; d_h^-, d_h^+, e_h)^{1/2},$$

wobei (d_h^-, d_h^+, e_h) das Problem (2.60) löst und die Notation $\mathcal{B}_K$ aus dem vorangehenden Beweis stammt. Mit diesen Hilfsmitteln können wir nun zeigen, daß der obige Ausdruck ein Fehlerschätzer ist, d.h. er ist äquivalent zu der Fehlernorm. Dies geschieht unter der für diese Art von Fehlerschätzern üblichen *Saturierungsbedingung*.

Satz 2.17 Unter der Annahme, daß die Saturierungsbedingung

$$\begin{aligned}
\inf_{(v^-, v^+, q) \in V_{h/2}^2 \times Q_{h/2}} |||(u_\tau^- - v^-, u_\tau^+ - v^+, p_\tau^+ - q)|||_{\mathcal{B}} \\
\leq \beta \inf_{(v^-, v^+, q) \in V_h^2 \times Q_h} |||(u_\tau^- - v^-, u_\tau^+ - v^+, p_\tau^+ - q)|||_{\mathcal{B}}
\end{aligned} \tag{2.66}$$

mit $\beta < 1$ gleichmäßig bzgl. h und τ gilt, ist

$$\sum_{K \in \mathcal{T}_h} \eta_K^2 \eqsim |||(u_\tau^- - u_{\tau,h}^-, u_\tau^+ - u_{\tau,h}^+, p_\tau^+ - p_{\tau,h}^+)|||_{\mathcal{B}}^2. \tag{2.67}$$

Beweis:
Der Beweis folgt analog zum Beweis für den hierarchischen Standard-Fehlerschätzer (vgl. z.B. [48, Section 1.4]). Im wesentlichen werden dabei die verschärfte Cauchy-Schwarzsche Ungleichung (Lemma 2.16) und die Saturierungsbedingung (2.66) verwendet.

■

Satz 2.18 Die zu der Variationsformulierung (2.60) gehörige Matrix B ist äquivalent zu ihrer Diagonale $\mathrm{diag}(B)$, unabhängig von h und τ.

Beweis:
Schreiben wir zunächst $z_h \in Z_h$ bzw. $y_h \in Y_h$ als

$$z_h = \sum_\mu z_h^{(\mu)} \Psi_h^{(\mu)} \text{ und } y_h = \sum_\mu y_h^{(\mu)} \Phi_h^{(\mu)} .$$

Wir benutzen wieder die Bilinearform $\mathcal{A}(\cdot,\cdot,\cdot;\cdot,\cdot,\cdot)$ aus dem Beweis vom Lemma 2.16. Wegen der Äquivalenz der Bilinearformen genügt es, zu zeigen, daß

$$\begin{aligned}\mathcal{A}(z_h^-, z_h^+, y_h; z_h^-, z_h^+, y_h) &\eqsim \sum_\mu (z_h^{-,(\mu)})^2 \mathcal{A}(\Psi_h^{(\mu)}, 0, 0; \Psi_h^{(\mu)}, 0, 0) \\ &+ \sum_\mu (z_h^{+,(\mu)})^2 \mathcal{A}(0, \Psi_h^{(\mu)}, 0; 0, \Psi_h^{(\mu)}, 0) + \sum_\mu (y_h^{(\mu)})^2 \mathcal{A}(0, 0, \Phi_h^{(\mu)}; 0, 0, \Phi_h^{(\mu)}) .\end{aligned}$$

Die Matrix

$$[\mathcal{A}_K(\Psi_h^{(\mu)}, 0, 0; \Psi_h^{(\nu)}, 0, 0)]_{1\le\mu,\nu\le 6} = [\mathcal{A}_K(0, \Psi_h^{(\mu)}, 0; 0, \Psi_h^{(\nu)}, 0)]_{1\le\mu,\nu\le 6}$$

des Referenzdreiecks (mit den Punkten $(-\sqrt{3}/2, -1/2), (\sqrt{3}/2, -1/2), (0, 1)$ (das Dreieck aus Abbildung 2.1) hat die Form

$$\frac{\sqrt{3}}{48}\tau \begin{pmatrix} 36 & -12 & -12 & 0 & 12 & -12 \\ -12 & 36 & -12 & -12 & 0 & 12 \\ -12 & -12 & 36 & 12 & -12 & 0 \\ 0 & -12 & 12 & 10 & -1 & -1 \\ 12 & 0 & -12 & -1 & 10 & -1 \\ -12 & 12 & 0 & -1 & -1 & 10 \end{pmatrix} + \frac{\sqrt{3}}{3}\tau^2 \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8 & 4 & 4 \\ 0 & 0 & 0 & 4 & 8 & 4 \\ 0 & 0 & 0 & 4 & 4 & 8 \end{pmatrix} .$$

Daß diese Matrix äquivalent zu ihrer Diagonalen unabhängig von τ ist, kann man anhand der Berechnung ihrer Eigenwerte zeigen. Diese Äquivalenz ist aber auch von h unabhängig, weil eine von h unabhängige Transformation zwischen beliebigen Dreiecken - der von uns betrachteten (nicht-entarteten) Familie von Triangulierungen - und dem Referenzdreieck möglich ist. Das Analogon dazu für die Matrix

$$[\mathcal{A}_K(0, 0, \Phi_h^{(\mu)}; 0, 0, \Phi_h^{(\nu)})]_{1\le\mu,\nu\le 3}$$

bzgl. der hierarchischen Standard-Basis ist in [7] zu finden.

■

Aus Satz 2.18 folgt, daß die Lösung von (2.60) durch das CG-Verfahren mit Jacobi-Vorkonditionierung (vgl. [17, Abschnitt 8.3 und 8.4]) mit einer von τ und h unabhängig beschränkten Iterationsanzahl approximiert werden kann.

2.2.3 Zeit- und ortsadaptiver Algorithmus

In Abschnitt 2.1 wurde die Norm

$$|||(v,q)|||_{\tau_j} = \left(\int_{t_{j-1}}^{t_j} (\|v(\sigma)\|_{0,\Omega}^2 + \tau_j \,\|\, \mathrm{div}\;\; v(\sigma)\|_{0,\Omega}^2 + \frac{1}{\tau_j}\,\|q(\sigma)\|_{0,\Omega}^2 + \|\nabla\, q(\sigma)\|_{0,\Omega}^2) d\sigma \right)^{1/2} \tag{2.68}$$

definiert (mit der Zeitschrittweite $\tau_j = t_j - t_{j-1}$), bzgl. der wir den Diskretisierungsfehler messen. Will man $|||(u-u_\tau, p-p_\tau)|||_{\tau_j}^2 \leq (\texttt{tol})^2$ in jedem Zeitschritt für eine bestimmte Toleranzgrenze `tol` erreichen, so sollte

$$|||(u-u_\tau, p-p_\tau)|||_{\tau_j}^2 \eqsim \mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), f) \leq \;\; (\texttt{tol})^2$$

gelten. Ist diese Bedingung erfüllt, so akzeptieren wir die Approximation. Dann muß ein Vorschlag für die nächste Zeitschrittweite gefunden werden. Andernfalls wird eine neue Zeitschrittweite benötigt, um den aktuellen Schritt zu wiederholen. In den nachfolgenden Überlegungen sollen beide Fälle berücksichtigt werden und wir führen dazu die Bezeichnungen τ_{alt} und τ_{neu} ein, um Missverständnisse zu vermeiden. Sei

$$\mathrm{Error}_{\mathrm{alt}} = \; \mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), f)\,.$$

In beiden Fällen soll bei der nächsten Berechnung $\mathrm{Error}_{\mathrm{neu}} \leq \texttt{tol}$ erreicht werden. Um den Aufwand zu minimieren, fordern wir zusätzlich

$$\mathrm{Error}_{\mathrm{neu}} \approx \texttt{tol}\,. \tag{2.69}$$

Satz 2.11 besagt, daß

$$\mathrm{Error}_{\mathrm{alt}} \approx C_1 \tau_{\mathrm{alt}}^{3/2}\,,\; \mathrm{Error}_{\mathrm{neu}} \approx C_2 \tau_{\mathrm{neu}}^{3/2}\,. \tag{2.70}$$

Dividieren wir nun die erste Beziehung in (2.70) durch die zweite, dann können wir unter der Annahme $C_1 \approx C_2$ folgern:

$$\left(\frac{\tau_{\mathrm{neu}}}{\tau_{\mathrm{alt}}}\right)^{3/2} \approx \frac{\mathrm{Error}_{\mathrm{neu}}}{\mathrm{Error}_{\mathrm{alt}}}. \tag{2.71}$$

Beachten wir nun (2.69) und führen wir einen *Sicherheitsfaktor* $\delta \in (0,1)$ ein, um Endlosschleifen zu vermeiden, so ist

$$\tau_{\mathrm{neu}} = \left(\frac{\texttt{tol}}{\mathrm{Error}_{\mathrm{alt}}}\right)^{2/3} \tau_{\mathrm{alt}} \delta. \tag{2.72}$$

Ferner ist es sinnvoll, die Zeitschrittweite nach oben und nach unten zu beschränken. Die obige Strategie kann ausführlich in [16, Section 5.1] nachgelesen werden.
In jedem Zeitschritt starten wir von einer vorgegebenen Triangulierung $\mathcal{T}_0$ und konstruieren eine Folge von adaptiv verfeinerten Triangulierungen $\mathcal{T}_l$, $l = 1, 2, \ldots, l_{\max}$, die mit Hilfe des hierarchischen Fehlerschätzers gebildet werden. Die zugehörigen FE-Räume bezeichnen wir mit V_l und Q_l, für $l = \; 0, 1, 2 \ldots, l_{\max}$. Die Räume des hierarchischen

Überschusses auf $\mathcal{T}_l$ bezeichnen wir mit Z_l und Y_l, $l = 0, 1, 2 \ldots, l_{\max}$. Dies sollte so lange durchgeführt werden, bis der Ortsdiskretisierungsfehler gegenüber dem Gesamtfehler nahezu bedeutungslos ist. Da wir auf dem Computer nur begrenzte Ressourcen haben, sollte die Anzahl der *Levels* in jedem Zeitschritt durch $l_{\max}$ beschränkt werden. Mit dem von Benutzer vorgegebenen Parameter `tol` (Toleranzgrenze), γ (Shift-Parameter), $\delta \in (0,1)$ (Sicherheitsfaktor), $\tau_{\max}$ (maximale Zeitschrittweite) und $\tau_{\min}$ (minimale Zeitschrittweite) und τ_0 (Anfangszeitschrittweite) lautet der adaptive Algorithmus:

Algorithmus 2.19

t $= 0\,; \tau = \tau_0\,; p^{\text{alt}} = p_0\,;$

`while` $t < T$,

berechne $(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}) \in V_0^2 \times Q_0$, die Lösung der Variationsformulierung (2.56) ;

berechne $(d^-_h, d^+_h, e_h) \in Z_0^2 \times Y_0$, die Lösung der Variationsformulierung (2.60) ;

berechne $\eta_K = \mathcal{B}_K(d^-_h, d^+_h, e_h; d^-_h, d^+_h, e_h)^{1/2}$ für jedes $K \in \mathcal{T}_0\,; \eta_0^2 = \sum_{K \in \mathcal{T}_0} \eta_K^2$;

$l = 0\,;$

`while` $l \le l_{\max} \ \wedge \ \neg\,[\eta_l \ll \mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p^{\text{alt}}, f)]\,,$

$l = l + 1;$

$\mathcal{T}_l =$ verfeinere $\hat{K} \in \mathcal{T}_{l-1}$ mit $\eta_{\hat{K}} > \gamma \max_{K \in \mathcal{T}_{l-1}} \eta_K\,;$

berechne $(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}) \in V_l^2 \times Q_l$, die Lösung der Variationsformulierung (2.56) ;

berechne $(d^-_h, d^+_h, e_h) \in Z_l^2 \times Y_l$, die Lösung der Variationsformulierung (2.60) ;

berechne $\eta_K = \mathcal{B}_K(d^-_h, d^+_h, e_h; d^-_h, d^+_h, e_h)^{1/2}$ für jedes $K \in \mathcal{T}_l\,; \eta_l^2 = \sum_{K \in \mathcal{T}_l} \eta_K^2$;

`end`

$\text{Error}_{\text{est}} = \mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p^{\text{alt}}, f)^{1/2}\,;$

`if` $\text{Error}_{\text{est}} <$ `tol` %% akzeptiere diesen Schritt

$$\text{t} = \text{t} + \tau\,;\ \tau = \min\left\{\tau_{\max}, \left(\frac{\texttt{tol}}{\text{Error}_{\text{est}}}\right)^{2/3} \tau\delta\right\}\,;\ p^{\text{alt}} = p^+_{\tau,h};$$

`else` %% die Berechnung dieses Schritts ist nicht akzeptabel

$$\tau = \max\left(\tau_{\min}, \left(\frac{\texttt{tol}}{\text{Error}_{\text{est}}}\right)^{2/3} \tau\delta\right)\,;$$

`end`

`end`

Die Zeitschrittstrategie des obigen Algorithmus kann in vielen Büchern nachgelesen werden (z.B. [16, Section 5.1] oder [26, Section II.4]). Ein volldiskreter adaptiver Algorithmus, der auf der linear-impliziten Zeitdiskretisierung mit einem hierarchischen Fehlerschätzer im Ort basiert, wird in [31] beschrieben. Die LSM mit der Wahl der linearen Ansatzfunktionen in der Zeit führt zu einer Methode der dritten Ordnung (im klassischen Sinne bzgl. p). Eine andere Methode mit derselben Genauigkeit ist z.B. das diskontinuierliche Galerkin-Verfahren mit stückweise linearen Ansatzfunktionen bzgl. der Zeit (vgl. unter anderem den Abschnitt 1.2.3, [18] und [46, Chapter 12]). Andere bekannte Methoden, die adaptive Zeitschrittweiten hervorbringen, sind Runge-Kutta-Verfahren wie z.B. implizite Runge-Kutta-Verfahren vom Typ Radau IIA, die auch Methoden der Ordnung drei liefern und ihrerseits mit einem Fehlerschätzer kombiniert werden können (vgl. z.B. [27, Section IV.5]).

2.2.4 Numerische Berechnungen

In diesem Abschnitt sollen einige numerische Beispiele die theoretischen Aussagen aus dem Abschnitt 2.1 unterstreichen. Der Einfachheit halber soll in allen Beispielen $a \equiv c \equiv 1$ gelten.

Beispiel 2.20 Auf $\Omega = (-1,1)^2$ berechnen wir eine Approximation zu der Lösung von (2.1) mit der rechten Seite $f \equiv 0$. Die Randwerte sollen wie folgt vorgeschrieben werden:

$$p\big|_{[-1,1]\times\{1,-1\}} = 0\,, \; < n, u >\big|_{\{1,-1\}\times[-1,1]} = 0\,,$$

d.h. homogene Dirichlet-Randbedingungen auf der oberen und unteren Seite und homogene Neumann-Randbedingungen auf der linken und der rechten Seite des Gebietes. Die Anfangsbedingung ist gegeben durch

$$p(0,x) = \cos\left(\frac{\pi}{2}(x_1-1)\right)\sin\left(\frac{\pi}{2}(x_2-1)\right).$$

Die exakte Lösung lautet

$$p(t,x) = \exp\left(-\frac{\pi^2}{2}\,t\right)\cos\left(\frac{\pi}{2}(x_1-1)\right)\sin\left(\frac{\pi}{2}(x_2-1)\right).$$

Die Berechnung einer Approximation ist unproblematisch. Dieses Beispiel soll lediglich die theoretisch bewiesene Ordnung des Algorithmus demonstrieren. Daher sollen äquidistante Zeitschritte und gleichmäßig verfeinerte Triangulierungen verwendet werden, um den Konvergenzsatz 2.11 zu illustrieren.
Tabelle 2.1 zeigt das Quadrat des Konsistenzfehlers im ersten Schritt, gemessen in der vom Funktional definierten Norm. Wir sehen die erwartete Reduktion des Funktionals proportional zu τ^3 (äquivalent zum Quadrat der Konsistenzfehlernorm $|||(\eta_u,\eta_p)|||_\tau^2$ wegen Satz 2.11). Die angesprochene Reduktion ist am besten in der letzten Spalte zu sehen, da dort der Fehler aus der Ortsdiskretisierung im Vergleich zu den Einträgen der anderen Spalten am kleinsten ist und gegenüber dem Zeitdiskretisierungsfehler nicht ins Gewicht fällt. Man beachte, daß für kleiner werdende Zeitschrittweiten τ die Approximationsordnung kleiner zu werden scheint, was daran liegt, daß wir für diese

h	1/8	1/16	1/32	1/64
$\tau = 0.2$	$3.22 \cdot 10^{-2}$	$2.10 \cdot 10^{-2}$	$1.82 \cdot 10^{-2}$	$1.75 \cdot 10^{-2}$
$\tau = 0.1$	$1.15 \cdot 10^{-2}$	$4.48 \cdot 10^{-3}$	$2.70 \cdot 10^{-3}$	$2.25 \cdot 10^{-3}$
$\tau = 0.05$	$5.54 \cdot 10^{-3}$	$1.55 \cdot 10^{-3}$	$5.36 \cdot 10^{-4}$	$2.81 \cdot 10^{-4}$
$\tau = 0.025$	$2.89 \cdot 10^{-3}$	$7.44 \cdot 10^{-4}$	$1.98 \cdot 10^{-4}$	$6.10 \cdot 10^{-5}$
$\tau = 0.0125$	$1.49 \cdot 10^{-3}$	$3.80 \cdot 10^{-4}$	$9.61 \cdot 10^{-5}$	$2.49 \cdot 10^{-5}$

Tabelle 2.1: Quadrat des Konsistenzfehlers (gemessen in der Norm des Funktionals $\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), f)$)

h	$\|p - p_{\tau,h}\|_{1,\Omega}$	$\|u - u_{\tau,h}\|_{\mathrm{div},\Omega}$
$\tau = 0.2$	$2.30 \cdot 10^{-3}$	$8.03 \cdot 10^{-3}$
$\tau = 0.1$	$3.45 \cdot 10^{-4}$	$3.00 \cdot 10^{-3}$
$\tau = 0.05$	$4.79 \cdot 10^{-5}$	$9.30 \cdot 10^{-4}$
$\tau = 0.025$	$6.42 \cdot 10^{-6}$	$2.88 \cdot 10^{-4}$
$\tau = 0.0125$	$1.31 \cdot 10^{-6}$	$1.55 \cdot 10^{-4}$

Tabelle 2.2: Fehler zum Zeitpunkt $t = 1$

Zeitschrittweiten nicht mehr in der Lage sind, den Ortsdiskretisierungsfehler so weit zu reduzieren, daß er im Vergleich zum Gesamtfehler nicht mehr ins Gewicht fällt.
Tabelle 2.2 zeigt jeweils den Fehler zum Zeitpunkt $t = 1$ für verschiedene Zeitschrittweiten τ (jeweils mit $h = 1/64$). Wir sehen die erwartete dritte Ordnung für die Approximation für p, die aus der Analyse im letzten Abschnitt klar wird. Die Approximation für u hat eine kleinere Ordnung.

Beispiel 2.21 Hier wählen wir wieder $\Omega = (-1, 1)^2$ und $f \equiv 0$ und nehmen homogene Dirichlet-Randbedingungen auf $\partial\Omega$ an. Die Anfangsbedingung ist durch

$$p(0, x) = \min\{1 - x_1, 1 + x_1, 1 - x_2, 1 + x_2\}$$

gegeben. Da die Anfangsbedingung nicht glatt ist, scheint es, daß die Methode eine niedrigere Ordnung während der ersten Zeitschritte (etwa für $t \in (0, 0.1)$) hat. Der wahre Grund ist aber, daß die elliptischen Subprobleme nicht so gut gelöst werden können. Die Ordnung „steigt“ jedoch nach einer Weile bis die von der Theorie vorhergesagte Ordnung erreicht wird. Abbildung 2.2 zeigt die Ordnung α in $\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t), f)^{1/2} = \mathcal{O}(\tau^\alpha)$ in Abhängigkeit der Zeit.

Beispiel 2.22 Das folgende Beispiel ist ein herausfordernder Test für unsere Einschrittmethode (vgl. [18, Example 9.2] und [10, Example 8.3]). Auf $\Omega = (-2, 2)^2$ berechnen wir eine Approximation zur Lösung von (2.1) mit der rechten Seite $f \equiv 0$ und schreiben homogene Dirichlet-Randbedingungen auf $\partial\Omega$ vor. Die Anfangsbedingung ist gegeben durch

$$p(0, x) = 250 \exp\left(-250 \|x\|^2\right),$$

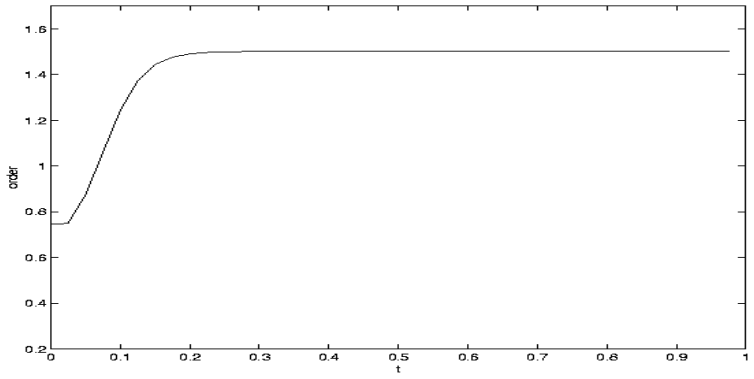

Abbildung 2.2: Die scheinbare Konsistenzordnung in Abhängigkeit der Zeit (Beispiel 2.21)

die als Approximation zu einer (punktierten) Wärmequelle oder zu der δ-Funktion angesehen werden kann. Die exakte Lösung ist durch

$$p(t,x) = \frac{1}{4t+1/250}\exp\left(-\frac{\|x\|^2}{4t+1/250}\right)$$

gegeben. Für $t \in (0,1]$, für die wir eine Approximation berechnen wollen, sind die Randbedingungen näherungsweise erfüllt. Man beachte, daß dieses Problem in den oben genannten Arbeiten nur auf $(0,1)^2$ behandelt wird, d.h. das Problem ist dort einfacher zu lösen als das vorliegende Problem. Das Program wurde mit einer Anfangszeitschrittweite $\tau_0 = 1.78\cdot 10^{-6}$ und einer Toleranzgrenze $\texttt{tol} = 0.4$ gestartet. Die Ortsdiskretisierung benutzt eine Triangulierung, die aus fünf adaptiven Verfeinerungen besteht. In Abbildung 2.3 (links) ist die Genauigkeit des Fehlerschätzers in Abhängigkeit von der Zeit (englisch: time) zu sehen.
Die Kurve, die mit „EXACT“ bezeichnet wird, misst den Fehler (englisch: error) in der Norm

$$|||(e_u,e_p)|||_\tau = \left(\int_0^\tau (\|e_u(t+\sigma)\|^2_{0,\Omega} + \tau \parallel \text{div } e_u(t+\sigma)\|^2_{0,\Omega} + \frac{1}{\tau}\|e_p(t+\sigma)\|^2_{0,\Omega} + \|\nabla e_p(t+\sigma)\|^2_{0,\Omega})d\sigma\right)^{1/2},$$

wobei $(e_u,e_p) = (u-u_{\tau,h}, p-p_{\tau,h})$. Man beachte, daß die Zeitintegration mit der Trapezregel approximiert wurde. Die linke Grafik in Abbildung 2.3 zeigt die Zeitschrittweite (englisch: time-step) in Abhängigkeit von der Zeit. Offensichtlich arbeitet der Fehlerschätzer sehr effizient. Man beachte insbesondere, daß $\tau_{\max} = 0.1$ gesetzt wurde. In Abbildung 2.4 wird die berechnete Approximation der Lösung und die dazugehörige Triangulierung zum Zeitpunkt $t = 2\cdot 10^{-6}$ dargestellt. Abbildung 2.5 zeigt die berechnete Approximation der Lösung und die zugehörige Triangulierung zum Zeitpunkt $t = 0.1$.

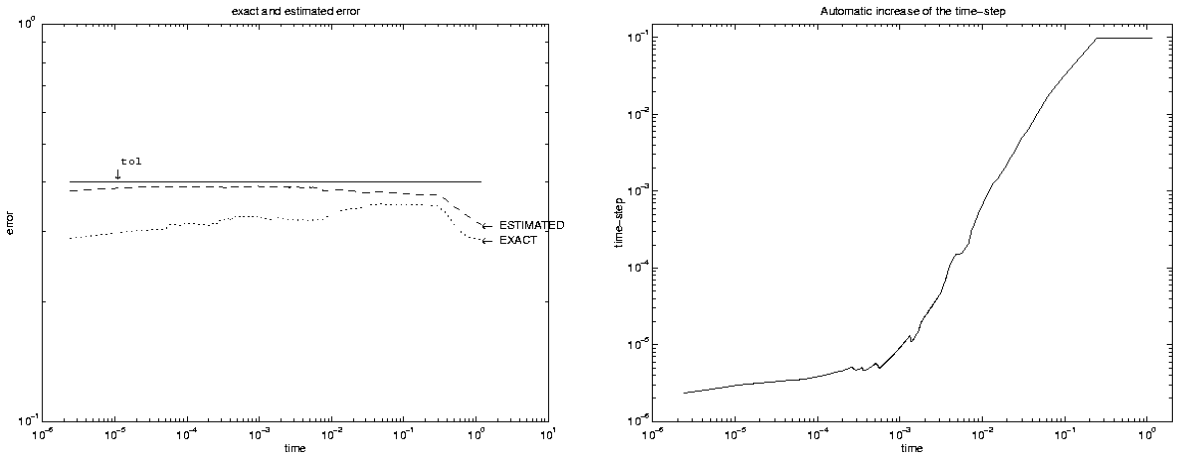

Abbildung 2.3: Qualität des Fehlerschätzers und die Zeitschrittweiten (Beispiel 2.22)

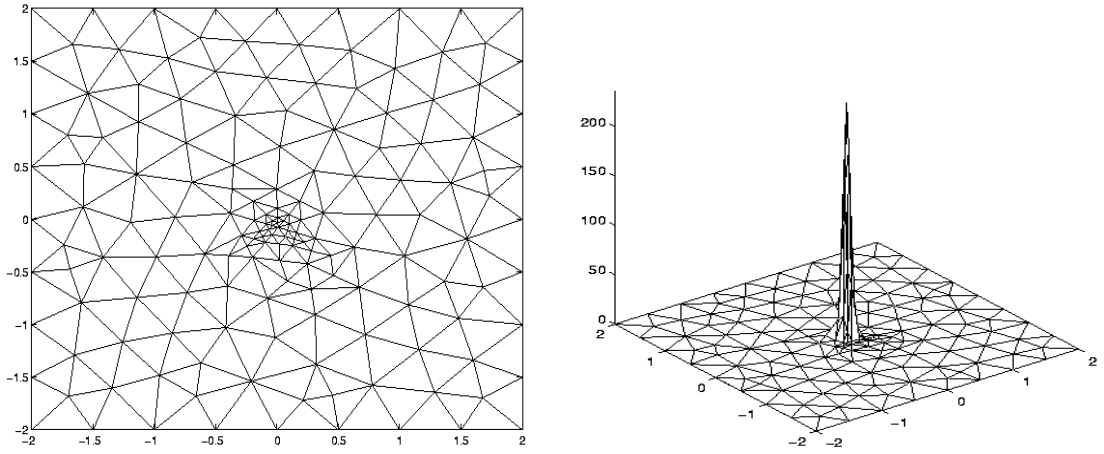

Abbildung 2.4: Triangulierung und Lösung zum Zeitpunkt $t = 2 \cdot 10^{-6}$ auf Level 3 (Beispiel 2.22)

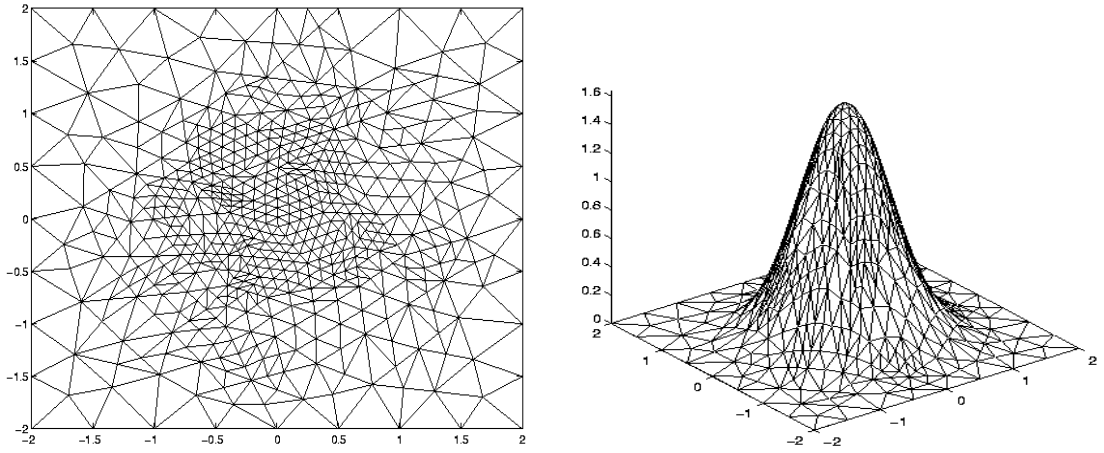

Abbildung 2.5: Triangulierung und Lösung zum Zeitpunkt $t = 0.1$ auf Level 3 (Beispiel 2.22)

Beispiel 2.23 Das nächste Beispiel stellt einen Test für die adaptive Verfeinerungsstrategie im Ort dar. (vgl. [2, Example 2] und [10, Example 8.4]).

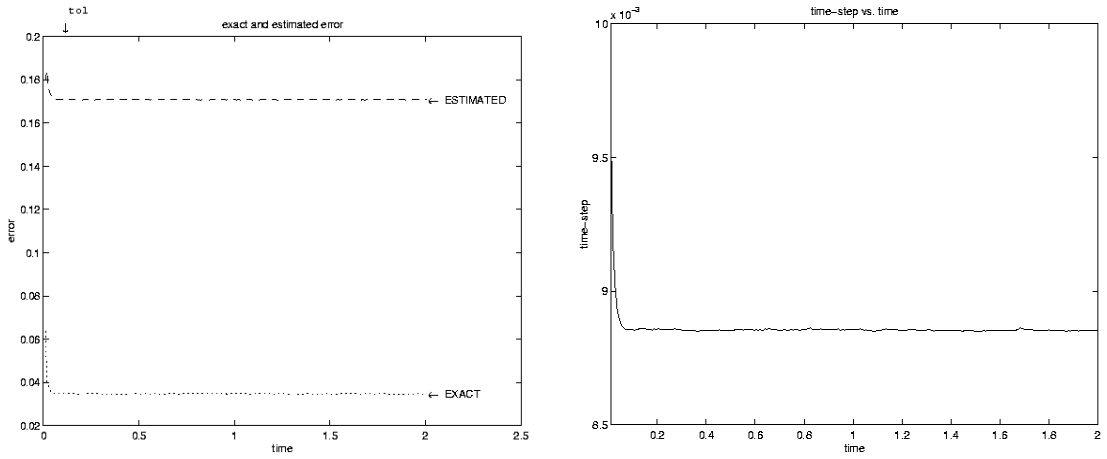

Abbildung 2.6: Qualität des Fehlerschätzers und die Zeitschrittweiten für `tol` $=$ 0.2 (Beispiel 2.23)

Man beachte, daß die rechte Seite f von t abhängt, d.h. unsere Theorie des letzten Abschnitts ist hier eigentlich nicht anwendbar. Daß f nicht nur von t sondern auch von p abhängen darf, werden wir im nächsten Kapitel sehen. Es sei $\Omega = (-1,1)^2$ und es gelten wieder homogene Dirichlet-Randbedingungen auf $\partial\Omega$. Die Anfangsbedingung sei

$$p(0,x) = 0.8 \exp\left(-80 \left\| x - \begin{pmatrix} 1/2 \\ 0 \end{pmatrix} \right\|^2\right).$$

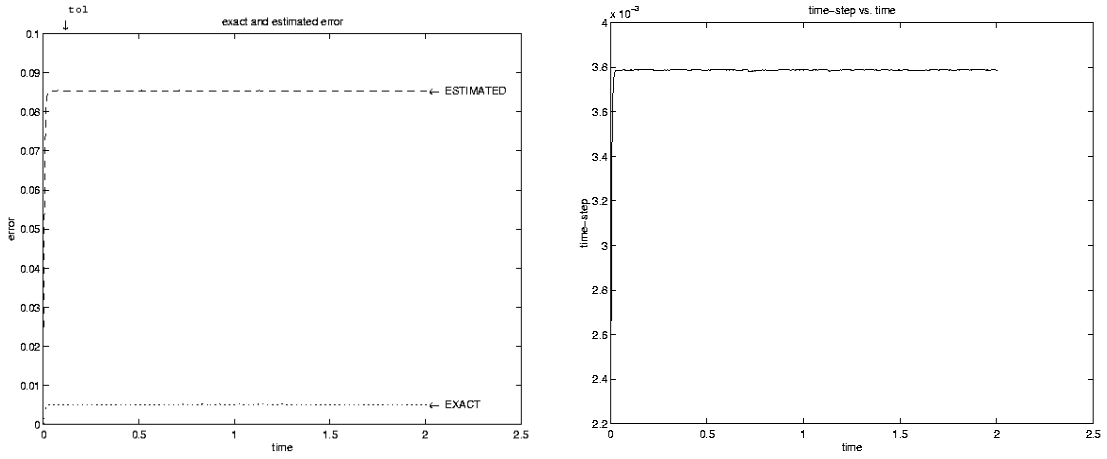

Abbildung 2.7: Qualität des Fehlerschätzers und die Zeitschrittweiten für `tol` $=$ 0.1 (Beispiel 2.23)

Definiert man

$$p(t,x) = 0.8 \exp\left(-80 \left\| x - \frac{1}{2} \begin{pmatrix} \cos(\pi t) \\ \sin(\pi t) \end{pmatrix} \right\|^2\right),$$

und setzt man

$$f(t,x) := \partial_t p - \Delta p, \tag{2.73}$$

so löst p offensichtlich die Gleichung (1.5).

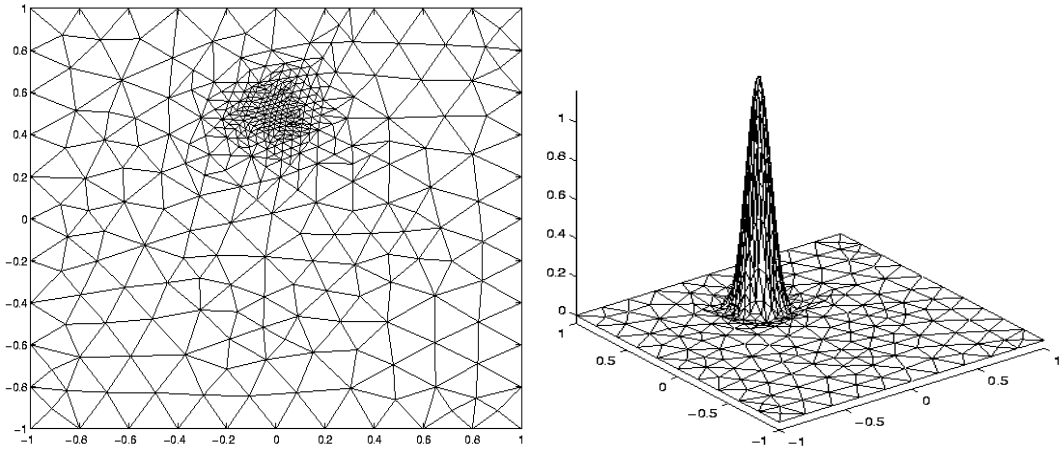

Abbildung 2.8: Triangulierung und Approximation zum Zeipunkt $t = 0.505$ auf Level 3 (Beispiel 2.23)

Gesucht ist eine Approximation für p, wobei $t \in (0, 2]$. Die Randbedingung ist wieder näherungsweise erfüllt. Das Programm wurde einmal mit der Anfangszeitschrittweite $\tau_0 = 0.01$ und `tol` $= 0.2$ gestartet (hier verkleinert sich die Zeitschrittweite, vgl. Abbildung 2.7 rechts). Ein zweites Mal wurde das Programm mit der Anfangszeitschrittweite $\tau_0 = 0.003$ und `tol` $= 0.1$ gestartet (hier vergrößert sich die Zeitschrittweite, vgl. Abbildung 2.6 rechts).

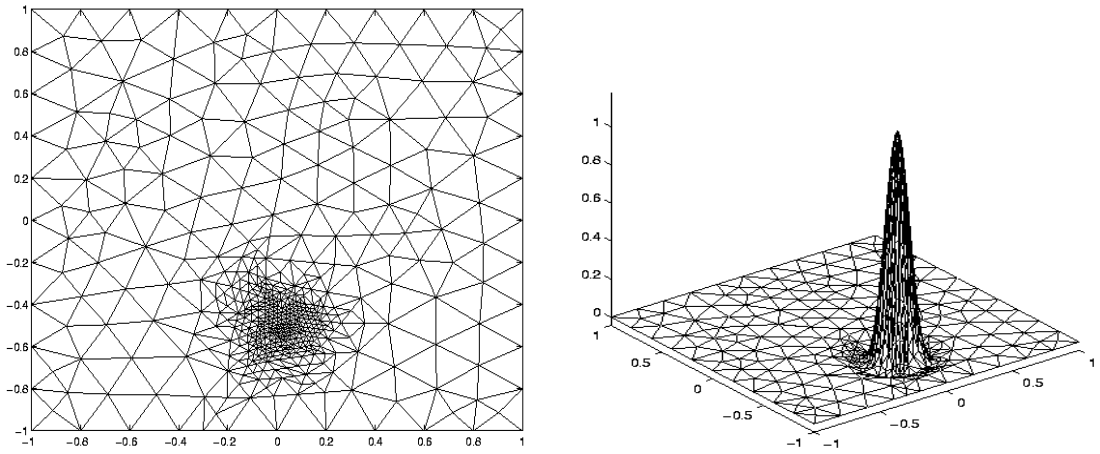

Abbildung 2.9: Triangulierung und Approximation zum Zeipunkt $t = 1.508$ auf Level 3 (Beispiel 2.23)

Wir berechnen bis $T = 2$, d.h. wenn eine volle Umdrehung erreicht worden ist. Wie man sieht, paßt Algorithmus 2.19 die Zeitschrittweite der geforderten Toleranzgerenze an. Danach bleibt die Schrittweite etwa konstant, da in jedem Zeitschritt das gleiche (um einige Grad gedrehte) elliptische Problem zu lösen ist. In den Abbildungen 2.7 und 2.6 (jeweils links) ist die Genauigkeit des Fehlerschätzers gut zu erkennen. In den Abbildungen 2.8 bzw. 2.9 wurden jeweils die berechnete Approximation der Lösung zum Zeitpunkt $t \approx 1/2$ (entspricht einer Viertel-Drehung) bzw. zum Zeitpunkt $t \approx 3/2$ (entspricht einer Dreiviertel-Drehung) und die zugehörige Triangulierung dargestellt.

Kapitel 3

Semilineare Probleme

In diesem Kapitel wollen wir die Methode aus dem letzten Kapitel auf semilineare parabolische Anfangs-Randwertprobleme verallgemeinern und einige Konvergenzaussagen beweisen. Anschließend betrachten wir einige numerische Beispiele.
Wir betrachten die semilineare Gleichung zweiter Ordnung

$$c(x)\partial_t p(t,x) - \text{ div } (a(x)\nabla p(t,x)) + \hat{f}(t,x,p(t,x),\nabla p(t,x)) = 0, \tag{3.1}$$

wobei $\hat{f} : [0,T] \times \Omega \times \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}$. Sei für $(t,q) \in [0,T] \times H^1_{\Gamma_D}(\Omega)$

$$\tilde{f}_{t,q} : \Omega \to \mathbb{R}$$

mit

$$\tilde{f}_{t,q}(x) = \hat{f}(t,x,q(x),\nabla q(x)).$$

Damit können wir $f(t,q) := \tilde{f}_{t,q}$ definieren. Wir fordern nun $f \in L^2((0,T); C^1(H^1_{\Gamma_D}(\Omega); L^2(\Omega)))$ mit

$$C^m(H^1_{\Gamma_D}(\Omega);L^2(\Omega)) = \left\{ \tilde{f} : H^1_{\Gamma_D}(\Omega) \to L^2(\Omega) \,\middle|\, \tilde{f} \text{ ist } m\text{-mal Fréchet-differenzierbar und } \tilde{f}^{(m)} \text{ ist stetig} \right\}.$$

Für fast alle $t \in [0,T]$ ist also $f(t,\cdot)$ Fréchet-differenzierbar. Aus der Mittelwertabschätzung (vgl. [5, Proposition 4.3.11]) folgt, daß $f(t,\cdot)$ für fast alle $t \in [0,T]$ lokal lipschitzstetig ist. Die Randbedingungen seien wie im letzten Kapitel. Für die Anfangsbedingung kann hier $P_0 \in H^1_{\Gamma_D}(\Omega)$ vorausgesetzt werden. Gesucht ist die schwache Lösung $p \in H^1((0,T); H^1_{\Gamma_D}(\Omega))$ von (3.1). Wir bezeichnen mit L_t die Lipschitzkonstante von $f(t,\cdot)$ in einer Umgebung $U(p(t),\delta_t) \subset H^1_{\Gamma_D}(\Omega)$ der Lösung $p(t)$. Es seien $L = \sup\limits_{t\in[0,T]} L_t < \infty$ und $\delta = \inf\limits_{t\in[0,T]} \delta_t > 0$.
Um das Problem mit der LSM zu behandeln, schreiben wir die Gleichung (3.1) in ein System erster Ordnung um:

$$\begin{aligned} c\partial_t p + \text{ div } u + f(\cdot,p) &= 0, \\ u + a\,\nabla p &= 0, \end{aligned} \tag{3.2}$$

Aus der Theorie der nichtlinearen partiellen Differentialgleichungen (vgl. z.B. [45, Section 15.1]) ist bekannt, daß eine eindeutige Lösung $(u,p) \in L^2((0,T); H_{\Gamma_N}(\text{div },\Omega)) \times$

$H^1((0,T);H^1_{\Gamma_D}(\Omega))$ von (3.2) existiert und daß diese stetig von den Anfangs- und Randdaten abhängt.
Aus Symmetriegründen wollen wir das System (3.2) in das äquivalente System

$$\begin{aligned} c^{1/2}\partial_t p + c^{-1/2}\ \text{div}\ u + c^{-1/2} f(\cdot,p) &= 0, \\ a^{-1/2}u + a^{1/2}\ \nabla p &= 0 \end{aligned}$$

umschreiben.
Nun diskretisieren wir in der Zeit, wobei wir die Bezeichnungen von Kapitel 2 übernehmen.
Analog zu Kapitel 2 definieren wir das LSF

$$\begin{aligned} \hat{\mathcal{F}}(u_\tau,p_\tau) = \int_0^\tau \Big(\tau \big\| c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}(\text{div}\ u_\tau(t+\sigma) + f(t+\sigma,p_\tau(t+\sigma))) \big\|^2_{0,\Omega} \\ + \big\| a^{-1/2}u_\tau(t+\sigma) + a^{1/2}\nabla p_\tau(t+\sigma) \big\|^2_{0,\Omega} \Big)\, d\sigma. \end{aligned} \tag{3.3}$$

3.1 Variationsformulierung für das nichtlineare LSF

Wir wollen wieder das LSF (3.3) im Raum $\tilde{V}_\tau \times \tilde{Q}_\tau$ minimieren (vgl. Kapitel 2).
Wir suchen $(u_\tau,\hat{p}_\tau) \in \tilde{V}_\tau \times \tilde{Q}_\tau$, so daß

$$\hat{\mathcal{F}}\,(u_\tau,p_\tau) = \min_{(v_\tau,\hat{q}_\tau)\in\tilde{V}_\tau\times\tilde{Q}_\tau} \hat{\mathcal{F}}\,(v_\tau,q_\tau) \tag{3.4}$$

mit $p_\tau(t+\sigma) = (1-\frac{\sigma}{\tau})p_\tau(t) + \hat{p}_\tau(\sigma)$ und $q_\tau(\sigma) = (1-\frac{\sigma}{\tau})p_\tau(t) + \hat{q}_\tau(\sigma)$.
Nun wollen wir den Evolutionsoperator definieren. Es sei $E(t+\sigma;t) : H^1_{\Gamma_D}(\Omega) \to H^1_{\Gamma_D}(\Omega)$, wobei $E(t+\sigma;t)p(t) = p(t+\sigma)$ die Lösung von (3.2) für $\sigma \in (0,\tau)$ bezeichnet. Dann ist (vgl. Bemerkung B.4.c)

$$\begin{aligned} p(t+\sigma) &= E(t+\sigma;t)\,p(t) \\ &= c^{-1/2}\exp(-\sigma A)\,c^{1/2}p(t) \\ &\quad - c^{-1/2}\int_0^\sigma (\exp(-(\sigma-\theta)A))c^{-1/2}f(t+\theta,p(t+\theta))\,d\theta\,, \end{aligned}$$

wobei A den in (2.9) definierten Operator bezeichnet.
Zunächst folgt ein Satz, welcher die Form von p^+_τ in Abhängigkeit von $p_\tau(t)$ angibt.

Satz 3.1 Der diskrete Evolutionsoperator $p^+_\tau = E_d(t+\tau,t)p_\tau(t)$ erfüllt die nichtlineare Gleichung (im Sinne der Distributionen (vgl. Abschnitt A.1)):

$$\begin{aligned} \int_0^\tau c^{1/2}\Big[I + \sigma\Big(c^{-1/2}\partial_p f(t+\sigma,p_\tau(t+\sigma))c^{-1/2} + A\Big)\Big]\Big[c^{1/2}\frac{p^+_\tau - p_\tau(t)}{\tau} \\ + Ac^{1/2}p_\tau(t+\sigma) + c^{-1/2}f(t+\sigma,p_\tau(t+\sigma))\Big]\,d\sigma = 0. \end{aligned}$$

Beweis:
Die zu der Minimierungsaufgabe (3.4) äquivalente Variationsformulierung lautet

$$\begin{aligned}\int_0^\tau \Big(\tau \Big(& c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}\mathrm{div}\, u_\tau(t+\sigma) + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma)), c^{1/2}\partial_t q(\sigma) \\ & + c^{-1/2}\ \mathrm{div}\ v(\sigma) + c^{-1/2}\partial_p f(t+\sigma, p_\tau(t+\sigma)) q(\sigma) \Big)_{0,\Omega} \\ & + \Big(a^{-1/2} u_\tau(t+\sigma) + a^{1/2}\nabla p_\tau(t+\sigma), a^{-1/2} v(\sigma) + a^{1/2}\nabla q(\sigma) \Big)_{0,\Omega} \Big) d\sigma = 0 \end{aligned} \tag{3.5}$$

für alle $(v, q) \in \tilde{V}_\tau \times \tilde{Q}_\tau$. Die obige Formel beinhaltet zwei verschiedene Variationsformulierungen in den Räumen $\tilde{V}_\tau$ und $\tilde{Q}_\tau$, die wir nun trennen. Aus (3.5) folgern wir zunächst

$$\begin{aligned}\int_0^\tau \Big(& \tau (c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}\ \mathrm{div}\ u_\tau(t+\sigma) \\ & + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma)), c^{-1/2}\ \mathrm{div}\ v(\sigma) \Big)_{0,\Omega} \\ & + \Big(a^{-1/2} u_\tau(t+\sigma) + a^{1/2}\nabla p_\tau(t+\sigma), a^{-1/2} v(\sigma)) \Big)_{0,\Omega} \Big) d\sigma = 0 \end{aligned}$$

für alle $v \in \tilde{V}_\tau$. Sei $\hat{D}_\tau = \left\{ \frac{\sigma}{\tau} q_1 + (1 - \frac{\sigma}{\tau}) q_2 \mid q_1, q_2 \in D_{A^2} \right\}$. Setzen wir nun $v = -a\nabla q$, wobei $q \in \hat{D}_\tau$ ist, so gilt: $c^{-1/2}\ \mathrm{div}\ v = Ac^{1/2} q$, und wir haben schließlich

$$\begin{aligned}\int_0^\tau \Big(\tau \Big(& c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}\ \mathrm{div}\ u_\tau(t+\sigma) \\ & + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma)), Ac^{1/2} q(\sigma) \Big)_{0,\Omega} \\ & - \Big(a^{-1/2} u_\tau(t+\sigma) + a^{1/2}\nabla p_\tau(t+\sigma), a^{1/2}\nabla q(\sigma)) \Big)_{0,\Omega} \Big) d\sigma = 0 \end{aligned} \tag{3.6}$$

für alle $q \in \hat{D}_\tau$. Nun folgern wir

$$\begin{aligned}\int_0^\tau \Big(\Big(& (c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma)) \\ & - Ac^{1/2} p_\tau(t+\sigma), \tau Ac^{1/2} q(\sigma) \Big)_{0,\Omega} \\ & + \Big(c^{-1/2}\ \mathrm{div}\ u_\tau(t+\sigma), (I+\tau A) c^{1/2} q(\sigma)) \Big)_{0,\Omega} \Big) d\sigma = 0 \end{aligned}$$

für alle $q \in \hat{D}_\tau$.
Es ergibt sich

$$\begin{aligned}\int_0^\tau \Big(\Big(& (I+\tau A)^{-1} [(c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma))) - \frac{1}{\tau}\, c^{1/2} p(t+\sigma)] \\ & + (\tau A)^{-1} c^{-1/2}\ \mathrm{div}\ u_\tau(t+\sigma), (I+\tau A)\tau Ac^{1/2} q(\sigma)) \Big)_{0,\Omega} \Big) d\sigma = 0 \end{aligned} \tag{3.7}$$

für alle $q \in \hat{D}_\tau$. Sei $\hat{L}_\tau = \left\{ \frac{\sigma}{\tau}\, q_1 + (1 - \frac{\sigma}{\tau}) q_2 \;\middle|\; q_1, q_2 \in L^2(\Omega) \right\}$. Da $(I + \tau A)\tau A c^{1/2}$ ein Isomorphismus von $\hat{D}_\tau$ nach $\hat{L}_\tau$ ist, folgt aus (3.7)

$$c^{-1/2} \text{ div } u_\tau = -(I + \tau A)^{-1} \left(\tau A (c^{1/2} \partial_t p_\tau + c^{-1/2} f(\cdot, p_\tau)) - A c^{1/2} p_\tau \right) \tag{3.8}$$

in $\hat{L}_\tau$. Die zweite Gleichung, die aus (3.5) folgt, lautet

$$\begin{aligned} \int_0^\tau \Bigg(\tau \Big(& c^{1/2} \partial_t p_\tau(t+\sigma) + c^{-1/2} \text{ div } u_\tau(t+\sigma) + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma)), \\ & c^{-1/2} \partial_t q(\sigma) + \partial_p f(t+\sigma, p_\tau(t+\sigma)) c^{1/2} q(\sigma) \Big)_{0,\Omega} \Bigg) d\sigma \\ + \underbrace{\int_0^\tau \left(\left(a^{-1/2} u_\tau(t+\sigma) + a^{1/2} \nabla p_\tau(t+\sigma), a^{1/2} \nabla q(\sigma) \right)_{0,\Omega} \right) d\sigma}_{=K} & = 0 \end{aligned} \tag{3.9}$$

für alle $q \in \tilde{Q}_\tau$. Fordern wir lediglich, daß (3.9) nur für $q \in \tilde{Q}_\tau \cap \hat{D}_\tau$ gilt, so liefert (3.6)

$$K = \int_0^\tau \Big(\tau \big(c^{1/2} \partial_t p_\tau(t+\sigma) + c^{-1/2} [\text{div}\, u_\tau(t+\sigma) + f(t+\sigma, p_\tau(t+\sigma))], A c^{1/2} q(\sigma) \big)_{0,\Omega} \Big)\, d\sigma$$

und damit insgesamt

$$\begin{aligned} \int_0^\tau \Bigg(\tau \Big(& c^{1/2} \partial_t p_\tau(t+\sigma) + c^{-1/2} \text{ div } u_\tau(t+\sigma) + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma)), \\ & c^{1/2} \partial_t q(\sigma) + c^{-1/2} \partial_p f(t+\sigma, p_\tau(t+\sigma)) q(\sigma) + A c^{1/2} q(\sigma) \Big)_{0,\Omega} \Bigg) = 0 \end{aligned}$$

für alle $q \in \tilde{Q}_\tau \cap \hat{D}_\tau$. Setzen wir nun $c^{-1/2}$ div u aus (3.8) ein und bedenken, daß $1 - \frac{z}{1+z} = \frac{1}{1+z}$ ist, so erhalten wir:

$$\begin{aligned} \int_0^\tau \Bigg(\tau \Big(& (I - \tau A)^{-1} [c^{1/2} \partial_t p_\tau(t+\sigma) + A c^{1/2} p_\tau(t+\sigma) + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma))], \\ & c^{1/2} \partial_t q(\sigma) + c^{-1/2} \partial_p f(t+\sigma, p_\tau(t+\sigma)) q(\sigma) + A\, c^{1/2} q(\sigma) \Big)_{0,\Omega} \Bigg) d\,\sigma = 0 \end{aligned}$$

für alle $q \in \tilde{Q}_\tau \cap \hat{D}_\tau$. Da $q \in \tilde{Q}_\tau \cap \hat{D}_\tau$, ergibt sich

$$\begin{aligned} \int_0^\tau \Bigg(\tau \Big(& (I - \tau A)^{-1} [c^{1/2} \partial_t p_\tau(t+\sigma) + A c^{1/2} p_\tau(t+\sigma) + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma))], \\ & \frac{1}{\tau}\, c^{1/2} \hat{q} + c^{-1/2} \partial_p f(t+\sigma, p_\tau(t+\sigma)) c^{-1/2} \frac{\sigma}{\tau}\, c^{1/2} \hat{q} + A\, \frac{\sigma}{\tau}\, c^{1/2} \hat{q} \Big)_{0,\Omega} \Bigg) d\,\sigma = 0 \end{aligned} \tag{3.10}$$

für alle $\hat{q} \in D_{A^2}$. Multipliziert man beide Seiten von (3.10) mit $(I - \tau A)$, so gilt (in Distributionen-Schreibweise (vgl. Abschnitt A.1)):

$$\begin{aligned} \Big\langle \int_0^\tau c^{1/2} \Big[I + \sigma \Big(c^{-1/2} \partial_p f(t+\sigma, p_\tau(t+\sigma)) c^{-1/2} + A \Big) \Big] & [c^{1/2} \partial_t p_\tau(t+\sigma) + A c^{1/2} p_\tau(t+\sigma) \\ & + c^{-1/2} f(t+\sigma, p_\tau(t+\sigma))] d\sigma, \hat{q} \Big\rangle = 0 \end{aligned}$$

für alle $\hat{q} \in C_0^\infty(\Omega)$, da $C_0^\infty(\Omega) \subset D_{A^2}$.

■

Um ein numerisches Verfahren konstruieren zu können, muß das Integral

$$\int_0^\tau c^{1/2}\Big[I+\sigma\Big(A+c^{-1/2}\partial_p f(t+\sigma,p_\tau(t+\sigma))c^{-1/2}\Big)\Big]\Big[c^{1/2}\partial_t p_\tau(t+\sigma) \\ +Ac^{1/2}p_\tau(t+\sigma)+c^{-1/2}f(t+\sigma,p_\tau(t+\sigma))\Big]d\sigma$$

durch berechenbare Terme approximiert werden.
Für die Analyse des Verfahrens lassen wir den Anteil mit ∂_p weg. Dies ist zwar nicht äquivalent zu (3.4), aber die so gelieferte Approximation erfüllt das Gewünschte (wie wir später sehen werden).
Somit ist ab jetzt $(u_\tau,\hat{p}_\tau)\in\tilde{V}_\tau\times\tilde{Q}_\tau$ gesucht (mit $p_\tau(t+\sigma)=(1-\frac{\sigma}{\tau})\,p_\tau(t)+\hat{p}_\tau(\sigma)$), so daß

$$\begin{aligned}\int_0^\tau\Big(&\tau(c^{1/2}\partial_t p_\tau(t+\sigma)+c^{-1/2}\ \mathrm{div}\ u_\tau(t+\sigma)\\ &+c^{-1/2}f(t+\sigma,p_\tau(t+\sigma)),c^{1/2}\partial_t q(\sigma)+c^{-1/2}\ \mathrm{div}\ v(\sigma)\Big)_{0,\Omega}\\ &+\Big(a^{-1/2}u_\tau(t+\sigma)+a^{1/2}\nabla p_\tau(t+\sigma),a^{-1/2}v(\sigma)+a^{1/2}\nabla q(\sigma)\Big)_{0,\Omega}\Big)d\sigma=0\end{aligned} \tag{3.11}$$

für alle $(v,q)\in\tilde{V}_\tau\times\tilde{Q}_\tau$. Die Lösung von (3.4) sei mit $(u_{\min},\hat{p}_{\min})$ bezeichnet. Die neue Lösung $(u_\tau,\hat{p}_\tau)$ genügt folgender Gleichung (der Beweis läuft analog zum Beweis von Satz 3.1, wobei man $\partial_p f$ durch 0 ersetzt):

$$\begin{aligned}0=&\int_0^\tau(I+\sigma A)[c^{1/2}\partial_t p_\tau(t+\sigma)+Ac^{1/2}p_\tau(t+\sigma)+c^{-1/2}f(t+\sigma,p_\tau(t+\sigma))]\,d\sigma\\ &=c^{1/2}p_\tau^+-c^{1/2}p_\tau(t)+\tau Ac^{1/2}p_\tau^++\frac{\tau^2A^2}{3}c^{1/2}\,p_\tau^++\frac{\tau^2A^2}{6}\,c^{1/2}p_\tau(t)\\ &\quad+\int_0^\tau(I+\sigma A)\,c^{-1/2}f(t+\sigma,p_\tau(t+\sigma))\,d\sigma.\end{aligned} \tag{3.12}$$

Man beachte, daß die obige Gleichung im Sinne der Distributionen gilt. Mit den Abkürzungen $r_{2,2}(z)=\frac{1-\frac{1}{6}z^2}{1+z+\frac{1}{3}z^2}$ (siehe (2.12)) und $r_{0,2}(z)=\frac{1}{1+z+\frac{1}{3}z^2}$ läßt sich das nichtlineare Einschrittverfahren angeben als

$$\begin{aligned}E_d(t+\tau,t)p_\tau(t)=&\,c^{-1/2}r_{2,2}(\tau A)c^{1/2}p_\tau(t)\\ &-\int_0^\tau c^{-1/2}r_{0,2}(\tau A)(I+\sigma A)c^{-1/2}f(t+\sigma,p_\tau(t+\sigma))\,d\sigma.\end{aligned} \tag{3.13}$$

Setzt man $f\equiv 0$, so erhält man die Aussage aus Lemma 2.2. Damit ist Satz 3.1 eine Verallgemeinerung von Lemma 2.2, d.h. man kann in Kapitel 2 statt $P_0\in D_A$ nur $P_0\in H^1_{\Gamma_D}(\Omega)$ fordern. Wir setzen ab jetzt aus Übersichtsgründen $a\equiv c\equiv 1$ voraus.

3.2 B-Stabilität

In diesem Abschnitt wird die B-Stabilität des angesprochenen Verfahrens bewiesen. Sei f bzgl. $(\cdot,\cdot)_{0,\Omega}$ dissipativ.

Im folgenden Satz werden wir sehen, daß die Methode (3.13) B-stabil ist. Zunächst machen wir aber eine wichtige Bemerkung.

Bemerkung 3.2 Sei $R(\sigma) = \sigma A$. Da A positiv definit ist, folgt $R(\sigma)$ ist positiv semidefinit für alle $\sigma \geq 0$. Da f dissipativ, folgt nach Lemma 1.10: $\partial_p f(t+\sigma, p)$ ist positiv semidefinit. Damit ist $\partial_p(R(\sigma)(Ap + f(t+\sigma, p))) = R(\sigma)(A + \partial_p f(t+\sigma, p))$ auch positiv semidefinit. Aus Lemma 1.10 folgt schließlich $R(\sigma)(A(\cdot) + f(t+\sigma, \cdot))$ ist dissipativ.

Satz 3.3 Das Verfahren (3.13) ist B-stabil.

Beweis:
Seien $p, q \in H^1_{\Gamma_D}(\Omega)$ beliebig. Es gelten folgende Abkürzungen:

$$\begin{aligned}
p^+ &= E_d(t+\tau, t)p\,, \\
q^+ &= E_d(t+\tau, t)q\,, \\
p_\tau(\sigma) &= \frac{\sigma}{\tau}\, p^+ + (1 - \frac{\sigma}{\tau}\,)p\,, \\
q_\tau(\sigma) &= \frac{\sigma}{\tau}\, q^+ + (1 - \frac{\sigma}{\tau}\,)q\,, \\
\xi^+ &= p^+ - q^+,\ \xi^- = p - q, \\
f_p(\sigma) &= f(t+\sigma, p(\sigma))\,, \\
f_q(\sigma) &= f(t+\sigma, q(\sigma))
\end{aligned}$$

und

$$\xi(\sigma) = \frac{\sigma}{\tau}\, \xi^+ + (1 - \frac{\sigma}{\tau}\,)\xi^-.$$

Wir werden in diesem Beweis die ursprüngliche Form des Verfahrens aus (3.12) verwenden. Setzt man in der Gleichung in (3.12) p bzw. q ein, so erhält man zwei Gleichungen. Zieht man von der ersten Gleichung die zweite ab, so ist

$$\int_0^\tau (I + \sigma A)(\partial_t \xi(\sigma) + A\xi(\sigma) + f_p(\sigma) - f_q(\sigma))\, d\sigma = 0.$$

Setzen wir nun ξ in die entprechende Variationsformulierung ein, so erhalten wir

$$\int_0^\tau ((I + \sigma A)(\partial_t \xi(\sigma) + A\xi(\sigma) + f_p(\sigma) - f_q(\sigma)), \xi(\sigma))_{0,\Omega}\, d\sigma = 0.$$

Damit folgt

$$\begin{aligned}
\int_0^\tau (\partial_t \xi(\sigma), \xi(\sigma))_{0,\Omega}\, d\sigma = &- \int_0^\tau (\sigma A \partial_t \xi(\sigma) + A\xi(\sigma), \xi(\sigma))_{0,\Omega} \\
&- \underbrace{(\sigma A(A\xi(\sigma) + f_p(\sigma) - f_q(\sigma)), \xi(\sigma))_{0,\Omega}}_{\geq 0 \text{ da } R(\sigma)\,(A(\cdot) + f(t+\sigma,\cdot)) \text{ dissipativ nach Bemerkung 3.2}} \\
&- \underbrace{(f_p(\sigma) - f_q(\sigma)), \xi(\sigma))_{0,\Omega}}_{\geq 0 \text{ da } f \text{ dissipativ}}\, d\sigma.
\end{aligned} \tag{3.14}$$

Berechnung beider Seiten von (3.14) liefert

$$\begin{aligned}
&\frac{1}{2}\,(\xi^+ - \xi^-, \xi^+ + \xi^-)_{0,\Omega} \le -\frac{1}{6}\,\tau\left(4\|A^{1/2}\xi^+\|^2_{0,\Omega} + \|A^{1/2}\xi^-\|^2_{0,\Omega} + (A\xi^+, \xi^-)_{0,\Omega}\right)\\
&= -\frac{1}{6}\,\tau\Bigg(3\|A^{1/2}\xi^+\|^2_{0,\Omega} + \frac{3}{4}\,\|A^{1/2}\xi^-\|^2_{0,\Omega} + (\|A^{1/2}\xi^+\|^2_{0,\Omega}\\
&\qquad\qquad + \frac{1}{4}\,\|A^{1/2}\xi^-\|^2_{0,\Omega} + (A^{1/2}\xi^+, A^{1/2}\xi^-)_{0,\Omega})\Bigg)\\
&= -\frac{1}{6}\,\tau\left(3\|A^{1/2}\xi^+\|^2_{0,\Omega} + \frac{3}{4}\,\|A^{1/2}\xi^-\|^2_{0,\Omega} + (\|A^{1/2}(\xi^+ + \frac{1}{2}\,\xi^-)\|^2_{0,\Omega})\right) \le 0.
\end{aligned} \tag{3.15}$$

Damit ist

$$\|\xi^+\|_{0,\Omega} \le \|\xi^-\|_{0,\Omega}.$$

■

3.3 Konvergenzaussagen

Sei für ein beliebiges Gitter

$$\tau_{\min,\Delta} := \min_{j=1}^{M_\Delta} \tau_j^\Delta.$$

Wir setzen nun zusätzlich für jede Folge von Gittern $\Delta \subset [0,T]$ mit $\tau_\Delta \to 0$ voraus:

$$\frac{\tau_\Delta}{\tau_{\min,\Delta}} \eqsim 1.$$

Hieraus folgt z.B., daß

$$\tau_\Delta M_\Delta \eqsim 1. \tag{3.16}$$

Es folgt ein Satz, der die Konvergenzordnung 2 für das Verfahren (3.13) garantiert. Bevor wir den Satz formulieren, erinnern wir an das berühmte

Lemma 3.4 (Lemma von Gronwall (L^1-Version))
Es sei $b > 0$ beliebig. Seien $v \in L^1(0,b)$ eine nicht negative Funktion, $\eta, \mu \ge 0$ und es gelte für fast alle $t \in (0,b)$

$$v(t) \le \eta + \mu \int_0^t v(\sigma)\, d\sigma. \tag{3.17}$$

Dann gilt für fast alle $t \in (0,b)$:

$$v(t) \le \eta\, e^{t\,\mu}.$$

Beweis:
Man findet in [49, Chap. 1.III, pp. 16, Absatz δ] die L^p-Version des Lemmas von Gronwall, sowie Literaturhinweise zum Beweis.

■

Satz 3.5 Sei $p \in C^2([0,T], H^1_{\Gamma_D}(\Omega))$ die exakte Lösung von (3.2). Sei p_τ die mit dem Verfahren (3.13) berechnete Approximation. Dann existiert ein τ^*, so daß für alle Gitter $\Delta \subset [0,T]$ mit $\tau_\Delta \le \tau^*$ gilt:

$$\|p_\tau(t_n) - p(t_n)\|_{1,\Omega} \le C\, \tau_\Delta^2\, e^{C(n+1)\tau_\Delta^{1/2}} \quad \text{für} \quad 0 \le n \le M_\Delta, \tag{3.18}$$

wobei C nur von $\sup\limits_{t\in[0,T]} \{\|\partial_t^2 p(t)\|_{1,\Omega}\}$ abhängt.

Beweis:
(a) Zunächst wollen wir zeigen, daß es $q_\tau \in Q_\tau$ gibt mit

$$\max_{t\in[0,T]} \|q_\tau(t) - p(t)\|_{1,\Omega} < \delta. \tag{3.19}$$

Zu jeder Zerlegung $\{0 < t_1 < \ldots < t_{M_\Delta - 1} < T\} = \Delta \subset [0,T]$ sei $I_\Delta p$ der lineare Spline zu p bzgl. Δ. Aus der Interpolationstheorie folgern wir $\max\limits_{t\in[0,T]} \|I_\Delta p(t) - p(t)\|_{1,\Omega} \le C_1 \tau_\Delta^2$, wobei $C_1 = \sup\limits_{t\in[0,T]} \|\partial_t^2\, p(t)\|_{1,\Omega} < \infty$ gesetzt wurde. Sei $\tau^* \le \min\{1, \frac{\delta}{2C_1}\} =: \varepsilon_0$. Damit gilt: $\max\limits_{t\in[0,T]} \|I_\Delta p(t) - p(t)\|_{1,\Omega} \le \frac{\delta}{2}$. Seien $q_j \in H^1_{\Gamma_D}(\Omega)$ mit $\|q_j - p(t_j)\|_{1,\Omega} < \frac{\delta}{2}$ für $j = 1, \ldots, M_\Delta$ und $q_0 = p_0$. Sei

$$q_\tau(t) = \frac{t - t_{j-1}}{\tau_j}\, q_j + \left(1 - \frac{t - t_{j-1}}{\tau_j}\right) q_{j-1}, \text{ falls } t \in [t_{j-1}, t_j],$$

wobei $1 \le j \le M_\Delta$. Dann gilt:

$$\max_{t\in[0,T]} \|q_\tau(t) - p(t)\|_{1,\Omega} \le \max_{t\in[0,T]} \|I_\Delta p(t) - p(t)\|_{1,\Omega} + \max_{t\in[0,T]} \|q_\tau(t) - I_\Delta p(t)\|_{1,\Omega} < \delta.$$

Angenommen, τ^* ist so gewählt, daß

$$\|p_\tau(t_j) - p(t_j)\|_{1,\Omega} \le j \frac{\delta}{3M_\Delta} \tag{3.20}$$

für $j = 0, \ldots, M_\Delta$, dann ist nach obiger Überlegung

$$\max_{t\in[0,T]} \|p_\tau(t) - p(t)\|_{1,\Omega} < \delta$$

für alle Δ mit $\tau_\Delta \le \tau^*$ erfüllt. Der Beweis von (3.20) wird in Teil (c) nachgeholt.

(b) (i) Seien $\hat{p}_j = p(t_j), j = 0, \cdots, M_\Delta$. Sei $0 < n \le M_\Delta$. Es gilt dann,

$$\hat{p}_n = r_{2,2}(\tau_n A)\hat{p}_{n-1} - \int_0^{\tau_n} (I + \sigma A)\, r_{0,2}(\tau_n A) f(t_{n-1} + \sigma, p(t_{n-1} + \sigma))\, d\sigma + d_n,$$

wobei d_n den *Defekt* darstellt. Der Defekt ist gegeben durch

$$\begin{aligned} d_n &= -r_{2,2}(\tau_n A)\hat{p}_{n-1} + \hat{p}_n \\ &\quad - \int_0^{\tau_n} (I + \sigma A)\, r_{0,2}(\tau_n A)(\partial_t p(t_{n-1} + \sigma) + Ap(t_{n-1} + \sigma))\, d\sigma. \end{aligned} \tag{3.21}$$

Durch partielle Integration erhält man

$$\begin{aligned}\int_0^{\tau_n} (I+\sigma A)\, r_{0,2}(\tau_n A)\partial_t p(t_n+\sigma)\, d\sigma = (I+\tau_n A)\, r_{0,2}(\tau_n A)\hat{p}_n \\ -r_{0,2}(\tau_n A)\hat{p}_{n-1} - \int_0^{\tau_n} A r_{0,2}(\tau_n A) p(t_{n-1}+\sigma)\, d\sigma.\end{aligned} \tag{3.22}$$

Setzen wir nun (3.22) in (3.21) ein und vereinfachen, so ergibt sich

$$\begin{aligned}d_n = (r_{0,2}(\tau_n A) - r_{2,2}(\tau_n A))\hat{p}_{n-1} + (I-(I+\tau_n A)\, r_{0,2}(\tau_n A))\hat{p}_n \\ -\int_0^{\tau_n} \sigma A^2 r_{0,2}(\tau_n A) p(t_{n-1}+\sigma)\, d\sigma.\end{aligned} \tag{3.23}$$

Beachten wir, daß

$$\int_0^{\tau_n} \sigma A^2 r_{0,2}(\tau_n A) I_\Delta p(t_{n-1}+\sigma)\, d\sigma = r_{0,2}(\tau_n A)(\frac{\tau_n^2}{3}\, A^2\hat{p}_n + \frac{\tau_n^2}{6}\, A^2\hat{p}_{n-1}), \tag{3.24}$$

so können wir (3.23) mit Hilfe von $I_\Delta p$ schreiben als

$$d_n = -\int_0^{\tau_n} \sigma A^2 r_{0,2}(\tau_n A)(p(t_{n-1}+\sigma) - I_\Delta p(t_{n-1}+\sigma))\, d\sigma. \tag{3.25}$$

Aus der Interpolationstheorie wissen wir, daß aus dem Mittelwertsatz für jedes $\sigma \in [0,\tau_n]$ folgt:

$$p(t_{n-1}+\sigma) - I_\Delta p(t_{n-1}+\sigma) = \frac{1}{2}\,\partial_t^2 p(t_{n-1}+\theta_\sigma)\sigma(\sigma-\tau_n) \tag{3.26}$$

mit einem $\theta_\sigma \in (0,\tau_n)$.
(ii) Seien

$$R(t_j,t_n) = \prod_{i=j+1}^{n} r_{2,2}(\tau_i A),$$

$$\Sigma_{j+1}(\sigma) = f(t_j+\sigma, p_\tau(t_j+\sigma)) - f(t_j+\sigma, p(t_j+\sigma)),$$

$$\tilde{\Sigma}_{j+1}(\sigma) = f(t_j+\sigma, p_\tau(t_j+\sigma)) - f(t_j+\sigma, I_\Delta p(t_j+\sigma))$$

und

$$\hat{\Sigma}_{j+1}(\sigma) = f(t_j+\sigma, I_\Delta p(t_j+\sigma)) - f(t_j+\sigma, p(t_j+\sigma)).$$

Bezeichnen wir den Fehler mit $e_j = p_\tau(t_j) - p(t_j)$, $j = 0,\dots,M_\Delta$, so erhalten wir rekursiv, wegen $e_0 = 0$

$$e_n = \sum_{j=0}^{n-1} R(t_{j+1},t_n)\left[d_{j+1} + \int_0^{\tau_{j+1}} r_{0,2}(\tau_{j+1}A)(I+\sigma A)\Sigma_{j+1}(\sigma)\right]. \tag{3.27}$$

Beachten wir, daß für $j = 1,\dots,M_\Delta$, $q \in H^1_{\Gamma_D}(\Omega)$ $\sigma \in [0,\tau]$ gilt

$$\|(I+\sigma A)\, r_{0,2}(\tau A)q\|_{1,\Omega} \le \|(I+\tau A)\, r_{0,2}(\tau A)q\|_{1,\Omega}$$

und setzen

$$S(t_{j+1},t_n) = (I+\tau_{j+1}A) r_{0,2}(\tau_{j+1}A) R(t_{j+1},t_n),$$

so folgt aus (3.27)

$$\begin{aligned}\|e_n\|_{1,\Omega} \leq & \left\|\sum_{j=0}^{n-1} S(t_{j+1},t_n)\int_0^{\tau_{j+1}} \hat{\Sigma}_{j+1}(\sigma)\,d\sigma\right\|_{1,\Omega} + \left\|\sum_{j=0}^{n-1} R(t_{j+1},t_n)d_{j+1}\right\|_{1,\Omega} \\ & + \left\|\sum_{j=0}^{n-1} S(t_{j+1},t_n)\int_0^{\tau_{j+1}} \tilde{\Sigma}_{j+1}(\sigma)\,d\sigma\right\|_{1,\Omega} =: \|I_1\|_{1,\Omega} + \|I_2\|_{1,\Omega} + \|I_3\|_{1,\Omega}\end{aligned} \tag{3.28}$$

Nun wollen wir $\|I_1\|_{1,\Omega}, \|I_2\|_{1,\Omega}$ und $\|I_3\|_{1,\Omega}$ abschätzen.
Um $\|I_1\|_{1,\Omega}$ abzuschätzen, definieren wir

$$\bar{r}_{2,2}(z) = \frac{1+\frac{5}{6}z+\frac{z^2}{6}}{1+z+\frac{z^2}{3}},$$

$$\hat{\bar{r}}_{2,2}(z) = \frac{1-\frac{z^2}{6}}{1+\frac{5}{6}z+\frac{z^2}{6}}$$

und

$$\bar{R}(t_j,t_n) = \prod_{i=j+1}^{n} \bar{r}_{2,2}(\tau_i A), j = 0,\ldots,M_\Delta - 1.$$

Aus der Poincaré-Friedrichs-Ungleichung folgern wir

$$\|I_1\|_{1,\Omega} \leq c_1\|\nabla I_1\|_{1,\Omega} \leq c_2\|AI_1\|_{0,\Omega}.$$

Wegen der lokalen Lipschitzstetigkeit von f, (3.19) und (3.26), gilt

$$\max_{\sigma\in[0,t_n]}\|\hat{\Sigma}_{j+1}(\sigma)\|_{0,\Omega} \leq L \max_{\sigma\in[0,t_n]}\|I_\Delta p(t_j+\sigma) - p(t_j+\sigma)\|_{1,\Omega} \lesssim \tau_\Delta^2. \tag{3.29}$$

Beachtet man, daß $\bar{r}_{2,2}$ auf $[0,\infty)$ positiv und monoton fallend ist, $\sup_{x\in(0,\infty)}|\hat{\bar{r}}_{2,2}(x)|=1$ und $r_{2,2} = \bar{r}_{2,2}\hat{\bar{r}}_{2,2}$, so folgt aus der Formel der geometrischen Reihe

$$\begin{aligned}\|I_1\|_{1,\Omega} &\lesssim \|AI_1\|_{0,\Omega} \lesssim \tau_\Delta^2 \left\|\sum_{j=0}^{n-1} S(t_{j+1},t_n)\tau_{j+1}A\right\|_{L^2(\Omega)\to L^2(\Omega)} \\ &\lesssim \tau_\Delta^2 \left\|\sum_{j=0}^{n-1} \bar{R}(t_{j+1},t_n) r_{0,2}(\tau_{\min,\Delta}A)\tau_\Delta A\right\|_{L^2(\Omega)\to L^2(\Omega)} \\ &\lesssim \tau_\Delta^2 \left\|\sum_{j=0}^{n-1} \bar{r}_{2,2}(\tau_{\min,\Delta}A)^{n-j-1} r_{0,2}(\tau_{\min,\Delta}A)\tau_\Delta A\right\|_{L^2(\Omega)\to L^2(\Omega)} \\ &\lesssim \tau_\Delta^2 \left\|(I-\bar{r}_{2,2}(\tau_{\min,\Delta}A))^{-1} r_{0,2}(\tau_{\min,\Delta}A)\tau_\Delta A\right\|_{L^2(\Omega)\to L^2(\Omega)} \\ &= \tau_\Delta^2 \left\|6(\tau_{\min,\Delta}A + \tau_{\min,\Delta}^2A^2)^{-1} r_{0,2}(\tau_{\min,\Delta}A)^{-1} r_{0,2}(\tau_{\min,\Delta}A)\tau_\Delta A\right\|_{L^2(\Omega)\to L^2(\Omega)} \\ &\lesssim \tau_\Delta^2 \left\|(I+\tau_{\min,\Delta}A)^{-1}\right\|_{L^2(\Omega)\to L^2(\Omega)} \lesssim \tau_\Delta^2.\end{aligned} \tag{3.30}$$

Für $\|I_2\|_{1,\Omega}$ seien nun

$$\tilde{r}_{2,2}(z) = \frac{1+z+\frac{z^2}{6}}{1+z+\frac{z^2}{3}},$$

$$\hat{r}_{2,2}(z) = \frac{1 - \frac{z^2}{6}}{1 + z + \frac{z^2}{6}}$$

und

$$\tilde{R}(t_j, t_n) = \prod_{i=j+1}^{n} \tilde{r}_{2,2}(\tau_i A), j = 0, \ldots, M_\Delta - 1$$

definiert. Man beachte, daß $\tilde{r}_{2,2}$ auf $[0, \infty)$ positiv und monoton fallend ist, $\sup_{x \in (0,\infty)} |\hat{r}_{2,2}(x)| = 1$ und $r_{2,2} = \tilde{r}_{2,2}\hat{r}_{2,2}$.

Die Gleichungen (3.25) und (3.26) liefern nun (mit der Formel der geometrischen Reihe)

$$\begin{aligned}
\|I_2\|_{1,\Omega} &\lesssim \left\| \sum_{j=0}^{n-1} R(t_{j+1}, t_n) r_{0,2}(\tau_{j+1}A) \int_0^{\tau_{j+1}} \frac{1}{2}\, \sigma A^2 \sigma(\sigma - \tau_{j+1}) d\sigma \right\|_{1,\Omega} \\
&= \left\| \sum_{j=0}^{n-1} R(t_{j+1}, t_n)\, r_{0,2}(\tau_{j+1}A)\, \frac{1}{24}\, \tau_{j+1}^4 A^2 \right\|_{H^1_{\Gamma_D}(\Omega) \to H^1_{\Gamma_D}(\Omega)} \\
&\lesssim \tau_\Delta^2 \left\| \sum_{j=0}^{n-1} \tilde{r}_{2,2}(\tau_{\min,\Delta}A)^{n-j-1}\, r_{0,2}(\tau_{\min,\Delta}A) \tau_\Delta^2 A^2 \right\|_{H^1_{\Gamma_D}(\Omega) \to H^1_{\Gamma_D}(\Omega)} \\
&\lesssim \tau_\Delta^2 \left\| (I - r_{2,2}((\tau_{\min,\Delta}A))^{-1}\, r_{0,2}(\tau_{\min,\Delta}A) \tau_\Delta^2 A^2 \right\|_{H^1_{\Gamma_D}(\Omega) \to H^1_{\Gamma_D}(\Omega)} \\
&= \tau_\Delta^2 \left\| 6(\tau_{\min}A)^{-2} r_{0,2}(\tau_{\min,\Delta}A)^{-1}\, r_{0,2}(\tau_{\min,\Delta}A) \tau_\Delta^2 A^2 \right\|_{H^1_{\Gamma_D}(\Omega) \to H^1_{\Gamma_D}(\Omega)} \\
&= \tau_\Delta^2\, 6(\tau_\Delta/\tau_{\min,\Delta})^2 \lesssim \tau_\Delta^2.
\end{aligned} \tag{3.31}$$

Für die Abschätzung von $\|I_3\|_{1,\Omega}$ beachte man, daß

$$\|(I + \tau_{j+1}A) r_{0,2}(\tau_{j+1}A)\|_{L^2(\Omega) \to H^1_{\Gamma_D}(\Omega)} \lesssim \tau_{j+1}^{-1/2}.$$

Aus der lokalen Lipschitzstetigkeit von f und (3.19) folgern wir

$$\begin{aligned}
\int_0^{\tau_{j+1}} \|\tilde{\Sigma}_{j+1}(\sigma)\|_{0,\Omega}\, d\sigma &\le L \int_0^{\tau_{j+1}} \|p_\tau(t_j + \sigma) - I_\Delta p(t_j + \sigma)\|_{1,\Omega}\, d\sigma \\
&= L\, \frac{\tau_{j+1}}{2} (\|e_j\|_{1,\Omega} + \|e_{j+1}\|_{1,\Omega}).
\end{aligned} \tag{3.32}$$

Wegen der A-Stabilität des Verfahrens folgt

$$\begin{aligned}
\|I_3\|_{1,\Omega} &\lesssim \left\| \sum_{j=0}^{n-1} \tau_{j+1}^{-1/2} \int_0^{\tau_{j+1}} \tilde{\Sigma}_{j+1}(\sigma)\, d\sigma \right\|_{0,\Omega} \\
&\lesssim \tau_\Delta^{1/2} \sum_{j=0}^{n-1} (\|e_{j+1}\|_{1,\Omega} + \|e_j\|_{1,\Omega}) \lesssim \tau_\Delta^{1/2} \sum_{j=1}^{n} \|e_j\|_{1,\Omega}.
\end{aligned} \tag{3.33}$$

Damit folgt insgesamt:

$$\|e_n\|_{1,\Omega} \leq (\|I_1\|_{1,\Omega} + \|I_2\|_{1,\Omega}) + \|I_3\|_{1,\Omega} \leq \tilde{C}\tau_\Delta^2 + \tilde{C}\tau_\Delta^{1/2} \sum_{j=1}^{n} \|e_j\|_{1,\Omega}.$$

Die Konstante $\tilde{C}$ hängt von C_1 und L ab. Wählen wir $\tau^* \leq \min\{\varepsilon_3, \frac{1}{4\tilde{C}^2}\}$ (wobei wir in Teil (c) ε_3 angeben), so gilt $(1 - \tilde{C}\tau_\Delta^{1/2}) > 0$ und (mit $C = 2\tilde{C}$):

$$\|e_n\|_{1,\Omega} \leq \frac{\tilde{C}}{1 - \tilde{C}\tau_\Delta^{1/2}}\, \tau_\Delta^2 + \frac{\tilde{C}\tau_\Delta^{1/2}}{1 - \tilde{C}\tau_\Delta^{1/2}} \sum_{j=1}^{n-1} \|e_j\|_{1,\Omega} \leq C\tau_\Delta^2 + C\tau_\Delta^{1/2} \sum_{j=1}^{n-1} \|e_j\|_{1,\Omega}. \quad (3.34)$$

Definieren wir $v(t) = \|e_j\|_{1,\Omega}$ für $t \in [j\ \tau_\Delta^{1/2}, (j+1)\tau_\Delta^{1/2}), j = 0, \ldots, M_\Delta - 1$ und $v(M_\Delta\, \tau_\Delta^{1/2}) = \|e_{M_\Delta}\|_{1,\Omega}$, so erhalten wir für $n = 1, \ldots, M_\Delta$ und $t \in [n\ \tau_\Delta^{1/2}, (n+1)\tau_\Delta^{1/2})$

$$\begin{aligned} \tau_\Delta \sum_{j=1}^{n-1} \|e_j\|_{1,\Omega} &= \int_0^{(n-1)\tau_\Delta^{1/2}} v(s)\, ds \\ &\leq \int_0^t v(s)\, ds. \end{aligned}$$

Es sei nun $t \in [0, M_\Delta\, \tau_\Delta^{1/2})$ beliebig. Wähle $0 \leq n \leq M_\Delta - 1$ mit $t \in [n\ \tau_\Delta^{1/2}, (n+1)\tau_\Delta^{1/2})$. Aus (3.34) folgern wir

$$v(t) = \|e_n\|_{1,\Omega} \leq C\, \tau_\Delta^2 + C\, \tau_\Delta^{1/2} \sum_{j=1}^{n-1} \|e_j\|_{1,\Omega} \leq C\, \tau_\Delta^2 + C \int_0^t v(s)\, ds.$$

Setzt man $\eta = C\, \tau_\Delta^2$, $\mu = C$ und beachtet, daß $t \leq (n+1)\tau_\Delta^{1/2}$, so liefert das Lemma von Gronwall (Lemma 3.4):

$$\|e_n\|_{1,\Omega} \leq C\, \tau_\Delta^2\, e^{C(n+1)\tau_\Delta^{1/2}}.$$

(c) Nun wollen wir durch geeignete Wahl von τ^* ereichen, daß (3.20) für $j = 0, \ldots, M_\Delta$ gilt. Das zeigen wir durch vollständige Induktion über j. Induktionsanfang ist für $j = 0$ wegen $p_\tau(0) = p(0) = p_0$ richtig. Für $j+1$ definieren wir zunächst: $\rho_1 := \frac{\delta}{6M_\Delta} < \frac{\delta}{2}$ und $\rho_2 := \frac{(2j+1)\delta}{6M_\Delta} < \frac{\delta}{2}$.
Sei die Abbildung

$$D : \overline{U(p(t_{j+1}), \rho_1)} \subset H^1_{\Gamma_D}(\Omega) \to \overline{U(p(t_{j+1}), \rho_1)}$$

mit

$$D(q) := r_{2,2}(\tau_{j+1}A)p(t_j) - \int_0^{\tau_{j+1}} r_{0,2}(\tau_{j+1}A)(I + \sigma A) f(t_j + \sigma, \tilde{q}(\sigma))d\sigma$$

definiert, wobei $\tilde{q}(\sigma) := (1 - \frac{\sigma}{\tau_{j+1}})p(t_j) + \frac{\sigma}{\tau_{j+1}}\, q$. Wegen

$$\begin{aligned} \|p(t_j + \sigma) - \tilde{q}(\sigma)\|_{1,\Omega} &\leq \|p(t_j + \sigma) - I_\Delta p(t_j + \sigma)\|_{1,\Omega} + \|\tilde{q}(\sigma) - I_\Delta p(t_j + \sigma)\|_{1,\Omega} \\ &\leq \underbrace{C_1\tau_\Delta^2}_{\leq \frac{\delta}{2}\text{ , da } \tau^* \leq \varepsilon_0} + \underbrace{\|q - p(t_{j+1})\|_{1,\Omega}}_{\leq \rho_1 < \frac{\delta}{2}} < \delta \end{aligned}$$

gilt:

$$\|p(t_{j+1}) - D(q)\|_{1,\Omega} \leq c_1 \int_0^{\tau_{j+1}} \|f(t_j+\sigma, p(t_j+\sigma)) - f(t_j+\sigma, \tilde{q}(\sigma))\|_{0,\Omega} + \underbrace{\|d_{j+1}\|_{1,\Omega}}_{\leq \hat{C}\tau_{j+1}^2 \text{ vgl. (3.31)}}$$

$$\leq c_1 L \int_0^{\tau_{j+1}} \underbrace{\|p(t_j+\sigma) - \tilde{q}(\sigma)\|_{1,\Omega}}_{<\delta} + \hat{C}\tau_{j+1} = \tau_{j+1}(c_1 L\delta + \hat{C}).$$

Für $\tau^* \leq \min\{\varepsilon_0, \frac{\rho_1}{(c_1 L\delta + \hat{C})}\} =: \varepsilon_1$ ist $D(q) \in \overline{U(p(t_{j+1}), \rho_1)}$ und damit ist D wohldefiniert. Ferner gilt:

$$\|D(q_1) - D(q_2)\|_{1,\Omega} \leq c_1 L\tau_{j+1}\|q_1 - q_2\|_{1,\Omega}.$$

Für $\tau^* \leq \min\{\varepsilon_1, \frac{1}{2c_1 L}\} =: \varepsilon_2$ gilt für beliebiges $q_1, q_2 \in \overline{U(p(t_{j+1}), \rho_1)}$

$$\|D(q_1) - D(q_2)\|_{1,\Omega} \leq \frac{1}{2}\|q_1 - q_2\|_{1,\Omega}.$$

Die Iteration $q_0 \in \overline{U(p(t_{j+1}), \rho_1)}$ beliebig, $q_{m+1} := D(q_m)$ ist nach dem Banach'schen Fixpunktsatz eine Fixpunktiteration. Für $p^\bullet := \lim_{m\to\infty} q_m$ gilt $\|p^\bullet - p(t_{j+1})\|_{1,\Omega} \leq \rho_1$. Als nächstes sei die Abbildung

$$\tilde{D} : \overline{U(p^\bullet, \rho_2)} \subset H^1_{\Gamma_D}(\Omega) \to \overline{U(p^\bullet, \rho_2)}$$

mit

$$\tilde{D} = r_{2,2}(\tau_{j+1}A)p_\tau(t_j) - \int_0^{\tau_{j+1}} r_{0,2}(\tau_{j+1}A)(I + \sigma A)f(t_j+\sigma, \hat{q}(\sigma))d\sigma$$

definiert, wobei $\hat{q}(\sigma) := (1 - \frac{\sigma}{\tau_{j+1}})p_\tau(t_j) + \frac{\sigma}{\tau_{j+1}}\, q$. Wegen Induktionsvoraussetzung folgt $\|\tilde{p}^\bullet(\sigma) - \hat{q}(\sigma)\|_{1,\Omega} \leq \max\{\|p(t_j) - p_\tau(t_j)\|_{1,\Omega}, \|p^\bullet - q\|_{1,\Omega}\} \leq \max\{\frac{j\delta}{3M_\Delta}, \rho_2\} = \rho_2 < \frac{\delta}{2}$. Es gilt also

$$\begin{aligned}\|p^\bullet - \tilde{D}(q)\|_{1,\Omega} =& \|D(p^\bullet) - \tilde{D}(q)\|_{1,\Omega} \\ \leq& \|p(t_j) - p_\tau(t_j)\|_{1,\Omega} + c_1 \int_0^{\tau_{j+1}} \|f(t_j+\sigma, \tilde{p}^\bullet(\sigma)) - f(t_j+\sigma, \hat{q}(\sigma))\|_{0,\Omega} \\ \leq& \frac{j\delta}{3M_\Delta} + c_1 L\tau_{j+1}\rho_2.\end{aligned}$$

Für $\tau^* \leq \min\{\varepsilon_2, \frac{\rho_1}{c_1 L\rho_2}\} =: \varepsilon_3$ gilt $\frac{j\delta}{3M_\Delta} + c_1 L\tau_{j+1}\rho_2 \leq \frac{j\delta}{3M_\Delta} + \frac{\delta}{6M_\Delta} = \frac{(2j+1)\delta}{6M_\Delta} = \rho_2$, d.h. $\tilde{D}$ ist wohldefiniert. Wegen $\tau^* \leq \varepsilon_2$ folgt: für beliebiges $q_1, q_2 \in \overline{U(p^\bullet, \rho_2)}$

$$\|\tilde{D}(q_1) - \tilde{D}(q_2)\|_{1,\Omega} \leq \frac{1}{2}\|q_1 - q_2\|_{1,\Omega}$$

und damit stellt die Folge $q_0 \in \overline{U(p^\bullet, \rho_2)}$ beliebig, $q_{m+1} := \tilde{D}(q_m)$ nach dem Banach'schen Fixpunktsatz eine Fixpunktiteration dar. Für $p_\tau(t_{j+1}) = \lim_{m\to\infty} q_{m+1}$ gilt $\|p^\bullet - p_\tau(t_{j+1})\|_{1,\Omega} \leq \rho_2$. Insgesamt folgt:

$$\|p(t_{j+1}) - p_\tau(t_{j+1})\|_{1,\Omega} \leq \|p^\bullet - p(t_{j+1})\|_{1,\Omega} + \|p^\bullet - p_\tau(t_{j+1})\|_{1,\Omega} \leq \rho_1 + \rho_2 = \frac{(j+1)\delta}{3M_\Delta}.$$

■

Bemerkung 3.6

(a) Für die Untersuchung des Minimierungsverfahren (3.4) haben wir folgende Schritte gemacht:

– Sei

$$\tilde{\mathcal{F}}(u_\tau, p_\tau) = \int_0^T \Big(\tau_\Delta \, \|\partial_t p_\tau(t) + \ \text{div}\ u_\tau(t) + f(t, p_\tau(t))\|_{0,\Omega}^2 + \|u_\tau(t) + \nabla p_\tau(t)\|_{0,\Omega}^2 \Big)\, dt. \tag{3.35}$$

Wir haben mit der Variationsformulierung (3.11) eine Approximation $(-\nabla p_\tau, p_\tau)$ der Lösung (u, p) berechnet. Es sei $u_\tau\big|_{[j\tau_\Delta,(j+1)\tau_\Delta)} \equiv -\nabla p_\tau(j\tau_\Delta)$, $j = 0, 1, \ldots, M_\Delta - 1$, dann gilt $u_\tau \in V_\tau$.
Es stellt sich die Frage: wie nah ist die für die theoretische Analyse konstruierte Approximation (u_τ, p_τ) an der Lösung $(u_{\min}, p_{\min})$ der Minimierungsaufgabe: $\tilde{\mathcal{F}}(u_{\min}, p_{\min}) = \min\limits_{(v_\tau, q_\tau) \in V_\tau \times Q_\tau} \tilde{\mathcal{F}}(v_\tau, q_\tau)$?

Zur einfacheren Übersicht lassen wir im Folgenden nur konstante Zeitschrittweiten zu:

– Angenommen, $(\tilde{\mathcal{F}}(v_\tau, q_\tau))^{1/2} \eqsim |||u - v_\tau, p - q_\tau|||$ ($|||\cdot|||$ wurde in Kapitel 2 definiert). Weiter sei angenommen, daß $\max\limits_{t\in[0,T]} \| \text{div}\ (u_\tau - u)\|_{0,\Omega} \lesssim \tau_\Delta^\alpha$ für ein $\alpha > 0$. Dann gilt:

$$\begin{aligned}
|||(u_{\min} - u, p_{\min} - p)|||^2 &\eqsim \tilde{\mathcal{F}}(u_{\min}, p_{\min}) \le \tilde{\mathcal{F}}(u_\tau, p_\tau) \eqsim |||(u_\tau - u, p_\tau - p)|||^2 \\
&= \int_0^T \Bigg(\underbrace{\|u_\tau(t) - u(t)\|_{0,\Omega}^2}_{\lesssim \tau_\Delta^4 \text{ Folgerung aus Satz 3.5}} + \underbrace{\tau_\Delta \| \text{div}\ (u_\tau(t) - u(t))\|_{0,\Omega}^2}_{\lesssim \tau_\Delta^{2\alpha+1} \text{ vgl. obige Annahme}} \\
&\quad + \underbrace{\frac{1}{\tau_\Delta} \|p_\tau(t) - p(t)\|_{0,\Omega}^2}_{\lesssim \tau_\Delta^3 \text{ Folgerung aus Satz 3.5}} + \underbrace{\|\nabla p_\tau(t) - \nabla p(t)\|_{0,\Omega}^2}_{\lesssim \tau_\Delta^4 \text{ Folgerung aus Satz 3.5}} \Bigg) dt \\
&\lesssim \tau_\Delta^{\min(3, 2\alpha+1)}
\end{aligned}$$

– Mit den obigen Annahmen gilt für genügend kleine τ_Δ $(u_\tau, p_\tau) \approx (u_{\min}, p_{\min})$. Damit kann einerseits gesichert werden, daß $(u_{\min}, p_{\min})$ die Konvergenzeigenschaften von (u_τ, p_τ) hat. Andererseits kann garantiert werden, daß $\hat{\mathcal{F}}(u_\tau, p_\tau)$ einen Fehlerschätzer hervorbringt, mit dessen Hilfe man optimale Zeitschrittweiten bestimmen kann.

(b) Zur Lösung von (3.4) bzw. (3.11) ist in jedem Zeitschritt mindestens ein nichtlineares Problem zu lösen, das mit Newton-ähnlichen Verfahren behandelt werden kann. Wir werden im nächsten Kapitel die Lösung von (3.4) in einem anderen nichtlinearen Fall (Gleichungen aus porösen Medien) mit Hilfe des Gauß-Newton-Algorithmus für nichtlineare Ausgleichprobleme (vgl. [17, Abschnitt 4.3]) lösen.

Ist das System (3.2) autonom, d.h. $f : H^1_{\Gamma_D}(\Omega) \to L^2(\Omega)$, so kann man für die Methode (3.4) unter bestimmten Bedingungen die Konvergenzordnung 2 ohne Weglassen des f'-Terms beweisen.

Für $\alpha \geq 0$ ist

$$H_\alpha := D_{A^{\frac{\alpha}{2}}} \subset L^2(\Omega)$$

mit dem Skalarprodukt

$$< q_1, q_2 >_\alpha := (A^{\frac{\alpha}{2}} q_1, A^{\frac{\alpha}{2}} q_2)_{0,\Omega}$$

ein Hilbertraum (vgl. [38, Seite 195 ff.]). Es sei $\|\cdot\|_\alpha := <\cdot,\cdot>_\alpha^{\frac{1}{2}}$ die zugehörige Norm. Außerdem sei

$$H_{-\alpha} := (H_\alpha)' = \left\{ \tilde{q} \in \mathcal{D}' \;\middle|\; \sup_{q \in H_\alpha} \frac{<\tilde{q}, q>}{\|q\|_\alpha} < \infty \right\}.$$

Definiert man $A^{-\frac{\alpha}{2}}\tilde{q}$ für stetige lineare Funktionale $\tilde{q} : H_\alpha \to \mathbb{R}$ durch die Vorschrift $< A^{-\frac{\alpha}{2}}\tilde{q}, q > := < \tilde{q}, \underbrace{A^{-\frac{\alpha}{2}} q}_{=:\hat{q}} >$ für $\hat{q} \in H_\alpha$, so ist $\underbrace{A^{\frac{\alpha}{2}}\hat{q}}_{=q} \in L^2(\Omega)$, d.h. $A^{-\frac{\alpha}{2}}\tilde{q} \in (L^2(\Omega))' = L^2(\Omega)$. Gilt umgekehrt für ein $\tilde{q} \in \mathcal{D}'$, daß $A^{-\frac{\alpha}{2}}\tilde{q} \in L^2(\Omega)$, so ist

$$\sup_{q \in H_\alpha} \frac{<\tilde{q}, q>}{\|q\|_\alpha} = \sup_{q \in H_\alpha} \frac{<A^{-\frac{\alpha}{2}}\tilde{q}, A^{\frac{\alpha}{2}} q>}{\|q\|_\alpha} = \sup_{q \in H_\alpha} \frac{(A^{-\frac{\alpha}{2}}\tilde{q}, A^{\frac{\alpha}{2}} q)_{0,\Omega}}{\|q\|_\alpha} \leq \|A^{-\frac{\alpha}{2}}\tilde{q}\|_{0,\Omega} < \infty.$$

Damit ist

$$H_{-\alpha} = \left\{ \tilde{q} \in \mathcal{D}' \;\middle|\; A^{-\frac{\alpha}{2}}\tilde{q} \in L^2(\Omega) \right\} =: D_{A^{-\frac{\alpha}{2}}}.$$

Wir setzen voraus, daß die Funktion

$$g(\tau, \sigma, q) = r_{0,2}(\tau A)\, \sigma f'(q)\, (Aq \;\; + f(q)) \;\; \forall\, q \in H^1_{\Gamma_D}(\Omega), \tag{3.36}$$

für alle $\sigma, \tau \in [0, T]$, $\sigma \leq \tau$ lokal lipschitzstetig ist. Dabei muß man $f : H_{-1} \to H_{-2}$ fordern mit $f(H_1) \subset H_0 = L^2(\Omega)$, damit $g(\tau, \sigma, \cdot)$ wohldefiniert ist. Es gilt dann,

$$H_1 \overset{A(\cdot)+f(\cdot)}{\longrightarrow} H_{-1} \overset{\sigma f'(q)[\cdot]}{\longrightarrow} H_{-2} \overset{r_{0,2}(\tau A)}{\longrightarrow} H_2 \subset H_1 \underbrace{=}_{\text{Satz B.7(c)}} H^1_{\Gamma_D}(\Omega).$$

Wir bezeichnen die Lipschitzkonstante für $g(\tau, \sigma, \cdot)$ in einer Umgebung $U(p(t), \tilde{\delta}_{t,\tau,\sigma}) \subset H^1_{\Gamma_D}(\Omega)$ der Lösung $p(t)$ mit $\tilde{L}_{t,\tau,\sigma}$. Seien $\tilde{L} = \max\limits_{t,\tau,\sigma \in [0,T]} \{\tilde{L}_{t,\tau,\sigma}\}$, $\tilde{\delta} = \min\limits_{t,\tau,\sigma \in [0,T]} \{\tilde{\delta}_{t,\tau,\sigma}\} > 0$, $\hat{\delta} = \min\{\tilde{\delta}, \delta\}$ und $\hat{L} = \max\{\tilde{L}, L\}$. Für $q \in H^1_{\Gamma_D}(\Omega)$ mit $\|q - p(t)\|_{1,\Omega} < \hat{\delta}$ folgt wegen der Konstruktion: $\|f(q) - f(p(t))\|_{0,\Omega} \leq \hat{L}\|q - p(t)\|_{1,\Omega}$ bzw. $\|g(\tau, \sigma, q) - g(\tau, \sigma, p(t))\|_{1,\Omega} \leq \hat{L}\|q - p(t)\|_{1,\Omega}$. Wir bezeichnen die Lösung von (3.4) wieder mit (u_τ, p_τ) statt $(u_{\min}, p_{\min})$, da das Verfahren (3.13) nicht mehr betrachtet wird. Aus Satz 3.1 folgern wir für (u_τ, p_τ), daß

$$\int_0^\tau \big[I + \sigma(A + f'(p_\tau(t+\sigma)))\big] \big[\partial_t p_\tau(t+\sigma) + A p_\tau(t+\sigma) + f(p_\tau(t+\sigma))\big]\, d\sigma = 0.$$

Ausrechnen des Integrals ergibt

$$\begin{aligned} p_\tau^+ &= r_{2,2}(\tau A)p_\tau^- - r_{0,2}(\tau A)\int_0^\tau (I+\sigma A)f(p_\tau(t+\sigma))\,d\sigma \\ &\quad - r_{0,2}(\tau A)\int_0^\tau \sigma f'(p_\tau(t+\sigma))(\partial_t p_\tau(t+\sigma) + Ap_\tau(t+\sigma) + f(p_\tau(t+\sigma)))\,d\sigma. \end{aligned} \tag{3.37}$$

Wir sind nun in der Lage den Konvergenzsatz für den autonomen Fall ohne Weglassen des f'-Terms zu zeigen.

Satz 3.7 Es sei $f : H_{-\frac{1}{2}} \to H_{-1}$ mit $f_{\big|_{H^1_{\Gamma_D}(\Omega)}} \in C^1(H^1_{\Gamma_D}(\Omega), L^2(\Omega))$, $p \in C^2([0,T], H^1_{\Gamma_D}(\Omega))$ die exakte Lösung von (3.2). Außerdem sei die in (3.36) definierte Funktion $g(\tau,\sigma,\cdot)$ für alle $\sigma,\tau \in [0,T]$, $\sigma \le \tau$ lokal lipschitzstetig. Für die Lösung von (3.4) gilt dann:
Es existiert ein τ^*, so daß für alle Gitter $\Delta \subset [0,T]$ mit $\tau_\Delta \le \tau^*$ gilt:

$$\|p_\tau(t_n) - p(t_n)\|_{1,\Omega} \le C\,\tau_\Delta^2\, e^{C(n+1)\tau_\Delta^{1/2}} \tag{3.38}$$

für $0 \le n \le M_\Delta$, wobei $C > 0$ nur von $\sup\limits_{t\in[0,T]}\{\|\partial_t^2 p(t)\|_{1,\Omega}\}$ abhängt.

Beweis:
Die Beweise der Teile (a) bzw. (c) verlaufen analog zum Beweis von Teil (a) bzw. (c) von Satz 3.5.
(b) Wir benutzen erneut die Bezeichnungen aus dem Beweis von Satz 3.5.
(i) Sei $0 < n \le M_\Delta$. Setzt man die exakte Lösung in (3.37) ein, so ist

$$\begin{aligned} \hat p_n &= r_{2,2}(\tau_n A)\hat p_{n-1} - r_{0,2}(\tau_n A)\int_0^{\tau_n}(I+\sigma A)f(p(t_{n-1}+\sigma))\,d\sigma \\ &\quad - r_{0,2}(\tau_n A)\int_0^{\tau_n}\sigma f'(p(t_n+\sigma))\Big(\partial_t p(t_n+\sigma) + Ap(t_n+\sigma) \\ &\qquad\qquad + f(p(t_n+\sigma))\Big)\,d\sigma + d_n, \end{aligned}$$

wobei d_n wieder den Defekt darstellt. Der Defekt ist gegeben durch:

$$\begin{aligned} d_n &= -\,r_{2,2}(\tau_n A)\hat p_{n-1} + \hat p_n \\ &\quad - \int_0^{\tau_n}(I+\sigma A)\,r_{0,2}(\tau_n A)(\partial_t p(t_{n-1}+\sigma) + Ap(t_{n-1}+\sigma))\,d\sigma \\ &\quad + r_{0,2}(\tau_n A)\int_0^{\tau_n}\sigma f'(p(t_n+\sigma))\,\underbrace{(\partial_t p(t_n+\sigma) + Ap(t_n+\sigma) + f(p(t_n+\sigma)))}_{=0}\,d\sigma. \end{aligned}$$

Dies ist genau der gleiche Defekt, der im Beweis von Satz 3.5 vorkommt. Also folgern wir (3.25).
(ii) Die Kettenregel liefert

$$\partial_t f(q(t)) = f'(q(t))\;\partial_t q(t)$$

für $q \in H^1((0,T); H^1_{\Gamma_D}(\Omega))$. Mit partieller Integration erhalten wir:

$$\begin{aligned}&\int_0^{\tau_n} \sigma f'(p_\tau(t_{n-1}+\sigma))\, \partial_t p_\tau(t_{n-1}+\sigma) - \sigma f'(p(t_{n-1}+\sigma))\, \partial_t p(t_{n-1}+\sigma)\, d\sigma \\ &= \tau_n \left(f(p_n) - f(\hat{p}_n)\right) - \int_0^{\tau_n} \tilde{\Sigma}_n(\sigma) + \hat{\Sigma}_n(\sigma)\, d\sigma.\end{aligned} \tag{3.39}$$

Insgesamt gilt also wegen (3.39) und (3.36):

$$\begin{aligned}e_n = \; & r_{2,2}(\tau_n A) e_{n-1} + \tau_n r_{0,2}(\tau_n A)(f(p_n) - f(\hat{p}_n)) \\ & + \int_0^{\tau_n} r_{0,2}(\tau_n A)\sigma A \tilde{\Sigma}_n(\sigma)\, d\sigma + \int_0^{\tau_n} r_{0,2}(\tau_n A)\sigma A \hat{\Sigma}_n(\sigma)\, d\sigma \\ & + \int_0^{\tau_n} \underbrace{g(\tau_n, \sigma, p_\tau(t_{n-1}+\sigma)) - g(\tau_n, \sigma, p(t_{n-1}+\sigma))}_{=:\mathcal{G}_n(\sigma)} d\sigma + d_n.\end{aligned}$$

Rekursiv erhalten wir

$$\begin{aligned}e_n = & \sum_{j=0}^{n-1} R(t_{j+1}, t_n) \Big[d_{j+1} + \tau_n r_{0,2}(\tau_n A)(f(p_n) - f(\hat{p}_n)) \\ & \int_0^{\tau_{j+1}} r_{0,2}(\tau_{j+1} A)(\sigma A)(\tilde{\Sigma}_{j+1}(\sigma) + \hat{\Sigma}_{j+1}(\sigma)) + \mathcal{G}_{j+1}(\sigma) d\sigma \Big].\end{aligned} \tag{3.40}$$

Beachten wir, daß für $j = 1, \ldots, M_\Delta$, $q \in H^1_{\Gamma_D}(\Omega)$ und $\sigma \in [0, \tau]$ gilt

$$\|\sigma A\, r_{0,2}(\tau A) q\|_{1,\Omega} \leq \|(I + \sigma A)\, r_{0,2}(\tau A) q\|_{1,\Omega} \leq \|(I + \tau A)\, r_{0,2}(\tau A) q\|_{1,\Omega}$$

und setzen wieder $S(t_{j+1}, t_n) = (I + \tau_{j+1} A) r_{0,2}(\tau_{j+1} A) R(t_{j+1}, t_n)$, so folgt aus (3.40), lokale Lipschitzstetigkeit von $g(\tau_j, \sigma, \cdot)$ und A-Stabilität des Verfahrens

$$\begin{aligned}\|e_n\|_{1,\Omega} \lesssim & \left\| \sum_{j=0}^{n-1} S(t_{j+1}, t_n) \Big(\int_0^{\tau_{j+1}} \hat{\Sigma}_{j+1}(\sigma) d\sigma + \tau_{j+1} \hat{\Sigma}_{j+1}(\tau_{j+1}) \Big) \right\|_{1,\Omega} \\ & + \left\| \sum_{j=0}^{n-1} R(t_{j+1}, t_n) d_{j+1} \right\|_{1,\Omega} \\ & + \left\| \sum_{j=0}^{n-1} S(t_{j+1}, t_n) \Big(\int_0^{\tau_{j+1}} \tilde{\Sigma}_{j+1}(\sigma)\, d\sigma + \tau_{j+1} \tilde{\Sigma}_{j+1}(\tau_{j+1}) \Big) \right\|_{1,\Omega} \\ & + \left\| \sum_{j=0}^{n-1} \int_0^{\tau_{j+1}} p_\tau(t_j + \sigma) - p(t_j + \sigma) d\sigma \right\|_{1,\Omega} \\ & =: \|\tilde{I}_1\|_{1,\Omega} + \|\tilde{I}_2\|_{1,\Omega} + \|I_3\|_{1,\Omega} + \|\tilde{I}_4\|_{1,\Omega}.\end{aligned}$$

Die Terme $\|\tilde{I}_i\|_{1,\Omega}, i = 1, 2$ sind analog zu $\|I_i\|_{1,\Omega}, i = 1, 2$ abzuschätzen (vgl. Beweis vonSatz 3.5). Für $\|\tilde{I}_4\|_{1,\Omega}$ gilt:

$$\|\tilde{I}_4\|_{1,\Omega} \lesssim \tau_\Delta^2 + \tau_\Delta \sum_{j=0}^{n-1} (\|e_j\|_{1,\Omega} + \|e_{j+1}\|_{1,\Omega}).$$

Der Rest des Beweises verläuft analog zum Beweis von Satz 3.5.

■

Wir wollen nun eine Konvergenzaussage für alle $t \in [0,T]$.

Korollar 3.8 Bezeichnungen und Voraussetzungen seien wie in Satz 3.7. Dann ist

$$\max_{t\in[0,T]} \|p(t) - p_\tau(t)\|_{1,\Omega} \leq C\tau_\Delta^2,$$

wobei C nur von $\sup_{t\in[0,T]} \{\|\partial_t^2 p(t)\|_{1,\Omega}\}$ abhängt.

Beweis:
Sei $\hat{t} \in [0,T]$ mit

$$\max_{t\in[0,T]} \|p(t) - p_\tau(t)\|_{1,\Omega} = \|p(\hat{t}) - p_\tau(\hat{t})\|_{1,\Omega}.$$

Aus der Dreiecksungleichung folgern wir

$$\|p(\hat{t}) - p_\tau(\hat{t})\|_{1,\Omega} \leq \|p(\hat{t}) - I_\Delta p(\hat{t})\|_{1,\Omega} + \|I_\Delta p(\hat{t}) - p_\tau(\hat{t})\|_{1,\Omega}.$$

Für den ersten Term folgt aus der Interpolationstheorie

$$\|p(\hat{t}) - I_\Delta p(\hat{t})\|_{1,\Omega} \leq C_1\tau_\Delta^2. \tag{3.41}$$

Wähle $0 \leq n \leq M_\Delta$ so, daß $\hat{t} \in [t_n, t_{n+1}]$. Wegen der Linearität der beiden Funktionen $I_\Delta p$ und p_τ auf $[t_n, t_{n+1}]$ gilt

$$\|I_\Delta p(\hat{t}) - p_\tau(\hat{t})\|_{1,\Omega} \leq \max\{\|e_n\|_{0,\Omega}, \|e_{n+1}\|_{1,\Omega}\} \underbrace{\leq}_{\text{Satz 3.7}} C^*\tau_\Delta^2.$$

■

Für den allgemeineren Fall kann man mit Satz 3.5 elementar auch zeigen:

$$\max_{t\in[0,T]} \|p(t) - p_\tau(t)\|_{1,\Omega} \leq C\tau_\Delta^2.$$

Der Beweis verläuft analog zum Beweis von Korollar 3.8.

3.4 Nähere Betrachtung der Minimierungsaufgabe

$\hat{\mathcal{F}}$ soll mit Hilfe der Simpsonregel wie folgt approximiert werden:

$$\begin{aligned}
\hat{\mathcal{F}}(u_\tau, p_\tau) \approx\ & \frac{1}{6}\|p_\tau^+ - p_\tau(t) + \tau\ \text{div}\ u_\tau^- + \tau\, f(t, p_\tau(t))\|_{0,\Omega}^2 \\
& + \frac{2}{3}\|p_\tau^+ - p_\tau(t) + \tau\,\frac{1}{2}\,(\ \text{div}\ u_\tau^- + \ \text{div}\ u_\tau^+) \\
& \qquad + \tau f(t + \frac{\tau}{2}\,, \frac{p_\tau^+ + p_\tau(t)}{2})\|_{0,\Omega}^2 \\
& + \frac{1}{6}\|p_\tau^+ - p_\tau(t) + \tau\ \text{div}\ u_\tau^+ + \tau f(t+\tau, p_\tau^+)\|_{0,\Omega}^2 \\
& + \frac{\tau}{6}\|u_\tau^- + a^{1/2}\nabla p_\tau(t)\|_{0,\Omega}^2 + \frac{\tau}{6}\|u_\tau^+ + \nabla p_\tau^+\|_{0,\Omega}^2 \\
& + \frac{\tau}{6}\|(u_\tau^- + u_\tau^+) + (\nabla p_\tau(t) + \nabla p_\tau^+)\|_{0,\Omega}^2 = \mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)).
\end{aligned}$$

Wir wollen nun statt $\hat{\mathcal{F}}$ das Funktional $\mathcal{F}$ in den entsprechenden Räumen $H_{\Gamma_N}(\text{div }, \Omega)^2 \times H^1_{\Gamma_D}(\Omega)$ minimieren.
Wie man leicht sieht, ist

$$\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) = \|\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t))\|^2_{0,\Omega}.$$

mit

$$\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) = \begin{pmatrix} \frac{1}{\sqrt{6}} \left(p_\tau^+ - p_\tau(t) + \tau \text{ div } u_\tau^- + \tau\, f(t, p_\tau(t))\right) \\ \sqrt{\frac{2}{3}} \left(p_\tau^+ - p_\tau(t) + \frac{\tau}{2} \text{ div } (u_\tau^- + u_\tau^+) + \tau\, f(t + \frac{\tau}{2}\,, \frac{p_\tau^+ + p_\tau(t)}{2})\right) \\ \frac{1}{\sqrt{6}} \left(p_\tau^+ - p_\tau(t) + \tau \text{ div } u_\tau^+ + \tau f(t + \tau, p_\tau^+)\right) \\ \sqrt{\frac{\tau}{6}} \left(\, u_\tau^- + \, \nabla p_\tau(t)\right) \\ \sqrt{\frac{\tau}{6}} \left(\, u_\tau^- + u_\tau^+ + \, \nabla p_\tau(t) + \nabla p_\tau^+\right) \\ \sqrt{\frac{\tau}{6}} \left(\, u_\tau^+ + \, \nabla p_\tau^+\right) \end{pmatrix}.$$

Betrachten wir das Funktional $\mathcal{F}$ auf $H_{\Gamma_N}(\text{div }, \Omega)^2 \times H^1_{\Gamma_D}(\Omega)$, so folgt aus (3.4), daß für die Fréchet-Ableitung

$$\mathcal{F}'(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) \equiv 0 \quad \text{auf } H_{\Gamma_N}(\text{div }, \Omega)^2 \times H^1_{\Gamma_D}(\Omega) \tag{3.42}$$

gilt, mit

$$\mathcal{F}'(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) \;:\; H_{\Gamma_N}(\text{div }, \Omega)^2 \times H^1_{\Gamma_D}(\Omega) \longrightarrow \mathbb{R},$$

$$\begin{aligned} &\mathcal{F}'(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) \begin{pmatrix} v_\tau^- \\ v_\tau^+ \\ q_\tau \end{pmatrix} \\ &\qquad = \left(\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)), \mathcal{J}_\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+) \begin{pmatrix} v_\tau^- \\ v_\tau^+ \\ q_\tau \end{pmatrix}\right)_{0,\Omega}, \end{aligned} \tag{3.43}$$

wobei

$$\mathcal{J}_\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+) = (\partial_j \mathcal{R}_i)_{i=1,\ldots,6, j=1,\ldots,3}$$

die *Jacobi-Matrix* von $\mathcal{R}$ an der Stelle $(u_\tau^-, u_\tau^+, p_\tau^+)$ bezeichnet ($\partial_j \mathcal{R}_i$ ist die Ableitung der i-ten Zeile von $\mathcal{R}$ nach der j-ten Variable). Mittels elementarer Analysis ergibt sich:

$$\mathcal{J}_\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+) \;:\; H_{\Gamma_N}(\text{div }, \Omega)^2 \times H^1_{\Gamma_D}(\Omega) \longrightarrow (L^2(\Omega))^9,$$

wobei

$$\mathcal{J}_{\mathcal{R}}(u_\tau^-, u_\tau^+, p_\tau^+) = \frac{1}{\sqrt{6}} \begin{pmatrix} \tau \text{ div} & 0 & 1 \\ \tau \text{ div} & \tau \text{ div} & 2 + \tau \partial_p f(t + \frac{\tau}{2}, \frac{p_\tau^+ + p_\tau(t)}{2}) \\ 0 & \tau \text{ div} & 1 + \tau \partial_p f(t + \tau, p_\tau^+) \\ \sqrt{\tau} & 0 & 0 \\ \sqrt{\tau} & \sqrt{\tau} & \sqrt{\tau}\nabla \\ 0 & \sqrt{\tau} & \sqrt{\tau}\nabla \end{pmatrix}.$$

Aus (3.42) und (3.43) folgt, daß (3.4) äquivalent ist zu

$$\left(\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)), \mathcal{J}_{\mathcal{R}}(u_\tau^-, u_\tau^+, p_\tau^+) \begin{pmatrix} v_\tau^- \\ v_\tau^+ \\ q_\tau \end{pmatrix} \right)_{0,\Omega} = 0 \tag{3.45}$$

für alle $(v_\tau^-, v_\tau^+, q_\tau) \in V_\tau^2 \times Q_\tau$. Es gibt eine Reihe von Verfahren, die sich direkt mit dieser nichtlinearen Variationsformulierung auseinandersetzen. Wir nehmen an, daß in jedem Zeitschritt das Problem (3.45) bzw. (3.4) exakt gelöst wird. Dazu verwenden wir die gedämpfte Gauß-Newton-Methode (vgl. [36]). Wir werden den volldiskreten Algorithmus erst in Kapitel 4 entwickeln, weil die Idee der Adaptivität im Ort in diesem und im nächsten Kapitel die gleiche ist.

3.5 Numerische Beispiele

In diesem Abschnitt wollen wir einige Testbeispiele betrachten. In einigen dieser Beispiele vergleichen wir den neuen Algorithmus mit der derzeit bekannten diskontinuierlichen Galerkin-Methode mit $l = 1$ und $l = 0$ (vgl. Abschnitt 1.2.3). Dabei werden wir für die LSM in der Zeit den volldiskreten Algorithmus 4.1 verwenden, der in Kapitel 4 erläutert wird. Mit den gleichen Zeitschrittweiten werden wir dann die diskontinuierliche Galerkin-Methode durchlaufen, um die beiden Methoden miteinander vergleichen zu können. Die anstehenden nichtlinearen Probleme bei der diskontinuierlichen Galerkin-Methode werden durch eine Newton-Methode mit Dämpfungsstrategie gemäß [17, Abschnitt 4.2] behandelt. Wir werden den Fehler $\|p(t_j) - p_{\tau,h}(t_j)\|_{0,\Omega}$, $j = 1, \ldots, M$ für die beiden Methoden darstellen und vergleichen. Anschließend werden wir den Fehler $|||u - u_\tau, p - p_\tau|||_{\tau_j}$, $j = 1, \ldots, M$ und das LSF (den Fehlerschätzer) $\hat{\mathcal{F}}(u_\tau, p_\tau)$ miteinander vergleichen, wobei

$$\begin{aligned} |||(v,q)|||_{\tau_j} = \Bigg(\int_{t_{j-1}}^{t_j} & (\|v(\sigma)\|_{0,\Omega}^2 + \tau_j \,\| \text{ div } v(\sigma)\|_{0,\Omega}^2 \\ & + \frac{1}{\tau_j} \|q(\sigma)\|_{0,\Omega}^2 + \|\nabla\, q(\sigma)\|_{0,\Omega}^2) d\sigma \Bigg)^{1/2}. \end{aligned} \tag{3.46}$$

nicht exakt auswertbar ist.

Beispiel 3.9 Sei $\Omega = \left\{ x \in \mathbb{R}^2 \mid \|x\| < 1 \right\}$ der Einheitskreis. Betrachte das Problem

$$\partial_t p - \Delta p + f(p) = 0, \tag{3.47}$$

wobei $f(t, x, p) = 2t\|x\|^2 - 2t - 4t^2 - (t^2\|x\|^2 - t^2)^2 + p^2$.

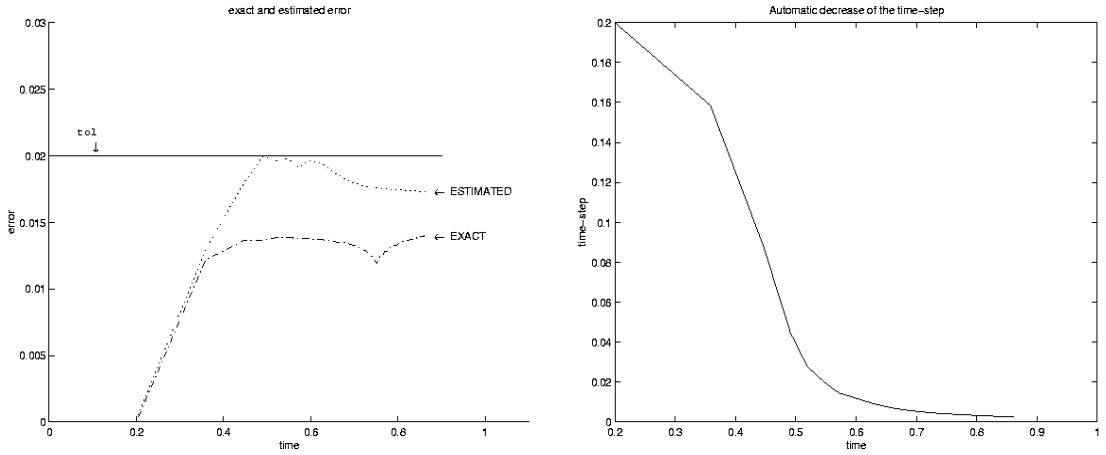

Abbildung 3.1: (links): Qualität des Fehlerschätzers und die Zeitschrittweiten für `tol` $= 0.02$; (rechts): die Abnahme der Zeitschrittweite im Laufe der Zeit (jeweils für Beispiel 3.9)

Wir starten mit $p(0, x) = 0$ und fordern $p\big|_{\partial\Omega} = 0$. Die exakte Lösung dieses Problems lautet:

$$p(t, x) = t^2\|x\|^2 - t^2.$$

Abbildung 3.1 zeigt auf der rechten Seite die Abnahme der Zeitschrittweiten. Auf der linken Seite der Abbildung 3.1 ist der exakte Fehler in der Norm aus (3.46) im Vergleich zu der Wurzel des LSF dargestellt, um die Qualität des Fehlerschätzers zu demonstrieren. In Abbildung 3.3 (links) ist die Lösung $p(1, x)$ auf der dritten Verfeinerungsstufe dargestellt.

Beispiel 3.10 Sei $\Omega = (-1, 1)^2$. Man betrachte das Problem (3.47), wobei

$$f(t, x, p) = -(x_1^2 - 1)(x_2^2 - 1) + 2t((x_1^2 - 1) + (x_2^2 - 1)) - t^3(x_1^2 - 1)^3(x_2^2 - 1)^3 + p^3.$$

Wir starten wieder mit $p(0, x) = 0$ und fordern $p\big|_{\partial\Omega} = 0$. Die exakte Lösung dieses Problems lautet:

$$p(t, x) = t(x_1^2 - 1)(x_2^2 - 1).$$

Abbildung 3.2 zeigt wieder auf der rechten Seite die Abnahme der Zeitschrittweiten. Auf der linken Seite der Abbildung 3.2 ist wieder der exakte Fehler in der Norm aus (3.46) im Vergleich zu der Wurzel des LSF dargestellt. Es wird nochmals betont, daß die Äquivalenz dieser beiden Größen noch nicht bewiesen ist.
In Tabelle 3.1 (1. bzw. 3. Spalte) wird der Fehler der Approximation an $p(0.2, x)$ mit verschiedenen Zeitschrittweiten aufgelistet; die 2. (bzw. 4.) Spalte listet den Quotienten aus dem L^2-Fehler (bzw. H^1-Fehler) des aktuellen Schritts (mit der Zeitschrittweite $\tilde{\tau}$)

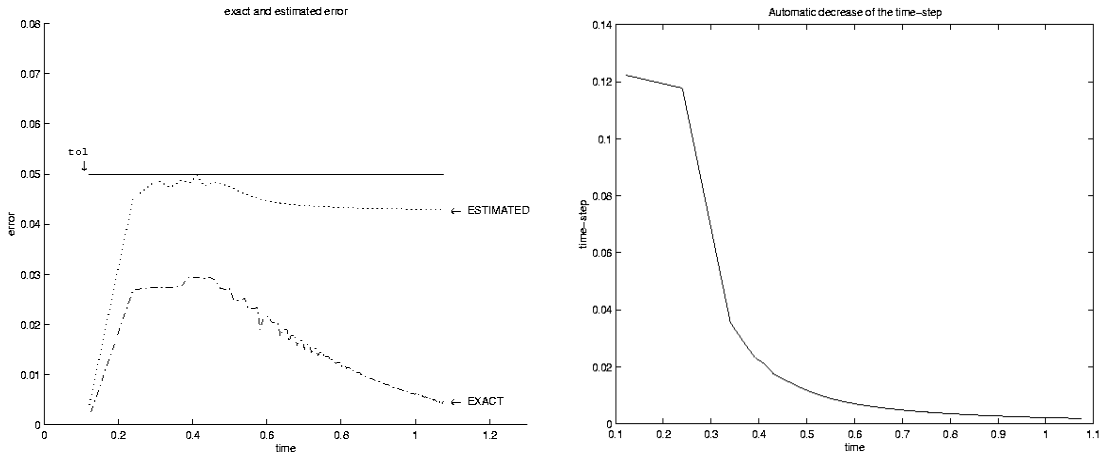

Abbildung 3.2: (links): Qualität des Fehlerschätzers und die Zeitschrittweiten für `tol` $= 0.05$; (rechts): die Abnahme der Zeitschrittweite im Laufe der Zeit (Beispiel 3.10)

Level=4	$\|p - p_{\tau,h}\|_{0,\Omega}$		$\|p - p_{\tau,h}\|_{1,\Omega}$	
$\tau = 0.2$	$31.472 \cdot 10^{-3}$	3.560	$94.417 \cdot 10^{-3}$	3.400
$\tau = 0.1$	$8.840 \cdot 10^{-3}$	3.599	$27.769 \cdot 10^{-3}$	3.826
$\tau = 0.05$	$2.465 \cdot 10^{-3}$	3.758	$7.258 \cdot 10^{-3}$	3.650
$\tau = 0.025$	$0.656 \cdot 10^{-3}$	3.905	$1.988 \cdot 10^{-3}$	3.921
$\tau = 0.0125$	$0.168 \cdot 10^{-3}$	–	$0.507 \cdot 10^{-3}$	–

Tabelle 3.1: Fehler zum Zeitpunkt $t = 0.2$ (Beispiel 3.10)

Level=4	$\mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))^{1/2}$	
$\tau = 0.2$	$3.994 \cdot 10^{-2}$	2.800
$\tau = 0.1$	$1.426 \cdot 10^{-2}$	2.597
$\tau = 0.05$	$0.549 \cdot 10^{-2}$	2.639
$\tau = 0.025$	$0.208 \cdot 10^{-2}$	2.773
$\tau = 0.0125$	$0.075 \cdot 10^{-2}$	–

Tabelle 3.2: LSF zum Zeitpunkt $t = 0.2$ (Beispiel 3.10)

Level=4	$\|\|\|u_{\tau,h} - u, p_{\tau,h} - p\|\|\|_\tau$	
$\tau = 0.2$	$2.567 \cdot 10^{-2}$	2.300
$\tau = 0.1$	$1.116 \cdot 10^{-2}$	2.542
$\tau = 0.05$	$0.439 \cdot 10^{-2}$	2.645
$\tau = 0.025$	$0.166 \cdot 10^{-2}$	2.690
$\tau = 0.0125$	$0.0617 \cdot 10^{-2}$	–

Tabelle 3.3: $|||\cdot|||_\tau$-Norm des Fehlers zum Zeitpunkt $t = 0.2$ (Beispiel 3.10)

und dem L^2-Fehler (bzw. H^1-Fehler) des nachfolgenden Schritts (mit der Zeitschrittweite $\frac{\tau}{2}$) auf. Es fällt auf, daß dieser Quotient gegen 2^2 strebt für $\tau \to 0$. Es scheint, daß die Konvergenzordnung 2 ist. Die Ordnung 2 wurde auch in Satz 3.5 (für die H^1-Norm des Fehlers) garantiert. Analog zu Tabelle 3.1 sind Tabelle 3.2 und Tabelle 3.3 zu verstehen. Allerdings wird diesmal $(\mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p_{\tau,h}; p_{\tau,h}(t)))^{1/2}$ bzw. $|||u_{\tau,h} - u, p_{\tau,h} - p|||_\tau$ aufge-

listet. Dabei strebt der ensprechende Quotient in beiden Fällen gegen $2^{3/2} \approx 2.828$. Es scheint, daß die Konvergenzordnung in der $|||\cdot|||_\tau$-Norm, wie im linearen Fall, $\frac{3}{2}$ ist. Andererseits zeigt sich hier, daß

$$(\mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p_{\tau,h}; p_{\tau,h}(t)))^{1/2} \eqsim |||u_{\tau,h} - u, p_{\tau,h} - p|||_\tau,$$

was auch zu erwarten ist. Es sei noch angemerkt, daß $|||u_{\tau,h} - u, p_{\tau,h} - p|||_\tau$ nicht exakt berechnet, sondern nur durch Quadraturformeln angenähert wird. In Abbildung 3.3 (rechts) ist die Lösung $p(1, x)$ auf der dritten Verfeinerungsstufe dargestellt.

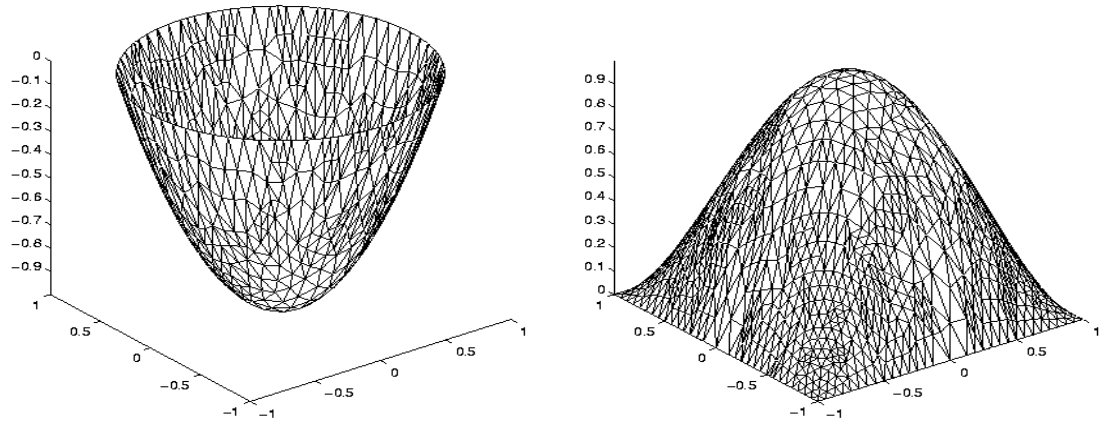

Abbildung 3.3: (links): Die Lösung von Beispiel 3.9 und (rechts): die Lösung von Beispiel 3.10 jeweils zum Zeitpunkt $t = 1$ auf Level 3

Beide Lösungen aus Beispiel 3.9 und 3.10 werden mit der Zeit größer.
Wie man leicht erkennt, gilt für beide Fälle $\lim_{t\to\infty} p(t, x) = \infty$. Diese beiden Beispiele haben wir gewählt, um zu testen, ob der Fehlerschätzer zuverlässig ist. Wir wollen nun eine Lösung approximieren, die im Laufe der Zeit gegen 0 strebt, um auch die Effizienz des Fehlerschätzers zu testen.

Beispiel 3.11 Sei $\Omega = (-1, 1)^2$. Man betrachte das Problem (3.47), wobei

$$f(t, x, p) = -e^{-t}(x_1^2 - 1)(x_2^2 - 1) + 2e^{-t}((x_1^2 - 1) + (x_2^2 - 1)) - e^{-3t}(x_1^2 - 1)^3(x_2^2 - 1)^3 + p^3.$$

Wir starten mit $p(0, x) = (x_1^2 - 1)(x_2^2 - 1)$ und fordern $p\big|_{\partial\Omega} = 0$. Die exakte Lösung dieses Problems lautet:

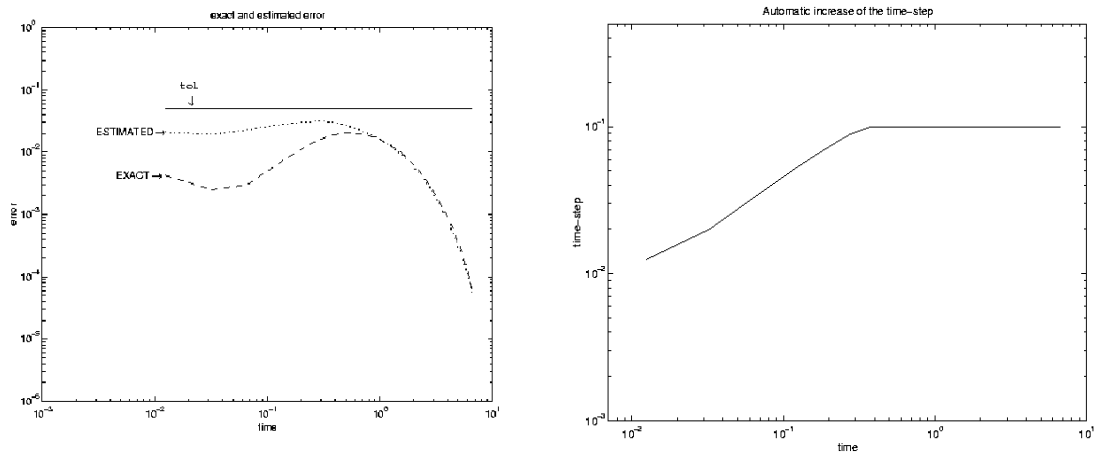

Abbildung 3.4: (links): Qualität des Fehlerschätzers und die Zeitschrittweiten für `tol` $= 0.05$; (rechts): die Zunahme der Zeitschrittweite im Laufe der Zeit ($\tau_{\max} = 0.1$) (Beispiel 3.11)

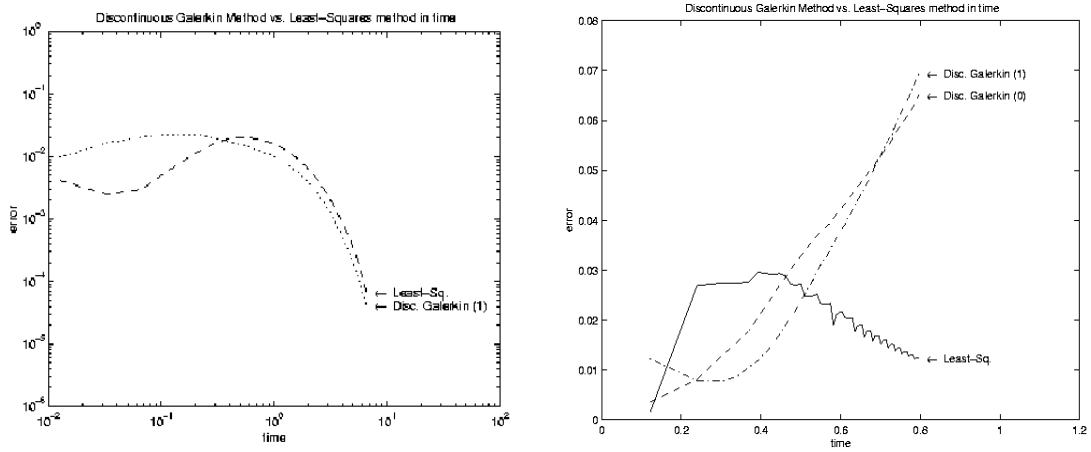

Abbildung 3.5: Vergleich zwischen der diskontinuierlichen Galerkin-Methode und dem FEM-Verfahren in der Zeit; (rechts): Beispiel 3.10; (links): Beispiel 3.11

Level=4	$\|p - p_{\tau,h}\|_{0,\Omega}$	
$\tau = 0.2$	$19.890 \cdot 10^{-3}$	3.135
$\tau = 0.1$	$6.344 \cdot 10^{-3}$	3.933
$\tau = 0.05$	$1.613 \cdot 10^{-3}$	3.700
$\tau = 0.025$	$0.436 \cdot 10^{-3}$	3.791
$\tau = 0.0125$	$0.115 \cdot 10^{-3}$	–

Tabelle 3.4: (LSM in der Zeit) Fehler zum Zeitpunkt $t = 0.2$ (Beispiel 3.11)

$$p(t, x) = e^{-t}(x_1^2 - 1)(x_2^2 - 1).$$

Wie aus den letzten Beispielen bekannt, wird in Abbildung 3.4 (links) der Fehler mit dem Fehlerschätzer verglichen. In Abbildung 3.4 (rechts) ist die Zunahme der Zeitschrittweite im Laufe der Zeit zu beobachten (bis die maximale Zeitschrittweite erreicht ist). Hier ist die Effizienz des Fehlerschätzers gut zu erkennen. In Abbildung 3.5 wird

Level=4	$\|p - p_{\tau,h}\|_{0,\Omega}$	
$\tau = 0.2$	$26.532 \cdot 10^{-3}$	3.096
$\tau = 0.1$	$8.569 \cdot 10^{-3}$	3.147
$\tau = 0.05$	$2.723 \cdot 10^{-3}$	3.518
$\tau = 0.025$	$0.774 \cdot 10^{-3}$	3.867
$\tau = 0.0125$	$0.270 \cdot 10^{-3}$	–

Tabelle 3.5: (diskontinuierliche Galerkin-Methode) Fehler zum Zeitpunkt $t = 0.2$ (Beispiel 3.11)

jeweils der Fehler in der $L^2(\Omega)$-Norm für die diskontinuierliche Galerkin-Methode und die LSM in der Zeit dargestellt (in Abhängigkeit der Zeit). In den Tabellen 3.4 bzw. 3.5 (1. Spalte) ist wieder der Fehler der Approximation an $p(0.2, x)$ mit verschiedenen Zeitschrittweiten für die LSM in der Zeit bzw. die diskontinuierliche Galerkin-Methode aufgelistet. Es scheint, daß beide Verfahren die Konvergenzordnung 2 aufweisen.

Kapitel 4

Nichtlineare Probleme zur Beschreibung von Strömungen in teilgesättigten porösen Medien

Im Folgenden stellen wir ein Modell aus der Bodenmechanik vor. Im Anschluß erweitern wir die LSM in der Zeit aus den letzten Kapiteln auf das vorgestellte Modell. Danach runden numerische Ergebnisse dieses Kapitel ab.

4.1 Herleitung der Gleichungen für die Beschreibung

Das Ziel ist es, für einen beschränkten Bereich $\Omega \in \mathbb{R}^d$, $d \in \{1, 2, 3\}$, bei Kenntnis der Bedingungen am Rande und der Gesetzmäßigkeiten im Inneren, gesuchte Größen wie z.B. den *Wassergehalt* Θ oder den *volumetrischen Fluß* u zu prognostizieren, und zwar zu jedem Zeitpunkt t und an jedem Ortspunkt $s = [x, y, z]^T$. Beispielsweise ist das Einsickern von Niederschlagswasser in den ungesättigten Boden und der resultierende Anstieg des Grundwasserspiegels zu beschreiben. Der Boden stellt ein poröses Medium dar, dessen Porenraum je nach Sättigungsgrad mehr oder weniger mit Wasser gefüllt ist.
Es sind also die Funktionen $\Theta(t, s)$ und $u(t, s)$ zu bestimmen. Zu diesem Zweck hat man zunächst festzulegen, wie man den Zustand des Systems beschreiben will und wie sich die gesuchten Größen daraus ableiten. Der Wassergehalt Θ eignet sich im allgemeinen nicht dazu, da sich daraus im gesättigten Bereich der Fluß u nicht berechnen läßt. Als geeignete Größe erweist sich das *hydraulische Potential* p, eine skalare Funktion, die sich zu jedem Zeitpunkt aus dem Gravitations- und dem Druckpotential zusammensetzt:

$$p(t, s) = z - \psi(t, s)$$

(hier bezeichnet z die vertikale Koordinatenrichtung (Gravitationspotential)). In ungesättigten Bereichen des Bodens ($\psi > 0$) wird ψ auch als *Saugspannung* bezeichnet, die durch Kapillarkräfte verursacht wird. Der volumetrische Fluß ist eine vektorwertige Funktion und bezeichnet die Volumenmenge Wasser, welche pro Zeit und Querschnitt in die entsprechende Richtung fließt. In der gesättigten Zone ($\psi \leq 0$) spiegelt ψ den hydrostatischen Druck wider. Ist diese Größe p bekannt, so lassen sich daraus, wie wir sehen werden, alle übrigen relevanten Größen herleiten. Das Aufstellen des angestrebten mathematischen Modells besteht nun darin, anzugeben, nach welchem mathematischen

Kalkül man die Zustandsfunktion $p(t,s)$ aus gegebenen Randbedingungen zu berechnen hat, und wie man daraus die übrigen Größen herleitet. Die Grundlagen hierfür sollen im Folgenden entwickelt werden. Eines der grundlegenden Prinzipien ist das Massenerhaltungsgesetz:

$$\partial_t \Theta(t,s) + \text{ div } u(t,s) = 0 \,, \tag{4.1}$$

wobei der div -Operator bzgl. der Ortsvariable gebildet werden soll. Diese Gleichung besagt, daß die zeitliche Änderung des Wassergehalts durch den Verlust an die lokale Umgebung ausgeglichen wird.

Die für die Bewegung des Wassers im Boden verantwortlichen Kräfte, die hier berücksichtigt werden sollen, werden durch das Spannungspotential p repräsentiert. Ist z die vertikale Orts-Komponente - positiv nach oben gerichtet - so wirkt die Gravitation in der negativen z-Richtung, und das Darcy-Buckingham-Gesetz schreibt sich in der Form

$$u = -K \, \nabla p \,, \tag{4.2}$$

wobei der Operator ∇ bzgl. der Ortsvariable gebildet werden soll. Im Darcy-Buckingham-Gesetz bezeichnet K die hydraulische Leitfähigkeit (Permeabilität) und hängt bei inhomogenen Medien vom Ort und in einer für das Medium typischen Weise vom Wassergehalt ab. Eine gängige Parametrisierung geht auf Mualem und van Genuchten [35, 47] zurück, die für den Wassergehalt besagt: $\Theta(t,s) = \tilde{\theta}(p(t,s),s)$, wobei

$$\tilde{\theta}(p(t,s),s) = \begin{cases} \theta_s & \text{für } p(t,s) \geq z \\ \theta_r + \frac{\theta_s - \theta_r}{(1+(\alpha(z-p(t,s))^\beta)^{1-1/\beta}} & \text{für } p(t,s) < z \end{cases} \,. \tag{4.3}$$

Hierbei bezeichnet θ_s bzw. θ_r den gesättigten bzw. residualen Wassergehalt, $\alpha > 0$ und $\beta > 1$ sind weitere, von der Bodenart abhängige Parameter. Man beachte, daß θ stetig und stückweise differenzierbar ist. Analog zu Kapitel 3 definieren wir für $q \in H^1((0,T); H^1_{\Gamma_D}(\Omega))$

$$f_q(t,s) := \tilde{\theta}(q(t,s),s)$$

und damit $\theta(q) := f_q$.
Für das Modell von Mualem und van Genuchten ist die Permeabilität K wie folgt parametrisiert:

$$K(p) = K_s \left(\frac{\theta(p) - \theta_r}{\theta_s - \theta_r} \right)^{1/2} \left(1 - \left(1 - \left(\frac{\theta(p) - \theta_r}{\theta_s - \theta_r} \right)^{\beta/(\beta-1)} \right)^{1-1/\beta} \right)^2 \,. \tag{4.4}$$

In Tabelle 4.1 sind die Parameter aus [15] aufgelistet, welche die Mittelwerte von Meßwerten aus einer Reihe von Experimenten sind.

Boden	θ_r [-]	θ_s [-]	α [1/m]	β [-]	K_s [m/s]
Sand	0.045	0.43	14.5	2.68	$8.25 \cdot 10^{-5}$
sandiger Lehm	0.065	0.41	7.5	1.89	$1.23 \cdot 10^{-5}$
Lehm	0.078	0.43	3.6	1.56	$2.89 \cdot 10^{-6}$
Ton	0.068	0.38	0.8	1.09	$5.56 \cdot 10^{-7}$

Tabelle 4.1: Parameter des Modells von Mualem und van Genuchten

Die resultierenden Funktionen $\theta(p)$ und $K(p)$ sind für Sand (durchgezogene Linie) und sandigen Lehm (gestrichelte Linie) in Abbildung 4.1 dargestellt.

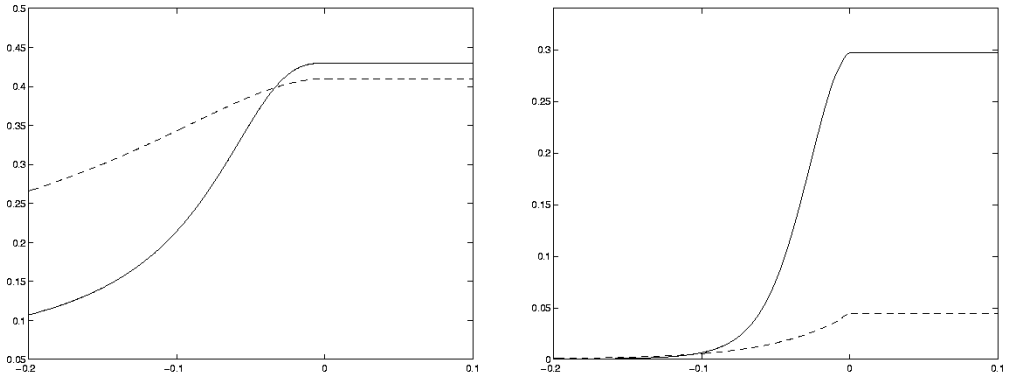

Abbildung 4.1: Wassergehalt (links) und Permeabilität (rechts) in Abhängigkeit von $-\psi$

4.2 Die Formulierung des LSF

Die Gleichungen (4.1) und (4.2) ergeben das folgende System

$$\begin{aligned} \partial_t \theta(p) + \ \text{div}\ u &= 0, \\ u + K(p)\nabla p &= 0, \end{aligned} \tag{4.5}$$

das zu lösen ist.
Wir setzen Anfangs- und Randbedingungen wie im letzten Kapitel voraus. Nach der Theorie von Alt und Luckhaus (vgl. [3]), existiert die Lösung $(p(t), u(t))$ von (4.5), ist eindeutig und hängt stetig von den Anfangs- und Randwerten ab.
Nun fahren wir fort, indem wir eine äquivalente Zeit-Ausgleichsformulierung definieren (vgl. auch Kapitel 2), die in $V_\tau \times Q_\tau$ minimiert werden soll. Analog zu den letzten Kapiteln erhalten wir in jedem Zeitschritt das LSF

$$\begin{aligned} &\tilde{\mathcal{F}}(u_\tau, p_\tau) \\ &\quad = \int_0^\tau \ \tau \| \partial_t \theta(p_\tau(t+\sigma)) + \left(\frac{\tau - \sigma}{\tau} \ \text{div}\ \ u_\tau^- + \frac{\sigma}{\tau} \ \text{div}\ \ u_\tau^+ \right) \|_{0,\Omega}^2 \, d\sigma \\ &\quad + \int_0^\tau \ \| \frac{\tau - \sigma}{\tau} (u_\tau^- + K(p_\tau(t))\nabla p_\tau(t)) + \frac{\sigma}{\tau} (u_\tau^+ + K(p_\tau^+)\nabla p_\tau^+) \|_{0,\Omega}^2 \, d\sigma \, , \end{aligned} \tag{4.6}$$

das in $\tilde{V}_\tau \times \tilde{Q}_\tau$ zu minimieren ist.
Wenn wir

$$\partial_t \theta(p_\tau(t+\sigma)) \approx \frac{\theta(p_\tau^+) - \theta(p_\tau(t))}{\tau}$$

annehmen, was für genügend kleine τ erfüllt ist, können wir $\tilde{\mathcal{F}}(u_\tau, p_\tau)$ mit Hilfe der Simpson-Regel durch

$$\begin{aligned}\mathcal{F}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) = &\frac{1}{6}\|\theta(p_\tau^+) - \theta(p_\tau(t)) + \tau \ \text{ div } \ u_\tau^-\|_{0,\Omega}^2 \\ &+ \frac{2}{3}\|\theta(p_\tau^+) - \theta(p_\tau(t)) + \tau\,\frac{1}{2}(\ \text{div } \ u_\tau^- + \ \text{div } \ u_\tau^+)\|_{0,\Omega}^2 \\ &+ \frac{1}{6}\|\theta(p_\tau^+) - \theta(p_\tau(t)) + \tau \ \text{ div } \ u_\tau^+\|_{0,\Omega}^2 \\ &+ \frac{\tau}{6}\|u_\tau^- + K(p_\tau(t))\nabla p_\tau(t)\|_{0,\Omega}^2 \\ &+ \frac{\tau}{6}\|u_\tau^- + u_\tau^+ + K(\frac{p_\tau(t)+p_\tau^+}{2})\left(\nabla p_\tau(t) + \nabla p_\tau^+\right)\|_{0,\Omega}^2 \\ &+ \frac{\tau}{6}\|u_\tau^+ + K(p_\tau^+)\nabla p_\tau^+\|_{0,\Omega}^2\end{aligned} \tag{4.7}$$

approximieren. Das System (4.5) besitzt eine eindeutige Lösung $(p(t), u(t))$, damit existiert das Minimum von $\tilde{\mathcal{F}}$. Dies führt dazu, daß $\mathcal{F}$ für genügend kleine Zeitschrittweite τ ein eindeutiges Minimum hat, das wir mit $(u_\tau^-, u_\tau^+, p_\tau^+)$ bezeichnen. Um die Lösung von (4.5) approximieren zu können, müssen wir nun auch bzgl. des Ortes diskretisieren. Wie im Abschnitt 2.2, wollen wir hier für V_h Raviart-Thomas Elemente niedrigster Ordnung und für Q_h stückweise lineare stetige finite Elemente wählen.
Unsere volldiskrete Minimierungsaufgabe lautet:
Man finde $(u_{\tau,h}^-, u_{\tau,h}^+, p_{\tau,h}^+) \in V_h^2 \times Q_h$, wobei $(u_{\tau,h}^-, u_{\tau,h}^+, p_{\tau,h}^+)$ die Gleichung

$$\mathcal{F}(u_{\tau,h}^-, u_{\tau,h}^+, p_{\tau,h}^+; p_\tau(t)) = \min_{(v_h^-, v_h^+) \in V_h^2, q_h \in Q_h} \mathcal{F}(v_h^-, v_h^+, q_h^+; p_\tau(t)) \tag{4.8}$$

erfüllen muß.

4.3 Zeit- und ortsadaptiver Algorithmus

Wir werden an dieser Stelle der Vollständigkeit halber einen volldiskreten Algorithmus vorstellen. In jedem Zeitschritt starten wir von einer vorgegebenen Triangulierung $\mathcal{T}_0$ und konstruieren eine Folge von adaptiv verfeinerten Triangulierungen $\mathcal{T}_l$ für $l = 1, 2, \ldots, l_{\max}$, die mit Hilfe eines Fehlerschätzers (im Ort) gebildet werden.
Die zugehörigen FE-Räume bezeichnen wir mit V_l und Q_l, für $l = 0, 1, 2 \ldots, l_{\max}$. Mit dem vom Benutzer vorgegebenen Parametern `tol` (Toleranzgrenze), $\delta \in (0, 1)$ (Sicherheitsfaktor), $\tau_{\max}$ (maximale Zeitschrittweite) und $\tau_{\min}$ (minimale Zeitschrittweite) und τ_0 (Anfangszeitschrittweite) lautet der adaptive Algorithmus

Algorithmus 4.1

$\mathrm{t} = 0\,; \tau = \tau_0\,; p^{\mathrm{alt}} = P_0\,;$
`while` $t < T$,
$l = 0\,;$
Berechne $(u^-_{\tau,0}, u^+_{\tau,0}, p^+_{\tau,0}) \in V_0^2 \times Q_0$ mit

$$\mathcal{F}(u^-_{\tau,0}, u^+_{\tau,0}, p^+_{\tau,0}; p^{\mathrm{alt}}) = \min_{(v^-_{\tau,0}, v^+_{\tau,0}, q^+_{\tau,0}) \in V_0^2 \times Q_0} \mathcal{F}(v^-_{\tau,0}, v^+_{\tau,0}, q^+_{\tau,0}; p^{\mathrm{alt}})\,;$$

Berechne den Fehlerschätzer η_0 für den Fehler $\|(u^-_{\tau,0}, u^+_{\tau,0}, p^+_{\tau,0}) - (u^-_\tau, u^+_\tau, p^+_\tau)\|$;
`while` $l \le l_{\max} \ \wedge \neg\ [\eta_l \ll \mathcal{F}(u^-_{\tau,l}, u^+_{\tau,l}, p^+_{\tau,l}; p^{\mathrm{alt}})]$,
$l = l + 1;$
$\mathcal{T}_l =$ verfeinere $\mathcal{T}_{l-1}$ mit Hilfe von η_{l-1} ;
Berechne $(u^-_{\tau,l}, u^+_{\tau,l}, p^+_{\tau,l}) \in V_l^2 \times Q_l$ mit

$$\mathcal{F}(u^-_{\tau,l}, u^+_{\tau,l}, p^+_{\tau,l}; p^{\mathrm{alt}}) = \min_{(v^-_{\tau,l}, v^+_{\tau,l}, q^+_{\tau,l}) \in V_0^2 \times Q_l} \mathcal{F}(v^-_{\tau,l}, v^+_{\tau,l}, q^+_{\tau,l}; p^{\mathrm{alt}})\,;$$

Berechne den Fehlerschätzer η_l für den Fehler $\|(u^-_{\tau,l}, u^+_{\tau,l}, p^+_{\tau,l}) - (u^-_\tau, u^+_\tau, p^+_\tau)\|$;
`end`
$\mathrm{Error}_{\mathrm{est}} = \mathcal{F}(u^-_{\tau,l}, u^+_{\tau,l}, p^+_{\tau,l}; p^{\mathrm{alt}})^{1/2}\,;$
`if` $\mathrm{Error}_{\mathrm{est}} <$ `tol` %% akzeptiere diesen Schritt

$$\mathrm{t} = \mathrm{t} + \tau\,;\ \tau = \min\left\{\tau_{\max}, \left(\frac{\texttt{tol}}{\mathrm{Error}_{\mathrm{est}}}\right)^{1/2} \tau\delta\right\}\,;\ p^{\mathrm{alt}} = p^+_{\tau,l}\,;$$

`else` %% die Berechnung dieses Schritts ist nicht akzeptabel

$$\tau = \max\left(\tau_{\min}, \left(\frac{\texttt{tol}}{\mathrm{Error}_{\mathrm{est}}}\right)^{1/2} \tau\delta\right)\,;$$

`end`
`end`

4.4 Numerische Experimente

In diesem Abschnitt wollen wir einige numerischen Ergebnisse präsentieren. Wir betrachten einen Sandkasten mit einer Höhe von zwei Metern (200 cm) und mit einer Breite von drei Metern (300 cm) (vgl. Abbildung 4.2).
Am Anfang geben wir die Anfangsbedingung $P_0 = -1.7m$ vor. Der Kasten wird an dem linken oberen Rand leicht bewässert, d.h. wir geben auf Γ_1 einen Fluß von $< n, u > = 0.148\ m/h$ vor. Γ_2, Γ_3 und Γ_4 sind so beschaffen, daß kein Wasser durchfließt, d.h. wir schreiben auf den genannten Randstücken $< n, u > = 0\ m/h$ vor. Schließlich lassen wir auf dem letzten Randstück Wasser herausfließen, d.h. wir geben auf Γ_5 $p = -1.7m$ vor (siehe Abbildung 4.2).

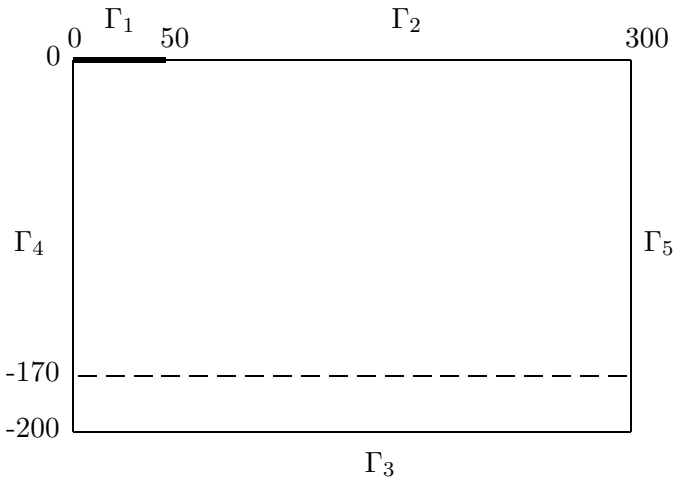

Abbildung 4.2: Testbeispiel eines zweidimensionalen Versickerungsproblems

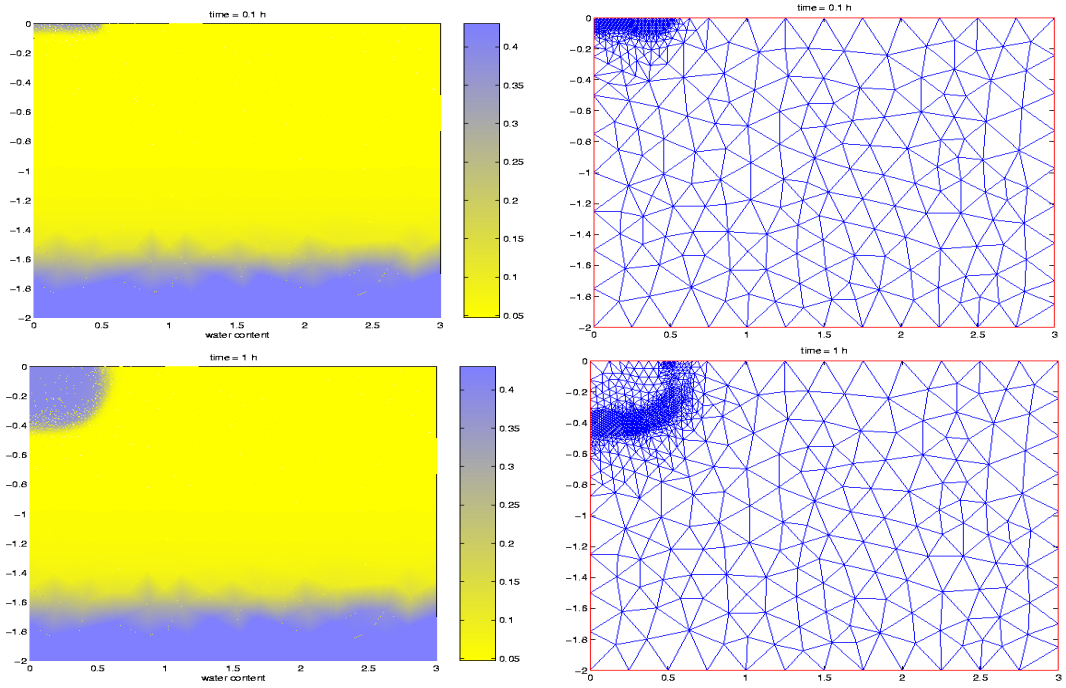

Abbildung 4.3: Wassergehalt und dazugehörige Gitter (auf Level 4) zu verschiedenen Zeitpunkten (Teil I)

Da wir die exakte Lösung nicht kennen, können wir Fehlerschätzer und Fehler (in der Zeit) nicht vergleichen, d.h. Ergebnisse wie in Kapitel 2 können wir an dieser Stelle nicht präsentieren.
Alternative: Man speichert den Fehlerschätzer in jedem Zeitschritt und auf jedem Level ab. Von $\mathcal{F}$ (dem Fehlerschätzer in der Zeit) erwartet man, daß er bei abnehmenden Zeitschrittweiten kleiner wird, d.h., falls das Funktional $\mathcal{F}$ auf feineren Levels kleiner wird, wird auch der Gesamtfehler kleiner (vorausgesetzt die Äquivalenz - analog zum linearen Fall - ist richtig). Wir wollen also wieder, wie im Kapitel 3, die Äquivalenz des LSF mit dem Fehler annehmen. Zur Approximation der Lösung wird Algorithmus

4.1 eingesetzt. Als Löser der Minimierungsaufgabe in Algorithmus 4.1 wird eine inexakte Gauß-Newton-Iteration mit einer Dämpfungsstrategie gewählt (vgl. z.B. [43]). Abbildung 4.5 zeigt Zeitschrittweiten in Abhängigkeit der Zeit.

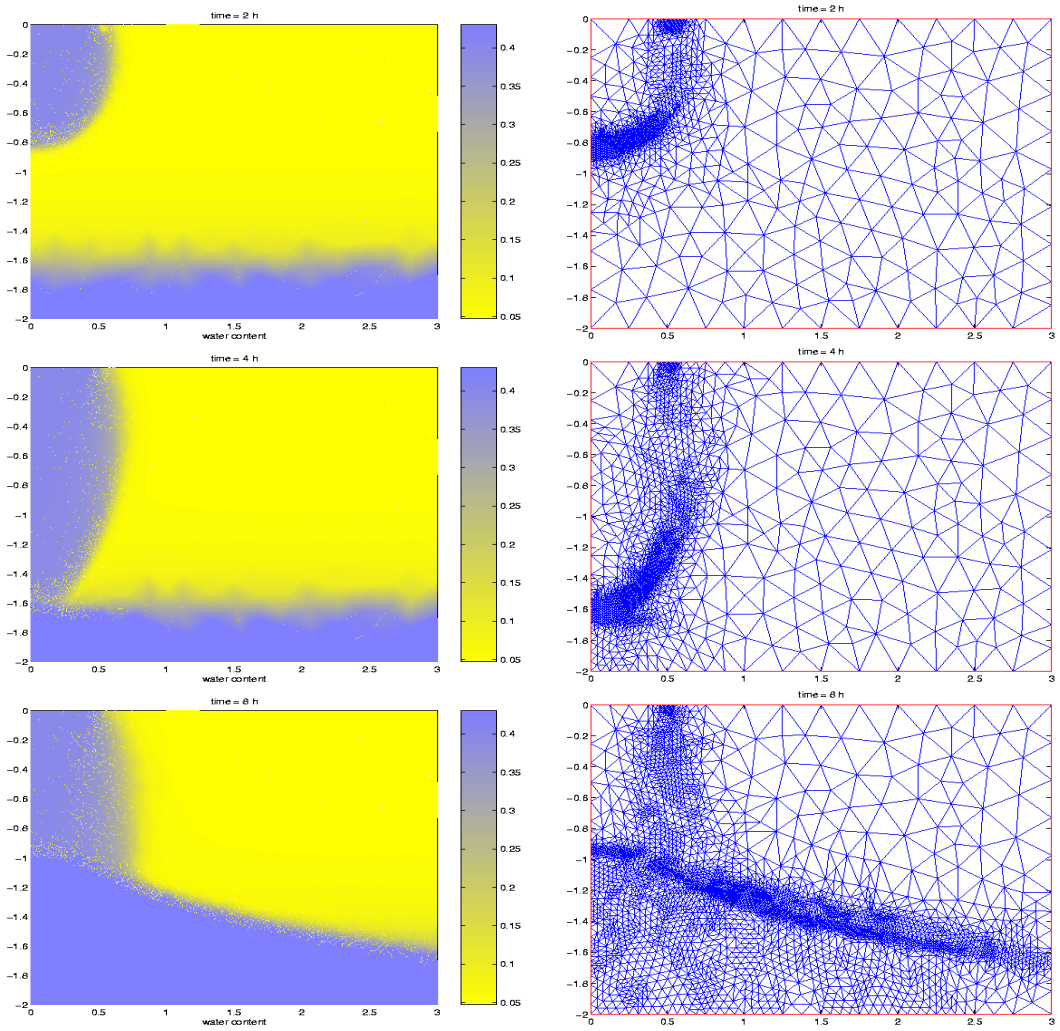

Abbildung 4.4: Wassergehalt und dazugehörige Gitter (auf Level 4) zu verschiedenen Zeitpunkten (Teil II)

Die Anfangszeitschrittweite beträgt $\tau_0 = 0.05$, wobei `tol`$= 0.02$. Die Zeitschrittweite wurde am Anfang größer. Nach einigen Stunden (vgl. Abbildung 4.4, mittlere Reihe), nachdem die obere Wasser-Front nach unten durchgesickert war, nahm die Zeitschrittweite langsam ab.
In Tabelle 4.2 ist der Wert des Funktionals $\mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))$ aufgelistet. Durch die Wahl der Zeitschrittweiten wird $\mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))^{1/2} <$ `tol` erfüllt (siehe die letzte Spalte in Tabelle 4.2). Man beachte, daß auch $\mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))^{1/2} \approx$ `tol` erfüllt ist.
Nun wollen wir experimentell die Ordnung des Verfahrens berechnen. In Tabelle 4.3 ist $\mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))$ zum Zeitpunkt $t = 0.8$ aufgelistet (gerechnet mit verschiedenen Zeitschrittweiten $\tau^{(j)} = 2^j \cdot 0.025$, $j = 0, \ldots, 5$). Damit der Fehler im Ort nicht so eine große Störung darstellt, ist immer der Wert auf dem letzten Level (Level 4) zu betrach-

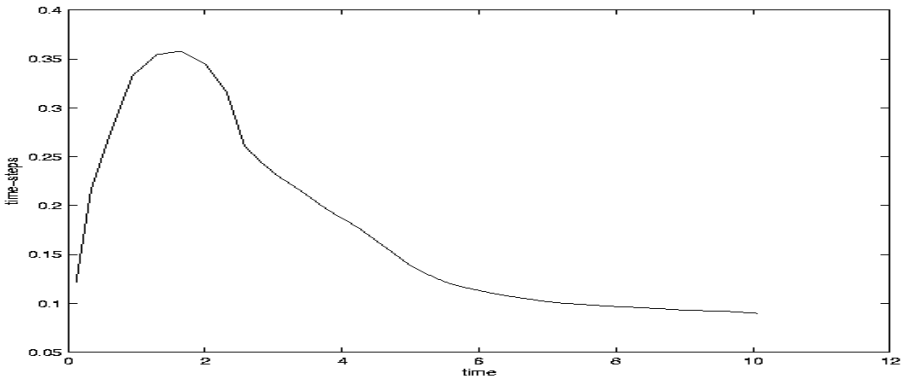

Abbildung 4.5: Zeitschrittweiten in Abhängigkeit der Zeit

Level / Zeitschritt	0	1	2	3	4
1	0.0053	$0.3517\ 10^{-4}$	$0.2707\ 10^{-4}$	$0.2167\ 10^{-4}$	$0.1929\ 10^{-4}$
10	$0.3909\ 10^{-3}$	$0.3891\ 10^{-3}$	$0.3819\ 10^{-3}$	$0.3811\ 10^{-3}$	$0.3681\ 10^{-3}$
20	$0.3696\ 10^{-3}$	$0.3596\ 10^{-3}$	$0.3483\ 10^{-3}$	$0.3411\ 10^{-3}$	$0.3314\ 10^{-3}$
40	$0.3195\ 10^{-3}$	$0.3187\ 10^{-3}$	$0.3099\ 10^{-3}$	$0.3021\ 10^{-3}$	$0.2967\ 10^{-3}$
60	$0.3385\ 10^{-3}$	$0.3133\ 10^{-3}$	$0.3060\ 10^{-3}$	$0.3034\ 10^{-3}$	$0.2957\ 10^{-3}$
71	$0.3218\ 10^{-3}$	$0.3028\ 10^{-3}$	$0.2988\ 10^{-3}$	$0.2960\ 10^{-3}$	$0.2924\ 10^{-3}$

Tabelle 4.2: $\mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))$

ten. Sei nun `ord` die Ordnung des Verfahrens. Dann gilt jeweils für zwei Zeitschrittweiten $\tau^{(j)}$ und $\tau^{(j-1)}$ $j = 1, \ldots, 5$ (man beachte, daß $\tau^{(j)} = 2\,\tau^{(j-1)}$) bzw. für die zugehörigen Funktionale:

$$\mathcal{F}_{\tau^{(j)}}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t)) \eqsim (\tau^{(j)})^{2\,\mathtt{ord}} = (2\,\tau^{(j-1)})^{2\,\mathtt{ord}}, \tag{4.9}$$

$$\mathcal{F}_{\tau^{(j-1)}}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t)) \eqsim (\tau^{(j-1)})^{2\,\mathtt{ord}}. \tag{4.10}$$

Dividieren von (4.9) durch (4.10) liefert

$$\frac{\mathcal{F}_{\tau^{(j)}}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))}{\mathcal{F}_{\tau^{(j-1)}}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))} \approx 2^{2\,\mathtt{ord}}, \tag{4.11}$$

wobei angenommen wurde, daß sich die in (4.9) und (4.10) vorkommende Konstanten gegenseitig kürzen. Damit ergibt sich

$$2\,\mathtt{ord} \approx \frac{\log\left(\frac{\mathcal{F}_{\tau^{(j)}}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))}{\mathcal{F}_{\tau^{(j-1)}}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))}\right)}{\log(2)}. \tag{4.12}$$

Bei den in Tabelle 4.3 aufgelisteten Werten ergibt sich im Durchschnitt $\mathtt{ord} = 0.51$. Diese Ordnung scheint zunächst sehr gering im Vergleich zu dem Aufwand, den man investiert. Man muß aber beachten, daß die Ordnung von $\frac{3}{2}$ (wie im linearen Fall) nicht zu erreichen ist.
Wenn man beachtet, daß wenige numerische Verfahren existieren, die überhaupt eine Ordnung bzgl. der Zeit für diese Art von nichtlinearen Probleme bieten, dann ist die

Level τ	2	3	4
0.025	$0.0145\ 10^{-3}$	$0.0118\ 10^{-3}$	$0.0103\ 10^{-3}$
0.05	$0.0279\ 10^{-3}$	$0.0213\ 10^{-3}$	$0.0202\ 10^{-3}$
0.1	$0.0619\ 10^{-3}$	$0.0538\ 10^{-3}$	$0.0503\ 10^{-3}$
0.2	$0.1578\ 10^{-3}$	$0.1432\ 10^{-3}$	$0.1234\ 10^{-3}$
0.4	$0.1958\ 10^{-3}$	$0.1705\ 10^{-3}$	$0.1734\ 10^{-3}$
0.8	$0.4059\ 10^{-3}$	$0.3282\ 10^{-3}$	$0.2781\ 10^{-3}$

Tabelle 4.3: $\mathcal{F}(u^-_{\tau,h}, u^+_{\tau,h}, p^+_{\tau,h}; p_{\tau,h}(t))$, berechnet jeweils mit verschiedenen Zeitschrittweiten τ

obige Ordnung interessant. Durch die Ordnung in der Zeit ist es möglich, den Fehler über die Zeit zu kontrollieren, was bei derartigen parabolischen Problemen oft die Schwierigkeit darstellt.

Anhang A

Bezeichnungen, Räume

In diesem Anhang wollen wir der Vollständigkeit halber die wichtigsten Räume und Bezeichnungen definieren, die in dieser Arbeit benutzt wurden.
Es sei $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$ ein Gebiet mit stückweise glattem Rand.

A.1 Distributionen

Wir definieren zuerst

$$C^\infty(\Omega) = \bigcap_{m \in \mathbb{N}_0} C^m(\Omega).$$

Für ein $\varphi \in C^\infty(\Omega)$ sei

$$\mathrm{supp}\varphi = \overline{\{x \in \Omega \mid f(x) \neq 0\}}.$$

Somit ergibt sich der Raum der *Testfunktionen*:

$$\mathcal{D}(\Omega) = C_0^\infty(\Omega) = \{f \in C^\infty(\Omega) \mid \mathrm{supp}\varphi \subset \Omega \text{ kompakt}\}.$$

Wir definieren, was man unter *Stetigkeit und Differenzierbarkeit bis zum Rand* versteht: Sei $\Gamma \subset \partial\Omega$ eine relativ offene Teilmenge. Für $m \in \mathbb{N}_0$ und $\tilde{\Omega} = \Omega \cup \Gamma$ sei

$$C^m(\tilde{\Omega}) = \left\{ f \in C^m(\Omega) \,\middle|\, \exists f_1 \in C^m(\Omega_1) \subset \mathbb{R}^d, \text{ mit } \tilde{\Omega} \subset \Omega_1 \text{ Gebiet in } \mathbb{R}^d \text{ und } f_1\big|_{\tilde{\Omega}} = f \right\}$$

und

$$\mathcal{D}(\tilde{\Omega}) = C_0^\infty(\tilde{\Omega}) = C_\Gamma^\infty(\Omega) = \left\{ f \in C^\infty(\tilde{\Omega}) \,\middle|\, \mathrm{supp}\varphi \subset \tilde{\Omega} \text{ kompakt} \right\}.$$

Auf $\mathcal{D}(\tilde{\Omega})$ führen wir folgenden Konvergenzbegriff ein: Für eine Folge $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{D}(\tilde{\Omega})$ und $\varphi \in \mathcal{D}(\tilde{\Omega})$ sagen wir, φ_n konvergiert in $\mathcal{D}(\tilde{\Omega})$ gegen φ oder $(\varphi_n \xrightarrow{\mathcal{D}} \varphi)$, falls ein kompaktes $K \subset \tilde{\Omega}$ existiert mit $\mathrm{supp}\varphi_n \subset K$ für alle $n \in \mathbb{N}$, und für alle $\alpha \in \mathbb{N}_0^d$ (Multiindex-Schreibweise!) gilt:

$$\partial^\alpha \varphi_n \longrightarrow \partial^\alpha \varphi \text{ gleichmäßig in } \tilde{\Omega}.$$

Mit diesem Konvergenzbegriff definieren wir den Raum $\mathcal{D}'(\tilde{\Omega})$ der stetigen linearen Funktionale auf $\mathcal{D}(\tilde{\Omega})$. Die Elemente aus diesem Raum werden *Distributionen* oder *verallgemeinerte Funktionen* genannt. Dabei gilt: Eine lineare Abbildung $F : \mathcal{D}(\tilde{\Omega}) \to \mathbb{R}$ nennen wir *stetig*, falls für $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{D}(\tilde{\Omega})$ und $\varphi \in \mathcal{D}(\Omega)$ aus $\varphi_n \xrightarrow{\mathcal{D}} \varphi$ stets

$F(\varphi_n) \xrightarrow{\mathrm{R}} F(\varphi)$ folgt.
Auf $\mathcal{D}'(\tilde{\Omega})$ definiert man nun Konvergenz: Seien $(F_n)_{n\in\mathbb{N}} \subset \mathcal{D}'(\tilde{\Omega})$ und $F \in \mathcal{D}'(\tilde{\Omega})$, dann sagen wir $F_n \xrightarrow{\mathcal{D}'} F$, falls $F_n(\varphi) \xrightarrow{\mathrm{R}} F(\varphi)$ für alle $\varphi \in \mathcal{D}(\tilde{\Omega})$. Damit wird $\mathcal{D}'(\tilde{\Omega})$ zu einem *topologischen* Vektorraum.
Man kann leicht sehen, daß jedes $f \in L^p(\Omega) \cup C^m(\Omega), 1 \le p \le \infty$, $m \in \mathbb{N}_0 \cup \{\infty\}$, die folgende Distribution erzeugt:

$$f : \mathcal{D}(\Omega) \to \mathbb{R},$$

wobei

$$f(\varphi) =< f, \varphi >= \int_\Omega f(x)\varphi(x)\,dx.$$

Es gibt aber auch Distributionen, die nicht von L^p-Funktionen erzeugt werden. Ein Beispiel dafür ist die Dirac-Distribution

$$\delta : \mathcal{D}(\mathbb{R}) \to \mathbb{R},$$

wobei

$$\delta(\varphi) =< \delta, \varphi >= \varphi(0).$$

Nun sind wir in der Lage, für alle Distributionen F und Multiindizes $\alpha \in \mathbb{N}_0^d$ die *(schwache) distributive Ableitung* $\partial^\alpha F$ als die Distribution

$$\partial^\alpha F : \mathcal{D}(\Omega) \to \mathbb{R}$$

zu definieren, wobei

$$\partial^\alpha F(\varphi) = (-1)^{|\alpha|} F(\partial^\alpha \varphi).$$

Man beachte, daß für im klassischen Sinne differenzierbare Funktionen die Distribution, die von der Ableitung erzeugt wird, mit der distributiven Ableitung übereinstimmt. Für $F \in \mathcal{D}'(\Omega)$ sagen wir: $\partial^\alpha F \in L^p(\Omega)$, falls es ein $g_\alpha \in L^p(\Omega)$ existiert mit

$$\partial^\alpha F(\varphi) = \int_\Omega g_\alpha(x)\varphi(x)\,dx \qquad \forall\ \varphi \in \mathcal{D}(\Omega).$$

A.2 Sobolevräume

Mit Hilfe der Distributionen sind wir in der Lage, die für uns wichtigen Sobolevräume ohne großen Aufwand zu definieren:
Für $k \in \mathbb{N}_0$ sei

$$H^k(\Omega) = \left\{ v \in L^2(\Omega) \;\middle|\; \partial^\alpha v \in L^2(\Omega)\ \forall\ \alpha \in \mathbb{N}_0^d,\ |\alpha| \le k \right\}.$$

Offensichtlich ist $H^0(\Omega) = L^2(\Omega)$. $H^k(\Omega)$ ist bzgl. des Skalarprodukts

$$(v, w)_{k,\Omega} = \sum_{|\alpha|\le k} (\partial^\alpha v, \partial^\alpha w)_{L^2(\Omega)}$$

ein Hilbertraum.

Für $v \in H^k(\Omega)$ ist

$$|v|_{k,\Omega} = \left(\sum_{|\alpha|=k} \|\partial^\alpha v\|_{L^2}^2 \right)^{1/2}$$

eine Halbnorm auf $H^k(\Omega)$.
Es sei für $k \in \mathbb{N}_0, 1 \leq p \leq \infty$, der Raum $H_0^k(\Omega)$ als Abschluß von $C_0^\infty(\Omega)$ in $H^k(\Omega)$ definiert, d.h.

$$H_0^k(\Omega) = \overline{(C_0^\infty(\Omega))}^{H^k}.$$

Für $k = 0$ kann man zeigen, daß

$$H_0^0(\Omega) = H^0(\Omega) = L^2(\Omega).$$

Mit anderen Worten

$$C_0^\infty(\Omega) \underset{\text{dicht}}{\subset} L^2(\Omega).$$

Sei

$$H^{-k}(\Omega) = \left(H_0^k(\Omega)\right)'$$

der Dualraum zu $H_0^k(\Omega)$.
Als letztes führen wir für eine Distribution $F \in \left(\mathcal{D}'(\tilde{\Omega})\right)^d$ die *Distributionen-Divergenz* $\text{div } F \in \mathcal{D}'(\tilde{\Omega})$ ein als

$$\text{div } F(\varphi) = -F(\nabla\varphi) \qquad \forall\ \varphi \in \mathcal{D}(\tilde{\Omega}).$$

Dann sei

$$H(\text{div}, \Omega) = \left\{ v \in (L^2(\Omega))^d \;\middle|\; \text{div } v \in L^2(\Omega) \right\}.$$

Mit dem Skalarprodukt

$$(v, w)_{H(\text{div},\Omega)} = (v, w)_{0,\Omega} + (\text{ div } v,\ \text{div } w)_{0,\Omega}$$

ist $H(\text{div}, \Omega)$ ein Hilbertraum. Nützliche Details über Sobolevräume und die zugehörigen Vollständigkeitsbeweise kann man in [1] nachlesen.

A.3 Das Bochner-Integral und Orts-Zeit-Funktionenräume

Für ein beliebiges Intervall $I \subset [0, T]$ sei $\mathcal{D}(I)$ der Raum der Testfunktionen.
Unter einer *Distribution T mit Werten in einem Hilbertraum H* verstehen wir eine stetige lineare Abbildung

$$T : \mathcal{D}(I) \longrightarrow H.$$

Dafür schreiben wir kurz $T \in \mathcal{D}'(I; H)$.
Dabei gilt: Eine lineare Abbildung $T : \mathcal{D}(I) \to H$ nennen wir *stetig*, falls für $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{D}(I)$ und $\varphi \in \mathcal{D}(I)$ aus $\varphi_n \xrightarrow{\mathcal{D}} \varphi$ stets $T(\varphi_n) \xrightarrow{H} T(\varphi)$ folgt.
Auf $\mathcal{D}'(I; H)$ definiert man nun Konvergenz: Seien $(T_n)_{n\in\mathbb{N}} \subset \mathcal{D}'(I; H)$ und $T \in \mathcal{D}'(I; H)$; dann sagen wir $T_n \xrightarrow{\mathcal{D}'} T$, falls $T_n(\varphi) \xrightarrow{H} T(\varphi)$ für alle $\varphi \in \mathcal{D}(I)$. Damit

wird $\mathcal{D}'(I;H)$ zu einem topologischen Vektorraum. Für $T \in \mathcal{D}'(I;H)$ und beliebigen $n \in \mathbb{N}_0$ sei die *(schwache) distributive Ableitung* $\partial_t^n T$ als die Distribution

$$\partial_t^n T : \mathcal{D}(I) \to H$$

definiert, wobei

$$\partial_t^n T(\varphi) = (-1)^n T(\partial_t^n \varphi).$$

Von nun an wollen wir unter $\partial_t f$ oder f' bzw. $\partial^\alpha f$ einer Funktion f immer die Ableitung im Distributionen-Sinne verstehen. Somit brauchen wir uns über deren Existenz keine Gedanken mehr zu machen. Sie existiert dann nämlich immer.
Sei H ein separabler Hilbertraum mit dem Skalarprodukt $\langle \cdot, \cdot \rangle_H$. Mit der Maßtheorie definiert man das Bochner-Integral, das jeder *Bochner-meßbaren* Funktion $f : \overset{\circ}{I} \to H$ ein Element $\int_{\overset{\circ}{I}} f(t)dt \in H$ zuordnet, das bestimmte Eigenschaften hat (vgl. hierzu [50, 24]). Für $1 \leq p < \infty$, $m \in \mathbb{N}_0$ und I (Intervall) $\subset [0,T]$ sei

$$L^p(\overset{\circ}{I}; H) = \left\{ f : \overset{\circ}{I} \to H \;\middle|\; f \text{ schwach meßbar und } \int_{\overset{\circ}{I}} \|f(t)\|_H^p \, dt < \infty \right\}.$$

Bzgl. der Norm

$$\|f\|_{L^p(\overset{\circ}{I};H)} = \left(\int_{\overset{\circ}{I}} \|f(t)\|_H^p \, dt \right)^{1/p}$$

ist $L^p(\overset{\circ}{I}; H)$ ein Banachraum. Ist $p = 2$, so ist $L^2(\overset{\circ}{I}; H)$ mit dem Skalarprodukt

$$(f,g)_{L^2(\overset{\circ}{I};H)} = \int_{\overset{\circ}{I}} \langle f(t), g(t) \rangle_H \, dt$$

ein Hilbertraum.
Der Raum der stetigen Funktionen von $\overset{\circ}{I}$ auf H sei definiert als

$$C(\overset{\circ}{I}; H) = \left\{ f : \overset{\circ}{I} \to H \;\middle|\; f \text{ ist stetig auf } \overset{\circ}{I} \right\}.$$

Analog sei der Raum der stetigen Funktionen von I auf H definiert als

$$C(I;H) = \{ f : I \to H \mid f \text{ ist stetig auf } I \}.$$

Die Elemente $f_1 \in L^p(\overset{\circ}{I}, H)$ und $f_2 \in C(\overset{\circ}{I}, H)$ bilden in natürlicher Weise eine Distribution:

$$f_i : \mathcal{D}(\overset{\circ}{I}) \longrightarrow H \,,$$

wobei

$$f_i(\varphi) = \int_{\overset{\circ}{I}} f_i(t)\varphi(t)dt, \quad i = 1, 2.$$

Ebenso bilden auch die Elemente $f \in C(I, H)$ in natürlicher Weise eine Distribution:

$$f : \mathcal{D}(I) \longrightarrow H \,,$$

wobei

$$f(\varphi) = \int_I f(t)\varphi(t)dt.$$

Nun wollen wir mit Hilfe dieser Räume weitere Funktionenräume definieren. Wegen $C(I;H) \subset \mathcal{D}'(I;H)$ kann man den Raum der differenzierbaren Funktionen von I auf H definieren als

$$C^1(I;H) = \{f \in C(I;H) \mid f' \in C(I;H)\}$$

und

$$C^m(I;H) = \left\{f \in C^{m-1}(I;H) \;\middle|\; f^{(m)} \in C(I;H)\right\}.$$

Schließlich sei

$$C^\infty(I;H) = \bigcap_{n\in\mathbf{N}} C^n(I;H).$$

Es ist auch möglich, Sobolevräume zu definieren: Es sei für $n \in \mathbb{N}$

$$H^m(\overset{\circ}{I};H) = \left\{f \in L^2(\overset{\circ}{I};H) \;\middle|\; \partial_t^n f \in L^2(\overset{\circ}{I};H)\;\forall\; n = 1\ldots m\right\}.$$

Als letztes sei angemerkt, daß die obigen Bemerkungen sämtlich in [50, IV] nachzulesen sind.

Anhang B

Funktionalanalytische Grundlagen

In diesem Anhang werden einige Begriffe definiert und wichtige Sätze zur Lösungstheorie der partiellen Differentialgleichungen zitiert, die in der Arbeit benutzt wurden.

B.1 Bilinearformen

Mit Hilfe des Satzes von Lax-Milgram und den Darstellungssätzen von Friedrichs (vgl. [28, Seite 322, Theorem 2.1, Seite 331, Theorem 2.23]) wird eine wichtige Aussage für Bilinearformen bewiesen.

Satz B.1 Sei $a : H^1_{\Gamma_D}(\Omega) \times H^1_{\Gamma_D}(\Omega) \to \mathbb{R}$ ein symmetrischer stetiger $H^1_{\Gamma_D}(\Omega)$-elliptischer Bilinearform. Dann gelten:

(a) Es existiert genau ein positiv definiter selbstadjungierter Operator $A : D_A \subset L^2(\Omega) \to L^2(\Omega)$ mit

(i) $D_A \subset H^1_{\Gamma_D}(\Omega)$,
(ii) $a(p,q) = (Ap,q)_{0,\Omega} \ \forall \ p \in D_A,\ q \in H^1_{\Gamma_D}(\Omega)$,
(iii) D_A liegt dicht in $H^1_{\Gamma_D}(\Omega)$ bzgl. $\|.\|_{1,\Omega}$.

(b) Für alle $f \in L^2(\Omega)$ existert genau eine Lösung p des Variationsproblems

$$a(p,q) = (p,q)_{0,\Omega} \qquad \text{für alle } v \in H^1_{\Gamma_D}(\Omega).$$

Es gilt $p \in D_A$ und $Ap = f$.

(c) $A^{1/2}$ ist wohldefiniert und es gilt $D_{A^{1/2}} = H^1_{\Gamma_D}(\Omega)$ mit $a(p,q) = (A^{1/2}p, A^{1/2}q)_{0,\Omega}$ für alle $p, q \in H^1_{\Gamma_D}(\Omega)$.

(d) $a(\cdot,\cdot)$ ist abgeschlossen.

Beweis:
Der Beweis läuft analog zu [32, Satz 1.1].

■

B.2 Der Lösungsoperator

In diesem Abschnitt wollen wir den Lösungsoperator für lineare parabolische Differentialgleichungen durch die sogenannte *Variation der Konstanten*-Formel definieren. Sei a eine Bilinearform mit den Voraussetzungen von Satz B.1 und A der zugehörige Operator aus Satz B.1.

Bemerkung B.2 (a) $T = A^{-1}$ ist wohldefiniert, da A positiv definit ist. T ist ein kompakter Operator (vgl. [41, Seite 165 ff.]), d.h. $\overline{T\left(\{f \in L^2(\Omega) : \|f\|_{0,\Omega} \leq 1\}\right)} \subset D_A$ ist kompakt in $H^1_{\Gamma_D}(\Omega)$.

Mit A ist auch T selbstadjungiert. Aus dem Spektralsatz für kompakte selbstadjungierte Operatoren (vgl. [6, Theorem 24.5, Seite 438]) folgt

$$Tp = \sum_{i \in \mathbb{N}} \lambda_i \left(p, \Phi_i\right)_{0,\Omega} \Phi_i,$$

wobei $\lambda_i > 0$, $i \in \mathbb{N}$, die abzählbar vielen Eigenwerte von T sind und $\{\Phi_i \mid i \in \mathbb{N}\}$, ein vollständiges Orthonormalsystem von $L^2(\Omega)$ bildet. Man definiert T^α für $\alpha > 0$ durch

$$T^\alpha p = \sum_{i \in \mathbb{N}} \lambda_i^\alpha \left(p, \Phi_i\right)_{0,\Omega} \Phi_i.$$

Dann ist $A^\alpha = (T^\alpha)^{-1}$ wohldefiniert.
Der Definitionsbereich von A^α wird mit D_{A^α} bezeichnet.
$\dot{H}^{2\alpha} = D_{A^\alpha}$ ist mit dem Skalarprodukt

$$(p, q)_{\dot{H}^{2\alpha}} = (A^\alpha p, A^\alpha q)_{0,\Omega}$$

ein Hilbertraum. Die Einbettung $\dot{H}^\alpha \hookrightarrow \dot{H}^\beta$ ist stetig für $\alpha > \beta$.

(b) Mit Hilfe des Operators A wird ein weiterer wichtiger Operator definiert. Für $t \in [0, T]$ sei der Operator $\mathcal{U}(t) : L^2(\Omega) \to L^2(\Omega)$ durch die folgende Vorschrift gegeben:

$$\mathcal{U}(t)(p) = \exp(-tA)(p) = \sum_{i \in \mathbb{N}} e^{(-t\lambda_i^{-1})} \left(p, \Phi_i\right)_{0,\Omega} \Phi_i \,.$$

Seine Stetigkeit folgt aus $\lambda_i^{-1} > 0$, denn es gilt:

$$
\begin{aligned}
&\left\| \sum_{i\in\mathbf{N}} e^{(-t\lambda_i^{-1})} (p,\Phi_i)_{0,\Omega}\Phi_i \right\|_{0,\Omega}^2 \\
&= \left(\sum_{i\in\mathbf{N}} e^{(-t\lambda_i^{-1})} \left(p,\Phi_i\right)_{0,\Omega} \Phi_i, \sum_{j\in\mathbf{N}} e^{(-t\lambda_j^{-1})} \left(p,\Phi_j\right)_{0,\Omega} \Phi_j \right)_{0,\Omega} \\
&= \sum_{i,j\in\mathbf{N}} e^{(-t\lambda_i^{-1})} e^{(-t\lambda_j^{-1})} \left(p,\Phi_i\right)_{0,\Omega} \left(p,\Phi_j\right)_{0,\Omega} \overbrace{\left(\Phi_i,\Phi_j\right)_{0,\Omega}}^{=\delta_{ij}\ \text{(Kronecker-Symbol)}} \\
&= \sum_{i\in\mathbf{N}} \left(e^{(-t\lambda_i^{-1})}\right)^2 \left((p,\Phi_i)_{0,\Omega} \right)^2 \\
&\leq \sum_{i\in\mathbf{N}} \left(e^{0}\right)^2 \left((p,\Phi_i)_{0,\Omega} \right)^2 \\
&= \sum_{i,j\in\mathbf{N}} \left(p,\Phi_i\right)_{0,\Omega} \left(p,\Phi_j\right)_{0,\Omega} \overbrace{\left(\Phi_i,\Phi_j\right)_{0,\Omega}}^{=\delta_{ij}} \\
&= \left(\sum_{i\in\mathbf{N}} \left(p,\Phi_i\right)_{0,\Omega} \Phi_i, \sum_{j\in\mathbf{N}} \left(p,\Phi_j\right)_{0,\Omega} \Phi_j \right)_{0,\Omega} \\
&= \left\| \sum_{i\in\mathbf{N}} \left(p,\Phi_i\right)_{0,\Omega} \Phi_i \right\|_{0,\Omega}^2 = \|p\|_{0,\Omega}^2 .
\end{aligned}
\tag{B.1}
$$

Es gilt weiterhin:

Da $\{\mathcal{U}(t) \mid t \geq 0\}$ eine *Halbgruppe* bzgl. der Verknüpfung $\cdot$ (wobei $\mathcal{U}(t)\cdot\mathcal{U}(s) = \mathcal{U}(t+s)$) bildet, nennt man die Theorie, die auf diesem Wege die Lösung konstruiert, die Halbgruppentheorie.
Analog zur obigen Überlegung, kann man für jede beschränkte Funktion
$\varphi : (0,\infty) \to \mathbb{R}$ den Operator $\varphi(tA)$ wie folgt definieren:

$$
\varphi(tA) = \sum_{i\in\mathbf{N}} \varphi(t\lambda_i^{-1}) \left(p,\Phi_i\right)_{0,\Omega} \Phi_i \,.
$$

Die Stetigkeit ergibt sich analog zu (B.1) aus der Beschränktheit von φ.

(c) Man kann also für $0 \leq s \leq 1$ (wegen $Ap = f$, $(p,p)_{\dot{H}^{2\alpha}} = (A^\alpha p, A^\alpha p)_{0,\Omega}$) $H^{s+1}(\Omega)$-Regularität durch die Existenz einer stetigen Einbettung
$\dot{H}^2 \hookrightarrow H^{1+s}(\Omega) \cap H_0^1(\Omega)$ charakterisieren.
Mit *stetiger Einbettung* ist dabei gemeint, daß $id : \dot{H}^2 \hookrightarrow H^{1+s}(\Omega) \cap H_0^1(\Omega)$ stetig ist. Da id eine lineare Abbildung ist, erhält man folgende äquivalente Aussage:
Es gibt ein C, so daß für alle $p \in \dot{H}^2$ gilt:

$$
\|p\|_{1+s,\Omega} \leq C \|p\|_{\dot{H}^2} \,.
$$

Beispiel B.3 Hat Ω einen Lipschitzrand und gilt außerdem $a^{\rho\sigma} \in C^{0,t}(\overline{\Omega})$ für $0 < t \leq 1$ (d.h. $a^{\rho\sigma} \in C(\overline{\Omega})$ und $\|a^{\rho\sigma}(x) - a^{\rho\sigma}(y)\| \leq L\|x-y\|^t, \ \forall \ x, y \in \Omega$), so erhält man: $\dot{H}^2 \hookrightarrow H^{1+s}(\Omega) \cap H_0^1(\Omega)$ ist stetig für alle $0 \leq s < \min(t, \frac{1}{2})$.
Gilt zusätzlich noch $t = 1$ und ist Ω konvex, so haben wir $H^2(\Omega)$-Regularität (vgl. z.B. [25]).
Weiteres über Sobolevräume kann der Leser in [12] erfahren.

Bemerkung B.4 (a) Man beachte, daß $p(t) \in H_0^1(\Omega)$ für alle $t > 0$, obwohl lediglich $p(0) = P_0 \in L^2(\Omega)$ vorausgesetzt wurde. Dies liegt an der *Glättungseigenschaft* der parabolischen Differentialgleichungen.

(b) Man beachte, daß für alle ρ, σ $a^{\rho\sigma}$ unabhängig von t sind. Eine allgemeinere Theorie findet man in [50, IV].

(c) Sei zunächst

$$E(t+\tau, t)p(t) = \mathcal{U}(t+\tau)p(t) + \int_t^{t+\tau} \mathcal{U}(t+\tau-s)f(s)ds. \tag{B.2}$$

Die exakte Lösung p von (1.5) ist für $t \in (0,T)$ gegeben durch (vgl. z.B. [39, p.370]):

$$p(t) = E(t,0)P_0 = \mathcal{U}(t)P_0 + \int_0^t \mathcal{U}(t-s)f(s)ds. \tag{B.3}$$

Diese Formel ist als Variation der Konstanten bekannt.

(d) Hängt f nicht von t ab, so kann man die Lösung p wie folgt konstruieren:
Es sei $w = A^{-1}f$. Sei

$$E(t; t+\tau)p(t) = [w - \mathcal{U}(t+\tau)w] + \mathcal{U}(t+\tau)p(t). \tag{B.4}$$

p läßt sich in jedem Zeitpunkt $t \in (0,T]$ durch

$$p(t) = E(t,0)P_0 = [w - \mathcal{U}(t)w] + \mathcal{U}(t)P_0 \tag{B.5}$$

angeben. Es gilt nämlich $p'(t) = A\mathcal{U}(t)w - A\mathcal{U}(t)P_0$. Insgesamt folgt also für $t \in (0,T]$:

$$p'(t) + Ap(t) = A\mathcal{U}(t)w + A[w - \mathcal{U}(t)w] = Aw = f.$$

Die Anfangsbedingung ist wegen $p(0) = [w - U(0)w] + U(0)P_0 = U(0)P_0 = P_0$ erfüllt.
Da $\mathcal{U}$ (bzgl. t) unendlich oft differenzierbar ist, ist somit auch $p \in C^\infty((0,T], \dot{H}^2) \cap C([0,T]; L^2(\Omega))$, d.h. p ist (bzgl. t) unendlich oft differenzierbar.

(e) Bei der Formulierung von (1.2) bzw. (1.5) haben wir nur *homogene Dirichlet*-Randwerte gefordert. Bei inhomogenen Dirichlet-Randwerten, also etwa durch die Forderung

$$p\big|_{\partial\Omega} = g\big|_{\partial\Omega},$$

wobei $g \in H^1(\Omega)$ vorgegeben ist, sucht man zunächst die schwache Lösung $\hat{p} \in H_0^1(\Omega)$ von (1.2) bzw. (1.5), wobei man jeweils f durch $\hat{f} = f - Ag$ ersetzt und A die schwache Repräsentation von $A(x, \partial)$ in (1.1) ist (vgl. Bemerkung B.2.(a)). Dann ist die Lösung der entsprechenden Randwertaufgabe bzw. der Anfangs-Randwertaufgabe mit inhomogenen Dirichlet-Randwerten gegeben durch $p = \hat{p} + g$.

Literaturverzeichnis

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] S. Adjerid and J. E. Flaherty, *A local refinement finite-element method for two-dimensional parabolic systems*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 792–811.

[3] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[4] H. Amann, *Linear and Quasilinear Parabolic Equations I*, Birkhäuser-Verlag, Berlin, 1995.

[5] K. E. Atkinson and W. Han, *Theoretical Numerical Analysis*, Springer, New York, 2001.

[6] G. Bachman and L. Narici, *Functional Analysis*, Academic Press, London, 1966.

[7] R. E. Bank, *Hierarchical bases and the finite element method*, Acta Numerica, 5 (1996), pp. 1–43.

[8] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Review, 40 (1998), pp. 789–837.

[9] F. A. Bornemann, *An adaptive multilevel approach to parabolic equations I*, Impact Comput. Sci. Engrg., 2 (1990), pp. 279–317.

[10] ——, *An adaptive multilevel approach to parabolic equations II*, Impact Comput. Sci. Engrg., 3 (1991), pp. 93–122.

[11] D. Braess, *Finite Elements*, Cambridge University Press, Cambridge, 1997.

[12] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[13] J. Butcher, *A stability property of implicit Runge-Kutta methods*, BIT, 15 (1975), pp. 358–361.

[14] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part I*, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[15] R. F. Carsel and R. S. Parrish, *Developing joint probability distributions of soil water retention characteristics*, Water Resources Research, 24 (1988), pp. 755–769.

[16] P. Deuflhard and F. Bornemann, *Numerische Mathematik II*, De Gruyter, Berlin, 1994.

[17] P. Deuflhard and A. Hohmann, *Numerische Mathematik I*, De Gruyter, Berlin, 1993.

[18] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems I: A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.

[19] ——, *Adaptive finite element methods for parabolic problems II: Optimal error estimates in $l_\infty l_2$ and $l_\infty l_\infty$*, SIAM J. Numer. Anal., 32 (1995), pp. 706–740.

[20] ——, *Adaptive finite element methods for parabolic problems IV: Nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.

[21] ——, *Adaptive finite element methods for parabolic problems V: Long-time integration*, SIAM J. Numer. Anal., 32 (1995), pp. 1750–1763.

[22] K. Eriksson, C. Johnson, and S. Larsson, *Adaptive finite element methods for parabolic problems VI: Analytic semigroups*, SIAM J. Numer. Anal., 35 (1998), pp. 1315–1325.

[23] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, Rhode Island, 1998.

[24] R. E. Ewing, R. D. Lazarov, J. E. Pasciak, and A. T. Vassilev, *Mathematical modeling, numerical techniques, and computer simulation of flows and transport in porous media*, in Proceedings of Computational Techniques and Applications: CTAC 95, World Scientific, 1995, pp. 13–30.

[25] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[26] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I*, Springer, Berlin, 2nd ed., 1993.

[27] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II*, Springer, Berlin, 2nd ed., 1996.

[28] T. Kato, *Perturbation Theory for Linear Operators*, Springer, New York, 1984.

[29] J. Korsawe and G. Starke, *Multilevel projection methods for nonlinear least-squares finite element computations*, Electr. Trans. Numer. Anal., 10 (2000), pp. 56–73.

[30] W. Kutta, *Beitrag zur nährungsweisen Integration totaler Differentialgleichungen*, Zeitschr. für Math. und Phys., 46 (1901), pp. 435–453.

[31] J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems*, Springer, Berlin, 2000.

[32] M. Majidi, *Adaptive Rothe-Verfahren für ein nichtlineares parabolisches Anfangs-Randwertproblem in der Hydrologie*. Diplomarbeit, Juli 1999.

[33] M. Majidi and G. Starke, *Least-squares Galerkin methods for parabolic problems I: Semi-discretization in time*, SIAM J. Numer. Anal., 39 (2001), pp. 1302–1323.

[34] ——, *Least-squares Galerkin methods for parabolic problems II: The fully discrete case and adaptive algorithms*, SIAM J. Numer. Anal., (2000). To Appear.

[35] Y. Mualem, *A new model for predicting the hydraulic conductivity of unsaturated porous media*, Water Resources Research, 12 (1976), pp. 513–522.

[36] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 1999.

[37] A. Ostermann and M. Thalhammer, *Convergence of Runge-Kutta methods for nonlinear parabolic equations*, Appl. Numer. Math., (2001). To Appear.

[38] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, Berlin- Heidelberg - New York, 1983.

[39] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer, Berlin, 1994.

[40] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method, E. M. I. Galligani, ed., Springer, 1977, pp. 292–315.

[41] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer, New York, 1993.

[42] C. Runge, *Ueber die numerische Auflösung von Differentialgleichungen*, Math. Ann., 46 (1895), pp. 167–178.

[43] G. Starke, *Gauss-Newton multilevel methods for least-squares finite element computations of variably saturated subsurface flow*, Computing, 64 (2000), pp. 323–338.

[44] ——, *Least-squares mixed finite element solution of variably saturated subsurface flow problems*, SIAM J. Sci. Comput., 21 (2000), pp. 1869–1885.

[45] M. E. Taylor, *Partial Differential Equations III (Nonlinear Equations)*, Applied Mathematical Sciences 117, Springer, USA, 1997.

[46] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer, Berlin, 1997.

[47] M. T. Van Genuchten, *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils*, Soil Sci. Soc. Am. J., 44 (1980), pp. 892–898.

[48] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester-Stuttgart, 1996.

[49] W. Walter, *Differential and integral inequalities*, Springer, Berlin, 1970.

[50] J. Wloka, *Partielle Differentialgleichungen*, Teubner, Stuttgart, 1982.

[51] B. I. Wohlmuth and R. H. W. Hoppe, *A comparison of a posteriori error estimators for mixed finite element discretizations by Raviart-Thomas elements*, Math. Comp., 68 (1999), pp. 1347–1378.

[52] C. S. Woodward and C. N. Dawson, *Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media*, SIAM J. Numer. Anal., 37 (2000), pp. 701–724.