Zeit- und ortsadaptive Verfahren angewandt auf Mehrphasenprobleme poröser Medien

Von der Fakultät Bauingenieur- und Vermessungswesen
der Universität Stuttgart zur Erlangung der Würde
eines Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Abhandlung

Vorgelegt von

Dipl.-Math. Peter Ellsiepen

aus

Konstanz

Hauptberichter: Prof. Dr.-Ing. Wolfgang Ehlers
Mitberichter: Prof. Dr. Karl Graf Finck von Finckenstein

Tag der mündlichen Prüfung: 9. Juli 1999

Institut für Mechanik (Bauwesen) der Universität Stuttgart
Lehrstuhl II, Prof. Dr.-Ing. W. Ehlers

Bericht Nr. II-3 aus dem Institut für Mechanik (Bauwesen), Lehrstuhl II,
Universität Stuttgart

Herausgeber:
Prof. Dr.-Ing. W. Ehlers

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Mechanik der Technischen Hochschule Darmstadt sowie anschließend am Institut für Mechanik (Bauwesen) der Universität Stuttgart.

Mein besonderer Dank gilt Herrn Prof. Dr.-Ing. Wolfgang Ehlers für die Anregung zu dieser Arbeit, für seine Unterstützung und Förderung sowie für seine stets vorhandene Bereitschaft zu Diskussionen über die Mechanik und angrenzende Gebiete wie die Numerische Mathematik. Seine Aufgeschlossenheit gegenüber interdisziplinären Fragestellungen hat dabei sehr zum Gelingen dieser Arbeit beigetragen. Desweiteren danke ich Herrn Ehlers für die Übernahme des Hauptberichts.

Herrn Prof. Dr. Karl Graf Finck von Finckenstein verdanke ich durch mein Studium in Darmstadt das Interesse und die Begeisterung für die Numerische Mathematik. Deren Anwendung auf ingenieurwissenschaftliche Probleme war letztlich die Motivation für meinen Wechsel in die Mechanik und die treibende Kraft für die vorliegende Arbeit. Ganz besonders danke ich Herrn von Finckenstein für sein großes Interesse an der Arbeit sowie für die bereitwillige und schnelle Übernahme des Mitberichts.

Zu großem Dank verpflichtet bin ich Herrn Dr.-Ing. Stefan Diebels für die vielen und zum Teil sehr ausführlichen Diskussionen und Gespräche über die Kontinuumsmechanik und angrenzende Gebiete. Dies hat mir den Einstieg in das für mich neue Thema der Kontinuumsmechanik enorm erleichtert. Außerdem hat mir Herr Diebels durch seine Offenheit und Geduld bei der Klärung von verschiedensten Fragestellungen immer wieder geholfen, die Freude an der Arbeit zu behalten.

An die zahlreichen Diskussionen mit Herrn Dr. Peter Fritzen über die Mechanik und die Numerische Mathematik denke ich mit Freude zurück. Durch den ständigen Austausch und den Vergleich von mathematischer und ingenieurwissenschaftlicher Sichtweise haben diese Diskussionen in vielen Punkten zu einer klareren Formulierung der Arbeit und zu dem nötigen Brückenschlag zwischen den beiden „Welten“ beigetragen. Daneben möchte ich mich ganz besonders für seine Unterstützung und Freundschaft bedanken.

Die langwierige Arbeit des Korrekturlesens haben vor allem Herr Martin Ammann, Herr Dr.-Ing. Stefan Diebels, Herr Johannes Dürr, Herr Dr. Peter Fritzen und Frau Claudia Mehlig übernommen. Dafür sei ihnen ganz herzlich gedankt. Durch zahlreiche Vorschläge und Hinweise haben sie zu vielen Verbesserungen und zur Klarheit des Textes beigetragen.

Nicht zuletzt gilt mein Dank allen Kollegen und Mitarbeitern am Institut für Mechanik (Bauwesen) der Universität Stuttgart für das angenehme und freundliche Arbeitsklima und für die in vielerlei Hinsicht hilfreiche Unterstützung.

Abschließend möchte ich mich bei meiner Familie und meinen Freunden für die fortwährende Unterstützung und für die entgegengebrachte Toleranz bedanken, ohne die diese Arbeit nicht zustande gekommen wäre.

Stuttgart, im Juli 1999

Peter Ellsiepen

Inhaltsverzeichnis

Einleitung und Übersicht

Motivation

In den letzten Jahren hat das Interesse an der Modellierung und numerischen Simulation poröser Materialien und Werkstoffe stark zugenommen. So werden etwa poröse Metallschäume in technischen Anwendungen als Energieabsorber oder Leichtbauwerkstoffe eingesetzt. Desweiteren spielen hochelastische Kunststoffschäume in der Automobilindustrie als Insassenaufprallschutz oder zu Dämpfungszwecken eine wichtige Rolle. In der Medizin treten poröse Materialien beispielsweise in Form von blutdurchströmtem Muskelgewebe oder Knorpel- bzw. Knochenmaterialien auf, deren numerische Simulation im Hinblick auf die immer häufiger eingesetzten Prothesen und Implantate zunehmend Bedeutung erlangt. Im Bauwesen findet man poröse Materialien in Form von nichtbindigen Böden (Sande) bzw. bindigen Böden (Schluffe, Tone) vor, die vor allem bei der Gründung von Bauwerken eine große Bedeutung haben. So erhalten etwa theoretische Erkenntnisse aus dem Bereich der Baugrund-Tragwerk-Interaktion in letzter Zeit verstärkt Einzug in die Planungs- und Bauausführungsphase größerer Bauwerke.

Die numerische Simulation derartiger Anwendungsprobleme gewinnt insbesondere durch die enorme Leistungssteigerung moderner Computer zunehmend an Bedeutung. Allerdings steigen gleichzeitig mit der Computerleistung auch die Anforderungen an die Komplexität der behandelten Probleme, so daß die Entwicklung effizienter numerischer Methoden auf der Basis moderner adaptiver Strategien einen hohen Stellenwert einnimmt. Die vom Computer-Pionier *Moore* vor etwa zwanzig Jahren empirisch aufgestellte These der Verdopplung der Prozessorleistung alle eineinhalb Jahre hat bis heute ihre Gültigkeit. Darüber hinaus bietet die Optimierung numerischer Methoden jedoch ein Vielfaches an Potential zur Leistungssteigerung. Halbiert man etwa durch Anwendung ortsadaptiver Strategien die Zahl der Freiheitsgrade, so führt dies bei heutigen Gleichungslösern grob zu einer Viertelung des Lösungsaufwands. Diese Effizienzsteigerung kann sich durch den gleichzeitigen Einsatz zeitadaptiver Strategien noch drastischer auswirken, da bei nichtlinearen Problemen je gerechnetem Zeitschritt mehrere lineare Gleichungssysteme gelöst werden müssen. Insbesondere bei Problemen mit starken Inhomogenitäten sowie Instabilitäten – wie z. B. bei zeitlich veränderlichen plastischen Zonen bzw. beim Scherbandproblem in Böden – gilt daher das Interesse einer möglichst ökonomischen Wahl der Zeitschrittweite und der Netzdichte zur Erreichung einer vorgegebenen Genauigkeit.

Problematik, Zielsetzung und Vorgehensweise

Aus Sicht der Kontinuumsmechanik ist es möglich, die o. g. porösen Materialien mit einer einheitlichen Theorie zu beschreiben, der *Theorie Poröser Medien* (TPM). Diese Art der Modellbildung führt zu volumengekoppelten Mehrfeldproblemen, die keiner der bei-

den klassischen Disziplinen der Festkörper- oder Fluidmechanik direkt zugeordnet werden können.

Ziel der vorliegenden Arbeit ist die numerische Simulation von Anfangs-Randwertproblemen auf Basis der TPM, wobei Phänomenen der Bodenmechanik besonderes Augenmerk gilt. Eine spezielle Problematik in diesem Bereich stellt die Scherbandbildung in Böden dar, bei deren numerischer Simulation der Auswahl der verwendeten Methoden entscheidende Bedeutung zukommt, da starke Inhomogenitäten in der Lösung auftreten. So führt im Zeitbereich etwa das Einsetzen der Plastizität zu Unstetigkeiten in der Ableitung der zugrundeliegenden partiellen Differentialgleichung. Im Ortsbereich treten insbesondere an den Rändern von Scherbändern große Lösungsgradienten oder gar Unstetigkeiten auf. Zur praxisnahen numerischen Simulation werden daher effiziente Lösungsverfahren benötigt, die diese Phänomene mit einer vorgegebenen Genauigkeit erfassen können.

Im Rahmen dieser Arbeit wird unter Verwendung moderner adaptiver Strategien in Zeit und Ort ein effizientes Gesamtverfahren auf Basis der Methode der finiten Elemente (FEM) entwickelt, das die numerische Simulation von Mehrphasenproblemen poröser Medien ermöglicht. Beispielhaft wird dabei ein inkompressibles Zweiphasenmodell, bestehend aus einem viskoplastischen Festkörperskelett und einer viskosen Fluidphase, in einer geometrisch linearen Formulierung zugrundegelegt.

Die Gesamtstruktur der entwickelten Algorithmen erlaubt darüber hinaus auch die Anwendung auf andere kontinuumsmechanische Modelle. So können aufgrund der Abwärtskompatibilität der TPM beim Übergang zu einphasigen Materialien auch klassische Modelle der Kontinuumsmechanik mit zugehörigen Plastizitätsformulierungen einbezogen werden. In Erweiterung des betrachteten inkompressiblen Zweiphasenmodells sind aber auch komplexere Mehrphasenmodelle denkbar. In dieser Hinsicht sind etwa Plastizitätsformulierungen mit Verfestigung sowie geometrisch nichtlineare Formulierungen zu nennen; außerdem können Dreiphasenmodelle – bestehend aus Festkörper, Flüssigkeit und Gas – zur Beschreibung von teilgesättigten Bodenzonen sowie Theorien mit Rotationsfreiheitsgraden (*Cosserat*) zur Beschreibung granularer Materialien einbezogen werden.

Die vorliegende Arbeit und die damit verbundene Entwicklung des Finite-Elemente-Programmsystems PANDAS[1] kann somit als Basis für Weiterentwicklungen auf dem Gebiet der numerischen Simulation von Modellen der Elastizitäts- und Plastizitätstheorie sowie von Mehrphasenproblemen poröser Medien dienen.

Gliederung und Umfang der Arbeit

Die Arbeit gliedert sich in fünf Kapitel, wobei zu Beginn jedes Kapitels ein Literaturüberblick über den Stand der Forschung des darin behandelten Fachgebiets gegeben wird.

In *Kapitel 1* werden die notwendigen kontinuumsmechanischen Grundlagen sowie die verwendete Notation bereitgestellt. Darauf aufbauend wird ein inkompressibles Zweiphasenmodell in einer geometrisch linearen Formulierung abgeleitet, das der weiteren Arbeit zugrundeliegt.

[1] **P**orous media **A**daptive **N**onlinear finite element solver based on **D**ifferential **A**lgebraic **S**ystems [50]

Kapitel 2 befaßt sich zunächst mit der starken Formulierung der Anfangs-Randwertprobleme im quasi-statischen und im dynamischen Fall sowie den zugehörigen Anfangs- und Randbedingungen. Eine konsequente mathematische Notation bei der schwachen Formulierung und der Ortsdiskretisierung mit finiten Elementen ermöglicht die einheitliche Darstellung verschiedenster Modelltypen als System differential-algebraischer Gleichungen (DAE) und eröffnet den Weg zur Anwendung moderner Zeitintegrationsverfahren.

Den Schwerpunkt in *Kapitel 3* bildet die adaptive Zeitintegration von DAE-Systemen. Nach einer Begriffsklärung und der Index-Bestimmung der ortsdiskreten Systeme aus Kapitel 2 wird die Problematik der steifen Differentialgleichungen im Zusammenhang mit einigen Stabilitätsbegriffen beleuchtet und deren Bedeutung für die behandelte Problemklasse diskutiert. Die Aufstellung einer Liste von Kriterien erlaubt anschließend die systematische Auswahl einer Verfahrensklasse, den diagonalimpliziten *Runge-Kutta*-Verfahren (DIRK). Insbesondere können bekannte Techniken zur Lösung der nichtlinearen Gleichungssysteme, die bei Anwendung des impliziten *Euler*-Verfahrens auf Plastizitätsprobleme auftreten (algorithmisch konsistente Linearisierung), systematisch auf die Klasse der DIRK-Verfahren verallgemeinert werden. Für die adaptive Zeitintegration wird im weiteren eine verläßliche Schätzung des lokalen Zeitfehlers benötigt. Hierfür bieten sich entweder die Verwendung einer lokalen *Richardson*-Extrapolation oder die Konstruktion eingebetteter Verfahren an, wobei letztere den Vorteil haben, daß die Fehlerschätzung nahezu ohne zusätzlichen Aufwand berechnet werden kann. Darauf aufbauend erhält man mit Hilfe üblicher Techniken der Schrittweitensteuerung aus dem Bereich der numerischen Mathematik sehr effiziente zeitadaptive Verfahren. Neben der Zusammenstellung bekannter Methoden wird zum Abschluß des Kapitels ein neues eingebettetes Verfahren konstruiert, das insbesondere auf Probleme mit geringer Lösungsregularität im Zeitbereich zugeschnitten ist, wie etwa den Grenzfall der idealen *Prandtl-Reuß*-Plastizität.

In *Kapitel 4* werden zunächst gängige Techniken zur Schätzung des Ortsfehlers im Rahmen ortsadaptiver FE-Strategien diskutiert und vergleichend gegenübergestellt. Dies bildet die Grundlage zur Konstruktion eines neuen Fehlerindikators, der alle treibenden Größen des behandelten Mehrphasenproblems berücksichtigt. Darauf aufbauend werden verschiedene Techniken zur Steuerung der Netzdichte vorgestellt und die Problematik der Datenhaltung und Netzgenerierung diskutiert. Vor allem in programmtechnischer Hinsicht stellt letzteres den wesentlichen Aufwand bei der Implementierung ortsadaptiver Strategien dar. Den Abschluß des Kapitels bildet die Kopplung der ortsadaptiven mit den zeitadaptiven Verfahren aus Kapitel 3. Bei Netzänderungen spielt in diesem Zusammenhang insbesondere der konsistente Transfer der diskreten Zustandsgrößen an den FE-Knoten und Integrationspunkten eine wichtige Rolle.

In *Kapitel 5* werden zunächst die verschiedenen Aspekte der zeit- und ortsadaptiven Verfahren anhand von Verifikationsbeispielen numerisch überprüft und bewertet. Es zeigt sich eine hervorragende Übereinstimmung zwischen den in Kapitel 3 und 4 dargestellten theoretischen Aussagen und den in den numerischen Rechnungen gewonnenen Ergebnissen. Zum Abschluß der Arbeit wird durch die Anwendung der entwickelten Verfahren auf typische Probleme der Bodenmechanik demonstriert, daß das eingangs gestellte Ziel der Simulation komplexer, praxisnaher Probleme auf Basis der TPM erreicht werden konnte.

Kapitel 1: Theorie Poröser Medien

Geomaterialien wie Böden, Sandsteine, Felsen oder Steinsalze, aber auch technische Materialien wie Elastomer- oder Metallschäume bestehen aus einem porösen Festkörperskelett, dessen Poren mit einem oder mehreren Fluiden gesättigt sind. Das Verhalten des Gesamtkörpers aus Festkörperskelett und Fluiden wird durch die Eigenschaften seiner Bestandteile bestimmt. Da die genaue Porenstruktur meist nicht bekannt ist, gelangt man erst nach einem gedachten oder real ausgeführten statistischen Mittelungsprozeß (Homogenisierung) zu einem kontinuumsmechanischen Modell. Eine Möglichkeit der Modellbildung im Rahmen einer makroskopischen Theorie ist die *Theorie Poröser Medien* (TPM), die auf der klassischen Mischungstheorie mit superponierten Kontinua basiert. Durch Einführung der Volumenanteile als Strukturvariablen (Konzept der Volumenanteile) kann die mikroskopische Zusammensetzung der Mischung erfaßt werden.

Erste Ansätze zur Entwicklung einer Theorie poröser Medien unter Berücksichtigung von Volumenanteilen gehen in die dreißiger Jahre zurück (*Fillunger* [55]). Diese Ideen wurden Ende der fünfziger Jahre von *Heinrich & Desoyer* [68, 69, 70] weiterentwickelt. Als Begründer der modernen TPM in ihrer heutigen Formulierung kann *Bowen* [23, 24] betrachtet werden, der erstmals ein Zweiphasenmodell mit inkompressiblen bzw. mit kompressiblen Konstituierenden angegeben hat. In den letzten Jahren wurde die Theorie von *de Boer & Ehlers* [21] und *Ehlers* [44, 45, 47] systematisch aufgearbeitet und weiterentwickelt. Die darin verwendete moderne, basisfreie Notation der Mehrphasen-Kontinuumsmechanik wird in der vorliegenden Arbeit zugrundegelegt und in diesem Kapitel knapp dargestellt. Am Ende des Kapitels wird ein Zweiphasenmodell in einer geometrisch linearen Formulierung abgeleitet, das als Grundlage für die weitere Arbeit dient.

1.1 Kinematik

Bevor auf die Kinematik von Mischungen eingegangen wird, sollen zunächst einige Begriffe aus der klassischen Kontinuumsmechanik eingeführt werden. Zentrale Bedeutung kommt dabei dem Begriff des Körpers zu, dessen Bewegung Gegenstand der Betrachtung in der Kontinuumsmechanik ist.

1.1.1 Kinematik eines Körpers

Def. 1.1: Ein *materieller Körper* $\mathcal{B} = \{\mathcal{X}\}$ ist eine zusammenhängende Menge von Elementen $\mathcal{X} \in \mathcal{B}$, die als *materielle Punkte* bezeichnet werden. ■

Ausgehend von dieser Definition eines materiellen Körpers gibt es verschiedene Möglichkeiten, die Bewegung eines Körpers zu beschreiben. Bei *Haupt* [67] werden den materiellen Punkten $\mathcal{X}$ des Körpers mit Hilfe einer Referenzplazierung zunächst Namen X zugeordnet. Als Spezialfall kann dann angenommen werden, daß zu einem festen Zeitpunkt t_0

jeder materielle Punkt eine ausgezeichnete Position $\mathbf{X}_0$ im *Euklid*schen Vektorraum $\mathbb{E}^3$ einnimmt[1]. Bei *Marsden & Hughes* [83] wird ein Körper als eine zwei- oder dreidimensionale Mannigfaltigkeit aufgefaßt, die nur in Sonderfällen mit einer offenen Teilmenge des Vektorraums $\mathbb{R}^3$ identifizierbar ist. Im Rahmen der vorliegenden Arbeit wird vereinfachend angenommen, daß die Lage des Körpers $\mathcal{B}$ als eine offene Teilmenge im dreidimensionalen Raum $\mathbb{E}^3$ zu einem festen Zeitpunkt t_0 bekannt ist. Zunächst wird definiert:

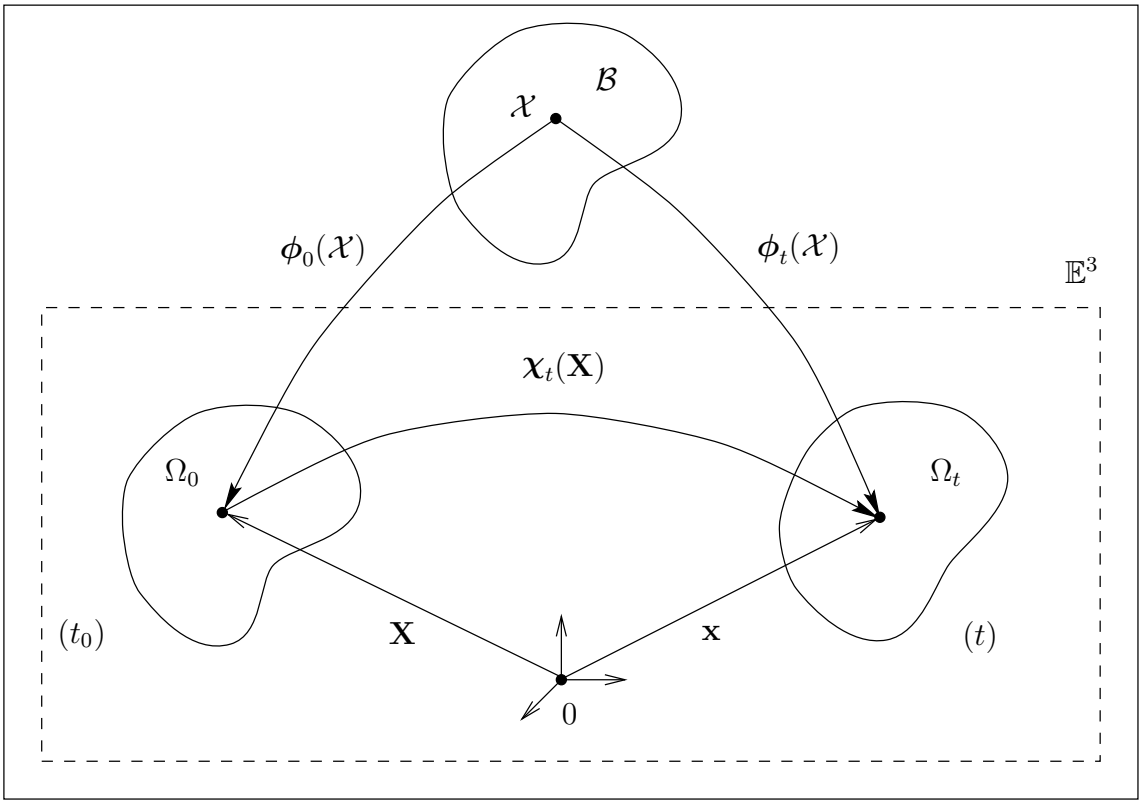


Abbildung 1.1: *Kinematik eines Körpers*

Def. 1.2: Eine *Plazierung* (*Konfiguration*) des Körpers $\mathcal{B}$ ist eine bijektive Abbildung[2]

$$\phi : \mathcal{B} \longrightarrow \Omega \subset \mathbb{E}^3 \,, \tag{1.1}$$

die jedem materiellen Punkt $\mathcal{X} \in \mathcal{B}$ einen Ortsvektor $\mathbf{x} = \phi(\mathcal{X}) \in \mathbb{E}^3$ zuordnet. Die Menge aller Konfigurationen des Körpers $\mathcal{B}$ wird mit $\mathcal{C}$ bzw. $\mathcal{C}(\mathcal{B})$ bezeichnet.

[1]Mit dem *Euklid*schen Vektorraum $\mathbb{E}^3$ ist im folgenden der dreidimensionale Anschauungsraum gemeint, dessen Elemente Ortsvektoren in einer basisfreien Notation sind. Je nach Sinnzusammenhang werden Ortsvektoren im Vektorraum $\mathbb{E}^3$ auch als Punkte bzw. Raumpunkte bezeichnet, was aufgrund der eineindeutigen Zuordnung kein Problem darstellt. Außerdem kann nach Festlegung einer Basis jedem Ortsvektor $\mathbf{x} \in \mathbb{E}^3$ eineindeutig ein Zahlentripel $(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3) \in \mathbb{R}^3$ zugeordnet werden, so daß $\mathbb{E}^3$ und $\mathbb{R}^3$ isomorphe Vektorräume sind. Die Bezeichnungen $\mathbb{E}^3$ und $\mathbb{R}^3$ werden daher in vielen Arbeiten synonym verwendet.

[2]Streng genommen ist hier wie im folgenden ein Homöomorphismus gemeint, also eine umkehrbar eindeutige (bijektive) stetige Abbildung mit stetiger Inverser.

Die *Referenzkonfiguration*

$$\boldsymbol{\phi}_0 : \mathcal{B} \longrightarrow \Omega_0 \subset \mathbb{E}^3 \tag{1.2}$$

ist eine fest gewählte Konfiguration des Körpers $\mathcal{B}$, die jedem materiellen Punkt $\mathcal{X} \in \mathcal{B}$ einen Ortsvektor $\mathbf{X} = \boldsymbol{\phi}_0(\mathcal{X}) \in \Omega_0$ zuordnet. Gleichbedeutend wird auch die offene Menge Ω_0 als Referenzkonfiguration bezeichnet. ■

Def. 1.3: Eine *Bewegung* ist eine Kurve in der Menge $\mathcal{C}$ aller Konfigurationen, d. h. eine Abbildung $\mathbb{R} \ni t \mapsto \boldsymbol{\phi}_t \in \mathcal{C}$, die jedem Zeitpunkt t aus einem reellen Zeitintervall eine Konfiguration $\boldsymbol{\phi}_t$ zuordnet. Anders ausgedrückt ist die Bewegung eine mit t parametrisierte Schar von Konfigurationen

$$\boldsymbol{\phi}_t : \mathcal{B} \longrightarrow \Omega_t \subset \mathbb{E}^3 \,, \tag{1.3}$$

die jedem materiellen Punkt $\mathcal{X}$ zu jedem Zeitpunkt $t \geq t_0$ einen Ortsvektor $\mathbf{x} = \boldsymbol{\phi}(\mathcal{X}, t) \in \Omega_t$ zuordnet. Die Menge $\Omega_t = \boldsymbol{\phi}_t(\mathcal{B})$ heißt *Momentankonfiguration* (*aktuelle Konfiguration*) des Körpers $\mathcal{B}$. Es wird angenommen, daß die Plazierung zum Zeitpunkt t_0 mit der Referenzkonfiguration zusammenfällt:

$$\Omega_0 = \boldsymbol{\phi}_0(\mathcal{B}) = \boldsymbol{\phi}_{t_0}(\mathcal{B}) = \Omega_{t_0} \,. \tag{1.4}$$

Ein tiefgestellter Index an der Funktion $\boldsymbol{\phi}$ bezeichnet das Festhalten einer der Variablen $\mathcal{X}$ oder t:

- $\boldsymbol{\phi}_t(\mathcal{X})$: Konfiguration zu einem festen Zeitpunkt t,
- $\boldsymbol{\phi}_{\mathcal{X}}(t)$: Bahnkurve des materiellen Punktes $\mathcal{X}$,
- $\boldsymbol{\phi}(\mathcal{X}, t)$: allgemeine (punkt- und zeitabhängige) Bewegung.

Aufgrund der eineindeutigen Zuordnung zwischen materiellen Punkten $\mathcal{X}$ und Ortsvektoren $\mathbf{X}$ der Referenzkonfiguration kann die Bewegung auch auf die Referenzkonfiguration bezogen werden. Man erhält dann die Bewegung $\boldsymbol{\chi}_t = \boldsymbol{\phi}_t \circ \boldsymbol{\phi}_0^{-1} : \Omega_0 \longrightarrow \Omega_t$ (Abbildung 1.1):

$$\boldsymbol{\chi}_t(\mathbf{X}) = \boldsymbol{\phi}_t(\boldsymbol{\phi}_0^{-1}(\mathbf{X})) \quad \text{bzw.} \quad \boldsymbol{\chi}(\mathbf{X}, t) = \boldsymbol{\phi}(\boldsymbol{\phi}_0^{-1}(\mathbf{X}), t) \,. \tag{1.5}$$

Analog zu $\boldsymbol{\phi}_t(\mathcal{X})$, $\boldsymbol{\phi}_{\mathcal{X}}(t)$ und $\boldsymbol{\phi}(\mathcal{X}, t)$ sind auch die Bezeichnungen $\boldsymbol{\chi}_t(\mathbf{X})$, $\boldsymbol{\chi}_{\mathbf{X}}(t)$ und $\boldsymbol{\chi}(\mathbf{X}, t)$ zu verstehen. ■

1.1.2 Kinematik von Mischungen

Die Mischungstheorie ist eine Erweiterung der klassischen Kontinuumsmechanik. Es werden Mischungen betrachtet, die aus mehreren Konstituierenden (Festkörpern oder Fluiden) bestehen, vgl. dazu *Truesdell & Toupin* [118, §158, 159], *Bowen* [22], *de Boer & Ehlers* [21], *Ehlers* [44].

Eine genaue Unterscheidung zwischen einer Phase als physikalischem Aggregatzustand (fest, flüssig oder gasförmig) und einer Komponente als physikalischem Stoff, der auch in mehreren Phasen vorkommen kann (z. B. die Komponente Wasser als Eis, Wasser oder

Dampf) wird im Rahmen dieser Arbeit nicht vorgenommen. Vielmehr werden alle vorkommenden Bestandteile einer Mischung (ob Phase oder Komponente) als Konstituierende bezeichnet.

Def. 1.4: Eine *Mischung* φ (auch *Mehrphasenkontinuum*) ist aus mehreren Bestandteilen zusammengesetzt, die als *Konstituierende* φ^α bezeichnet werden ($\alpha = 1, \ldots, n$). Synonym wird auch der Begriff der *Phase* φ^α verwendet. Es wird im weiteren angenommen, daß die einzelnen Phasen in der Mischung immer identifizierbar bleiben, d. h. daß keine chemischen Reaktionen zwischen den Phasen stattfinden. ■

In der Mischungstheorie geht man davon aus, daß die in einem repräsentativen Elementarvolumen (REV) enthaltenen Konstituierenden durch einen gedachten oder real ausgeführten statistischen Mittelungsprozeß (Homogenisierung) über das gesamte Elementarvolumen „verschmiert“ sind. Daher überlagert man in der kontinuumsmechanischen Betrachtung die Kontinua der einzelnen Konstituierenden, so daß sich in der Momentankonfiguration an einem Raumpunkt gleichzeitig materielle Punkte aller Konstituierenden befinden. Jede Konstituierende folgt dabei ihrer eigenen Bewegung, so daß die zum Zeitpunkt t am gleichen Ort befindlichen materiellen Punkte von verschiedenen Orten der jeweiligen Referenzkonfiguration stammen können.

Eine analoge Erweiterung der im vorherigen Abschnitt eingeführten Grundbegriffe der Kontinuumsmechanik auf Mischungen führt zu den folgenden Definitionen:

Def. 1.5: Ein *Mischungskörper* ist eine Überlagerung von *Teilkörpern* $\mathcal{B}^\alpha$. Man spricht daher in der Mischungstheorie von *superponierten Kontinua*. ■

Def. 1.6: Eine *Plazierung* (*Konfiguration*) des Teilkörpers $\mathcal{B}^\alpha$ ist eine bijektive Abbildung

$$\overset{\alpha}{\boldsymbol{\phi}} : \mathcal{B}^\alpha \longrightarrow \Omega \subset \mathbb{E}^3 \,, \tag{1.6}$$

die jedem materiellen Punkt $\mathcal{X}^\alpha \in \mathcal{B}^\alpha$ einen Ortsvektor $\mathbf{x} = \overset{\alpha}{\boldsymbol{\phi}}(\mathcal{X}^\alpha) \in \mathbb{E}^3$ zuordnet. Die Menge aller Konfigurationen des Körpers $\mathcal{B}^\alpha$ wird mit $\mathcal{C}^\alpha$ bzw. $\mathcal{C}^\alpha(\mathcal{B}^\alpha)$ bezeichnet.

Die *Referenzkonfiguration*

$$\overset{\alpha}{\boldsymbol{\phi}}_0 : \mathcal{B}^\alpha \longrightarrow \Omega_{0\alpha} \subset \mathbb{E}^3 \tag{1.7}$$

des Teilkörpers $\mathcal{B}^\alpha$ ist eine fest gewählte Konfiguration, die jedem materiellen Punkt $\mathcal{X}^\alpha \in \mathcal{B}^\alpha$ einen Ortsvektor $\mathbf{X}_\alpha = \overset{\alpha}{\boldsymbol{\phi}}_0(\mathcal{X}^\alpha) \in \Omega_{0\alpha}$ zuordnet. Gleichbedeutend wird auch die offene Menge $\Omega_{0\alpha}$ als Referenzkonfiguration bezeichnet. ■

Def. 1.7: Eine *Bewegung* ist eine Kurve in der Menge $\mathcal{C}^\alpha$ aller Konfigurationen, d. h. eine Abbildung $\mathbb{R} \ni t \mapsto \overset{\alpha}{\boldsymbol{\phi}}_t \in \mathcal{C}^\alpha$, die jedem Zeitpunkt t aus einem reellen Zeitintervall eine Konfiguration $\overset{\alpha}{\boldsymbol{\phi}}_t$ zuordnet. Anders ausgedrückt ist die Bewegung eine mit t parametrisierte Schar von Konfigurationen

$$\overset{\alpha}{\boldsymbol{\phi}}_t : \mathcal{B}^\alpha \longrightarrow \Omega_t \subset \mathbb{E}^3 \,, \tag{1.8}$$

die jedem materiellen Punkt $\mathcal{X}^\alpha$ zu jedem Zeitpunkt $t \geq t_0$ einen Ortsvektor $\mathbf{x} = \overset{\alpha}{\boldsymbol{\phi}}(\mathcal{X}^\alpha, t) \in \Omega_t$ zuordnet. Die Menge $\Omega_t = \overset{\alpha}{\boldsymbol{\phi}}_t(\mathcal{B}^\alpha)$ heißt *Momentankonfiguration* (*aktuelle*

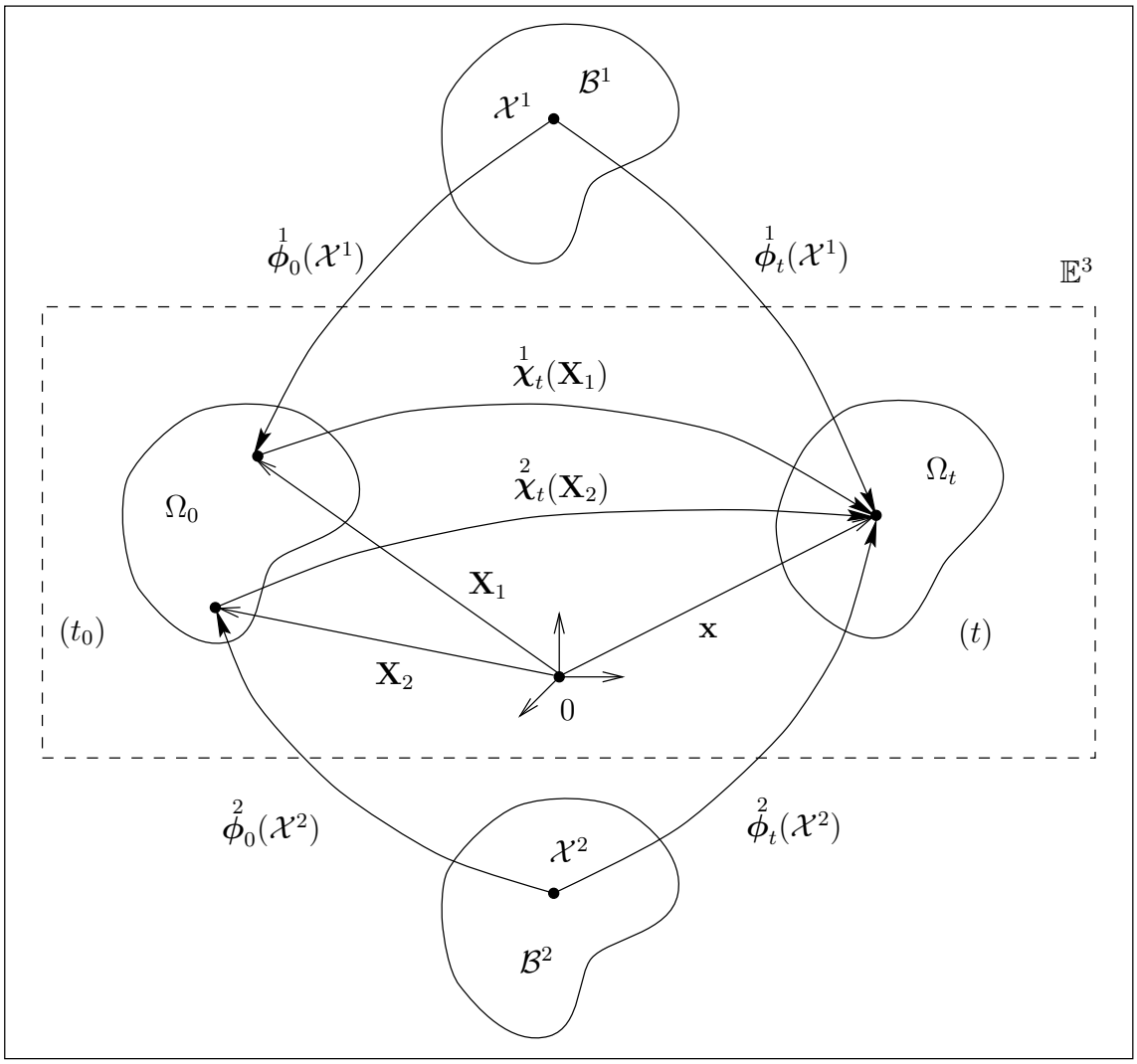


Abbildung 1.2: *Kinematik eines Zweiphasenkontinuums*

Konfiguration) der Mischung. Es wird angenommen, daß die Plazierung zum Zeitpunkt t_0 mit der jeweiligen Referenzkonfiguration zusammenfällt:

$$\Omega_{0\alpha} = \overset{\alpha}{\phi}_0(\mathcal{B}^\alpha) = \overset{\alpha}{\phi}_{t_0}(\mathcal{B}^\alpha) = \Omega_{t_0} =: \Omega_0 \,. \tag{1.9}$$

Aufgrund der eineindeutigen Zuordnung zwischen materiellen Punkten $\mathcal{X}^\alpha$ und Ortsvektoren $\mathbf{X}_\alpha$ der Referenzkonfiguration kann die Bewegung auch auf die Referenzkonfiguration bezogen werden. Man erhält dann die Bewegungen $\overset{\alpha}{\boldsymbol{\chi}}_t = \overset{\alpha}{\boldsymbol{\phi}}_t \circ \overset{\alpha}{\boldsymbol{\phi}}_0^{-1} : \Omega_0 \longrightarrow \Omega_t$ mit

$$\overset{\alpha}{\boldsymbol{\chi}}_t(\mathbf{X}_\alpha) = \overset{\alpha}{\boldsymbol{\phi}}_t(\overset{\alpha}{\boldsymbol{\phi}}_0^{-1}(\mathbf{X}_\alpha)) \quad \text{bzw.} \quad \overset{\alpha}{\boldsymbol{\chi}}(\mathbf{X}_\alpha, t) = \overset{\alpha}{\boldsymbol{\phi}}(\overset{\alpha}{\boldsymbol{\phi}}_0^{-1}(\mathbf{X}_\alpha), t) \,. \tag{1.10}$$

Dieser Zusammenhang ist in Abbildung 1.2 graphisch veranschaulicht. ■

Bemerkung: Die Annahme einer gemeinsamen Referenzkonfiguration aller Phasen zum Zeitpunkt t_0 bedeutet, daß die Mischung von Anfang an besteht. Die materiellen Punkte

$\mathcal{X}^\alpha$ der einzelnen Phasen bewegen sich aber auf jeweils eigenen Bahnlinien $\overset{\alpha}{\boldsymbol{\phi}}_{\mathcal{X}^\alpha}(t)$. □

1.1.3 Konzept der Volumenanteile

Die Mischungstheorie erlaubt keine Beschreibung von inneren Strukturen des betrachteten Mischungskörpers. In der *Theorie Poröser Medien* werden deshalb als statistisch gemittelte Strukturvariablen die Volumenanteile eingeführt, vgl. *Ehlers* [44]. Damit kann z. B. die Porenraumverteilung eines mit einem Fluid angefüllten Festkörpers dargestellt werden (Abbildung 1.3).

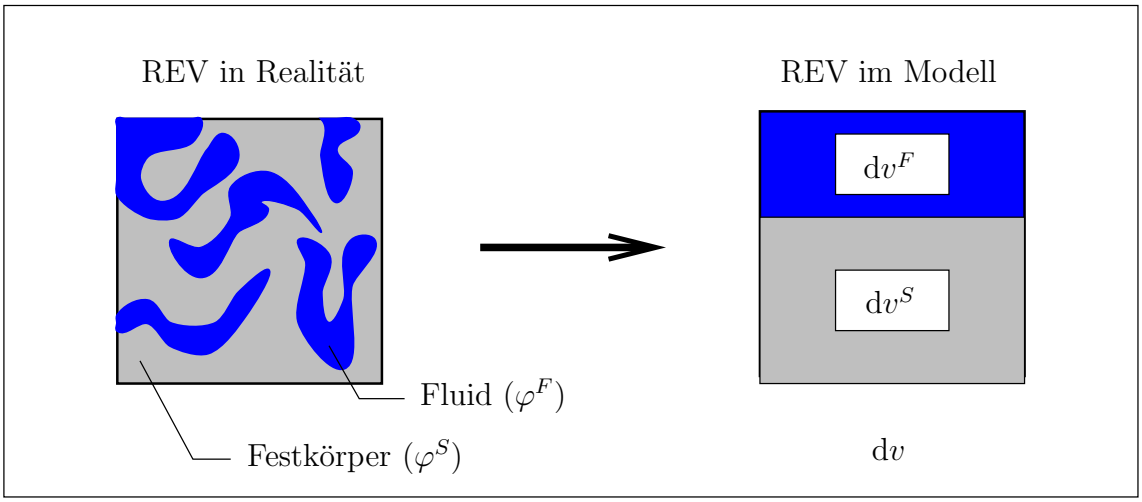


Abbildung 1.3: *Homogenisierung und Konzept der Volumenanteile, schematische Darstellung anhand eines repräsentativen Elementarvolumens (REV)*

Def. 1.8: Der *Volumenanteil* ist ein Skalarfeld $n^\alpha : \Omega_t \longrightarrow \mathbb{R}$, das jedem Raumpunkt $\mathbf{x} \in \Omega_t$ der Momentankonfiguration den lokalen Anteil des Volumens der Phase φ^α am Gesamtvolumen der Mischung zuordnet. Das *Partialvolumen* V^α der Phase φ^α ist damit:

$$V^\alpha = \int\limits_{\Omega_t} n^\alpha(\mathbf{x})\,\mathrm{d}v\,. \tag{1.11}$$

Das *Volumen* V der Mischung ist die Summe der Partialvolumina V^α:

$$V = \int\limits_{\Omega_t} \mathrm{d}v \;= \sum_\alpha V^\alpha = \sum_\alpha \int\limits_{\Omega_t} n^\alpha\,\mathrm{d}v \;= \int\limits_{\Omega_t} \sum_\alpha n^\alpha\,\mathrm{d}v \;=: \int\limits_{\Omega_t} \sum_\alpha \mathrm{d}v^\alpha\,. \tag{1.12}$$

Das *Partialvolumenelement* $\mathrm{d}v^\alpha$ der Phase φ^α ist das mit dem Volumenanteil gewichtete Volumenelement der Mischung:

$$\mathrm{d}v^\alpha = n^\alpha\,\mathrm{d}v\,. \tag{1.13}$$

Aus Gleichung (1.12) ergibt sich sofort, daß die betrachtete Mischung *gesättigt*[3] ist:

$$\sum_{\alpha} n^{\alpha} = 1\,. \tag{1.14}$$

Analog zu den Volumenelementen wird das *partiale Oberflächenelement*

$$\mathrm{d}a^{\alpha} = n^{\alpha}\,\mathrm{d}a \tag{1.15}$$

eingeführt. ■

Mit dem Konzept der Volumenanteile ergeben sich zwei verschiedene Dichtefunktionen, die das Massenelement einer Phase entweder auf das Gesamtvolumen oder auf das Partialvolumen beziehen.

Def. 1.9: Die *Partialdichte* $\rho^{\alpha} : \Omega_t \longrightarrow \mathbb{R}$ ist das lokale Verhältnis der Masse der Phase φ^{α} zum Gesamtvolumen:

$$\rho^{\alpha} = \frac{\mathrm{d}m^{\alpha}}{\mathrm{d}v}\,. \tag{1.16}$$

Die *realistische Dichte* oder *effektive Dichte* $\rho^{\alpha R} : \Omega_t \longrightarrow \mathbb{R}$ ist das lokale Verhältnis der Masse der Phase φ^{α} zum Partialvolumen:

$$\rho^{\alpha R} = \frac{\mathrm{d}m^{\alpha}}{\mathrm{d}v^{\alpha}}\,. \tag{1.17}$$

Die *Mischungsdichte* $\rho : \Omega_t \longrightarrow \mathbb{R}$ ist die Summe der Partialdichten:

$$\rho = \sum_{\alpha} \rho^{\alpha} = \sum_{\alpha} \frac{\mathrm{d}m^{\alpha}}{\mathrm{d}v} = \frac{\mathrm{d}m}{\mathrm{d}v}\,. \tag{1.18}$$

Darin ist $\mathrm{d}m = \sum_{\alpha} \mathrm{d}m^{\alpha}$ das Massenelement der Mischung. ■

Eine direkte Folgerung aus (1.13), (1.16) und (1.17) ist der Zusammenhang:

$$\rho^{\alpha} = n^{\alpha}\,\rho^{\alpha R}\,. \tag{1.19}$$

Hieraus ist ersichtlich, daß Änderungen der Partialdichte sowohl durch Änderung des Volumenanteils als auch durch Änderung der realistischen Dichte stattfinden können.

1.2 Kinematische Größen

Nach der Einführung der Bewegungsfunktionen ist es nun möglich, kinematische und physikalische Eigenschaften der materiellen Punkte des Mischungskörpers zu untersuchen, etwa Geschwindigkeiten oder Verzerrungen. Diese Eigenschaften können entweder auf die Referenz- oder die Momentankonfiguration des Mischungskörpers bezogen werden. Die

Größe	Notation	Beispiel
Skalare	Klein- und Großbuchstaben	β, f, Γ, G
Vektoren	fette, gerade Kleinbuchstaben	$\mathbf{v}, \mathbf{w}$
Ortsvektor Referenzkonfiguration	fettes, gerades X	$\mathbf{X}$
Ortsvektor Momentankonfiguration	fettes, gerades x	$\mathbf{x}$
Tensoren 2. Stufe	fette, gerade Großbuchstaben	$\mathbf{S}, \mathbf{T}$
Tensoren n. Stufe ($n > 2$)	fette, gerade Großbuchstaben	$\overset{n}{\mathbf{A}}$
Zweipunkttensoren	fette, römische Großbuchstaben	$\mathbf{F}, \mathbf{R}$
Operation	**Ergebnis**	**Beispiel**
Skalarprodukt von Vektoren	Skalar	$\mathbf{v} \cdot \mathbf{w}$
Dyadisches Produkt von Vektoren	Tensor 2. Stufe	$\mathbf{v} \otimes \mathbf{w}$
Skalarprodukt von Tensoren 2. Stufe	Skalar	$\mathbf{S} \cdot \mathbf{T}$
Produkt von Tensor 2. Stufe und Vektor	Vektor	$\mathbf{T}\,\mathbf{v}$
Produkt von Tensoren 2. Stufe	Tensor 2. Stufe	$\mathbf{S}\,\mathbf{T}$
Produkt von Tensor 4. und 2. Stufe	Tensor 2. Stufe	$\overset{4}{\mathbf{C}}\,\mathbf{E}$

Tabelle 1.1: *Konventionen der Vektor- und Tensornotation*

Größen werden hier in einer basisfreien Tensornotation dargestellt, vgl. *de Boer* [20], *Ehlers* [44]. In Tabelle 1.1 sind einige Konventionen der Notation zusammengefaßt.

Wird eine Größe in Koordinaten der Referenzkonfiguration parametrisiert (*materielle Darstellung*), so spricht man auch von einer *Lagrangeschen Beschreibung*, bei Parametrisierung in Koordinaten der Momentankonfiguration (*räumliche Darstellung*) von einer *Eulerschen Beschreibung*.

Bemerkung: In Anhang A wird die hier verwendete Notation einer differentialgeometrischen Formulierung gegenübergestellt. Dabei spielt der Begriff des Tangentialraums eine tragende Rolle, und Tensoren werden als multilineare Abbildungen aufgefaßt. Außerdem wird dort auf die Basisdarstellung vektorieller und tensorieller Größen in beliebigen, krummlinigen Koordinatensystemen eingegangen. □

1.2.1 Geschwindigkeiten und Beschleunigungen

Da jede Phase eines Mehrphasenkontinuums ihrer eigenen Bewegung folgt, besitzt jede Phase auch eine eigene Geschwindigkeit und Beschleunigung. Zusätzlich wird noch die Geschwindigkeit der Mischung eingeführt.

[3] In der Literatur werden z. T. auch „ungesättigte“ poröse Medien behandelt. In diesem Fall sind die Porenräume teilweise materiefrei: $\sum_\alpha n^\alpha < 1$.

Def. 1.10: Die *materielle Geschwindigkeit* $\mathbf{v}_{0\alpha}$ der Phase φ^α ist die Zeitableitung[4] der Bewegung von φ^α:

$$\mathbf{v}_{0\alpha}(\mathbf{X}_\alpha, t) = \frac{\mathrm{d}\overset{\alpha}{\boldsymbol{\chi}}_{\mathbf{X}_\alpha}(t)}{\mathrm{d}t} = \frac{\mathrm{d}\overset{\alpha}{\boldsymbol{\chi}}(\mathbf{X}_\alpha, t)}{\mathrm{d}t} = \frac{\partial\overset{\alpha}{\boldsymbol{\chi}}(\mathbf{X}_\alpha, t)}{\partial t}\,. \tag{1.20}$$

Die Verkettung mit der inversen Bewegung führt auf $\mathbf{v}_\alpha = \mathbf{v}_{0\alpha} \circ \overset{\alpha}{\boldsymbol{\chi}}{}_t^{-1}$, die *Euler*sche Darstellung der Geschwindigkeit (*räumliche Geschwindigkeit*):

$$\mathbf{v}_\alpha(\mathbf{x}, t) = \mathbf{v}_{0\alpha}(\overset{\alpha}{\boldsymbol{\chi}}{}_t^{-1}(\mathbf{x}), t) =: \frac{\mathrm{d}_\alpha\mathbf{x}}{\mathrm{d}t}(\mathbf{x}, t)\,. \tag{1.21}$$

Hierin kennzeichnet die Symbolik „$\frac{\mathrm{d}_\alpha(\ldots)}{\mathrm{d}t}$“ die Zeitableitung, die der Bewegung der Phase φ^α folgt. Als Abkürzung für diese Zeitableitung dient außerdem das Symbol „$(\ldots)'_\alpha$“, so daß man für die Geschwindigkeit auch kurz schreibt:

$$\mathbf{v}_\alpha = \frac{\mathrm{d}_\alpha\mathbf{x}}{\mathrm{d}t} = \mathbf{x}'_\alpha\,. \tag{1.22}$$

Die *Mischungsgeschwindigkeit* (auch *Massenmittelpunktsgeschwindigkeit* bzw. *baryzentrische Geschwindigkeit*) ist das dichtegewichtete Mittel der Geschwindigkeiten der einzelnen Phasen:

$$\mathbf{v} = \dot{\mathbf{x}} = \frac{1}{\rho}\sum_\alpha \rho^\alpha \mathbf{x}'_\alpha\,. \tag{1.23}$$

Die *Diffusionsgeschwindigkeit*

$$\mathbf{d}_\alpha = \mathbf{x}'_\alpha - \dot{\mathbf{x}}\,, \tag{1.24}$$

ist die Relativgeschwindigkeit der Phase φ^α gegenüber der Mischung φ. ■

Das Verschwinden der Summe der *Diffusionsmassenströme*,

$$\sum_\alpha \rho^\alpha \mathbf{d}_\alpha \;=\; \mathbf{0}\,, \tag{1.25}$$

ist eine direkte Folgerung aus (1.18), (1.23) und (1.24).

Bemerkung: Die oben eingeführte symbolische Schreibweise $(\ldots)'_\alpha = \frac{\mathrm{d}_\alpha(\ldots)}{\mathrm{d}t}$ für die Zeitableitung, die der Bewegung der Phase φ^α folgt, ist in allen Fällen die totale Ableitung nach der Zeit t. Im Spezialfall der Zeitableitung von Größen in materieller Darstellung – wie etwa der Bewegung selbst – fallen totale und partielle Ableitung zusammen, da die Referenzlage $\mathbf{X}_\alpha$ nicht von der Zeit t abhängt. □

Def. 1.11: Die *materielle Beschleunigung* $\mathbf{a}_{0\alpha}$ der Phase φ^α ist die zweite Zeitableitung der Bewegung von φ^α:

$$\mathbf{a}_{0\alpha}(\mathbf{X}_\alpha, t) = \frac{\mathrm{d}^2\overset{\alpha}{\boldsymbol{\chi}}_{\mathbf{X}_\alpha}(t)}{\mathrm{d}t^2} = \frac{\mathrm{d}^2\overset{\alpha}{\boldsymbol{\chi}}(\mathbf{X}_\alpha, t)}{\mathrm{d}t^2} = \frac{\partial^2\overset{\alpha}{\boldsymbol{\chi}}(\mathbf{X}_\alpha, t)}{\partial t^2}\,. \tag{1.26}$$

[4] Da es sich bei der Bewegung um eine mit der Zeit t parametrisierte Schar von Konfigurationen handelt, ist die Zeitableitung definiert als Ableitung der Schar nach dem Scharparameter t. Die Schreibweise $\mathrm{d}\overset{\alpha}{\boldsymbol{\chi}}_{\mathbf{X}_\alpha}(t)/\mathrm{d}t$ bedeutet, daß die Bahnlinie eines festen Punktes $\mathbf{X}_\alpha$ der Referenzkonfiguration nach der Zeit abgeleitet wird; man erhält also die (Momentan-)Geschwindigkeit als Tangente an die Bahnlinie von $\mathbf{X}_\alpha$.

Die Verkettung mit der inversen Bewegung führt auf $\mathbf{a}_\alpha = \mathbf{a}_{0\alpha} \circ \overset{\alpha}{\boldsymbol{\chi}}{}_t^{-1}$, die *Euler*sche Darstellung der Beschleunigung (*räumliche Beschleunigung*):

$$\mathbf{a}_\alpha(\mathbf{x},t) = \mathbf{a}_{0\alpha}(\overset{\alpha}{\boldsymbol{\chi}}{}_t^{-1}(\mathbf{x}),t) = \frac{\mathrm{d}_\alpha^2\mathbf{x}}{\mathrm{d}t^2}(\mathbf{x},t). \tag{1.27}$$

Man schreibt auch kurz für die Beschleunigung:

$$\mathbf{a}_\alpha = \frac{\mathrm{d}_\alpha^2\mathbf{x}}{\mathrm{d}t^2} = \mathbf{x}''_\alpha\,. \tag{1.28}$$

■

Mit Hilfe der Kettenregel kann man die räumliche Beschleunigung $\mathbf{a}_\alpha$ ganz ohne Kenntnis der Referenzkonfiguration aus der räumlichen Geschwindigkeit $\mathbf{v}_\alpha$ berechnen[5]:

$$\mathbf{a}_\alpha(\mathbf{x},t) = \frac{\mathrm{d}_\alpha\mathbf{v}_\alpha(\mathbf{x},t)}{\mathrm{d}t} = \frac{\partial\mathbf{v}_\alpha}{\partial t} + (\operatorname{grad}\mathbf{v}_\alpha)\,\mathbf{v}_\alpha\,. \tag{1.29}$$

Darin wird der erste Ausdruck als *lokaler Anteil* und der zweite als *konvektiver Anteil* der Beschleunigung bezeichnet, der Operator „grad“ ist die partielle Ableitung nach $\mathbf{x}$. Analoge Ausdrücke erhält man für beliebige Skalar- und Vektorfelder, die in Koordinaten der Momentankonfiguration parametrisiert sind, und es wird definiert:

Def. 1.12: Für ein Skalarfeld $\Gamma(\mathbf{x},t)$ und ein Vektorfeld $\boldsymbol{\Gamma}(\mathbf{x},t)$ heißen die Ausdrücke

$$\Gamma'_\alpha = \frac{\partial\Gamma}{\partial t} + \operatorname{grad}\Gamma\cdot\mathbf{x}'_\alpha, \qquad \boldsymbol{\Gamma}'_\alpha = \frac{\partial\boldsymbol{\Gamma}}{\partial t} + (\operatorname{grad}\boldsymbol{\Gamma})\,\mathbf{x}'_\alpha \tag{1.30}$$

materielle Zeitableitung bzgl. der Bewegung der Phase φ^α. ■

Der in (1.29) auftretende Gradient der räumlichen Geschwindigkeit wird bei der Formulierung der Bilanzgleichungen ebenfalls benötigt. Es wird definiert:

Def. 1.13: Der *räumliche Geschwindigkeitsgradient* $\mathbf{L}_\alpha$ ist das Tensorfeld

$$\mathbf{L}_\alpha = \operatorname{grad}\mathbf{x}'_\alpha = \operatorname{grad}\mathbf{v}_\alpha \tag{1.31}$$

der Momentankonfiguration. Seine symmetrischen und schiefsymmetrischen Anteile,

$$\mathbf{D}_\alpha = \frac{1}{2}(\mathbf{L}_\alpha + \mathbf{L}_\alpha^T) \qquad \text{und} \qquad \mathbf{W}_\alpha = \frac{1}{2}(\mathbf{L}_\alpha - \mathbf{L}_\alpha^T)\,, \tag{1.32}$$

heißen *Deformationsgeschwindigkeitstensor* und *Drehgeschwindigkeitstensor* (auch *Wirbeltensor* bzw. *Spintensor*). ■

Der symmetrische Anteil $\mathbf{D}_\alpha$ ist eine wichtige Größe bei der Formulierung von Materialgesetzen für viskose Materialien, insbesondere für Fluide. Außerdem spielt er – allerdings auf der *plastischen Zwischenkonfiguration* – eine wichtige Rolle bei der Formulierung von finiten Plastizitätsgesetzen.

[5] Dies ist u. a. wichtig zur Modellierung von Fluiden, da letztere keine ausgezeichnete Referenzkonfiguration besitzen und daher i. a. mit der räumlichen Geschwindigkeit beschrieben werden.

1.2.2 Deformationsgradient

Das wichtigste Deformationsmaß in der Kontinuumsmechanik ist der Deformationsgradient. Er ist das Differential an die Bewegungsfunktion und damit in natürlicher Weise ein Zweipunkttensorfeld. Zunächst wird für eine beliebige Bewegung $\boldsymbol{\chi}$ der Begriff des Differentials eingeführt. Alle Betrachtungen beziehen sich auf einen festen Zeitpunkt t, der Index t entfällt jedoch im folgenden aus Gründen der Übersichtlichkeit.

Def. 1.14: Das *Differential* $\mathrm{D}\boldsymbol{\chi}$ einer Bewegung $\boldsymbol{\chi} : \Omega_0 \longrightarrow \Omega_t$ ist ein Zweipunkttensorfeld $\mathrm{D}\boldsymbol{\chi}$, das Vektoren der Referenzkonfiguration auf Vektoren der Momentankonfiguration abbildet. Für das Differential an einem festen Punkt $\mathbf{X} \in \Omega_0$ schreibt man gleichwertig:

$$\mathrm{D}\boldsymbol{\chi}(\mathbf{X}) = \mathrm{d}_{\mathbf{X}}\boldsymbol{\chi} = \frac{\mathrm{d}\boldsymbol{\chi}(\mathbf{X})}{\mathrm{d}\mathbf{X}}\,. \tag{1.33}$$

Die Anwendung des Differentials an einem Punkt $\mathbf{X}$ ordnet einem Vektor $\mathbf{w}_0$ der Referenzkonfiguration mittels

$$\mathbf{w}_0 \mapsto \mathbf{w} := \mathrm{D}\boldsymbol{\chi}(\mathbf{X})(\mathbf{w}_0) = \mathrm{d}_{\mathbf{X}}\boldsymbol{\chi}(\mathbf{w}_0) = \frac{\mathrm{d}\boldsymbol{\chi}(\mathbf{X})}{\mathrm{d}\mathbf{X}}\,\mathbf{w}_0 \tag{1.34}$$

den Vektor $\mathbf{w}$ der Momentankonfiguration zu, was gerade der Richtungsableitung der Bewegung $\boldsymbol{\chi}$ in Richtung des Vektors $\mathbf{w}_0$ entspricht. Man nennt dies auch die *Vorwärtstransformation* (engl.: *push-forward*) des Vektors $\mathbf{w}_0$. Umgekehrt bezeichnet man die Abbildung

$$\mathbf{w} \mapsto \mathbf{w}_0 = \mathrm{D}\boldsymbol{\chi}^{-1}(\mathbf{x})(\mathbf{w})$$

als die *Rückwärtstransformation* (engl: *pull-back*) des Vektors $\mathbf{w}$. ■

Bemerkung: Das Differential $\mathrm{d}_{\mathbf{X}}\boldsymbol{\chi}$ in (1.33) wird auch als *Frechet-Differential* im Punkt $\mathbf{X}$ bezeichnet, die Richtungsableitung $\mathrm{d}_{\mathbf{X}}\boldsymbol{\chi}(\mathbf{w}_0)$ in (1.34) als *Gateaux-Ableitung*. □

Da jede Phase eines Mehrphasenkontinuums ihrer eigenen Bewegung $\overset{\alpha}{\boldsymbol{\chi}}$ folgt, besitzt jede Phase auch ein eigenes Differential. Man definiert:

Def. 1.15: Der *Deformationsgradient* der Phase φ^α ist das Differential der zugehörigen Bewegung, also das Zweipunkttensorfeld $\mathbf{F}_\alpha = \mathrm{D}\overset{\alpha}{\boldsymbol{\chi}}_t$ mit:

$$\mathbf{F}_\alpha(\mathbf{X}_\alpha, t) = \mathrm{d}_{\mathbf{X}_\alpha}\overset{\alpha}{\boldsymbol{\chi}}_t = \frac{\mathrm{d}\overset{\alpha}{\boldsymbol{\chi}}_t(\mathbf{X}_\alpha)}{\mathrm{d}\mathbf{X}_\alpha} = \frac{\partial\overset{\alpha}{\boldsymbol{\chi}}(\mathbf{X}_\alpha, t)}{\partial\mathbf{X}_\alpha}\,. \tag{1.35}$$

Als Abkürzung für die partielle Ableitung nach der Referenzlage $\mathbf{X}_\alpha$ wird das Symbol „Grad_α“ eingeführt. Man schreibt daher mit $\mathbf{x} = \overset{\alpha}{\boldsymbol{\chi}}_t(\mathbf{X}_\alpha)$ auch kurz für den Deformationsgradienten in einem Punkt $\mathbf{X}_\alpha \in \Omega_0$:

$$\mathbf{F}_\alpha = \frac{\mathrm{d}\overset{\alpha}{\boldsymbol{\chi}}_t(\mathbf{X}_\alpha)}{\mathrm{d}\mathbf{X}_\alpha} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{X}_\alpha} = \mathrm{Grad}_\alpha\,\mathbf{x}\,. \tag{1.36}$$

■

Bemerkung: Der Deformationsgradient „transportiert“ Vektoren $\mathrm{d}\mathbf{X}_\alpha$ (*Linienelemente*) der Referenzkonfiguration auf Vektoren $\mathrm{d}\mathbf{x}$ der Momentankonfiguration:

$$\mathrm{d}\mathbf{x} = \mathbf{F}_\alpha \, \mathrm{d}\mathbf{X}_\alpha \, .$$

Man spricht daher auch vom *kovarianten Vorwärtstransport* von Linienelementen. □

Def. 1.16: Die Determinante des Deformationsgradienten wird mit

$$J_\alpha = \det \mathbf{F}_\alpha \tag{1.37}$$

bezeichnet. Die Determinante ist strikt positiv,

$$J_\alpha > 0 \, ,$$

da $J_\alpha(\mathbf{X}_\alpha, t_0) = 1$ ist und die Bewegung $\overset{\alpha}{\boldsymbol{\chi}}_t$ als eineindeutig angenommen wurde. ■

1.2.3 Transport von Volumenelementen

Im Gegensatz zu Linienelementen werden Volumenelemente nicht direkt mit dem Deformationsgradienten, sondern mit seiner Determinante transportiert. Zur Herleitung dieses Zusammenhangs wird eine Transformationsregel für Integrale von Funktionen mehrerer Veränderlicher benötigt. Alle Größen werden dazu im festen, aber möglicherweise nicht orthogonalen und nicht normierten Koordinatensystem $\{0, \mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$ dargestellt.

Def. 1.17: Für eine reellwertige Funktion $f : \mathbb{R}^3 \longrightarrow \mathbb{R}$ und eine reguläre Transformation $\mathbb{R}^3 \ni \mathbf{z} = \sum_{i=1}^{3} \mathrm{z}^i \mathbf{g}_i \mapsto \mathbf{x} = \sum_{i=1}^{3} \mathrm{x}^i \mathbf{g}_i = \mathbf{h}(\mathbf{z}) \in \mathbb{R}^3$ gilt mit einer Menge $U \subset \mathbb{R}^3$ die *Substitutionsregel* für bestimmte Integrale:

$$\int_U f(\mathbf{z}) \, \mathrm{dz} = \int_{\mathbf{h}(U)} f(\mathbf{h}^{-1}(\mathbf{x})) \left| \det \frac{\mathrm{d}\mathbf{z}}{\mathrm{d}\mathbf{x}} \right| \mathrm{dx} \, ,$$

worin $\mathrm{dz} = \mathrm{dz}^1\mathrm{dz}^2\mathrm{dz}^3$ und $\mathrm{dx} = \mathrm{dx}^1\mathrm{dx}^2\mathrm{dx}^3$ gesetzt wurden und $\det \frac{\mathrm{d}\mathbf{z}}{\mathrm{d}\mathbf{x}} = \frac{\partial(\mathrm{z}^1,\mathrm{z}^2,\mathrm{z}^3)}{\partial(\mathrm{x}^1,\mathrm{x}^2,\mathrm{x}^3)}$ die Funktionaldeterminante der inversen Transformation $\mathbf{h}^{-1}$ bezeichnet. ■

Diese Regel kann direkt zur Herleitung des Transformationsverhaltens von Volumenelementen verwendet werden. Der Wert der Funktion f wird dann als das (konstante) Volumen des von der Basis aufgespannten Quaders gewählt (Spatprodukt),

$$f(\mathbf{X}_\alpha) := (\mathbf{g}_1 \times \mathbf{g}_2) \cdot \mathbf{g}_3 \, ,$$

die Transformation $\mathbf{h}$ ist die Bewegung $\overset{\alpha}{\boldsymbol{\chi}}_t$, und die Vektoren $\mathbf{z}$ und $\mathbf{x}$ werden mit $\mathbf{X}_\alpha$ und $\mathbf{x}$ identifiziert. Man erhält für eine beliebige zusammenhängende Teilmenge $U_0 \subset \Omega_0$ der Referenzkonfiguration:

$$\int_{U_0} f(\mathbf{X}_\alpha) \, \mathrm{dX}_\alpha = \int_{\overset{\alpha}{\boldsymbol{\chi}}_t(U_0)} f(\overset{\alpha}{\boldsymbol{\chi}}_t^{-1}(\mathbf{x})) \underbrace{\left| \det \frac{\mathrm{d}\mathbf{X}_\alpha}{\mathrm{d}\mathbf{x}} \right|}_{= J_\alpha^{-1}} \mathrm{dx} = \int_{\overset{\alpha}{\boldsymbol{\chi}}_t(U_0)} J_\alpha^{-1} \, f(\overset{\alpha}{\boldsymbol{\chi}}_t^{-1}(\mathbf{x})) \, \mathrm{dx} \, . \tag{1.38}$$

Aufgrund der Identitäten

$$\begin{aligned} \mathrm{d}V_\alpha &= (\mathrm{d}\mathbf{X}_{\alpha 1} \times \mathrm{d}\mathbf{X}_{\alpha 2}) \cdot \mathrm{d}\mathbf{X}_{\alpha 3} = (\mathrm{d}X^1_\alpha \mathbf{g}_1 \times \mathrm{d}X^2_\alpha \mathbf{g}_2) \cdot \mathrm{d}X^3_\alpha \mathbf{g}_3 &&= f(\mathbf{X}_\alpha)\,\mathrm{d}X_\alpha\,, \\ \mathrm{d}v &= (\mathrm{d}\mathbf{x}_1 \times \mathrm{d}\mathbf{x}_2) \cdot \mathrm{d}\mathbf{x}_3 = (\mathrm{d}x^1 \mathbf{g}_1 \times \mathrm{d}x^2 \mathbf{g}_2) \cdot \mathrm{d}x^3 \mathbf{g}_3 &&= f(\overset{\alpha}{\boldsymbol{\chi}}{}^{-1}_t(\mathbf{x}))\,\mathrm{d}x \end{aligned}$$

erhält man aus (1.38) durch Grenzübergang lokal den Zusammenhang

$$\mathrm{d}V_\alpha = J_\alpha^{-1} \mathrm{d}v \quad \Longrightarrow \quad \mathrm{d}v = J_\alpha\,\mathrm{d}V_\alpha\,, \tag{1.39}$$

der den *Transport von Volumenelementen* beschreibt.

1.2.4 Deformations- und Verzerrungsmaße

Nach der Einführung des Deformationsgradienten können nun Deformations- und Verzerrungsmaße definiert werden. Anschaulich gibt ein Deformationsmaß Auskunft darüber, wie sich der Körper während der Bewegung lokal deformiert, d. h. das Deformationsmaß im Ausgangszustand ist die Identität. Ein Verzerrungsmaß vergleicht den deformierten Zustand mit dem Ausgangszustand, d. h. sein Wert im Ausgangszustand ist der Nulltensor.

Zum Verständnis der Deformation erweist sich die polare Zerlegung als ein nützliches Hilfsmittel. Bekanntlich kann man einen beliebigen invertierbaren Tensor eindeutig multiplikativ in einen orthogonalen und einen symmetrisch positiv definiten Tensor zerlegen.

Def. 1.18: Die *polare Zerlegung* des Deformationsgradienten $\mathbf{F}_\alpha$ ist gegeben durch

$$\mathbf{F}_\alpha \;=\; \mathbf{R}_\alpha\,\mathbf{U}_\alpha \;=\; \mathbf{V}_\alpha\,\mathbf{R}_\alpha\,. \tag{1.40}$$

Darin ist $\mathbf{R}_\alpha$ ein orthogonaler Zweipunkttensor (Drehung des Körpers aus der Referenz- in die Momentankonfiguration), $\mathbf{U}_\alpha$ ist der symmetrisch positiv definite *Rechtsstrecktensor* der Referenzkonfiguration und $\mathbf{V}_\alpha$ der symmetrisch positiv definite *Linksstrecktensor* der Momentankonfiguration. ■

Die Betrachtung der Längenänderung von Linienelementen während der Deformation des Körpers,

$$\|\mathrm{d}\mathbf{x}\|^2 \;=\; \mathrm{d}\mathbf{x}\cdot\mathrm{d}\mathbf{x} \;=\; (\mathbf{F}_\alpha\,\mathrm{d}\mathbf{X}_\alpha)\cdot(\mathbf{F}_\alpha\,\mathrm{d}\mathbf{X}_\alpha) \;=\; \mathrm{d}\mathbf{X}_\alpha\cdot(\mathbf{F}_\alpha^T\,\mathbf{F}_\alpha)\,\mathrm{d}\mathbf{X}_\alpha\,, \tag{1.41}$$

führt auf eine quadratische Form mit einem symmetrisch positiv definiten Tensor $\mathbf{F}_\alpha^T\,\mathbf{F}_\alpha$. Umgekehrt führt die Betrachtung aus dem Blickwinkel der Momentankonfiguration[6],

$$\|\mathrm{d}\mathbf{X}_\alpha\|^2 \;=\; \mathrm{d}\mathbf{X}_\alpha\cdot\mathrm{d}\mathbf{X}_\alpha \;=\; (\mathbf{F}_\alpha^{-1}\,\mathrm{d}\mathbf{x})\cdot(\mathbf{F}_\alpha^{-1}\,\mathrm{d}\mathbf{x}) \;=\; \mathrm{d}\mathbf{x}\cdot(\mathbf{F}_\alpha^{T-1}\,\mathbf{F}_\alpha^{-1})\,\mathrm{d}\mathbf{x}\,, \tag{1.42}$$

zu einer analogen quadratischen Form mit einem Tensor $\mathbf{F}_\alpha^{T-1}\,\mathbf{F}_\alpha^{-1} = (\mathbf{F}_\alpha\,\mathbf{F}_\alpha^T)^{-1}$. Dies führt zur folgenden Definition:

[6] Die inverse Transponierte eines Tensors $\mathbf{A}$ wird hier übereinstimmend mit *Ehlers* [44] als $\mathbf{A}^{T-1}$ bezeichnet. In der mathematischen Literatur (z. B. *Ciarlet* [32]) findet man dafür auch häufig die Schreibweise $\mathbf{A}^{-T} = (\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$.

Def. 1.19: Der *rechte Cauchy-Green-Deformationstensor*

$$\mathbf{C}_\alpha \;=\; \mathbf{F}_\alpha^T\,\mathbf{F}_\alpha \tag{1.43}$$

ist ein Deformationsmaß der Referenzkonfiguration.
Der *linke Cauchy-Green-Deformationstensor* (auch *Finger-Tensor*)

$$\mathbf{B}_\alpha \;=\; \mathbf{F}_\alpha\,\mathbf{F}_\alpha^T \tag{1.44}$$

ist ein Deformationsmaß der Momentankonfiguration. ■

Zwischen den Deformationsmaßen bestehen die Zusammenhänge

$$\begin{aligned} \mathbf{C}_\alpha &= \mathbf{U}_\alpha^2\,, & \mathbf{U}_\alpha &= \sqrt{\mathbf{C}_\alpha}\,, \\ \mathbf{B}_\alpha &= \mathbf{V}_\alpha^2\,, & \mathbf{V}_\alpha &= \sqrt{\mathbf{B}_\alpha}\,, \end{aligned} \tag{1.45}$$

wobei die Wurzel über die spektrale Zerlegung positiv definiter Tensoren erklärt ist.

Verzerrungsmaße vergleichen den deformierten mit dem undeformierten Zustand des Körpers. Die Herleitung der beiden wichtigsten Verzerrungsmaße erfolgt durch Betrachtung der Differenz der Quadrate der Linienelemente auf der Referenz- und Momentankonfiguration. Mit (1.41) und (1.42) ergibt sich

$$\begin{aligned} \|\mathrm{d}\mathbf{x}\|^2 - \|\mathrm{d}\mathbf{X}_\alpha\|^2 &= \mathrm{d}\mathbf{X}_\alpha\cdot\mathbf{C}_\alpha\,\mathrm{d}\mathbf{X}_\alpha - \mathrm{d}\mathbf{X}_\alpha\cdot\mathrm{d}\mathbf{X}_\alpha &&= \mathrm{d}\mathbf{X}_\alpha\cdot(\mathbf{C}_\alpha-\mathbf{I})\,\mathrm{d}\mathbf{X}_\alpha\,, \\ \|\mathrm{d}\mathbf{x}\|^2 - \|\mathrm{d}\mathbf{X}_\alpha\|^2 &= \mathrm{d}\mathbf{x}\cdot\mathrm{d}\mathbf{x} - \mathrm{d}\mathbf{x}\cdot\mathbf{B}_\alpha^{-1}\,\mathrm{d}\mathbf{x} &&= \mathrm{d}\mathbf{x}\cdot(\mathbf{I}-\mathbf{B}_\alpha^{-1})\,\mathrm{d}\mathbf{x}\,, \end{aligned}$$

so daß die folgende Definition naheliegt:

Def. 1.20: Der *Greensche Verzerrungstensor*

$$\mathbf{E}_\alpha \;=\; \frac{1}{2}(\mathbf{C}_\alpha - \mathbf{I})\,, \tag{1.46}$$

ist ein Verzerrungsmaß der Referenzkonfiguration, wobei der Vorfaktor 1/2 historische Gründe in der Interpretation der Koeffizienten des Tensors hat (Ingenieur-Verzerrungen).
Der *Almansische Verzerrungstensor*

$$\mathbf{A}_\alpha \;=\; \frac{1}{2}(\mathbf{I} - \mathbf{B}_\alpha^{-1}) \tag{1.47}$$

ist ein Verzerrungsmaß der Momentankonfiguration. Es besteht der Zusammenhang

$$\mathbf{E}_\alpha = \mathbf{F}_\alpha^T\,\mathbf{A}_\alpha\,\mathbf{F}_\alpha\,, \tag{1.48}$$

d. h. man erhält den *Green*schen Verzerrungstensor durch Rückwärtstransformation (pull-back) des *Almansi*schen Verzerrungstensors. ■

Bemerkung: Im Sinne der in Anhang A vorgestellten differentialgeometrischen Notation kann der *Green*sche Verzerrungstensor auch als Vergleich des pull-backs $\mathbf{C}_\alpha = \mathbf{F}_\alpha^T\,\mathbf{I}\,\mathbf{F}_\alpha$

des Metriktensors der Momentankonfiguration mit dem Metriktensor der Referenzkonfiguration verstanden werden. Entsprechend vergleicht der *Almansi*sche Verzerrungstensor den push-forward $\mathbf{B}_\alpha^{-1} = \mathbf{F}_\alpha^{T-1}\,\mathbf{I}\,\mathbf{F}_\alpha^{-1}$ des Metriktensors der Referenzkonfiguration mit dem Metriktensor der Momentankonfiguration. □

Neben dem *Green*schen und dem *Almansi*schen Verzerrungstensor werden in der Materialtheorie häufig weitere Verzerrungstensoren verwendet – etwa die *Karni-Reiner*schen Verzerrungstensoren. Diese können der Literatur entnommen werden, z. B. *Ehlers* [44].

1.2.5 Geometrisch lineare Theorie

In den Anwendungen hat man es häufig mit Problemen zu tun, bei denen von kleinen Verzerrungen ausgegangen werden kann. In diesem Fall kann man die kinematischen Größen linearisieren und erhält eine geometrisch lineare Theorie. Dies hat den Vorteil, daß sowohl die Formulierung der Modellgleichungen als auch die numerische Implementierung vereinfacht werden.

Mittels *Taylor*-Entwicklung in mehreren Raumdimensionen kann man eine beliebige Feldfunktion $f : \Omega_t \longrightarrow \mathbb{R}$ bei festgehaltener Zeit t wie folgt linearisieren:

$$f(\mathbf{x}) \;=\; \underbrace{f(\bar{\mathbf{x}}) + \mathrm{D}f(\bar{\mathbf{x}}) \cdot \Delta\mathbf{x}}_{=:\,\mathrm{lin}(f)} + \mathcal{O}(|\Delta\mathbf{x}|^2)\,.$$

Darin bezeichnet $\bar{\mathbf{x}}$ den Entwicklungspunkt und $\Delta\mathbf{x} = \mathbf{x} - \bar{\mathbf{x}}$ die Richtung, in der die Funktion f linearisiert wird. Außerdem ist $\mathrm{D}f(\bar{\mathbf{x}}) = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}(\bar{\mathbf{x}})$ das Differential von f bzgl. der Ortsvariablen $\mathbf{x}$ (*Frechet*-Differential, vgl. (1.33)) und $\mathrm{D}f(\bar{\mathbf{x}}) \cdot \Delta\mathbf{x} = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\,\{f(\bar{\mathbf{x}} + \epsilon\,\Delta\mathbf{x}\}|_{\epsilon=0}$ die Richtungsableitung in Richtung $\Delta\mathbf{x}$ (*Gateaux*-Differential, vgl. (1.34)), siehe *Marsden & Hughes* [83, S. 183 ff.].

Es werden zunächst einige kinematische Größen in Richtung eines Deformationsinkrements $\Delta\mathbf{u}_\alpha = \mathbf{u}_\alpha - \bar{\mathbf{u}}_\alpha$ linearisiert (*Eipper* [53]):

$$\begin{aligned}
\mathrm{D}\mathbf{F}_\alpha \cdot \Delta\mathbf{u}_\alpha &= \mathrm{Grad}_\alpha\,\Delta\mathbf{u}_\alpha\,,\\
\mathrm{D}J_\alpha \cdot \Delta\mathbf{u}_\alpha &= J_\alpha\,\mathrm{div}\,\Delta\mathbf{u}_\alpha\,,\\
\mathrm{D}\mathbf{E}_\alpha \cdot \Delta\mathbf{u}_\alpha &= \frac{1}{2}(\mathbf{F}_\alpha^T\;\mathrm{Grad}_\alpha\,\Delta\mathbf{u}_\alpha + \mathrm{Grad}_\alpha^T\,\Delta\mathbf{u}_\alpha\,\mathbf{F}_\alpha)\,.
\end{aligned}$$

Linearisiert man um den natürlichen, verzerrungsfreien Zustand mit $\mathrm{Grad}_\alpha\,\bar{\mathbf{u}}_\alpha = \mathbf{0}$ und $\mathrm{Grad}_\alpha\,\Delta\mathbf{u}_\alpha = \mathrm{Grad}_\alpha\,\mathbf{u}_\alpha$, so erhält man die Ausdrücke der geometrisch linearen Theorie:

$$\begin{aligned}
\mathrm{lin}(\mathbf{F}_\alpha) &= \mathbf{I} \;+\; \mathrm{Grad}_\alpha\,\mathbf{u}_\alpha\,,\\
\mathrm{lin}(J_\alpha) &= 1 \;+\; \mathrm{Div}_\alpha\,\mathbf{u}_\alpha\,,\\
\mathrm{lin}(\mathbf{E}_\alpha) &= \frac{1}{2}(\mathrm{Grad}_\alpha\,\mathbf{u}_\alpha + \mathrm{Grad}_\alpha^T\,\mathbf{u}_\alpha)\,.
\end{aligned}$$

Außerdem gestattet die Voraussetzung kleiner Verzerrungen die Annahme

$$\mathrm{Grad}_\alpha\,\mathbf{u}_\alpha \;\approx\; \mathbf{0} \quad\Longrightarrow\quad \mathbf{F}_\alpha \;\approx\; \mathbf{I} \quad\Longrightarrow\quad J_\alpha \;\approx\; 1\,,$$

so daß keine Unterscheidung mehr zwischen Referenz- und Momentankonfiguration erforderlich ist. Für Integrale und Differentialoperatoren kann dann jeweils die Schreibweise der Momentankonfiguration verwendet werden:

$$\int_{\Omega_0}(\dots)\,\mathrm{d}V_\alpha \;\approx\; \int_{\Omega_t}(\dots)\,\mathrm{d}v\,, \quad \mathrm{Grad}_\alpha(\dots) \;\approx\; \mathrm{grad}(\dots)\,, \quad \mathrm{Div}_\alpha(\dots) \;\approx\; \mathrm{div}(\dots)\,. \tag{1.49}$$

Der Verzerrungstensor wird in der geometrisch linearen Theorie mit

$$\boldsymbol{\varepsilon}_\alpha \;=\; \frac{1}{2}(\mathrm{Grad}_\alpha\,\mathbf{u}_\alpha + \mathrm{Grad}_\alpha^T\,\mathbf{u}_\alpha) \tag{1.50}$$

bezeichnet.

1.3 Mechanische Bilanzgleichungen

In der Kontinuumsmechanik werden axiomatisch verschiedene *Erhaltungsgleichungen* eingeführt, die sich auf die Erfahrung stützen, daß gewisse Größen in einem abgeschlossenen physikalischen System weder produziert werden noch verloren gehen. Berücksichtigt man zusätzlich Einflüsse der Umgebung auf das betrachtete System, so spricht man von *Bilanzgleichungen*, wobei die Interaktion mit der Umgebung mittels Fluß- und Zufuhrtermen einbezogen wird (äußere Nah- und Fernwirkung). Im einzelnen wird die Erhaltung der mechanischen Größen Masse, Impuls (Bewegungsgröße) und Drall (Drehimpuls) axiomatisch eingeführt. Weiter benennt der erste Hauptsatz der Thermodynamik die Energie als Erhaltungsgröße (Energieerhaltungssatz). Der zweite Hauptsatz der Thermodynamik (Entropieungleichung) besagt, daß die Entropieproduktion nie negativ sein kann, d. h. daß die Entropie in einem abgeschlossenen physikalischen System niemals abnehmen, sondern höchstens zunehmen kann.

In diesem Abschnitt werden zunächst die allgemeinen Formen der Bilanzgleichungen vorgestellt und anschließend für die mechanischen Größen Masse, Impuls und Drall formuliert. Die Bilanzgleichungen für die thermodynamischen Größen Energie und Entropie werden im darauffolgenden Abschnitt diskutiert.

1.3.1 Allgemeine Struktur der Bilanzgleichungen

Die Beschreibung von Mehrphasenkontinua basiert auf den von *Truesdell* [116, 117] formulierten *metaphysischen Prinzipien*:

- Alle Eigenschaften der Mischung müssen als mathematische Folgerungen aus den Eigenschaften der einzelnen Konstituierenden ableitbar sein.
- Zur Beschreibung der Bewegung einer Konstituierenden kann man diese in Gedanken vom Rest der Mischung trennen, sofern die Einwirkungen anderer Konstituierender auf die betrachtete korrekt berücksichtigt werden.

- Die Bewegung der Mischung als Ganzes wird mit denselben Gleichungen beschrieben wie die Bewegung eines einfachen Körpers.

Man stellt also die Bilanzgleichungen sowohl für jede Konstituierende φ^α als auch für die Mischung φ auf. In den Bilanzgleichungen der Phasen φ^α sind dabei jeweils Produktionsterme enthalten, die die Interaktion zwischen den Phasen beschreiben.

Es erweist sich bei der Aufstellung von Bilanzrelationen als nützlich, zunächst die allgemeine Struktur der Gleichungen für eine beliebige zu bilanzierende physikalische Größe zu formulieren, und in einem zweiten Schritt die einzelnen Terme der allgemeinen Bilanzrelation für spezielle Bilanzrelationen zu identifizieren. Die vorliegende Darstellung folgt dabei in weiten Teilen der Formulierung, wie sie von *Ehlers* [44, 47] angegeben wurde.

Gemäß den obigen Prinzipien kann die Struktur der Bilanzgleichungen der Mischung aus der klassischen Kontinuumsmechanik des einfachen Körpers übernommen werden. Dazu sei Ψ (bzw. $\boldsymbol{\Psi}$) die volumenspezifische skalarwertige (bzw. vektorwertige) Dichte der zu bilanzierenden physikalischen Größe (d. h. Massendichte, Impulsdichte, Dralldichte, spezifische innere Energie oder spezifische Entropie). Man erhält für eine beliebige zusammenhängende Teilmenge $U_t \subset \Omega_t$ der Momentankonfiguration die *integrale (*auch *globale) Form der allgemeinen Bilanzgleichung der Mischung*:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{U_t} \Psi \,\mathrm{d}v &= \int\limits_{\partial U_t} \boldsymbol{\phi}\cdot\mathbf{n}\,\mathrm{d}a + \int\limits_{U_t} \sigma\,\mathrm{d}v + \int\limits_{U_t} \hat{\Psi}\,\mathrm{d}v\,, \\ \frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{U_t} \boldsymbol{\Psi} \,\mathrm{d}v &= \int\limits_{\partial U_t} \boldsymbol{\Phi}\,\mathbf{n}\,\mathrm{d}a + \int\limits_{U_t} \boldsymbol{\sigma}\,\mathrm{d}v + \int\limits_{U_t} \hat{\boldsymbol{\Psi}}\,\mathrm{d}v\,. \end{aligned} \tag{1.51}$$

Darin ist $\mathbf{n}$ der nach außen orientierte Normaleneinheitsvektor auf der Oberfläche ∂U_t des Gebietes, $\boldsymbol{\phi}\cdot\mathbf{n}$ (bzw. $\boldsymbol{\Phi}\,\mathbf{n}$) ist der Ausfluß über die Oberfläche infolge äußerer Nahwirkung (z. B. die Spannung in der Impulsbilanz), und σ (bzw. $\boldsymbol{\sigma}$) ist die Zufuhr infolge äußerer Fernwirkung (z. B. die Schwerkraft in der Impulsbilanz). Die Produktion $\hat{\Psi}$ (bzw. $\hat{\boldsymbol{\Psi}}$) entfällt bei allen Bilanzgleichungen außer im Fall der Entropiebilanz.

Durch Differentiation des zeitabhängigen Integrals auf der linken Seite (*Reynolds*sches Transporttheorem) und Anwendung des *Gauß*schen Integralsatzes zur Transformation des Oberflächenintegrals auf der rechten Seite in ein Volumenintegral erhält man Gleichungen, die nur noch Volumenintegrale über dasselbe Gebiet $U_t \subset \Omega_t$ enthalten. Die Voraussetzung stetig differenzierbarer Integranden führt auf die *lokale Form der allgemeinen Bilanzgleichung der Mischung*:

$$\begin{aligned} \dot{\Psi} + \Psi\,\operatorname{div}\dot{\mathbf{x}} &= \operatorname{div}\boldsymbol{\phi} + \sigma + \hat{\Psi}\,, \\ \dot{\boldsymbol{\Psi}} + \boldsymbol{\Psi}\,\operatorname{div}\dot{\mathbf{x}} &= \operatorname{div}\boldsymbol{\Phi} + \boldsymbol{\sigma} + \hat{\boldsymbol{\Psi}}\,. \end{aligned} \tag{1.52}$$

Die Struktur der Bilanzgleichungen für die einzelnen Phasen φ^α der Mischung ergibt sich gemäß den *Truesdell*schen Prinzipien auf analoge Weise. Alle Größen werden über den hochgestellten Index $(...)^\alpha$ als zur Phase φ^α gehörig gekennzeichnet. Die Produktionsterme

$\hat{\Psi}^\alpha$ (bzw. $\hat{\mathbf{\Psi}}^\alpha$) beinhalten jetzt die Kopplung zwischen den einzelnen Konstituierenden, können also zur Beschreibung von Austauschvorgängen zwischen den Phasen verwendet werden (Massen-, Impuls-, Drall- oder Energieaustausch). Man erhält die *integrale Form der allgemeinen Bilanzgleichung der Phase* φ^α:

$$\begin{aligned} \frac{\mathrm{d}_\alpha}{\mathrm{d}t}\int\limits_{U_t}\Psi^\alpha\,\mathrm{d}v &= \int\limits_{\partial U_t}\boldsymbol{\phi}^\alpha\cdot\mathbf{n}\,\mathrm{d}a + \int\limits_{U_t}\sigma^\alpha\,\mathrm{d}v + \int\limits_{U_t}\hat{\Psi}^\alpha\,\mathrm{d}v\,, \\ \frac{\mathrm{d}_\alpha}{\mathrm{d}t}\int\limits_{U_t}\mathbf{\Psi}^\alpha\,\mathrm{d}v &= \int\limits_{\partial U_t}\mathbf{\Phi}^\alpha\,\mathbf{n}\,\mathrm{d}a + \int\limits_{U_t}\boldsymbol{\sigma}^\alpha\,\mathrm{d}v + \int\limits_{U_t}\hat{\mathbf{\Psi}}^\alpha\,\mathrm{d}v\,. \end{aligned} \tag{1.53}$$

Wie oben gelangt man zur *lokalen Form der allgemeinen Bilanzgleichung der Phase* φ^α:

$$\begin{aligned} (\Psi^\alpha)'_\alpha + \Psi^\alpha\,\mathrm{div}\,\mathbf{x}'_\alpha &= \mathrm{div}\,\boldsymbol{\phi}^\alpha + \sigma^\alpha + \hat{\Psi}^\alpha\,, \\ (\mathbf{\Psi}^\alpha)'_\alpha + \mathbf{\Psi}^\alpha\,\mathrm{div}\,\mathbf{x}'_\alpha &= \mathrm{div}\,\mathbf{\Phi}^\alpha + \boldsymbol{\sigma}^\alpha + \hat{\mathbf{\Psi}}^\alpha\,. \end{aligned} \tag{1.54}$$

Da sich nach *Truesdell* die Bilanzgleichung der Mischung als Summe der Bilanzgleichungen der einzelnen Phasen ergeben muß, sind die auftretenden Größen nicht unabhängig, sondern müssen gewissen Zwangsbedingungen genügen. Man erhält diese durch Vergleich der Summe der Partialbilanzen mit der Mischungsbilanz. Für den Fall einer skalarwertigen physikalischen Größe lauten die Zwangsbedingungen:

$$\begin{aligned} &\text{Physikalische Größe:} && \Psi &&= \sum_\alpha \Psi^\alpha\,, \\ &\text{Fluß:} && \boldsymbol{\phi} &&= \sum_\alpha \left(\boldsymbol{\phi}^\alpha - \Psi^\alpha\,\mathbf{d}_\alpha\right)\,, \\ &\text{Zufuhr:} && \sigma &&= \sum_\alpha \sigma^\alpha\,, \\ &\text{Produktion:} && \hat{\Psi} &&= \sum_\alpha \hat{\Psi}^\alpha\,. \end{aligned} \tag{1.55}$$

Darin ist $\mathbf{d}_\alpha = \mathbf{x}'_\alpha - \dot{\mathbf{x}}$ die in Gleichung (1.24) eingeführte Diffusionsgeschwindigkeit. Im Fall einer vektorwertigen physikalischen Größe erhält man analoge Zwangsbedingungen.

1.3.2 Massenbilanzen

Für die Mischung als Ganzes lautet das Axiom der Massenerhaltung

$$m = \int\limits_{\Omega_t}\mathrm{d}m = \text{konst.} \qquad \text{bzw.} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{\Omega_t}\mathrm{d}m = 0\,, \tag{1.56}$$

d. h. die Gesamtmasse des Mischungskörpers bleibt für alle Zeiten konstant. Unter Berücksichtigung von $\mathrm{d}m = \rho\,\mathrm{d}v$ liefert der Vergleich mit $(1.51)_1$ die Massendichte ρ als physikalische Größe, wobei Fluß, Zufuhr und Produktion entfallen:

$$\Psi = \rho\,, \qquad \boldsymbol{\phi} = \mathbf{0}\,, \qquad \sigma = 0\,, \qquad \hat{\Psi} = 0\,.$$

Damit lautet die *Massenbilanz der Mischung* in lokaler Form:

$$\boxed{\dot{\rho} + \rho \operatorname{div} \dot{\mathbf{x}} = 0}\,. \tag{1.57}$$

Für die einzelnen Konstituierenden wird keine Erhaltungsgleichung gefordert, sondern es werden Produktionsterme $\hat{\rho}^\alpha$ eingeführt, mit denen Massenaustauschprozesse beschrieben werden können (z. B. der Übergang von flüssigem Wasser in Eis oder Wasserdampf):

$$\frac{\mathrm{d}_\alpha}{\mathrm{d}t} \int\limits_{\Omega_t} \mathrm{d}m^\alpha \;=\; \int\limits_{\Omega_t} \hat{\rho}^\alpha \,\mathrm{d}v\,. \tag{1.58}$$

Durch Vergleich mit $(1.53)_1$ können physikalische Größe, Fluß, Zufuhr und Produktion wie folgt identifiziert werden:

$$\Psi^\alpha \;=\; \rho^\alpha\,, \qquad \boldsymbol{\phi}^\alpha \;=\; \mathbf{0}\,, \qquad \sigma^\alpha \;=\; 0\,, \qquad \hat{\Psi}^\alpha \;=\; \hat{\rho}^\alpha\,.$$

Damit lautet die *Massenbilanz der Phase* φ^α in lokaler Form:

$$\boxed{(\rho^\alpha)'_\alpha + \rho^\alpha \operatorname{div} \mathbf{x}'_\alpha = \hat{\rho}^\alpha}\,. \tag{1.59}$$

Die Auswertung der Zwangsbedingungen (1.55) liefert neben der bekannten Gleichung (1.18) für die Mischungsdichte und dem Verschwinden der Diffusionsmassenströme (1.25) die naheliegende Forderung, daß die Summe aller Massenaustauschterme verschwindet:

$$\sum_\alpha \hat{\rho}^\alpha = 0\,. \tag{1.60}$$

1.3.3 Impulsbilanzen

Die Impulsbilanz ist auch als der *Satz von der Erhaltung der Bewegungsgröße* bekannt. Es wird die zeitliche Änderung des Impulses mit inneren und äußeren Kräften bilanziert:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int\limits_{\Omega_t} \rho \dot{\mathbf{x}} \,\mathrm{d}v \;=\; \int\limits_{\partial\Omega_t} \mathbf{T}\,\mathbf{n} \,\mathrm{d}a \;+\; \int\limits_{\Omega_t} \rho \mathbf{b} \,\mathrm{d}v\,. \tag{1.61}$$

Darin sind $\mathbf{T}$ der *Cauchy*sche Spannungstensor und $\rho\mathbf{b}$ die Volumenkraftdichte, die häufig im Sinne einer *A-priori*-Konstitutivgleichung mit der Schwerkraft identifiziert wird; sie kann aber auch eine anderweitig verursachte Volumenkraft darstellen, etwa die durch ein Magnetfeld in einem ferromagnetischen Körper induzierte Magnetkraft.

Der Vergleich mit $(1.51)_2$ liefert als physikalische Größe die Impulsdichte, als Flußtensor den *Cauchy*schen Spannungstensor und als Zufuhrvektor die Volumenkraftdichte:

$$\boldsymbol{\Psi} \;=\; \rho\dot{\mathbf{x}}\,, \qquad \boldsymbol{\Phi} \;=\; \mathbf{T}\,, \qquad \boldsymbol{\sigma} \;=\; \rho\mathbf{b}\,, \qquad \hat{\boldsymbol{\Psi}} \;=\; \mathbf{0}\,.$$

Der Produktionsvektor entfällt, da es sich um eine Erhaltungsgleichung handelt. Nach Einarbeitung der Massenbilanz erhält man die *Impulsbilanz der Mischung* in lokaler Form:

$$\boxed{\rho\ddot{\mathbf{x}} = \operatorname{div} \mathbf{T} + \rho\mathbf{b}}\,. \tag{1.62}$$

Wieder wird für die einzelnen Konstituierenden ein Produktionsterm $\hat{\mathbf{s}}^\alpha$ eingeführt, der sich als innere Interaktionskraft mit den anderen Konstituierenden interpretieren läßt:

$$\frac{\mathrm{d}_\alpha}{\mathrm{d}t}\int\limits_{\Omega_t}\rho^\alpha\mathbf{x}'_\alpha\,\mathrm{d}v \;=\; \int\limits_{\partial\Omega_t}\mathbf{T}^\alpha\,\mathbf{n}\,\mathrm{d}a \;+\; \int\limits_{\Omega_t}\rho^\alpha\mathbf{b}^\alpha\,\mathrm{d}v \;+\; \int\limits_{\Omega_t}\hat{\mathbf{s}}^\alpha\,\mathrm{d}v\,. \tag{1.63}$$

Der *Cauchy*sche Spannungstensor $\mathbf{T}^\alpha$ ist in diesem Fall ein partialer Spannungstensor, da er nur die Spannung einer Phase φ^α darstellt. Durch Vergleich mit $(1.53)_2$ identifiziert man wie oben die physikalische Größe, den Flußtensor, den Zufuhrvektor und den Produktionsvektor:

$$\boldsymbol{\Psi}^\alpha \;=\; \rho^\alpha\mathbf{x}'_\alpha\,, \qquad \boldsymbol{\Phi}^\alpha \;=\; \mathbf{T}^\alpha\,, \qquad \boldsymbol{\sigma}^\alpha \;=\; \rho^\alpha\mathbf{b}^\alpha\,, \qquad \hat{\boldsymbol{\Psi}}^\alpha \;=\; \hat{\mathbf{s}}^\alpha\,.$$

Die gesamte Impulsproduktion $\hat{\mathbf{s}}^\alpha = \hat{\rho}^\alpha\mathbf{x}'_\alpha + \hat{\mathbf{p}}^\alpha$ teilt sich auf in einen Anteil $\hat{\rho}^\alpha\mathbf{x}'_\alpha$, der durch den Produktionsterm einer „niedrigeren“ Bilanz, hier der Massenbilanz, hervorgerufen wird, und einen direkten Anteil $\hat{\mathbf{p}}^\alpha$. Durch Einarbeitung der „niedrigeren“ Massenbilanz erhält man aus der allgemeinen Bilanzgleichung mit den obigen Definitionen die *Impulsbilanz der Phase* φ^α in lokaler Form:

$$\boxed{\rho^\alpha\mathbf{x}''_\alpha = \operatorname{div}\mathbf{T}^\alpha + \rho^\alpha\mathbf{b}^\alpha + \hat{\mathbf{p}}^\alpha}\,. \tag{1.64}$$

Die vektoriellen Zwangsbedingungen lauten:

$$\begin{aligned}
\rho\,\dot{\mathbf{x}} &= \sum_\alpha \rho^\alpha\mathbf{x}'_\alpha\,,\\
\mathbf{T} &= \sum_\alpha (\mathbf{T}^\alpha - \rho^\alpha\mathbf{d}_\alpha\otimes\mathbf{d}_\alpha)\,,\\
\rho\,\mathbf{b} &= \sum_\alpha \rho^\alpha\mathbf{b}^\alpha\,,\\
\mathbf{0} &= \sum_\alpha \hat{\mathbf{s}}^\alpha\,.
\end{aligned} \tag{1.65}$$

1.3.4 Drallbilanzen

Die Drallbilanz ist auch als der *Drehimpulserhaltungssatz* bekannt. Es wird die zeitliche Änderung des Dralls mit den von inneren und äußeren Kräften verursachten Momenten bilanziert[7]:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{\Omega_t}(\mathbf{x}\times\rho\dot{\mathbf{x}})\,\mathrm{d}v \;=\; \int\limits_{\partial\Omega_t}(\mathbf{x}\times\mathbf{T})\,\mathbf{n}\,\mathrm{d}a \;+\; \int\limits_{\Omega_t}(\mathbf{x}\times\rho\mathbf{b})\,\mathrm{d}v\,. \tag{1.66}$$

[7] Hier wird das Vektorprodukt zwischen einem Vektor $\mathbf{v}$ und einem Tensor $\mathbf{T}$ sowie das Vektorprodukt zwischen zwei Tensoren $\mathbf{S}$ und $\mathbf{T}$ benutzt (vgl. *de Boer* [20, §4.9.2 ff.]). Ersteres liefert einen Tensor zweiter Stufe und ist über die drei Gleichungen $(\mathbf{v}\times\mathbf{T})\,\mathbf{w} = \mathbf{v}\times(\mathbf{T}\,\mathbf{w})$, $(\mathbf{T}\times\mathbf{v})\,\mathbf{w} = (\mathbf{T}\,\mathbf{w})\times\mathbf{v}$, $\mathbf{T}\cdot(\mathbf{S}\times\mathbf{v}) = -\mathbf{S}\cdot(\mathbf{T}\times\mathbf{v})$ definiert. Letzteres liefert einen Vektor und wird über die Gleichung $\mathbf{v}\cdot(\mathbf{S}\times\mathbf{T}) = -\mathbf{S}\cdot(\mathbf{v}\times\mathbf{T})$ eingeführt. Insbesondere liefert $\frac{1}{2}(\mathbf{I}\times\mathbf{T}) = \operatorname{axl}\mathbf{T}$ den axialen Vektor eines Tensors, der sich auf den schiefsymmetrischen Anteil bezieht, also für symmetrische Tensoren verschwindet: $\mathbf{T} = \mathbf{T}^T \iff \operatorname{axl}\mathbf{T} = \mathbf{0}$.

Der Vergleich mit $(1.51)_2$ liefert als physikalische Größe die Dralldichte, als Flußtensor den Momentenfluß und als Zufuhrvektor das Volumenkraftmoment:

$$\boldsymbol{\Psi} \;=\; \mathbf{x}\times\rho\dot{\mathbf{x}}\,,\quad \boldsymbol{\Phi} \;=\; \mathbf{x}\times\mathbf{T}\,,\quad \boldsymbol{\sigma} \;=\; \mathbf{x}\times\rho\mathbf{b}\,,\quad \hat{\boldsymbol{\Psi}} \;=\; \mathbf{0}\,.$$

Nach Einarbeitung „niedrigerer“ Bilanzen (Massenbilanz und Impulsbilanz) lautet die *Drallbilanz der Mischung* in lokaler Form:

$$\boxed{\mathbf{0}=\mathbf{I}\times\mathbf{T} \qquad\Longleftrightarrow\qquad \mathbf{T}=\mathbf{T}^T}\,. \tag{1.67}$$

Wie in der Theorie der Einphasenkontinua ist das Ergebnis lediglich die Symmetrie des *Cauchy*schen Spannungstensors. Man beachte, daß dies unabhängig von der Symmetrie bzw. Unsymmetrie der Partialspannungstensoren $\mathbf{T}^\alpha$ der Fall ist.

Für die einzelnen Konstituierenden wird ein Produktionsterm $\hat{\mathbf{h}}^\alpha$ eingeführt, mit dem ein Drallaustausch zwischen den Konstituierenden modelliert werden kann:

$$\frac{\mathrm{d}_\alpha}{\mathrm{d}t}\int\limits_{\Omega_t}(\mathbf{x}\times\rho^\alpha\mathbf{x}'_\alpha)\,\mathrm{d}v \;=\; \int\limits_{\partial\Omega_t}(\mathbf{x}\times\mathbf{T}^\alpha)\,\mathbf{n}\,\mathrm{d}a \;+\; \int\limits_{\Omega_t}(\mathbf{x}\times\rho^\alpha\mathbf{b}^\alpha)\,\mathrm{d}v \;+\; \int\limits_{\Omega_t}\hat{\mathbf{h}}^\alpha\,\mathrm{d}v\,. \tag{1.68}$$

Der Vergleich mit $(1.53)_2$ liefert hier:

$$\boldsymbol{\Psi}^\alpha \;=\; \mathbf{x}\times\rho^\alpha\mathbf{x}'_\alpha\,,\quad \boldsymbol{\Phi}^\alpha \;=\; \mathbf{x}\times\mathbf{T}^\alpha\,,\quad \boldsymbol{\sigma}^\alpha \;=\; \mathbf{x}\times\rho^\alpha\mathbf{b}^\alpha\,,\quad \hat{\boldsymbol{\Psi}}^\alpha \;=\; \hat{\mathbf{h}}^\alpha\,.$$

Die gesamte Drallproduktion $\hat{\mathbf{h}}^\alpha=\mathbf{x}\times(\hat{\rho}^\alpha\mathbf{x}'_\alpha+\hat{\mathbf{p}}^\alpha)+\hat{\mathbf{m}}^\alpha$ setzt sich wieder aus einem durch „niedrigere“ Bilanzen hervorgerufenen Term $\mathbf{x}\times(\hat{\rho}^\alpha\mathbf{x}'_\alpha+\hat{\mathbf{p}}^\alpha)$ und der direkten Drallproduktion $\hat{\mathbf{m}}^\alpha$ zusammen. Durch Einarbeitung der „niedrigeren“ Bilanzen (Massenbilanz und Impulsbilanz) erhält man aus der allgemeinen Bilanzgleichung mit den obigen Definitionen die *Drallbilanz der Phase* φ^α in lokaler Form:

$$\boxed{\mathbf{0}=\mathbf{I}\times\mathbf{T}^\alpha+\hat{\mathbf{m}}^\alpha}\,. \tag{1.69}$$

Bei verschwindender direkter Drallproduktion, $\hat{\mathbf{m}}^\alpha=\mathbf{0}$, erhält man also symmetrische Partialspannungstensoren, $\mathbf{T}^\alpha=(\mathbf{T}^\alpha)^T$, ansonsten können diese auch unsymmetrisch sein. Die Summe der Drallbilanzen der einzelnen Phasen liefert die Zwangsbedingung

$$\sum_\alpha\hat{\mathbf{m}}^\alpha \;=\; \mathbf{0}\,. \tag{1.70}$$

Bemerkung: Im Rahmen dieser Arbeit wird die Drallbilanz nicht weiter berücksichtigt, da die Partialspannungstensoren bei dem später betrachteten Materialmodell stets als symmetrisch angenommen werden können (vgl. Abschnitt 1.5). □

1.4 Thermodynamische Bilanzgleichungen

Die Energie- und die Entropiebilanz sind auch als der *erste und zweite Hauptsatz der Thermodynamik* bekannt. In bezug auf die richtige Formulierung der Entropieungleichung für

Mischungen gab es in den sechziger Jahren lange Zeit kontroverse Diskussionen unter den beteiligten Wissenschaftlern (vgl. die historische Zusammenstellung bei *Ehlers* [44]). Hier wird die inzwischen allgemein anerkannte Form der Entropiebilanz für Mischungen dargestellt. Die Auswertung der Entropieungleichung zur thermodynamisch konsistenten Formulierung von Materialgesetzen bei Mischungen ist aufgrund der Vielzahl der auftretenden Terme sehr aufwendig, weshalb hier nicht darauf eingegangen werden kann. Eine ausführliche Darstellung der Thematik kann ebenfalls bei *Ehlers* [44] nachgelesen werden. In jüngerer Zeit wurden von *Diebels* [39] Ergebnisse bei der thermodynamischen Betrachtung einer Mischung, bestehend aus einem mikropolaren Festkörper (*Cosserat*-Theorie) und einem viskosen Porenfluid, erzielt.

1.4.1 Energiebilanzen

Beim *Energieerhaltungssatz* wird die zeitliche Änderung der inneren und kinetischen Energie mit der Leistung der äußeren Nah- und Fernwirkungskräfte sowie dem Wärmezufluß und der Wärmezufuhr bilanziert:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{\Omega_t}\rho(\varepsilon+\tfrac{1}{2}\,\dot{\mathbf{x}}\cdot\dot{\mathbf{x}})\,\mathrm{d}v \;=\; \int\limits_{\partial\Omega_t}(\mathbf{T}^T\,\dot{\mathbf{x}}-\mathbf{q})\cdot\mathbf{n}\,\mathrm{d}a \;+\; \int\limits_{\Omega_t}\rho(\mathbf{b}\cdot\dot{\mathbf{x}}+r)\,\mathrm{d}v\,. \tag{1.71}$$

Darin sind ε die spezifische innere Energie, $\mathbf{q}$ der Wärmezufluß über die Oberfläche infolge äußerer Nahwirkung und r die Wärmezufuhr im Volumen infolge äußerer Fernwirkung (z. B. die im Inneren durch Mikrowellen zugeführte Wärme).

Der Vergleich mit $(1.51)_1$ liefert:

$$\begin{aligned}\Psi &= \rho(\varepsilon+\tfrac{1}{2}\,\dot{\mathbf{x}}\cdot\dot{\mathbf{x}})\,, & \boldsymbol{\phi} &= \mathbf{T}^T\,\dot{\mathbf{x}}-\mathbf{q}\,,\\ \sigma &= \rho(\mathbf{b}\cdot\dot{\mathbf{x}}+r)\,, & \hat{\Psi} &= \mathbf{0}\,.\end{aligned}$$

Nach Einarbeitung „niedrigerer“ Bilanzen erhält man die *Energiebilanz der Mischung* in lokaler Form:

$$\boxed{\rho\dot{\varepsilon}=\mathbf{T}\cdot\mathbf{L}-\operatorname{div}\mathbf{q}+\rho r}\,. \tag{1.72}$$

Darin ist $\mathbf{L}=\operatorname{grad}\dot{\mathbf{x}}$ der räumliche Geschwindigkeitsgradient des Mischungskörpers.

Für die einzelnen Konstituierenden wird ein Produktionsterm $\hat{e}^\alpha$ eingeführt, mit dem Energieaustauschvorgänge zwischen den Konstituierenden modelliert werden können. Ein analoges Vorgehen wie oben und der Vergleich mit $(1.53)_1$ führt auf die physikalische Größe, den Fluß, die Zufuhr und die Produktion:

$$\begin{aligned}\Psi^\alpha &= \rho^\alpha(\varepsilon^\alpha+\tfrac{1}{2}\,\mathbf{x}'_\alpha\cdot\mathbf{x}'_\alpha)\,, & \boldsymbol{\phi}^\alpha &= \mathbf{T}^{\alpha T}\,\mathbf{x}'_\alpha-\mathbf{q}^\alpha\,,\\ \sigma^\alpha &= \rho^\alpha(\mathbf{b}^\alpha\cdot\mathbf{x}'_\alpha+r^\alpha)\,, & \hat{\Psi}^\alpha &= \hat{e}^\alpha\,.\end{aligned}$$

Die gesamte Energieproduktion $\hat{e}^\alpha=\hat{\mathbf{p}}^\alpha\cdot\mathbf{x}'_\alpha+\hat{\rho}^\alpha(\varepsilon^\alpha+\frac{1}{2}\,\mathbf{x}'_\alpha\cdot\mathbf{x}'_\alpha)+\hat{\varepsilon}^\alpha$ setzt sich wieder aus Anteilen mit Produktionstermen „niedrigerer“ Bilanzen sowie dem direkten Anteil $\hat{\varepsilon}^\alpha$ zusammen. Mit dem in Gleichung (1.31) eingeführten räumlichen Geschwindigkeitsgradienten $\mathbf{L}_\alpha=\operatorname{grad}\mathbf{x}'_\alpha$ erhält man nach Einarbeitung „niedrigerer“ Bilanzen die *Energiebilanz*

der Phase φ^α in lokaler Form:

$$\boxed{\rho^\alpha(\varepsilon^\alpha)'_\alpha = \mathbf{T}^\alpha \cdot \mathbf{L}_\alpha - \operatorname{div}\mathbf{q}^\alpha + \rho^\alpha r^\alpha + \hat{\varepsilon}^\alpha}\,. \tag{1.73}$$

Auf die Auswertung der bei der Energiebilanz auftretenden Zwangsbedingungen wird an dieser Stelle verzichtet.

1.4.2 Entropieungleichung

Der Ansatz von *A-priori*-Konstitutivannahmen (*Ehlers* [47]) für den Entropiefluß und die Entropiezufuhr mit der absoluten *Kelvin*schen Temperatur Θ^α der Phase φ^α führt auf die folgende Identifikation von physikalischer Größe, Fluß, Zufuhr und Produktion:

$$\begin{aligned} \Psi^\alpha &= \rho^\alpha\eta^\alpha\,, & \boldsymbol{\phi}^\alpha &= -\frac{1}{\Theta^\alpha}\mathbf{q}^\alpha\,,\\ \sigma^\alpha &= \frac{1}{\Theta^\alpha}\rho^\alpha r^\alpha\,, & \hat{\Psi}^\alpha &= \hat{\eta}^\alpha\,. \end{aligned}$$

Die Entropieproduktion läßt sich in einen indirekten und einen direkten Anteil aufteilen: $\hat{\eta}^\alpha = \hat{\rho}^\alpha\eta^\alpha + \hat{\zeta}^\alpha$. Unter Berücksichtigung der Massenbilanzen der Phasen φ^α lautet damit die *Entropiebilanz der Phase* φ^α in lokaler Form:

$$\rho^\alpha(\eta^\alpha)'_\alpha = \operatorname{div}(-\frac{1}{\Theta^\alpha}\mathbf{q}^\alpha) + \frac{1}{\Theta^\alpha}\rho^\alpha r^\alpha + \hat{\zeta}^\alpha\,. \tag{1.74}$$

Das Ergebnis der oben erwähnten Diskussion war, daß die Annahme einer Entropieungleichung für jede einzelne Phase (also eine Ungleichung der Form $\hat{\eta}^\alpha \geq 0$) eine zu starke Einschränkung an die möglichen thermodynamischen Prozesse bedeutet. Lediglich für die Mischung als Ganzes ist der zweite Hauptsatz der Thermodynamik zu fordern ($\hat{\eta} \geq 0$). Die *Entropieungleichung der Mischung* kann demnach wie folgt formuliert werden:

$$\rho\dot{\eta} \geq \operatorname{div}(-\frac{1}{\Theta}\mathbf{q}) + \frac{1}{\Theta}\rho r\,. \tag{1.75}$$

Die Zwangsbedingungen (1.55) lauten im Falle der Entropiebilanz:

$$\begin{aligned} \rho\eta &= \sum_\alpha \rho^\alpha\eta^\alpha\,,\\ \frac{1}{\Theta}\mathbf{q} &= \sum_\alpha \frac{1}{\Theta^\alpha}\mathbf{q}^\alpha + \rho^\alpha\eta^\alpha\mathbf{d}_\alpha\,,\\ \frac{1}{\Theta}\rho r &= \sum_\alpha \frac{1}{\Theta^\alpha}\rho^\alpha r^\alpha\,,\\ \hat{\eta} &= \sum_\alpha \hat{\eta}^\alpha\,. \end{aligned} \tag{1.76}$$

Mit $(1.76)_4$ kann die Entropieungleichung der Mischung auch in der Form

$$\boxed{\hat{\eta} = \sum_\alpha \hat{\eta}^\alpha = \sum_\alpha \left[\rho^\alpha(\eta^\alpha)'_\alpha + \hat{\rho}^\alpha\eta^\alpha + \operatorname{div}(\frac{1}{\Theta^\alpha}\mathbf{q}^\alpha) - \frac{1}{\Theta^\alpha}\rho^\alpha r^\alpha\right] \geq 0} \tag{1.75*}$$

dargestellt werden.

1.5 Ein inkompressibles Zweiphasenmodell

Die *Theorie Poröser Medien* soll nun zur Beschreibung von bindigen Böden herangezogen werden. In der Bodenmechanik unterscheidet man in Tiefenrichtung des Bodens im wesentlichen drei Bereiche:

- Die *ungesättigte Zone*: In den oberen Erdschichten ist der Porenraum des Bodens mit Luft gefüllt.
- Die *teilgesättigte Zone*: In darunter liegenden Erdschichten sind die Poren zum Teil mit Luft, zum Teil mit Wasser gefüllt.
- Die *gesättigte Zone*: Unterhalb des Grundwasserspiegels sind alle Poren des Bodens mit Wasser gefüllt.

Der Begriff der Sättigung ist in der Bodenmechanik also immer auf das Porenwasser bezogen. Im Gegensatz dazu werden im Rahmen der Theorie Poröser Medien alle drei Bereiche als gesättigt bezeichnet (vgl. Gleichung (1.14)); es findet lediglich ein Wechsel des Porenfluids bzw. der Porenfluide statt.

Bei der Modellierung bodenmechanischer Fragestellungen sind also je nach der betrachteten Zone zwei oder drei Phasen einzubeziehen. Für die ungesättigte Zone sind dies das Festkörperskelett des Bodens und ein Gas (Zweiphasenmodell mit inkompressiblem Festkörper und kompressiblem Fluid), für die teilgesättigte Zone kommt als dritte Phase Wasser hinzu (Dreiphasenmodell), und in der gesättigten Zone wird der Boden durch ein wassergefülltes Festkörperskelett dargestellt (Zweiphasenmodell mit inkompressiblem Festkörper und inkompressiblem Fluid).

Im folgenden wird ein Zweiphasenmodell für die gesättigte Zone vorgestellt. Das Bodenmaterial ist aus dem Blickwinkel der *Theorie Poröser Medien* ein wassergesättigter poröser Festkörper, wobei das Festkörperskelett φ^S mit dem Index $\alpha = S$ (engl.: solid) und das Porenfluid φ^F mit dem Index $\alpha = F$ (engl.: fluid) bezeichnet wird. Das *inkompressible Zweiphasenmodell* basiert auf den folgenden konstitutiven Annahmen, die in den nächsten Abschnitten detailliert diskutiert werden:

- Gemeinsame, konstante Temperatur beider Phasen,
- gemeinsame, konstante Volumenkraft beider Phasen (Schwerkraft),
- materielle Inkompressibilität beider Phasen,
- kein Massenaustausch zwischen den Phasen,
- symmetrische Partialspannungen,
- elastisch-viskoplastisches Festkörperskelett,
- viskoses Porenfluid,
- geometrisch lineare Theorie (kleine Verzerrungen).

Die Annahme einer gemeinsamen, konstanten Temperatur beider Phasen führt auf ein rein mechanisches Modell ohne explizite Berücksichtigung der Energiebilanz. Desweiteren stellt die Schwerkraft in der Bodenmechanik die einzige Volumenkraft dar. Die materielle Inkompressibilität ist sowohl für das Festkörpermaterial des bindigen Bodens (z. B. Ton) als auch für das Porenfluid (Wasser) eine realistische Annahme; bei einer Mischung aus Ton und Wasser gibt es zudem keinen Massenaustauch und es kann von symmetrischen Partialspannungen ausgegangen werden. Die Plastizitätsformulierung für den Festkörper muß die speziellen Eigenschaften von Reibungsmaterialien berücksichtigen, da ansonsten nicht mit realistischen Ergebnissen zu rechnen ist. Ein *Reibungsmaterial* (engl.: frictional material) kann auch unter hydrostatischen Spannungszuständen bleibende Deformationen erleiden. Schließlich liegt der Anwendungsbereich des Modells in der Berechnung des Bodenverhaltens bei Bauingenieur-Fragestellungen, weshalb grundsätzlich von kleinen Deformationen ausgegangen wird. Die Voraussetzung einer geometrisch linearen Theorie führt zu den in Abschnitt 1.2.5 hergeleiteten Vereinfachungen. Die in den beiden folgenden Abschnitten beschriebenen Umformungen und Gleichungen gelten jedoch grundsätzlich in der allgemeinen, nichtlinearen Formulierung; linearisierte Formen sind jeweils gesondert vermerkt.

1.5.1 Volumenbilanzen

Da beide Phasen als *materiell inkompressibel* angenommen werden, ist die realistische Dichte $\rho^{\alpha R} : \Omega_t \longrightarrow \mathbb{R}$ (siehe (1.17)) zeitlich konstant. Mit Hilfe der realistischen Dichte $\rho_0^{\alpha R} : \Omega_0 \longrightarrow \mathbb{R}$ der Referenzkonfiguration wird definiert:

$$\rho^{\alpha R}(\mathbf{x}, t) := \rho_0^{\alpha R}(\overset{\alpha}{\boldsymbol{\chi}}{}_t^{-1}(\mathbf{x})) = \rho_0^{\alpha R}(\mathbf{X}_\alpha)\,.$$

Bemerkung: In der Referenzkonfiguration kann die Dichteverteilung durchaus inhomogen sein ($\mathrm{Grad}_\alpha\, \rho_0^{\alpha R} \neq 0$), jedoch „haftet“ die in der Referenzkonfiguration gegebene Dichte am jeweiligen materiellen Punkt, verändert sich also zeitlich nicht. Die Partialdichte $\rho^\alpha = n^\alpha\, \rho^{\alpha R}$ ist jedoch aufgrund möglicher Änderungen der Volumenanteile trotz materieller Inkompressibilität *nicht* konstant. □

Da die materielle Zeitableitung der realistischen Dichte verschwindet,

$$(\rho^{\alpha R})'_\alpha = \frac{\mathrm{d}_\alpha}{\mathrm{d}t}\left[\rho^{\alpha R}(\mathbf{x}, t)\right] = \frac{\mathrm{d}_\alpha}{\mathrm{d}t}\left[\rho_0^{\alpha R}(\mathbf{X}_\alpha)\right] = 0\,,$$

reduziert sich die Massenbilanz (1.59) aufgrund

$$(\rho^\alpha)'_\alpha = (n^\alpha\, \rho^{\alpha R})'_\alpha = (n^\alpha)'_\alpha\, \rho^{\alpha R} + n^\alpha \underbrace{(\rho^{\alpha R})'_\alpha}_{=0}$$

unter der Voraussetzung positiver Dichte ($\rho^{\alpha R} > 0$) und unter Vernachlässigung des Massenaustauschs ($\hat{\rho}^\alpha = 0$) zur *Volumenbilanz der Phase* φ^α:

$$\boxed{(n^\alpha)'_\alpha + n^\alpha \operatorname{div} \mathbf{x}'_\alpha = 0}\,. \tag{1.77}$$

Die Volumenbilanz des Festkörpers kann man analytisch integrieren und erhält mit dem Anfangswert n^S_{0S} (Volumenanteil des Festkörpers in der Referenzkonfiguration von φ^S):

$$n^S = n^S_{0S}\ \det \mathbf{F}_S^{-1} = \frac{1}{J_S}\ n^S_{0S}\,. \tag{1.78}$$

Mit der formalen Linearisierung der inversen *Jacobi*-Determinante um den natürlichen Zustand (Referenzkonfiguration) erhält man die linearisierte Volumenbilanz:

$$n^S = n^S_{0S}\ (1 - \mathrm{Div}_S\, \mathbf{u}_S)\,. \tag{1.79}$$

Die *Sickergeschwindigkeit* wird definiert als die Differenzgeschwindigkeit des Fluids relativ zum sich deformierenden Festkörper (modifizierte *Euler*sche Beschreibung des Fluids):

$$\mathbf{w}_F = \mathbf{x}'_F - \mathbf{x}'_S\,. \tag{1.80}$$

Mit der Sickergeschwindigkeit und (1.30) kann man alle auftretenden Zeitableitungen bzgl. der Festkörperbewegung darstellen,

$$\Gamma'_F = \Gamma'_S + \mathrm{grad}\,\Gamma \cdot \mathbf{w}_F\,, \qquad\qquad \boldsymbol{\Gamma}'_F = \boldsymbol{\Gamma}'_S + (\mathrm{grad}\,\boldsymbol{\Gamma})\,\mathbf{w}_F\,, \tag{1.81}$$

und erhält aus der Summe der Volumenbilanzen (1.77) des Festkörpers und des Fluids mit Hilfe der Sättigungsbedingung (1.14):

$$\begin{aligned} & (n^S)'_S + n^S \,\mathrm{div}\,\mathbf{x}'_S + (n^F)'_S + \mathrm{grad}\,n^F \cdot \mathbf{w}_F + n^F\,\mathrm{div}(\mathbf{x}'_S + \mathbf{w}_F) &&= 0 \\ \Longleftrightarrow \quad & \underbrace{(n^S + n^F)'_S}_{=0} + \underbrace{(n^S + n^F)}_{=1}\,\mathrm{div}\,\mathbf{x}'_S + \mathrm{grad}\,n^F \cdot \mathbf{w}_F + n^F\,\mathrm{div}\,\mathbf{w}_F &&= 0\,. \end{aligned}$$

Die Anwendung der Produktregel $\mathrm{div}(n^F\mathbf{w}_F) = \mathrm{grad}\,n^F \cdot \mathbf{w}_F + n^F\,\mathrm{div}\,\mathbf{w}_F$ führt schließlich mit $\mathbf{v}_S = \mathbf{x}'_S$ auf die *Volumenbilanz der Mischung*:

$$\boxed{\mathrm{div}(\mathbf{v}_S + n^F\mathbf{w}_F) = 0}\,. \tag{1.82}$$

1.5.2 Konstitutivgleichungen

Die Sättigungsbedingung (1.14) führt zu einer additiven Aufteilung der Konstitutivgleichungen in zwei Anteile (*Ehlers* [44]). Ein Summand ist bestimmt durch den *Lagrange*-Parameter p, der als effektiver Porenfluiddruck identifiziert werden kann. Ein zweiter Summand ist der sogenannte Extraanteil, im folgenden durch $(...)_E$ gekennzeichnet, der durch konstitutive Annahmen im Rahmen der Materialtheorie bestimmt wird. Demnach haben die Partialspannungstensoren und der Impulsaustauschterm die folgende Form:

$$\mathbf{T}^\alpha = -n^\alpha p\,\mathbf{I} + \mathbf{T}^\alpha_E\,, \qquad \hat{\mathbf{p}}^\alpha = p\ \mathrm{grad}\,n^\alpha + \hat{\mathbf{p}}^\alpha_E\,. \tag{1.83}$$

Im Rahmen der geometrisch linearen Theorie wird der linearisierte Verzerrungstensor des Festkörpers additiv in einen elastischen und einen plastischen Anteil aufgespalten:

$$\boldsymbol{\varepsilon}_S = \frac{1}{2}\left(\mathrm{Grad}_S\,\mathbf{u}_S + \mathrm{Grad}_S^T\,\mathbf{u}_S\right) = \boldsymbol{\varepsilon}_{Se} + \boldsymbol{\varepsilon}_{Sp}\,. \tag{1.84}$$

Der *Cauchy*sche Extraspannungstensor hängt nach dem *Hooke*schen Gesetz

$$\mathbf{T}_E^S = \boldsymbol{\sigma}_E^S = 2\mu^S \boldsymbol{\varepsilon}_{Se} + \lambda^S \ \mathrm{tr}\, \boldsymbol{\varepsilon}_{Se}\, \mathbf{I} \tag{1.85}$$

mit den *Lamé*schen Konstanten μ^S und λ^S nur von der elastischen Verzerrung $\boldsymbol{\varepsilon}_{Se}$ ab. Man beachte, daß die *Lamé*schen Konstanten bei der Behandlung von porösen Materialien als Strukturparameter zu verstehen sind, so daß trotz materieller Inkompressibilität des Festkörpers eine Volumendeformation möglich ist. Im hier nicht betrachteten Bereich großer Deformationen erfordert dies die Berücksichtigung des *Kompressionspunkts* im elastischen Gesetz (*Ehlers & Eipper* [48]), der erreicht wird, wenn alle Poren geschlossen sind.

In Abschnitt 1.5.4 wird auf die Entwicklungsgleichung für die plastische Verzerrung $\boldsymbol{\varepsilon}_{Sp}$ eingegangen, die ebenfalls konstitutiv vorzugeben ist. Wie in der Grundwasserhydraulik üblich, wird die Extraspannung des Fluids (Reibungsspannung) $\mathbf{T}_E^F = 2\mu^F \mathbf{D}_F \approx \mathbf{0}$ *a priori* vernachlässigt, da die Fluidviskosität implizit in den *Darcy*schen *Permeabilitätsbeiwert* k^F des Extraanteils $\hat{\mathbf{p}}_E^F$ der Impulsproduktion eingeht:

$$\hat{\mathbf{p}}_E^F = \mathbf{S}_V\, \mathbf{w}_F\,, \quad \mathbf{S}_V = -\frac{(n^F)^2\, \gamma^{FR}}{k^F}\, \mathbf{I}\,, \qquad \hat{\mathbf{p}}_E^S + \hat{\mathbf{p}}_E^F = \mathbf{0}\,. \tag{1.86}$$

Darin ist $\gamma^{FR} = \rho^{FR}\, g$ die *reale Wichte* des Fluids mit dem Betrag g der Gravitationsbeschleunigung. Der *Permeabilitätstensor* $\mathbf{S}_V$ wird hier als isotrop und unabhängig von der Deformation angenommen.

1.5.3 Impulsbilanzen

Ausgehend von der Impulsbilanz des Fluids sowie den im letzten Abschnitt beschriebenen Konstitutivgleichungen,

$$\begin{aligned}
&(1.64): && \rho^F \mathbf{x}_F'' &&= \operatorname{div} \mathbf{T}^F + \rho^F \mathbf{b} + \hat{\mathbf{p}}^F\,,\\
&(1.81): && \mathbf{x}_F'' &&= (\mathbf{v}_S + \mathbf{w}_F)_S' + \operatorname{grad}(\mathbf{v}_S + \mathbf{w}_F)\, \mathbf{w}_F\,,\\
&(1.83),\ \mathbf{T}_E^F = \mathbf{0}: && \operatorname{div} \mathbf{T}^F &&= \operatorname{div}(-n^F p\, \mathbf{I}) \ = \ -n^F \operatorname{grad} p - p\, \operatorname{grad} n^F\,,\\
&(1.83),\ (1.86): && \hat{\mathbf{p}}^F &&= p\, \operatorname{grad} n^F - \frac{(n^F)^2\, \gamma^{FR}}{k^F}\, \mathbf{w}_F\,,
\end{aligned}$$

erhält man die *Impulsbilanz des Fluids im Zweiphasenmodell*:

$$\boxed{\rho^F \left([(\mathbf{v}_S + \mathbf{w}_F)_S' + \operatorname{grad}(\mathbf{v}_S + \mathbf{w}_F)\, \mathbf{w}_F] - \mathbf{b} \right) = -n^F \operatorname{grad} p - \frac{(n^F)^2\, \gamma^{FR}}{k^F}\, \mathbf{w}_F}\,. \tag{1.87}$$

In einer quasi-statischen Formulierung (Vernachlässigung des Trägheitsterms $\rho^F \mathbf{x}_F''$) kann man (1.87) nach der *Filtergeschwindigkeit* $n^F \mathbf{w}_F$ auflösen und erhält das bekannte *Darcysche Gesetz* der Durchströmung eines porösen Mediums:

$$n^F \mathbf{w}_F = -\frac{k^F}{\gamma^{FR}} \left(\operatorname{grad} p - \rho^{FR}\, \mathbf{b} \right)\,. \tag{1.88}$$

Addiert man die Impulsbilanzen beider Phasen, Fluid und Festkörper, so erhält man mit

$$\begin{aligned}
&(1.64),\,(1.83)_2,\,(1.86)_3: && \rho^S \mathbf{x}_S'' + \rho^F \mathbf{x}_F'' &&= \operatorname{div}(\mathbf{T}^S + \mathbf{T}^F) + (\rho^S + \rho^F)\,\mathbf{b}\,,\\
&(1.81): && \mathbf{x}_F'' &&= (\mathbf{v}_S + \mathbf{w}_F)_S' + \operatorname{grad}(\mathbf{v}_S + \mathbf{w}_F)\,\mathbf{w}_F\,,\\
&(1.83),\,\mathbf{T}_E^F = \mathbf{0}: && \operatorname{div}(\mathbf{T}^S + \mathbf{T}^F) &&= \operatorname{div}(\mathbf{T}_E^S - p\,\mathbf{I})
\end{aligned}$$

die *Impulsbilanz der Mischung im Zweiphasenmodell*:

$$\boxed{\rho^S\Big(\mathbf{x}_S'' - \mathbf{b}\Big) + \rho^F\Big([(\mathbf{v}_S + \mathbf{w}_F)_S' + \operatorname{grad}(\mathbf{v}_S + \mathbf{w}_F)\,\mathbf{w}_F] - \mathbf{b}\Big) = \operatorname{div}(\mathbf{T}_E^S - p\,\mathbf{I})}\,. \tag{1.89}$$

Die obigen Gleichungen gelten auch im Rahmen einer geometrisch nichtlinearen Formulierung, wobei insbesondere der *Cauchy*sche Extraspannungstensor $\mathbf{T}_E^S$ mittels eines nichtlinearen Elastizitätsgesetzes bestimmt werden könnte. Zur Kennzeichnung des linearen Spannungstensors wird in den folgenden Abschnitten $\boldsymbol{\sigma}_E^S$ statt $\mathbf{T}_E^S$ verwendet.

1.5.4 Viskoplastizität

Im Gegensatz zur Metallplastizität muß ein Fließkriterium für Geomaterialien eine geschlossene Form im Hauptspannungsraum besitzen, da plastisches Fließen auch unter rein hydrostatischen Spannungszuständen stattfinden kann, vgl. *Ehlers* [46]. Das dort angegebene Einflächen-Fließkriterium kann in Abhängigkeit der ersten Invarianten I sowie der zweiten und dritten deviatorischen Invarianten II_D und III_D des Spannungstensors $\boldsymbol{\sigma}_E^S$ ausgedrückt werden[8]:

$$F(\boldsymbol{\sigma}_E^S) = \sqrt{\mathrm{II}_D\left(1 + \gamma\,\mathrm{III}_D/\mathrm{II}_D^{3/2}\right)^m + \frac{1}{2}\alpha\mathrm{I}^2 + \delta^2\mathrm{I}^4} + \beta\mathrm{I} + \varepsilon\mathrm{I}^2 - \kappa\,. \tag{1.90}$$

Dieses Fließkriterium enthält 7 Materialparameter $\alpha, \beta, \gamma, \delta, \varepsilon, \kappa$ und m, die Versuchsergebnissen angepaßt werden müssen. Neben der geschlossenen Fließfläche benötigt man im Kontext von Geomaterialien eine nicht-assoziierte Fließregel, d. h. das plastische Potential G ist verschieden vom Fließkriterium F. Hier wird

$$G(\boldsymbol{\sigma}_E^S) = \sqrt{\mathrm{II}_D + \frac{1}{2}\alpha\mathrm{I}^2 + \delta^2\mathrm{I}^4} + \beta\mathrm{I} + \varepsilon\mathrm{I}^2 + g(\mathrm{I}) \tag{1.91}$$

aus *Diebels, Ellsiepen & Ehlers* [40] verwendet, bei dem mit Hilfe der Funktion $g(\mathrm{I})$ die Abweichung der gemessenen Fließrichtung von der assoziierten Richtung angepaßt werden kann (Modellierung des Dilatanzwinkels). Damit ergibt sich die Fließregel für den plastischen Anteil des linearisierten Verzerrungstensors (vgl. (1.84)) zu

$$(\boldsymbol{\varepsilon}_{Sp})_S' = \Lambda \frac{\partial G}{\partial \boldsymbol{\sigma}_E^S}\,, \tag{1.92}$$

[8] Der *Deviator* eines Tensors $\mathbf{T}$ ist definiert als: $\operatorname{dev}\mathbf{T} := \mathbf{T}^D := \mathbf{T} - (\frac{1}{3}\operatorname{tr}\mathbf{T})\,\mathbf{I}$. Bei Deformations- und Verzerrungstensoren beschreibt der Deviator den „gestaltändernden“ Anteil des Tensors $\mathbf{T}$, während der *Kugeltensor* $(\frac{1}{3}\operatorname{tr}\mathbf{T})\,\mathbf{I}$ die Volumenänderung beschreibt.

wobei der Proportionalitätsfaktor Λ im Fall des hier betrachteten viskoplastischen Modells nach einem Überspannungsansatz vom *Perzyna*-Typ [95]

$$\Lambda = \frac{1}{\eta} \left\langle \frac{F(\boldsymbol{\sigma}_E^S)}{\sigma_0} \right\rangle^r \tag{1.93}$$

mit der Relaxationszeit η und einem Exponenten r angenommen wird, vgl. *Hartmann, Lührs & Haupt* [66]. Das *Föppl*-Symbol (*Macauley*-Klammer) ist darin definiert durch $\langle x \rangle = (x + |x|)/2$, so daß im elastischen Bereich mit $F(\boldsymbol{\sigma}_E^S) \leq 0$ kein plastisches Fließen stattfindet. Dies entspricht einem Knick in der rechten Seite der gewöhnlichen Differentialgleichung für die plastische Verzerrung (1.92), was bei der numerischen Behandlung berücksichtigt werden muß. Im Fall ratenunabhängiger Plastizität (Elastoplastizität) wird der Proportionalitätsfaktor Λ anstatt über (1.93) aus den *Kuhn-Tucker*-Bedingungen

$$F \leq 0, \quad \Lambda \geq 0, \quad \Lambda F = 0 \tag{1.94}$$

bestimmt. Die ersten beiden Bedingungen werden dabei üblicherweise durch die algorithmische Vorgehensweise (elastischer Prädiktorschritt) automatisch erfüllt, so daß im plastischen Bereich (im Prädiktorschritt ist $F \geq 0$) der Proportionalitätsfaktor Λ über die Nebenbedingung $F = 0$ bestimmt wird. Im viskoplastischen Ansatz entspricht dies einem Grenzübergang $\eta \to 0$ (Relaxation in beliebig kurzer Zeit). Für kleine η hat man also eine singulär gestörte Differentialgleichung (1.92), (1.93), die im Grenzfall $\eta \to 0$ in das differential-algebraische System aus (1.92) und $F = 0$ übergeht.

1.6 Zusammenstellung der Modellgleichungen

In diesem Abschnitt werden zunächst die Variablen, Gleichungen und Parameter des inkompressiblen Zweiphasenmodells zusammengestellt, die in den vorherigen Abschnitten hergeleitet wurden. Anschließend werden Formulierungen für geeignete Primärvariablen

Variablen des inkompressiblen Zweiphasenmodells

1.	Volumenanteil (F):	n^F
2.	Volumenanteil (S):	n^S
3.	Porenfluiddruck (F):	p
4.	Sickergeschwindigkeit (F):	$\mathbf{w}_F = \mathbf{x}'_F - \mathbf{x}'_S$
5.	Verschiebung (S):	$\mathbf{u}_S = \mathbf{x} - \mathbf{X}_S$
6.	Geschwindigkeit (S):	$\mathbf{v}_S = (\mathbf{u}_S)'_S$
7.	Gesamtverzerrungen (S):	$\boldsymbol{\varepsilon}_S$
8.	Extraspannungen (S):	$\boldsymbol{\sigma}_E^S$
9.	Plastische Verzerrungen (S):	$\boldsymbol{\varepsilon}_{Sp}$
10.	Proportionalitätsfaktor (S):	Λ

(1.95)

angegeben, die eine numerische Umsetzung des Modells im Rahmen der Methode der finiten Elemente (FEM) gestatten.

Die Variablen des inkompressiblen Zweiphasenmodells sind in Kasten (1.95) zusammengestellt. In Klammern ist jeweils vermerkt, ob sich die Variable auf den Festkörper (S) oder das Fluid (F) bezieht. Die Gleichungen des inkompressiblen Zweiphasenmodells sind in

Gleichungen des inkompressiblen Zweiphasenmodells

1. Sättigung (M): $1 = n^S + n^F$

2. Volumenbilanz (S): $n^S = n^S_{0S}(1 - \operatorname{div} \mathbf{u}_S)$

3. Volumenbilanz (M): $0 = \operatorname{div}(\mathbf{v}_S + n^F \mathbf{w}_F)$

4. Impulsbilanz (F):

$$n^F \rho^{FR} \left[(\mathbf{v}_S + \mathbf{w}_F)'_S + \operatorname{grad}(\mathbf{v}_S + \mathbf{w}_F)\, \mathbf{w}_F \right] = -n^F \operatorname{grad} p - \frac{(n^F)^2\, \gamma^{FR}}{k^F}\, \mathbf{w}_F + n^F \rho^{FR}\, \mathbf{b}$$

quasi-statisch: $n^F \mathbf{w}_F = -\dfrac{k^F}{\gamma^{FR}} \left(\operatorname{grad} p - \rho^{FR}\, \mathbf{b} \right)$

5. Impulsbilanz (M):

$$n^S \rho^{SR}\, (\mathbf{v}_S)'_S + n^F \rho^{FR} \left[(\mathbf{v}_S + \mathbf{w}_F)'_S + \operatorname{grad}(\mathbf{v}_S + \mathbf{w}_F)\, \mathbf{w}_F \right] = \operatorname{div}(\boldsymbol{\sigma}^S_E - p\, \mathbf{I}) + (n^S \rho^{SR} + n^F \rho^{FR})\, \mathbf{b}$$

quasi-statisch: $\mathbf{0} = \operatorname{div}(\boldsymbol{\sigma}^S_E - p\, \mathbf{I}) + (n^S \rho^{SR} + n^F \rho^{FR})\, \mathbf{b}$

6. Geschwindigkeit (S): $\mathbf{v}_S = (\mathbf{u}_S)'_S$

7. Verzerrungen (S): $\boldsymbol{\varepsilon}_S = \dfrac{1}{2} \left(\operatorname{grad} \mathbf{u}_S + \operatorname{grad}^T \mathbf{u}_S \right) = \boldsymbol{\varepsilon}_{Se} + \boldsymbol{\varepsilon}_{Sp}$

8. Elastizität (S): $\boldsymbol{\sigma}^S_E = 2\mu^S \boldsymbol{\varepsilon}_{Se} + \lambda^S\ \operatorname{tr} \boldsymbol{\varepsilon}_{Se}\, \mathbf{I}$

9. Fließregel (S): $(\boldsymbol{\varepsilon}_{Sp})'_S = \Lambda \dfrac{\partial G}{\partial \boldsymbol{\sigma}^S_E}$

10a. Viskoplastizität (S): $\Lambda = \dfrac{1}{\eta} \left\langle \dfrac{F(\boldsymbol{\sigma}^S_E)}{\sigma_0} \right\rangle^r$

10b. Elastoplastizität (S): $\Lambda\, F = 0, \quad F \leq 0, \quad \Lambda \geq 0$

(1.96)

Kasten (1.96) zusammengefaßt, wobei alle in diesem Abschnitt angegebenen Annahmen und Konstitutivgleichungen eingearbeitet sind. Insbesondere sind alle örtlichen Ableitungen aufgrund der geometrischen Linearität auf die Referenzkonfiguration bezogen, jedoch mit den in der linearen Theorie gebräuchlichen Schreibweisen grad(...) und div(...) statt $\operatorname{Grad}_S(...)$ und $\operatorname{Div}_S(...)$ bezeichnet. Alle zeitlichen Ableitungen sind auf die Festkörperbewegung bezogen, so daß nur noch eine Zeitableitung $(...)'_S$ vorkommt. Bei langsam veränderlichen Prozessen gelangt man durch Vernachlässigung der Beschleunigungsterme in den Impulsbilanzen zu einer quasi-statischen Formulierung, die in Kasten (1.96) gesondert angegeben ist. In Klammern ist jeweils vermerkt, ob die Gleichung sich auf den Festkörper (S), das Fluid (F) oder die Mischung (M) bezieht.

Die Formulierung der Plastizität erlaubt die Auswahl zwischen einem viskoplastischen Modell (Verwendung der Gleichung 10a zur Bestimmung des Proportionalitätsfaktors Λ) und einem elastoplastischen Modell (Verwendung der *Kuhn-Tucker*-Bedingungen 10b zur Bestimmung des Proportionalitätsfaktors Λ). In beiden Fällen hat das Modell gleich viele Variablen wie Gleichungen, ist also in sich geschlossen.

Parameter des inkompressiblen Zweiphasenmodells

Anfangs-Volumenanteil (S):	n_{0S}^S
Darcy-Permeabilität (F):	k^F
Reale Wichte (F):	γ^{FR}
Reale Dichte (F):	ρ^{FR}
Volumenkraft (S+F):	$\mathbf{b}$
Reale Dichte (S):	ρ^{SR}
Lamé-Konstanten (S):	μ^S, λ^S
Fließbedingung (S):	$F(\boldsymbol{\sigma}_E^S)$
Plastisches Potential (S):	$G(\boldsymbol{\sigma}_E^S)$
Viskoplastizität (S):	η, σ_0, r

(1.97)

Die nun noch unbestimmten Größen sind die Materialparameter bzw. Materialfunktionen, die in Kasten (1.97) zusammengestellt sind. Die Parameter zur Beschreibung der Plastizität werden hier nicht mehr im einzelnen aufgeführt, vielmehr wird auf die Definition der Funktionen F und G in Abschnitt 1.5.4 verwiesen.

Das Zweiphasenmodell wird abschließend noch in zwei speziellen Formulierungen angegeben, die als Grundlage für die weitere Arbeit dienen. Die dynamische Formulierung beinhaltet alle dargestellten Gleichungen und Variablen und stellt damit die allgemeinste Version des Modells dar. Treten in einer Anwendung jedoch nur kleine Beschleunigungen auf, so können diese *a priori* vernachlässigt werden, und man gelangt zu einer quasistatischen Formulierung des Modells. Insbesondere im Hinblick auf die Behandlung mit numerischen Diskretisierungsmethoden stellt diese Vereinfachung eine erhebliche Ersparnis im Gesamtrechenaufwand dar, wie sich im nächsten Kapitel zeigen wird; so reduziert sich beispielsweise in zwei Raumdimensionen die Anzahl der Primärvariablen in der Finite-Elemente-Formulierung von sieben auf drei.

1.6.1 Dynamische Formulierung

Das dynamische Zweiphasenmodell (1.98) in den Primärvariablen ($\mathbf{u}_S$, $\mathbf{w}_F$, p) ist ein System quasi-linearer partieller Differentialgleichungen von zweiter Ordnung in Ort und Zeit. Die Gleichungen 1–6 aus Kasten (1.96) wurden so umgestellt, daß alle Zeitableitungen auf der linken Seite stehen. Der Übersichtlichkeit halber sind die Gleichungen 1 und 2 zur Berechnung der Volumenanteile nicht überall eingesetzt, obwohl die Volumenanteile selbst nicht zu den Primärvariablen zählen. Außer den Materialparametern ist in (1.98) nur noch die Extraspannung $\boldsymbol{\sigma}_E^S$ des Festkörpers unbestimmt. Als sekundäre Variable kann diese

Dynamische Formulierung in den Variablen $\mathbf{u}_S$, $\mathbf{w}_F$, p

$$
\begin{aligned}
n^S\rho^{SR}\,(\mathbf{u}_S)''_S + n^F\rho^{FR}\,[(\mathbf{u}_S)''_S \;+\; & (\mathbf{w}_F)'_S + \operatorname{grad}((\mathbf{u}_S)'_S + \mathbf{w}_F)\,\mathbf{w}_F] \\
&= \operatorname{div}(\boldsymbol{\sigma}^S_E - p\,\mathbf{I}) + (n^S\rho^{SR} + n^F\rho^{FR})\,\mathbf{b} \\
n^F\rho^{FR}\,[(\mathbf{u}_S)''_S \;+\; & (\mathbf{w}_F)'_S + \operatorname{grad}((\mathbf{u}_S)'_S + \mathbf{w}_F)\,\mathbf{w}_F] \\
&= -n^F \operatorname{grad} p - \frac{(n^F)^2\,\gamma^{FR}}{k^F}\,\mathbf{w}_F + n^F\rho^{FR}\,\mathbf{b} \\
\operatorname{div}(\mathbf{u}_S)'_S \;&= \operatorname{div}(-n^F\mathbf{w}_F)
\end{aligned}
\tag{1.98}
$$

zusammen mit den *internen Variablen*[9] $\boldsymbol{\varepsilon}_{S_p}$ und Λ aus den Gleichungen 7–10 von Kasten (1.96) bestimmt werden, wenn der Verschiebungszustand $\mathbf{u}_S$ bekannt ist. Diese Entkopplung der Gleichungen wird später bei der numerischen Umsetzung mit finiten Elementen zur effizienten Lösung der entstehenden nichtlinearen Gleichungssysteme ausgenutzt.

1.6.2 Quasi-statische Formulierung

Bei langsam veränderlichen Prozessen können die Beschleunigungsterme in den Impulsbilanzen vernachlässigt werden. Die quasi-statische Impulsbilanz des Fluids $(1.96)_4$ kann

Quasi-statische Verschiebungs-Druck-Formulierung

$$
\begin{aligned}
\mathbf{0} \;&= \operatorname{div}\left(\boldsymbol{\sigma}^S_E - p\,\mathbf{I}\right) + \left(n^S\rho^{SR} + n^F\rho^{FR}\right)\mathbf{b} \\
\operatorname{div}(\mathbf{u}_S)'_S \;&= \operatorname{div}\Big[\frac{k^F}{\gamma^{FR}}\left(\operatorname{grad} p - \rho^{FR}\,\mathbf{b}\right)\Big].
\end{aligned}
\tag{1.99}
$$

man in diesem Fall nach der Filtergeschwindigkeit $n^F\mathbf{w}_F$ auflösen und in Gleichung $(1.96)_3$ einsetzen, weshalb die Sickergeschwindigkeit $\mathbf{w}_F$ nicht mehr als Primärvariable benötigt wird. Damit gelangt man zur quasi-statischen Verschiebungs-Druck-Formulierung des Zweiphasenmodells (1.99) in den freien Variablen $\mathbf{u}_S$ und p. Im Gegensatz zur dynamischen Formulierung ist dies ein System von erster Ordnung in der Zeit.

[9] Interne Variablen im Sinne der Thermodynamik sind Bestandteil einer Konstitutivgleichung, jedoch nicht primäre Variablen des thermodynamischen Prozesses.

Kapitel 2:
Ortsdiskretisierung und Finite Elemente

2.1 Grundlagen

2.1.1 Rechenregeln aus der Vektoranalysis

Für die Herleitung von schwachen Formulierungen werden einige Rechenregeln aus der Vektoranalysis gebraucht. Dabei werden jeweils die benötigten Stetigkeitsanforderungen der Funktionen sowie Regularitätseigenschaften des betrachteten Gebiets vorausgesetzt. Für ein Skalarfeld $\alpha(\mathbf{x},t)$, ein Vektorfeld $\mathbf{v}(\mathbf{x},t)$ und ein Tensorfeld $\mathbf{T}(\mathbf{x},t)$ gelten die folgenden Produktregeln für den Divergenz-Operator:

$$\begin{aligned} \operatorname{div}(\alpha\mathbf{v}) &= \operatorname{grad}\alpha\cdot\mathbf{v} + \alpha\ \operatorname{div}\mathbf{v}\,, \\ \operatorname{div}(\mathbf{T}^T\mathbf{v}) &= \operatorname{grad}\mathbf{v}\cdot\mathbf{T} + \mathbf{v}\cdot\operatorname{div}\mathbf{T}\,. \end{aligned} \tag{2.1}$$

Der *Integralsatz von Gauß* erlaubt die Transformation eines Volumenintegrals über das Gebiet Ω auf die Oberfläche $\Gamma = \partial\Omega$ mit dem Normalenvektor $\mathbf{n}$:

$$\int_\Omega \operatorname{div}\mathbf{v}\,\mathrm{d}v = \int_\Gamma \mathbf{v}\cdot\mathbf{n}\,\mathrm{d}a\,. \tag{2.2}$$

In bezug auf die schwache Formulierung sind Kombinationen aus den obigen Produktregeln und dem *Gauß*schen Integralsatz von Bedeutung (*partielle Integration*):

$$\begin{aligned} \int_\Omega \alpha\ \operatorname{div}\mathbf{v}\,\mathrm{d}v &= \int_\Gamma \alpha\,\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}a - \int_\Omega \operatorname{grad}\alpha\cdot\mathbf{v}\,\mathrm{d}v\,, \\ \int_\Omega \mathbf{v}\cdot\operatorname{div}\mathbf{T}\,\mathrm{d}v &= \int_\Gamma \mathbf{v}\cdot\mathbf{T}\,\mathbf{n}\,\mathrm{d}a - \int_\Omega \operatorname{grad}\mathbf{v}\cdot\mathbf{T}\,\mathrm{d}v\,. \end{aligned} \tag{2.3}$$

Im Oberflächenintegral der zweiten Gleichung wurde $(\mathbf{T}^T\mathbf{v})\cdot\mathbf{n} = \mathbf{v}\cdot\mathbf{T}\,\mathbf{n}$ ausgenutzt. Dies erlaubt später die Identifikation von $\mathbf{T}\,\mathbf{n}$ auf dem Rand mit dem Spannungsvektor $\mathbf{t}$ (*Cauchy-Theorem*).

2.1.2 *Sobolev*-Räume

Bei der schwachen Formulierung wird außerdem noch der Begriff des *Sobolev*-Raums benötigt, der eng mit dem Begriff der schwachen Ableitung verknüpft ist. Für weitere Details sei an dieser Stelle auf die Literatur zur mathematischen Theorie finiter Elemente

verwiesen (*Strang & Fix* [109], *Oden & Reddy* [90], *Ciarlet* [31], *Brezzi & Fortin* [28], *Brenner & Scott* [27], *Braess* [25]). Zunächst wird definiert:

Def. 2.1: Eine Funktion $f : \Omega \longrightarrow \mathbb{R}$ auf einer Menge $\Omega \subset \mathbb{R}^d$ heißt *quadratintegrierbar*, wenn das (*Lebesgue*-)Integral des Quadrats der Funktion beschränkt ist:

$$\int_\Omega f^2 \, \mathrm{d}v < \infty \,. \tag{2.4}$$

Der Raum aller quadratintegrierbaren Funktionen über Ω wird mit $L^2(\Omega)$ bezeichnet. Mit dem Skalarprodukt zweier Funktionen $f, g \in L^2(\Omega)$,

$$(f, g) = \int_\Omega f \, g \, \mathrm{d}v \,, \tag{2.5}$$

und der dadurch induzierten Norm

$$\|f\|_0 = \|f\|_{L^2} = \sqrt{(f, f)} = \left(\int_\Omega f^2 \, \mathrm{d}v \right)^{1/2} \tag{2.6}$$

ist der Vektorraum $L^2(\Omega)$ ein vollständiger normierter Raum mit Skalarprodukt, also ein *Hilbert*-Raum. ■

In bezug auf die Mechanik spielen die quadratintegrierbaren Funktionen eine herausragende Rolle. Tritt nämlich das Integral in obiger Definition im Rahmen eines Energiefunktionals auf, so beschreiben die L^2-Funktionen gerade diejenigen Lösungen, deren Energie beschränkt bleibt, was eine für physikalisch sinnvolle Lösungen wichtige Forderung ist.

Def. 2.2: Für ganzzahliges $m \geq 0$ wird die Menge aller Funktionen $f \in L^2(\Omega)$, deren Ableitungen bis zur Ordnung m quadratintegrierbar sind, mit $H^m(\Omega)$ bezeichnet. Die *schwache Ableitung* zu einem ganzzahligen Multiindex α der Länge $|\alpha| \leq m$ wird mit $\partial^\alpha f$ bezeichnet. Zusammen mit dem Skalarprodukt

$$(f, g)_m = \sum_{|\alpha| \leq m} (\partial^\alpha f, \partial^\alpha g) \tag{2.7}$$

und der dadurch induzierten Norm (*Sobolev-Norm*)

$$\|f\|_m = \|f\|_{H^m} = \sqrt{(f, f)_m} = \left(\sum_{|\alpha| \leq m} \|\partial^\alpha f\|^2_{L^2} \right)^{1/2} \tag{2.8}$$

ist der *Sobolev-Raum* $H^m(\Omega)$ ein *Hilbert*-Raum. Der Unterraum

$$H^m_0(\Omega) = \{ \, f \in H^m(\Omega) \; : \; f = 0 \text{ auf } \partial\Omega \, \} \; \subset \; H^m(\Omega) \,, \tag{2.9}$$

dessen Funktionen auf dem Rand des Gebietes Ω verschwinden, liegt *dicht* in $H^m(\Omega)$, d. h. jede Funktion aus $H^m(\Omega)$ kann beliebig genau durch Funktionen aus $H^m_0(\Omega)$ approximiert werden. ■

Man erkennt durch Vergleich der beiden Definitionen, daß $H^0(\Omega) = L^2(\Omega)$ ist. Im Zusammenhang mit finiten Elementen spielen vor allem der *Sobolev*-Raum $H^1(\Omega)$ und dessen

Unterräume eine Rolle. Die Funktionen in $H^1(\Omega)$ sind nämlich genau diejenigen stetigen Funktionen, deren erste Ableitungen noch im schwachen Sinne existieren. Bei finiten Elementen werden üblicherweise stetige, stückweise polynomiale Funktionen betrachtet, die an Elementkanten unstetige, aber noch quadratintegrierbare Ableitungen besitzen. Die mit diesen Funktionen gebildeten Finite-Elemente-Räume sind Unterräume von $H^1(\Omega)$.

2.1.3 Notation

Zur Unterscheidung der in Kapitel 1 verwendeten Notation der Vektor- und Tensorrechnung und der in diesem und folgenden Kapiteln benötigten Matrizenrechnung werden die folgenden Konventionen getroffen.

Gemäß Tabelle 1.1 auf Seite 12 werden Symbole der Vektor- und Tensorrechnung mit fetten, geraden Buchstaben bezeichnet (z. B. $\mathbf{u}, \mathbf{x}, \mathbf{T}$).

Bei der allgemeingültigen Darstellung der Ortsdiskretisierung mit finiten Elementen in Abschnitt 2.4 und der dabei eingeführten abstrakten Formulierung der behandelten Anfangs-Randwertprobleme werden Vektoren und Matrizen mit fetten, serifenlosen Buchstaben bezeichnet (z. B. $\mathsf{u}, \mathsf{q}, \mathsf{A}$). Dabei handelt es sich um „kleine“ Vektoren und Matrizen, deren Dimension durch die Anzahl der Freiheitsgrade bzw. die Anzahl der internen Variablen des kontinuierlichen Problems bestimmt ist.

Durch Zusammenfassung aller Koeffizienten der Ortsdiskretisierung entstehen in Abschnitt 2.5 „große“ Vektoren und Matrizen, deren Dimension durch die Anzahl der Freiheitsgrade an allen FE-Knoten bzw. die Anzahl der internen Variablen an allen Integrationspunkten des semidiskreten Gesamtproblems bestimmt ist. Diese Vektoren und Matrizen werden mit fetten, schräggestellten Buchstaben bezeichnet (z. B. $\boldsymbol{u}, \boldsymbol{q}, \boldsymbol{A}$).

2.2 Das quasi-statische Anfangs-Randwertproblem

Das in Abschnitt 1.6.2 angegebene quasi-statische Modell in der kontinuumsmechanischen Formulierung wird jetzt als mathematisches Anfangs-Randwertproblem mit allen Abhängigkeiten, Anfangs- und Randbedingungen formuliert, zunächst in starker und dann in schwacher Form. Dabei wird im folgenden nur noch das viskoplastische Modell be-

Das quasi-statische Anfangs-Randwertproblem

Primäre Variablen: $\mathbf{x} \in \Omega, \quad t \in [0,T]$

$$\mathbf{u}_S(\mathbf{x},t), \qquad p(\mathbf{x},t)$$

Sekundäre Variablen:

$$\begin{aligned}
n^S(\mathbf{x},t) &= n^S_{0S}(1 - \operatorname{div}\mathbf{u}_S(\mathbf{x},t)) \\
n^F(\mathbf{x},t) &= 1 - n^S(\mathbf{x},t) \\
(\rho^S + \rho^F)(\mathbf{x},t) &= n^S(\mathbf{x},t)\,\rho^{SR}(\mathbf{x}) + n^F(\mathbf{x},t)\,\rho^{FR}(\mathbf{x}) \\
\boldsymbol{\varepsilon}_S(\mathbf{x},t) &= \tfrac{1}{2}\left(\operatorname{grad}\mathbf{u}_S(\mathbf{x},t) + \operatorname{grad}^T\mathbf{u}_S(\mathbf{x},t)\right) \\
\boldsymbol{\varepsilon}_{Se}(\mathbf{x},t) &= \boldsymbol{\varepsilon}_S(\mathbf{x},t) - \boldsymbol{\varepsilon}_{Sp}(\mathbf{x},t) \\
\boldsymbol{\sigma}^S_E(\mathbf{x},t) &= 2\mu^S\boldsymbol{\varepsilon}_{Se}(\mathbf{x},t) + \lambda^S\ \operatorname{tr}\boldsymbol{\varepsilon}_{Se}(\mathbf{x},t)\,\mathbf{I} =: \overset{4}{\mathbf{C}}\,\boldsymbol{\varepsilon}_{Se}
\end{aligned}$$

Bilanzgleichungen:

$$\begin{aligned}
\mathbf{0} &= \operatorname{div}(\boldsymbol{\sigma}^S_E - p\,\mathbf{I}) + (\rho^S + \rho^F)\,\mathbf{b} \\
\operatorname{div}(\mathbf{u}_S)'_S &= \operatorname{div}\Big[\frac{k^F}{\gamma^{FR}}\left(\operatorname{grad}p - \rho^{FR}\,\mathbf{b}\right)\Big]
\end{aligned}$$

Plastische Entwicklungsgleichungen:

$$\begin{aligned}
(\boldsymbol{\varepsilon}_{Sp})'_S &= \Lambda\frac{\partial G}{\partial\boldsymbol{\sigma}^S_E} \\
\Lambda &= \frac{1}{\eta}\left\langle\frac{F(\boldsymbol{\sigma}^S_E)}{\sigma_0}\right\rangle^r
\end{aligned}$$

(2.10)

Anfangsbedingungen: $\mathbf{x} \in \Omega$

$$\begin{aligned}
\mathbf{u}_S(\mathbf{x},0) &= \mathbf{u}_{S0}(\mathbf{x}) \\
p(\mathbf{x},0) &= p_0(\mathbf{x}) \\
\boldsymbol{\varepsilon}_{Sp}(\mathbf{x},0) &= \mathbf{0} \\
\Lambda(\mathbf{x},0) &= 0
\end{aligned}$$

Randbedingungen:

$$\begin{aligned}
\mathbf{u}_S(\mathbf{x},t) &= \bar{\mathbf{u}}_S(\mathbf{x},t) && \mathbf{x} \in \Gamma_{\mathrm{u}}, \quad t \in [0,T] \\
p(\mathbf{x},t) &= \bar{p}(\mathbf{x},t) && \mathbf{x} \in \Gamma_p, \quad t \in [0,T] \\
(\boldsymbol{\sigma}\,\mathbf{n})(\mathbf{x},t) = \mathbf{t}(\mathbf{x},t) &= \bar{\mathbf{t}}(\mathbf{x},t) && \mathbf{x} \in \Gamma_{\mathrm{t}}, \quad t \in [0,T] \\
(n^F\mathbf{w}_F \cdot \mathbf{n})(\mathbf{x},t) = v(\mathbf{x},t) &= \bar{v}(\mathbf{x},t) && \mathbf{x} \in \Gamma_v, \quad t \in [0,T] \\
\partial\Omega = \Gamma &= \Gamma_{\mathrm{u}} \cup \Gamma_{\mathrm{t}} = \Gamma_p \cup \Gamma_v \\
\emptyset &= \Gamma_{\mathrm{u}} \cap \Gamma_{\mathrm{t}} = \Gamma_p \cap \Gamma_v
\end{aligned}$$

trachtet; ein Übergang zum elastoplastischen Modell ist stets durch Austausch der Gleichung 10a durch Gleichung 10b in Kasten (1.96) möglich. Wegen der geometrischen Linearität (keine Unterscheidung von Referenz- und Momentankonfiguration) werden alle Ortskoordinaten mit $\mathbf{x}$ und das d-dimensionale Rechengebiet mit $\Omega \subset \mathbb{R}^d$ bezeichnet.

2.2.1 Starke Formulierung

Die *starke Formulierung* des quasi-statischen Anfangs-Randwertproblems ist in Kasten (2.10) zusammengefaßt. In den Gleichungen wurde aus Gründen der Übersichtlichkeit auf die explizite Notierung der Orts- und Zeitabhängigkeiten verzichtet. Bei den Randbedingungen bezeichnen $\bar{\mathbf{u}}_S$ und $\bar{p}$ die *Dirichlet-* oder *essentiellen Randbedingungen* von Verschiebung und Druck, $\bar{\mathbf{t}}$ den *Cauchy*schen Spannungsvektor, der auf die Oberfläche der gesamten Mischung wirkt ($\boldsymbol{\sigma} = -p\,\mathbf{I} + \boldsymbol{\sigma}_E^S$ ist der *Cauchy*sche Spannungstensor der Mischung), und $\bar{v}$ den Volumenstrom des Fluids über die Oberfläche (*Neumann-* oder *natürliche Randbedingungen*).

2.2.2 Schwache Formulierung

Zunächst werden für die Verschiebung $\mathbf{u}_S$ und den Druck p die beiden verschobenen *Sobolev*-Räume (auch: affine *Sobolev*-Räume)

$$\begin{aligned} \mathcal{S}_{\mathrm{u}}(t) &= \{\, \mathbf{u}_S \in H^1(\Omega)^d \;:\; \mathbf{u}_S(\mathbf{x}) = \bar{\mathbf{u}}_S(\mathbf{x},t) \text{ auf } \Gamma_{\mathrm{u}} \,\} \subset H^1(\Omega)^d\,, \\ \mathcal{S}_p(t) &= \{\, p \in H^1(\Omega) \;:\; p(\mathbf{x}) = \bar{p}(\mathbf{x},t) \text{ auf } \Gamma_p \,\} \subset H^1(\Omega) \end{aligned} \tag{2.11}$$

eingeführt, deren *Ansatzfunktionen* jeweils die *Dirichlet*-Randbedingungen des Problems (2.10) erfüllen. Die zugehörigen *Testfunktionen*[1] erfüllen homogene *Dirichlet*-Randbedingungen und liegen daher in den in $\mathcal{S}_{\mathrm{u}}(t)$ bzw. $\mathcal{S}_p(t)$ dicht liegenden *Sobolev*-Räumen

$$\begin{aligned} \mathcal{T}_{\mathrm{u}} &= \{\, \delta\mathbf{u}_S \in H^1(\Omega)^d \;:\; \delta\mathbf{u}_S(\mathbf{x}) = \mathbf{0} \text{ auf } \Gamma_{\mathrm{u}} \,\}\,, \\ \mathcal{T}_p &= \{\, \delta p \in H^1(\Omega) \;:\; \delta p(\mathbf{x}) = 0 \text{ auf } \Gamma_p \,\}\,. \end{aligned} \tag{2.12}$$

Zur Herleitung der schwachen Formulierung werden zunächst die Bilanzgleichungen in (2.10) skalar mit den Testfunktionen $\delta\mathbf{u}_S \in \mathcal{T}_{\mathrm{u}}$ und $\delta p \in \mathcal{T}_p$ multipliziert und über das Gebiet Ω integriert:

$$\begin{aligned} 0 &= \int_\Omega \delta\mathbf{u}_S \cdot \big(\operatorname{div}(\boldsymbol{\sigma}_E^S - p\,\mathbf{I}) + (\rho^S + \rho^F)\,\mathbf{b}\big)\,\mathrm{d}v\,, \\ \int_\Omega \delta p\;\operatorname{div}(\mathbf{u}_S)'_S\,\mathrm{d}v &= \int_\Omega \delta p\;\operatorname{div}\Big[\frac{k^F}{\gamma^{FR}}\left(\operatorname{grad} p - \rho^{FR}\,\mathbf{b}\right)\Big]\,\mathrm{d}v\,. \end{aligned} \tag{2.13}$$

[1] In der Literatur werden die Testfunktionen auch als *virtuelle Verrückungen* (*virtuelle Verschiebungen*) und die schwache Formulierung als das *Prinzip der virtuellen Verrückungen* bezeichnet.

Anwendung der Formeln (2.3) auf die Divergenz-Terme der rechten Seite und Berücksichtigung der homogenen *Dirichlet*-Randbedingungen der Testfunktionen liefert die *schwache Formulierung* des quasi-statischen Anfangs-Randwertproblems.

Def. 2.3: Eine Funktionenschar $(\mathbf{u}_S(t),\, p(t))$, $0 \le t \le T$ mit $\mathbf{u}_S(t) \in \mathcal{S}_{\mathrm{u}}(t)$ und $p(t) \in \mathcal{S}_p(t)$ heißt *schwache Lösung* von (2.10), wenn für alle Zeiten $t \in [0,T]$ die beiden Gleichungen

$$\begin{aligned} \int_\Omega \operatorname{grad}\delta\mathbf{u}_S \cdot (\boldsymbol{\sigma}_E^S - p\,\mathbf{I})\,\mathrm{d}v &= \int_\Omega \delta\mathbf{u}_S \cdot (\rho^S + \rho^F)\,\mathbf{b}\,\mathrm{d}v \\ &\quad + \int_{\Gamma_{\mathrm{t}}} \delta\mathbf{u}_S \cdot \bar{\mathbf{t}}\,\mathrm{d}a\,, \\ \int_\Omega \delta p\ \operatorname{div}(\mathbf{u}_S)'_S\,\mathrm{d}v + \int_\Omega \operatorname{grad}\delta p \cdot \frac{k^F}{\gamma^{FR}}\operatorname{grad} p\,\mathrm{d}v &= \int_\Omega \operatorname{grad}\delta p \cdot \frac{k^F}{\gamma^{FR}}\,\rho^{FR}\,\mathbf{b}\,\mathrm{d}v \\ &\quad - \int_{\Gamma_v} \delta p\,\bar{v}\,\mathrm{d}a \end{aligned} \tag{2.14}$$

sowie für $t = 0$ die Anfangsbedingungen

$$\int_\Omega \delta\mathbf{u}_S \cdot (\mathbf{u}_S(0) - \mathbf{u}_{S0})\,\mathrm{d}v \;=\; 0\,, \qquad \int_\Omega \delta p\,(p(0) - p_0)\,\mathrm{d}v \;=\; 0 \tag{2.15}$$

jeweils für beliebige Testfunktionen $\delta\mathbf{u}_S \in \mathcal{T}_{\mathrm{u}}$ und $\delta p \in \mathcal{T}_p$ erfüllt sind und gleichzeitig die plastischen Entwicklungsgleichungen aus Kasten (2.10) sowie deren Anfangsbedingungen in starker Form gelten. ∎

Bemerkung: Im rein elastischen Fall mit $\boldsymbol{\sigma}_E^S = \overset{4}{\mathbf{C}}\,\boldsymbol{\varepsilon}_S = 2\mu^S\boldsymbol{\varepsilon}_S + \lambda^S\,\operatorname{tr}\boldsymbol{\varepsilon}_S\,\mathbf{I}$ und dem symmetrischen Gradienten-Operator $\boldsymbol{L} = \frac{1}{2}(\operatorname{grad} + \operatorname{grad}^T)$ kann man das Problem durch Einführung der symmetrischen Bilinearformen

$$\begin{aligned} a(\mathbf{v},\mathbf{w}) &= \int_\Omega \boldsymbol{L}\,\mathbf{v} \cdot \overset{4}{\mathbf{C}}\,\boldsymbol{L}\,\mathbf{w}\,\mathrm{d}v \qquad && \text{für } \mathbf{v},\mathbf{w} \in H^1(\Omega)^d\,, \\ b(p,q) &= \int_\Omega \operatorname{grad} p \cdot \frac{k^F}{\gamma^{FR}}\operatorname{grad} q\,\mathrm{d}v \qquad && \text{für } p,q \in H^1(\Omega) \end{aligned} \tag{2.16}$$

auch wie folgt formulieren: Gesucht ist eine Funktionenschar $(\mathbf{u}_S(t),\, p(t))$, $0 \le t \le T$ mit $\mathbf{u}_S(t) \in \mathcal{S}_{\mathrm{u}}(t)$ und $p(t) \in \mathcal{S}_p(t)$, so daß die Gleichungen

$$\begin{aligned} a(\delta\mathbf{u}_S,\,\mathbf{u}_S) - (\operatorname{div}\delta\mathbf{u}_S,\,p) &= (\delta\mathbf{u}_S,\,(\rho^S+\rho^F)\,\mathbf{b}) \quad + \int_{\Gamma_{\mathrm{t}}} \delta\mathbf{u}_S \cdot \bar{\mathbf{t}}\,\mathrm{d}a\,, \\ (\delta p,\,\operatorname{div}(\mathbf{u}_S)'_S) + \quad b(\delta p,\,p) &= (\operatorname{grad}\delta p,\,\frac{k^F}{\gamma^{FR}}\,\rho^{FR}\,\mathbf{b}) - \int_{\Gamma_v} \delta p\,\bar{v}\,\mathrm{d}a \end{aligned} \tag{2.17}$$

sowie die Anfangsbedingungen

$$(\delta\mathbf{u}_S,\ \mathbf{u}_S(0)) \;=\; (\delta\mathbf{u}_S,\ \mathbf{u}_{S0})\,, \qquad (\delta p,\ p(0)) \;=\; (\delta p,\ p_0) \tag{2.18}$$

für beliebige Testfunktionen $\delta\mathbf{u}_S \in \mathcal{T}_{\mathrm{u}}$ und $\delta p \in \mathcal{T}_p$ erfüllt sind. Darin ist $(\cdot,\cdot)$ das L^2-Skalarprodukt bzw. dessen natürliche Erweiterung auf vektorwertige Funktionen. □

2.3 Das dynamische Anfangs-Randwertproblem

Wie zuvor beim quasi-statischen Modell wird hier das in Abschnitt 1.6.1 formulierte dynamische Zweiphasenmodell als mathematisches Anfangs-Randwertproblem in starker und in schwacher Form angegeben. Aufgrund der geometrischen Linearität (keine Unterscheidung von Referenz- und Momentankonfiguration) werden wieder alle Ortskoordinaten mit $\mathbf{x}$ und das d-dimensionale Rechengebiet mit $\Omega \subset \mathbb{R}^d$ bezeichnet.

2.3.1 Starke Formulierung

Für das dynamische Modell sind die primären Variablen die Verschiebung $\mathbf{u}_S$, die Sickergeschwindigkeit $\mathbf{w}_F$ und der Druck p (vgl. Kasten (1.98) auf Seite 36). Die *starke Formulierung* des dynamischen Anfangs-Randwertproblems ist in Kasten (2.19) zusammengefaßt. In den Gleichungen wurde aus Gründen der Übersichtlichkeit auf die explizite Notierung der Orts- und Zeitabhängigkeiten verzichtet. Bei den Randbedingungen bezeichnen $\bar{\mathbf{u}}_S$, $\bar{\mathbf{w}}_F$ und $\bar{p}$ die *Dirichlet*-Randbedingungen von Verschiebung, Sickergeschwindigkeit und Druck (*Dirichlet*-Ränder $\Gamma_{\mathrm{u}}, \Gamma_{\mathrm{w}}, \Gamma_p$), $\bar{\mathbf{t}}$ den *Cauchy*schen Spannungsvektor, der auf die Oberfläche der gesamten Mischung wirkt (*Neumann*-Rand Γ_{t}), $\bar{\mathbf{t}}^F$ den *Cauchy*schen Spannungsvektor, der nur auf das Fluid wirkt (*Neumann*-Rand Γ_{t^F}), und $\bar{v}$ den Volumenstrom des Fluids über die Oberfläche (*Neumann*-Rand Γ_v). Wie zuvor beim quasi-statischen Modell ist dabei $\boldsymbol{\sigma} = -p\,\mathbf{I} + \boldsymbol{\sigma}_E^S$ der *Cauchy*sche Spannungstensor der Mischung.

Bemerkung: Die Randbedingungen sind i. a. nicht unabhängig voneinander wählbar. So muß auf überlappenden Randbereichen $\Gamma_{\mathrm{w}} \cap \Gamma_v$ die Vorgabe der Sickergeschwindigkeit (*Dirichlet*-Randbedingung Impulsbilanz Fluid) zur Vorgabe des Volumenstroms (*Neumann*-Randbedingung Volumenbilanz) kompatibel sein, d. h. $\bar{v} = n^F \bar{\mathbf{w}}_F \cdot \mathbf{n}$. Außerdem muß auf $\Gamma_p \cap \Gamma_{\mathrm{t}^F}$ die Vorgabe des Drucks (*Dirichlet*-Randbedingung Volumenbilanz) zur Vorgabe der Fluidspannung (*Neumann*-Randbedingung Impulsbilanz Fluid) kompatibel sein: $\bar{\mathbf{t}}^F = (-n^F \bar{p}\,\mathbf{I})\,\mathbf{n} = -n^F \bar{p}\,\mathbf{n}$. Die Lösungsabhängigkeit der Porosität $n^F(\mathbf{u}_S)$ auf dem Rand wird hier wegen der Betrachtung geometrisch linearer Probleme (kleine Deformationen) vernachlässigt; bei großen Deformationen erhält man an dieser Stelle nichtlineare *Dirichlet*- und *Neumann*-Randbedingungen. □

Das dynamische Anfangs-Randwertproblem

Primäre Variablen: $\mathbf{x} \in \Omega, \quad t \in [0,T]$

$$\mathbf{u}_S(\mathbf{x},t), \quad \mathbf{w}_F(\mathbf{x},t), \quad p(\mathbf{x},t)$$

Sekundäre Variablen:

$$\begin{aligned}
\mathbf{v}_S(\mathbf{x},t) &= (\mathbf{u}_S(\mathbf{x},t))'_S \\
n^S(\mathbf{x},t) &= n^S_{0S}(1 - \operatorname{div}\mathbf{u}_S(\mathbf{x},t)) \\
n^F(\mathbf{x},t) &= 1 - n^S(\mathbf{x},t) \\
(\rho^S + \rho^F)(\mathbf{x},t) &= n^S(\mathbf{x},t)\,\rho^{SR}(\mathbf{x}) + n^F(\mathbf{x},t)\,\rho^{FR}(\mathbf{x}) \\
\boldsymbol{\varepsilon}_S(\mathbf{x},t) &= \tfrac{1}{2}\left(\operatorname{grad}\mathbf{u}_S(\mathbf{x},t) + \operatorname{grad}^T\mathbf{u}_S(\mathbf{x},t)\right) \\
\boldsymbol{\varepsilon}_{Se}(\mathbf{x},t) &= \boldsymbol{\varepsilon}_S(\mathbf{x},t) - \boldsymbol{\varepsilon}_{Sp}(\mathbf{x},t) \\
\boldsymbol{\sigma}^S_E(\mathbf{x},t) &= 2\mu^S\boldsymbol{\varepsilon}_{Se}(\mathbf{x},t) + \lambda^S\ \operatorname{tr}\boldsymbol{\varepsilon}_{Se}(\mathbf{x},t)\,\mathbf{I} \\
\mathbf{c}(\mathbf{x},t) &= \operatorname{grad}(\mathbf{v}_S(\mathbf{x},t) + \mathbf{w}_F(\mathbf{x},t))\,\mathbf{w}_F(\mathbf{x},t)
\end{aligned}$$

Bilanzgleichungen:

$$\begin{aligned}
\rho^S\,(\mathbf{u}_S)''_S + \rho^F\,[((\mathbf{u}_S)'_S + \mathbf{w}_F)'_S + \mathbf{c}] &= \operatorname{div}(\boldsymbol{\sigma}^S_E - p\,\mathbf{I}) + (\rho^S + \rho^F)\,\mathbf{b} \\
\rho^F\,[((\mathbf{u}_S)'_S + \mathbf{w}_F)'_S + \mathbf{c}] &= -n^F \operatorname{grad} p - \frac{(n^F)^2\,\gamma^{FR}}{k^F}\,\mathbf{w}_F + \rho^F\,\mathbf{b} \\
\operatorname{div}(\mathbf{u}_S)'_S &= \operatorname{div}\left(-n^F\mathbf{w}_F\right)
\end{aligned}$$

Plastische Entwicklungsgleichungen:

$$\begin{aligned}
(\boldsymbol{\varepsilon}_{Sp})'_S &= \Lambda\frac{\partial G}{\partial \boldsymbol{\sigma}^S_E} \\
\Lambda &= \frac{1}{\eta}\left\langle\frac{F(\boldsymbol{\sigma}^S_E)}{\sigma_0}\right\rangle^r
\end{aligned} \tag{2.19}$$

Anfangsbedingungen: $\mathbf{x} \in \Omega$

$$\begin{aligned}
\mathbf{u}_S(\mathbf{x},0) &= \mathbf{u}_{S0}(\mathbf{x}) \\
\mathbf{v}_S(\mathbf{x},0) &= \mathbf{v}_{S0}(\mathbf{x}) \\
\mathbf{w}_F(\mathbf{x},0) &= \mathbf{w}_{F0}(\mathbf{x}) \\
p(\mathbf{x},0) &= p_0(\mathbf{x}) \\
\boldsymbol{\varepsilon}_{Sp}(\mathbf{x},0) &= \mathbf{0} \\
\Lambda(\mathbf{x},0) &= 0
\end{aligned}$$

Randbedingungen:

$$\begin{aligned}
\mathbf{u}_S(\mathbf{x},t) &= \bar{\mathbf{u}}_S(\mathbf{x},t) && \mathbf{x} \in \Gamma_{\mathrm{u}}, \quad t \in [0,T] \\
\mathbf{w}_F(\mathbf{x},t) &= \bar{\mathbf{w}}_F(\mathbf{x},t) && \mathbf{x} \in \Gamma_{\mathrm{w}}, \quad t \in [0,T] \\
p(\mathbf{x},t) &= \bar{p}(\mathbf{x},t) && \mathbf{x} \in \Gamma_p, \quad t \in [0,T] \\
(\boldsymbol{\sigma}\,\mathbf{n})(\mathbf{x},t) = \mathbf{t}(\mathbf{x},t) &= \bar{\mathbf{t}}(\mathbf{x},t) && \mathbf{x} \in \Gamma_{\mathrm{t}}, \quad t \in [0,T] \\
(\boldsymbol{\sigma}^F\,\mathbf{n})(\mathbf{x},t) = \mathbf{t}^F(\mathbf{x},t) &= \bar{\mathbf{t}}^F(\mathbf{x},t) && \mathbf{x} \in \Gamma_{\mathrm{t}^F}, \quad t \in [0,T] \\
(n^F\mathbf{w}_F \cdot \mathbf{n})(\mathbf{x},t) = v(\mathbf{x},t) &= \bar{v}(\mathbf{x},t) && \mathbf{x} \in \Gamma_v, \quad t \in [0,T]
\end{aligned}$$

$$\partial\Omega = \Gamma = \Gamma_{\mathrm{u}} \cup \Gamma_{\mathrm{t}} = \Gamma_{\mathrm{w}} \cup \Gamma_{\mathrm{t}^F} = \Gamma_p \cup \Gamma_v$$

$$\emptyset = \Gamma_{\mathrm{u}} \cap \Gamma_{\mathrm{t}} = \Gamma_{\mathrm{w}} \cap \Gamma_{\mathrm{t}^F} = \Gamma_p \cap \Gamma_v$$

2.3.2 Schwache Formulierung

Analog zum quasi-statischen Anfangs-Randwertproblem werden für die Verschiebung $\mathbf{u}_S$, die Sickergeschwindigkeit $\mathbf{w}_F$ und den Druck p die verschobenen *Sobolev*-Räume

$$\begin{aligned}
\mathcal{S}_{\mathrm{u}}(t) &= \{\ \mathbf{u}_S \in H^1(\Omega)^d \ :\ \mathbf{u}_S(\mathbf{x}) = \bar{\mathbf{u}}_S(\mathbf{x},t) \ \text{ auf } \Gamma_{\mathrm{u}}\ \} \subset H^1(\Omega)^d\,, \\
\mathcal{S}_{\mathrm{w}}(t) &= \{\ \mathbf{w}_F \in H^1(\Omega)^d \ :\ \mathbf{w}_F(\mathbf{x}) = \bar{\mathbf{w}}_F(\mathbf{x},t) \ \text{ auf } \Gamma_{\mathrm{w}}\ \} \subset H^1(\Omega)^d\,, \\
\mathcal{S}_{p}(t) &= \{\ p \in H^1(\Omega) \ :\ p(\mathbf{x}) = \bar{p}(\mathbf{x},t) \ \text{ auf } \Gamma_{p}\ \} \subset H^1(\Omega)
\end{aligned} \tag{2.20}$$

eingeführt, deren Funktionen jeweils die *Dirichlet*-Randbedingungen in Kasten (2.19) erfüllen. Die zugehörigen Testfunktionen erfüllen wieder homogene *Dirichlet*-Randbedingungen und liegen daher in den in $\mathcal{S}_{\mathrm{u}}(t)$, $\mathcal{S}_{\mathrm{w}}(t)$ bzw. $\mathcal{S}_p(t)$ dicht liegenden *Sobolev*-Räumen

$$\begin{aligned}
\mathcal{T}_{\mathrm{u}} &= \{\ \delta\mathbf{u}_S \in H^1(\Omega)^d \ :\ \delta\mathbf{u}_S(\mathbf{x}) = \mathbf{0} \ \text{ auf } \Gamma_{\mathrm{u}}\ \}\,, \\
\mathcal{T}_{\mathrm{w}} &= \{\ \delta\mathbf{w}_F \in H^1(\Omega)^d \ :\ \delta\mathbf{w}_F(\mathbf{x}) = \mathbf{0} \ \text{ auf } \Gamma_{\mathrm{w}}\ \}\,, \\
\mathcal{T}_{p} &= \{\ \delta p \in H^1(\Omega) \ :\ \delta p(\mathbf{x}) = 0 \ \text{ auf } \Gamma_{p}\ \}\,.
\end{aligned} \tag{2.21}$$

Die Bilanzgleichungen in (2.19) werden erneut skalar mit Testfunktionen $\delta\mathbf{u}_S \in \mathcal{T}_{\mathrm{u}}$, $\delta\mathbf{w}_F \in \mathcal{T}_{\mathrm{w}}$ und $\delta p \in \mathcal{T}_p$ multipliziert und über das Gebiet Ω integriert. Anschließend liefert die Anwendung der Formeln (2.3) auf die Divergenzterme – nach Einsetzen der Identität $\operatorname{grad} p = \operatorname{div}(p\,\mathbf{I})$ in die Impulsbilanz des Fluids – die *schwache Formulierung* des dynamischen Anfangs-Randwertproblems.

Def. 2.4: Eine Funktionenschar $(\mathbf{u}_S(t),\, \mathbf{w}_F(t),\, p(t))$, $0 \le t \le T$ mit $\mathbf{u}_S(t) \in \mathcal{S}_{\mathrm{u}}(t)$, $\mathbf{w}_F(t) \in \mathcal{S}_{\mathrm{w}}(t)$ und $p(t) \in \mathcal{S}_p(t)$ heißt *schwache Lösung* von (2.19), wenn für alle Zeiten $t \in [0,T]$ die Gleichungen

$$\begin{aligned}
0 &= \int_\Omega \delta\mathbf{u}_S \cdot \Big(\rho^S\,[(\mathbf{u}_S)'_S{}' - \mathbf{b}] + \rho^F\,[((\mathbf{u}_S)'_S + \mathbf{w}_F)'_S + \operatorname{grad}((\mathbf{u}_S)'_S + \mathbf{w}_F)\,\mathbf{w}_F - \mathbf{b}]\Big)\,\mathrm{d}v \\
&\quad + \int_\Omega \operatorname{grad}\delta\mathbf{u}_S \cdot (\boldsymbol{\sigma}^S_E - p\,\mathbf{I})\,\mathrm{d}v \;-\; \int_{\Gamma_{\mathrm{t}}} \delta\mathbf{u}_S \cdot \bar{\mathbf{t}}\,\mathrm{d}a\,, \\
0 &= \int_\Omega \delta\mathbf{w}_F \cdot \Big(\rho^F\,[((\mathbf{u}_S)'_S + \mathbf{w}_F)'_S + \operatorname{grad}((\mathbf{u}_S)'_S + \mathbf{w}_F)\,\mathbf{w}_F - \mathbf{b}] + \frac{(n^F)^2\,\gamma^{FR}}{k^F}\,\mathbf{w}_F\Big)\,\mathrm{d}v \\
&\quad - \int_\Omega (\operatorname{div}\delta\mathbf{w}_F)\; n^F p\,\mathrm{d}v \;-\; \int_{\Gamma_{\mathrm{t}F}} \delta\mathbf{w}_F \cdot \bar{\mathbf{t}}^F\,\mathrm{d}a\,, \\
0 &= \int_\Omega \delta p\,\operatorname{div}(\mathbf{u}_S)'_S\,\mathrm{d}v \;-\; \int_\Omega \operatorname{grad}\delta p \cdot n^F\mathbf{w}_F\,\mathrm{d}v \;+\; \int_{\Gamma_v} \delta p\,\bar{v}\,\mathrm{d}a
\end{aligned} \tag{2.22}$$

sowie für $t = 0$ die Anfangsbedingungen

$$\begin{aligned} \int_\Omega \delta\mathbf{u}_S \cdot (\mathbf{u}_S(0) - \mathbf{u}_{S0})\,\mathrm{d}v &= 0\,, & \int_\Omega \delta\mathbf{w}_F \cdot (\mathbf{w}_F(0) - \mathbf{w}_{F0})\,\mathrm{d}v &= 0\,, \\ \int_\Omega \delta\mathbf{u}_S \cdot ((\mathbf{u}_S)'_S(0) - \mathbf{v}_{S0})\,\mathrm{d}v &= 0\,, & \int_\Omega \delta p\,(p(0) - p_0)\,\mathrm{d}v &= 0 \end{aligned} \tag{2.23}$$

jeweils für beliebige Testfunktionen $\delta\mathbf{u}_S \in \mathcal{T}_\mathrm{u}$, $\delta\mathbf{w}_F \in \mathcal{T}_\mathrm{w}$ und $\delta p \in \mathcal{T}_p$ erfüllt sind und gleichzeitig die plastischen Entwicklungsgleichungen aus Kasten (2.19) sowie deren Anfangsbedingungen in starker Form gelten. ■

2.4 Finite Elemente

Die Ortsdiskretisierung mit der *Methode der finiten Elemente* (FEM) wird in diesem Abschnitt anhand einer allgemeingültigen Operatordarstellung einer beliebigen schwachen Formulierung vorgestellt. Dabei wird die Zeitableitung mit einem Punkt bezeichnet[2]. Die Beschreibung der Methode der finiten Elemente selbst wird hier sehr knapp gehalten. Für Details sei auf die Standardliteratur verwiesen, z. B. *Oden* [88], *Strang & Fix* [109], *Oden & Reddy* [89], *Ciarlet* [31], *Bathe* [18], *Hughes* [73], *Brezzi & Fortin* [28], *Zienkiewicz & Taylor* [126, 127], *Schwarz* [104], *Brenner & Scott* [27], *Braess* [25].

2.4.1 Operatordarstellung einer allgemeinen schwachen Formulierung

Es wird nun ein Anfangs-Randwertproblem mit D *Freiheitsgraden* (Dimension der vektorwertigen gesuchten Lösungsfunktion bzw. Anzahl der Primärvariablen) und Q internen Variablen sowie beliebigen *Dirichlet*- und *Neumann*-Randbedingungen betrachtet. Je Freiheitsgrad und Randbereich dürfen entweder *Dirichlet*- oder *Neumann*-Randbedingungen[3] vorgegeben werden, der gesamte Rand $\partial\Omega$ wird also je Freiheitsgrad d in einen *Dirichlet*-Randbereich Γ_{u_d} und einen *Neumann*-Randbereich Γ_{f_d} aufgeteilt:

$$\partial\Omega = \Gamma = \Gamma_{\mathsf{u}_d} \cup \Gamma_{\mathsf{f}_d}\,, \qquad \emptyset = \Gamma_{\mathsf{u}_d} \cap \Gamma_{\mathsf{f}_d}\,, \qquad d = 1, \ldots, \mathrm{D}\,. \tag{2.24}$$

Die *Dirichlet*-Vorgaben für die d-te Komponente u_d der gesuchten Lösung $\mathbf{u}(\mathbf{x}, t)$ werden durch Überstreichung gekennzeichnet:

$$\mathsf{u}_d(\mathbf{x}, t) = \bar{\mathsf{u}}_d(\mathbf{x}, t) \;\text{ auf }\; \Gamma_{\mathsf{u}_d}\,, \qquad d = 1, \ldots, \mathrm{D}\,. \tag{2.25}$$

[2] Eine Verwechslung mit der in Gleichung (1.23) eingeführten Zeitableitung für die Schwerpunktsgeschwindigkeit ist ausgeschlossen, da diese im jetzigen und folgenden Kapiteln nicht mehr auftaucht. In bezug auf die behandelten Zweiphasenmodelle bezeichnet der Punkt im folgenden also immer die einzig vorkommende Zeitableitung $(\cdot)'_S$, die der Festkörperbewegung folgt.

[3] Die *Neumann*-Randbedingungen werden in dieser allgemeinen Darstellung nicht explizit angegeben, da sie stark von der Art des betrachteten Anfangs-Randwertproblems abhängen und bei der Herleitung der schwachen Form (s. o.) entsprechend berücksichtigt werden müssen. Hier soll nur die Vorgehensweise der Ortsdiskretisierung – ausgehend von einer gegebenen schwachen Formulierung – dargestellt werden.

Außerdem werden wieder Ansatz- und Testräume für die schwache Formulierung benötigt:

$$\begin{aligned} \mathcal{S}(t) &= \{\ \mathbf{u} \in H^1(\Omega)^{\mathrm{D}} \ :\ \mathsf{u}_d(\mathbf{x}) = \bar{\mathsf{u}}_d(\mathbf{x},t) \ \text{ auf } \Gamma_{\mathsf{u}_d},\ \ d = 1,\ldots,\mathrm{D}\ \}\,, \\ \mathcal{T} &= \{\ \boldsymbol{\eta} \in H^1(\Omega)^{\mathrm{D}} \ :\ \eta_d(\mathbf{x}) = 0 \qquad\ \ \text{auf } \Gamma_{\mathsf{u}_d},\ \ d = 1,\ldots,\mathrm{D}\ \}\,. \end{aligned} \tag{2.26}$$

Der verschobene *Sobolev*-Raum $\mathcal{S}(t)$ mit den Ansatzfunktionen $\mathbf{u}(t) = (\mathsf{u}_1(t),\ldots,\mathsf{u}_{\mathrm{D}}(t))^T$ erfüllt die *Dirichlet*-Randbedingungen der D Freiheitsgrade auf den Randbereichen Γ_{u_d}. Der *Sobolev*-Raum $\mathcal{T}$ mit den Testfunktionen $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_{\mathrm{D}})^T$ erfüllt entsprechende homogene *Dirichlet*-Randbedingungen, d. h. die schwache Formulierung ist auf *Dirichlet*-Rändern automatisch erfüllt, da die Testfunktionen dort verschwinden.

Die internen Variablen werden zu einem Vektor $\mathbf{q}(t) = (\mathsf{q}_1(t),\ldots,\mathsf{q}_{\mathrm{Q}}(t))^T$ zusammengefaßt. Da die internen Variablen über *punktweise* im Gebiet Ω gegebene gewöhnliche Differentialgleichungen in der Zeit bestimmt werden, können sie als Funktionen des Ortes nur in einem Vektorraum $L^\infty(\Omega)$ erwartet werden. Damit kann die allgemeine schwache Formulierung in Operatordarstellung formuliert werden.

Def. 2.5: Gesucht ist eine Funktionenschar $(\mathbf{u}(t),\, \mathbf{q}(t))$, $0 \le t \le T$ mit $\mathbf{u}(t) \in \mathcal{S}(t)$ und $\mathbf{q}(t) \in (L^\infty(\Omega))^{\mathrm{Q}}$, die für alle $t \in [0,T]$ und alle $\boldsymbol{\eta} \in \mathcal{T}$ die Operatorgleichung

$$\mathcal{G}[\boldsymbol{\eta}, \mathbf{u};\, \mathbf{q}] \ \equiv\ \int_\Omega \left(\mathcal{M}(\boldsymbol{\eta}, \mathbf{u}, \ddot{\mathbf{u}}) + \mathcal{D}(\boldsymbol{\eta}, \mathbf{u}, \dot{\mathbf{u}}) + \mathcal{K}(\boldsymbol{\eta}, \mathbf{u}, \mathbf{q})\right)\, \mathrm{d}v - \int_\Gamma \mathcal{F}(\boldsymbol{\eta}, \mathbf{u})\, \mathrm{d}a \ =\ 0 \tag{2.27}$$

sowie für alle $t \in [0,T]$ die Entwicklungsgleichungen

$$\mathbf{A}\,\dot{\mathbf{q}} - \mathbf{r}(\mathbf{q}, \mathbf{u}) = \mathbf{0} \tag{2.28}$$

und außerdem für $t = 0$ und alle $\boldsymbol{\eta} \in \mathcal{T}$ die Anfangsbedingungen

$$\int_\Omega \boldsymbol{\eta} \cdot (\mathbf{u}(0) - \mathbf{u}_0)\, \mathrm{d}v = 0\,, \qquad \int_\Omega \boldsymbol{\eta} \cdot (\dot{\mathbf{u}}(0) - \mathbf{v}_0)\, \mathrm{d}v = 0\,, \qquad \mathbf{q}(0) = \mathbf{q}_0 \tag{2.29}$$

erfüllt. Darin sind $\mathcal{M}$, $\mathcal{D}$, $\mathcal{K}$ und $\mathcal{F}$ nichtlineare reellwertige Operatoren der jeweiligen Funktionen. Bei den in dieser Arbeit behandelten Problemen sind $\mathcal{M}$ und $\mathcal{D}$ linear in den Zeitableitungen $\ddot{\mathbf{u}}$ und $\dot{\mathbf{u}}$ der Freiheitsgrade. Ist die Matrix $\mathbf{A} \in \mathbb{R}^{\mathrm{Q}} \times \mathbb{R}^{\mathrm{Q}}$ regulär, so handelt es sich bei den Entwicklungsgleichungen um ein System gewöhnlicher Differentialgleichungen, andernfalls um ein System differential-algebraischer Gleichungen. ■

2.4.2 *Petrov-Galerkin*-Verfahren

Da die Ansatz- und Testräume als Funktionenräume unendliche Dimension besitzen, ist man für die Berechnung von numerischen Näherungen darauf angewiesen, Lösungen in gewissen endlichdimensionalen Teilräumen zu suchen. Man approximiert also die Ansatz- und Testräume $\mathcal{S}(t)$ und $\mathcal{T}$ durch N-dimensionale Teilräume $\mathcal{S}^h(t)$ und $\mathcal{T}^h$, wobei die

Funktion $\bar{\mathbf{u}}^h \in \mathcal{S}(t)$ näherungsweise die *Dirichlet*-Randbedingungen (2.25) erfüllt[4]:

$$\begin{aligned} \mathcal{S}^h(t) &= \{\ \mathbf{u}^h \in H^1(\Omega^h)^{\mathrm{D}} \ :\ \mathbf{u}^h(\mathbf{x},t) = \sum_{j=1}^{\mathrm{N}} \boldsymbol{\phi}^j(\mathbf{x})\,\mathbf{u}_j(t) + \bar{\mathbf{u}}^h(\mathbf{x},t)\ \}\,, \\ \mathcal{T}^h &= \{\ \boldsymbol{\eta}^h \in H^1(\Omega^h)^{\mathrm{D}} \ :\ \boldsymbol{\eta}^h(\mathbf{x}) = \sum_{i=1}^{\mathrm{N}} \boldsymbol{\psi}^i(\mathbf{x})\,\boldsymbol{\eta}_i \qquad\ \}\,. \end{aligned} \tag{2.30}$$

Der hochgestellte Index h symbolisiert einen Diskretisierungsparameter. Die matrixwertigen[5] Funktionen $\boldsymbol{\phi}^j = \mathrm{diag}(\phi_1^j, \ldots, \phi_{\mathrm{D}}^j)$ und $\boldsymbol{\psi}^i = \mathrm{diag}(\psi_1^i, \ldots, \psi_{\mathrm{D}}^i)$ sind jeweils linear unabhängig und erfüllen homogene *Dirichlet*-Randbedingungen:

$$\phi_d^j \equiv 0 \ \text{ auf } \ \Gamma^h_{\mathsf{u}_d}\,, \qquad \psi_d^i \equiv 0 \ \text{ auf } \ \Gamma^h_{\mathsf{u}_d}\,, \qquad i,j = 1,\ldots,\mathrm{N},\ \ d = 1,\ldots,\mathrm{D}\,. \tag{2.31}$$

Def. 2.6: Eine *Petrov-Galerkin-Approximation* ist eine Funktionenschar $(\mathbf{u}^h(t),\, \mathbf{q}^h(t))$, $0 \le t \le T$ mit $\mathbf{u}^h(t) \in \mathcal{S}^h(t)$ und $\mathbf{q}^h(t) \in (L^\infty(\Omega^h))^{\mathrm{Q}}$, die

$$\begin{aligned} \mathcal{G}^h[\boldsymbol{\eta}^h, \mathbf{u}^h;\, \mathbf{q}^h] \ \equiv\ & \int_{\Omega^h} \left(\mathcal{M}(\boldsymbol{\eta}^h, \mathbf{u}^h, \ddot{\mathbf{u}}^h) + \mathcal{D}(\boldsymbol{\eta}^h, \mathbf{u}^h, \dot{\mathbf{u}}^h) + \mathcal{K}(\boldsymbol{\eta}^h, \mathbf{u}^h, \mathbf{q}^h)\right)\,\mathrm{d}v\, - \\ & - \int_{\Gamma^h} \mathcal{F}(\boldsymbol{\eta}^h, \mathbf{u}^h)\,\mathrm{d}a \ = \ 0 \end{aligned} \tag{2.32}$$

und

$$\mathbf{A}\,\dot{\mathbf{q}}^h - \mathbf{r}(\mathbf{q}^h, \mathbf{u}^h) = \mathbf{0} \tag{2.33}$$

für beliebige $t \in [0,T]$ und $\boldsymbol{\eta}^h \in \mathcal{T}^h$ sowie

$$\int_{\Omega^h} \boldsymbol{\eta}^h \cdot (\mathbf{u}^h(0) - \mathbf{u}_0^h)\,\mathrm{d}v = 0\,, \quad \int_{\Omega^h} \boldsymbol{\eta}^h \cdot (\dot{\mathbf{u}}^h(0) - \mathbf{v}_0^h)\,\mathrm{d}v = 0\,, \qquad \mathbf{q}^h(0) = \mathbf{q}_0^h \tag{2.34}$$

für beliebige $\boldsymbol{\eta}^h \in \mathcal{T}^h$ erfüllt. Dabei symbolisiert der Ersatz des Operators $\mathcal{G}[\cdot\,,\cdot\,;\cdot]$ aus (2.27) durch den Operator $\mathcal{G}^h[\cdot\,,\cdot\,;\cdot]$ die Approximation des Gebiets Ω durch das Gebiet Ω^h mit entsprechender Veränderung der auftretenden Integrale. Die approximierten Anfangsbedingungen sind analog mit $\mathbf{u}_0^h, \mathbf{v}_0^h$ und $\mathbf{q}_0^h$ bezeichnet. ■

Eine Methode zur Bestimmung einer *Petrov-Galerkin*-Approximation (das *Petrov-Galerkin-Verfahren* oder die *Methode der gewichteten Residuen*) ist nun naheliegend. Da Gleichung (2.32) für beliebige Testfunktionen $\boldsymbol{\eta}^h \in \mathcal{T}^h$ erfüllt sein muß und die Funktionen $\boldsymbol{\psi}^i$ eine Basis des Raumes $\mathcal{T}^h$ bilden, erhält man für jede Basisfunktion (jedes Freiheitsgrades) eine Gleichung. Mit der Bezeichung $\boldsymbol{\psi}_d^i := (0,\ldots,0,\psi_d^i,0,\ldots,0)^T$ für die

[4] In der Praxis wird meist auch das Rechengebiet Ω durch ein Gebiet Ω^h approximiert, zusammen mit einer entsprechenden Approximation der Anfangs- und Randbedingungen.

[5] Es wird hier die bei finiten Elementen übliche Matrix-Vektor-Notation verwendet, bei der die Ansatz- bzw. Testfunktionen Diagonalmatrizen mit Ansatz- bzw. Testfunktionen für die jeweiligen Freiheitsgrade sind. Das Produkt $\boldsymbol{\phi}^j\,\mathbf{u}_j = (\phi_1^j \mathsf{u}_j^1, \ldots, \phi_{\mathrm{D}}^j \mathsf{u}_j^{\mathrm{D}})^T$ liefert also eine D-dimensionale Vektorfunktion.

i-te Testfunktion am d-ten Freiheitsgrad erhält man die *Petrov-Galerkin*-Approximation durch Lösen des Systems aus $\mathrm{N} \cdot \mathrm{D}$ gewöhnlichen Differentialgleichungen zweiter Ordnung

$$\mathcal{G}^h[\boldsymbol{\psi}_d^i, \mathbf{u}^h;\, \mathbf{q}^h] \;=\; 0, \qquad\qquad i = 1, \ldots, \mathrm{N},\;\; d = 1, \ldots, \mathrm{D} \tag{2.35}$$

mit den $2 \cdot \mathrm{N} \cdot \mathrm{D}$ Anfangsbedingungen

$$\int_{\Omega^h} \boldsymbol{\psi}_d^i \cdot (\mathbf{u}^h(0) - \mathbf{u}_0^h)\, \mathrm{d}v = 0\,, \qquad \int_{\Omega^h} \boldsymbol{\psi}_d^i \cdot (\dot{\mathbf{u}}^h(0) - \mathbf{v}_0^h)\, \mathrm{d}v = 0 \tag{2.36}$$

unter Beibehaltung der Entwicklungsgleichungen und Anfangsbedingungen

$$\mathbf{A}\, \dot{\mathbf{q}}^h - \mathbf{r}(\mathbf{q}^h, \mathbf{u}^h) = \mathbf{0}\,, \qquad\qquad \mathbf{q}^h(0) = \mathbf{q}_0^h \tag{2.37}$$

für die Näherung $\mathbf{q}^h(t)$ der internen Variablen. Da das ursprüngliche Problem nur bzgl. der Ortsvariablen $\mathbf{x}$ diskretisiert wurde, aber noch kontinuierlich in der Zeitvariablen t ist, spricht man von einer *Semidiskretisierung*.

2.4.3 *Galerkin*-Verfahren

Ein Spezialfall des allgemeinen *Petrov-Galerkin*-Verfahrens ist das *Galerkin-Verfahren* (auch *Bubnov-Galerkin-Verfahren*). In diesem Fall wählt man für die endlichdimensionalen Ansatz- und Testräume $\mathcal{S}^h(t)$ und $\mathcal{T}^h$ dieselben Basisfunktionen: $\boldsymbol{\phi}^i \equiv \boldsymbol{\psi}^i$. Der Ansatzraum ist also gerade der um die *Dirichlet*-Randbedingungen verschobene Testraum: $\mathcal{S}^h(t) = \bar{\mathbf{u}}^h(t) + \mathcal{T}^h$. Eine schöne Eigenschaft des *Galerkin*-Verfahrens ist (z. B. im Fall eines linear elastischen Problems), daß der Fehler $\mathbf{u} - \mathbf{u}^h$ senkrecht auf dem Ansatzraum steht, d. h. daß das Energie-Skalarprodukt $B(\mathbf{u} - \mathbf{u}^h, \boldsymbol{\eta}^h)$ zwischen dem Fehler und einer beliebigen Funktion $\boldsymbol{\eta}^h \in \mathcal{T}^h$ verschwindet. Das *Galerkin*-Verfahren liefert also die *Best-approximation* bzgl. des gewählten Ansatzraumes (vgl. z. B. *Strang & Fix* [109, §1.6]).

Bemerkung: In dieser Arbeit wird im folgenden nur das *Galerkin*-Verfahren betrachtet. Die oben dargestellte allgemeine Formulierung eines *Petrov-Galerkin*-Verfahrens und die damit verbundene Implementierung erlauben es jedoch, stabilisierte Verfahren mit lösungsabhängigen Testfunktionen einzubeziehen, wie etwa die *Galerkin*-Least-Squares-Methoden (GLS) oder das Streamline-Upwind-*Petrov-Galerkin*-Verfahren (SUPG). So wurde im Rahmen eines DFG-Forschungsvorhabens [51] vom Verfasser eine stabilisierte FE-Formulierung entwickelt, die bei dem in dieser Arbeit vorgestellten inkompressiblen Zweiphasenmodell die Verwendung von linearen Ansätzen sowohl für die Verschiebung als auch für den Druck gestattet. Dabei werden die bei unstabilisierten Elementen auftretenden unphysikalischen Oszillationen im Druckfeld vermieden, wobei gleichzeitig die Konsistenz des Verfahrens sichergestellt ist. Dies bedeutet insbesondere, daß die Stabilisierungsterme im Grenzfall beliebig feiner Netze ($h \to 0$) verschwinden. □

2.4.4 Wahl der Ansatz- und Testfunktionen – Finite Elemente

Bisher wurde noch nichts über die Wahl der Ansatz- und Testfunktionen ausgesagt. Das (*Petrov-*)*Galerkin*-Verfahren kann prinzipiell mit beliebigen Ansatz- und Testfunktionen

durchgeführt werden. In der Praxis gilt das Interesse jedoch den folgenden Eigenschaften der Basis:

1. einfache Struktur der linearen Gleichungssysteme (dünnbesetzte Matrizen),
2. einfache Auswertbarkeit der auftretenden Integrale,
3. direkt interpretierbare Koeffizienten $\mathbf{u}_j$ der Näherungslösung.

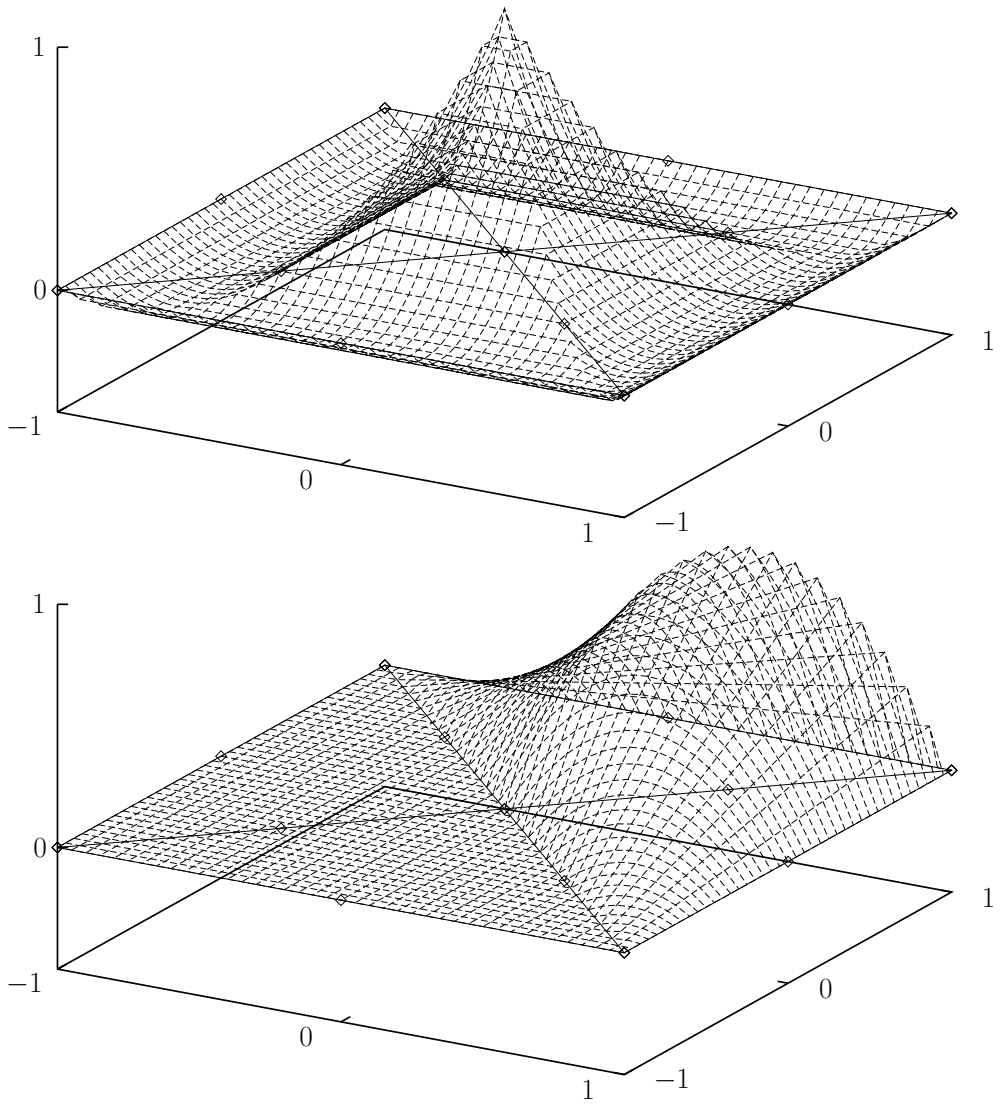


Abbildung 2.1: *Quadratische Ansatzfunktionen auf Dreiecken*

Diese Forderungen führten zur Entwicklung der Methode der finiten Elemente (FEM), die sich in den letzten Jahren insbesondere für Anwendungen mit komplizierter Geometrie durchgesetzt hat. Die erste Forderung wird dadurch erfüllt, daß man Basisfunktionen mit *kompaktem Träger* wählt, wodurch die (räumliche) Kopplung der Koeffizienten minimiert wird. Dadurch erhält man nach Zeitdiskretisierung und Linearisierung Gleichungssysteme

mit dünnbesetzten Matrizen. Zur Erfüllung der zweiten Forderung approximiert man das Rechengebiet Ω durch ein Netz aus E *Elementen* Ω_e:

$$\Omega \approx \Omega^h = \bigcup_{e=1}^{\mathrm{E}} \Omega_e \,. \tag{2.38}$$

Die Elemente werden dabei *konform* mit N *Knoten* $\mathbf{x}_j$ verbunden, d. h. zwei Elemente schneiden sich entweder gar nicht, in einem Knoten, in einer Kante oder – in drei Dimensionen – in einer Fläche. Jedem Knoten wird genau eine Basisfunktion $\boldsymbol{\phi}^j$ zugeordnet, deren Träger der *Patch* aller an den Knoten angrenzenden Elemente ist,

$$\boldsymbol{\phi}^j(\mathbf{x}) = \mathbf{0}, \quad \text{falls } \mathbf{x} \notin \bigcup_{e \in E_j} \Omega_e \,, \qquad j = 1, \ldots, \mathrm{N} \,, \tag{2.39}$$

wobei die Indexmenge E_j alle Elemente beinhaltet, die an den Knoten $\mathbf{x}_j$ grenzen. Diese Wahl der Basisfunktionen erlaubt die Auswertung der auftretenden Integrale auf Elementbasis und damit eine weitgehende Vereinfachung bei der programmtechnischen Umsetzung. Die dritte Forderung läßt sich durch eine Normierung der Basisfunktionen erfüllen:

$$\phi_d^i(\mathbf{x}_j) = \delta_j^i \,, \qquad i, j = 1, \ldots, \mathrm{N} \,, \quad d = 1, \ldots, \mathrm{D} \,. \tag{2.40}$$

Dadurch ist sichergestellt, daß der Koeffizient $\mathbf{u}_j$ am Knoten j genau dem Funktionswert der Näherungslösung an diesem Knoten entspricht (ausgenommen *Dirichlet*-Randknoten):

$$\mathbf{u}^h(\mathbf{x}_j) \;=\; \sum_{i=1}^{\mathrm{N}} \boldsymbol{\phi}^i(\mathbf{x}_j)\, \mathbf{u}_i \;=\; \mathbf{u}_j \,. \tag{2.41}$$

An einem *Dirichlet*-Rand $\Gamma^h_{\mathsf{u}_d}$ verschwinden alle Ansatzfunktionen ϕ^j_d, so daß man dort den Wert der vorgegebenen Randbedingung erhält: $\mathsf{u}^h_d(\mathbf{x}_j) = \bar{\mathsf{u}}^h_d(\mathbf{x}_j)$. Eine weitere wichtige Eigenschaft der Finite-Elemente-Ansatzfunktionen ist die *Teilung der Eins*, d. h. in jedem Punkt ist die Summe der Ansatzfunktionen gleich Eins. Dies bedeutet, daß die Approximation über das ganze Gebiet gleichmäßig ist. Alle Eigenschaften der FEM-Basisfunktionen sind in Kasten (2.42) zusammengefaßt. Die auf ein Element bezogenen Funktionen sind mit dem Index (e) überschrieben, $\mathbf{x}_m^{(e)}$ bezeichnet den Knoten m des Elementes e, und N_e ist die Anzahl der Knoten von Element e. Der Operator $\mathbf{A}$ symbolisiert die Assemblierung, d. h. er leistet die Umindizierung von elementbezogenen zu globalen Knotennummern (dem Elementknoten n wird der globale Knoten j zugeordnet). Im Fall verschiedener Ansatz- und Testfunktionen (*Petrov-Galerkin*-Verfahren) müssen die Testfunktionen $\boldsymbol{\psi}^i$ dieselben Bedingungen erfüllen wie die Ansatzfunktionen $\boldsymbol{\phi}^j$.

Als Beispiel sind in Abbildung 2.1 quadratische Ansatzfunktionen auf Dreiecken (6-Knoten-Dreieckselement) dargestellt, im oberen Bild für einen Eckknoten, im unteren Bild für einen Mittelknoten.

Bemerkung: In einem Finite-Elemente-Programm kann eine *Dirichlet*-Randbedingung explizit erfüllt werden, indem die aus der schwachen Formulierung resultierende Gleichung für den entsprechenden Freiheitsgrad durch die Randbedingung ersetzt wird. Dies kann

Eigenschaften der FEM-Basisfunktionen

Lokale Ansatzfunktionen $\overset{(e)}{\boldsymbol{\phi}}{}^n = \mathrm{diag}(\overset{(e)}{\phi}{}^n_1, \ldots, \overset{(e)}{\phi}{}^n_{\mathrm{D}})$ je Element e:

$$\begin{aligned}
&(i) && \overset{(e)}{\phi}{}^n_d(\mathbf{x}) &&= 0 && \mathbf{x} \notin \Omega_e \\
&(ii) && \overset{(e)}{\phi}{}^n_d(\mathbf{x}^{(e)}_m) &&= \delta^n_m && m,n = 1,2,\ldots,\mathrm{N}_e \\
&(iii) && \sum_{n=1}^{\mathrm{N}_e} \overset{(e)}{\phi}{}^n_d(\mathbf{x}) &&= 1 && \mathbf{x} \in \Omega_e
\end{aligned} \tag{2.42}$$

Globale Ansatzfunktionen $\boldsymbol{\phi}^j = \mathrm{diag}(\phi^j_1, \ldots, \phi^j_{\mathrm{D}})$ je Knoten j:

$$\begin{aligned}
&(i) && \boldsymbol{\phi}^j(\mathbf{x}) &&= \sum_{e=1}^{\mathrm{E}} \sum_{n=1}^{\mathrm{N}_e} \mathbf{A}^j_n \overset{(e)}{\boldsymbol{\phi}}{}^n(\mathbf{x}) && \mathbf{x} \in \Omega^h \\
&(ii) && \phi^i_d(\mathbf{x}_j) &&= \delta^i_j, && i,j = 1,2,\ldots,\mathrm{N} \\
&(iii) && \sum_{j=1}^{\mathrm{N}} \phi^j_d(\mathbf{x}) &&= 1 && \mathbf{x} \in \Omega^h
\end{aligned}$$

innerhalb der Programm-Funktion geschehen, die die Anteile des Integrals aus den einzelnen finiten Elementen zusammensetzt (Assemblierung). Daher müssen die Randbedingungen bei der Programmierung einer Auswertungsfunktion für ein finites Element nicht berücksichtigt werden. Die Assemblierungsfunktion ersetzt einfach vor der Auswertung der Element-Funktion die durch *Dirichlet*-Randbedingungen festgelegten Freiheitsgrade; innerhalb der Element-Funktion kann damit immer ein Ansatz ohne Berücksichtigung der Randbedingungen verwendet werden. □

2.4.5 Geometrie-Transformation und numerische Integration

Im letzten Abschnitt war eine wichtige Forderung die einfache Berechnung der Integrale. Aufgrund der speziellen Wahl des Trägers der Finite-Elemente-Ansatzfunktionen ist es möglich, die Integrale Element für Element – jeweils mit allen Ansatzfunktionen der Knoten des betrachteten Elementes – zu berechnen und danach zu einem Gesamtsystem zusammenzusetzen (*Assemblierung*). Außerdem ist man für eine allgemeingültige Formulierung innerhalb eines FEM-Programmes daran interessiert, die Ansatzfunktionen möglichst einfach zu implementieren. Daher werden die Ansatzfunktionen üblicherweise auf ein Referenzelement $\hat{\Omega}_e$ bezogen, in Abbildung 2.2 ist dies für den quadratischen Ansatz im Dreieck dargestellt. Wird die Element-Geometrie mit denselben Ansatzfunktionen transformiert wie die Näherungslösung, so spricht man von einem *isoparametrischen Ansatz*, bei Ansätzen höherer Ordnung für die Geometrie von einem *superparametrischen Ansatz*. Bei Problemen mit mehreren Freiheitsgraden können auch Mischformen auftreten, wenn z. B. die Geometrie und die Verschiebungen quadratisch und der Druck linear angesetzt werden (Verschiebungen isoparametrisch, Druck superparametrisch).

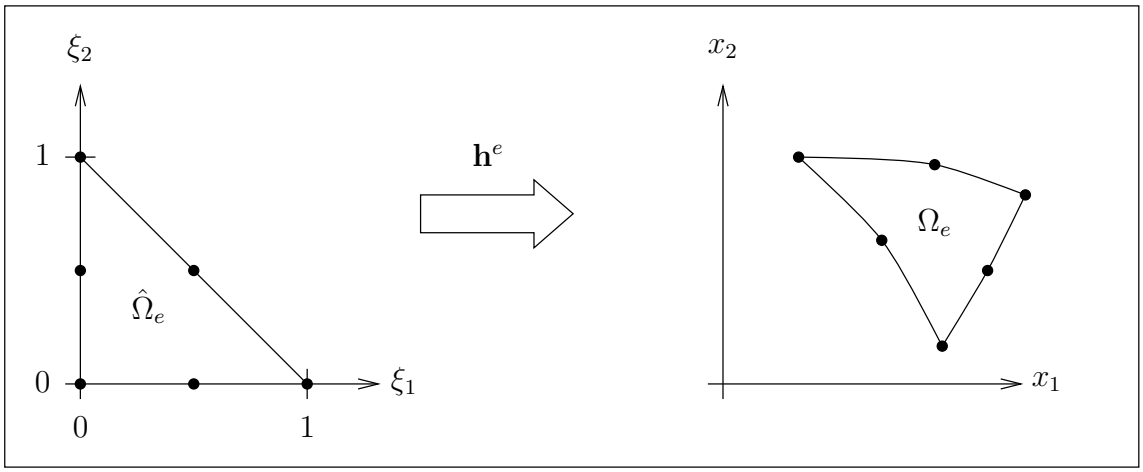


Abbildung 2.2: *Geometrie-Transformation (Dreieck, quadratischer Ansatz)*

Bei der Berechnung der auftretenden Integrale ist man i. a. auf *numerische Integrationsverfahren* (*Quadratur* bzw. *Kubatur*) angewiesen. Spezielle Formeln für die verschiedenen Elementtypen können der Literatur entnommen werden, z. B. *Schwarz* [104], *Zienkiewicz & Taylor* [126]. Das zu berechnende Integral über einem Element wird zunächst mit der Substitutionsregel auf das Referenzelement bezogen, und man erhält mit der *Jacobi*-Determinante

$$J_e = \left| \det \frac{\mathrm{d}\mathbf{h}^e}{\mathrm{d}\boldsymbol{\xi}} \right| = \left| \det \left(\frac{\partial \mathrm{x}_i}{\partial \xi_j} \right) \right| \tag{2.43}$$

der Geometrie-Transformation $\boldsymbol{\xi} \mapsto \mathbf{x} = \mathbf{h}^e(\boldsymbol{\xi})$:

$$\int\limits_{\Omega_e} f(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int\limits_{\hat{\Omega}_e} f(\mathbf{h}^e(\boldsymbol{\xi})) \; J_e(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} \, . \tag{2.44}$$

Die numerische Integrationsformel besteht aus K *Integrationspunkten* $\boldsymbol{\xi}_k$ innerhalb des Referenzelements (bei *Gauß*-Quadraturen heißen die Integrationspunkte auch *Gauß-Punkte*) mit zugehörigen Gewichten w_k. Damit erhält man als Näherung für das Integral über einem Element:

$$\int\limits_{\Omega_e} f(\mathbf{x}) \, \mathrm{d}\mathbf{x} \approx \sum_{k=1}^{\mathrm{K}} f(\mathbf{h}^e(\boldsymbol{\xi}_k)) \; J_e(\boldsymbol{\xi}_k) \; w_k \, . \tag{2.45}$$

Für die Berechnung des Integrals über das gesamte Rechengebiet Ω, das durch das FE-Netz Ω^h approximiert wird, nutzt man die Aufteilung in finite Elemente aus und erhält:

$$\int\limits_{\Omega} f(\mathbf{x}) \, \mathrm{d}\mathbf{x} \approx \int\limits_{\Omega^h} f(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \sum_{e=1}^{\mathrm{E}} \int\limits_{\Omega_e} f(\mathbf{x}) \, \mathrm{d}\mathbf{x} \approx \sum_{e=1}^{\mathrm{E}} \sum_{k=1}^{\mathrm{K}} f(\mathbf{h}^e(\boldsymbol{\xi}_k)) \; J_e(\boldsymbol{\xi}_k) \; w_k \, . \tag{2.46}$$

Bei allgemeinen, krummlinig berandeten Elementen müssen auch für die Randintegrale numerische Integrationsformeln verwendet werden.

2.4.6 Diskretisierung der plastischen Entwicklungsgleichungen

Die Werte der internen Variablen werden innerhalb der schwachen Formulierung nur im Integral mit dem Operator $\mathcal{K}(\boldsymbol{\eta}^h, \mathbf{u}^h, \mathbf{q}^h)$ benötigt. Es ist daher bei Verwendung einer numerischen Integrationsformel zur Berechnung dieses Integrals naheliegend, die Gültigkeit der Entwicklungsgleichungen genau an den Integrationspunkten $\mathbf{x}_k^e := \mathbf{h}^e(\boldsymbol{\xi}_k)$ zu fordern (*Kollokation*). Man erhält dadurch neben den $\mathrm{N} \cdot \mathrm{D}$ Gleichungen aus der Ortsdiskretisierung der schwachen Formulierung weitere $\mathrm{E} \cdot \mathrm{K} \cdot \mathrm{Q}$ (Anzahl Elemente mal Anzahl Integrationspunkte je Element mal Anzahl interne Variablen[6]) Gleichungen aus der Ortsdiskretisierung der Entwicklungsgleichungen für die internen Variablen:

$$\mathbf{A}\,\dot{\mathbf{q}}^h(\mathbf{x}_k^e) - \mathbf{r}(\mathbf{q}^h(\mathbf{x}_k^e), \mathbf{u}^h(\mathbf{x}_k^e)) = \mathbf{0}\,, \qquad e = 1, \ldots, \mathrm{E}\,, \quad k = 1, \ldots, \mathrm{K}\,, \tag{2.47}$$

mit den zugehörigen Anfangsbedingungen

$$\mathbf{q}^h(\mathbf{x}_k^e)\big|_{t=0} = \mathbf{q}_0^h(\mathbf{x}_k^e)\,, \qquad e = 1, \ldots, \mathrm{E}\,, \quad k = 1, \ldots, \mathrm{K}\,. \tag{2.48}$$

2.5 Struktur der ortsdiskreten Systeme

Die allgemeine Form der Ortsdiskretisierung wird nun sowohl für das quasi-statische als auch für das dynamische Zweiphasenmodell spezialisiert, und es werden geeignete Ansatzfunktionen für die verschiedenen zu approximierenden Felder angegeben.

Es wird im folgenden der räumlich zweidimensionale Fall mit einem kartesischen Koordinatensystem unter Voraussetzung des ebenen Verzerrungszustandes betrachtet. Eine Formulierung für rotationssymmetrische Probleme (Zylinderkoordinaten) erfordert nur wenige Änderungen bei den auftretenden Gradienten-Operatoren, wird aber im Rahmen dieser Arbeit nicht behandelt. Auch der dreidimensionale Fall wird nicht behandelt; die Formulierung stellt kein Problem dar, jedoch steigt der Rechenaufwand insbesondere beim Lösen der linearen Gleichungssysteme immens an, und adaptive Verfahren sind – insbesondere in Hinblick auf die Erzeugung und Anpassung der Netze – ungleich schwieriger zu implementieren als in zwei Dimensionen.

In bezug auf die plastischen Entwicklungsgleichungen beschränkt sich die folgende Darstellung auf das viskoplastische Modell. Der Fall der Elastoplastizität ist als Grenzfall enthalten, wird aber nicht mehr mit angegeben.

2.5.1 Quasi-statische Formulierung

Der Vektor $\mathbf{u}$ besteht aus $\mathrm{D} = 3$ Freiheitsgraden, nämlich den Verschiebungen $\mathbf{u}_S$ in den beiden Raumrichtungen und dem Porenfluiddruck p. Entsprechend enthält der Vektor $\boldsymbol{\eta}$

[6] Bei einem Netz aus verschiedenartigen Elementen (z. B. Dreiecken und Vierecken) müssen an dieser Stelle statt der Multiplikation entsprechende Summen gebildet werden (unterschiedliche Zahl von Integrationspunkten in den Elementen).

die Testfunktionen $\delta\mathbf{u}_S$ und δp:

$$\mathbf{u} = (\mathrm{u}_S^1,\ \mathrm{u}_S^2,\ p)^T, \qquad \boldsymbol{\eta} = (\delta\mathrm{u}_S^1,\ \delta\mathrm{u}_S^2,\ \delta p)^T. \tag{2.49}$$

Durch Vergleich der schwachen Formulierung in Definition 2.3 mit der Operatordarstellung in Definition 2.5 erhält man die Operatoren $\mathcal{M}$, $\mathcal{D}$, $\mathcal{K}$ und $\mathcal{F}$:

$$\begin{aligned}
\mathcal{M}(\boldsymbol{\eta}, \mathbf{u}, \ddot{\mathbf{u}}) &\equiv 0\,, \\
\mathcal{D}(\boldsymbol{\eta}, \mathbf{u}, \dot{\mathbf{u}}) &= \delta p\ \mathrm{div}(\mathbf{u}_S)'_S\,, \\
\mathcal{K}(\boldsymbol{\eta}, \mathbf{u}, \mathbf{q}) &= \mathrm{grad}\,\delta\mathbf{u}_S \cdot (\boldsymbol{\sigma}_E^S - p\,\mathbf{I}) \;-\; \delta\mathbf{u}_S \cdot (\rho^S + \rho^F)\,\mathbf{b} \\
&+ \mathrm{grad}\,\delta p \cdot \frac{k^F}{\gamma^{FR}}\,\mathrm{grad}\,p \;-\; \mathrm{grad}\,\delta p \cdot \frac{k^F}{\gamma^{FR}}\,\rho^{FR}\,\mathbf{b}\,, \\
\mathcal{F}(\boldsymbol{\eta}, \mathbf{u}) &= \delta\mathbf{u}_S \cdot \bar{\mathbf{t}} \\
&- \delta p\,\bar{v}\,.
\end{aligned} \tag{2.50}$$

Dabei wurden die beiden skalaren Gleichungen der schwachen Formulierung des quasi-statischen Modells addiert. Die Summe ist wegen der linearen Unabhängigkeit der Testfunktionen $\delta\mathbf{u}_S$ und δp äquivalent zu den beiden einzelnen Gleichungen.

Bemerkung: Streng genommen muß man an dieser Stelle die schwachen Formulierungen vor der Addition entdimensionieren bzw. auf die gleiche physikalische Dimension (Einheit) bringen. Dies stellt jedoch in der Praxis kein Problem dar, da man jede Gleichung entsprechend skalieren kann. Beim quasi-statischen Modell hat die schwache Formulierung der Impulsbilanz der Mischung die Einheit $1\,\mathrm{Nm} = 1\,\mathrm{J}$ (Energie) und die schwache Formulierung der Volumenbilanz die Einheit $1\,\mathrm{Nm/s} = 1\,\mathrm{J/s}$ (Leistung). Man multipliziert also die schwache Formulierung der Impulsbilanz der Mischung mit dem Faktor $1/\mathrm{Nm}$ und die schwache Form der Volumenbilanz mit $1\,\mathrm{s/Nm}$. Danach sind beide Gleichungen dimensionslos und können addiert werden wie in Gleichung (2.50). Man beachte, daß die Operatoren $\mathcal{M}, \mathcal{D}$ und $\mathcal{K}$ dann die gemeinsame Einheit $1/\mathrm{m}^3$ haben und der Operator $\mathcal{F}$ die Einheit $1/\mathrm{m}^2$. Erst das Ergebnis der Integration gemäß (2.27) ist also dimensionslos ($[\mathrm{d}v] = 1\,\mathrm{m}^3$, $[\mathrm{d}a] = 1\,\mathrm{m}^2$). □

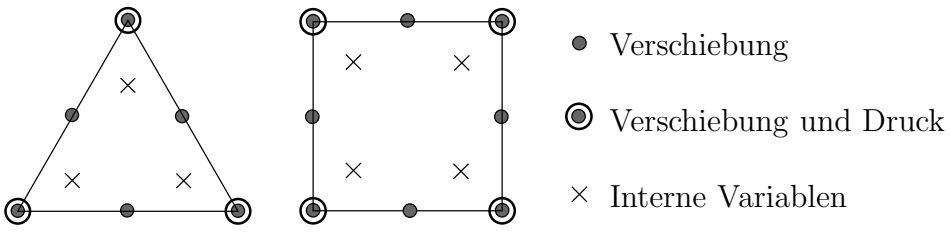


Abbildung 2.3: *FEM-Ansatz im Dreieck und im Viereck*

Der Vektor $\mathbf{q}$ der internen Variablen besteht aus vier Komponenten des Tensors $\boldsymbol{\varepsilon}_{Sp}$ der plastischen Verzerrungen und dem Proportionalitätsfaktor Λ. Entsprechend ist auch die

Entwicklungsgleichung eine Vektorfunktion mit fünf Komponenten, da der Spannungstensor im ebenen Verzerrungszustand die gleichen vier nicht-verschwindenden Einträge besitzt:

$$\mathbf{q} = \begin{pmatrix} \varepsilon_{Sp}^{11} \\ \varepsilon_{Sp}^{22} \\ \varepsilon_{Sp}^{33} \\ \varepsilon_{Sp}^{12} \\ \Lambda \end{pmatrix}, \qquad \mathbf{A}\,\dot{\mathbf{q}} - \mathbf{r}(\mathbf{q},\mathbf{u}) \equiv \begin{pmatrix} (\boldsymbol{\varepsilon}_{Sp})'_S \\ 0 \end{pmatrix} - \begin{pmatrix} \Lambda \dfrac{\mathrm{d}G}{\mathrm{d}\boldsymbol{\sigma}_E^S} \\ \Lambda\,\eta\,\sigma_0^r - \left\langle F(\boldsymbol{\sigma}_E^S)\right\rangle^r \end{pmatrix} = \mathbf{0}\,. \quad (2.51)$$

Bemerkung: Im hier betrachteten viskoplastischen Fall könnte man die Variable Λ direkt eliminieren, indem man die zweite Gleichung nach Λ auflöst und in die erste einsetzt. Der Vorteil der o. a. Darstellung der Bestimmungsgleichung für den plastischen Multiplikator Λ, die man durch Multiplikation von (1.93) mit $\eta\,\sigma_0^r$ erhält, ist jedoch darin zu sehen, daß eine einheitliche Formulierung sowohl für den viskoplastischen Fall ($\eta > 0$) als auch für den elasto-plastischen Grenzfall ($\eta = 0$, $r = 1$) vorliegt. Die weiteren *Kuhn-Tucker*-Bedingungen des elasto-plastischen Modells werden dabei durch einen geeigneten Algorithmus auf Element-Ebene (elastischer Trial-State) automatisch eingehalten. □

Im folgenden werden quadratische Ansätze für die Verschiebung und lineare Ansätze für den Druck verwendet (siehe Abbildung 2.3). Diese Wahl kann wie folgt motiviert werden:

- In der schwachen Formulierung der Impulsbilanz der Mischung wird ein Verschiebungsgradient (Spannungen) zum Druck addiert. Der diskrete Spannungstensor ist also mit dem gewählten Ansatz genau wie der diskrete Druck eine lineare Funktion der Ortsvariablen $\mathbf{x}$.
- Bei gemischten Methoden ist die inf-sup-Bedingung (*Babuska-Brezzi*-Bedingung) ein entscheidendes Kriterium für die Stabilität der numerischen Lösung (*Brezzi & Fortin* [28]). Die gewählte Kombination der Ansätze erfüllt diese Bedingung und ist bei Dreiecken unter dem Namen *Taylor-Hood-Element* bekannt (*Braess* [25]).

Durch Nullsetzen der linearen Ansatzfunktionen an den Mittelknoten kann die Summe formal bei allen Ansätzen von 1 bis N laufen[7]. Damit ergibt sich für die Approximation von Verschiebung und Druck[8]:

$$\mathbf{u}_S^h(\mathbf{x},t) = \sum_{j=1}^{\mathrm{N}} \begin{pmatrix} \phi_1^j(\mathbf{x})\ \mathrm{u}_{Sj}^1(t) \\ \phi_2^j(\mathbf{x})\ \mathrm{u}_{Sj}^2(t) \end{pmatrix}, \qquad p^h(\mathbf{x},t) = \sum_{j=1}^{\mathrm{N}} \phi_3^j(\mathbf{x})\ p_j(t)\,. \quad (2.52)$$

Setzt man die Ansätze (2.52) in die Operatoren (2.50) ein (*Petrov-Galerkin*-Verfahren), so erhält man je Knoten i die drei Gleichungen in Kasten (2.53). Darin bezeichnen $(...)_{,k} = \partial(...)/\partial x_k$ die partiellen Ableitungen nach den Koordinaten und σ_{kl} die von der

[7] Im Programm wird natürlich aus Effizienzgründen für die linearen Ansätze nur über die M < N Eckknoten summiert.

[8] Wie in der Bemerkung auf Seite 51 erläutert, können die *Dirichlet*-Randbedingungen bei den folgenden Betrachtungen ignoriert werden.

Knoten-Gleichungen des ortsdiskreten quasi-statischen Modells

$$\int_{\Omega^h} \sum_{j=1}^{\mathrm{N}} \Bigg[\underbrace{\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \psi_3^i \phi_{1,1}^j & \psi_3^i \phi_{2,2}^j & 0 \end{pmatrix}}_{\mathbf{M}_{ij}} \underbrace{\begin{pmatrix} \dot{\mathrm{u}}_{Sj}^1 \\ \dot{\mathrm{u}}_{Sj}^2 \\ \dot{p}_j \end{pmatrix}}_{\dot{\mathbf{u}}_j} + \underbrace{\begin{pmatrix} \psi_{1,1}^i(\sigma_{11} - \phi_3^j p_j) + \psi_{1,2}^i \sigma_{12} - \psi_1^i(\rho^S + \rho^F) b^1 \\ \psi_{2,1}^i \sigma_{21} + \psi_{2,2}^i(\sigma_{22} - \phi_3^j p_j) - \psi_2^i(\rho^S + \rho^F) b^2 \\ \dfrac{k^F}{\gamma^{FR}} [\psi_{3,1}^i(\phi_{3,1}^j p_j - \rho^{FR} b^1) + \psi_{3,2}^i(\phi_{3,2}^j p_j - \rho^{FR} b^2)] \end{pmatrix}}_{\mathbf{k}_i(\boldsymbol{u}, \boldsymbol{q})} \Bigg] \mathrm{d}v = \underbrace{\begin{pmatrix} \mathsf{f}_i^1 \\ \mathsf{f}_i^2 \\ \mathsf{f}_i^3 \end{pmatrix}}_{\mathbf{f}_i} \tag{2.53}$$

Verschiebung und den internen Variablen abhängigen Koeffizienten des Spannungstensors $\boldsymbol{\sigma}_E^S$. Die Einträge f_i^d auf der rechten Seite stammen aus der Berechnung der Randintegrale mit dem Operator $\mathcal{F}$ (*Neumann*-Randbedingungen). Das Volumenintegral wird gemäß (2.46) über eine numerische Integrationsformel berechnet, und die Summe in (2.52) und (2.53) erstreckt sich über alle N Knoten des FE-Netzes (globale Sicht).

Bemerkung: Die hier dargestellte globale Sicht ist hilfreich zum Verständnis der Struktur des Gesamtsystems, das durch die FEM-Ortsdiskretisierung entsteht. In einem Finite-Elemente-Programm geschieht die Auswertung der Integrale jedoch nicht knoten-orientiert, sondern element-orientiert. Dies führt zu den üblichen Begriffen von *Element-Massenmatrix*, *Element-Steifigkeitsvektor* und *Element-Steifigkeitsmatrix*. Nach der elementweisen Auswertung werden dann die einzelnen Anteile der Integrale mit dem in Kasten (2.42) angegebenen Zusammenhang zwischen lokalen und globalen Basisfunktionen assembliert. Der Umkehroperator des Assemblierungsoperators $\mathbf{A}$ dient dabei zur Auswahl der elementweisen Unbekannten aus dem globalen Vektor der Unbekannten, die zur Auswertung des nichtlinearen Element-Steifigkeitsvektors benötigt werden. □

Es werden nun die 3 N Unbekannten[9] an den Knoten (FEM-Freiheitsgrade) im Vektor

$$\boldsymbol{u} := \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_{\mathrm{N}} \end{pmatrix} \qquad \text{mit} \qquad \mathbf{u}_j = \begin{pmatrix} \mathrm{u}_{Sj}^1 \\ \mathrm{u}_{Sj}^2 \\ p_j \end{pmatrix}, \quad j = 1, \ldots, \mathrm{N} \tag{2.54}$$

[9] Die Druck-Freiheitsgrade treten aufgrund des linearen Ansatzes nur an den M Eckknoten auf, so daß im Programm auch nur 2 N + M Variablen behandelt werden. Zur Vereinfachung der Darstellung wird dies hier jedoch nicht berücksichtigt.

und die $5\,\mathrm{E}\cdot\mathrm{K}$ Unbekannten an den Integrationspunkten (interne Variablen) im Vektor

$$\boldsymbol{q} := \begin{pmatrix} \mathbf{q}_1^1 \\ \vdots \\ \mathbf{q}_{\mathrm{K}}^{\mathrm{E}} \end{pmatrix} \qquad \text{mit} \qquad \mathbf{q}_k^e = \left.\begin{pmatrix} \varepsilon_{Sp}^{11} \\ \varepsilon_{Sp}^{22} \\ \varepsilon_{Sp}^{33} \\ \varepsilon_{Sp}^{12} \\ \Lambda \end{pmatrix}\right|_{\mathbf{x}=\mathbf{x}_k^e}, \quad e = 1,\ldots,\mathrm{E},\ k = 1,\ldots,\mathrm{K} \tag{2.55}$$

zusammengefaßt (vgl. Gleichungen (2.47) und (2.51)). Die verallgemeinerte *Massenmatrix* $\boldsymbol{M}$ ist eine Blockmatrix

$$\boldsymbol{M} = \begin{pmatrix} \mathbf{M}_{11} & \cdots & \mathbf{M}_{1\mathrm{N}} \\ \vdots & \ddots & \vdots \\ \mathbf{M}_{\mathrm{N}1} & \cdots & \mathbf{M}_{\mathrm{NN}} \end{pmatrix}, \tag{2.56}$$

wobei jeder Block $\mathbf{M}_{ij}$ gemäß Gleichung (2.53) die Kopplungsterme zwischen Knoten i und Knoten j enthält, die aus der Ortsdiskretisierung der Volumenbilanz stammen.

Damit erhält man insgesamt mit $\boldsymbol{y} := (\boldsymbol{u}^T, \boldsymbol{q}^T)^T$ für das ortsdiskrete quasi-statische Modell das folgende Anfangswertproblem in der Zeit:

$$\boldsymbol{F}(t, \boldsymbol{y}, \dot{\boldsymbol{y}}) \equiv \begin{bmatrix} \boldsymbol{g}(t, \boldsymbol{u}, \dot{\boldsymbol{u}}, \boldsymbol{q}) \\ \boldsymbol{l}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}) \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{M}\dot{\boldsymbol{u}} + \boldsymbol{k}(\boldsymbol{u}, \boldsymbol{q}) - \boldsymbol{f} \\ \boldsymbol{A}\,\dot{\boldsymbol{q}} - \boldsymbol{r}(\boldsymbol{q}, \boldsymbol{u}) \end{bmatrix} \stackrel{!}{=} \boldsymbol{0}\,, \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0\,. \tag{2.57}$$

Die Vektorfunktion $\boldsymbol{g}$ stellt die $3\,\mathrm{N}$ globalen FEM-Bestimmungsgleichungen (2.53) dar, während $\boldsymbol{l}$ die $5\,\mathrm{E}\cdot\mathrm{K}$ lokalen plastischen Entwicklungsgleichungen (2.47) an den Integrationspunkten repräsentiert. In Anlehnung an die klassischen Bezeichnungen der FEM für linear elastische Probleme wird $\boldsymbol{M}$ die verallgemeinerte *Massenmatrix* (*Systemmatrix*), $\boldsymbol{k}$ der verallgemeinerte *Steifigkeitsvektor* und $\boldsymbol{f}$ der verallgemeinerte *Kraftvektor* genannt. Die verallgemeinerte *Steifigkeitsmatrix* $\boldsymbol{K}$ erhält man aufgrund der Nichtlinearitäten im Vektor $\boldsymbol{k}$ erst nach der Linearisierung, die wiederum erst nach der Zeitdiskretisierung stattfindet. Die Matrix $\boldsymbol{K}$ hat dann dieselbe Struktur wie die Matrix $\boldsymbol{M}$ in Gleichung (2.56), jedoch mit den 3×3-Blöcken $\mathbf{K}_{ij} = \partial\mathbf{k}_i/\partial\mathbf{u}_j$.

Da die Matrix $\boldsymbol{M}$ nicht den vollen Rang besitzt, handelt es sich bei (2.57) nicht um ein System gewöhnlicher Differentialgleichungen (ODE), sondern um ein System differential-algebraischer Gleichungen (DAE). Hierfür können zwei Gründe genannt werden:

1. In der quasi-statischen Formulierung werden die Beschleunigungsterme $\rho^\alpha \mathbf{x}''_\alpha$ vernachlässigt, so daß die Matrix $\mathbf{M}_{ij}$ in (2.53) im 2×2-Block links oben keine Einträge enthält.

2. Die Inkompressibilität beider Phasen führt dazu, daß für den Druck p keine Entwicklungsgleichung vorliegt, so daß der Eintrag rechts unten in $\mathbf{M}_{ij}$ verschwindet.

Der zweite Punkt legt nahe, warum auch in der dynamischen Formulierung ein DAE-System vorliegt. Dies ist Thema des folgenden Abschnitts.

2.5.2 Dynamische Formulierung

Der Vektor $\mathbf{u}$ besteht im Fall der dynamischen Formulierung aus fünf Freiheitsgraden, nämlich der Verschiebung $\mathbf{u}_S$ und der Sickergeschwindigkeit $\mathbf{w}_F$, jeweils in beiden Raumrichtungen, sowie dem Porenfluiddruck p. Entsprechend enthält der Vektor $\boldsymbol{\eta}$ die Testfunktionen $\delta\mathbf{u}_S$, $\delta\mathbf{w}_F$ und δp:

$$\mathbf{u} = \left(\mathrm{u}_S^1,\, \mathrm{u}_S^2,\, \mathrm{w}_F^1,\, \mathrm{w}_F^2,\, p\right)^T, \qquad \boldsymbol{\eta} = \left(\delta\mathrm{u}_S^1,\, \delta\mathrm{u}_S^2,\, \delta\mathrm{w}_F^1,\, \delta\mathrm{w}_F^2,\, \delta p\right)^T. \tag{2.58}$$

Wieder erhält man durch Vergleich der schwachen Formulierung in Definition 2.4 mit der Operatordarstellung in Definition 2.5 die Operatoren $\mathcal{M}$, $\mathcal{D}$, $\mathcal{K}$ und $\mathcal{F}$:

$$\begin{aligned}
\mathcal{M}(\boldsymbol{\eta}, \mathbf{u}, \ddot{\mathbf{u}}) &= \delta\mathbf{u}_S \cdot (\rho^S + \rho^F)\,(\mathbf{u}_S)_S'' \\
&+ \delta\mathbf{w}_F \cdot \rho^F\,(\mathbf{u}_S)_S'', \\
\mathcal{D}(\boldsymbol{\eta}, \mathbf{u}, \dot{\mathbf{u}}) &= \delta\mathbf{u}_S \cdot \rho^F[(\mathbf{w}_F)_S' + (\operatorname{grad}(\mathbf{u}_S)_S')\,\mathbf{w}_F] \\
&+ \delta\mathbf{w}_F \cdot \rho^F[(\mathbf{w}_F)_S' + (\operatorname{grad}(\mathbf{u}_S)_S')\,\mathbf{w}_F] \\
&+ \delta p\, \operatorname{div}(\mathbf{u}_S)_S', \\
\mathcal{K}(\boldsymbol{\eta}, \mathbf{u}, \mathbf{q}) &= \delta\mathbf{u}_S \cdot (\rho^F[(\operatorname{grad}\mathbf{w}_F)\,\mathbf{w}_F - \mathbf{b}] - \rho^S\,\mathbf{b}) + \operatorname{grad}\delta\mathbf{u}_S \cdot (\boldsymbol{\sigma}_E^S - p\,\mathbf{I}) \\
&+ \delta\mathbf{w}_F \cdot (\rho^F[(\operatorname{grad}\mathbf{w}_F)\,\mathbf{w}_F - \mathbf{b}] + \tfrac{(n^F)^2\gamma^{FR}}{k^F}\,\mathbf{w}_F) - (\operatorname{div}\delta\mathbf{w}_F)\,n^F p \\
&- \operatorname{grad}\delta p \cdot n^F\mathbf{w}_F, \\
\mathcal{F}(\boldsymbol{\eta}, \mathbf{u}) &= \delta\mathbf{u}_S \cdot \bar{\mathbf{t}} \\
&+ \delta\mathbf{w}_F \cdot \bar{\mathbf{t}}^F \\
&- \delta p\,\bar{v}\,.
\end{aligned} \tag{2.59}$$

Die Entwicklungsgleichungen können aus Gleichung (2.51) der quasi-statischen Formulierung unverändert übernommen werden. Die Verschiebungen werden wie in (2.52) wieder durch quadratische, die Sickergeschwindigkeit und der Durck durch lineare Ansatzfunktionen approximiert:

$$\mathbf{w}_F^h(\mathbf{x}, t) = \sum_{j=1}^{\mathrm{N}} \begin{pmatrix} \phi_3^j(\mathbf{x})\ \mathrm{w}_{Fj}^1(t) \\ \phi_4^j(\mathbf{x})\ \mathrm{w}_{Fj}^2(t) \end{pmatrix}, \qquad p^h(\mathbf{x}, t) = \sum_{j=1}^{\mathrm{N}} \phi_5^j(\mathbf{x})\ p_j(t)\,. \tag{2.60}$$

Die Anwendung des *Petrov-Galerkin*-Verfahrens führt auf einen ähnlichen Gleichungssatz wie in (2.53), jedoch mit fünf Gleichungen je Knoten i und – nach Zusammenfassung aller Unbekannten in Vektoren $\boldsymbol{u}$ und $\boldsymbol{q}$ – zu einem System

$$\boldsymbol{F}(t, \boldsymbol{y}, \dot{\boldsymbol{y}}, \ddot{\boldsymbol{y}}) \equiv \begin{bmatrix} \boldsymbol{g}(t, \boldsymbol{u}, \dot{\boldsymbol{u}}, \ddot{\boldsymbol{u}}, \boldsymbol{q}) \\ \boldsymbol{l}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}, \quad \boldsymbol{u}) \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{M}\ddot{\boldsymbol{u}} + \boldsymbol{C}(\boldsymbol{u})\,\dot{\boldsymbol{u}} + \boldsymbol{k}(\boldsymbol{u}, \boldsymbol{q}) - \boldsymbol{f} \\ \boldsymbol{A}\,\dot{\boldsymbol{q}} - \boldsymbol{r}(\boldsymbol{q}, \boldsymbol{u}) \end{bmatrix} \stackrel{!}{=} \boldsymbol{0} \tag{2.61}$$

von zweiter Ordnung in der Zeit mit den Anfangsbedingungen

$$\boldsymbol{y}(0) = \boldsymbol{y}_0\,, \qquad \dot{\boldsymbol{y}}(0) = \dot{\boldsymbol{y}}_0\,. \tag{2.62}$$

Die verallgemeinerte Massenmatrix steht nun vor den Größen mit zweiter Zeitableitung, zusätzlich tritt noch die in Analogie zur klassischen FEM mit $\boldsymbol{C}$ bezeichnete verallgemeinerte *Dämpfungsmatrix*[10] auf.

Im Rahmen dieser Arbeit werden nur Zeitintegrationsverfahren für Systeme erster Ordnung behandelt. Daher werden je Knoten j zusätzlich zu den Verschiebungen u^d_{Sj} die Festkörpergeschwindigkeiten als Variablen v^d_{Sj} eingeführt. Diese sind allerdings keine Freiheitsgrade im Sinne einer schwachen Formulierung, sondern dienen lediglich der Transformation des Systems (2.61) auf ein System erster Ordnung in der Zeit. Damit erhält man die 2 N zusätzlichen Bestimmungsgleichungen für die Festkörpergeschwindigkeiten an den Knoten:

$$\dot{\mathrm{u}}^d_{Sj} - \mathrm{v}^d_{Sj} = 0\,, \qquad\qquad j = 1, \ldots, \mathrm{N}\,,\ \ d = 1, 2\,. \tag{2.63}$$

Setzt man nun für den Vektor der Variablen an jedem Knoten

$$\tilde{\mathbf{u}}_j = (\mathrm{u}^1_{Sj},\, \mathrm{u}^2_{Sj},\, \mathrm{v}^1_{Sj},\, \mathrm{v}^2_{Sj},\, \mathrm{w}^1_{Fj},\, \mathrm{w}^2_{Fj},\, p_j)^T, \tag{2.64}$$

so erhält man ein System von der gleichen Struktur wie (2.57), allerdings mit einer von der Lösung abhängigen[11] verallgemeinerten Massenmatrix $\tilde{\boldsymbol{M}}(\tilde{\boldsymbol{u}})$. Mit der Abkürzung

$$\begin{pmatrix} c^1_j \\ c^2_j \end{pmatrix} = \sum_{k=1}^{\mathrm{N}} \begin{pmatrix} (\phi^j_{1,1}\,\mathrm{v}^1_{Sj} + \phi^j_{3,1}\,\mathrm{w}^1_{Fj})\,\phi^k_3\,\mathrm{w}^1_{Fk} \;+\; (\phi^j_{1,2}\,\mathrm{v}^1_{Sj} + \phi^j_{3,2}\,\mathrm{w}^1_{Fj})\,\phi^k_4\,\mathrm{w}^2_{Fk} \\ (\phi^j_{2,1}\,\mathrm{v}^2_{Sj} + \phi^j_{4,1}\,\mathrm{w}^2_{Fj})\,\phi^k_3\,\mathrm{w}^1_{Fk} \;+\; (\phi^j_{2,2}\,\mathrm{v}^2_{Sj} + \phi^j_{4,2}\,\mathrm{w}^2_{Fj})\,\phi^k_4\,\mathrm{w}^2_{Fk} \end{pmatrix} \tag{2.65}$$

für den Anteil des Knotens j am Konvektionsterm $\mathbf{c} = \operatorname{grad}(\mathbf{v}_S + \mathbf{w}_F)\,\mathbf{w}_F$ sowie der Mischungsdichte $\rho = \rho^S + \rho^F$ (vgl. (1.18)) erhält man je Knoten i die sieben Gleichungen in Kasten (2.66).

Die analog zu (2.56) aufgebaute Blockmatrix $\tilde{\boldsymbol{M}}$ ist ebenfalls singulär, so daß man es auch im Fall des dynamischen Modells mit einem System differential-algebraischer Gleichungen zu tun hat. Da die letzte Gleichung von (2.66) nicht von den Druck-Variablen p_j abhängt, führt diese Formulierung allerdings bei der Zeitintegration zu Schwierigkeiten (DAE von höherem Index, vgl. Kapitel 3). Analog zum quasi-statischen Fall liefert die Auflösung der starken Form der Impulsbilanz des Fluids nach der Filtergeschwindigkeit $n^F\mathbf{w}_F$ das *Darcy*sche Gesetz, hier jedoch mit Berücksichtigung der Trägheitsterme (vgl. Gleichung (1.88) auf Seite 31):

$$n^F\mathbf{w}_F = -\frac{k^F}{\gamma^{FR}}\left(\operatorname{grad} p - \rho^{FR}\Big[\mathbf{b} - \big((\mathbf{v}_S + \mathbf{w}_F)'_S + \operatorname{grad}(\mathbf{v}_S + \mathbf{w}_F)\,\mathbf{w}_F\big)\Big]\right). \tag{2.67}$$

Setzt man (2.67) in die dritte Gleichung von (2.22) ein, so erhält man mit der Beziehung

[10] Die Dämpfungsmatrix erhält man hier wegen der Anwesenheit eines viskosen Porenfluids auf natürliche Weise. Im Gegensatz dazu wird bei der FEM-Berechnung von dynamischen einphasigen Modellen bei rein mechanischer Betrachtung die in der Realität auftretende und im Modell wegen fehlender Energiebilanz nicht vorhandene Dämpfung meist durch die sog. *Rayleigh*-Dämpfung simuliert (Linearkombination $\boldsymbol{C} = \alpha\boldsymbol{M} + \beta\boldsymbol{K}$ aus Massen- und Steifigkeitsmatrix mit heuristischen Faktoren α und β).

[11] Dies ist nur eine schwache Nichtlinearität aufgrund der Abhängigkeit der Dichten von der Festkörperverschiebung, siehe Gleichung (2.66).

Knoten-Gleichungen des ortsdiskreten dynamischen Modells

$$
\int_{\Omega^h} \sum_{j=1}^{\mathrm{N}} \Bigg[\underbrace{\begin{pmatrix}
\delta_i^j & & & & & & \\
\hdashline
 & \delta_i^j & & & & & \\
\hdashline
 & & \psi_1^i\,\rho\,\phi_1^j & & \psi_1^i\,\rho^F\phi_3^j & & \\
\hdashline
 & & & \psi_2^i\,\rho\,\phi_2^j & & \psi_2^i\,\rho^F\phi_4^j & \\
\hdashline
 & & \psi_3^i\,\rho^F\phi_1^j & & \psi_3^i\,\rho^F\phi_3^j & & \\
\hdashline
 & & & \psi_4^i\,\rho^F\phi_2^j & & \psi_4^i\,\rho^F\phi_4^j & \\
\hdashline
 & & & & & &
\end{pmatrix}}_{\tilde{\mathbf{M}}_{ij}(\tilde{\boldsymbol{u}})} \underbrace{\begin{pmatrix}
\dot{\mathrm{u}}_{Sj}^1 \\ \hdashline \dot{\mathrm{u}}_{Sj}^2 \\ \hdashline \dot{\mathrm{v}}_{Sj}^1 \\ \hdashline \dot{\mathrm{v}}_{Sj}^2 \\ \hdashline \dot{\mathrm{w}}_{Fj}^1 \\ \hdashline \dot{\mathrm{w}}_{Fj}^2 \\ \hdashline \dot{p}_j
\end{pmatrix}}_{\dot{\tilde{\mathbf{u}}}_j} \Bigg] \mathrm{d}v \;+\; \underbrace{\begin{pmatrix}
-\,\delta_i^j\,\mathrm{v}_{Sj}^1 \\ \hdashline
-\,\delta_i^j\,\mathrm{v}_{Sj}^2 \\ \hdashline
\psi_1^i\,(\rho^F[c_j^1-b^1]-\rho^S b^1)+\psi_{1,1}^i(\sigma_{11}-\phi_5^j p_j)+\psi_{1,2}^i\sigma_{12} \\ \hdashline
\psi_2^i\,(\rho^F[c_j^2-b^2]-\rho^S b^2)+\psi_{2,1}^i\sigma_{21}+\psi_{2,2}^i(\sigma_{22}-\phi_5^j p_j) \\ \hdashline
\psi_3^i\,(\rho^F[c_j^1-b^1]+\frac{(n^F)^2\,\gamma^{FR}}{k^F}\,\phi_3^j\mathrm{w}_{Fj}^1)-(\psi_{3,1}^i+\psi_{3,2}^i)\,n^F\phi_5^j p_j \\ \hdashline
\psi_4^i\,(\rho^F[c_j^2-b^2]+\frac{(n^F)^2\,\gamma^{FR}}{k^F}\,\phi_4^j\mathrm{w}_{Fj}^2)-(\psi_{4,1}^i+\psi_{4,2}^i)\,n^F\phi_5^j p_j \\ \hdashline
\psi_5^i\,(\phi_{1,1}^j\mathrm{v}_{Sj}^1+\phi_{2,2}^j\mathrm{v}_{Sj}^2)-(\psi_{5,1}^i\,n^F\phi_3^j\mathrm{w}_{Fj}^1+\psi_{5,2}^i\,n^F\phi_4^j\mathrm{w}_{Fj}^2)
\end{pmatrix}}_{\tilde{\mathbf{k}}_i(\tilde{\boldsymbol{u}},\boldsymbol{q})} = \underbrace{\begin{pmatrix}
0 \\ \hdashline 0 \\ \hdashline \mathrm{f}_i^1 \\ \hdashline \mathrm{f}_i^2 \\ \hdashline \mathrm{f}_i^3 \\ \hdashline \mathrm{f}_i^4 \\ \hdashline \mathrm{f}_i^5
\end{pmatrix}}_{\tilde{\mathbf{f}}_i} \tag{2.66}
$$

$\gamma^{FR} = g\,\rho^{FR}$ eine veränderte schwache Formulierung der Volumenbilanz:

$$
\begin{aligned}
0 \;=\; & \int_{\Omega} \delta p\,\operatorname{div}\mathbf{v}_S\,\mathrm{d}v \\
& + \int_{\Omega} \operatorname{grad}\delta p \cdot \left(\frac{k^F}{\gamma^{FR}}\,\operatorname{grad} p + \frac{k^F}{g}\left[(\mathbf{v}_S+\mathbf{w}_F)'_S + \operatorname{grad}(\mathbf{v}_S+\mathbf{w}_F)\,\mathbf{w}_F - \mathbf{b} \right] \right) \mathrm{d}v \\
& + \int_{\Gamma_v} \delta p\,\bar{v}\,\mathrm{d}a\,.
\end{aligned}
$$

Für die letzte Gleichung von (2.66) ergibt sich damit:

$$\int_{\Omega^h} \sum_{j=1}^{\mathrm{N}} \Bigg[\underbrace{\begin{pmatrix} \delta_i^j & & & & & & \\ \hline \text{s. o.} & \vdots & \text{s. o.} & \vdots & \text{s. o.} & \vdots & \text{s. o.} \\ \hline & & \psi_{5,1}^i \frac{k^F}{g} \phi_1^j & \psi_{5,2}^i \frac{k^F}{g} \phi_2^j & \psi_{5,1}^i \frac{k^F}{g} \phi_3^j & \psi_{5,2}^i \frac{k^F}{g} \phi_4^j & \end{pmatrix}}_{\tilde{\tilde{\mathbf{M}}}_{ij}(\tilde{\boldsymbol{u}})} \underbrace{\begin{pmatrix} \dot{\mathrm{u}}_{Sj}^1 \\ \hline \vdots \\ \hline \dot{p}_j \end{pmatrix}}_{\dot{\tilde{\mathbf{u}}}_j}$$

$$+ \underbrace{\begin{pmatrix} -\delta_i^j \, \mathrm{v}_{Sj}^1 \\ \hline \vdots \quad \text{s. o.} \quad \vdots \\ \hline \psi_5^i \left(\phi_{1,1}^j \mathrm{v}_{Sj}^1 + \phi_{2,2}^j \mathrm{v}_{Sj}^2\right) + \psi_{5,1}^i \left(\frac{k^F}{\gamma^{FR}} \phi_{5,1}^j p_j + \frac{k^F}{g} \left[c_j^1 - b^1\right]\right) \\ + \psi_{5,2}^i \left(\frac{k^F}{\gamma^{FR}} \phi_{5,2}^j p_j + \frac{k^F}{g} \left[c_j^2 - b^2\right]\right) \end{pmatrix}}_{\tilde{\tilde{\mathbf{k}}}_i(\tilde{\boldsymbol{u}}, \boldsymbol{q})} \Bigg] \mathrm{d}v = \underbrace{\begin{pmatrix} 0 \\ \hline \vdots \\ \hline \mathsf{f}_i^5 \end{pmatrix}}_{\tilde{\mathbf{f}}_i} . \tag{2.68}$$

In Kapitel 3 werden die ortsdiskreten Systeme genauer untersucht (Index der DAE). Es wird sich zeigen, daß das System (2.68) wegen seiner numerisch angenehmeren Eigenschaften der ursprünglichen Formulierung (2.66) vorzuziehen ist.

Kapitel 3: Adaptive Zeitintegration

Die Ortsdiskretisierung des inkompressiblen Zweiphasenmodells (Kapitel 2) liefert sowohl im quasi-statischen wie im dynamischen Fall ein System von differential-algebraischen Gleichungen in der Zeit. In diesem Kapitel werden nun geeignete Zeitintegrationsverfahren für derartige Systeme vorgestellt, die zudem bei nur geringem zusätzlichem Aufwand eine Schrittweitensteuerung mit Fehlerkontrolle erlauben (adaptive Zeitintegration).

Die numerische Integration von *gewöhnlichen Differentialgleichungen* (engl. ordinary differential equations, kurz ODE) und *differential-algebraischen Gleichungen* (*Algebrodifferentialgleichungen*, engl. differential-algebraic equations, kurz DAE) ist seit geraumer Zeit intensives Forschungsthema innerhalb der numerischen Mathematik, so daß man auf umfangreiche Literatur zurückgreifen kann. Zu nennen ist etwa das zweibändige Standardwerk von *Hairer et al.* (*Hairer, Nørsett & Wanner* [63], *Hairer & Wanner* [64]), in dem sowohl nicht-steife als auch steife Differentialgleichungen und DAE behandelt werden. *Hairer, Lubich & Roche* [62] beschäftigen sich vor allem mit Konvergenz- und Stabilitätsfragen bei der Anwendung von *Runge-Kutta*-Verfahren auf DAE mit höherem Index, *Brenan, Campbell & Petzold* [26] legen den Schwerpunkt diesbezüglich auf Mehrschrittverfahren (speziell BDF), verbunden mit dem Programmpaket `DASSL`. Aus der deutschsprachigen Literatur sind beispielsweise die Bücher von *Törnig & Spellucci* [115] sowie von *Strehmel & Weiner* [110] zu nennen, die beide einen umfassenden Überblick über das Thema geben.

Im Rahmen dieser Arbeit werden nur *Einschrittverfahren* betrachtet, also Zeitintegrationsverfahren, bei denen die numerische Lösung zur Zeit t_{n+1} lediglich von der Lösung des vorhergehenden Zeitpunkts t_n abhängt. Diese Wahl ist insbesondere im Hinblick auf die adaptive Ortsdiskretisierung (Netzverfeinerung und -vergröberung) von entscheidender Bedeutung, da der Transfer der numerischen Lösung auf diese Weise lediglich zwei Netze einbezieht. Im Fall von Mehrschrittverfahren benötigt man bei der Berechnung des Zeitschritts von t_n nach t_{n+1} die numerische Lösung an zurückliegenden Zeitschritten t_{n-k}, deren Daten alle auf das aktuelle Netz bezogen werden müssen. Dies erfordert die gleichzeitige Vorhaltung mehrerer Netze und die Verwendung mehrerer zugeordneter Transferoperatoren, was einerseits den Speicheraufwand stark erhöht und andererseits bei schnell veränderlicher Ortsdiskretisierung wegen der auftretenden Interpolationsfehler nicht unproblematisch ist.

Die Behandlung von Problemen aus der Plastizitätstheorie mit Mehrschrittverfahren ist jedoch prinzipiell möglich. So werden etwa von *Kirchner & Simeon* [75] Probleme aus dem Bereich der Viskoplastizität auf Basis vereinheitlichter Werkstoffmodelle mit linearen Mehrschrittverfahren und fester Ortsdiskretisierung behandelt.

3.1 Differential-algebraische Gleichungen (DAE)

Treten in einem System sowohl gewöhnliche Differentialgleichungen als auch lineare oder nichtlineare algebraische Gleichungen auf, so spricht man von einem *System differential-*

algebraischer Gleichungen (DAE) oder von einem *Algebrodifferentialgleichungssystem*. Der einfachste Fall kann in der allgemeinen Form (Index-1-System, vgl. Definition 3.1)

$$\begin{array}{rcl} \dot{\boldsymbol{y}} & = & \boldsymbol{f}(t,\boldsymbol{y},\boldsymbol{z}) \\ \mathbf{0} & = & \boldsymbol{g}(t,\boldsymbol{y},\boldsymbol{z}) \end{array}, \qquad \boldsymbol{y}(t_0)=\boldsymbol{y}_0, \quad \boldsymbol{z}(t_0)=\boldsymbol{z}_0, \qquad t \geq t_0 \tag{3.1}$$

charakterisiert werden. Darin sind die differentiellen Variablen im Vektor $\boldsymbol{y}$ und die algebraischen Variablen im Vektor $\boldsymbol{z}$ zusammengefaßt. Nach dem Satz über implizite Funktionen kann die zweite Gleichung lokal eindeutig nach $\boldsymbol{z}$ aufgelöst werden, wenn die Matrix $\partial\boldsymbol{g}/\partial\boldsymbol{z}$ eine beschränkte Inverse besitzt. Außerdem muß für die Anfangswerte gefordert werden, daß sie die Gleichung $\mathbf{0}=\boldsymbol{g}(t_0,\boldsymbol{y}_0,\boldsymbol{z}_0)$ erfüllen. Allgemeiner wird definiert:

Def. 3.1: Ein implizites Anfangswertproblem

$$\boldsymbol{F}(t,\boldsymbol{y},\dot{\boldsymbol{y}})=\mathbf{0}, \qquad \boldsymbol{y}(t_0)=\boldsymbol{y}_0, \qquad t \geq t_0 \tag{3.2}$$

wird als *System differential-algebraischer Gleichungen (DAE)* bezeichnet, wenn die Matrix $\partial\boldsymbol{F}/\partial\dot{\boldsymbol{y}}$ singulär ist. Der *differentielle Index*[1] einer DAE ist die kleinste natürliche Zahl k, für die die Gleichungen

$$\boldsymbol{F}(t,\boldsymbol{y},\dot{\boldsymbol{y}}) \;=\; \mathbf{0}, \quad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}(t,\boldsymbol{y},\dot{\boldsymbol{y}}) \;=\; \mathbf{0}, \quad \ldots \quad \frac{\mathrm{d}^k}{\mathrm{d}t^k}\boldsymbol{F}(t,\boldsymbol{y},\dot{\boldsymbol{y}}) \;=\; \mathbf{0} \tag{3.3}$$

mit algebraischen Operationen in ein System $\dot{\boldsymbol{y}}=\boldsymbol{f}(t,\boldsymbol{y})$ gewöhnlicher Differentialgleichungen überführt werden können (*zugrundeliegendes Differentialgleichungssystem*). ■

Der Index ist eine wichtige Kenngröße zur Charakterisierung von DAE. Als Faustregel gilt: Je höher der Index der DAE, desto schwieriger ist die numerische Lösung des Anfangswertproblems (3.2). Für die Existenz von Lösungen bei DAE sind konsistente Anfangswerte eine unverzichtbare Voraussetzung.

Def. 3.2: Die Anfangswerte des DAE-Anfangswertproblems (3.2) mit dem differentiellen Index k sind *konsistent*, wenn das System

$$\boldsymbol{F}(t_0,\boldsymbol{y}_0,\dot{\boldsymbol{y}}) \;=\; \mathbf{0}, \quad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}(t_0,\boldsymbol{y}_0,\dot{\boldsymbol{y}}) \;=\; \mathbf{0}, \quad \ldots \quad \frac{\mathrm{d}^k}{\mathrm{d}t^k}\boldsymbol{F}(t_0,\boldsymbol{y}_0,\dot{\boldsymbol{y}}) \;=\; \mathbf{0} \tag{3.4}$$

eine Lösung $\dot{\boldsymbol{y}}(t_0,\boldsymbol{y}_0)=\dot{\boldsymbol{y}}_0$ hat. ■

3.1.1 Differentieller Index der ortsdiskreten Systeme

Für die Auswahl geeigneter numerischer Verfahren spielt der Index des behandelten DAE-Systems eine entscheidende Rolle. In diesem Abschnitt wird daher der Index des ortsdiskreten Zweiphasenmodells aus Kapitel 2 im elastischen Fall bei Verwendung des *Galerkin*-Verfahrens bestimmt.

[1] Von *Hairer, Lubich & Roche* [62] wurde neben dem differentiellen Index *di* der *Störungsindex pi* eingeführt, der die Anfälligkeit des diskretisierten Systems gegenüber Rundungsfehlern berücksichtigt. Im allgemeinen sind Ungleichungsaussagen über den Zusammenhang zwischen *di* und *pi* schwierig. Von *Campbell & Gear* [29] wurden dazu detaillierte Untersuchungen durchgeführt.

Quasi-statisches Modell

Es wird zunächst Gleichung (2.53) für $\boldsymbol{q} \equiv \boldsymbol{0}$ (rein elastisches Problem) untersucht, also das System der FEM-Freiheitsgrade ohne Berücksichtigung der internen Variablen. Im folgenden wird gezeigt, daß dieses DAE-System den differentiellen Index 1 besitzt.

Dazu werden die ersten beiden Gleichungen (ortsdiskrete Impulsbilanz der Mischung) nach der Zeit t abgeleitet und die dritte Gleichung (ortsdiskrete Volumenbilanz) mit -1 multipliziert. Berücksichtigt man dann noch $\boldsymbol{\psi}^i = \boldsymbol{\phi}^i$ (*Galerkin*-Verfahren), so erhält man das zugrundeliegende Differentialgleichungssystem ($i = 1, \ldots, \mathrm{N}$):

$$\underbrace{\sum_{j=1}^{\mathrm{N}} \left(\begin{array}{cc:c} \tilde{\mathsf{K}}_{ij}^{11} & \tilde{\mathsf{K}}_{ij}^{12} & -(\phi_{1,1}^i,\, \phi_3^j) \\ \tilde{\mathsf{K}}_{ij}^{21} & \tilde{\mathsf{K}}_{ij}^{22} & -(\phi_{2,2}^i,\, \phi_3^j) \\ \hdashline -(\phi_{1,1}^i,\, \phi_3^j) & -(\phi_{2,2}^i,\, \phi_3^j) & 0 \end{array} \right)}_{\tilde{\mathbf{M}}_{ij}} \underbrace{\left(\begin{array}{c} \dot{\mathrm{u}}_{Sj}^1 \\ \dot{\mathrm{u}}_{Sj}^2 \\ \dot{p}_j \end{array} \right)}_{\dot{\mathbf{u}}_j} = \underbrace{\left(\begin{array}{c} \mathsf{r}_i^1(\boldsymbol{u}) \\ \mathsf{r}_i^2(\boldsymbol{u}) \\ \mathsf{r}_i^3(\boldsymbol{u}) \end{array} \right)}_{\mathbf{r}_i(\boldsymbol{u})}, \tag{3.5}$$

wobei in $\boldsymbol{r}(\boldsymbol{u})$ alle verbleibenden Terme aus dem Vektor $\boldsymbol{k}(\boldsymbol{u})$ sowie der rechten Seite $\boldsymbol{f}$ zusammengefaßt sind. Es muß also gezeigt werden, daß die Matrix $\tilde{\boldsymbol{M}}$ regulär ist.

Für die Integrale wurde zur Abkürzung das L^2-Skalarprodukt $(\cdot, \cdot)$ aus Gleichung (2.5) verwendet. Die 2×2-Blockmatrix $\tilde{\mathbf{K}}_{ij}$ links oben entsteht durch Ableitung der Spannungsterme nach den Verschiebungen, entspricht also im hier betrachteten elastischen Fall genau dem Anteil einer klassischen Steifigkeitsmatrix, der die Knoten i und j koppelt. Es gilt mit dem Operator $\boldsymbol{L} = \frac{1}{2}(\mathrm{grad} + \mathrm{grad}^T)$ sowie $\tilde{\boldsymbol{\phi}}^i := \mathrm{diag}(\phi_1^i, \phi_2^i)$ und $\tilde{\mathbf{u}}_j := (u_{Sj}^1,\, u_{Sj}^2)^T$:

$$\tilde{\mathbf{K}}_{ij} = \frac{\partial}{\partial \tilde{\mathbf{u}}_j} \left[\sum_{k=1}^{\mathrm{N}} (\boldsymbol{L}\, \tilde{\boldsymbol{\phi}}^i,\, \overset{4}{\mathbf{C}}\, \boldsymbol{L}\, \tilde{\boldsymbol{\phi}}^k\, \tilde{\mathbf{u}}_k) \right] = (\boldsymbol{L}\, \tilde{\boldsymbol{\phi}}^i,\, \overset{4}{\mathbf{C}}\, \boldsymbol{L}\, \tilde{\boldsymbol{\phi}}^j).$$

Numeriert man die Unbekannten $\boldsymbol{u} = (\mathrm{u}_{S1}^1, \mathrm{u}_{S1}^2, \ldots, \mathrm{u}_{S\mathrm{N}}^1, \mathrm{u}_{S\mathrm{N}}^2,\ p_1, \ldots, p_{\mathrm{N}})^T$ entsprechend der Blockstruktur von (3.5), so erhält man für die Gesamtmatrix

$$\tilde{\boldsymbol{M}} = \begin{pmatrix} \tilde{\boldsymbol{K}} & \boldsymbol{P} \\ \boldsymbol{P}^T & \boldsymbol{0} \end{pmatrix}.$$

Die Matrix $\tilde{\boldsymbol{K}}$ ist die Gesamtsteifigkeitsmatrix eines linear elastischen Problems und somit bei geeigneten Materialparametern und Randbedingungen positiv definit. Die Matrix $\boldsymbol{P}$ ist eine $2\,\mathrm{N} \times \mathrm{M}$-Matrix mit Skalarprodukten $(\phi_{k,k}^i,\, \phi_3^j)$ aus Verschiebungs-Testfunktionen (N Verschiebungs-Knoten) und Druck-Ansatzfunktionen (M Druck-Knoten).

Es soll nun mit Hilfe einer Schurkomplement-Darstellung gezeigt werden, daß die Matrix $\tilde{\boldsymbol{M}}$ regulär ist. Das Schurkomplement einer Blockmatrix kann als blockweises Anwenden des *Gauß*-Algorithmus interpretiert werden und ist wie folgt definiert:

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} \boldsymbol{I} & \boldsymbol{B}' \\ \boldsymbol{0} & \boldsymbol{S} \end{pmatrix} \quad \text{mit} \quad \boldsymbol{B}' := \boldsymbol{A}^{-1}\, \boldsymbol{B}, \quad \boldsymbol{S} := \boldsymbol{D} - \boldsymbol{C}\, \boldsymbol{A}^{-1}\, \boldsymbol{B}.$$

Die Blockmatrix auf der linken Seite ist genau dann regulär, wenn die Matrix $\boldsymbol{S}$ (das *Schurkomplement*) regulär ist. Im vorliegenden Fall gilt:

$$\boldsymbol{S} = -\boldsymbol{P}^T \tilde{\boldsymbol{K}}^{-1} \boldsymbol{P}.$$

Statt der Regularität von $\boldsymbol{S}$ wird die positive Definitheit von $-\boldsymbol{S}$ gezeigt. Dazu muß gelten:

$$\underbrace{\boldsymbol{x}^T \boldsymbol{P}^T}_{\boldsymbol{y}^T} \tilde{\boldsymbol{K}}^{-1} \underbrace{\boldsymbol{P}\,\boldsymbol{x}}_{\boldsymbol{y}} > 0 \qquad \text{für} \qquad \boldsymbol{x} \neq \boldsymbol{0}.$$

Da die Matrix $\tilde{\boldsymbol{K}}^{-1}$ als Inverse einer positiv definiten Matrix positiv definit ist, gilt obige Aussage für alle $\boldsymbol{y} \neq \boldsymbol{0}$. Es bleibt also zu zeigen, daß die Matrix $\boldsymbol{P}$ von Null verschiedene Vektoren $\boldsymbol{x}$ nicht auf den Nullvektor abbildet. Anders ausgedrückt muß $\boldsymbol{P}$ den Zeilenrang M (Anzahl der Druckfreiheitsgrade) besitzen. Dies folgt aber gerade aus der linearen Unabhängigkeit der Ansatz- und Testfunktionen des FEM-Ansatzes. Betrachtet man nämlich die Zeilen der Matrix $\boldsymbol{P}$ in einer Raumrichtung $d \in \{1, 2\}$, so lautet eine beliebige Linearkombination mit einer Indexmenge I der Größe $|I| = \mathrm{M} < \mathrm{N}$ dieser Zeilen:

$$\sum_{i \in I} \alpha_i \left((\phi^i_{d,d},\ \phi^j_3) \right)_{j=1,\ldots,\mathrm{M}} = \left((\sum_{i \in I} \alpha_i\, \phi^i_{d,d},\ \phi^j_3) \right)_{j=1,\ldots,\mathrm{M}}.$$

Dies kann wegen der linearen Unabhängigkeit der ϕ^i_d bzw. ϕ^j_3 (der Träger liegt jeweils über unterschiedlichen Elementen) nur dann der Nullvektor sein, wenn alle α_i verschwinden.

Die plastischen Entwicklungsgleichungen (Viskoplastizität) sind gewöhnliche Differentialgleichungen, stellen also für die Betrachtung des Index kein Problem dar[2]. Allerdings muß vorausgesetzt werden, daß der Materialtensor $\mathbf{C}$ auch im viskoplastischen Fall positiv definit bleibt. Dies ist aber ohnehin eine notwendige Bedingung für die Lösbarkeit des Problems. Damit ist gezeigt, daß das ortsdiskrete quasi-statische Modell in der Formulierung (2.53) den differentiellen Index 1 besitzt.

Dynamisches Modell

Faßt man im ortsdiskreten dynamischen Modell (2.66) alle Variablen außer dem Druck im Vektor $\boldsymbol{y}$ und die Druck-Variablen im Vektor $\boldsymbol{z}$ zusammen, so ist (2.66) nach geeigneter Umsortierung der Variablen und Gleichungen (jeweils alle gleichartigen Knoten-Variablen und -Gleichungen hintereinander, also zuerst alle Verschiebungen, dann alle Geschwindigkeiten usw.) ein System der Struktur

$$\begin{aligned} \boldsymbol{M}\,\dot{\boldsymbol{y}} &= \boldsymbol{f}(t, \boldsymbol{y}, \boldsymbol{z})\,, \\ \boldsymbol{0} &= \boldsymbol{g}(t, \boldsymbol{y}) \end{aligned} \tag{3.6}$$

[2]Beim elastoplastischen Modell erhält man formal ein System vom differentiellen Index 2, da der Multiplikator Λ in der Fließbedingung (algebraische Gleichung) nicht vorkommt. Nach Wissen des Verfassers gibt es bisher in der Literatur keine mathematischen Untersuchungen, die den für die Numerik wichtigen Störungsindex von Systemen der Elastoplastizität bestimmen. Diese Problematik wird hier jedoch nicht weiter betrachtet, da im Rahmen dieser Arbeit stets das viskoplastische Modell eingesetzt wird.

mit einer regulären Matrix $\boldsymbol{M}$. Nach *Hairer & Wanner* [64] besitzt ein solches System den differentiellen Index 2, wenn die Matrix $(\partial \boldsymbol{g}/\partial \boldsymbol{y})\,\boldsymbol{M}^{-1}\,(\partial \boldsymbol{f}/\partial \boldsymbol{z})$ regulär ist. Ohne weitere Untersuchungen läßt sich also feststellen, daß das dynamische Modell in der Formulierung (2.66) mindestens den Index 2 besitzt.

Im folgenden soll nun der Weg zur Index-Bestimmung der alternativen Formulierung (2.68) aufgezeigt werden. Dazu wird zunächst die siebte Gleichung (ortsdiskrete Volumenbilanz, im folgenden abgekürzt mit F_7^i) mit Hilfe der Gleichungen 5 und 6 (ortsdiskrete Impulsbilanz des Fluids, abgekürzt mit F_5^i und F_6^i) so umgeformt, daß der Beschleunigungsterm $\dot{\mathbf{v}}_S + \dot{\mathbf{w}}_F + \mathbf{c} - \mathbf{b}$ herausfällt. Dies erreicht man durch die algebraische Operation

$$\tilde{F}_7^i := F_7^i - \frac{k^F}{n^F \gamma^{FR}} \left[\frac{\psi_{5,1}^i}{\psi_3^i}\, F_5^i + \frac{\psi_{5,2}^i}{\psi_4^i}\, F_6^i \right], \tag{3.7}$$

so daß die transformierte ortsdiskrete Volumenbilanz nach einigen Umformungen lautet:

$$\begin{aligned} \int_{\Omega^h} \sum_{j=1}^{\mathrm{N}} \Bigg[& \psi_5^i \left(\phi_{1,1}^j \mathrm{v}_{Sj}^1 + \phi_{2,2}^j \mathrm{v}_{Sj}^2\right) \\ & + \psi_{5,1}^i \left(\frac{k^F}{\gamma^{FR}} \left(\phi_{5,1}^j p_j + \frac{\psi_{3,1}^i + \psi_{3,2}^i}{\psi_3^i}\, \phi_5^j p_j \right) - n^F\, \phi_3^j \mathrm{w}_{Fj}^1 \right) \\ & + \psi_{5,2}^i \left(\frac{k^F}{\gamma^{FR}} \left(\phi_{5,2}^j p_j + \frac{\psi_{4,1}^i + \psi_{4,2}^i}{\psi_4^i}\, \phi_5^j p_j \right) - n^F\, \phi_4^j \mathrm{w}_{Fj}^2 \right) \Bigg] \,\mathrm{d}v \;=\; \tilde{\mathsf{f}}_5^i \,. \end{aligned} \tag{3.8}$$

Darin ist $\tilde{\mathsf{f}}_5^i$ die entsprechend (3.7) transformierte rechte Seite. Mit der gleichen Wahl für $\boldsymbol{y}$ und $\boldsymbol{z}$ wie oben hat das resultierende System die Struktur

$$\begin{aligned} \boldsymbol{M}\,\dot{\boldsymbol{y}} &= \boldsymbol{f}(t, \boldsymbol{y}, \boldsymbol{z}) \,, \\ \mathbf{0} &= \boldsymbol{g}(t, \boldsymbol{y}, \boldsymbol{z}) \,. \end{aligned} \tag{3.9}$$

DAE-Systeme dieser Struktur besitzen den Index 1, wenn die Matrizen $\boldsymbol{M}$ und $\partial \boldsymbol{g}/\partial \boldsymbol{z}$ regulär sind. In der Praxis zeigte sich, daß die Formulierung (2.68) mit den später vorgestellten diagonal-impliziten Verfahren problemlos behandelt werden konnte, während die ursprüngliche Formulierung (2.66) zum Abbruch des Verfahrens führte. Daher liegt die Vermutung nahe, daß der Index der DAE durch die o. g. Transformation tatsächlich reduziert wird. Auf detaillierte Untersuchungen wird an dieser Stelle jedoch verzichtet.

Fazit

Bei der Integration von DAE mit höherem Index ($di \geq 2$) kommt es bei vielen Verfahren zur *Ordnungsreduktion* in den algebraischen Variablen. Dieses Phänomen wurde in bezug auf *Runge-Kutta*-Verfahren insbesondere von *Hairer, Lubich & Roche* [62] untersucht und führte zur Entwicklung spezieller Verfahren für solche DAE (insbesondere implizite *Runge-Kutta*-Verfahren vom *Gauß*-Typ).

Im Fall des vorliegenden quasi-statischen Modells konnte oben der Index 1 nachgewiesen werden. Beim dynamischen Modell wird stets die Formulierung (2.68) verwendet, deren Index 1 (ohne Nachweis, s. o.) im folgenden angenommen wird. Bei der späteren Auswahl geeigneter Verfahren kann daher eine große Klasse impliziter *Runge-Kutta*-Verfahren in die Überlegungen einbezogen werden.

3.2 Steife Differentialgleichungen

Bevor auf die numerische Behandlung von Index-1-Systemen mit impliziten *Runge-Kutta*-Verfahren eingegangen wird, soll zunächst noch das wichtige Thema der *steifen Differentialgleichungen* angesprochen werden, das insbesondere bei Ortsdiskretisierungen im Rahmen der Linienmethode eine wichtige Rolle spielt und bei der Auswahl geeigneter Verfahren (Stabilität) berücksichtigt werden muß.

Kurz gesagt sind steife Differentialgleichungen solche, bei denen explizite Verfahren „versagen“. Versucht man etwa, die sehr einfache Differentialgleichung $\dot{y}(t) = -50(y(t) - \cos t)$ ausgehend von $y(0) = 0$ mit dem expliziten *Euler*-Verfahren zu integrieren, so erhält man erst für sehr kleine Zeitschrittweiten stabile Lösungen. Das Beispiel in Abbildung 3.1 ist *Hairer & Wanner* [64] entnommen und zeigt links das Feld der Trajektorien mit der Näherung durch das implizite *Euler*-Verfahren mit verhältnismäßig großer Schrittweite und rechts zwei Lösungen des expliziten *Euler*-Verfahrens mit kleinen Schrittweiten. Man erkennt bereits, daß die numerische Lösung mit der kleineren Schrittweite weniger stark schwingt. Wie sich noch zeigen wird, ist das explizite *Euler*-Verfahren bei diesem Problem für Schrittweiten $h > 2/50$ instabil.

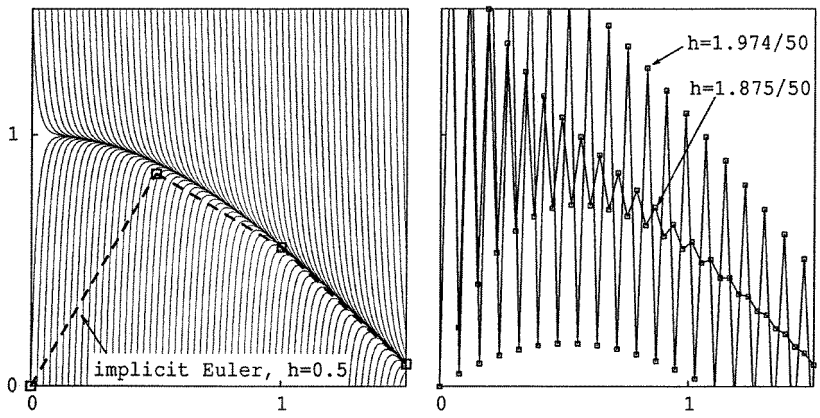


Abbildung 3.1: *Eingeschränkte Stabilität des expliziten Euler-Verfahrens* (Beispiel aus *Hairer & Wanner* [64])

Steife Differentialgleichungen treten in vielen Anwendungsbereichen auf. Die ersten Beispiele entstammten der numerischen Simulation chemischer Reaktionen. Charakteristisch ist bei Systemen von Differentialgleichungen in jedem Fall, daß es sowohl schnell veränderliche (transiente) als auch träge Lösungskomponenten gibt, was meist durch stark unterschiedliche Beträge der Eigenwerte gekennzeichnet ist.

Bei der Methode der finiten Elemente führt die Anwendung der *Linienmethode* (engl. method of lines, kurz MOL), also die schrittweise Zeitintegration einer festen Ortsdiskretisierung, ebenfalls zu steifen Differentialgleichungen. Die Steifheit („Bösartigkeit“) wird z. B. im Fall der Ortsdiskretisierung einer partiellen Differentialgleichung 2. Ordnung in zwei Raumdimensionen maßgeblich durch den Faktor $1/h^2$ bestimmt, wobei h ein Maß für die Feinheit der Ortsdiskretisierung darstellt. Mit feiner werdenden Netzen nimmt also die Steifheit zu, so daß man insbesondere bei adaptiver Netzverfeinerung auf robuste

Verfahren zur Zeitintegration der ortsdiskreten Gleichungen angewiesen ist.

3.2.1 A-, L- und S-Stabilität

Zur Stabilitätsuntersuchung von Zeitintegrationsverfahren bei steifen Differentialgleichungen wurde von *Dahlquist* das lineare Modellproblem

$$\dot{y}(t) = \lambda\, y(t)\,, \quad y(0) = 1\,, \qquad t \geq 0 \tag{3.10}$$

mit $\lambda \in \mathbb{C}$ eingeführt. Die Anwendung eines Einschrittverfahrens vom *Runge-Kutta*-Typ auf dieses Problem führt zu einer Rekursionsgleichung

$$y_{n+1} = R(z)\, y_n\,, \quad y_0 = 1\,, \qquad n = 0, 1, 2, \ldots\,, \tag{3.11}$$

mit $z = h\,\lambda$. Die rationale Funktion $R(z)$ heißt *Stabilitätsfunktion* und charakterisiert das Verfahren in bezug auf seine Stabilitätseigenschaften.

Def. 3.3: Ein Einschrittverfahren heißt *A-stabil*, wenn der *Stabilitätsbereich*

$$S := \{\, z \in \mathbb{C} \;:\; |R(z)| \leq 1 \,\} \tag{3.12}$$

die komplexe linke Halbebene umfaßt:

$$S \supset \mathbb{C}^- := \{\, z \in \mathbb{C} \;:\; \Re(z) \leq 0 \,\}\,. \tag{3.13}$$

Darin bezeichnet $\Re(z)$ den Realteil einer komplexen Zahl z. ■

Die exakte Lösung $y(t) = e^{\lambda t}$ des Modellproblems (3.10) klingt für komplexe λ mit negativem Realteil im Laufe der Zeit ab. Die A-Stabilität beschreibt nun die wünschenswerte Eigenschaft, daß die numerische Lösung ebenfalls abklingt bzw. wenigstens beschränkt bleibt. Dies wird gerade dadurch ausgedrückt, daß der Betrag der Stabilitätsfunktion in der Rekursionsgleichung (3.11) kleiner oder gleich Eins ist.

Bemerkung: Das „Versagen“ der expliziten Verfahren bei steifen Differentialgleichungen liegt daran, daß sie nicht A-stabil sind. So ist z. B. die Stabilitätsfunktion des expliziten *Euler*-Verfahrens wegen $y_{n+1} = y_n + h\lambda\, y_n = (1 + h\lambda)\, y_n$ die rationale Funktion $R(z) = 1 + z$. Der Stabilitätsbereich $S = \{\, z \in \mathbb{C} \;:\; |1 + z| \leq 1 \,\}$ ist der Kreis mit Radius 1 um den Punkt -1. Man erhält stabile Lösungen für $h\lambda \in S$, also für λ mit negativem Realteil unter der Bedingung $0 < h \leq 2/|\lambda|$. Im Beispiel in Abbildung 3.1 ist $\lambda = -50$, so daß man stabile Lösungen für $h \leq 2/50$ erhält. Solche Schrittweiteneinschränkungen als Bedingung für stabile Lösungen sind typisch für alle expliziten Verfahren (*Hairer & Wanner* [64, IV.2]), bei denen der Grad des Zählerpolynoms der rationalen Funktion $R(z)$ immer größer als der Grad des Nennerpolynoms ist. □

Bei Problemen mit betragsmäßig großen Eigenwerten (sehr steife Probleme) ist außerdem eine gewisse numerische Dämpfung zur Stabilisierung erwünscht. So führt beispielsweise die Anwendung der (A-stabilen) Trapezregel auf solche Probleme dazu, daß die steifen

Komponenten der Lösung nur sehr langsam abklingen. Das Resultat ist, daß die numerische Lösung in der transienten Phase große Schwingungen um die exakte Lösung herum aufweist. Dies führte zur Einführung eines weiteren Stabilitätsbegriffs.

Def. 3.4: Ein Einschrittverfahren heißt *L-stabil* (auch *stark A-stabil*, *steif A-stabil*), wenn es A-stabil ist und die Bedingung

$$\lim_{\Re(z)\to-\infty} R(z) = 0 \tag{3.14}$$

erfüllt. ■

Das eben beschriebene Verhalten der Trapezregel bei Anwendung auf sehr steife Probleme liegt darin begründet, daß die Trapezregel nicht L-stabil ist. Mit der Stabilitätsfunktion $R(z) = \frac{1+z/2}{1-z/2}$ erhält man nämlich $\lim_{\Re(z)\to-\infty} R(z) = -1$.

Neben diesen beiden für steife Differentialgleichungen unverzichtbaren Stabilitätseigenschaften wurden in den letzten Jahren weitere Stabilitätsbegriffe entwickelt. Insbesondere bei nichtlinearen Problemen ist das lineare Modellproblem (3.10) zu einfach, um die erforderlichen Eigenschaften numerischer Verfahren charakterisieren zu können. Für dissipative Probleme wurde daher das Konzept der A-Stabilität zu einem nichtlinearen Stabilitätsbegriff, der B-Stabilität, verallgemeinert. Im Zusammenhang mit großen nichtlinearen steifen Systemen spielen noch die Begriffe der S-Stabilität und der B-Konvergenz eine wichtige Rolle. Diese und weitere Stabilitätsbegriffe können der Literatur entnommen werden, z. B. *Hairer & Wanner* [64]. Hier wird wegen seiner Bedeutung für große nichtlineare Systeme nur noch der von *Prothero & Robinson* [96] eingeführte Begriff der S-Stabilität behandelt.

Def. 3.5: Betrachtet wird das gegenüber (3.10) um eine Funktion $g \in C^1([0,T])$ erweiterte Modellproblem

$$\dot{y}(t) = \dot{g}(t) + \lambda(y(t) - g(t))\,, \quad y(0) = y_0\,, \qquad \Re(\lambda) < 0\,. \tag{3.15}$$

Ein Einschrittverfahren heißt *S-stabil*, wenn es für jedes g und jede Konstante $\lambda_0 < 0$ eine Konstante $h_0 > 0$ gibt, so daß für die numerische Lösung $\{y_n\}$ von (3.15) auf dem Intervall $[0,T]$ für alle $0 < h < h_0$ mit $[t_n, t_n + h] \subset [0,T]$ und alle $\lambda \in \mathbb{C}$ mit $\Re(\lambda) \le \lambda_0$ gilt:

$$\left|\frac{y_{n+1} - g(t_{n+1})}{y_n - g(t_n)}\right| < 1\,, \qquad \text{falls } y_n \neq g(t_n)\,. \tag{3.16}$$

Das Verfahren heißt *stark S-stabil* (*steif S-stabil*), wenn zusätzlich gilt:

$$\lim_{\Re(\lambda)\to-\infty} \frac{y_{n+1} - g(t_{n+1})}{y_n - g(t_n)} = 0\,. \tag{3.17}$$

■

Die exakte Lösung von (3.15) ist $y(t) = g(t) + (y_0 - g(0))\,e^{\lambda t}$, der Ausdruck $y(t) - g(t)$ klingt also für $\Re(\lambda) < 0$ ab. Wie bei der A-Stabilität ist nun die Aussage der S-Stabilität, daß auch die numerische Lösung dieses Abklingverhalten aufweist. Die Eigenschaft der S-Stabilität ist also eine Erweiterung der A-Stabilität (man setze $g \equiv 0$):

$$\text{S-Stabilität} \quad \Longrightarrow \quad \text{A-Stabilität}.$$

3.3 *Runge-Kutta*-Verfahren

Die wichtigste Klasse der Einschrittverfahren für Differentialgleichungen erster Ordnung bilden die *Runge-Kutta*-Verfahren. Zunächst werden die Begriffe für Systeme gewöhnlicher Differentialgleichungen eingeführt und anschließend auf DAE-Systeme verallgemeinert, bei denen insbesondere steif genaue Verfahren eine wichtige Rolle spielen.

3.3.1 Formulierung für gewöhnliche Differentialgleichungen

Am Beispiel eines Anfangswertproblems gewöhnlicher Differentialgleichungen,

$$\dot{\boldsymbol{y}}(t) = \boldsymbol{f}(t, \boldsymbol{y}(t))\,, \quad \boldsymbol{y}(0) = \boldsymbol{y}_0\,, \qquad \boldsymbol{y} \in \mathbb{R}^m\,, \quad t \in [0, T]\,, \tag{3.18}$$

soll kurz das Konstruktionsprinzip für diese Verfahren erläutert werden. Ausgehend von einem gedachten Punkt $(t_n, \boldsymbol{y}(t_n))$ auf der exakten Lösung und einer *Schrittweite* $h_n > 0$ wird das Integral im Hauptsatz der Differential- und Integralrechnung auf das Intervall $[0, 1]$ transformiert und $\dot{\boldsymbol{y}}$ durch die rechte Seite $\boldsymbol{f}$ ersetzt:

$$\boldsymbol{y}(t_n + h_n) \;=\; \boldsymbol{y}(t_n) + \int_{t_n}^{t_n+h_n} \dot{\boldsymbol{y}}(t)\,\mathrm{d}t \;=\; \boldsymbol{y}(t_n) + h_n \int_0^1 \boldsymbol{f}(t_n + \tau h_n,\, \boldsymbol{y}(t_n + \tau h_n))\,\mathrm{d}\tau\,. \tag{3.19}$$

Die Näherung des Integrals durch eine Quadraturformel mit Stützstellen c_i und Gewichten b_i führt dann zu neuen unbekannten Größen $\boldsymbol{y}(t_n + c_i h_n)$. Man könnte nun in (3.19) h_n durch $c_i h_n$ ersetzen und die neuen Integrale wieder durch Quadraturformeln nähern, was aber erneut zu unbekannten Größen führen würde. Deshalb wählt man für die Näherung der „inneren“ Integrale Quadraturformeln mit den gleichen Stützstellen c_i, aber neuen Gewichten a_{ij}.

Def. 3.6: Das Intervall $[0, T]$ wird mit *Schrittweiten* $h_n = t_{n+1} - t_n > 0$ in Teilintervalle $[t_n, t_{n+1}]$ zerlegt: $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$. Ein s-stufiges *Runge-Kutta-Verfahren* berechnet für $n = 0, 1, 2, \ldots, N-1$ ausgehend von der Näherungslösung $\boldsymbol{y}_n$ zur Zeit t_n die Näherungslösung $\boldsymbol{y}_{n+1}$ zur Zeit $t_{n+1} = t_n + h_n$ wie folgt:

1. Löse das (nichtlineare) Gleichungssystem der Dimension $s \cdot m$ in den Größen $\dot{\boldsymbol{Y}}_{ni}$:

$$\dot{\boldsymbol{Y}}_{ni} = \boldsymbol{f}(t_n + c_i h_n,\, \boldsymbol{y}_n + h_n \sum_{j=1}^{s} a_{ij} \dot{\boldsymbol{Y}}_{nj})\,, \qquad i = 1, \ldots, s\,. \tag{3.20}$$

2. Berechne die neue Näherungslösung:

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h_n \sum_{i=1}^{s} b_i \dot{\boldsymbol{Y}}_{ni}\,. \tag{3.21}$$

Die Größen $\dot{\boldsymbol{Y}}_{ni}$ sind Näherungen für $\dot{\boldsymbol{y}}(t_n + c_i h_n)$ und werden als *Stufenableitungen* bezeichnet (in der Literatur findet man diese auch als *Steigungswerte* $\boldsymbol{k}_{ni}$). ■

Bemerkung: Alternativ können auf den Stufen statt der Stufenableitungen auch die *Stufenlösungen* $\boldsymbol{Y}_{ni}$ berechnet werden. Das Verfahren lautet dann:

1. Löse das (nichtlineare) Gleichungssystem der Dimension $s \cdot m$ in den Größen $\boldsymbol{Y}_{ni}$:

$$\boldsymbol{Y}_{ni} = \boldsymbol{y}_n + h_n \sum_{j=1}^{s} a_{ij} \boldsymbol{f}(t_n + c_j h_n,\ \boldsymbol{Y}_{nj})\,, \qquad i = 1, \ldots, s\,. \tag{3.22}$$

2. Berechne die neue Näherungslösung:

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h_n \sum_{i=1}^{s} b_i \boldsymbol{f}(t_n + c_i h_n,\ \boldsymbol{Y}_{ni})\,. \tag{3.23}$$

Sind die inneren Gewichte der letzten Stufe verschieden von den äußeren Gewichten, $a_{si} \neq b_i$, so erfordert dieses Vorgehen gegenüber der obigen Berechnung der Stufenableitungen s zusätzliche Funktionsauswertungen. Dies kann man nur mit zusätzlichem Speicheraufwand vermeiden, indem während der Lösung des Gleichungssystems die Größen $\dot{\boldsymbol{Y}}_{nj} = \boldsymbol{f}(t_n + c_j h_n,\ \boldsymbol{Y}_{nj})$ abgespeichert werden. □

Def. 3.7: Die Koeffizienten eines *Runge-Kutta*-Verfahrens werden üblicherweise im *Butcher-Schema* zusammengefaßt:

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} \qquad \text{oder kurz} \qquad \begin{array}{c|c} \boldsymbol{c} & \boldsymbol{A} \\ \hline & \boldsymbol{b}^T \end{array}. \tag{3.24}$$

Ein *Runge-Kutta*-Verfahren heißt ...

... *explizit* (ERK), wenn $\boldsymbol{A}$ strikte untere Dreiecksmatrix ist: $a_{ij} = 0$ für $j \geq i$,

... *implizit* (IRK), wenn mindestens ein $a_{ij} \neq 0$ für $j \geq i$.

Ein implizites *Runge-Kutta*-Verfahren heißt ...

... *diagonal-implizit* (DIRK), wenn $\boldsymbol{A}$ untere Dreiecksmatrix ist: $a_{ij} = 0$ für $j > i$,

... *einfach diagonal-implizit* (SDIRK), wenn zusätzlich alle a_{ii} gleich sind,

... *einfach implizit* (SIRK), wenn alle Eigenwerte von $\boldsymbol{A}$ gleich sind.

In Klammern sind jeweils gebräuchliche Abkürzungen der englischen Bezeichnungen angegeben (SDIRK steht z. B. für Singly Diagonally Implicit *Runge Kutta*). ■

Die expliziten Verfahren haben den Vorteil, daß alle Stufen nacheinander direkt berechnet werden können, was allerdings mit dem Nachteil einer bedingten Stabilität erkauft wird (vgl. Abschnitt 3.2). Bei impliziten Verfahren sind grundsätzlich Gleichungssysteme zu lösen, wobei im diagonal-impliziten Fall die Stufengrößen entkoppelt berechnet werden können (s Gleichungssysteme der Dimension m statt eines Gleichungssystems der Dimension $s \cdot m$). Dies ist gerade im Hinblick auf große Systeme (z. B. finite Elemente) ein entscheidender Vorteil gegenüber voll-impliziten Verfahren. Die SDIRK-Verfahren erlauben zudem bei (schwach) nichtlinearen Systemen die Anwendung des vereinfachten *Newton*-Verfahrens, also die Verwendung einer einmal faktorisierten Matrix zum Lösen aller Stufen-Systeme eines Zeitschritts.

Bei der Wahl der Koeffizienten im *Butcher*-Schema hat man viele Freiheiten. Zunächst wird man versuchen, mittels Taylorentwicklungen eine möglichst hohe Ordnung des Verfahrens zu erreichen, wobei man bei impliziten Verfahren außerdem an geeigneten Stabilitätseigenschaften interessiert ist. Weitere Kriterien sind ein möglichst kleiner Vor-

Name	*Butcher*-Schema	Typ	s	p	Stabilität
Euler explizit	$\begin{array}{c\|c} 0 & 0 \\ \hline & 1 \end{array}$	ERK	1	1	–
Modif. *Euler*	$\begin{array}{c\|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$	ERK	2	2	–
Runge-Kutta	$\begin{array}{c\|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$	ERK	4	4	–
Euler implizit	$\begin{array}{c\|c} 1 & 1 \\ \hline & 1 \end{array}$	IRK	1	1	A,L
Implizite Mittelpunktregel	$\begin{array}{c\|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$	DIRK	1	2	A
Trapezregel (*Crank-Nicholson*)	$\begin{array}{c\|cc} 0 & 0 & \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$	IRK (nur 1 implizite Stufe!)	2	2	A

Tabelle 3.1: *Einige gebräuchliche Runge-Kutta-Verfahren (Stufenzahl s, Ordnung p)*

faktor vor dem Restglied in der Taylorentwicklung bzw. die Reduktion des numerischen Aufwands bei der Berechnung der Stufengrößen. Auf die Konstruktion spezieller Verfahren wird hier jedoch nicht weiter eingegangen. In Tabelle 3.1 sind eine Reihe bekannter

Verfahren jeweils mit Angabe der Stufenzahl s und der Ordnung p sowie der Stabilitätseigenschaften angegeben. Die Trapezregel ist dabei vom Rechenaufwand her ein einstufiges Verfahren, die erste Stufe dient nur der Notation im *Butcher*-Schema.

Bemerkung: In der Literatur findet man *Runge-Kutta*-Verfahren häufig auch unter anderem Namen. Beispiele hierfür sind die „θ-Methode“ sowie die „α-Methode“ (auch „verallgemeinerte α-Methode“). Das *Butcher*-Schema und die Verfahrensvorschrift der θ-Methode lauten

$$\begin{array}{c|c} \theta & \theta \\ \hline & 1 \end{array} \qquad \boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h_n \boldsymbol{f}(t_n + \theta h_n, (1-\theta)\,\boldsymbol{y}_n + \theta\,\boldsymbol{y}_{n+1})\,.$$

Für $\theta = 0$ erhält man das explizite *Euler*-Verfahren (Ordnung 1), für $\theta = 1$ das implizite *Euler*-Verfahren (Ordnung 1) und für $\theta = 1/2$ die implizite Mittelpunktregel (Ordnung 2), vgl. Tabelle 3.1. Aus der Stabilitätsfunktion $R(z) = (1 + (1-\theta)z)/(1-\theta z)$ liest man ab, daß das Verfahren A-stabil ist für $\theta \geq 1/2$ und L-stabil nur für $\theta = 1$.

Die α-Methode ist gegeben durch

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1-\alpha & \alpha \\ \hline & 1-\alpha & \alpha \end{array} \qquad \boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h_n \left((1-\alpha)\,\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \alpha\,\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1})\right)\,.$$

Für $\alpha = 0$ erhält man wieder das explizite *Euler*-Verfahren, für $\alpha = 1$ das implizite *Euler*-Verfahren und für $\alpha = 1/2$ die Trapezregel (Ordnung 2). Die Stabilitätsfunktion $R(z) = (1 + (1-\alpha)z)/(1-\alpha z)$ stimmt mit derjenigen der θ-Methode überein, so daß beide Verfahren dieselben Stabilitätseigenschaften besitzen. □

Die in Abschnitt 3.2.1 angegebenen Stabilitätseigenschaften kann man bei impliziten *Runge-Kutta*-Verfahren direkt an den Koeffizienten des *Butcher*-Schemas ablesen. Mit dem Vektor $\mathbb{1} := (1, \ldots, 1)^T$ ist die Stabilitätsfunktion eines IRK (*Hairer & Wanner* [64, IV.3])

$$R(z) \;=\; 1 + z\boldsymbol{b}^T(\boldsymbol{I} - z\boldsymbol{A})^{-1}\,\mathbb{1} \;=\; \frac{\det(\boldsymbol{I} - z\boldsymbol{A} + z\mathbb{1}\boldsymbol{b}^T)}{\det(\boldsymbol{I} - z\boldsymbol{A})}\,, \tag{3.25}$$

und es gilt $R(\infty) = 1 - \boldsymbol{b}^T\boldsymbol{A}^{-1}\,\mathbb{1}$. Ein IRK mit nichtsingulärer Koeffizientenmatrix $\boldsymbol{A}$ ist L-stabil, wenn es A-stabil ist und eine der beiden Bedingungen

$$\begin{array}{lllll} (i) & a_{sj} & = & b_j, & j = 1, \ldots, s\,, \\ (ii) & a_{i1} & = & b_1, & i = 1, \ldots, s \end{array} \tag{3.26}$$

erfüllt. Bei DIRK-Verfahren mit $a_{ii} > 0$ ist zudem die S-Stabilität äquivalent zur Bedingung $|1 - \boldsymbol{b}^T\boldsymbol{A}^{-1}\,\mathbb{1}| < 1$ (*Alexander* [5, Theorem 4]).

3.3.2 Formulierung für implizite DAE

Bei der Zeitintegration von DAE müssen algebraische Nebenbedingungen erfüllt werden, so daß sich i. a. nur implizite Verfahren anwenden lassen. Ist allerdings der differentielle

Anteil der DAE nicht-steif, so besteht die Möglichkeit, die differentiellen Variablen mit einem expliziten Verfahren zu integrieren und die algebraischen Nebenbedingungen durch Lösen eines Gleichungssystems der Dimension der algebraischen Variablen zu erfüllen (*halb-explizite Runge-Kutta-Verfahren*, vgl. *Hairer, Lubich & Roche* [62, S. 20]). Die im Rahmen dieser Arbeit behandelten DAE stammen aus FEM-Ortsdiskretisierungen partieller Differentialgleichungen und haben daher stets steife differentielle Anteile, so daß im folgenden nur implizite Verfahren betrachtet werden.

Die Verallgemeinerung von *Runge-Kutta*-Verfahren für DAE der Form (3.2) mit konsistenten Anfangswerten gemäß Definition 3.2 ist nun naheliegend. Statt (3.20) wird mit den Abkürzungen

$$T_{ni} = t_n + c_i h_n \qquad \text{und} \qquad \boldsymbol{Y}_{ni} = \boldsymbol{y}_n + h_n \sum_{j=1}^{s} a_{ij} \dot{\boldsymbol{Y}}_{nj} \tag{3.27}$$

für die Stufenzeiten und Stufenlösungen das nichtlineare Gleichungssystem

$$\boldsymbol{F}(T_{ni},\ \boldsymbol{Y}_{ni},\ \dot{\boldsymbol{Y}}_{ni}) = \boldsymbol{0}\,, \qquad i = 1, \ldots, s \tag{3.28}$$

der Dimension $s \cdot m$ nach den Stufenableitungen $\dot{\boldsymbol{Y}}_{ni}$ gelöst. Die Berechnung der neuen Näherungslösung gemäß (3.21) sichert dann allerdings nicht unbedingt die Einhaltung der algebraischen Nebenbedingungen. Bei den *projizierten Runge-Kutta-Verfahren* (*Hairer & Wanner* [64]) wird dies durch Lösen eines zusätzlichen nichtlinearen Systems erzwungen. Man ist jedoch daran interessiert, den Zusatzaufwand zu vermeiden, und berücksichtigt diese Forderung daher schon bei der Konstruktion. Von *Prothero & Robinson* [96] wurde der folgende Begriff in bezug auf steife Differentialgleichungen eingeführt.

Def. 3.8: Ein *Runge-Kutta*-Verfahren heißt *steif genau*, wenn

$$a_{si} = b_i\,, \qquad i = 1, \ldots, s \tag{3.29}$$

für die Koeffizienten im *Butcher*-Schema gilt. ■

Im Zusammenhang mit DAE bedeutet dies, daß die neue Näherungslösung zum Zeitpunkt t_{n+1} gleich der Stufenlösung auf der letzten Stufe ist, $\boldsymbol{y}_{n+1} = \boldsymbol{Y}_{ns}$, so daß stets die algebraischen Nebenbedingungen eingehalten werden. Ein steif genaues A-stabiles IRK mit nicht-singulärer Koeffizientenmatrix $\boldsymbol{A}$ ist wegen (3.26) auch L-stabil. Außerdem gewährleisten steif genaue Verfahren bei Index-1-Systemen, daß die bei gewöhnlichen Differentialgleichungen erzielte Verfahrensordnung sowohl für die differentiellen als auch für die algebraischen Variablen erhalten bleibt[3].

3.3.3 Kriterien zur Auswahl der Verfahrensklasse

Die Auswahl von geeigneten Zeitintegrationsverfahren für die in der vorliegenden Arbeit auftretenden DAE-Systeme kann damit zusammenfassend wie folgt motiviert werden:

[3]Das Phänomen der *Ordnungsreduktion* bei steifen Differentialgleichungen wurde erstmals von *Prothero & Robinson* [96] beschrieben und führte zur Entwicklung der S-Stabilität sowie steif genauer Verfahren. Die Verfahrensordnung von IRK bei Anwendung auf DAE mit höherem Index wurde ausführlich von *Hairer, Lubich & Roche* [62] untersucht.

- Ortsadaptivität, Plastizität $\rightsquigarrow$ Einschrittverfahren
- Flexibilität, effiziente Schrittweitensteuerung (s. u.) $\rightsquigarrow$ *Runge-Kutta*-Verfahren
- Behandlung von DAE $\rightsquigarrow$ steif genaue Verfahren
- Große DAE-Systeme mit Index 1 $\rightsquigarrow$ DIRK
- Steifheit aus FEM und DAE $\rightsquigarrow$ A-, L- und evtl. S-stabile Verfahren
- Vereinfachtes *Newton*-Verfahren $\rightsquigarrow$ SDIRK

Es werden daher im folgenden nur noch diagonal-implizite Verfahren (DIRK und SDIRK) behandelt.

3.3.4 Diagonal-implizite *Runge-Kutta*-Verfahren (DIRK)

Zur Reduktion der Rundungsfehler bei der Lösung der nichtlinearen Systeme werden als Unbekannte statt der Stufenlösungen die *Stufeninkremente* $\Delta\boldsymbol{Y}_{ni} := \boldsymbol{Y}_{ni} - \boldsymbol{y}_n$ verwendet (*Hairer & Wanner* [64, IV.8]). Bei den im folgenden betrachteten diagonal-impliziten Verfahren läuft die Summe in (3.27) jeweils nur bis zur aktuellen Stufe i, so daß man die Stufenableitung $\dot{\boldsymbol{Y}}_{ni}$ als Funktion des Stufeninkrements $\Delta\boldsymbol{Y}_{ni}$ darstellen kann:

$$\boldsymbol{Y}_{ni} \;=\; \boldsymbol{y}_n + h_n \sum_{j=1}^{i} a_{ij}\dot{\boldsymbol{Y}}_{nj} \quad\Longrightarrow\quad \dot{\boldsymbol{Y}}_{ni} \;=\; \frac{1}{h_n\, a_{ii}}\Bigg[\underbrace{\boldsymbol{Y}_{ni} - \boldsymbol{y}_n}_{\Delta\boldsymbol{Y}_{ni}} - \underbrace{h_n \sum_{j=1}^{i-1} a_{ij}\dot{\boldsymbol{Y}}_{nj}}_{\overline{\boldsymbol{Y}}_{ni}}\Bigg]. \tag{3.30}$$

Die *akkumulierte Stufenableitung* $\overline{\boldsymbol{Y}}_{ni}$ ist dabei nur von schon berechneten Größen auf vorhergehenden Stufen abhängig und stellt für die aktuelle Stufe i eine konstante Größe dar. Damit lautet das nichtlineare Gleichungssystem zur Bestimmung des Inkrements $\Delta\boldsymbol{Y}_{ni}$ auf der Stufe i mit der in (3.27) eingeführten Stufenzeit $T_{ni} = t_n + c_i h_n$:

$$\boldsymbol{R}_{ni}(\Delta\boldsymbol{Y}_{ni}) \equiv \boldsymbol{F}\Big(T_{ni},\; \underbrace{\boldsymbol{y}_n + \Delta\boldsymbol{Y}_{ni}}_{\boldsymbol{Y}_{ni}},\; \underbrace{\frac{1}{h_n\, a_{ii}}[\Delta\boldsymbol{Y}_{ni} - \overline{\boldsymbol{Y}}_{ni}]}_{\dot{\boldsymbol{Y}}_{ni}}\Big) = \boldsymbol{0}\,. \tag{3.31}$$

Insgesamt führt dies auf den in Kasten (3.33) angegebenen Algorithmus zur Berechnung eines Zeitschritts. Bei Verwendung des *Newton*-Verfahrens zur Lösung von Schritt 1b in Kasten (3.33) benötigt man die Ableitung der nichtlinearen Vektorfunktion $\boldsymbol{R}_{ni}$ nach den Stufeninkrementen $\Delta\boldsymbol{Y}_{ni}$ (*Jacobi*-Matrix). Diese hat die Form

$$\boldsymbol{J}_{ni} \;=\; \frac{\mathrm{d}\boldsymbol{R}_{ni}}{\mathrm{d}\Delta\boldsymbol{Y}_{ni}} \;=\; \left.\frac{\partial\boldsymbol{F}}{\partial\boldsymbol{y}}\right|_{\boldsymbol{z}} \;+\; \frac{1}{h_n\, a_{ii}}\left.\frac{\partial\boldsymbol{F}}{\partial\dot{\boldsymbol{y}}}\right|_{\boldsymbol{z}}, \tag{3.32}$$

wobei $\boldsymbol{z} = (T_{ni}, \boldsymbol{Y}_{ni}, \dot{\boldsymbol{Y}}_{ni})$ die aktuellen Argumente von $\boldsymbol{F}$ in (3.31) bezeichnet.

Bemerkung: Bei einem rein elastischen Problem mit FEM-Ortsdiskretisierung entspricht die partielle Ableitung nach $\dot{\boldsymbol{y}}$ der Massen- oder Systemmatrix und die partielle Ableitung nach $\boldsymbol{y}$ der Steifigkeitsmatrix. □

Zeitschritt-Algorithmus eines steif genauen DIRK für DAE

Gegeben: Koeffizienten c_i, a_{ij}, b_j eines steif genauen DIRK,
Näherungslösung $\boldsymbol{y}_n$ zum Zeitpunkt t_n, Schrittweite h_n.

Gesucht: Näherungslösung $\boldsymbol{y}_{n+1}$ zum Zeitpunkt t_{n+1}.

Schritt 1: Für $i = 1, \ldots, s$

(a) setze $T_{ni} := t_n + c_i h_n$ und $\overline{\boldsymbol{Y}}_{ni} := h_n \sum_{j=1}^{i-1} a_{ij} \dot{\boldsymbol{Y}}_{nj}$, (3.33)

(b) löse $\boldsymbol{R}_{ni}(\Delta \boldsymbol{Y}_{ni}) = \boldsymbol{0}$ nach $\Delta \boldsymbol{Y}_{ni}$ mit $\boldsymbol{R}_{ni}$ gemäß (3.31),

(c) setze $\dot{\boldsymbol{Y}}_{ni} := \dfrac{1}{h_n\, a_{ii}} [\Delta \boldsymbol{Y}_{ni} - \overline{\boldsymbol{Y}}_{ni}]$.

Schritt 2: Setze $\boldsymbol{y}_{n+1} := \boldsymbol{Y}_{ns}$ und $t_{n+1} := T_{ns}$.

3.3.5 Lösung der nichtlinearen Gleichungssysteme

Der Rechenaufwand des in Kasten (3.33) angegebenen Algorithmus besteht im wesentlichen in der Lösung der nichtlinearen Gleichungssysteme in Schritt 1b. In diesem Abschnitt wird daher ein effizientes Lösungsverfahren angegeben, das die Struktur der DAE-Systeme ausnutzt, die bei der FE-Ortsdiskretisierung von Plastizitätsproblemen entstehen. Grundlage der folgenden Überlegungen ist das DAE-Anfangswertproblem

$$\boldsymbol{F}(t, \boldsymbol{y}, \dot{\boldsymbol{y}}) \;\equiv\; \begin{bmatrix} \boldsymbol{g}(t, \boldsymbol{u}, \dot{\boldsymbol{u}}; \boldsymbol{q}) \\ \boldsymbol{l}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}; \boldsymbol{u}) \end{bmatrix} \;=\; \boldsymbol{0}\,, \quad \boldsymbol{y}(0) \;=\; \boldsymbol{y}_0\,, \quad t \geq 0\,, \tag{3.34}$$

bei dem $\boldsymbol{g}$ die globalen Gleichungen aus der FE-Ortsdiskretisierung der schwachen Formulierung der Bilanzgleichungen und $\boldsymbol{l}$ die lokalen plastischen Entwicklungsgleichungen an den Integrationspunkten repräsentiert. Im Fall des ortsdiskreten quasi-statischen Modells erhält man direkt ein System dieser Struktur (vgl. Gleichung (2.57)). Zur Behandlung des dynamischen Modells (2.61) mit DIRK-Verfahren erfolgt zunächst eine Transformation auf ein System erster Ordnung in der Zeit, indem gemäß (2.63) die Geschwindigkeiten an den Knoten als neue Variablen eingeführt werden.

Das nichtlineare Gleichungssystem (3.31) lautet im Fall des DAE-Systems (3.34):

$$\begin{bmatrix} \boldsymbol{G}(\boldsymbol{U}; \boldsymbol{Q}) \\ \boldsymbol{L}(\boldsymbol{Q}; \boldsymbol{U}) \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{g}\Big(T_{ni},\ \boldsymbol{u}_n + \boldsymbol{U},\ \alpha_{ni}[\boldsymbol{U} - \overline{\boldsymbol{U}}_{ni}];\ \boldsymbol{q}_n + \boldsymbol{Q}\Big) \\ \boldsymbol{l}\Big(T_{ni},\ \boldsymbol{q}_n + \boldsymbol{Q},\ \alpha_{ni}[\boldsymbol{Q} - \overline{\boldsymbol{Q}}_{ni}];\ \boldsymbol{u}_n + \boldsymbol{U}\Big) \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \tag{3.35}$$

wobei die Abkürzungen $\boldsymbol{U} := \Delta \boldsymbol{U}_{ni}$, $\boldsymbol{Q} := \Delta \boldsymbol{Q}_{ni}$ und $\alpha_{ni} = 1/(h_n a_{ii})$ eingeführt und die Indizes n und i bei $\boldsymbol{G}$ und $\boldsymbol{L}$ zur Vereinfachung der Notation fortgelassen wurden. Die

Jacobi-Matrix von (3.35) ergibt sich zu[4]:

$$\boldsymbol{J}_{ni} = \left[\begin{array}{c|c} \dfrac{\partial \boldsymbol{G}}{\partial \boldsymbol{U}} & \dfrac{\partial \boldsymbol{G}}{\partial \boldsymbol{Q}} \\ \hline \dfrac{\partial \boldsymbol{L}}{\partial \boldsymbol{U}} & \dfrac{\partial \boldsymbol{L}}{\partial \boldsymbol{Q}} \end{array}\right] = \left[\begin{array}{c|c} \dfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}} + \alpha_{ni}\dfrac{\partial \boldsymbol{g}}{\partial \dot{\boldsymbol{u}}} & \dfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}} \\ \hline \dfrac{\partial \boldsymbol{l}}{\partial \boldsymbol{u}} & \dfrac{\partial \boldsymbol{l}}{\partial \boldsymbol{q}} + \alpha_{ni}\dfrac{\partial \boldsymbol{l}}{\partial \dot{\boldsymbol{q}}} \end{array}\right]. \tag{3.36}$$

Die linke obere Blockmatrix ist eine schwach besetzte FE-Matrix, bestehend aus der verallgemeinerten Steifigkeitsmatrix $\boldsymbol{K} = \partial \boldsymbol{g}/\partial \boldsymbol{u}$ und der verallgemeinerten Massenmatrix $\boldsymbol{M} = \partial \boldsymbol{g}/\partial \dot{\boldsymbol{u}}$. Sie ist das Ergebnis der Linearisierung der zeitdiskreten FE-Knotengleichungen (vgl. (2.53), (2.66)). Die rechte untere Blockmatrix besteht aus kleinen Diagonalblöcken, wobei jeder Block die plastischen Entwicklungsgleichungen eines Integrationspunkts repräsentiert (vgl. (2.51), (2.55)).

Diese Überlegungen zeigen, daß eine direkte Lösung des linearen Systems mit der Matrix (3.36) aus numerischer Sicht nicht vertretbar ist, da die schwach besetzte Struktur komplett zerstört würde. Es wird also ein geeignetes Block-Lösungsverfahren benötigt, bei dem einerseits die schwach besetzte Strutur der FE-Matrix erhalten bleibt, und das andererseits die entkoppelte Lösung der Gleichungen an den einzelnen Integrationspunkten erlaubt.

Beim *Newton*-Block-*Gauß-Seidel*-Verfahren (NBGS) wird eine Block-*Gauß-Seidel*-Iteration für das lineare System mit der Matrix (3.36) und der rechten Seite (3.35) in eine übergeordnete *Newton*-Iteration zur Lösung von (3.35) eingebettet, siehe *Wittekind* [124], *Fritzen* [57]. Umgekehrt wird beim ebenfalls dort diskutierten Block-*Gauß-Seidel-Newton*-Verfahren (BGSN) zunächst eine nichtlineare Block-*Gauß-Seidel*-Iteration auf das System (3.35) angewandt, wobei die nichtlinearen Block-Gleichungssysteme $\boldsymbol{G}(\boldsymbol{U};\boldsymbol{Q})$ bei festem $\boldsymbol{Q}$ und $\boldsymbol{L}(\boldsymbol{Q};\boldsymbol{U})$ bei festem $\boldsymbol{U}$ mit dem *Newton*-Verfahren gelöst werden. In der Linearisierung werden also nur die Diagonalblöcke von (3.36) benutzt. Beide Verfahren konvergieren, wenn die Schrittweite h_n „genügend klein“ gewählt wird (*Fritzen* [57]). Diese Schrittweitenbeschränkung kann jedoch in der Praxis in Bereichen mit starken Nichtlinearitäten, z. B. bei plastischen Lokalisierungsproblemen, den Vorteil der A-stabilen diagonalimpliziten Verfahren vollständig zunichte machen, bei denen die Möglichkeit der Verwendung „großer“ Zeitschritte ein wesentlicher Effizienzaspekt ist.

Eine andere Möglichkeit besteht in der Verallgemeinerung des BGSN-Verfahrens, bei der man bei der Linearisierung der globalen $\boldsymbol{G}$-Gleichungen die Abhängigkeit von den lokalen $\boldsymbol{L}$-Gleichungen berücksichtigt (*Ehlers & Ellsiepen* [49], *Diebels, Ellsiepen & Ehlers* [42]). Im Fall des impliziten *Euler*-Verfahrens bei Problemen der Elastoplastizität ist dieser Ansatz als *algorithmisch konsistente Linearisierung* bekannt (*Simo & Taylor* [107]). Die Methode kann auch als zwei-stufiges *Newton*-Verfahren interpretiert werden (*Hartmann* [65]). Ein Schritt der Iteration besteht aus den folgenden Teilschritten, wobei aus Gründen der Übersichtlichkeit kein Iterationsindex notiert wird:

[4] Im Gegensatz zur sonst verwendeten Notation $\boldsymbol{L}(\boldsymbol{Q};\boldsymbol{U})$, die ausdrückt, daß die Primärvariablen $\boldsymbol{U}$ für die lokalen Gleichungen im Rahmen des zweistufigen *Newton*-Verfahrens als Parameter zu verstehen sind, wurden zur Darstellung der *Jacobi*-Matrix die Variablen der Funktion $\boldsymbol{L}$ vertauscht, so daß die erste Spalte die Ableitungen nach $\boldsymbol{U}$ und die zweite Spalte die Ableitungen nach $\boldsymbol{Q}$ enthält.

1. Löse die lokalen Integrationspunkt-Gleichungen $\boldsymbol{L}(\boldsymbol{Q};\boldsymbol{U})$ nach $\boldsymbol{Q}$ bei festgehaltenen globalen Variablen $\boldsymbol{U}$,
2. berechne die *Jacobi*-Matrix des globalen Systems $\boldsymbol{G}(\boldsymbol{U};\boldsymbol{Q}(\boldsymbol{U}))$,
3. löse das globale schwach besetzte FE-Gleichungssystem nach $\Delta\boldsymbol{U}$ und
4. aktualisiere die globalen Variablen $\boldsymbol{U}$.

Im ersten Teilschritt der Iteration werden die lokalen Gleichungen an den Integrationspunkten bei gegebenem, festem $\boldsymbol{U}$ aus dem letzten Iterationsschritt mit dem *Newton*-Verfahren nach $\boldsymbol{Q}$ gelöst:

$$\boldsymbol{L}(\boldsymbol{Q};\boldsymbol{U}) = \boldsymbol{0} \quad \Longrightarrow \quad \boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{U})\,. \tag{3.37}$$

Aufgrund der Block-Diagonalstruktur von $\boldsymbol{l}$ und damit von $\boldsymbol{L}$ kann dies entkoppelt je Integrationspunkt geschehen. Nach dem Satz über implizite Funktionen ist $\boldsymbol{Q}$ lokal eine Funktion von $\boldsymbol{U}$, sofern (3.37) exakt gelöst wird[5]. Dann ist das Differential von $\boldsymbol{L}(\boldsymbol{Q}(\boldsymbol{U});\boldsymbol{U})$ nach $\boldsymbol{U}$ ebenfalls Null, und man erhält als Ergebnis das Differential der impliziten Funktion $\boldsymbol{Q}(\boldsymbol{U})$:

$$\frac{\mathrm{d}\boldsymbol{L}}{\mathrm{d}\boldsymbol{U}} = \frac{\partial\boldsymbol{L}}{\partial\boldsymbol{Q}}\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\boldsymbol{U}} + \frac{\partial\boldsymbol{L}}{\partial\boldsymbol{U}} = \boldsymbol{0} \quad \Longrightarrow \quad \frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\boldsymbol{U}} = -\left[\frac{\partial\boldsymbol{L}}{\partial\boldsymbol{Q}}\right]^{-1}\frac{\partial\boldsymbol{L}}{\partial\boldsymbol{U}}\,. \tag{3.38}$$

Dies kann ebenfalls lokal je Integrationspunkt durch Lösung kleiner linearer System von der Größe der Anzahl interner Variablen berechnet werden (im Fall des viskoplastischen Zweiphasenmodells sind dies 5×5-Systeme, vgl. (2.51)).

Der zweite Teilschritt besteht in der Berechnung des totalen Differentials der globalen Gleichung $\boldsymbol{G}(\boldsymbol{U};\boldsymbol{Q}(\boldsymbol{U}))$ nach $\boldsymbol{U}$, d. h. der algorithmisch konsistenten Linearisierung:

$$\boldsymbol{J}_{\boldsymbol{G}} := \frac{\mathrm{d}\boldsymbol{G}}{\mathrm{d}\boldsymbol{U}} = \frac{\partial\boldsymbol{G}}{\partial\boldsymbol{U}} + \frac{\partial\boldsymbol{G}}{\partial\boldsymbol{Q}}\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\boldsymbol{U}} = \frac{\partial\boldsymbol{G}}{\partial\boldsymbol{U}} - \frac{\partial\boldsymbol{G}}{\partial\boldsymbol{Q}}\left[\frac{\partial\boldsymbol{L}}{\partial\boldsymbol{Q}}\right]^{-1}\frac{\partial\boldsymbol{L}}{\partial\boldsymbol{U}}\,. \tag{3.39}$$

Dabei repräsentiert der erste Term im wesentlichen die Steifigkeitsmatrix aus der Linearisierung des elastischen Materialgesetzes. Bei den hier betrachteten Mehrphasenmodellen enthält dieser Term zusätzlich die Linearisierung der anderen diskreten Bilanzgleichungen, etwa der Volumenbilanz beim quasi-statischen Zweiphasenmodell. Der zweite Term ist nur im plastischen Bereich „aktiv“ und berücksichtigt die Nichtlinearität, die aus dem plastischen Modell nach Zeit- und Ortsdiskretisierung resultiert.

Im dritten Teilschritt der Iteration wird das schwach besetzte lineare System

$$\boldsymbol{J}_{\boldsymbol{G}}\,\Delta\boldsymbol{U} = \boldsymbol{G}(\boldsymbol{U};\boldsymbol{Q}(\boldsymbol{U})) \tag{3.40}$$

[5] Im Rahmen eines numerischen Verfahrens ist dies natürlich nicht möglich. Es wird hier jedoch davon ausgegangen, daß die lokalen Gleichungen mit ausreichender Genauigkeit gelöst werden, so daß diese Annahme gerechtfertigt ist. Für eine Fehlerbetrachtung bei zwei-stufigen *Newton*-Prozessen sei z. B. auf *Rabbat et al.* [97] verwiesen.

nach dem globalen *Newton*-Inkrement $\Delta \boldsymbol{U}$ gelöst. Schließlich wird der globale Lösungsvektor aktualisiert:

$$\boldsymbol{U} \leftarrow \boldsymbol{U} - \Delta \boldsymbol{U} \,. \tag{3.41}$$

Diese Lösungsstrategie erlaubt auch in Bereichen mit starken Nichtlinearitäten die Verwendung relativ großer Schrittweiten und hat sich in der Praxis gut bewährt. Anschaulich ist dies darauf zurückzuführen, daß durch die Berücksichtigung der lokalen Systeme bei der Linearisierung des globalen Systems die im Modell enthaltenen Nichtlinearitäten aufgrund plastischer Deformationen besser erfaßt werden als dies beim NBGS- oder BGSN-Verfahren der Fall ist.

Für ein besseres Verständnis des Verfahrens und Hinweise zur geeigneten Wahl von enthaltenen relativen und absoluten Toleranzen wäre eine detaillierte mathematische Untersuchung solcher mehrstufiger *Newton*-Verfahren unter Berücksichtigung der speziellen Struktur der hier betrachteten Plastizitätsprobleme sehr hilfreich. Da das Verfahren jedoch im Rahmen dieser Arbeit nur einen von vielen Teilaspekten zur Lösung des Gesamtproblems der numerischen Simulation von Mehrphasenproblemen darstellt, muß an dieser Stelle auf die Beantwortung derartiger Fragestellungen verzichtet werden.

3.4 Fehlerschätzung und Schrittweitensteuerung

Zur effizienten Steuerung der Schrittweite wird eine Schätzung des lokalen Fehlers benötigt. Zunächst werden dazu einige Begriffe eingeführt, vgl. etwa *Hairer, Nørsett & Wanner* [63], *Strehmel & Weiner* [110], *Törnig & Spellucci* [115]. Für alle Funktionen wird genügende Glattheit (Differenzierbarkeit) vorausgesetzt.

Def. 3.9: Gegeben sei ein Anfangswertproblem (3.18) gewöhnlicher Differentialgleichungen. Das Differentialgleichungssystem heißt *dissipativ* bzgl. einer Norm $\|\cdot\|$, wenn für je zwei Lösungen $\boldsymbol{y}(t)$ und $\boldsymbol{z}(t)$ zu verschiedenen Anfangswerten $\boldsymbol{y}_0$ und $\boldsymbol{z}_0$ gilt:

$$\|\boldsymbol{y}(t_2) - \boldsymbol{z}(t_2)\| \leq \|\boldsymbol{y}(t_1) - \boldsymbol{z}(t_1)\| \qquad \text{für } t_0 \leq t_1 \leq t_2 < \infty \,. \tag{3.42}$$

Diese Eigenschaft wird auch als *Kontraktivität* bezeichnet und bedeutet, daß sich Lösungen zu verschiedenen Anfangswerten mit der Zeit annähern. ■

Bemerkung: Bei den hier betrachteten Problemen aus der Plastizitätstheorie poröser Medien kann davon ausgegangen werden, daß die zugehörigen semidiskreten Anfangswertprobleme dissipativ sind. Dies kann damit begründet werden, daß zum einen das Vorhandensein eines Porenfluids zu einem diffusiven Charakter der beschreibenden Gleichungen führt, und daß zum anderen mit den plastischen Deformationen eine interne Dissipation verbunden ist (plastische Leistung). □

Def. 3.10: Jedes *Einschrittverfahren* (ESV) zur Lösung von (3.18) kann in der Form

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h_n \boldsymbol{\Phi}(t_n, \boldsymbol{y}_n; h_n) \,, \quad h_n = t_{n+1} - t_n \,, \quad n = 0, \ldots, N \,, \quad t_0 = 0, \; t_N = T \tag{3.43}$$

geschrieben werden. Darin bezeichnet $\boldsymbol{\Phi}$ die *Verfahrensfunktion* (*Inkrementfunktion*). Die Schreibweise ist nur formal explizit, im Falle impliziter Verfahren beinhaltet die Verfahrensfunktion die Lösung der auftretenden (nichtlinearen) Gleichungssysteme.

Ausgehend von einem Punkt $(t, \boldsymbol{y}(t))$ auf der exakten Lösung werde mit (3.43) zur Zeit $t + h$ eine Näherungslösung $\boldsymbol{y}_h$ berechnet:

$$\boldsymbol{y}_h = \boldsymbol{y}(t) + h\boldsymbol{\Phi}(t, \boldsymbol{y}(t); h)\,. \tag{3.44}$$

Die Größe

$$\boldsymbol{\delta}(t + h) = \boldsymbol{y}(t + h) - \boldsymbol{y}_h \tag{3.45}$$

wird als *lokaler Diskretisierungsfehler* bezeichnet. Da $\boldsymbol{\delta}$ sowohl von der exakten Lösung $\boldsymbol{y}$ als auch von der Schrittweite h abhängt, schreibt man auch $\boldsymbol{\delta}(t + h, \boldsymbol{y}; h)$.

Das ESV (3.43) heißt *konsistent* mit dem Anfangswertproblem (3.18), wenn gilt:

$$\lim_{h\to 0} \max_{t\in[0,T]} \|\boldsymbol{f}(t, \boldsymbol{y}(t)) - \boldsymbol{\Phi}(t, \boldsymbol{y}(t); h)\| = 0\,. \tag{3.46}$$

Dies ist gleichwertig mit $\boldsymbol{\Phi}(t, \boldsymbol{y}(t); 0) = \boldsymbol{f}(t, \boldsymbol{y}(t))$ für alle $t \in [0, T]$.

Das ESV besitzt die *Konsistenzordnung* p, wenn gilt:

$$\max_{t\in[0,T]} \|\boldsymbol{\delta}(t + h)\| \le C\, h^{p+1} \quad \text{für alle } h \in (0, h_{\max}]\,, \tag{3.47}$$

mit einer von h unabhängigen Konstante C. ■

Die Konsistenzordnung p eines Verfahrens ist also ein Maß dafür, wie gut die Verfahrensfunktion $\boldsymbol{\Phi}$ lokal die Differentialgleichung $\dot{\boldsymbol{y}}(t) = \boldsymbol{f}(t, \boldsymbol{y}(t))$ approximiert. Gemäß (3.45) gilt nämlich

$$\boldsymbol{\Phi}(t, \boldsymbol{y}(t); h) = \frac{\boldsymbol{y}(t + h) - \boldsymbol{y}(t)}{h} - \frac{1}{h}\,\boldsymbol{\delta}(t + h)\,. \tag{3.48}$$

Für $h \to 0$ strebt der Differenzenquotient in (3.48) gegen die Ableitung $\dot{\boldsymbol{y}}(t)$ (rechte Seite $\boldsymbol{f}(t, \boldsymbol{y}(t))$), und für den *lokalen Abschneidefehler*

$$\boldsymbol{\tau}(t) = \frac{1}{h}\,\boldsymbol{\delta}(t + h) \quad \text{bzw.} \quad \boldsymbol{\tau}(t, \boldsymbol{y}; h) = \frac{1}{h}\,\boldsymbol{\delta}(t + h, \boldsymbol{y}; h) \tag{3.49}$$

gilt wegen (3.47): $\|\boldsymbol{\tau}(t, \boldsymbol{y}; h)\| = \mathcal{O}(h^p)$.

Natürlich muß jedes sinnvolle numerische Verfahren *konvergent* sein. Dies besagt, daß der *globale Diskretisierungsfehler* $\boldsymbol{e}_N(t) := \boldsymbol{y}(t) - \boldsymbol{y}_N(t)$ zu einem (festen) Zeitpunkt t nach N gerechneten Schritten für $N \to \infty$ (und damit gleichzeitig $h_{\max} \to 0$) verschwinden muß. Man kann zeigen, daß bereits jedes konsistente und stabile ESV auch konvergent ist und daß die Konvergenzordnung dann mit der Konsistenzordnung übereinstimmt.

Betrachtet man nur dissipative Probleme, bei denen zurückliegende Fehler mit der Zeit gedämpft werden, so genügt es zudem, den lokalen Fehler zu kontrollieren, um den globalen Fehler beschränkt zu halten. Für viele ESV, insbesondere für *Runge-Kutta*-Verfahren, kann man aufgrund der Konsistenz für den lokalen Diskretisierungsfehler $\boldsymbol{\delta}$ eine Darstellung der Form

$$\boldsymbol{\delta}(t + h, \boldsymbol{y}; h) = \boldsymbol{\psi}(t, \boldsymbol{y}(t))\, h^{p+1} + \mathcal{O}(h^{p+2}) \tag{3.50}$$

herleiten, wobei $\boldsymbol{\psi}(t, \boldsymbol{y}(t))$ eine Linearkombination der elementaren Differentiale $(p+1)$-ter Ordnung im Punkt $(t, \boldsymbol{y}(t))$ ist, also insbesondere nicht von der Schrittweite h abhängt. Der Fehler teilt sich also auf in den *Hauptteil* $\boldsymbol{\psi}(t, \boldsymbol{y}(t))\, h^{p+1}$ und in Terme höherer Ordnung $\mathcal{O}(h^{p+2})$. Das Ziel ist es nun, den Hauptteil des lokalen Fehlers verläßlich zu schätzen und aufgrund dieser Schätzung die Schrittweite entsprechend zu steuern. In den folgenden beiden Abschnitten werden dazu zwei gängige Methoden vorgestellt.

1. Bei der *Richardson*-Extrapolation werden mit ein und demselben Verfahren zwei Lösungen mit verschiedenen Schrittweiten berechnet.

2. Bei eingebetteten *Runge-Kutta*-Verfahren werden die Koeffizienten so gewählt, daß man nahezu ohne Zusatzaufwand zwei Lösungen verschiedener Ordnung erhält.

3.4.1 *Richardson*-Extrapolation

Ausgehend von der Darstellung (3.50) für den lokalen Fehler geht man wie folgt vor:

1. Man berechnet zwei (kleine) Schritte mit der Schrittweite $h/2$ und bezeichnet die Lösungen mit $\boldsymbol{y}_{h/2}$ und $\boldsymbol{y}_{2\mathrm{x}h/2}$.

2. Man berechnet einen (großen) Schritt mit der Schrittweite h und bezeichnet die Lösung mit $\boldsymbol{y}_h$.

Der Fehler $\boldsymbol{\delta}_{2\mathrm{x}h/2}$ nach dem zweiten der beiden kleinen Schritte setzt sich additiv aus dem mit dem Faktor $(1 + \mathcal{O}(h))$ „transportierten“ Fehler $\boldsymbol{\delta}_{h/2}$ des ersten Schritts sowie dem Fehler des zweiten Schritts zusammen:

$$\boldsymbol{\delta}_{2\mathrm{x}h/2} = (1 + \mathcal{O}(h))\, \boldsymbol{\delta}_{h/2} + \boldsymbol{\psi}(t + h/2, \boldsymbol{y}_{h/2}) \left(\frac{h}{2}\right)^{p+1} + \mathcal{O}(h^{p+2})\,.$$

Mit Hilfe der Beziehung $\boldsymbol{\psi}(t + h/2, \boldsymbol{y}_{h/2}) = (1 + \mathcal{O}(h))\, \boldsymbol{\psi}(t, \boldsymbol{y}(t))$ erhält man nach einigen Umformungen für den Fehler nach zwei Schritten mit der Schrittweite $h/2$:

$$\boldsymbol{\delta}_{2\mathrm{x}h/2} = \boldsymbol{y}(t + h) - \boldsymbol{y}_{2\mathrm{x}h/2} = \frac{1}{2^p}\, \boldsymbol{\psi}(t, \boldsymbol{y}(t))\, h^{p+1} + \mathcal{O}(h^{p+2})\,. \tag{3.51}$$

Andererseits gilt gemäß (3.50) für den Fehler nach einem Schritt mit der Schrittweite h:

$$\boldsymbol{\delta}_h = \boldsymbol{y}(t + h) - \boldsymbol{y}_h = \boldsymbol{\psi}(t, \boldsymbol{y}(t))\, h^{p+1} + \mathcal{O}(h^{p+2})\,. \tag{3.52}$$

Aus (3.51) und (3.52) kann man die Fehlerfunktion $\boldsymbol{\psi}(t, \boldsymbol{y}(t))$ durch Multiplikation von (3.51) mit 2^p und Subtraktion von (3.52) eliminieren:

$$\boldsymbol{y}(t + h) - \boldsymbol{y}_{2\mathrm{x}h/2} = \frac{\boldsymbol{y}_{2\mathrm{x}h/2} - \boldsymbol{y}_h}{2^p - 1} + \mathcal{O}(h^{p+2})\,.$$

Man erhält also mittels

$$\hat{\boldsymbol{y}}_h = \boldsymbol{y}_{2\mathrm{x}h/2} + \frac{\boldsymbol{y}_{2\mathrm{x}h/2} - \boldsymbol{y}_h}{2^p - 1} \tag{3.53}$$

eine verbesserte Näherungslösung der Konsistenzordnung $p+1$. Dieser Vorgang wird als *Richardson-Extrapolation* bezeichnet. Andererseits kann man auch durch Subtraktion von (3.52) und (3.51) nach Weglassen der Terme $\mathcal{O}(h^{p+2})$ höherer Ordnung in erster Näherung den Hauptteil des lokalen Fehlers schätzen:

$$\boldsymbol{\psi}(t,\boldsymbol{y}(t))\,h^{p+1} \approx \frac{2^p}{2^p-1}\,(\boldsymbol{y}_{2\mathrm{x}h/2}-\boldsymbol{y}_h) = \hat{\boldsymbol{y}}_h - \boldsymbol{y}_h\,. \tag{3.54}$$

Bemerkung: Man beachte, daß dies eine lokale Fehlerschätzung für den Fehler in der numerischen Lösung $\boldsymbol{y}_h$ bei Rechnung mit der Schrittweite h ist, so daß man streng genommen mit $\boldsymbol{y}_h$ als neuer Näherungslösung weiter rechnen muß. In der Praxis verwendet man jedoch üblicherweise die „genauere“ Lösung[6] $\hat{\boldsymbol{y}}_h$ als neue Näherungslösung, obwohl (3.54) dafür keine Fehlerschätzung darstellt.

Die obige Herleitung geht wegen der Übersichtlichkeit der Darstellung jeweils von einem Punkt $(t,\boldsymbol{y}(t))$ auf der exakten Lösung aus. Während der numerischen Rechnung steht dieser natürlich nicht zur Verfügung, so daß man statt dessen ausgehend von einem Punkt $(t_n,\boldsymbol{y}_n)$ auf der numerischen Lösung das Anfangswertproblem

$$\dot{\boldsymbol{y}}_{[n]}(t) = \boldsymbol{f}(t,\boldsymbol{y}_{[n]}(t))\,,\quad \boldsymbol{y}_{[n]}(t_n)=\boldsymbol{y}_n\,,\quad t\ge t_n \tag{3.55}$$

betrachtet, bei dem die numerische Lösung des letzten Zeitschritts als Anfangswert zum Zeitpunkt t_n vorgegeben wird. In den obigen Formeln muß man also die exakte Lösung $\boldsymbol{y}$ durch die Lösung $\boldsymbol{y}_{[n]}$ von (3.55) ausgehend von $(t_n,\boldsymbol{y}_n)$ und die Näherungen $\boldsymbol{y}_h$, $\boldsymbol{y}_{h/2}$, $\boldsymbol{y}_{2\mathrm{x}h/2}$ und $\hat{\boldsymbol{y}}_h$ durch die Näherungen $\boldsymbol{y}_{n+1}$, $\boldsymbol{y}_{n+1/2}$, $\boldsymbol{y}_{n+2\mathrm{x}1/2}$ und $\hat{\boldsymbol{y}}_{n+1}$ ersetzen. □

3.4.2 Eingebettete *Runge-Kutta*-Verfahren

Zunächst wird das Prinzip der Fehlerschätzung anhand von zwei beliebigen Verfahren verschiedener Ordnung erläutert. Es seien also $\boldsymbol{\Phi}$ und $\hat{\boldsymbol{\Phi}}$ die Inkrementfunktionen (vgl. Definition 3.10) von zwei *Runge-Kutta*-Verfahren der Ordnungen p und $\hat{p}\neq p$ (normalerweise ist $\hat{p}=p-1$ oder $\hat{p}=p+1$) sowie $\boldsymbol{y}_h$ und $\hat{\boldsymbol{y}}_h$ die zugehörigen Näherungslösungen ausgehend von einem Punkt $(t,\boldsymbol{y}(t))$ auf der exakten Lösung:

$$\begin{aligned}\boldsymbol{y}_h &= \boldsymbol{y}(t)+h\,\boldsymbol{\Phi}(t,\boldsymbol{y}(t);h)\,,\\ \hat{\boldsymbol{y}}_h &= \boldsymbol{y}(t)+h\,\hat{\boldsymbol{\Phi}}(t,\boldsymbol{y}(t);h)\,.\end{aligned}$$

Gemäß (3.45) und (3.50) gilt für die zugehörigen Fehler

$$\begin{aligned}\boldsymbol{\delta}(t+h) &= \boldsymbol{y}(t+h)-\boldsymbol{y}_h &&= h^{p+1}\,\boldsymbol{\psi}(t,\boldsymbol{y}(t))+\mathcal{O}(h^{p+2})\,,\\ \hat{\boldsymbol{\delta}}(t+h) &= \boldsymbol{y}(t+h)-\hat{\boldsymbol{y}}_h &&= h^{\hat{p}+1}\,\hat{\boldsymbol{\psi}}(t,\boldsymbol{y}(t))+\mathcal{O}(h^{\hat{p}+2})\,,\end{aligned}$$

so daß man als Differenz dieser beiden Gleichungen erhält:

$$\boldsymbol{\delta}(t+h)-\hat{\boldsymbol{\delta}}(t+h) = \hat{\boldsymbol{y}}_h-\boldsymbol{y}_h = h^{p+1}\,\boldsymbol{\psi}(t,\boldsymbol{y}(t))+\mathcal{O}(h^{p+2})-\left[h^{\hat{p}+1}\,\hat{\boldsymbol{\psi}}(t,\boldsymbol{y}(t))+\mathcal{O}(h^{\hat{p}+2})\right].$$

[6] Wenn das betrachtete Anfangswertproblem nicht genügend glatt ist (etwa bei Knicken in der rechten Seite wie im Fall von Plastizität), kann $\hat{\boldsymbol{y}}_h$ durchaus eine schlechtere Näherung sein als $\boldsymbol{y}_h$.

Gilt nun o. B. d. A. $\hat{p} > p$ (ansonsten vertausche man die Rollen von p und $\hat{p}$), dann ist der Term in eckigen Klammern von der Ordnung h^{p+2}, so daß insgesamt gilt:

$$h^{p+1}\,\boldsymbol{\psi}(t,\boldsymbol{y}(t)) = \hat{\boldsymbol{y}}_h - \boldsymbol{y}_h + \mathcal{O}(h^{p+2})\,. \tag{3.56}$$

Die Differenz der beiden Lösungen verschiedener Ordnung ist also in erster Näherung (Weglassen der Terme höherer Ordnung) eine Schätzung des Hauptteils des lokalen Diskretisierungsfehlers für das Verfahren niedrigerer Ordnung.

Bemerkung: Wieder wurde der Übersichtlichkeit halber von einem Punkt $(t, \boldsymbol{y}(t))$ auf der exakten Lösung ausgegangen. Wie bereits in der Bemerkung am Ende des letzten Abschnitts erwähnt, sind eigentlich alle Formeln bezüglich des veränderten Anfangswertproblems (3.55) mit Anfangswert auf der numerischen Lösung zu betrachten. □

Bisher wurde davon ausgegangen, daß zwei Verfahren verschiedener Ordnung zur Verfügung stehen. In der Praxis ist man nun daran interessiert, die für die Fehlerschätzung benötigten Lösungen verschiedener Ordnung mit möglichst geringem Aufwand zu bestimmen. Dieser Ansatz führte zur Konstruktion von *eingebetteten Runge-Kutta-Verfahren*, bei denen neben der Näherungslösung $\boldsymbol{y}_{n+1}$ der Ordnung p mittels zusätzlicher äußerer Gewichte $\hat{b}_i$ eine Näherungslösung $\hat{\boldsymbol{y}}_{n+1}$ der Ordnung $\hat{p}$ berechnet wird. Die ersten eingebetteten (expliziten) Verfahren wurden Ende der fünfziger und Anfang der sechziger Jahre entwickelt, wobei $\hat{p} > p$ gewählt wurde, da die Fehlerschätzung gemäß (3.56) das Ziel war.

Von *Dormand & Prince* [43] stammt die Idee, umgekehrt vorzugehen, d. h. ein Verfahren niedrigerer Ordnung $\hat{p} < p$ einzubetten. Da die Formel (3.56) eine Schätzung für den Fehler des Verfahrens niedrigerer Ordnung ist, verliert man bei diesem Vorgehen zwar den Vorteil einer echten Fehlerabschätzung für das Verfahren der Ordnung p, kann aber durch geeignete Wahl der Koeffizienten eine wesentlich höhere Effizienz erreichen. *Hairer, Nørsett & Wanner* [63, II.4] haben numerisch verschiedene eingebettete explizite Verfahren verglichen, was die hervorragende Effizienz der Formeln von *Dormand* und *Prince* zeigt.

Def. 3.11: Das *Butcher*-Schema eines eingebetteten diagonal-impliziten *Runge-Kutta*-Verfahrens, abgekürzt durch „DIRK $p(\hat{p})$“, lautet:

$$\begin{array}{c|ccc} c_1 & a_{11} & & \\ \vdots & \vdots & \ddots & \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \\ & \hat{b}_1 & \cdots & \hat{b}_s \end{array} \qquad \text{oder kurz} \qquad \begin{array}{c|c} \boldsymbol{c} & \boldsymbol{A} \\ \hline & \boldsymbol{b}^T \\ & \hat{\boldsymbol{b}}^T \end{array}\,. \tag{3.57}$$

Nach Lösung der s nichtlinearen Systeme in den $\dot{\boldsymbol{Y}}_{ni}$ stehen zwei Näherungslösungen

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h_n \sum_{i=1}^{s} b_i \dot{\boldsymbol{Y}}_{ni} \qquad \text{und} \qquad \hat{\boldsymbol{y}}_{n+1} = \boldsymbol{y}_n + h_n \sum_{i=1}^{s} \hat{b}_i \dot{\boldsymbol{Y}}_{ni} \tag{3.58}$$

der Ordnungen p und $\hat{p} \neq p$ zur Verfügung. Gleichzeitig erhält man mittels

$$\boldsymbol{\delta}_{n+1} \;\approx\; \hat{\boldsymbol{y}}_{n+1} - \boldsymbol{y}_{n+1} \;=\; h_n \sum_{i=1}^{s} (\hat{b}_i - b_i)\dot{\boldsymbol{Y}}_{ni} \tag{3.59}$$

eine Fehlerschätzung für das Verfahren niedrigerer Ordnung. ■

In Tabelle 3.2 sind einige eingebettete diagonal-implizite *Runge-Kutta*-Verfahren aufgeführt. Da sowohl die implizite Mittelpunktregel als auch die Trapezregel einstufige Verfahren mit optimaler Ordnung $p = 2$ sind, gelingt es nicht, ohne Zusatzaufwand ein Verfahren zur Fehlerschätzung einzubetten. In beiden Fällen kann man aber formal ein eingebettetes Verfahren erhalten, indem man das implizite *Euler*-Verfahren (Ordnung $\hat{p} = 1$) zur Fehlerschätzung verwendet.

Von *Cash* [30] wurde ein eingebettetes SDIRK 3(2) angegeben, das auf dem 3-stufigen SDIRK-Verfahren optimaler Ordnung $p = 3$ von *Alexander* [5] beruht. Darauf aufbauend hat *Fritzen* [57] ein SDIRK 3(2) Verfahren mit 4 impliziten Stufen konstruiert, bei dem auch das eingebettete Verfahren steif genau ist.

Name	*Butcher*-Schema						s	p	$\hat{p}$	Stabilität
Implizite Mittelpunktregel	1	1					2	2	1	A
(Implizites *Euler*-Verfahren)	$\frac{1}{2}$	0	$\frac{1}{2}$							
		0	$\frac{1}{2}$							
		1	0							
Trapezregel	0	0					2	2	1	A
(Implizites *Euler*-Verfahren)	1	0	1							
	1	$\frac{1}{2}$	0	$\frac{1}{2}$						
		$\frac{1}{2}$	0	$\frac{1}{2}$						
		0	1	0						
Cash [30], *Alexander* [5]	γ	γ					3	3	2	A,L,S
	$\frac{1+\gamma}{2}$	$\frac{1-\gamma}{2}$	γ							
$\beta = \frac{1}{4}(6\gamma^2 - 20\gamma + 5)$	1	α	β	γ						
$\alpha = 1 - \beta - \gamma$		α	β	γ						
		$\frac{\gamma}{1-\gamma}$	$\frac{1-2\gamma}{1-\gamma}$	0						
$\gamma = 0{,}4358...$ ist Nullstelle von $6x^3 - 18x^2 + 9x - 1 = 0$ im Intervall $[\frac{1}{6}, \frac{1}{2}]$										
Fritzen [57]	0	0					4	3	2	A,L,S
	γ	0	γ							
α, β, γ wie oben	$\frac{1+\gamma}{2}$	0	$\frac{1-\gamma}{2}$	γ						
$\hat{\beta} = \frac{1}{2\gamma} - 1$	1	$\hat{\alpha}$	$\hat{\beta}$	0	γ					
$\hat{\alpha} = 1 - \hat{\beta} - \gamma$	1	0	α	β	0	γ				
		0	α	β	0	γ				
		$\hat{\alpha}$	$\hat{\beta}$	0	γ	0				

Tabelle 3.2: *Einige eingebettete Runge-Kutta-Verfahren (Stufenzahl s, Ordnungen p und $\hat{p}$)*

3.4.3 Fehlerschätzung bei DAE-Systemen

Die obigen Fehlerschätzungen gehen jeweils von einem System gewöhnlicher Differentialgleichungen aus. Bei der Behandlung von differential-algebraischen Systemen ist es daher fraglich, ob die Ordnungs- und Fehleraussagen ihre Gültigkeit behalten.

Es stellt sich heraus, daß für DAE-Systeme vom Index 1 die meisten Ergebnisse unverändert übernommen werden können (*Hairer & Wanner* [64, VI.1]). Für DAE-Systeme mit höherem Index ist dies nicht der Fall; dazu wurden in bezug auf *Runge-Kutta*-Verfahren von *Hairer, Lubich & Roche* [62] umfangreiche theoretische Untersuchungen angestellt. Die Hauptschwierigkeit ist, wie bereits erwähnt, die Reduktion der Ordnung in den algebraischen Variablen.

Für die in der vorliegenden Arbeit behandelten DAE-Systeme vom Index 1 kann man zusammenfassen:

- Die Konsistenzordnung p bleibt für die differentiellen Komponenten der Lösung erhalten.
- Die Verwendung von steif genauen Verfahren sichert die Konsistenzordnung p auch für die algebraischen Komponenten der Lösung.
- Bei L-stabilen Verfahren ($R(\infty) = 0$) ist die Ordnung r in den algebraischen Variablen $r = \min(p, q+1)$, wobei q die größte natürliche Zahl ist, so daß mit

$$C(q): \quad \sum_{j=1}^{s} a_{ij}\, c_j^{k-1} = \frac{c_i^k}{k}\,, \qquad i = 1, \ldots, s\,, \quad k = 1, \ldots, q$$

  die Bedingungen $C(q)$ erfüllt sind (Stufenordnung q).

Will man also auch für die eingebettete Fehlerschätzung die Ordnung $\hat{p}$ erhalten, so sollte auch das eingebettete Verfahren steif genau oder zumindest L-stabil sein und zusätzlich die Bedingung $C(\hat{p}-1)$ erfüllen. Ansonsten kann sich die Ordnung des Verfahrens entsprechend reduzieren, was die Qualität der Fehlerschätzung beeinträchtigen bzw. die Fehlerschätzung unbrauchbar machen kann.

3.4.4 Schrittweitensteuerung

Ausgehend von einer Schätzung des Hauptteils des lokalen Diskretisierungsfehlers mittels *Richardson*-Extrapolation (Gleichung (3.54) in Abschnitt 3.4.1) bzw. eingebetteter Verfahren (Gleichung (3.59) in Abschnitt 3.4.2) soll die Schrittweite nun so gesteuert werden, daß der lokale Fehler eine vorzugebende Toleranz nicht überschreitet. Dieses Vorgehen hält auch den globalen Fehler beschränkt, wenn das Problem gemäß (3.42) dissipativ ist.

Da die Fehlerschätzungen (3.54) und (3.59) vektorielle Größen sind, muß man zunächst eine geeignete Norm einführen, die ein skalares Maß für den Fehler liefert. Übliche Vorgehensweisen sind die komponentenweise Betrachtung von absoluten oder/und relativen

Schrittweitensteuerung

Gegeben: Koeffizienten eines IRK, Toleranzen ϵ_a und ϵ_r,
Näherungslösung $\boldsymbol{y}_n$ zum Zeitpunkt t_n,
Vorschlagsschrittweite h_{new},
erlaubter Schrittweitenbereich $[h_{\min},\, h_{\max}]$,
erlaubter Änderungsbereich $[fac_{\min} < 1,\, fac_{\max} > 1]$,
Sicherheitsfaktor $fac_{\text{safety}} < 1$.

Gesucht: Schrittweite h_n und Näherungslösung $\boldsymbol{y}_{n+1}$ zum Zeitpunkt $t_{n+1} = t_n + h_n$, so daß die gegebenen Toleranzen erfüllt sind.

Schritt 1: Setze die Schrittweite: $h_n := \min\left\{h_{\max},\, \max\left\{h_{\min},\, h_{\text{new}}\right\}\right\}$.

Schritt 2: Berechne mit der Schrittweite h_n zwei Lösungen $\boldsymbol{y}_{n+1}$ und $\hat{\boldsymbol{y}}_{n+1}$ der Ordnungen p und $\hat{p} < p$ (*Richardson*-Extrapolation oder eingebettetes Verfahren, vgl. Kasten (3.33)).

Schritt 3: Berechne ein gewichtetes Fehlermaß *errtol* („Fehler / Toleranz“) gemäß (3.61).

Schritt 4: Berechne den Schrittweiten-Änderungsfaktor

$$fac := \min\left\{fac_{\max},\, \max\left\{fac_{\min},\, fac_{\text{safety}} \cdot errtol^{-1/(\hat{p}+1)}\right\}\right\}$$

und setze die neue Vorschlagsschrittweite $h_{\text{new}} := fac \cdot h_n$.

Schritt 5: Wenn $errtol \leq 1$ ist, dann akzeptiere den Zeitschritt (Ende). Ansonsten verwerfe den Zeitschritt und prüfe, ob $h_n > h_{\min}$ ist. Wenn ja, dann gehe zu Schritt 1 (wiederhole den Zeitschritt mit kleinerer Schrittweite), ansonsten brich das Verfahren ab (Toleranz kann mit Schrittweite $h_{\min}$ nicht erreicht werden).

(3.60)

Fehlern sowie die Verwendung von toleranz-gewichteten Normen. Letzteres führt auf gewichtete Fehlermaße und soll nun genauer erläutert werden.

Man gibt absolute und relative Toleranzen ϵ_a und ϵ_r vor und definiert gewichtete Fehlermaße basierend auf der *Euklid*schen Norm (diskrete 2-Norm) oder der Maximum-Norm (diskrete Unendlich-Norm):

$$\begin{aligned} errtol_2 &= \|\hat{\boldsymbol{y}}_{n+1} - \boldsymbol{y}_{n+1}\|_{2,w} = \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left[\frac{\hat{y}_{n+1}^k - y_{n+1}^k}{\epsilon_r \cdot |y_n^k| + \epsilon_a}\right]^2}, \\ errtol_\infty &= \|\hat{\boldsymbol{y}}_{n+1} - \boldsymbol{y}_{n+1}\|_{\infty,w} = \max_{1\leq k\leq m}\left|\frac{\hat{y}_{n+1}^k - y_{n+1}^k}{\epsilon_r \cdot |y_n^k| + \epsilon_a}\right|. \end{aligned} \tag{3.61}$$

Die Fehlermaße gewichten den Fehler also komponentenweise mit den gegebenen Toleranzen, so daß für $errtol \leq 1$ der Fehler (zumindest im Mittel) kleiner als die geforderte Toleranz ist. Welches der beiden Fehlermaße besser geeignet ist, muß problemabhängig entschieden werden, wobei auch Mischformen möglich sind.

Bemerkung: In der Praxis werden häufig komponentenweise verschiedene absolute Toleranzen $\epsilon_{a,i}$ statt einer gemeinsamen absoluten Toleranz ϵ_a vorgegeben, was den verschiedenen Größenordnungen der einzelnen Variablen Rechnung trägt. □

Die Bestimmung einer neuen Schrittweite wird zunächst für den absoluten Fehler *err* und eine absolute Toleranz *tol* erläutert. Aufgrund von (3.50) gilt für den absoluten Fehler $err \approx C\, h^{\hat{p}+1}$, wobei $\hat{p}$ die niedrigere der beiden Ordnungen ist. Man fordert nun zur Bestimmung einer neuen Schrittweite, daß der Fehler gerade gleich der vorgegebenen Toleranz ist: $tol \approx C\, h_{\text{new}}^{\hat{p}+1}$. Durch Auflösen der ersten Gleichung nach C und Einsetzen in die zweite erhält man dann die neue Schrittweite

$$h_{\text{new}} = h \cdot \left(\frac{tol}{err} \right)^{1/(\hat{p}+1)} . \tag{3.62}$$

Der Quotient in (3.62) entspricht bei toleranz-gewichteten Fehlermaßen gerade dem inversen Fehlermaß $1/errtol$. Da in der Fehlerschätzung die Glieder höherer Ordnung vernachlässigt wurden, multipliziert man in der Praxis die neue Schrittweite noch mit einem Sicherheitsfaktor kleiner als Eins und begrenzt zur Stabilisierung der Steuerung die Schrittweitenänderungen nach oben und nach unten. Dies führt insgesamt zu dem in Kasten (3.60) dargestellten Algorithmus für einen gesteuerten Zeitschritt.

Bemerkung: Bei der Behandlung von Plastizitätsproblemen ist die folgende Wahl der o. g. Normen sinnvoll (*Ehlers & Ellsiepen* [49]). Für die FE-Freiheitsgrade $\boldsymbol{u}$, die aufgrund der schwachen Formulierung im L^2-Sinn berechnet werden, wird die diskrete 2-Norm verwendet. Damit erhält man das Fehlermaß

$$e_u := \|\hat{\boldsymbol{u}}_{n+1} - \boldsymbol{u}_{n+1}\|_{2,w} .$$

Die internen Variablen $\boldsymbol{q}$ werden an den Integrationspunkten der numerischen Quadratur im Sinne eines Kollokationsverfahrens berechnet. Dies legt die Verwendung einer diskreten Maximumnorm zur Messung des Fehlers nahe:

$$e_q := \|\hat{\boldsymbol{q}}_{n+1} - \boldsymbol{q}_{n+1}\|_{\infty,w} .$$

Insgesamt kann damit das Fehlermaß des volldiskreten Problems definiert werden:

$$errtol := \max\{e_u, e_q\} .$$

Durch diese Wahl ist zum einen sichergestellt, daß der Fehler in den diskreten Primärvariablen $\boldsymbol{u}$ im Mittel innerhalb der geforderten Toleranzen liegt. Zum anderen dominiert im plastischen Bereich derjenige Integrationspunkt mit dem größten Fehler die Wahl der Zeitschrittweite, so daß der für die Gesamtrechnung entscheidende Zeitpunkt des Einsetzens plastischer Deformationen zuverlässig lokalisiert werden kann. □

3.5 Ein neues eingebettetes SDIRK-Verfahren

In diesem Abschnitt wird auf der Basis des zweistufigen stark S-stabilen Verfahrens von *Alexander* [5] mit Ordnung $p = 2$ ein eingebettetes Verfahren der Ordnung $\hat{p} = 1$ konstruiert, das zur Fehlerschätzung gemäß Abschnitt 3.4.2 verwendet werden kann. Die beiden

möglichen stark S-stabilen zweistufigen Verfahren der Ordnung $p = 2$ lauten [5]:

$$\begin{array}{c|cc} \alpha & \alpha & \\ 1 & 1-\alpha & \alpha \\ \hline & 1-\alpha & \alpha \\ & \hat{b}_1 & \hat{b}_2 \end{array} \qquad \alpha = 1 \pm \frac{1}{2}\sqrt{2}\,. \tag{3.63}$$

Als Ausgangspunkt wird das Verfahren mit $\alpha = 1 - \frac{1}{2}\sqrt{2}$ gewählt, bei dem alle Gewichte positiv und zudem die Stufenpositionen c_i im Intervall $[0, 1]$ aufsteigend sortiert sind. Dies ist insbesondere bei der Zeitintegration von Plastizitätsproblemen mit Fließbedingung von Vorteil, da auf diese Weise stets „in der Zeit vorwärts“ integriert wird.

Damit das eingebettete Verfahren zu den Gewichten $\hat{b}_i$ die Ordnung $\hat{p} = 1$ besitzt, muß gelten:

$$\hat{b}_1 + \hat{b}_2 = 1 \qquad \Longrightarrow \qquad \hat{b}_1 := 1 - \hat{\alpha}, \quad \hat{b}_2 := \hat{\alpha}\,. \tag{3.64}$$

Es verbleibt also nur noch ein freier Parameter $\hat{\alpha}$, der durch eine zusätzliche Stabilitätsforderung bestimmt werden kann. Fordert man, daß das eingebettete Verfahren ebenfalls L-stabil ist, so erhält man die Bedingung (vgl. (3.25))

$$\alpha_0 := R(\infty) = 1 - \boldsymbol{b}^T \boldsymbol{A}^{-1} \mathbb{1} \stackrel{!}{=} 0 \quad \Longrightarrow \quad \alpha_0 = \frac{\alpha(\alpha-1) - \hat{\alpha}(\alpha-1)}{\alpha^2} \stackrel{!}{=} 0\,. \tag{3.65}$$

Diese ist aber nur für $\hat{\alpha} = \alpha$ erfüllbar, so daß man das ursprüngliche Verfahren erhält, was aber für eine Fehlerschätzung unbrauchbar ist. Daher wird statt der L-Stabilität nun lediglich die S-Stabilität[7] des eingebetteten Verfahrens gefordert:

$$|\alpha_0| < 1 \quad \Longrightarrow \quad 3 - 2\sqrt{2} = \frac{-\alpha(2\alpha-1)}{1-\alpha} \;<\; \hat{\alpha} \;<\; \frac{\alpha}{1-\alpha} = \sqrt{2} - 1\,. \tag{3.66}$$

Die Mitte des zulässigen Intervalls für $\hat{\alpha}$ ist α. Zum Zweck der Fehlerschätzung ist ein möglichst großer Abstand zwischen den Koeffizienten α und $\hat{\alpha}$ sinnvoll. Andererseits ist der optimale Wert für $\hat{\alpha}$ im Hinblick auf Stabilität der Wert α selbst. Es wird daher willkürlich der Mittelwert zwischen der unteren Stabilitätsgrenze $3-2\sqrt{2}$ und $\alpha = 1-\frac{1}{2}\sqrt{2}$ gewählt, so daß die Koeffizienten des neuen SDIRK-2(1)-Verfahrens der Ordnung $p = 2$ mit eingebettetem Verfahren der Ordnung $\hat{p} = 1$ lauten:

$$\begin{array}{c|cc} \alpha & \alpha & \\ 1 & 1-\alpha & \alpha \\ \hline & 1-\alpha & \alpha \\ & 1-\hat{\alpha} & \hat{\alpha} \end{array} \qquad \alpha = 1 - \frac{1}{2}\sqrt{2}\,, \quad \hat{\alpha} = 2 - \frac{5}{4}\sqrt{2}\,. \tag{3.67}$$

Die in diesem Kapitel angegebenen adaptiven Zeitintegrationsverfahren werden in Kapitel 5 im Rahmen der numerischen Beispielrechnungen verglichen und bewertet. Bei der Konstruktion des Gesamtverfahrens ist der nächste Schritt die Einbeziehung der adaptiven Ortsdiskretisierung. Dies ist Thema des folgenden Kapitels.

[7] Man beachte, daß dies automatisch die A-Stabilität einschließt, vgl. Seite 70.

Kapitel 4: Adaptive Ortsdiskretisierung

In Kapitel 2 wurde bei der Darstellung der Methode der finiten Elemente vereinfachend davon ausgegangen, daß die Ortsdiskretisierung mit einem fest gewählten Finite-Elemente-Netz durchgeführt wird. Bei „einfachen“ Problemstellungen ist diese Annahme durchaus gerechtfertigt, da sich mit Hilfe der ungefähren Kenntnis der zu erwartenden Lösung ein Netz erzeugen läßt, das der Problemstellung angemessen ist. Im allgemeinen kennt man aber das Verhalten der Lösung nicht im voraus, so daß es im Sinne der Effizienz numerischer Verfahren wünschenswert ist, neben der Zeitdiskretisierung auch die Ortsdiskretisierung variabel an das lokale Verhalten der Lösung anzupassen.

Für das volldiskrete Verfahren spielt dabei die Kopplung der adaptiven Zeitdiskretisierung mit der adaptiven Ortsdiskretisierung eine wichtige Rolle. Eine naheliegende Möglichkeit der Kopplung besteht darin, jeden Zeitschritt als ein stationäres Teilproblem zu betrachten, so daß sich die numerische Lösung des Anfangs-Randwertproblems durch eine Aneinanderreihung von stationären Problemen ergibt.

Es wird daher in diesem Kapitel zunächst auf die adaptive Ortsdiskretisierung stationärer Probleme eingegangen, bevor die speziellen Aspekte zeitabhängiger Probleme behandelt werden. Man unterscheidet bei der adaptiven Ortsdiskretisierung die folgenden Varianten:

- *h-Adaptivität*: Die Netzdichte – repräsentiert durch den Diskretisierungsparameter h (Elementradius) – wird an das lokale Verhalten der Lösung angepaßt, wobei im gesamten Netz mit Elementen gleicher Ansatzordnung gearbeitet wird.

- *p-Adaptivität*: Die Finite-Elemente-Ansätze – repräsentiert durch den Polynomgrad p der Ansatzfunktionen – werden an das lokale Verhalten der Lösung angepaßt, wobei das Netz selbst nicht verändert wird.

- *h-p-Adaptivität*: Sowohl die Netzdichte h als auch der Polynomgrad p werden an das lokale Verhalten der Lösung angepaßt.

Bemerkung: In manchen Arbeiten werden noch andere adaptive Strategien betrachtet, etwa die bestmögliche Ausrichtung eines Netzes mit vorgegebener Anzahl von Freiheitsgraden gemäß einem benutzerdefinierten Kriterium (r-Adaptivität).

Außerdem findet man in der Literatur (z. B. *Stein & Ohnimus* [108]) adaptive Strategien, die nicht die effiziente numerische Lösung eines gegebenen Anfangs-Randwertproblems zum Ziel haben, sondern vielmehr die automatische Auswahl einer geeigneten Raumdimension (d-Adaptivität, z. B. Übergang vom zwei- zum dreidimensionalen Modell bei Auftreten gewisser Schubspannungen) bzw. eines geeigneten Modells (m-Adaptivität, z. B. Übergang vom elastischen zum elastoplastischen Modell oder umgekehrt). □

Die Verwendung von p-adaptiven Strategien setzt eine hohe Regularität der Lösung voraus. So erhält man beispielsweise an Sprungstellen der Lösung die bei Polynomen bekannten „Überschwinger“ (*Gibbs*sches Phänomen), die dazu führen, daß trotz des großen

Aufwands für den Polynomansatz höherer Ordnung nur eine geringe Approximationsordnung erreicht wird. Ein weiterer Nachteil von p-Methoden in bezug auf die Effizienz beim Lösen der resultierenden linearen Gleichungssysteme ist das starke Anwachsen der Bandbreite mit zunehmender Ansatzordnung, was insbesondere beim Übergang auf den räumlich 3-dimensionalen Fall eine dramatische Erhöhung des Lösungsaufwands zur Folge hat. Um dies zu verhindern, sind stark spezialisierte Lösungsverfahren in Verbindung mit hierarchisch organisierten Ansatzfunktionen notwendig; Details zur effizienten Implementierung der p-Version sowie der h-p-Version der adaptiven FEM können z. B. den Arbeiten von *Demkowicz, Oden et al.* [37, 38, 98] oder *Ainsworth & Senior* [3] entnommen werden.

Szabó & Babuška [112, Kap. 16] klassifizieren zu behandelnde Randwertprobleme in die drei Kategorien A, B und C mit zunehmendem Schwierigkeitsgrad. Die im Rahmen der vorliegenden Arbeit betrachteten Probleme mit lösungs- und pfadabhängigen Singularitäten sind demnach *stark in Kategorie C*. Für derartige Probleme empfehlen die Autoren die Verwendung der h-Version der FEM.

Im Rahmen dieser Arbeit wird aus den o. g. Gründen nur die h-Adaptivität eingesetzt, die zudem als Analogon zur Schrittweitensteuerung im Zeitbereich betrachtet werden kann. Im Gegensatz zur Bestimmung einer geeigneten Zeitschrittweite ist die Konstruktion eines möglichst ökonomischen Netzes jedoch u. a. aus den folgenden Gründen ungleich schwieriger:

- Das Problem ist grundsätzlich 2- bzw. 3-dimensional.
- Die Abschätzung des Fehlers der FE-Diskretisierung ist schwierig und problemabhängig. Für die in dieser Arbeit betrachteten Anfangs-Randwertprobleme ist man nach heutigem Wissensstand auf heuristische Fehlerindikatoren angewiesen, da bisher keine echten Fehlerabschätzungen bekannt sind.
- Zur Verwaltung der verschiedenen Netze sind umfangreiche Datenstrukturen und Algorithmen notwendig.

Die adaptive Ortsdiskretisierung eines stationären Problems mit finiten Elementen läßt sich in die folgenden Teilaufgaben zerlegen:

1. *Fehlerindikator*: Berechnung von lokalen und globalen Kenngrößen zur Beurteilung des Fehlers einer gegebenen Ortsdiskretisierung.
2. *Netzanpassungs-Strategie*: Umsetzung der berechneten Fehler-Kenngrößen in eine Soll-Dichtefunktion für das neue Netz (Elementgröße als Funktion des Ortes) auf der Grundlage gewisser (heuristischer) Optimalitätskriterien.
3. *Netzgenerierung*: Erzeugung eines neuen Netzes bzw. hierarchische Verfeinerung und Vergröberung des bestehenden Netzes gemäß der zuvor berechneten Dichtefunktion.

Bei zeitabhängigen Problemen ist zusätzlich noch die folgende Teilaufgabe notwendig, damit die Rechnung nicht bei jeder Netzänderung neu gestartet werden muß:

4. *Datentransfer*: Geeigneter Transfer der diskreten Zustandsgrößen (Freiheitsgrade und interne Variablen) vom alten zum neuen Netz.

In den folgenden Abschnitten werden nun zunächst diese Teilaufgaben behandelt, bevor auf den Gesamtalgorithmus sowie die Kopplung mit den adaptiven Zeitintegrationsverfahren aus Kapitel 3 eingegangen wird.

4.1 Fehlerschätzer und Fehlerindikatoren

Die Abschätzung von Fehlern gehört zu den zentralen Aufgaben der numerischen Mathematik, da nur so die Qualität von numerischen Lösungen beurteilt werden kann. Man unterscheidet generell *A-priori*-Fehlerabschätzungen, die man vor der eigentlichen Rechnung ausschließlich mit Hilfe bekannter Daten (etwa Anfangs- und Randwerten sowie Materialparametern) auswerten kann, und *A-posteriori*-Fehlerabschätzungen, die man nach der Rechnung bzw. nach einem Teilschritt der Rechnung auswertet.

A-priori-Fehlerabschätzungen spielen vor allem bei theoretischen Aussagen eine Rolle, etwa bei der Untersuchung von Konvergenzeigenschaften, und werden in der Praxis selten eingesetzt. Dies liegt auch daran, daß sie i. a. weniger scharfe Fehlerschranken liefern als *A-posteriori*-Fehlerabschätzungen, da man ja keinerlei Zusatzinformationen aus der numerischen Rechnung verwendet. Im Zusammenhang mit finiten Elementen findet man Standard-Resultate z. B. bei *Strang & Fix* [109], *Brezzi & Fortin* [28], *Brenner & Scott* [27], *Braess* [25].

Der Einsatz von *A-posteriori*-Fehlerabschätzungen zur Steuerung der Netzdichte bei Finite-Elemente-Diskretisierungen ist seit dem Ende der siebziger Jahre intensives Forschungsthema, sowohl im mathematischen als auch im ingenieurwissenschaftlichen Bereich. Der weitaus größte Teil der mathematisch orientierten Arbeiten beschäftigte sich zunächst mit der Fehlerschätzung bei elliptischen Problemen anhand des Modellproblems der *Poisson*-Gleichung bzw. leichten Abwandlungen davon, wie etwa Problemen aus der linearen Elastizitätstheorie oder dem *Stokes*-Problem aus der Fluidmechanik. Eine aktuelle Übersicht über bestehende Ansätze zur Gewinnung von *A-posteriori*-Fehlerabschätzungen bei finiten Elementen gibt *Verfürth* [120]. Erst in den letzten Jahren werden auch schwierigere Probleme angegangen, wie etwa die finite Elastizitätstheorie (*Mücke & Whiteman* [85]) sowie die stationären, inkompressiblen *Navier-Stokes*-Gleichungen bzw. nichtlineare elliptische Probleme (*Verfürth* [119, 120] und Referenzen darin). Probleme aus der Plastizitätstheorie werden z. B. von *Suttmeier* [111], *Rannacher & Suttmeier* [99] behandelt. Hier wird ein Fehlerschätzer für die Deformationstheorie der Plastizität (*Hencky*-Plastizität) sowie dessen Übertragung auf die Fließtheorie der Plastizität (*Prandtl-Reuß*-Plastizität) im Sinne eines Fehlerindikators angegeben.

In den Anwendungen werden aufgrund der o. g. fehlenden mathematischen Fundierung häufig Fehlerindikatoren verwendet, die entweder physikalisch motiviert sind oder durch Analogiebetrachtungen zu einfacheren Modellen mit existierender Theorie hergeleitet werden. Zur numerischen Behandlung von Problemen aus der Plastizitätstheorie mit ortsadaptiven Methoden findet man verschiedene Ansätze u. a. bei *Ortiz & Quigley* [91],

Pastor, Peraire & Zienkiewicz [92], *Perić, Yu & Owen* [94] (statische Verzerrungslokalisierung), *Deb, Prevost & Loret* [36] (dynamische Verzerrungslokalisierung), *Wriggers & Scherf* [125] (Kontaktprobleme).

Im wesentlichen existieren folgende Ansätze zur Gewinnung von *A-posteriori*-Fehlerschätzern bzw. Fehlerindikatoren:

1. *Residuen-basierte Fehlerschätzer*: Der Fehler in der berechneten numerischen Lösung wird mit Hilfe des Residuums bzgl. der starken Form der Differentialgleichung abgeschätzt.

2. *Fehlerschätzung durch Lösung lokaler Probleme*: Mit Hilfe lokaler Probleme, die ähnlich, aber einfacher als das ursprüngliche Problem sind, berechnet man lokale Vergleichslösungen, die zur Fehlerschätzung herangezogen werden.

3. *Hierarchische Fehlerschätzer*: Die berechnete numerische Lösung wird zur Fehlerabschätzung mit einer Lösung verglichen, die mit einem Finite-Elemente-Raum höherer Ordnung berechnet wird. Letzteren erhält man entweder durch Verwendung hierarchischer Basen – und damit Ansatzfunktionen höherer Ordnung – oder durch eine Netzverfeinerung (*Richardson*-Extrapolation im Ortsbereich).

4. *Gradienten-basierte Fehlerindikatoren*: Für gewisse Gradienten der numerischen Lösung wird eine lokal „verbesserte“ Lösung berechnet und mit den aus den Ansatzfunktionen berechneten Gradienten verglichen. Da die Gradienten von FE-Lösungen i. a. Sprünge an Elementgrenzen aufweisen, berechnet man die „verbesserten“ Gradienten durch lokale Glättungs- bzw. Extrapolationsoperationen. Unter Ausnutzung von Superkonvergenzeigenschaften an ausgezeichneten Punkten im finiten Element kann auf diese Weise eine Vergleichslösung höherer Ordnung erzeugt werden. Da damit in den meisten Fällen allerdings keine echte Fehlerabschätzung vorliegt, spricht man hier von Fehlerindikatoren.

In den folgenden Abschnitten werden die verschiedenen Fehlerschätzer und Fehlerindikatoren exemplarisch anhand des Randwertproblems der linearen Elastostatik in zwei Raumdimensionen in einer vereinheitlichten Notation zusammengefaßt. Aus dem Verschiebungsvektor $\mathbf{u}$ ergibt sich der lineare Verzerrungstensor $\boldsymbol{\varepsilon}$ sowie über das *Hooke*sche Gesetz mit den *Lamé*-Konstanten $\lambda, \mu > 0$ der Spannungstensor $\boldsymbol{\sigma}$:

$$\begin{aligned} \boldsymbol{\varepsilon}(\mathbf{u}) &= \frac{1}{2}(\operatorname{grad}\mathbf{u} + \operatorname{grad}^T\mathbf{u}), \\ \boldsymbol{\sigma}(\mathbf{u}) &= 2\,\mu\,\boldsymbol{\varepsilon}(\mathbf{u}) + \lambda\,(\operatorname{tr}\boldsymbol{\varepsilon}(\mathbf{u}))\,\mathbf{I}. \end{aligned} \tag{4.1}$$

In der starken Formulierung des Randwertproblems ist das Verschiebungsfeld $\mathbf{u} : \Omega \longrightarrow \mathbb{R}^2$ auf dem Gebiet $\Omega \subset \mathbb{R}^2$ gesucht, so daß gilt:

$$\left.\begin{array}{rcll} -\operatorname{div}\boldsymbol{\sigma}(\mathbf{u}) &=& \mathbf{f} & \text{in} \quad \Omega \\ \mathbf{u} &=& \bar{\mathbf{u}} & \text{auf} \quad \Gamma_{\mathrm{u}} \\ \boldsymbol{\sigma}(\mathbf{u})\,\mathbf{n} &=& \bar{\mathbf{t}} & \text{auf} \quad \Gamma_{\mathrm{t}} \end{array}\right\} \quad \partial\Omega = \Gamma = \Gamma_{\mathrm{u}} \cup \Gamma_{\mathrm{t}}, \;\; \emptyset = \Gamma_{\mathrm{u}} \cap \Gamma_{\mathrm{t}}. \tag{4.2}$$

Dabei werden die üblichen Annahmen bzgl. der Regularität des Gebiets (*Lipschitz*-Berandung), der Vorgabe der Randwerte (Γ_u mit nicht-verschwindendem $(d-1)$-dimensionalen *Lebesgue*-Maß) sowie der Glattheit der rechten Seite und der *Neumann*-Randbedingungen gemacht.

Für die schwache Formulierung des Randwertproblems werden zunächst mit den Test- und Ansatzfunktionen $\boldsymbol{\eta}$, $\mathbf{u} \in [H^1(\Omega)]^2$ die bei geeignet gewählten Materialparametern positiv definite Bilinearform (virtuelle Arbeit im Inneren des Körpers)

$$B(\boldsymbol{\eta}, \mathbf{u}) \;\equiv\; \int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{\eta}) \cdot \boldsymbol{\sigma}(\mathbf{u}) \,\mathrm{d}v \;=\; \int_\Omega (\lambda \,\mathrm{div}\,\boldsymbol{\eta}\,\mathrm{div}\,\mathbf{u} + 2\,\mu\,\boldsymbol{\varepsilon}(\boldsymbol{\eta}) \cdot \boldsymbol{\varepsilon}(\mathbf{u})) \,\mathrm{d}v \tag{4.3}$$

sowie das lineare Funktional (virtuelle Arbeit der Volumen- und Oberflächenkräfte)

$$F(\boldsymbol{\eta}) \;\equiv\; \int_\Omega \boldsymbol{\eta} \cdot \mathbf{f} \,\mathrm{d}v + \int_{\Gamma_\mathrm{t}} \boldsymbol{\eta} \cdot \bar{\mathbf{t}} \,\mathrm{d}a \tag{4.4}$$

eingeführt. Gesucht ist dann eine Funktion $\mathbf{u} \in \mathcal{S}_\mathrm{u}$, so daß

$$B(\boldsymbol{\eta}, \mathbf{u}) = F(\boldsymbol{\eta}) \qquad\qquad \forall\, \boldsymbol{\eta} \in \mathcal{T} \tag{4.5}$$

gilt, wobei die Funktionen im Ansatzraum $\mathcal{S}_\mathrm{u}$ wie üblich die *Dirichlet*-Randbedingungen des Problems (4.2) und die Funktionen im Testraum $\mathcal{T}$ entsprechende homogene *Dirichlet*-Randbedingungen erfüllen (vgl. dazu auch die Herleitung der schwachen Formulierung des quasi-statischen Zweiphasenmodells in Abschnitt 2.2.2 sowie die Darstellung einer allgemeinen schwachen Formulierung in Abschnitt 2.4.1). Das in Gleichung (4.3) definierte Skalarprodukt $B(\cdot,\cdot)$ induziert eine Norm $|||\cdot|||$, die aus Dimensionsgründen[1] als *Energienorm* bezeichnet wird:

$$|||\mathbf{u}||| = \sqrt{B(\mathbf{u}, \mathbf{u})}, \qquad\qquad \mathbf{u} \in [H^1(\Omega)]^2. \tag{4.6}$$

Als natürliche Norm bei der Berechnung von schwachen Lösungen innerhalb der FEM wird diese häufig zur Messung von Fehlern herangezogen.

Die mit dem *Galerkin*-Verfahren auf einem endlichdimensionalen Finite-Elemente-Raum $\mathcal{S}_\mathrm{u}^h = \bar{\mathbf{u}} + \mathcal{T}^h \subset \mathcal{S}_\mathrm{u}$ berechnete FE-Näherungslösung wird wieder mit $\mathbf{u}^h$ bezeichnet:

$$B(\boldsymbol{\eta}^h, \mathbf{u}^h) = F(\boldsymbol{\eta}^h) \qquad\qquad \forall\, \boldsymbol{\eta}^h \in \mathcal{T}^h \subset \mathcal{T}. \tag{4.7}$$

Für den Fehler

$$\mathbf{e} := \mathbf{u} - \mathbf{u}^h \tag{4.8}$$

gilt mit (4.5) und (4.7) die bekannte *Galerkin*-Orthogonalitätsbeziehung

$$B(\boldsymbol{\eta}^h, \mathbf{e}) = \underbrace{B(\boldsymbol{\eta}^h, \mathbf{u})}_{F(\boldsymbol{\eta}^h)} - \underbrace{B(\boldsymbol{\eta}^h, \mathbf{u}^h)}_{F(\boldsymbol{\eta}^h)} = 0 \qquad\qquad \forall\, \boldsymbol{\eta}^h \in \mathcal{T}^h, \tag{4.9}$$

d. h. der Fehler steht „senkrecht“ (bzgl. des Skalarprodukts $B(\cdot,\cdot)$) auf dem Ansatz- bzw. Testraum (Bestapproximations-Eigenschaft des *Galerkin*-Verfahrens).

[1] Gemäß (4.3) ist $B(\mathbf{u}, \mathbf{u})$ das Volumenintegral über Verzerrungen mal Spannungen mit der physikalischen Dimension Energie: $[\boldsymbol{\varepsilon} \cdot \boldsymbol{\sigma} \,\mathrm{d}v] = 1\,\mathrm{N/m^2} \cdot \mathrm{m^3} = 1\,\mathrm{Nm} = 1\,\mathrm{J}$.

Man kann statt des Fehlers (4.8) auch den Fehler $\mathbf{e}_\sigma$ in den Spannungen betrachten, der üblicherweise in der L^2-Norm gemessen wird:

$$\mathbf{e}_\sigma := \boldsymbol{\sigma}(\mathbf{e}) = \boldsymbol{\sigma}(\mathbf{u}) - \boldsymbol{\sigma}(\mathbf{u}^h)\,, \qquad \|\mathbf{e}_\sigma\|_{L^2(\Omega)} = \left(\int_\Omega \mathbf{e}_\sigma \cdot \mathbf{e}_\sigma \,\mathrm{d}v\right)^{1/2}. \tag{4.10}$$

Im Vergleich mit dem Fehler $\mathbf{e}$ in der Energienorm,

$$|||\mathbf{e}||| = \left(\int_\Omega \boldsymbol{\sigma}(\mathbf{e}) \cdot \boldsymbol{\varepsilon}(\mathbf{e}) \,\mathrm{d}v\right)^{1/2} = \left(\int_\Omega \mathbf{e}_\sigma \cdot \overset{4}{\mathbf{C}}{}^{-1}\mathbf{e}_\sigma \,\mathrm{d}v\right)^{1/2}, \tag{4.11}$$

entfällt lediglich die Multiplikation mit dem inversen Materialtensor.

Der Finite-Elemente-Raum $\mathcal{S}^h_\mathrm{u}$ wird durch ein Netz $\mathcal{T}^h$ aus Elementen T dargestellt, deren Ansatzfunktionen jeweils Polynome sind. Im folgenden werden noch einige Bezeichnungen benötigt:

$\mathcal{T}^h = \{T\}$	Menge aller Elemente (Netz)		
T	Element		
h_T	Elementdurchmesser		
$\mathcal{E}^h = \{E\}$	Menge aller Kanten	$\mathcal{N}^h = \{N\}$	Menge aller Knoten
E	Kante	N	Knoten
h_E	Kantenlänge		
$\mathcal{E}^h_\Omega$	Innere Kanten	$\mathcal{N}^h_\Omega$	Innere Knoten
$\mathcal{E}^h_{\Gamma_\mathrm{u}}$	*Dirichlet*-Randkanten	$\mathcal{N}^h_{\Gamma_\mathrm{u}}$	*Dirichlet*-Randknoten
$\mathcal{E}^h_{\Gamma_\mathrm{t}}$	*Neumann*-Randkanten	$\mathcal{N}^h_{\Gamma_\mathrm{t}}$	*Neumann*-Randknoten

4.1.1 Residuen-basierte Fehlerschätzer

Dieser Fehlerschätzer ist der erste in der Literatur dokumentierte *A-posteriori*-Fehlerschätzer für finite Elemente und wurde von *Babuška & Rheinboldt* [8] für ein eindimensionales Modellproblem angegeben. Die Ideen wurden danach vielfach aufgegriffen und erweitert, z. B. geben *Babuška & Miller* [7] einen Schätzer für das Randwertproblem der linearen Elastostatik an.

Das zugrundeliegende Prinzip ist die Abschätzung des Fehlers in der Energienorm durch das Residuum der FE-Lösung im Inneren und am Rand. Man gelangt dadurch zu Abschätzungen, die nur noch berechenbare Größen enthalten. Von *Johnson* und Mitarbeitern (z. B. *Johnson & Hansbo* [74], *Eriksson et. al* [54]) stammt die Idee, die in den Fehlerabschätzungen auftretenden Konstanten quantitativ zu bestimmen, um auf diese Weise eine Garantie für das Einhalten des Toleranzkriteriums durch den adaptiven Algorithmus zu erhalten. Sie unterscheiden Interpolationskonstanten C^i, die nur von den gewählten Elementen abhängen, und Stabilitätskonstanten C^s, die nur vom betrachteten Problem und der gewählten Norm abhängen (im Fall des hier behandelten Problems und der Energienorm ist $C^s = 1$). Bei komplexeren Problemen werden letztere durch Lösung von diskreten dualen Problemen bestimmt. Im Gegensatz zu den Arbeiten von *Babuška* und Mitarbeitern, die fast ausschließlich Abschätzungen in der für die Praxis

recht unanschaulichen Energienorm betrachten, gestattet das Vorgehen von *Johnson* und Mitarbeitern die Fehlerabschätzung in beliebigen Normen und ist zudem auf eine größere Problemklasse, angefangen von linear elliptischen Problemen über parabolische Probleme bis hin zu nichtlinearen hyperbolischen Problemen, anwendbar.

Im folgenden wird die Herleitung des Fehlerschätzers in der Energienorm in Anlehnung an *Johnson & Hansbo* [74] und *Verfürth* [120] in Kurzform dargestellt. Wegen (4.5) gilt mit $\boldsymbol{\eta} := \mathbf{e}$ für den Fehler in der Energienorm:

$$|||\mathbf{e}|||^2 \;=\; B(\mathbf{e},\mathbf{e}) \;=\; B(\mathbf{e},\mathbf{u}-\mathbf{u}^h) \;=\; B(\mathbf{e},\mathbf{u}) - B(\mathbf{e},\mathbf{u}^h) \;=\; F(\mathbf{e}) - B(\mathbf{e},\mathbf{u}^h).$$

Die Anwendung des Interpolations-Operators $I^h : [H^1(\Omega)]^2 \longrightarrow \mathcal{S}^h_{\mathrm{u}}$ (lokale L^2-Projektion auf die Knoten, vgl. *Verfürth* [120, §1.2]) auf den Fehler $\mathbf{e}$ führt mit $\boldsymbol{\eta}^h = I^h\mathbf{e}$ in (4.7) und der Abkürzung $\tilde{\mathbf{e}} := \mathbf{e} - I^h\mathbf{e}$ für den Interpolationsfehler auf:

$$|||\mathbf{e}|||^2 \;=\; F(\mathbf{e}-I^h\mathbf{e}) - B(\mathbf{e}-I^h\mathbf{e},\mathbf{u}^h) \;=\; F(\tilde{\mathbf{e}}) - B(\tilde{\mathbf{e}},\mathbf{u}^h).$$

An dieser Stelle kommt nun die FE-Diskretisierung direkt ins Spiel, indem in allen Elementen T der zweite Term auf der rechten Seite partiell integriert wird. Dies liefert mit $\boldsymbol{\sigma}^h := \boldsymbol{\sigma}(\mathbf{u}^h)$ zunächst

$$\begin{aligned}
|||\mathbf{e}|||^2 \;&=\; \underbrace{\int_\Omega \tilde{\mathbf{e}}\cdot\mathbf{f}\,\mathrm{d}v \;+\; \int_{\Gamma_{\mathrm{t}}} \tilde{\mathbf{e}}\cdot\bar{\mathbf{t}}\,\mathrm{d}a}_{F(\tilde{\mathbf{e}})} \;-\; \underbrace{\int_\Omega \boldsymbol{\varepsilon}(\tilde{\mathbf{e}})\cdot\boldsymbol{\sigma}^h\,\mathrm{d}v}_{B(\tilde{\mathbf{e}},\mathbf{u}^h)} \\
&=\; \sum_{T\in\mathcal{T}^h}\int_T \tilde{\mathbf{e}}\cdot(\mathbf{f}+\operatorname{div}\boldsymbol{\sigma}^h)\,\mathrm{d}v \;+\; \int_{\Gamma_{\mathrm{t}}} \tilde{\mathbf{e}}\cdot\bar{\mathbf{t}}\,\mathrm{d}a \;-\; \sum_{T\in\mathcal{T}^h}\int_{\partial T} \tilde{\mathbf{e}}\cdot\boldsymbol{\sigma}^h\,\mathbf{n}_T\,\mathrm{d}a\,,
\end{aligned}$$

wobei $\mathbf{n}_T$ den nach außen gerichteten Normaleneinheitsvektor auf dem Rand ∂T des Elementes T bezeichnet. Durch Umsortieren der Terme auf innere Kanten und Randkanten sowie Ausnutzen der Tatsache, daß $\tilde{\mathbf{e}}$ per Definition auf *Dirichlet*-Randkanten verschwindet, erhält man

$$\begin{aligned}
|||\mathbf{e}|||^2 \;=\; &\sum_{T\in\mathcal{T}^h}\int_T \tilde{\mathbf{e}}\cdot(\mathbf{f}+\operatorname{div}\boldsymbol{\sigma}^h)\,\mathrm{d}v \;-\; \sum_{E\in\mathcal{E}^h_\Omega}\int_E \tilde{\mathbf{e}}\cdot\left[\!\left[\boldsymbol{\sigma}^h\,\mathbf{n}_E\right]\!\right]\,\mathrm{d}a \\
&+\; \sum_{E\in\mathcal{E}^h_{\Gamma_{\mathrm{t}}}}\int_E \tilde{\mathbf{e}}\cdot(\bar{\mathbf{t}}-\boldsymbol{\sigma}^h\,\mathbf{n}_E)\,\mathrm{d}a.
\end{aligned}$$

Darin ist $\left[\!\left[\boldsymbol{\sigma}^h\,\mathbf{n}_E\right]\!\right]$ der Sprung des Spannungsvektors in Normalenrichtung $\mathbf{n}_E$ über die Kante E. Es werden nun die elementbezogenen *Residuen* der FE-Näherungslösung im Inneren und am Rand eingeführt[2]:

$$\begin{aligned}
\mathbf{R}_1(\mathbf{u}^h) \;&:=\; \mathbf{f}+\operatorname{div}\boldsymbol{\sigma}^h \qquad \text{auf Elementen } T\ , \\
\mathbf{R}_2(\mathbf{u}^h) \;&:=\; \begin{cases} -\frac{1}{2}\left[\!\left[\boldsymbol{\sigma}^h\,\mathbf{n}_E\right]\!\right] & \text{auf inneren Kanten } E\subset\partial T, E\in\mathcal{E}^h_\Omega\,, \\ (\bar{\mathbf{t}}-\boldsymbol{\sigma}^h\,\mathbf{n}_E) & \text{auf } \textit{Neumann}\text{-Randkanten } E\subset\partial T, E\in\mathcal{E}^h_{\Gamma_{\mathrm{t}}}\,. \end{cases}
\end{aligned} \tag{4.12}$$

[2]Man beachte, daß in den obigen Summen jede Kante nur einmal vorkommt, in einer Summe über alle Kanten aller Elemente jedoch zweimal. Dies wird im Residuum durch einen Faktor 1/2 berücksichtigt.

Damit gelangt man schließlich zur folgenden Darstellung des Fehlers in der Energienorm:

$$\begin{aligned} |||\mathbf{e}|||^2 &= \sum_{T\in\mathfrak{T}^h} \int_T \tilde{\mathbf{e}} \cdot \mathbf{R}_1(\mathbf{u}^h)\,\mathrm{d}v \;+\; \sum_{T\in\mathfrak{T}^h}\sum_{E\subset\partial T} \int_E \tilde{\mathbf{e}} \cdot \mathbf{R}_2(\mathbf{u}^h)\,\mathrm{d}a \\ &= \sum_{T\in\mathfrak{T}^h} \big(\tilde{\mathbf{e}},\, \mathbf{R}_1(\mathbf{u}^h)\big)_{L^2(T)} \;+\; \sum_{T\in\mathfrak{T}^h}\sum_{E\subset\partial T} \big(\tilde{\mathbf{e}},\, \mathbf{R}_2(\mathbf{u}^h)\big)_{L^2(E)}\,. \end{aligned}$$

Bis hierher wurden noch keine Abschätzungen vorgenommen; der Fehler wurde lediglich in elementbezogenen residualen Größen und dem Interpolationsfehler $\tilde{\mathbf{e}}$ dargestellt. Man kann zeigen, daß die L^2-Norm des Interpolationsfehlers durch die Energienorm des Gesamtfehlers auf einem Patch von Knoten-Nachbarn ω_T bzw. ω_E abgeschätzt werden kann (*Clément* [33], *Ciarlet* [31, Beispiel 3.2.3], *Verfürth* [120, §1.1]):

$$\begin{aligned} \|\tilde{\mathbf{e}}\|_{L^2(T)} &= \|\mathbf{e} - I^h\mathbf{e}\|_{L^2(T)} \;\le\; C_1\, h_T\, |||\mathbf{e}|||_{\omega_T}\,, \\ \|\tilde{\mathbf{e}}\|_{L^2(E)} &= \|\mathbf{e} - I^h\mathbf{e}\|_{L^2(E)} \;\le\; C_2\, h_E^{1/2}\, |||\mathbf{e}|||_{\omega_E}\,. \end{aligned} \tag{4.13}$$

Die beiden Konstanten C_1 und C_2 sind dabei von den Diskretisierungsparametern h_T bzw. h_E unabhängig. Damit erhält man schließlich durch zweimalige Anwendung der *Cauchy-Schwarz*schen (C.-S.) Ungleichung $|(f,g)| \le \|f\|\,\|g\|$, einmal für das L^2-Skalarprodukt und einmal für endliche Summen, die folgende Abschätzung:

$$\begin{aligned} |||\mathbf{e}|||^2 &\overset{\text{C.-S.}}{\le} \sum_{T\in\mathfrak{T}^h} \|\tilde{\mathbf{e}}\|_{L^2(T)}\,\|\mathbf{R}_1\|_{L^2(T)} \;+\; \sum_{T\in\mathfrak{T}^h}\sum_{E\subset\partial T} \|\tilde{\mathbf{e}}\|_{L^2(E)}\,\|\mathbf{R}_2\|_{L^2(E)} \\ &\overset{(4.13)}{\le} \sum_{T\in\mathfrak{T}^h} C_1\, h_T\,\|\mathbf{R}_1\|_{L^2(T)}\,|||\mathbf{e}|||_{\omega_T} \;+\; \sum_{T\in\mathfrak{T}^h}\sum_{E\subset\partial T} C_2\, h_E^{1/2}\,\|\mathbf{R}_2\|_{L^2(E)}\,|||\mathbf{e}|||_{\omega_E} \\ &\overset{\text{C.-S.}}{\le} \max\{C_1,C_2\}\cdot\left[\sum_{T\in\mathfrak{T}^h} h_T^2\,\|\mathbf{R}_1\|^2_{L^2(T)} \;+\; \sum_{T\in\mathfrak{T}^h}\sum_{E\subset\partial T} h_E\,\|\mathbf{R}_2\|^2_{L^2(E)}\right]^{1/2}\cdot \\ &\qquad\qquad \cdot\left[\sum_{T\in\mathfrak{T}^h} |||\mathbf{e}|||^2_{\omega_T} \;+\; \sum_{T\in\mathfrak{T}^h}\sum_{E\subset\partial T} |||\mathbf{e}|||^2_{\omega_E}\right]^{1/2} \\ &\le \overline{C}\cdot\underbrace{\left[\sum_{T\in\mathfrak{T}^h} h_T^2\,\|\mathbf{R}_1\|^2_{L^2(T)} \;+\; \sum_{T\in\mathfrak{T}^h}\sum_{E\subset\partial T} h_E\,\|\mathbf{R}_2\|^2_{L^2(E)}\right]^{1/2}}_{\eta_R}\cdot |||\mathbf{e}|||\,. \end{aligned} \tag{4.14}$$

Die letzte Ungleichung folgt durch eine Abschätzung der Summe aller Energienormen von $\mathbf{e}$ auf den Gebieten ω_T und ω_E durch eine Konstante C_3 mal der Energienorm auf dem Gesamtgebiet. Damit ist die Konstante $\overline{C} = C_3 \cdot \max\{C_1, C_2\}$. Eine Division durch $|||\mathbf{e}|||$ liefert schließlich den residuen-basierten Fehlerschätzer η_R.

Dies ist zunächst eine globale Fehleraussage; für ein adaptives Verfahren sind jedoch elementbezogene Größen zur Steuerung der Adaptivität unerläßlich. Diese erhält man sehr leicht aus (4.14), indem man die Anteile der einzelnen Elemente in den Summen zusammenfaßt. Mit

$$\eta_{R,T}^2 \;:=\; h_T^2\,\|\mathbf{R}_1\|^2_{L^2(T)} \;+\; \sum_{E\subset\partial T} h_E\,\|\mathbf{R}_2\|^2_{L^2(E)} \tag{4.15}$$

gilt nämlich

$$|||\mathbf{e}||| \;\leq\; \overline{C} \cdot \eta_R \;=\; \overline{C} \cdot \left(\sum_{T \in \mathcal{T}^h} \eta_{R,T}^2 \right)^{1/2} . \tag{4.16}$$

Eine wichtige Eigenschaft eines Fehlerschätzers ist die *Effizienz*. Das bedeutet, daß der Fehler nicht zu stark überschätzt werden darf, damit keine unnötigen Verfeinerungen erfolgen. Die Effizienz wird i. a. durch den Nachweis einer umgekehrten Ungleichung zu (4.16) mit einer Konstanten $\underline{C}$ gezeigt. Insgesamt erhält man damit eine Eingrenzung des wahren Fehlers $\mathbf{e} = \mathbf{u} - \mathbf{u}^h$ in der Energienorm durch den Fehlerschätzer η_R und Konstanten $\underline{C}$ und $\overline{C}$, die nicht von der Netzdichte h abhängen:

$$\underline{C} \cdot \eta_R \;\leq\; |||\mathbf{u} - \mathbf{u}^h||| \;\leq\; \overline{C} \cdot \eta_R \tag{4.17}$$

Weitere Details hierzu können z. B. *Eriksson et. al* [54] oder *Verfürth* [120] entnommen werden.

4.1.2 Fehlerschätzung durch Lösung lokaler Probleme

Das Residuum der FE-Näherungslösung in schwacher Form

$$R(\boldsymbol{\eta}) \;:=\; B(\boldsymbol{\eta}, \mathbf{u} - \mathbf{u}^h) \;=\; F(\boldsymbol{\eta}) - B(\boldsymbol{\eta}, \mathbf{u}^h), \qquad\qquad \boldsymbol{\eta} \in \mathcal{T} \tag{4.18}$$

verschwindet per Definition der Lösung $\mathbf{u}^h$ auf dem Raum $\mathcal{T}^h$ (d. h. für $\boldsymbol{\eta} \in \mathcal{T}^h$), jedoch i. a. nicht auf dem Raum $\mathcal{T}$, da es sich bei $\mathbf{u}^h$ um eine Näherungslösung der schwachen Lösung $\mathbf{u}$ handelt. Durch (4.18) ist ersichtlich, daß das Residuum als ein lineares Funktional auf dem Testraum $\mathcal{T}$ interpretiert werden kann. Dies bedeutet, daß das Residuum ein Element des Dualraums $\mathcal{T}^*$ ist.

Bemerkung: In der Interpretation aus Sicht der Mechanik ist dies ein ganz natürlicher Zusammenhang. Das primale Problem ist in Verschiebungen formuliert, so daß die zugehörigen dualen Größen Spannungen sind. Das Residuum (in starker Form) stellt aber gerade die Spannungen dar, die verbleiben, wenn man die Näherungslösung in die starke Form der Differentialgleichung einsetzt.

Eine erneute Betrachtung der Herleitung des residuen-basierten Fehlerschätzers im letzten Abschnitt macht diesen Zusammenhang deutlich: Durch Abschätzen des Residuums in schwacher Form mit dem Fehler als Testfunktion ergeben sich gerade Ausdrücke, die das Residuum gegenüber der starken Form der Differentialgleichung darstellen. Aufgrund der geringeren Regularität der FE-Näherungslösung treten dabei sowohl das Residuum im Inneren als auch Sprünge über Elementkanten auf. □

Zur Gewinnung eines Fehlerschätzers in der Energienorm kann man also das Residuum in der Norm des Dualraums abschätzen, wobei als spezielle Testfunktion $\boldsymbol{\eta} =: \mathbf{e}$ verwendet wird (man spricht in diesem Zusammenhang auch vom komplementären Variationsproblem, vgl. z. B. *Ainsworth & Craig* [2]). Mit Hilfe dieses *Dualitätsarguments* kann man auf zwei verschiedene Arten zu einem Fehlerschätzer gelangen:

- Man löst lokale *Dirichlet*-Probleme mit Randvorgaben aus der FE-Näherungslösung $\mathbf{u}^h$. Dabei verwendet man Finite-Elemente-Räume höherer Ordnung. Dieser Fehlerschätzer geht auf *Babuška & Rheinboldt* [9] zurück.
- Man löst lokale *Neumann*-Probleme, wobei die rechte Seite durch das Residuum im Inneren und die *Neumann*-Randbedingungen durch die Kantensprünge der Normalableitungen definiert sind. Dabei verwendet man wiederum Finite-Elemente-Räume höherer Ordnung. Dieser Fehlerschätzer geht auf *Bank & Weiser* [16] zurück.

Prinzipiell könnte man obige Probleme auch global formulieren. Der Lösungsaufwand ist dann aber aufgrund der Finite-Elemente-Räume höherer Ordnung wesentlich größer als der des ursprünglichen Problems. *Verfürth* [120] gibt Kriterien an, die für den praktischen Nutzen eines Fehlerschätzers durch Lösung lokaler Zusatzprobleme wichtig sind:

- Um Informationen über das lokale Verhalten des Fehlers zu gewinnen, sollten die Zusatzprobleme nur kleine Teilgebiete von Ω einbeziehen.
- Damit diese Informationen von Nutzen sind, sollten die Zusatzprobleme mit Finite-Elemente-Räumen arbeiten, die genauer sind als der ursprüngliche.
- Um den Rechenaufwand klein zu halten, sollten möglichst wenige Freiheitsgrade verwendet werden.
- Jeder Kante und jedem Element sollte mindestens ein Freiheitsgrad von mindestens einem der Zusatzprobleme zugeordnet sein.

Fehlerschätzung durch Lösung lokaler *Dirichlet*-Probleme

Für einen Knoten $N \in \mathcal{N}^h \backslash \mathcal{N}^h_{\Gamma_\mathrm{u}}$ (Knoten, der nicht an einem *Dirichlet*-Rand liegt) sei ω_N der Patch aller an den Knoten angrenzenden Elemente und $\mathcal{S}^h_N$ ein Finite-Elemente-Raum höherer Ordnung auf ω_N (z. B. können Bubble-Funktionen[3] verwendet werden). Nun löst man auf ω_N ausgehend von der Residuumsgleichung (4.18) das folgende Randwertproblem in schwacher Formulierung:

$$\begin{aligned} B(\boldsymbol{\eta}^h, \mathbf{z}^h_N)\big|_{\omega_N} &= \underbrace{\sum_{T \subset \omega_N} \int_T \boldsymbol{\eta}^h \cdot \mathbf{f} \,\mathrm{d}v + \sum_{E \subset \partial\omega_N \cap \Gamma_\mathrm{t}} \int_E \boldsymbol{\eta}^h \cdot \bar{\mathbf{t}} \,\mathrm{d}a}_{\widehat{=}\; F(\boldsymbol{\eta}^h)\big|_{\omega_N}} \\ &- B(\boldsymbol{\eta}^h, \mathbf{u}^h)\big|_{\omega_N} \qquad \forall\, \boldsymbol{\eta}^h \in \mathcal{S}^h_N. \end{aligned}$$

Die Energienorm der Lösung $\mathbf{z}^h_N$ dient als lokaler Fehlerindikator

$$\eta_{D,N} := |||\mathbf{z}^h_N|||, \tag{4.19}$$

[3] Bubble-Funktionen „leben“ nur im Inneren des Elements und Verschwinden an den Elementrändern. Dadurch wird erreicht, daß keine zusätzliche Kopplung durch die Finite-Elemente-Räume höherer Ordnung entsteht. Bei linearen Ansätzen im Grundproblem verwendet man z. B. quadratische Bubble-Funktionen, bei quadratischen Ansätzen kubische Bubble-Funktionen usw.

und man kann zeigen, daß

$$\eta_D := \left(\sum_{N \in \mathcal{N}^h \setminus \mathcal{N}^h_{\Gamma_u}} \eta^2_{D,N} \right)^{1/2} \tag{4.20}$$

einen globalen Fehlerschätzer darstellt, der bis auf multiplikative Konstanten gleichwertig mit dem residuen-basierten Schätzer η_R ist. Setzt man $\mathbf{u}^h_N := \mathbf{u}^h + \mathbf{z}^h_N$, so entspricht das obige Vorgehen der Lösung des diskreten Analogons zu dem *Dirichlet*-Problem

$$\begin{aligned} -\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}_N) &= \mathbf{f} && \text{in} && \omega_N \\ \mathbf{u}_N &= \mathbf{u}^h && \text{auf} && \partial\omega_N \setminus \Gamma_t \\ \boldsymbol{\sigma}(\mathbf{u}_N)\,\mathbf{n} &= \bar{\mathbf{t}} && \text{auf} && \partial\omega_N \cap \Gamma_t \end{aligned}$$

auf dem Finite-Elemente-Raum $\mathcal{S}^h_N$, und man erhält den Fehlerschätzer durch Vergleich der verbesserten Näherungslösung $\mathbf{u}^h_N$ mit der ursprünglichen Näherungslösung $\mathbf{u}^h$.

Bemerkung: Verwendet man statt der knotenbezogenen Teilgebiete ω_N elementbezogene Teilgebiete ω_T, so gelangt man analog zu einem Fehlerindikator $\eta_{D,T}$ je Element, der wieder durch Summieren zu einem globalen Fehlerschätzer $\tilde{\eta}_D$ führt (*Verfürth* [120]). □

Fehlerschätzung durch Lösung lokaler *Neumann*-Probleme

Bei diesem Schätzer kann man die Elemente T selbst als Gebiete für die Lösung der Zusatzprobleme wählen. Man verwendet also auf T einen Finite-Elemente-Raum $\mathcal{S}^h_T$ höherer Ordnung (z. B. wieder mit den Bubble-Funktionen) und löst das folgende Randwertproblem in schwacher Formulierung (vgl. (4.12)):

$$B(\boldsymbol{\eta}^h, \mathbf{u}^h_T)\big|_T = \int_T \boldsymbol{\eta}^h \cdot \mathbf{R}_1(\mathbf{u}^h)\,\mathrm{d}v + \sum_{E \subset \partial T} \int_E \boldsymbol{\eta}^h \cdot \mathbf{R}_2(\mathbf{u}^h)\,\mathrm{d}a \quad \forall\, \boldsymbol{\eta}^h \in \mathcal{S}^h_T.$$

Die Energienorm der Lösung $\mathbf{u}^h_T$ dient als lokaler Fehlerindikator

$$\eta_{N,T} := |||\mathbf{u}^h_T|||, \tag{4.21}$$

und erneut läßt sich zeigen, daß

$$\eta_N := \left(\sum_{T \in \mathfrak{T}^h} \eta^2_{N,T} \right)^{1/2} \tag{4.22}$$

einen globalen Fehlerschätzer darstellt, der bis auf multiplikative Konstanten gleichwertig mit dem residuen-basierten Schätzer η_R ist. Das obige Vorgehen ist das diskrete Analogon zur Lösung des *Neumann*-Problems

$$\begin{aligned} -\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}_T) &= \mathbf{R}_1(\mathbf{u}^h) && \text{in} && T \\ \mathbf{u}_T &= \mathbf{0} && \text{auf} && \partial T \cap \Gamma_u \\ \boldsymbol{\sigma}(\mathbf{u}_T)\,\mathbf{n} &= \mathbf{R}_2(\mathbf{u}^h) && \text{auf} && \partial T \setminus \Gamma_u \end{aligned}$$

auf dem Finite-Elemente-Raum $\mathcal{S}^h_T$.

4.1.3 Hierarchische Fehlerschätzer

Hierarchische Basissysteme sind schon seit den Anfängen der Entwicklung der Methode der finiten Elemente bekannt. Die Idee, solche Basissysteme zur Berechnung von *A-posteriori*-Fehlerschätzern zu verwenden, stammt von *Bank & Smith* [15]. Dazu wird ein Finite-Elemente-Raum höherer Ordnung $\bar{\mathcal{T}}^h$ als direkte Summe[4] des Finite-Elemente-Raums $\mathcal{T}^h$ und einer hierarchischen Erweiterung $\mathcal{Z}^h$ definiert:

$$\bar{\mathcal{T}}^h = \mathcal{T}^h \oplus \mathcal{Z}^h.$$

Im folgenden werden zwei Beispiele hierarchischer Basen in einer und zwei Raumdimensionen angegeben.

- 1-d: Die linearen Ansatzfunktionen auf dem Referenzelement $\hat{\Omega}_e = [-1, 1]$

  $$\begin{aligned} N_1(\xi) &= \tfrac{1}{2}(1-\xi) \\ N_2(\xi) &= \tfrac{1}{2}(1+\xi) \end{aligned}$$

  bilden zusammen mit der FE-Geometrie-Transformation (vgl. Abschnitt 2.4.5) den Raum $\mathcal{T}^h$, und die dem Mittelpunkt zugeordnete quadratische Ansatzfunktion

  $$N_3(\xi) = \tfrac{1}{4}(1-\xi)(1+\xi)$$

  bildet die hierarchische Erweiterung $\mathcal{Z}^h$. Die lineare Unabhängigkeit ist sofort ersichtlich, und die drei Basisfunktionen N_i spannen denselben Raum auf wie die Basisfunktionen $\tilde{N}_i$ des üblichen quadratischen FE-Ansatzes[5]. Es bestehen die Zusammenhänge:

  $$\begin{aligned} \tilde{N}_1(\xi) &= \tfrac{1}{2}\xi(\xi-1) &&= N_1(\xi) - \tfrac{1}{2}N_3(\xi), \\ \tilde{N}_2(\xi) &= \tfrac{1}{2}\xi(\xi+1) &&= N_2(\xi) - \tfrac{1}{2}N_3(\xi), \\ \tilde{N}_3(\xi) &= &&= N_3(\xi). \end{aligned}$$

  Durch analoge Erweiterung dieser Vorgehensweise erhält man auch für die 2-d-Rechteckelemente und 3-d-Quaderelemente hierarchische Basen. Außerdem können auch hierarchische Basen höherer Ordnung konstruiert werden, etwa die hierarchische Erweiterung quadratischer Ansätze durch kubische Funktionen.

- 2-d: Die linearen Ansatzfunktionen im Dreieck mit den baryzentrischen Koordinaten u, v und $w = 1 - u - v$,

  $$N_1(u,v) = w, \qquad N_2(u,v) = u, \qquad N_3(u,v) = v,$$

  bilden den Raum $\mathcal{T}^h$, und die quadratischen Ansatzfunktionen der Kantenmittelpunkte (Bubble-Funktionen der Kanten),

  $$N_4(u,v) = 4\,w\,u, \qquad N_5(u,v) = 4\,u\,v, \qquad N_6(u,v) = 4\,v\,w,$$

[4] Der Schnitt der beiden Räume besteht nur aus dem Nullvektor, und die Vereinigung liefert den gesamten Raum. Dies bedeutet insbesondere, daß die beiden Räume in der Summe linear unabhängig sind.

[5] Allerdings gehen bei der hierarchischen Version einige Eigenschaften der FE-Ansatzfunktionen verloren, z. B. die Teilung der Eins oder die direkte Interpretierbarkeit der Koeffizienten als Knotenwerte.

bilden die hierarchische Erweiterung $\mathcal{Z}^h$. Wieder kann der übliche quadratische FE-Ansatz $\tilde{N}_i$ durch die linearen Ansatzfunktionen und die hierarchischen Erweiterungsfunktionen dargestellt werden. Es bestehen die Zusammenhänge

$$\begin{aligned}
\tilde{N}_1(u,v) &= (2w-1)\,w &&= N_1(u,v) - \tfrac{1}{2}N_4(u,v) - \tfrac{1}{2}N_6(u,v),\\
\tilde{N}_2(u,v) &= (2u-1)\,u &&= N_2(u,v) - \tfrac{1}{2}N_4(u,v) - \tfrac{1}{2}N_5(u,v),\\
\tilde{N}_3(u,v) &= (2v-1)\,v &&= N_3(u,v) - \tfrac{1}{2}N_5(u,v) - \tfrac{1}{2}N_6(u,v).
\end{aligned}$$

Die Funktionen $\tilde{N}_4$ bis $\tilde{N}_6$ stimmen wie im 1-d Fall mit den entsprechenden Funktionen N_4 bis N_6 überein. □

Das Prinzip zur Gewinnung eines Fehlerschätzers mit Hilfe hierarchischer Basen besteht nun darin, daß man auf dem Finite-Elemente-Raum höherer Ordnung $\bar{\mathcal{T}}^h$ aufgrund der Konvergenz der p-Methode der FEM eine genauere Lösung als auf dem eingebetteten Finite-Elemente-Raum $\mathcal{T}^h \subset \bar{\mathcal{T}}^h$ erwartet. Man könnte also das diskrete Randwertproblem einfach auf dem Raum $\bar{\mathcal{T}}^h$ lösen und die so erhaltene Lösung mit der ursprünglichen Lösung auf dem Raum $\mathcal{T}^h$ vergleichen, hätte dadurch jedoch einen unvertretbar hohen Aufwand.

Da man davon ausgeht, daß sich die durch die Funktionen aus $\mathcal{T}^h$ repräsentierten Lösungsanteile auch auf dem Raum $\bar{\mathcal{T}}^h$ nicht wesentlich verändern, löst man statt dessen ein Problem, das lediglich die Basisfunktionen der hierarchischen Erweiterung $\mathcal{Z}^h$ einbezieht:

$$B(\boldsymbol{\eta}^h, \mathbf{z}^h) \;=\; F(\boldsymbol{\eta}^h) - B(\boldsymbol{\eta}^h, \mathbf{u}^h) \qquad\qquad \forall\, \boldsymbol{\eta}^h \in \mathcal{Z}^h. \tag{4.23}$$

Die Grundlage bildet also wiederum die Residuumsgleichung (4.18). Allerdings ist auch die Lösung des Problems (4.23) mit etwa dem gleichen Aufwand verbunden wie die Lösung des Ausgangsproblems. Daher ersetzt man die linke Seite von (4.23) durch eine zu $B(\cdot,\cdot)$ äquivalente Bilinearform $\tilde{B}(\cdot,\cdot)$ mit

$$c_0 \le \frac{\tilde{B}(\mathbf{z},\mathbf{z})}{B(\mathbf{z},\mathbf{z})} \le c_1, \qquad\qquad \forall\, \mathbf{z} \in \mathcal{Z}^h,\ \mathbf{z} \neq \mathbf{0},\ c_0, c_1 > 0,$$

die zu einem linearen Gleichungssystem in Diagonalgestalt führt, das somit fast ohne Zusatzaufwand lösbar ist. Man löst also das veränderte System

$$\tilde{B}(\boldsymbol{\eta}^h, \tilde{\mathbf{z}}^h) \;=\; F(\boldsymbol{\eta}^h) - B(\boldsymbol{\eta}^h, \mathbf{u}^h) \qquad\qquad \forall\, \boldsymbol{\eta}^h \in \mathcal{Z}^h \tag{4.24}$$

und erhält nach *Bank & Smith* [15] über

$$\eta_H := \sqrt{\tilde{B}(\tilde{\mathbf{z}}^h, \tilde{\mathbf{z}}^h)} \tag{4.25}$$

einen Schätzer für den Fehler $\mathbf{u}-\mathbf{u}^h$ in der Energienorm. Die Aufteilung in elementbezogene Fehlerindikatoren $\eta_{H,T}$ für die Steuerung eines adaptiven Algorithmus ist abhängig von der gewählten hierarchischen Basis. *Verfürth* [120, §1.4] gibt einige Beispiele für hierarchische Basen und zeigt, daß die elementbezogenen Indikatoren im wesentlichen wieder durch das Residuum im Inneren und die Kantensprünge dargestellt werden, beides ausgewertet mit den Testfunktionen der hierarchischen Erweiterung $\mathcal{Z}^h$.

4.1.4 Gradienten-basierte Fehlerindikatoren

Schon für relativ einfache Probleme – etwa dem oben behandelten Problem der linearen Elastostatik – ist die Herleitung von mathematisch fundierten *A-posteriori*-Fehlerschätzern eine komplexe Aufgabe. Man ist daher an alternativen Techniken zur Bereitstellung lokaler Fehlerindikatoren interessiert, die bei solchen Problemen zu ähnlichen Ergebnissen führen wie mathematisch fundierte Fehlerschätzer, die aber außerdem auch auf komplexere Probleme – etwa aus der Plastizitätstheorie – übertragen werden können.

Von *Zienkiewicz & Zhu* [128] wurde 1987 ein solcher Fehlerindikator vorgestellt, der auf einem Vergleich der Spannungen aus der FE-Lösung mit geeignet geglätteten Spannungen beruht. Man spricht daher in der Literatur häufig von Z^2-Fehlerindikatoren (nach den Initialen der Autoren benannt), wenn Gradienten der Lösung zur lokalen Fehlerindikation eingesetzt werden. Die Idee gradienten-basierter Fehlerindikatoren wird im folgenden wieder am Beispiel der linearen Elastostatik erläutert.

Die Berechnung des Fehlers $\mathbf{e} = \mathbf{u} - \mathbf{u}^h$ aus (4.8) setzt die Kenntnis der exakten Lösung $\mathbf{u}$ voraus und ist daher i. a. nicht möglich. Sowohl im Energienorm-Fehler $|||\mathbf{e}|||$ gemäß Gleichung (4.11) als auch im L^2-Norm-Fehler $\|\mathbf{e}_\sigma\|_{L^2(\Omega)}$ gemäß Gleichung (4.10) spielt der Fehler in den Spannungen

$$\mathbf{e}_\sigma = \boldsymbol{\sigma} - \boldsymbol{\sigma}^h \tag{4.26}$$

die entscheidende Rolle. Da die FE-Ansatzfunktionen über Elementgrenzen hinweg nur C^0-stetig sind, erhält man wegen (4.1) sowohl in den FE-Verzerrungen $\boldsymbol{\varepsilon}^h = \boldsymbol{\varepsilon}(\mathbf{u}^h)$ als auch in den FE-Spannungen $\boldsymbol{\sigma}^h = \boldsymbol{\sigma}(\mathbf{u}^h)$ i. a. Sprünge über die Elementkanten. Die Idee besteht nun darin, diese Sprünge geeignet zu glätten und damit eine „verbesserte“ Spannungsnäherung $\boldsymbol{\sigma}^*$ zu berechnen, für die in einer gewählten Norm $\|\cdot\|$ die folgende Annahme gerechtfertigt ist:

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^*\| \ll \|\boldsymbol{\sigma}^* - \boldsymbol{\sigma}^h\| \qquad \text{bzw.} \qquad \|\boldsymbol{\sigma} - \boldsymbol{\sigma}^*\| \le C\,\|\boldsymbol{\sigma}^* - \boldsymbol{\sigma}^h\|, \quad C \ll 1. \tag{4.27}$$

Dies ist insbesondere dann der Fall, wenn $\boldsymbol{\sigma}^*$ *superkonvergent* ist. Dies bedeutet, daß $\boldsymbol{\sigma}^*$ eine Approximation von höherer Ordnung für das wahre Spannungsfeld $\boldsymbol{\sigma}$ darstellt als $\boldsymbol{\sigma}^h$. Gilt also eine Ungleichung der Form (4.27), so kann man mit der Dreiecksungleichung folgern:

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\| \le \|\boldsymbol{\sigma} - \boldsymbol{\sigma}^*\| + \|\boldsymbol{\sigma}^* - \boldsymbol{\sigma}^h\| \le (1 + C)\,\|\boldsymbol{\sigma}^* - \boldsymbol{\sigma}^h\|, \tag{4.28}$$

und erhält durch Vernachlässigung der Konstanten C die folgende Schätzung (keine Abschätzung!) des Fehlers in den Spannungen:

$$\|\mathbf{e}_\sigma\| = \|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\| \approx \|\boldsymbol{\sigma}^* - \boldsymbol{\sigma}^h\| =: \eta_Z. \tag{4.29}$$

Anders ausgedrückt wird die unbekannte exakte Spannung $\boldsymbol{\sigma}$ in (4.26) durch die geglättete Spannung $\boldsymbol{\sigma}^*$ ersetzt.

Die Zuordnung von elementbezogenen Fehlerindikatoren $\eta_{Z,T}$ geschieht wieder auf naheliegende Weise durch den Anteil des Elements T an der Integralnorm $\|\cdot\|$, also z. B. im Fall der L^2-Norm:

$$\eta_{Z,T} := \left(\int_T (\boldsymbol{\sigma}^* - \boldsymbol{\sigma}^h) \cdot (\boldsymbol{\sigma}^* - \boldsymbol{\sigma}^h)\, \mathrm{d}v \right)^{1/2} \quad \Longrightarrow \quad \eta_Z = \left(\sum_{T \in \mathcal{T}^h} \eta_{Z,T}^2 \right)^{1/2}. \tag{4.30}$$

Es bleibt noch die Frage offen, wie man das „verbesserte“ Spannungsfeld $\boldsymbol{\sigma}^*$ berechnet. Die gebräuchlichsten Methoden werden im folgenden beschrieben. Dabei wird jeweils in zwei Schritten vorgegangen. Zunächst werden die Spannungswerte von den Integrationspunkten auf die Knoten projiziert („Knoten-Spannungen“). Anschließend wird das kontinuierliche Spannungsfeld $\boldsymbol{\sigma}^*(\mathbf{x})$ mit Hilfe der Ansatzfunktionen ausgewertet.

- *Globale L^2-Projektion.* Seitdem man die FEM als Berechnungsmethode in den Ingenieurwissenschaften einsetzt, ist die Berechnung von „guten“ Spannungen ein entscheidender Punkt für die Anwendbarkeit (z. B. zur Berechnung von Spannungs-Intensitätsfaktoren und zur Visualisierung und Auswertung der Ergebnisse). Erste Arbeiten zur globalen L^2-Projektion stammen aus den frühen siebziger Jahren, z. B. *Oden & Brauchli* [87], *Hinton & Campbell* [71]. Dabei werden die Spannungen von den Integrationspunkten an die Knoten projiziert, indem das L^2-Fehlerintegral über das gesamte Rechengebiet Ω minimiert wird, so daß das erhaltene Spannungsfeld $\boldsymbol{\sigma}^L$ in diesem Sinn optimal ist. Es entsteht ein lineares Gleichungssystem $\boldsymbol{AX} = \boldsymbol{B}$, wobei die Größe der quadratischen Matrix $\boldsymbol{A}$ durch die Anzahl der FE-Knoten und die Spaltenzahl der rechten Seite $\boldsymbol{B}$ durch die Anzahl der Spannungs-Komponenten bestimmt ist. Der Aufwand entspricht also der Lösung eines FE-Gleichungssystems.

- *Lokale L^2-Projektion und Mittelung.* Statt eines globalen Ansatzes werden hier lokale Ansätze in jedem Element verwendet. Damit erhält man je Knoten soviele Werte wie angrenzende Elemente, so daß eine anschließende Mittelung erforderlich ist. Dieses Verfahren wurde von *Hinton & Campbell* [71] vorgeschlagen und für den Fall quadratischer Ansätze in Vierecken untersucht. Aufgrund seiner Lokalität ist dieses Verfahren wesentlich effizienter als die globale L^2-Projektion, jedoch verliert man die Eigenschaft der globalen Bestapproximation.

  Eine ähnliche Vorgehensweise wird auch von *Zienkiewicz & Zhu* [128] in ihrem ersten Artikel über den Z^2-Fehlerindikator verwendet. Wie bei der lokalen L^2-Projektion werden die Spannungswerte aus den angrenzenden Elementen eines Knotens lokal gemittelt, jedoch mit gewichteten Mittelungsverfahren.

- *Superconvergent Patch Recovery (SPR).* Unter bestimmten Bedingungen besitzt die FE-Lösung lokal (an ausgewählten Punkten innerhalb eines Elements) eine höhere Konvergenzordnung als global (gemessen in der natürlichen Norm). Dieser Effekt wird als *Superkonvergenz* bezeichnet und ist bereits sehr lange bekannt (z. B. *Strang & Fix* [109]). Diese Eigenschaft der FE-Lösung liegt der von *Zienkiewicz & Zhu* [129] vorgeschlagenen Glättungsmethode zugrunde, die unter der englischen Bezeichnung „Superconvergent Patch Recovery“ bekannt ist. Je Knoten wird auf dem „Patch“ aller angrenzenden Elemente eine lokale L^2-Projektion durchgeführt, wobei Polynomansätze abhängig von der Ansatzordnung der Elemente verwendet werden. Dabei werden die FE-Spannungen als Stützwerte an den *Superkonvergenzpunkten* der Elemente ausgewertet. Insgesamt erhält man mit diesem Verfahren ein geglättetes Spannungsfeld $\boldsymbol{\sigma}^*$, das um eine bzw. zwei Ordnungen besser ist als $\boldsymbol{\sigma}^h$. Da der theoretische Nachweis der Superkonvergenz nur in sehr speziellen Fällen gelingt (reguläre Netze, spezielle Anordnung der Elemente), geben die Autoren eine

Reihe numerischer Beispiele an, die auch bei allgemeineren Aufgabenstellungen auf die Superkonvergenz schließen lassen.

Die SPR-Methode hat sich im Zusammenhang mit dem Z^2-Fehlerindikator durchgesetzt, da sie numerisch die besten Ergebnisse liefert. Von *Zienkiewicz & Zhu* [129] wurde numerisch die Konvergenzordnung der obigen Methoden anhand der *Poisson*-Gleichung im Einheitsquadrat ermittelt. Dabei wurden die numerischen Lösungen zu verschiedenen Diskretisierungsparametern h an einem ausgewählten Punkt bzw. in der Energienorm mit der analytischen Lösung verglichen. Einige der so ermittelten Resultate sind in Tabelle 4.1 zusammengefaßt, wobei $\boldsymbol{\sigma}^h$ die FE-Spannung, $\boldsymbol{\sigma}^L$ die mit globaler L^2-Projektion, $\boldsymbol{\sigma}^{HC}$ die mit lokaler L^2-Projektion und $\boldsymbol{\sigma}^*$ die mit SPR berechnete Spannung bezeichnen.

Ansatz	$\boldsymbol{\sigma}^h$	$\boldsymbol{\sigma}^L$	$\boldsymbol{\sigma}^{HC}$	$\boldsymbol{\sigma}^*$
Viereck, bilinear (4 Knoten)	1(1)	2(1.5)	-	2(2)
Viereck, quadratisch (8 Knoten)	2(2)	-	2(2)	4(3)
Viereck, quadratisch (9 Knoten)	2(2)	2(2)	2(2)	4(3)
Dreieck, linear (3 Knoten)	1(1)	2(1.5)	-	2(1.5)
Dreieck, quadratisch (6 Knoten)	2(2)	2(2)	-	4(2.5)

Tabelle 4.1: *Konvergenzordnung nach Zienkiewicz & Zhu [129], punktweise und in der Energienorm (in Klammern)*

Man erkennt, das die SPR-Methode den anderen Verfahren klar überlegen ist, und daß die Superkonvergenz in allen Fällen erreicht wird.

4.1.5 Vergleichende Gegenüberstellung

Den in den vorangegangenen Abschnitten vorgestellten Ansätzen zur Gewinnung von *A-posteriori*-Fehlerschätzern bzw. Fehlerindikatoren liegen häufig ähnliche Ideen zugrunde, so daß sie in vielen Fällen qualitativ zu denselben Ergebnissen führen.

So hat *Verfürth* [120] im Fall der *Poisson*-Gleichung gezeigt, daß die Fehlerschätzer η_R (residuen-basiert), η_D (lokale *Dirichlet*-Probleme), η_N (lokale *Neumann*-Probleme) und η_H (hierarchisch) im folgenden Sinn äquivalent sind: Zu je zwei Fehlerschätzern η_1 und η_2 existieren Konstanten $\underline{C}$ und $\overline{C}$, so daß gilt:

$$\underline{C}\,\eta_1 \leq \eta_2 \leq \overline{C}\,\eta_1\,. \tag{4.31}$$

In einem qualitativen Vergleich verschiedener Fehlerschätzer ist es also ausreichend, einen der o. a. Fehlerschätzer mit dem Z^2-Fehlerindikator zu vergleichen. Von *Babuška & Rodriguez* [10] und *Babuška, Strouboulis et al.* [11, 12, 13] sind Vergleiche von residuen-basierten Fehlerschätzern für verschiedene Differentialgleichungen mit dem Z^2-Fehlerindikator auf

der Basis von numerischen Testrechnungen durchgeführt worden. Dabei wurde insbesondere die Robustheit der Fehlerschätzer in bezug auf Netzverzerrungen, Singularitäten und Randwertvorgaben untersucht. Das wesentliche Ergebnis der Tests ist, daß der Z^2-Fehlerindikator in den meisten Fällen robuster gegenüber den o. a. Einflüssen ist als die residuen-basierten Fehlerschätzer.

Ainsworth et al. [4] und *Ainsworth & Craig* [2] vergleichen verschiedene *A-posteriori*-Fehlerschätzer für FE-Diskretisierungen in der linearen Elastizitätstheorie bzw. von allgemeinen linearen elliptischen Operatoren. Dabei kommen sie zu dem Schluß, daß gradienten-basierte Fehlerindikatoren – wie der Z^2-Fehlerindikator zusammen mit der SPR-Glättungsmethode – unter geeigneten Voraussetzungen äquivalent zu residuen-basierten Fehlerschätzern vom *Babuška-Rheinboldt*-Typ sind. Außerdem wird gezeigt, daß derartige Fehlerindikatoren unter geeigneten Voraussetzungen den wahren Fehler für genügend kleine h eher überschätzen, und daß sie unter weiteren Voraussetzungen sogar asymptotisch exakt sind.

Insgesamt kann man als Ergebnis festhalten, daß gradienten-basierte Fehlerindikatoren trotz geringer theoretischer Fundierung in vielen Fällen numerisch gute bis sehr gute Ergebnisse liefern, die z. T. sogar die theoretisch gut fundierten residuen-basierten Fehlerschätzer übertreffen. Dies ist insbesondere bei adaptiv verfeinerten Netzen mit stark nicht-regulären Eigenschaften der Fall (*Babuška, Strouboulis et al.* [13]), da die theoretischen Aussagen häufig starke Anforderungen an die Netzregularität stellen, die in der Praxis nicht erfüllt sind.

4.1.6 Konstruktion eines neuen Fehlerindikators für die TPM

In diesem Abschnitt soll ein Fehlerindikator für quasi-statische Probleme elastisch-viskoplastischer poröser Medien konstruiert werden. Da für diese Modellklasse bislang keine mathematisch fundierten Abschätzungen des Ortsfehlers bekannt sind, wird wie folgt vorgegangen: Ein Zeitschritt des instationären Problems wird – wie bereits eingangs erwähnt – als ein stationäres Problem mit Anfangsbedingungen aus dem letzten Zeitschritt betrachtet. Zur Beurteilung des Ortsfehlers wird der von *Zienkiewicz & Zhu* [128, 129, 130] vorgeschlagene gradienten-basierte Fehlerindikator derart erweitert, daß alle treibenden Größen des elastisch-viskoplastischen Zweiphasenmodells eingehen.

Im Fall der Elastostatik bewertet der Z^2-Fehlerindikator den Fehler in den Spannungen. Die Spannungen sind im wesentlichen Verschiebungsgradienten und stellen den Flußterm in der zugrundeliegenden Impulsbilanz dar. Es ist daher naheliegend, in der TPM ebenfalls die Flußterme bzw. Gradienten der Primärvariablen der einzelnen Bilanzgleichungen zu betrachten. Diese Überlegung führt auf die folgenden Größen zur Bewertung der Fehler bei elastischen Berechnungen:

- Die Impulsbilanz der Mischung $(1.99)_1$ bestimmt im wesentlichen die Festkörperverschiebung $\mathbf{u}_S$, so daß als Flußgröße die Extraspannung $\mathbf{T}_E^S$ des Festkörpers maßgeblich ist und analog zur Elastostatik in den Fehlerindikator eingeht.[6]

[6] Man könnte an dieser Stelle auch die Gesamtspannung $\mathbf{T}_E^S - p\mathbf{I}$ wählen, erhielte auf diese Weise aber eine gekoppelte Größe aus Festkörper- und Fluid-Eigenschaften.

- Die Volumenbilanz mit der eingesetzten Impulsbilanz des Fluids $(1.99)_2$ bestimmt im wesentlichen den Porenfluiddruck p, so daß der Druckgradient die maßgebliche Größe darstellt. Dieser wird im Fehlerindikator durch die Sickergeschwindigkeit $\mathbf{w}_F$ repräsentiert, die gemäß dem *Darcy*schen Gesetz (1.88) – also der umgeformten quasi-statischen Impulsbilanz des Fluids – proportional zum Druckgradienten ist.

Zur adaptiven Berechnung von elastisch-viskoplastischen Problemen müssen noch geeignete plastische Größen in den Fehlerindikator aufgenommen werden. Die Ergebnisse von *Rannacher & Suttmeier* [99] zeigen im Fall der *Hencky*-Plastizität (Deformationstheorie), daß eine Netzverfeinerung in der Übergangszone zwischen elastischem und plastischem Materialverhalten entscheidend für die Genauigkeit der Gesamtlösung ist. Diese Beobachtung legt die Verwendung der plastischen Verzerrung $\boldsymbol{\varepsilon}_{Sp}$ zur Fehlerindikation nahe, deren Gradienten vor allem an den Rändern der plastischen Zone groß sind.

Bemerkung: Die in Kapitel 5 dargestellten numerischen Ergebnisse rechtfertigen diese Annahme. Insbesondere werden die Ränder von Scherbändern angemessen verfeinert, wogegen das Innere der plastischen Zone erwartungsgemäß relativ grob vernetzt wird. □

Zunächst werden ausgehend von den FE-Größen $(\cdot)^h$ mit Hilfe der in Abschnitt 4.1.4 beschriebenen SPR-Methode geglättete Größen $(\cdot)^*$ berechnet. Mit der L^2-Norm $\|\cdot\|_{L^2(\Omega^h)}$ sowie deren Anteil $\|\cdot\|_{L^2(T)}$ im Element $T \in \mathfrak{T}^h$ werden anschließend die elementbezogenen Fehlerindikatoren $\eta_{i,T}$ definiert:

$$\begin{aligned}
&\text{Festkörper-Elastizität:} && \eta_{1,T} := \|\mathbf{T}_E^{S*} - \mathbf{T}_E^{Sh}\|_{L^2(T)}, && W_1 := \|\mathbf{T}_E^{Sh}\|_{L^2(\Omega^h)},\\
&\text{Festkörper-Plastizität:} && \eta_{2,T} := \|\boldsymbol{\varepsilon}_{Sp}^{*} - \boldsymbol{\varepsilon}_{Sp}^{h}\|_{L^2(T)}, && W_2 := \|\boldsymbol{\varepsilon}_{Sp}^{h}\|_{L^2(\Omega^h)},\\
&\text{Porenfluid-Strömung:} && \eta_{3,T} := \|\mathbf{w}_F^{*} - \mathbf{w}_F^{h}\|_{L^2(T)}, && W_3 := \|\mathbf{w}_F^{h}\|_{L^2(\Omega^h)}.
\end{aligned} \tag{4.32}$$

Man beachte, daß es sich bei den $\eta_{i,T}$ um absolute Größen handelt, so daß ein direkter Vergleich nicht sinnvoll ist. So liegen z. B. Spannungen und plastische Verzerrungen i. a. in völlig verschiedenen Größenordnungen. Für die Anwendung sind außerdem relative Fehlerindikatoren vorteilhaft, und es wird je Element genau ein skalarer Fehlerindikator benötigt. Man muß also eine geeignete Kombination der drei Fehlerindikatoren $\eta_{i,T}$ berechnen, wobei die Gebietsintegrale W_i als Bezugsgrößen und zur Entdimensionierung dienen.

In Analogie zur Zeitadaptivität (vgl. Abschnitt 3.4) werden toleranz-gewichtete Fehlermaße („Fehler / Toleranz“) verwendet, die mit vorgegebenen absoluten Toleranzen $\epsilon_{a,i} > 0$ und der relativen Toleranz $\epsilon_r \geq 0$ wie folgt definiert werden:

$$\xi_{i,T} := \frac{\eta_{i,T}}{TOL_i} \quad \text{mit} \quad TOL_i := \epsilon_r \cdot W_i + \epsilon_{a,i}\,, \qquad i = 1, \ldots, 3. \tag{4.33}$$

Die mit benutzerdefinierbaren Faktoren $\alpha_i \geq 0$ gewichtete Summe der dimensionslosen Fehlermaße $\xi_{i,T}$ liefert schließlich elementbezogene Fehlermaße ξ_T sowie ein globales skalares Fehlermaß ξ:

$$\xi_T := \left(\sum_{i=1}^{3} \alpha_i \cdot \xi_{i,T}^2\right)^{1/2} \text{ mit } \sum_{i=1}^{3} \alpha_i = 1, \qquad \xi := \left(\sum_{T \in \mathfrak{T}^h} \xi_T^2\right)^{1/2}. \tag{4.34}$$

Gilt $\xi \leq 1$, so ist der durch die absoluten Fehlermaße $\eta_{i,T}$ geschätzte Fehler (zumindest im gewichteten Mittel) kleiner als die geforderten Toleranzen, so daß die Lösung auf dem aktuellen Netz akzeptiert werden kann. Die Faktoren α_i können dabei vom Benutzer entsprechend der Zielvorstellung bei der adaptiven Berechnung frei gewählt werden. So ist etwa bei der Berechnung rein elastischer Probleme die Ausblendung des Fehlerindikators $\xi_{2,T}$ mittels $\alpha_1 = \alpha_3 = 1/2$, $\alpha_2 = 0$ sinnvoll.

4.2 Strategien der Netzanpassung

Auf der Basis von gegebenen Fehlerindikatoren müssen im Rahmen eines ortsadaptiven Verfahrens geeignete Strategien zum Auffinden eines „optimalen“ Netzes bereitgestellt werden. Nach der Begriffsklärung in Abschnitt 4.2.1 werden in Abschnitt 4.2.2 Strategien zur Umsetzung der lokalen Fehlerinformation in lokale Netzdichtefunktionen vorgestellt. Dabei werden jeweils toleranz-gewichtete Fehlermaße

$$\xi_T := \frac{\eta_T}{TOL} \tag{4.35}$$

betrachtet (vgl. (4.33), (4.34)), die aus den absoluten Fehlermaßen η_T (vgl. (4.15), (4.21), (4.25), (4.30), (4.32)) durch Gewichtung mit einer gegebenen Toleranz TOL entstehen, die aus relativen und absoluten Anteilen bestehen kann (vgl. (4.33)). Das Gesamtfehlermaß ergibt sich dann aus (4.35) durch Summation über alle Elemente T des Netzes $\mathfrak{T}^h$:

$$\xi = \left(\sum_{T \in \mathfrak{T}^h} \xi_T^2 \right)^{1/2} . \tag{4.36}$$

Die Anzahl der Elemente im Netz $\mathfrak{T}^h$ wird wie in Kapitel 2 mit E bezeichnet.

4.2.1 Optimalitätskriterien

Die Bestimmung eines optimalen Netzes im Rahmen eines adaptiven Verfahrens ist i. a. ein komplexes nichtlineares Optimierungsproblem. Der Begriff des „optimalen Netzes“ kann dabei wie folgt verstanden werden:

(O1) Das Netz ist optimal, wenn mit geringstmöglicher Anzahl von Freiheitsgraden (oder Elementen) der geschätzte Fehler kleiner ist als die vorgegebenen Toleranzen.

(O2) Das Netz ist optimal, wenn bei gegebener Maximalzahl von Freiheitsgraden (oder Elementen) eine Lösung mit geringstmöglichem geschätztem Fehler erzielt wird.

Die beiden Kriterien sind im folgenden Sinn dual zueinander: beim Kriterium (O1) wird der Fehler vorgegeben und der numerische Aufwand minimiert; beim Kriterium (O2) wird der numerische Aufwand vorgegeben und der Fehler minimiert.

Zur Lösung des nichtlinearen Optimierungsproblems (O1) bzw. (O2) wird üblicherweise ein Iterationsalgorithmus wie in Kasten (4.37) eingesetzt. Da die Anzahl der Knoten oder

Iterationsalgorithmus zur Bestimmung eines optimalen Netzes

Schritt 1: Löse das diskrete Problem auf dem gegebenen Netz $\mathfrak{T}^h$.

Schritt 2: Werte den Fehlerindikator (4.36) aus.

Schritt 3: Falls $\xi > 1$ (O1) bzw. falls $\mathrm{E} < \mathrm{E}_{\max}$ (O2), dann
- bestimme ein geeignetes neues Netz $\hat{\mathfrak{T}}^h$,
- setze $\mathfrak{T}^h := \hat{\mathfrak{T}}^h$,
- gehe zu Schritt 1.

Schritt 4: Ende.

(4.37)

der Freiheitsgrade aufgrund der Abhängigkeit von der Topologie des Netzes nicht direkt zugänglich ist, wird im Fall des Kriteriums (O2) die Anzahl E der Elemente durch $\mathrm{E}_{\max}$ beschränkt. Die Größen des neu zu bestimmenden Netzes $\hat{\mathfrak{T}}^h$ werden im weiteren mit einem Dach gekennzeichnet.

Bemerkung: Neben (O1) und (O2) bei der Abfrage in Schritt 3 sind auch Mischformen möglich. Man kann z. B. Toleranzen vorgeben, jedoch die maximale Elementzahl auf $\mathrm{E}_{\max}$ beschränken. Der Algorithmus ist beendet, wenn entweder die Toleranz erfüllt ist oder die Elementzahl $\mathrm{E}_{\max}$ überschritten wird. Dies entspricht dann zwar keinem der beiden Optimalitätskriterien mehr, ist aber in der Praxis aufgrund von Speicher- und Rechenzeitbeschränkungen ein sinnvolles Vorgehen. □

4.2.2 Dichtefunktionen

Eine *Netzdichtefunktion* dient bei der Bestimmung des neuen Netzes in Schritt 3 von Algorithmus (4.37) als Vorgabe für einen Netzgenerator (Wiedervernetzung) bzw. zur Steuerung der hierarchischen Netzanpassung. Für Elemente des neuen Netzes $\hat{\mathfrak{T}}^h$, die innerhalb des Elements $T \in \mathfrak{T}^h$ des alten Netzes liegen, wird der Elementradius $\hat{h}_T$ wie folgt bestimmt:

$$\hat{h}_T = h_T \; r_T(\xi)\,. \tag{4.38}$$

Die Dichtefunktion r_T hängt dabei vom Fehlerindikator ξ bzw. von den elementbezogenen Indikatoren ξ_T ab. Ein Wert $r_T < 1$ bedeutet eine Verfeinerung des Elements T, während im Fall $r_T > 1$ eine Vergröberung möglich ist. Im Rahmen einer vom Verfasser der vorliegenden Arbeit betreuten Diplomarbeit (*Ammann* [6]) wurde eine vergleichende Untersuchung verschiedener in der Literatur aufgeführter Dichtefunktionen zur Approximation des Optimierungsproblems (O1) durchgeführt. Weitere Details der im folgenden vorgestellten Ansätze können dieser Arbeit sowie den folgenden darin betrachteten Artikeln entnommen werden: *Zienkiewicz & Zhu* [128], *Ladevèze, Pelle & Rougeot* [78], *Oñate & Bugeda* [86], *Li & Bettess* [79], *Gallimard, Ladevèze & Pelle* [59, 60], *Fuenmayor & Oliver* [58].

Gleichverteilung des Fehlers auf Basis des alten Netzes

Es ist naheliegend, zur Ermittlung der Dichtefunktion die Gleichverteilung der Elementfehler η_T zu fordern. Da die Anzahl der Elemente $\hat{\mathrm{E}}$ im neuen Netz von der Wahl der Dichtefunktion abhängt, wird von *Zienkiewicz & Zhu* [128] die Forderung

$$TOL \stackrel{!}{=} \left(\sum_{T \in \mathfrak{T}^h} \hat{\eta}_T^2 \right)^{1/2} = \sqrt{\mathrm{E}}\, \hat{\eta}_T \quad \Longrightarrow \quad \hat{\eta}_T \stackrel{!}{=} \frac{TOL}{\sqrt{\mathrm{E}}} \tag{4.39}$$

an den Elementfehler $\hat{\eta}_T$ von Elementen im neuen Netz gestellt, die innerhalb des Elements T aus dem alten Netz liegen. Um dies explizit in Abhängigkeit von Größen des alten Netzes darstellen zu können, ist eine weitere Annahme über das Fehlerverhalten in Abhängigkeit von der Netzdichte erforderlich.

Aufgrund von theoretischen Konvergenzuntersuchungen bei Anwendung der FEM auf elliptische Randwertprobleme (z. B. *Strang & Fix* [109]) erhält man üblicherweise asymptotische *A-priori*-Fehlerabschätzungen

$$|||\mathbf{u} - \mathbf{u}^h||| \leq C\, h^p \quad \text{mit} \quad h = \max_{T \in \mathfrak{T}^h} h_T\,, \quad h \to 0\,. \tag{4.40}$$

Dabei ist C eine von der Diskretisierung h unabhängige Konstante und p die asymptotische Konvergenzordnung. Zur Herleitung der Dichtefunktion wird diese Aussage auf den lokalen Fehlerindikator innerhalb eines Elements T bezogen, so daß man näherungsweise die folgenden Zusammenhänge im alten Netz $\mathfrak{T}^h$ und im neuen Netz $\hat{\mathfrak{T}}^h$ erhält:

$$\eta_T \approx C\, h_T^p \quad \text{und} \quad \hat{\eta}_T \approx C\, \hat{h}_T^p\,. \tag{4.41}$$

Dies führt dann mit (4.35) zu der Dichtefunktion

$$r_T = \frac{\hat{h}_T}{h_T} = \left[\frac{\hat{\eta}_T}{\eta_T}\right]^{1/p} = \left[\frac{TOL}{\eta_T \sqrt{\mathrm{E}}}\right]^{1/p} = \xi_T^{-1/p}\, \mathrm{E}^{-1/(2p)}\,. \tag{4.42}$$

Minimierung der Elementanzahl im neuen Netz

Ein Schritt zur Lösung des Problems (O1) kann als diskretes Optimierungsproblem

$$\hat{\mathrm{E}} \to \text{Min.} \quad \text{mit} \quad \hat{\xi} = 1 \tag{4.43}$$

aufgefaßt werden (*Ladevèze, Pelle & Rougeot* [78], *Gallimard, Ladevèze & Pelle* [59, 60]), d. h. die Anzahl der Elemente im neuen Netz wird unter der Nebenbedingung der Einhaltung der Toleranzen minimiert. Da $\hat{\mathrm{E}}$ und $\hat{\xi}$ von der Wahl des neuen Netzes abhängen, liegt ein implizites Problem vor, das nur mittels zusätzlicher Annahmen explizit lösbar ist.

In zwei Raumdimensionen kann $\hat{\mathrm{E}}$ mit Hilfe der Elementradien wie folgt abgeschätzt werden:

$$\hat{\mathrm{E}} \approx \sum_{T \in \mathfrak{T}^h} \frac{h_T^2}{\hat{h}_T^2} = \sum_{T \in \mathfrak{T}^h} \frac{1}{r_T^2}\,. \tag{4.44}$$

Wird ein Element verfeinert ($h_T^2 > \hat{h}_T^2$), so liefert das Flächenverhältnis in (4.44) eine Zahl größer als Eins, die die Anzahl der Elemente im neuen Netz innerhalb des Elements T aus dem alten Netz abschätzt. Umgekehrtes gilt bei einer Vergröberung.

Setzt man zur Abschätzung von $\hat{\xi}$ wieder (4.41) für die FEM-Konvergenzrate voraus, so erhält man mit (4.35) und (4.36):

$$\hat{\xi}^2 \;=\; \sum_{T\in\mathfrak{T}^h} \hat{\xi}_T^2 \;=\; \sum_{T\in\mathfrak{T}^h} r_T^{2p}\,\xi_T^2\,. \tag{4.45}$$

Löst man damit das diskrete Optimierungsproblem (4.43) nach der Methode der *Lagrange*-Multiplikatoren, so ergibt sich die Dichtefunktion

$$r_T \;=\; \xi_T^{-1/(p+1)} \left[\sum_{\tau\in\mathfrak{T}^h} \xi_\tau^{2/(p+1)}\right]^{-1/(2p)}\,. \tag{4.46}$$

Man erhält also eine zu (4.42) sehr ähnliche Struktur mit einem vom Elementindikator ξ_T abhängigen Term und einem Zusatzterm, der hier im Gegensatz zu (4.42) nicht durch die Anzahl der Elemente E im *alten* Netz, sondern durch die Anzahl der Elemente $\hat{\mathrm{E}}$ im *neuen* Netz bestimmt ist (vgl. auch (4.49)).

Bemerkung: Von *Li & Bettess* [79] wird eine Dichtefunktion vorgeschlagen, die dieselbe Formel wie (4.46) liefert, jedoch auf völlig andere Weise hergeleitet wird (*Ammann* [6]). Zunächst wird im Unterschied zu (4.39) die Gleichverteilung des Fehlers im *neuen* Netz gefordert:

$$\hat{\eta}_{\hat{T}} \;=\; \frac{TOL}{\sqrt{\hat{\mathrm{E}}}} \qquad \forall\,\hat{T}\in\hat{\mathfrak{T}}^h\,. \tag{4.47}$$

Die unbekannte Elementzahl $\hat{\mathrm{E}}$ wird dann wieder gemäß (4.44) abgeschätzt. Anstelle von (4.41) setzen *Li & Bettess* voraus, daß die lokale Konvergenzordnung um Eins höher ist als die globale Konvergenzordnung (4.40):

$$\eta_T \approx C\,h_T^{p+1} \quad \text{und} \quad \hat{\eta}_T \approx C\,\hat{h}_T^{p+1}\,. \tag{4.48}$$

Dies begründen sie damit, daß die Fläche eines Elements in 2-d von der Ordnung h_T^2 ist. Damit kann schließlich die unbekannte Elementzahl $\hat{\mathrm{E}}$ bestimmt werden, und Einsetzen in die aus (4.47), (4.48) hergeleitete Formel

$$r_T \;=\; \frac{\hat{h}_T}{h_T} \;=\; \left[\frac{TOL}{\eta_T\sqrt{\hat{\mathrm{E}}}}\right]^{1/(p+1)} \;=\; \xi_T^{-1/(p+1)}\;\hat{\mathrm{E}}^{-1/(2(p+1))} \tag{4.49}$$

liefert die Dichtefunktion (4.46). □

Von *Ammann* [6] wurde gezeigt, daß die Dichtefunktion (4.46) in bezug auf die Anzahl der Elemente wesentlich bessere Netze liefert als (4.42).

„Fixed fraction“ Strategie

Die beiden obigen Dichtefunktionen sind auf das Optimierungsproblem (O1) zugeschnitten, d. h. es wird versucht, mit möglichst wenigen Elementen die geforderte Toleranz einzuhalten. Im Gegensatz dazu versucht die „fixed fraction“ Strategie (*Suttmeier* [111], *Verfürth* [120]), das Optimierungsziel (O2) zu erreichen. Zunächst werden die Elemente aufsteigend nach η_T sortiert. Anschließend wird ein fester Anteil $\gamma_C \in (0,1)$, z. B. $\gamma_C = 0.2$, von Elementen mit geringen Fehlern zur Vergröberung markiert. Dann wird ein fester Anteil $\gamma_R \in (0,1)$ von Elementen mit großen Fehlern zur Verfeinerung markiert, wobei die maximale Elementzahl $\mathrm{E}_{\max}$ nicht überschritten werden darf. Die Verfeinerungs-Iteration in (4.37) wird beendet, wenn entweder die Maximalzahl $\mathrm{E}_{\max}$ der Elemente erreicht oder der Fehler ξ_T bereits näherungsweise gleichverteilt ist.

Die „fixed fraction“ Strategie ist zunächst nicht im Fall der Wiedervernetzung einsetzbar, da keine explizite Dichtefunktion berechnet wird. Sie liefert lediglich die Information, ob verfeinert oder vergröbert werden soll, jedoch nicht in welchem Maße. Naheliegend ist daher die folgende Definition einer Dichtefunktion:

$$r_T := \begin{cases} 2 & \text{falls } T \text{ zur Vergröberung markiert ist,} \\ 1/2 & \text{falls } T \text{ zur Verfeinerung markiert ist,} \\ 1 & \text{sonst.} \end{cases} \tag{4.50}$$

4.3 Netzgenerierung und Datenhaltung

Das Kernstück eines adaptiven FE-Systems bildet die Verwaltung der Geometriedaten, die im Laufe einer adaptiven Berechnung anfallen. Dabei lassen sich grundsätzlich zwei verschiedene Vorgehensweisen unterscheiden.

- Bei der *hierarchischen Netzanpassung* wird das aktuelle Netz in jedem Adaptionsschritt lokal entsprechend den Vorgaben der Dichtefunktion verändert. Dadurch entsteht ausgehend von einem Startnetz eine Netzhierarchie, die im Laufe der Rechnung neben lokalen Verfeinerungen auch lokale Vergröberungen erlaubt (Abschnitt 4.3.1).

- Bei der Methode der *Wiedervernetzung* (engl.: remeshing) wird in jedem Adaptionsschritt auf der Basis der Dichtefunktion ein komplett neues Netz generiert (Abschnitt 4.3.2).

In beiden Fällen muß zunächst eine Beschreibung der Geometrie des Randwertproblems sowie der dazugehörigen Randbedingungen vorhanden sein. In dem im Rahmen der vorliegenden Arbeit entwickelten Programmsystem PANDAS ist eine Geometriebeschreibung auf Basis von Makrokanten implementiert, die neben polygonalen Rändern auch die Vorgabe von gekrümmten Randstücken zuläßt, die mit Hilfe von Splinekurven und Kreissegmenten modelliert werden. Der Vorteil der auf diese Weise definierten gekrümmten Ränder gegenüber einer rein polygonalen Randapproximation besteht darin, daß ein vergleichsweise grobes Startnetz gewählt werden kann, ohne auf eine gute Approximation der wahren Randgeometrie im Laufe der adaptiven Berechnung zu verzichten.

Die Randbedingungen werden in PANDAS je Freiheitsgrad des betrachteten Modells als orts- und zeitabhängige Funktionen auf den Makrokanten definiert. Dadurch ist sichergestellt, daß die vorgegebenen Randbedingungen zu jedem Zeitpunkt einer adaptiven Rechnung ausgewertet werden können.

Im Rahmen dieser Arbeit werden lediglich Dreiecksnetze betrachtet. Im Fall der hierarchischen Netzanpassung hat dies den Vorteil, daß ohne großen Zusatzaufwand konforme adaptive Verfeinerungen erzielt werden können, was bei Viereckснetzen nicht ohne weiteres möglich ist. Dort muß das Problem der „hängenden Knoten“[7] entweder durch Zusatzgleichungen im Rahmen der FE-Lösung oder durch einen konformen Abschluß des Netzes mit Dreiecken behandelt werden. In bezug auf die Wiedervernetzung ist die Beschränkung auf Dreiecksnetze jedoch rein technischer Natur, da kein Vierecks-Netzgenerator zur Verfügung stand, der über eine Dichtefunktion gesteuert werden kann.

Als Dreiecks-Vernetzer wird Triangle von *Shewchuk* [105] eingesetzt, der im C-Quellcode über die Numerik-Bibliothek *Netlib* im Internet verfügbar ist. Triangle ist ein *Delaunay*-Triangulierer, der Qualitäts-Nebenbedingungen wie Minimalwinkel oder Maximalfläche berücksichtigen kann. Dieser Vernetzer wurde derart angepaßt, daß die Wiedervernetzung durch Dichtefunktionen gesteuert werden kann, die innerhalb von PANDAS gemäß Abschnitt 4.2.2 bezogen auf ein bestehendes Netz berechnet werden.

4.3.1 Hierarchische Netzverfeinerung und -vergröberung

Eine aktuelle Zusammenfassung von zwei- und dreidimensionalen Verfeinerungsalgorithmen für Simplizes (Dreiecke und Tetraeder) kann z. B. *Bey* [19] entnommen werden. Zur hierarchischen Verfeinerung von Dreiecks-Netzen existieren im wesentlichen zwei Klassen von Algorithmen:

- *Rot-Grün-Verfeinerung*. Durch Kombination von *regulären* (Rot-) und *irregulären* (Grün-) Verfeinerungen wird ein Algorithmus konstruiert, der die Stabilität der Verfeinerung sicherstellt (keine Entartung der Elemente). Die Rot-Verfeinerung wird dabei auf die vom Fehlerschätzer markierten Elemente angewandt, während mit der Grün-Verfeinerung die Konformität des Netzes wiederhergestellt wird. Dieser Verfeinerungs-Algorithmus stammt von *R. E. Bank* und ist u. a. in seinem Buch über das FE-Mehrgitterverfahren PLTMG [14] dargestellt.

- *Bisektionsverfahren*. Bei diesen Verfahren werden ausschließlich Bisektionen (Halbierungen von Dreiecken) zur Verfeinerung verwendet. Für zwei Raumdimensionen werden im folgenden die beiden wichtigsten Verfahren aufgeführt.

  Longest Side Bisection. Beim Verfahren von *Rivara* [100, 101] wird grundsätzlich die längste Kante eines Dreiecks halbiert. Die Anzahl der Ähnlichkeitsklassen, die

[7] Als „hängende Knoten“ werden Knoten auf Elementkanten bezeichnet, die durch Verfeinerung des Nachbarelements entstehen und im betrachteten Element keine konforme Entsprechung haben. Für die Freiheitsgrade dieser Knoten müssen innerhalb der FEM zusätzliche Gleichungen eingeführt werden, die den Wert je nach den gewählten Ansatzfunktionen als geeignete Linearkombination von anderen Freiheitsgraden des Elements bestimmen.

dadurch aus jedem Ausgangsdreieck T entstehen können, ist in jedem Fall endlich und das Verfahren folglich stabil.

Newest Vertex Bisection. Beim Verfahren von *Mitchell* [84] wird in der Ausgangstriangulierung je Dreieck eine Kante markiert, die als erstes zur Verfeinerung ansteht. Im Prinzip kann hier eine beliebige Kante gewählt werden, wobei in der Praxis etwa die längste Kante eine sinnvolle Wahl darstellt. Anschließend wird immer diejenige Kante eines Dreiecks halbiert, die dem zuletzt erzeugten Eckpunkt (engl.: newest vertex) gegenüberliegt, wobei diese Bedingung in beiden Nachbarn der zu teilenden Kante erfüllt sein muß. Dabei entstehen je Ausgangsdreieck T höchstens vier Ähnlichkeitsklassen (Abbildung 4.1), so daß die Verfeinerung stabil ist.

Im Rahmen dieser Arbeit wurde aus den folgenden Gründen das Verfahren von *Mitchell* mit Hilfe des von *Kossaczký* [76] angegebenen rekursiven Algorithmus implementiert:

- Der Algorithmus ist vergleichsweise einfach zu implementieren, da im Gegensatz zur Rot-Grün-Verfeinerung keine Abschlußregeln notwendig sind.
- Durch Speicherung der Netzhierarchie ist auf vergleichsweise einfache Art und Weise eine Vergröberung möglich. Dazu muß lediglich sichergestellt sein, daß das „Bruderdreieck“ – also das Element, das aus dem gleichen „Vaterdreieck“ durch Bisektion entstanden ist – ebenfalls zur Vergröberung markiert ist. In diesem Fall werden beide Bruderdreiecke aus der Triangulierung entfernt und durch das Vaterdreieck ersetzt. Der Vergröberungsalgorithmus bei einer Rot-Grün-Triangulierung ist im Gegensatz dazu wesentlich aufwendiger, da zum einen die Abschlußregeln invers behandelt werden müssen und sich zum anderen die Vergröberung weniger lokal auswirkt.
- Das Verfahren läßt sich prinzipiell auf den dreidimensionalen Fall übertragen (*Bey* [19]), was im Hinblick auf die spätere Erweiterung der vorgestellten Adaptivitätstechniken zur Berechnung räumlich dreidimensionaler Randwertprobleme von Vorteil ist.

In Abbildung 4.1 sind für ein Dreieck die bei der Verfeinerung entstehenden vier Ähnlichkeitsklassen dargestellt. Dabei ist jeweils der „neueste Eckpunkt“ mit einem schwarzen Kreis markiert.

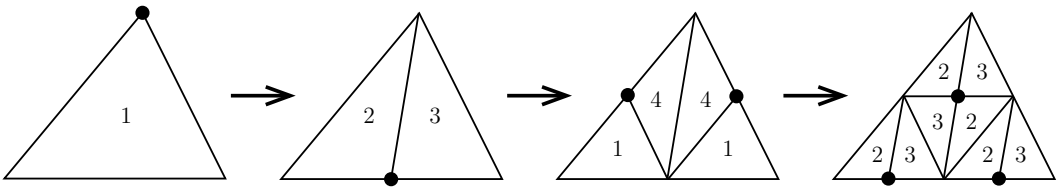


Abbildung 4.1: *Ähnlichkeitsklassen bei der Newest Vertex Bisection*

Abbildung 4.2 zeigt beispielhaft die auftretende Rekursion bei Verfeinerung des grau schraffierten Dreiecks. Da im Nachbardreieck nicht die gleiche Kante zur Verfeinerung markiert ist (Pfeilmarkierungen am gegenüberliegenden Eckpunkt), wird rekursiv zuerst

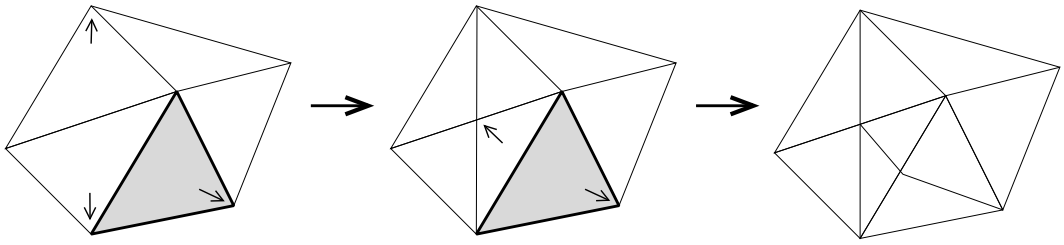


Abbildung 4.2: *Rekursive Bisektion mit der Newest Vertex Bisection*

das Nachbardreieck und anschließend das grau schraffierte Dreieck zusammen mit dem neu entstandenen Nachbardreieck halbiert. Von *Mitchell* [84] wird gezeigt, daß diese Rekursion endlich ist. Dies liegt im wesentlichen darin begründet, daß im ungünstigsten Fall nach endlich vielen Rekursionsschritten der Rand des Gebiets erreicht wird.

Die Möglichkeit der Vergröberung ist insbesondere bei zeitabhängigen adaptiven Rechnungen wichtig, wenn sich die örtlichen Gradienten der Lösung zeitlich verlagern, wie z. B. bei entstehenden Scherbändern, oder sich nach und nach abbauen, wie z. B. beim Konsolidationsproblem.

4.3.2 Wiedervernetzung

Im Gegensatz zur hierarchischen Verfeinerung und Vergröberung wird im Fall der Wiedervernetzung je Adaptionsschritt ein völlig neues Netz generiert. Die Netzdichte wird dem Netzgenerator dabei in Form von Soll-Volumina (in zwei Raumdimensionen: Soll-Flächeninhalten) der Elemente im neuen Netz bezogen auf das vorangegangene Netz übergeben. Es ergibt sich also eine stückweise konstante Dichtefunktion, die ggf. innerhalb des Netzgenerators durch einfache oder flächenbezogene Knotenmittelung geglättet werden kann. Nach der Generierung des neuen Netzes wird die Rechnung mit diesem Netz wiederholt bzw. fortgesetzt.

4.4 Behandlung zeitabhängiger Probleme

Normalerweise wird bei zeitabhängigen Problemen im Zusammenhang mit FE-Ortsdiskretisierungen die Linienmethode (engl.: method of lines, MOL) angewandt, bei der eine feste Ortsdiskretisierung mit Hilfe eines Zeitintegrationsverfahrens von Zeitschicht zu Zeitschicht numerisch integriert wird (Kapitel 3). Zur Kopplung mit ortsadaptiven Verfahren ergeben sich damit zwei grundsätzlich verschiedene Ansätze.

1. *Globale Fehlerbetrachtung*. Die Rechnung wird mit einem festen Netz bis zur geforderten Zielzeit im Sinne der Linienmethode durchgeführt. Im Nachhinein wird anhand von gesammelten Fehlerinformationen ein neues Netz generiert, das für den gesamten Zeitbereich „geeignet“ ist. Die Rechnung wird mit dem neuen Netz komplett wiederholt, und dieser Vorgang wird solange iteriert, bis eine geforderte Toleranz für den Gesamtfehler unterschritten ist.

Ein solches Vorgehen wird z. B. von *Ladevèze* und Mitarbeitern [78, 59, 60] zur Behandlung von elasto-plastischen Problemen vorgeschlagen. Allerdings wird dort der Gesamtfehler mit Hilfe von heuristischen Fehlerindikatoren beurteilt, so daß keine wirkliche Fehlerkontrolle im Sinne einer Abschätzung nach oben vorliegt.

2. *Lokale Fehlerbetrachtung*. Ein Zeitschritt wird als stationäres Teilproblem angesehen, und der Gesamtfehler aus Zeit- und Ortsfehler wird getrennt betrachtet. Mit den in Kapitel 3 behandelten zeitadaptiven Verfahren wird der (lokale) Zeitfehler kontrolliert, während der Ortsfehler im stationären Teilproblem eines Zeitschritts mit den oben beschriebenen Fehlerschätzern und Fehlerindikatoren bewertet wird. Nach der Netzanpassung werden die diskreten Zustandsgrößen (Freiheitsgrade und interne Variablen) vom alten auf das neue Netz übertragen, und die Rechnung wird mit dem neuen Netz fortgesetzt.

Generell wäre es wünschenswert, den globalen Gesamtfehler des numerischen Verfahrens unterhalb einer gegegebenen Toleranz zu halten (Strategie 1). Zum einen setzt dies aber die Kenntnis eines globalen Fehlerschätzers im Orts-Zeit-Kontinuum voraus, der i. a. nicht zur Verfügung steht. Zum anderen ist der Rechenaufwand bei einem solchen Vorgehen extrem hoch, da die gesamte Rechnung bei jeder Netzänderung wiederholt werden muß.

Im folgenden wird daher die zweite Strategie eingesetzt, bei der man zudem durch die gedankliche Entkopplung[8] der Fehleranteile auf bewährte Fehlerkriterien im Zeit- wie im Ortsbereich zurückgreifen kann. Nach der Beschreibung des adaptiven Gesamtalgorithmus in Abschnitt 4.4.1 behandelt Abschnitt 4.4.2 einige Aspekte des notwendigen Datentransfers der diskreten Zustandsgrößen.

4.4.1 Zeit- und ortsadaptiver Gesamtalgorithmus

In Abbildung 4.3 ist der Algorithmus eines zeit- und ortsadaptiven Schritts als Flußdiagramm dargestellt. Die linke Prozedur (`FullStep`) zeigt die Berechnung eines gekoppelten zeit- und ortsadaptiven Schritts, während die rechte Prozedur (`TimeStep`) die Berechnung eines adaptiven Zeitschritts gemäß Kasten (3.60) in Kurzform zusammenfaßt.

Innerhalb eines zeit- und ortsadaptiven Schritts (Prozedur `FullStep`) wird zunächst das vorhandene Netz vergröbert, falls dies aufgrund der Fehlerindikatoren aus dem letzten Zeitschritt möglich ist. Dann wird im wesentlichen der Algorithmus aus Kasten (4.37) zur Bestimmung eines optimalen Netzes mit dem Kriterium (O1) ausgeführt (Schleife). Der Lösung des diskreten Problems in (4.37) entspricht hier die Berechnung eines adaptiven Zeitschritts $t_n \rightsquigarrow t_{n+1}$ (Prozedur `TimeStep`). Nachdem der Zeitfehler mit dem aktuellen Netz die gegebenen Toleranzen erfüllt (erfolgreiche Rückkehr aus der Prozedur `TimeStep`), wird der Ortsfehler mit dem Fehlermaß ξ gemäß (4.36) berechnet. Erfüllt auch dieser die Toleranzen ($\xi \leq 1$), so wird der gerechnete Schritt akzeptiert, und es erfolgt eine

[8] Im Fall der Elastodynamik führt die zeitliche Semidiskretisierung des kontinuierlichen Problems zu einem stationären, elliptischen Problem analog zur Elastostatik, dessen Ortsfehler mit den bekannten Fehlerschätzern für elliptische Probleme bewertet werden kann. Dieses Vorgehen wird an dieser Stelle ohne weitere Untersuchungen auf die behandelten Gleichungen des viskoplastischen Zweiphasenmodells übertragen. Die Entkopplung der einzelnen Fehleranteile ist daher als Arbeitshypothese zu verstehen.

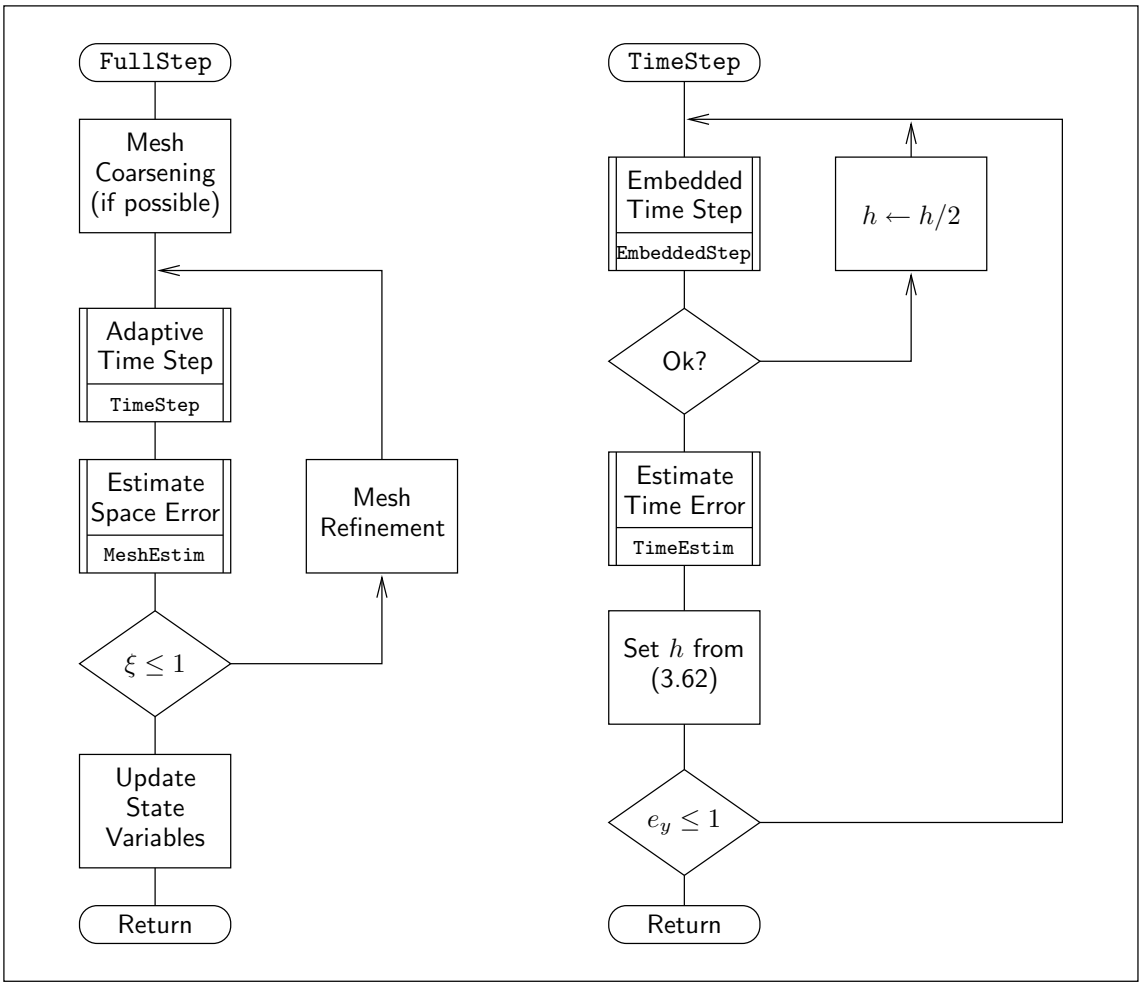


Abbildung 4.3: *Algorithmus für einen zeit- und ortsadaptiven Schritt*

Aktualisierung der Zustandsgrößen und die Weiterschaltung zum nächsten Zeitschritt ($n \leftarrow n+1$). Andernfalls wird das aktuelle Netz verfeinert, und der gesamte Zeitschritt wird auf Basis der Daten zum Zeitpunkt t_n wiederholt.

Bei jeder Netzänderung wird gemäß einer gewählten Dichtefunktion (Abschnitt 4.2.2) ein neues Netz mittels hierarchischer Netzanpassung (Abschnitt 4.3.1) bzw. Wiedervernetzung (Abschnitt 4.3.2) generiert. Die diskreten Zustandsgrößen (Freiheitsgrade und interne Variablen) zum Zeitpunkt t_n werden anschließend vom alten auf das neue Netz übertragen. Dies ist Thema des folgenden Abschnitts.

4.4.2 Transfer der diskreten Zustandsgrößen

Für das zeit- und ortsadaptive Gesamtverfahren spielt der Transfer der diskreten Zustandsgrößen bei Netzänderungen eine entscheidende Rolle. Wesentlich dabei ist, daß der Transfer konsistent mit der zugrundeliegenden schwachen Formulierung ist. Diese Eigenschaft wird auch als *variationelle Konsistenz* bezeichnet. Mit Hilfe eines Dreifeldfunk-

tionals vom *Hu-Washizu*-Typ wird von *Ortiz & Quigley* [91] eine Vorgehensweise beim Datentransfer hergeleitet, die die variationelle Konsistenz garantiert. Dabei wird lediglich vorausgesetzt, daß die schwache Formulierung auf einem Variationsfunktional beruht, und daß die Geschichtsvariablen der Plastizitätsformulierung an den Integrationspunkten der numerischen Quadratur berechnet werden, so daß die im folgenden skizzierte Vorgehensweise direkt auf die im Rahmen dieser Arbeit betrachteten Probleme übertragen werden kann:

- Die Geschichtsvariablen an den Integrationspunkten werden zum Zeitpunkt t_n vom alten Netz $\mathfrak{T}^h$ auf das neue Netz $\hat{\mathfrak{T}}^h$ übertragen; diese Werte dienen dann als Anfangswerte für den Zeitschritt von t_n nach t_{n+1}.
- Die FE-Knotenvariablen werden ebenfalls zum Zeitpunkt t_n vom alten auf das neue Netz übertragen; sofern die globalen Bilanzgleichungen keine Zeitabhängigkeit enthalten, sind die übertragenen Werte lediglich als Startwerte der anschließenden Gleichgewichtsiteration zu verstehen.
- Die diskrete schwache Formulierung der Bilanzgleichungen wird zum Zeitpunkt t_{n+1} bezüglich des neuen Netzes $\hat{\mathfrak{T}}^h$ mit Hilfe der üblichen Gleichgewichtsiteration (*Newton*-Verfahren) erfüllt.

Insgesamt führt dieses Vorgehen dazu, daß zum Zeitpunkt t_{n+1} auf dem neuen Netz $\hat{\mathfrak{T}}^h$ wieder ein Gleichgewichtszustand vorliegt; dem Transfer liegen dabei jeweils die Daten des letzten Gleichgewichtszustands zum Zeitpunkt t_n zugrunde. Daher müssen bis zur vollständigen Akzeptanz eines Zeitschritts das Netz und die Daten des letzten Zeitschritts vorgehalten werden.

Im Zusammenhang mit adaptiven Verfahren für Anfangs-Randwertprobleme hyperbolischer Differentialgleichungen wird ein analoges Vorgehen auch von *Franz* [56] angewandt.

Transfer der FE-Knotenvariablen

Die Freiheitsgrade der FE-Semidiskretisierung können direkt durch Auswertung der Ansatzfunktionen vom alten auf das neue Netz übertragen werden. Im Fall der hierarchischen Netzanpassung werden bei Verfeinerung eines Dreiecks zwei neue Dreiecke erzeugt. Dabei sind die lokalen Koordinaten der neu erzeugten Knoten aufgrund der hierarchischen Struktur bekannt, so daß direkt die Ansatzfunktionen ausgewertet werden können. Bei einer Vergröberung werden die Daten der entfernten Knoten anschließend nicht mehr benötigt, so daß kein Transfer erfolgen muß.

Im Fall der Wiedervernetzung ist das Vorgehen prinzipiell identisch; es muß jedoch zunächst für jeden Knoten des neuen Netzes dasjenige Element im alten Netz gefunden werden, das diesen Knoten enthält. Anschließend müssen die zu den globalen Koordinaten des Knotens im neuen Netz gehörigen lokalen Koordinaten bzgl. des gefundenen Elements im alten Netz bestimmt werden. Sowohl zur Elementsuche als auch zur Invertierung der Ansatzfunktionen wurden im Rahmen einer vom Verfasser der vorliegenden Arbeit betreuten Diplomarbeit (*Ammann* [6]) Untersuchungen durchgeführt.

Zur Elementsuche wird vorteilhaft zunächst ein *Quadtree* aufgebaut, bei dem mit Hilfe einer Baumstruktur das Rechengebiet rekursiv in Quadranten unterteilt wird, die jeweils maximal einen Elementmittelpunkt enthalten. Die Algorithmen zur Behandlung von Quadtree-Datenstrukturen können z. B. *Samet* [103] bzw. *Krause & Rank* [77] entnommen werden. Der Aufwand zum Aufbau des Baumes beträgt $\mathcal{O}(\mathrm{E} \cdot \log \mathrm{E})$, wobei E die Anzahl der Elemente im alten Netz ist. Der anschließende Suchaufwand je Knoten des neuen Netzes wird aufgrund der Baumstruktur auf $\mathcal{O}(\log \mathrm{E})$ reduziert, so daß der Gesamtaufwand zur Elementsuche $\mathcal{O}((\mathrm{E}+\hat{\mathrm{N}}) \log \mathrm{E})$ beträgt, worin $\hat{\mathrm{N}}$ die Anzahl der Knoten im neuen Netz bezeichnet. Dies stellt eine erhebliche Ersparnis gegenüber dem naiven Vergleich aller Elemente mit allen Knoten dar, der einen Suchaufwand von $\mathcal{O}(\mathrm{E} \cdot \hat{\mathrm{N}})$ zur Folge hätte.

Die Invertierung der Geometrie-Transformation zur Berechnung der lokalen Koordinaten aus den globalen Koordinaten erfordert bei Ansätzen höherer Ordnung i. a. die Lösung eines nichtlinearen Gleichungssystems. Eine Möglichkeit stellt hierbei die iterative Lösung mit Hilfe des *Newton*-Verfahrens dar. Für quadratische Ansätze in Dreiecken und Vierecken gibt es jedoch wesentlich effizientere Verfahren (*Ammann* [6], *Crawford, Anderson & Waggenspack* [34]), die spezielle Eigenschaften der Ansatzpolynome ausnutzen.

Transfer der Geschichtsvariablen an den Integrationspunkten

Beim Transfer der Geschichtsvariablen ist es üblich, diese zunächst im alten Netz geeignet auf Knotenwerte umzurechnen und dann genau wie beim Transfer der FE-Knotenvariablen zu verfahren (siehe z. B. *Cuitiño & Ortiz* [35], *Perić, Hochard, Dutko & Owen* [93]). Von *Ammann* [6] wurden noch weitere Möglichkeiten aus dem Bereich der numerischen Interpolation und Approximation untersucht, die jedoch nicht zu besseren Ergebnissen führten.

Der Transfer im Fall der hierarchischen Netzanpassung ist sehr lokal, d. h. bei einer Verfeinerung werden die Daten an den Integrationspunkten des Vater-Dreiecks mittels einer linearen Extrapolation auf die FE-Knoten übertragen; anschließend kann die dadurch gegebene Funktion an den Integrationspunkten der neu erzeugten Dreiecke ausgewertet werden. Im Fall einer Vergröberung muß statt der linearen Extrapolation eine L^2-Projektion der Daten an den Integrationspunkten der Kind-Dreiecke auf die FE-Knoten des Vater-Dreiecks durchgeführt werden.

Bei der Wiedervernetzung gestaltet sich das Problem etwas schwieriger, da keine direkten Beziehungen zwischen Elementen des alten und neuen Netzes bestehen. Es werden daher zunächst alle Integrationspunkt-Variablen im alten Netz mit Hilfe des SPR-Verfahrens auf die FE-Knoten projiziert. Anschließend kann analog zum Transfer der FE-Knotenvariablen vorgegangen werden, der im vorigen Abschnitt behandelt wurde.

Kapitel 5: Numerische Beispielrechnungen

Die in diesem Kapitel präsentierten numerischen Beispielrechnungen sind in zwei Klassen unterteilt. Zunächst werden anhand von Verifikationsbeispielen die verschiedenen Aspekte der entwickelten zeit- und ortsadaptiven Verfahren mittels analytischer Lösungen bzw. numerisch berechneter Referenzlösungen überprüft und bewertet. Auf dieser Grundlage werden anschließend praxisnahe Anwendungsbeispiele aus der Bodenmechanik präsentiert, die das Potential der entwickelten Methoden aufzeigen.

Zur Verifikation der zeitadaptiven Verfahren werden zwei Anfangs-Randwertprobleme bei fester Ortsdiskretisierung betrachtet – zum einen das elastische Konsolidationsproblem als ein klassisches Problem der Bodenmechanik und zum anderen ein Zugversuch aus dem Bereich der Metallplastizität (ideale *Prandtl-Reuß*-Plastizität). Diese stellt im Sinne eines singulär gestörten Grenzübergangs vom viskoplastischen zum elastoplastischen Fall einen schwierigen Testfall für numerische Zeitintegrationsverfahren dar, da i. a. nur mit einer geringen Lösungsregularität im Zeitbereich gerechnet werden kann. Die Verifikation der ortsadaptiven Strategien erfolgt danach anhand eines Beispiels aus der linearen Elastostatik, bei dem eine analytische Lösung vorliegt. Zur Überprüfung der Kombination der zeit- und ortsadaptiven Verfahren dient ein Biaxialversuch, bei dem durch Vorgabe einer kleinen Störung eine Scherbandbildung initiiert wird. Abschließend wird die Leistungsfähigkeit der gekoppelten zeit- und ortsadaptiven Verfahren anhand der klassischen bodenmechanischen Probleme des Böschungsbruchs und des Grundbruchs gezeigt, bei denen keine Referenzlösungen zur Verfügung stehen. Hier ist man auf den Vergleich mit klassischen Theorien oder mit Experimenten angewiesen.

5.1 Verifikation Zeitadaptivität

In den folgenden Beispielen werden die in Tabelle 5.1 aufgeführten SDIRK-Verfahren

Bezeichnungen		Verweis /	Parameter			Stabilität
kurz	lang	Quelle	s	p	$\hat{p}$	
`Euler`	Implizites *Euler*-Verfahren	Tabelle 3.1	1	1	–	A,L
`Trapez`	Trapezregel	Tabelle 3.1	2	2	–	A
`Midpnt`	Implizite Mittelpunktregel	Tabelle 3.1	1	2	–	A
`Alexander-2`	*Alexander* 2. Ordnung	*Alexander* [5]	2	2	–	A,L,S
`Alexander-3`	*Alexander* 3. Ordnung	*Alexander* [5]	3	3	–	A,L,S
`SDIRK-2(1)`	SDIRK 2(1)	Abschnitt 3.5	2	2	1	A,L,S
`Cash-3(2)`	*Cash* 3(2)	*Cash* [30]	3	3	2	A,L,S

Tabelle 5.1: *Verwendete SDIRK-Zeitintegrationsverfahren (Stufenzahl s, Ordnungen p und $\hat{p}$)*

miteinander verglichen. Die ersten drei Verfahren sind gebräuchliche *Runge-Kutta*-Verfahren, die hier nur zu Vergleichszwecken betrachtet werden. Die beiden Verfahren von *Alexander* [5] eignen sich wegen der guten Stabilitätseigenschaften hervorragend zur Zeitintegration von DAE mit Index 1. Außerdem kommen zwei eingebettete zeitadaptive SDIRK-Verfahren zum Einsatz, zum einen das in Abschnitt 3.5 konstruierte `SDIRK-2(1)`-Verfahren auf Basis von `Alexander-2` und zum anderen das von *Cash* [30] angegebene `Cash-3(2)`-Verfahren auf Basis von `Alexander-3`.

5.1.1 Elastische Konsolidation

Als erstes Beispiel zur Verifikation der Zeitintegrationsverfahren wird das klassische Konsolidationsproblem herangezogen. Ein dichter Behälter ist mit einem Zweiphasenmaterial angefüllt (linear elastischer poröser Festkörper und viskoses Porenfluid) und wird mittig mit einer Streckenlast beaufschlagt, wobei aus Symmetriegründen nur die Hälfte des Problems behandelt werden muß (Abbildung 5.1). Es wird das quasi-statische inkompressible

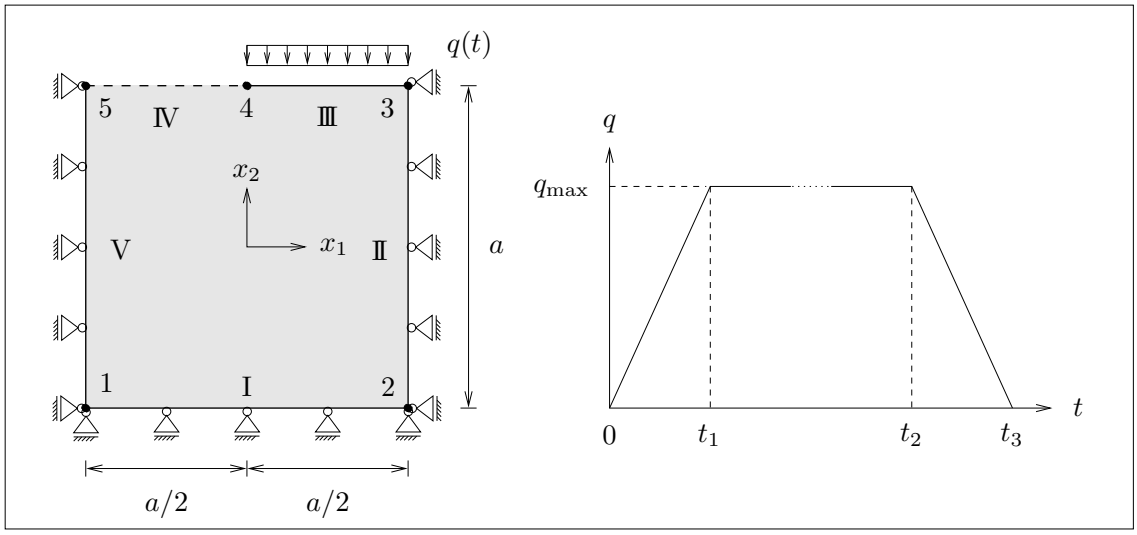


Abbildung 5.1: *Konsolidationsproblem; links: Skizze des Anfangs-Randwertproblems mit Makroknoten und Makrokanten, rechts: Lastverlauf*

Zweiphasenmodell zugrundegelegt (Abschnitt 1.6.2). Die verwendeten Materialparameter sind in Anhang B angegeben. Der linke Teil des oberen Randes (Makrokante IV) ist drainiert. Im Rahmen des inkompressiblen Zweiphasenmodells wird dies über eine homogene *Dirichlet*-Randbedingung für den Porenfluiddruck p modelliert ($p = 0$). Alle anderen Ränder sind undrainiert, d. h. die Sickergeschwindigkeit kann dort nur tangential zum Rand verlaufen ($n^F \, \mathbf{w}_F \cdot \mathbf{n} = 0$ in der schwachen Formulierung, vgl. (2.10), (2.14)). Dies entspricht dem Verschwinden der *Neumann*-Randbedingung in der Volumenbilanz. Die Last $q(t)$ wird als *Neumann*-Randbedingung entlang der Makrokante III in der schwachen Form der Impulsbilanz der Mischung berücksichtigt. Insgesamt führt dies zu den folgenden

Randbedingungen:

I:	Vertikalverschiebung	$\bar{u}_2$	$=$	0	Ⅲ:	Last	$\bar{t}_2$	$=$ $-q(t)$
Ⅱ, Ⅴ:	Horizontalverschiebung	$\bar{u}_1$	$=$	0	Ⅳ:	Druck	$\bar{p}$	$=$ 0

(5.1)

Die Kantenlänge des Gebiets wird zu $a = 10\,\mathrm{m}$ gesetzt. Die Last $q(t)$ wird linear bis zu einer Maximallast von $q_{\max} = 1\,\mathrm{MPa}$ innerhalb der Belastungszeit von $t_1 = 1\,\mathrm{d}$ aufgebracht. Während der anschließenden Haltephase von $t_2 - t_1 = 21\,\mathrm{d}$ (drei Wochen) kann eine Ausströmung des Porenfluids über den oberen Rand stattfinden, die mit einer Setzung verbunden ist (Konsolidation). Schließlich führt die Entlasung innerhalb von einem Tag $(t_3 - t_2)$ zu einem leichten Rückgang der Setzung.

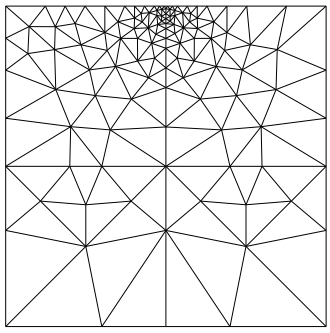


Abbildung 5.2: *FE-Netz*

Das Problem wird mit einem festen FE-Netz berechnet, das im Bereich des Übergangs vom belasteten zum unbelasteten Bereich etwas verfeinert ist (Abbildung 5.2, 462 Knoten, 213 *Taylor-Hood*-Elemente (P2P1), 1049 Freiheitsgrade). Die numerische Ermittlung der Konvergenzordnung setzt die Kenntnis einer Referenzlösung voraus. Da für dieses Beispiel keine analytische Lösung vorliegt, wurde zunächst mit dem `Alexander-3`-Verfahren eine Referenzlösung mit 6624 Zeitschritten im betrachteten Zeitintervall von 23 Tagen berechnet. Dies entspricht einer Schrittweite h von 5 Minuten. Als Referenzwerte zum späteren Vergleich wurden dabei die Vertikalverschiebung u_2 sowie der Porenfluiddruck p am Punkt 3 (Lastmitte, siehe Abbildung 5.1) nach der Zeit von 23 Tagen aufgenommen, so daß sowohl der Fehler in den differentiellen Variablen (Verschiebungen) als auch in den algebraischen Variablen (Druck) des DAE-Systems (2.57) beurteilt werden kann.

Mit den ersten fünf Verfahren aus Tabelle 5.1 wurden Rechnungen mit den konstanten Schrittweiten $h = 1\,\mathrm{d},\ 1/2\,\mathrm{d}, \ldots, 1/128\,\mathrm{d}$ durchgeführt. Da es sich um ein lineares Problem handelt, erhält man mit der Trapezregel und der impliziten Mittelpunktregel identische Ergebnisse. In Abbildung 5.3 ist der relative Fehler im Porenfluiddruck über der Schrittweite h (in Tagen) aufgetragen. Durch den doppelt logarithmischen Maßstab kann als Steigung direkt die numerisch erreichte Ordnung p des Verfahrens abgelesen werden. Man erkennt, daß von allen Verfahren die jeweilige Ordnung p erreicht wird. Dieses Ergebnis zeigt eine gute Übereinstimmung mit den theoretischen Aussagen aus Kapitel 3, wonach bei DAE-Systemen vom Index 1 keine Ordnungsreduktion auftritt. Im Fall der Vertikalverschiebung ergibt sich ein analoges Verhalten, wobei eine etwas höhere Genauigkeit als im Druck erreicht wird.

Zur Verifikation der Effizienz der zeitadaptiven Verfahren `SDIRK-2(1)` (auf Basis von `Alexander-2`) und `Cash-3(2)` (auf Basis von `Alexander-3`) wurden Rechnungen mit den relativen und absoluten Toleranzen

$$\epsilon_r = \epsilon_a = 10^{-2},\ 10^{-3}, \ldots,\ 10^{-8}$$

durchgeführt. In Abbildung 5.4 ist der jeweilige Rechenaufwand über der erreichten Genauigkeit aufgetragen. Zum Vergleich sind die Ergebnisse aus den obigen Rechnungen mit konstanter Schrittweite eingetragen.

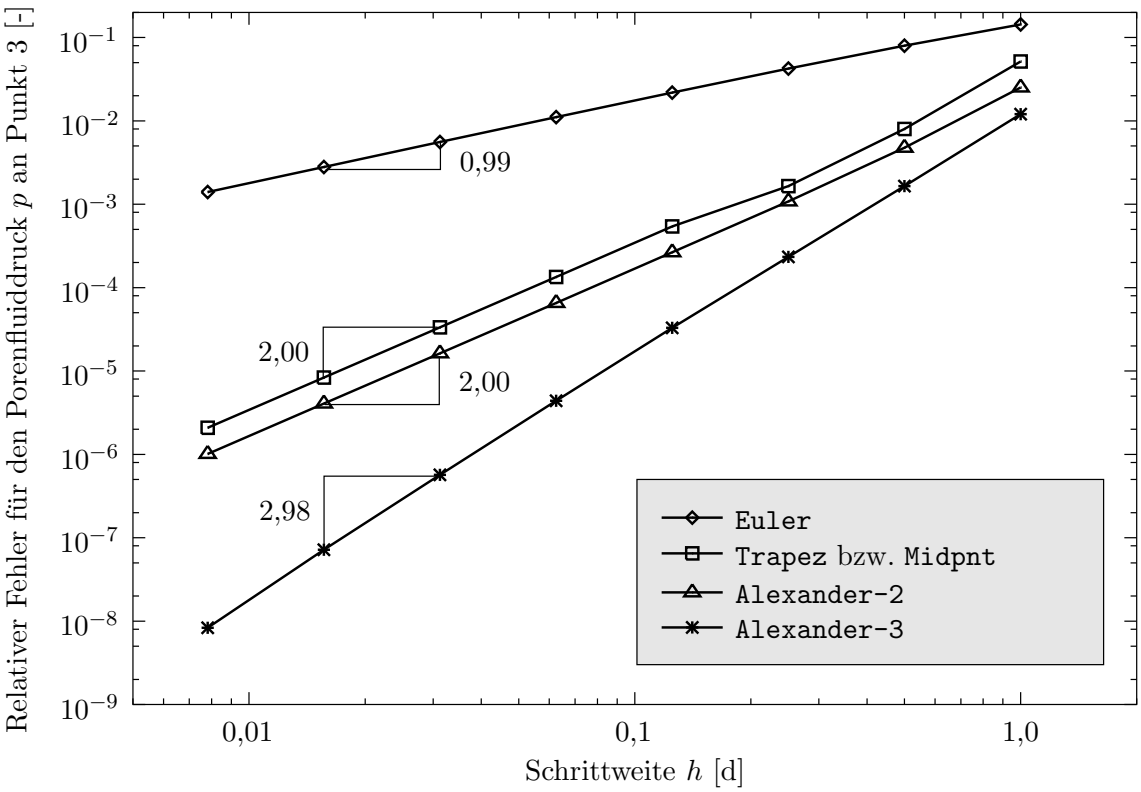


Abbildung 5.3: *Konsolidationsproblem: Konvergenzordnungen für den Porenfluiddruck*

Die Ersparnis durch Verwendung der zeitadaptiven Verfahren (gestrichelte Linien) ist deutlich erkennbar. Je nach Verfahren und Toleranz ist der Aufwand zur Erreichung einer gegebenen Genauigkeit bei konstanter Schrittweite um einen Faktor 3 bis 8 höher als bei zeitadaptiver Rechnung.

Es kann nicht erwartet werden, daß die gegebenen Toleranzen exakt eingehalten werden, da der Fehler über eine gewichtete Norm beurteilt wird. Außerdem wird nur der lokale Fehler kontrolliert, und im Diagramm ist der globale Fehler aufgetragen. Da es sich jedoch um ein stark dissipatives Problem handelt, werden zurückliegende Fehler gedämpft, und ein Vergleich von gegebener Toleranz und erreichtem Fehler ist sinnvoll. Dieser Vergleich zeigt, daß beide Verfahren die Toleranz in etwa einhalten – `Cash-3(2)` ist dabei etwas genauer als `SDIRK-2(1)`. Der Anwender kann somit *a priori* eine Genauigkeit wählen, und das zeitadaptive Verfahren liefert mit sehr hoher Effizienz eine akzeptable Lösung.

In Abbildung 5.5 sind die Schrittweitenverläufe aufgetragen, die sich mit dem Verfahren `SDIRK-2(1)` bei verschiedenen Toleranzen ergeben. Im Belastungsbereich wird die Schrittweite zunächst stückweise vergrößert, fällt dann am Lastknick ($t = 1\,\mathrm{d}$) stark ab, und wird während der Konsolidationsphase exponentiell vergrößert. Beim Knick zur Entlastungsphase ($t = 22\,\mathrm{d}$) wird die Schrittweite wieder drastisch verkleinert und zeigt dann ein ähnliches Verhalten wie in der Belastungsphase.

Insgesamt konnte mit diesem Beispiel gezeigt werden, daß einerseits die *Runge-Kutta*-Verfahren höherer Ordnung die theoretische Ordnung erreichen, und daß andererseits

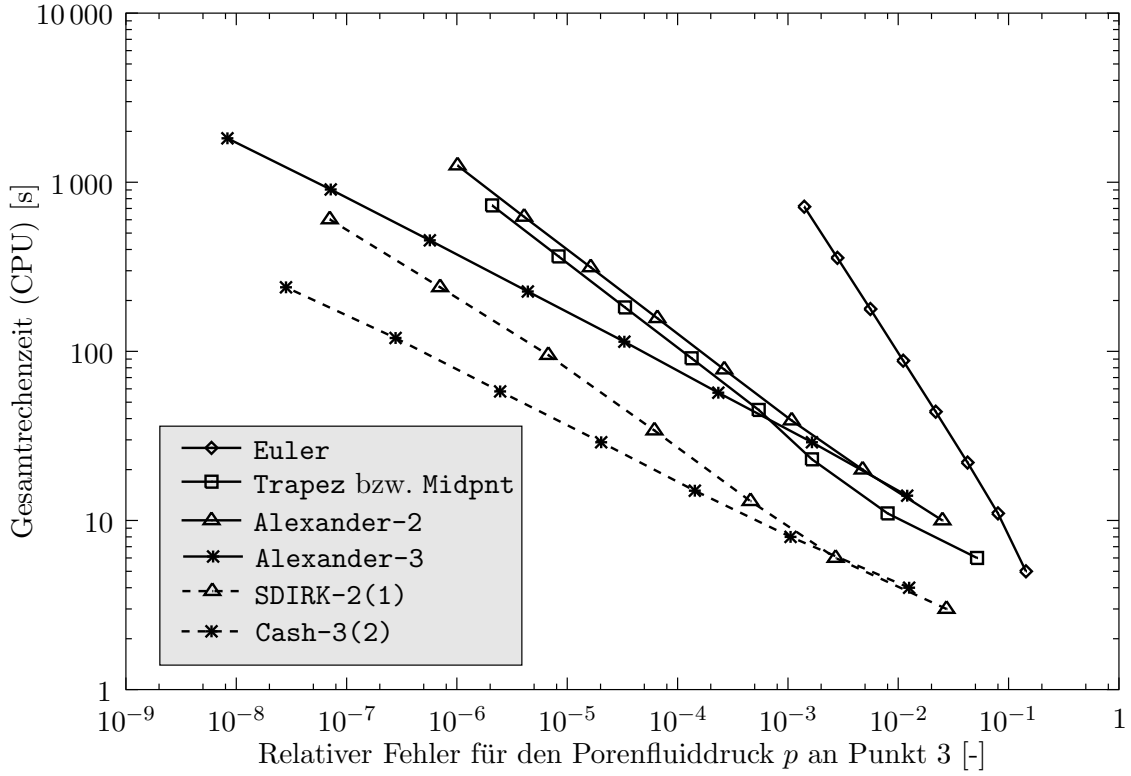


Abbildung 5.4: *Konsolidationsproblem: Aufwand über erreichter Genauigkeit*

durch Verwendung der Schrittweitensteuerung (Zeitadaptivität) eine enorme Ersparnis im Gesamtrechenaufwand möglich ist.

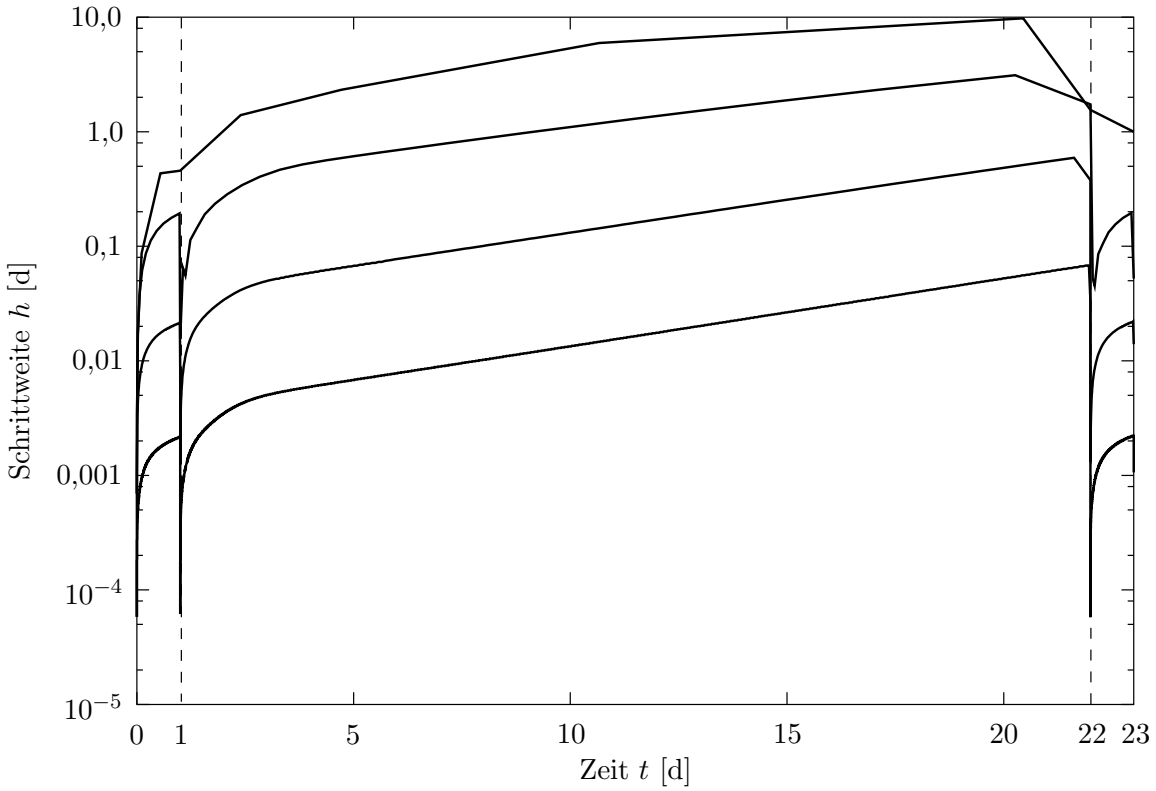


Abbildung 5.5: *Konsolidationsproblem: Schrittweitenverlauf bei* `SDIRK-2(1)` *bei verschiedenen Toleranzen* $\epsilon_r = \epsilon_a = 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}$ *(von oben nach unten)*

5.1.2 DFG-Benchmark: *Prandtl-Reuß*-Plastizität

Problemstellung

Als zweites Beispiel zur Verifikation der zeitadaptiven Verfahren wird ein „numerisch hartes“ Problem[1] gewählt, bei dem als Materialmodell die ideale Elastoplastizität (*Prandtl-Reuß*-Plastizität ohne Verfestigung) zusammen mit der *von-Mises*-Fließbedingung für Stahl zugrundegelegt wird. Das Modell wird im folgenden knapp beschrieben, für Details sei auf die Literatur verwiesen, z. B. *Lubliner* [80], *Simo & Hughes* [106].

Die Verzerrungen werden wie üblich in elastische und plastische Anteile zerlegt:

$$\boldsymbol{\varepsilon} = \frac{1}{2}(\operatorname{grad}\mathbf{u} + \operatorname{grad}^T\mathbf{u}) = \boldsymbol{\varepsilon}_e + \boldsymbol{\varepsilon}_p\,. \tag{5.2}$$

Die Verschiebungen $\mathbf{u}$ stellen die Primärvariablen $\mathbf{u}$ des Problems dar, die plastischen Verzerrungen $\boldsymbol{\varepsilon}_p$ sind interne Variablen im Sinne der Thermodynamik. Der Vektor aller internen Variablen $\mathbf{q} = (\boldsymbol{\varepsilon}_p, \Lambda)$ enthält zusätzlich den plastischen Multiplikator Λ. Das *von-Mises*-Fließkriterium mit der Fließgrenze k_0 lautet im Fall der hier betrachteten idealen Plastizität:

$$F(\boldsymbol{\sigma}) \;=\; \|\boldsymbol{\sigma}^D\| - \sqrt{\frac{2}{3}}\,k_0\,. \tag{5.3}$$

[1] Damit ist gemeint, daß die Lösung im Zeitbereich i. a. nur eine geringe Regularität besitzt.

Wie in Kapitel 2 kann die statische Impulsbilanz in eine schwache Formulierung überführt werden,

$$\mathcal{G}[\boldsymbol{\eta}, \mathbf{u}; \mathbf{q}] \;\equiv\; \int_{\Omega} \operatorname{grad} \boldsymbol{\eta} \cdot \boldsymbol{\sigma} \, \mathrm{d}v - \int_{\Omega} \boldsymbol{\eta} \cdot \rho \mathbf{b} \, \mathrm{d}v - \int_{\Gamma_\mathrm{t}} \boldsymbol{\eta} \cdot \bar{\mathbf{t}} \, \mathrm{d}a \;=\; 0\,, \tag{5.4}$$

wobei der Spannungstensor $\boldsymbol{\sigma}$ im elastischen Fall ($F < 0$) über das *Hooke*sche Gesetz

$$\boldsymbol{\sigma} \;=\; 2\,\mu\,\boldsymbol{\varepsilon}_e + \lambda\,(\operatorname{tr}\boldsymbol{\varepsilon}_e)\,\mathbf{I} \tag{5.5}$$

direkt aus den elastischen Verzerrungen und im plastischen Fall ($F \geq 0$) über das System differential-algebraischer Gleichungen

$$\begin{aligned} \dot{\boldsymbol{\varepsilon}}_p &= \Lambda \frac{\partial F}{\partial \boldsymbol{\sigma}}\,, \\ 0 &= F(\boldsymbol{\sigma}) \end{aligned} \tag{5.6}$$

zusammen mit dem plastischen Multiplikator Λ bestimmt wird. Die Ortsdiskretisierung mit der Methode der finiten Elemente erfolgt völlig analog zum viskoplastischen Zweiphasenmodell (vgl. Kapitel 2). Die zweite Gleichung in (5.6) bestimmt den plastischen Multiplikator Λ, so daß es sich beim Modell der *Prandtl-Reuß*-Plastizität im Zeitbereich formal um ein DAE-System vom differentiellen Index 2 handelt[2]. Bei SDIRK-Verfahren muß daher mit einer Ordnungsreduktion gerechnet werden.

Im Rahmen des DFG-Schwerpunkts „Adaptivität“ wurde ein Benchmark-Problem definiert, anhand dessen verschiedene ortsadaptive Verfahren untersucht und einander gegenübergestellt werden können (siehe z. B. *Barthold, Schmidt & Stein* [17], *Rannacher & Suttmeier* [99], *Wieners* [122]). Dieses Beispiel wird hier nur in bezug auf die adaptive Zeitdiskretisierung, d. h. mit fester Ortsdiskretisierung, untersucht.

Eine quadratische Scheibe mit kreisförmigem Loch wird mit einer zeitabhängigen Last

$$q(t) \;=\; 100\ \mathrm{MPa} \cdot t\,\frac{1}{\mathrm{s}} \tag{5.7}$$

in x_2-Richtung auf Zug belastet (Abbildung 5.6). Dadurch entsteht ausgehend vom Rand des Lochs eine plastische Zone, die sich nach und nach ins Innere der Scheibe ausbreitet (Abbildung 5.7). Die im Benchmark definierten Materialparameter und Geometriedaten sind in Tabelle 5.2 angegeben.

Referenzlösung

Das Problem wird im Zeitbereich $[0;\ 4{,}5\,\mathrm{s}]$ betrachtet. Zum Zeitpunkt $t = 4{,}66\,\mathrm{s}$ ist bei der kritischen Last $q_\mathrm{krit} = 466\,\mathrm{MPa}$ der theoretische Versagenszustand der gesamten Probe erreicht, so daß der o. a. Zeitbereich für den ideal-plastischen Fall des Benchmark-Problems festgelegt wurde.

[2]Wie bereits in Kapitel 3 erwähnt, gibt es nach Wissen des Verfassers bisher keine Untersuchungen über den für die Numerik wichtigen Störungsindex solcher Systeme. Die Anwendbarkeit des impliziten *Euler*-Verfahrens und der Erhalt der Ordnung bei einigen SDIRK-Verfahren zweiter Ordnung (s. u.) legt allerdings die Vermutung nahe, daß das System numerisch nicht den Index 2 besitzt [123].

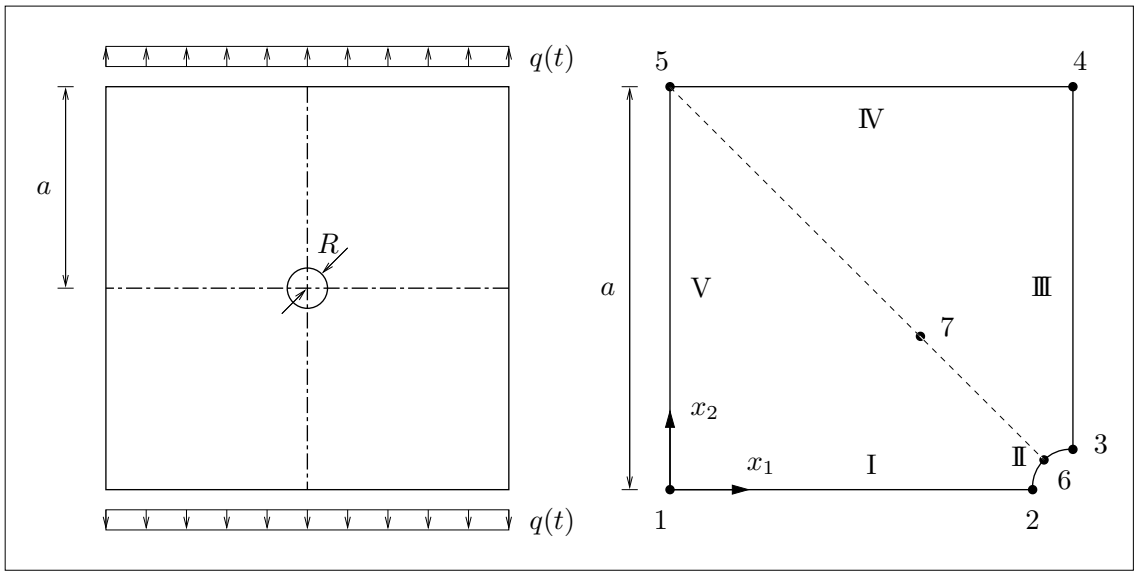


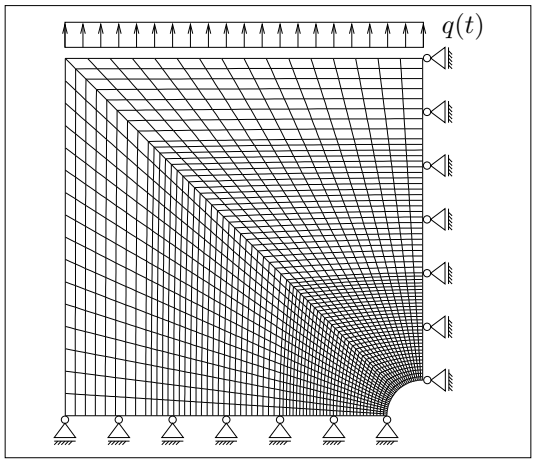
Abbildung 5.6: *Quadratische Scheibe mit kreisförmigem Loch unter Zugbelastung; links: Skizze des Randwertproblems, rechts: Makroknoten und Makrokanten für die FE-Diskretisierung*

Parameter	Symbol	Wert
Elastizitätsmodul	E	206 900,0 MPa
Querkontraktionszahl	ν	0,29
Fließgrenze	k_0	450,0 MPa
Lochradius	R	10,0 mm
Kantenlänge des Quadrats	a	200,0 mm
Position von Punkt 7	p_7	$1/3\,p_5 + 2/3\,p_6$

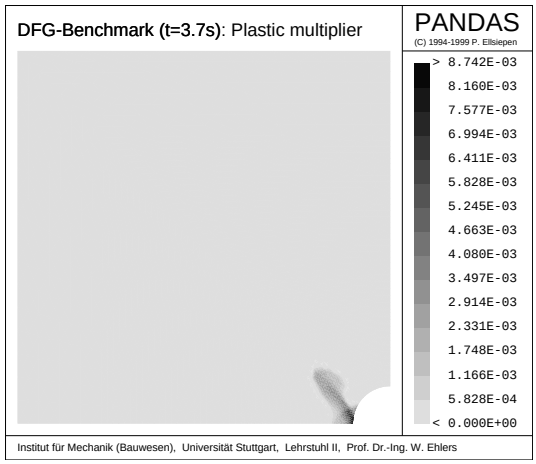
Tabelle 5.2: *DFG-Benchmark: Materialparameter und Geometriedaten*

Bei der Zeit t handelt es sich im Fall der Elastoplastizität um eine Pseudo-Zeit, da die Materialgleichungen unabhängig von der Zeit sind (ratenunabhängige Plastizität). Die Zeit muß daher als Lastfaktor verstanden werden; eine Skalierung der Lastgeschwindigkeit in (5.7) hat keinen Einfluß auf die Lösung des Problems.

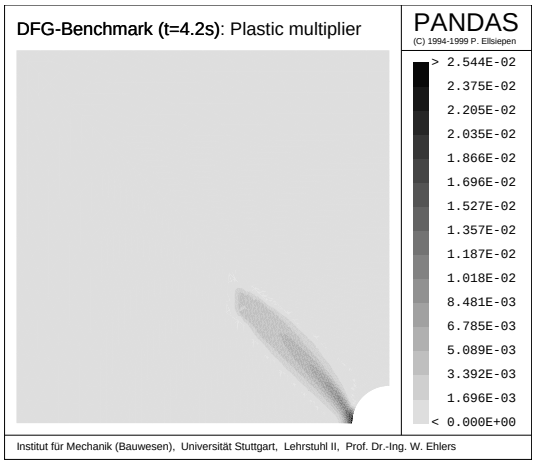
Im Benchmark werden Spannungs- und Verschiebungsdaten an einigen der Punkte p_1 bis p_7 als Vergleichswerte gefordert (siehe Abbildung 5.6). *Wieners* [122] gibt dafür Referenzwerte an, die mit einer „Overkill"-Lösung (4 Mio. Freiheitsgrade, Mehrgitterverfahren) berechnet wurden. Die Referenzwerte zum Zeitpunkt $t = 4{,}5\,\mathrm{s}$ sind in Tabelle 5.3 angegeben. Da hier die Verifikation der Zeitintegrationsverfahren mit dem fest gewählten Netz aus Abbildung 5.7(a) durchgeführt werden soll, kann die o. g. Referenzlösung nicht direkt zum Vergleich herangezogen werden. Es ist vielmehr eine Referenzlösung im Zeitbereich bei festgehaltener Ortsdiskretisierung erforderlich. Diese wurde mit dem Verfahren zweiter Ordnung von *Alexander* (`Alexander-2`) und der sechs mal je Zeitintervall halbierten Zeitschrittreihe aus Tabelle 5.4 (insgesamt 1793 an den Lösungsverlauf angepaßte Zeit-



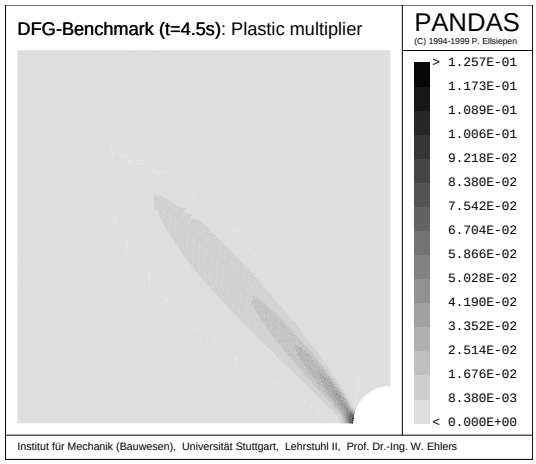
(a) Netz und Randbedingungen



(b) $t = 3{,}7$ s, Last $q(t) = 370$ MPa



(c) $t = 4{,}2$ s, Last $q(t) = 420$ MPa



(d) $t = 4{,}5$ s, Last $q(t) = 450$ MPa

Abbildung 5.7: *DFG-Benchmark: (a): Netz aus 8-knotigen Viereckelementen mit quadratischem Ansatz (Q2),* 4 *Gauß-Punkte je Element,* 6337 *Knoten,* 2048 *Elemente,* 12674 *Freiheitsgrade,* 40960 *interne Variablen; (b)-(d): Entwicklung der plastischen Zone (plastischer Multiplikator* Λ*) mit zunehmender Belastung* $q(t)$

Quelle	$\mathrm{u}_1(p_2)$ [mm]	$\mathrm{u}_2(p_4)$ [mm]	$\mathrm{u}_1(p_5)$ [mm]	$\sigma_{22}(p_2)$ [MPa]	$\sigma_{11}(p_7)$ [MPa]	$\int_{p_4}^{p_5} \lvert\mathrm{u}_2\rvert ds$ [mm^2]
Wieners [122]	0,019228	0,24577	0,061830	519,59	−52,585	22,36367
hier (s. Text)	0,01922209	0,2457789	0,06182936	520,3594	−52,61786	22,363826

Tabelle 5.3: *DFG-Benchmark: Referenzwerte*

schritte) berechnet und ist in der zweiten Zeile von Tabelle 5.3 angegeben. Eine weitere Halbierung der Zeitintervalle lieferte identische Ergebnisse, so daß von einer Konvergenz im Zeitbereich ausgegangen wird.

Nr.	Zeitpunkt t [s]	Schrittweite h [s]
0	0,0	1,0
1	1,0	0,5
5	3,0	0,25
7	3,5	0,1

Nr.	Zeitpunkt t [s]	Schrittweite h [s]
9	3,7	0,075
13	4,0	0,05
17	4,2	0,025
29	4,5	

Tabelle 5.4: *DFG-Benchmark: Zeitschrittreihe mit 29 Zeitschritten (aus Wieners [122])*

Konvergenzordnung

Zunächst werden die ersten fünf Zeitintegrationsverfahren aus Tabelle 5.1 bei Vorgabe der Schrittweite gemäß der Zeitschrittreihe in Tabelle 5.4 verglichen. Dabei wird die Rechnung jeweils einmal mit 29 Zeitschritten durchgeführt, beim nächsten Mal mit 57 Zeitschritten (halbierte Zeitintervalle mit Ausnahme des ersten Schritts), und so weiter bis zu 449 Zeitschritten. Am Ende jeder Rechnung ($t = 4{,}5\,\mathrm{s}$) wird beispielhaft der Spannungswert $\sigma_{11}(p_7)$ innerhalb der plastischen Zone mit der oben ermittelten zeitlichen Referenzlösung verglichen. In Abbildung 5.8 sind die relativen Fehler über der Schrittweite des letzten

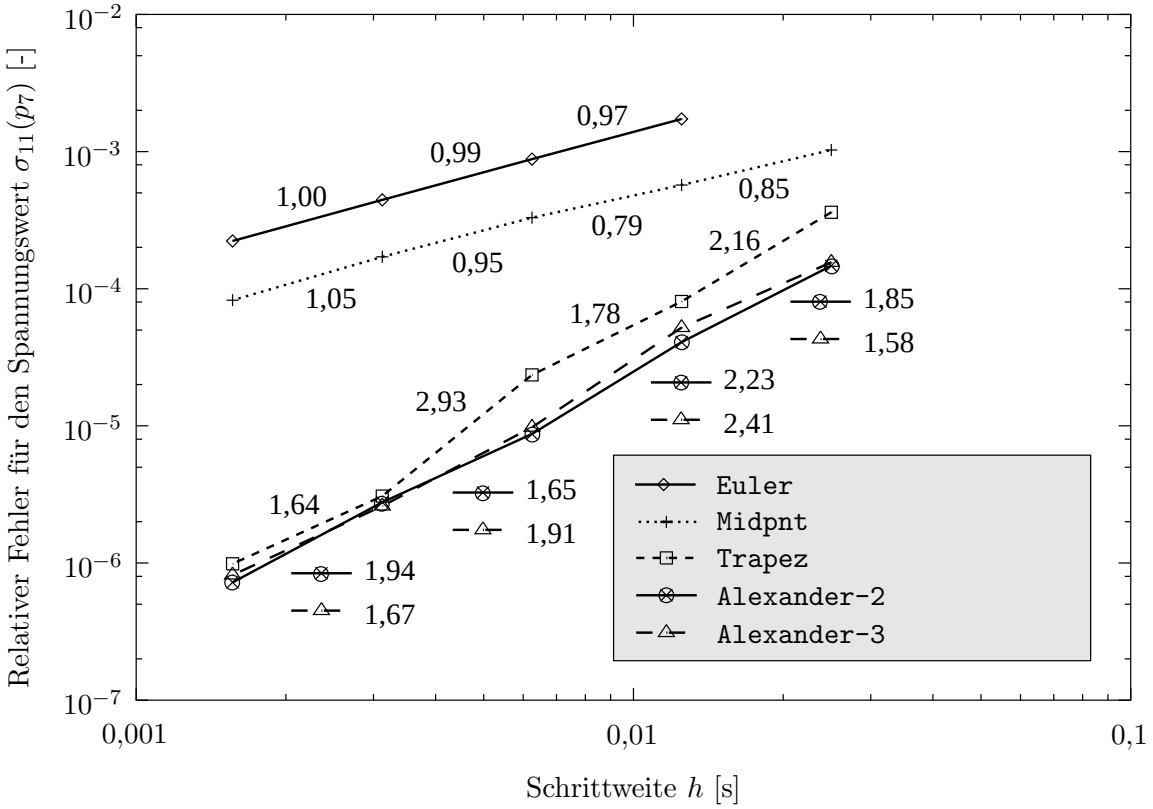


Abbildung 5.8: *DFG-Benchmark: Konvergenzordnung einiger Zeitintegrationsverfahren*

Zeitintervalls aufgetragen. Durch den doppelt logarithmischen Maßstab kann als Steigung direkt die numerisch erreichte Ordnung p des Verfahrens abgelesen werden. Die jeweiligen

Zahlenwerte sind an den Kurvenstücken vermerkt. Beim impliziten *Euler*-Verfahren war die Rechnung mit 29 Schritten nicht möglich, da im letzten Schritt keine Konvergenz im *Newton*-Verfahren erzielt werden konnte.

Die Ergebnisse lassen die folgenden Interpretationen zu:

- Das implizite *Euler*-Verfahren erreicht wie erwartet die Konvergenzordnung $p = 1$.
- Die implizite Mittelpunktregel erfährt eine Ordnungsreduktion um Eins. Dies ist darauf zurückzuführen, daß die implizite Mittelpunktregel nicht steif genau ist, und daß somit häufig die Nebenbedingung (Fließbedingung) verletzt wird.

  Fazit: Die implizite Mittelpunktregel ist zur Behandlung von Problemen der Elastoplastizität nicht zu empfehlen.
- Die Trapezregel erreicht im Mittel die Ordnung $p = 2$. Allerdings ist das Verfahren nicht L-stabil und daher nur bedingt für Plastizitätsprobleme geeignet, die per se stark dissipativ sind. Dies könnte die Ursache für das nicht-monotone Konvergenzverhalten sein.
- Das SDIRK-Verfahren zweiter Ordnung von *Alexander* [5] (`Alexander-2`) erreicht die theoretische Verfahrensordnung $p = 2$ und liefert von allen Verfahren die höchste Genauigkeit.

  Fazit: Das Verfahren ist sehr gut zur Zeitintegration von Problemen der Elastoplastizität geeignet.
- Das SDIRK-Verfahren dritter Ordnung von *Alexander* [5] (`Alexander-3`) erleidet eine Ordnungsreduktion um Eins und liegt bei der Genauigkeit in etwa gleichauf mit dem Verfahren zweiter Ordnung. Der höhere Aufwand für die dritte Stufe zahlt sich im vorliegenden Fall also nicht aus.

  Fazit: Bei „numerisch harten“ Problemen wie der hier betrachteten idealen Elastoplastizität lohnt sich der Einsatz des Verfahrens dritter Ordnung nicht.

Insgesamt kann festgestellt werden, daß sich die in Kapitel 3 getroffene Auswahl von Verfahren aufgrund von theoretischen Aussagen in der numerischen Praxis bewährt. Es besteht eine gute Übereinstimmung zwischen den von der Theorie her zu erwartenden Ordnungs-Aussagen und den numerisch ermittelten Ergebnissen. Insbesondere ist die Verwendung von steif genauen Verfahren wesentlich, wie sich am schlechten Abschneiden der impliziten Mittelpunktregel zeigt.

Zeitadaptive Rechnungen

Das gleiche Problem wird nun mit den beiden eingebetteten Verfahren `SDIRK-2(1)` (Abschnitt 3.5, Basis: `Alexander-2`) und `Cash-3(2)` (*Cash* [30], Basis: `Alexander-3`) unter Verwendung der Schrittweitensteuerung berechnet (Zeitadaptivität). Hier kann nicht erwartet werden, daß im Gesamtaufwand die gleiche Effizienz erreicht wird wie mit der hand-

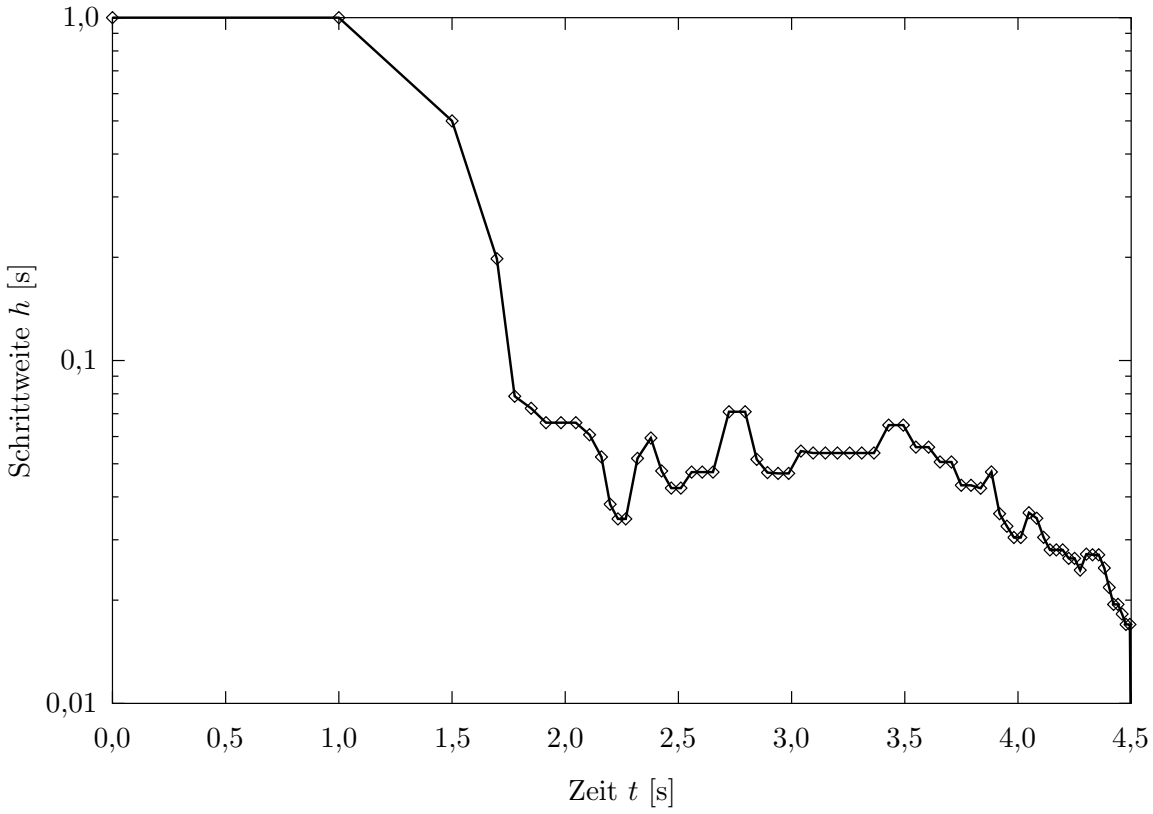


Abbildung 5.9: *DFG-Benchmark: Schrittweitenverlauf bei* `SDIRK-2(1)`, $\epsilon_r = \epsilon_a = 10^{-4}$

optimierten Schrittweitenreihe aus Tabelle 5.4, die in den obigen Rechnungen zur Anwendung kam. In Abbildung 5.9 ist beispielhaft ein Schrittweitenverlauf beim `SDIRK-2(1)`-Verfahren aufgetragen. Man erkennt, daß im elastischen Bereich zunächst mit der Maximalschrittweite gerechnet wird (hier: $h_{\max} = 1\,\mathrm{s}$). Das Einsetzen der Plastizität bei $t \approx 1{,}7\,\mathrm{s}$ führt zur drastischen Verkleinerung der Schrittweite, die anschließend bis $t \approx 3{,}5\,\mathrm{s}$ etwa auf dem gleichen Niveau verbleibt. Ab diesem Zeitpunkt beginnt die weitere Ausbreitung der plastischen Zone ins Innere der Scheibe (vgl. Abbildung 5.7), so daß durch die Schrittweitensteuerung nach und nach kleinere Schrittweiten gewählt werden.

Im Aufwands-Genauigkeits-Diagramm (Abbildung 5.10) sind sowohl die Ergebnisse der obigen Rechnungen mit fester Schrittweitenreihe als auch die zeitadaptiven Ergebnisse der beiden eingebetteten Verfahren dargestellt. Die zeitadaptiven Rechnungen wurden jeweils mit den relativen und absoluten Toleranzen

$$\epsilon_r = \epsilon_a = 10^{-2},\ 10^{-3},\ 10^{-4},\ 10^{-5} \tag{5.8}$$

durchgeführt. Im Diagramm ist für jede durchgeführte Rechnung ein Punkt auf der entsprechenden Kurve dargestellt. Aus den Ergebnissen des Aufwands-Genauigkeits-Vergleichs lassen sich einige Schlußfolgerungen ziehen:

- In Abbildung 5.10 ist sofort der höhere Aufwand beim `Alexander-3`-Verfahren gegenüber dem `Alexander-2`-Verfahren ersichtlich (drei statt zwei Stufen je Zeit-

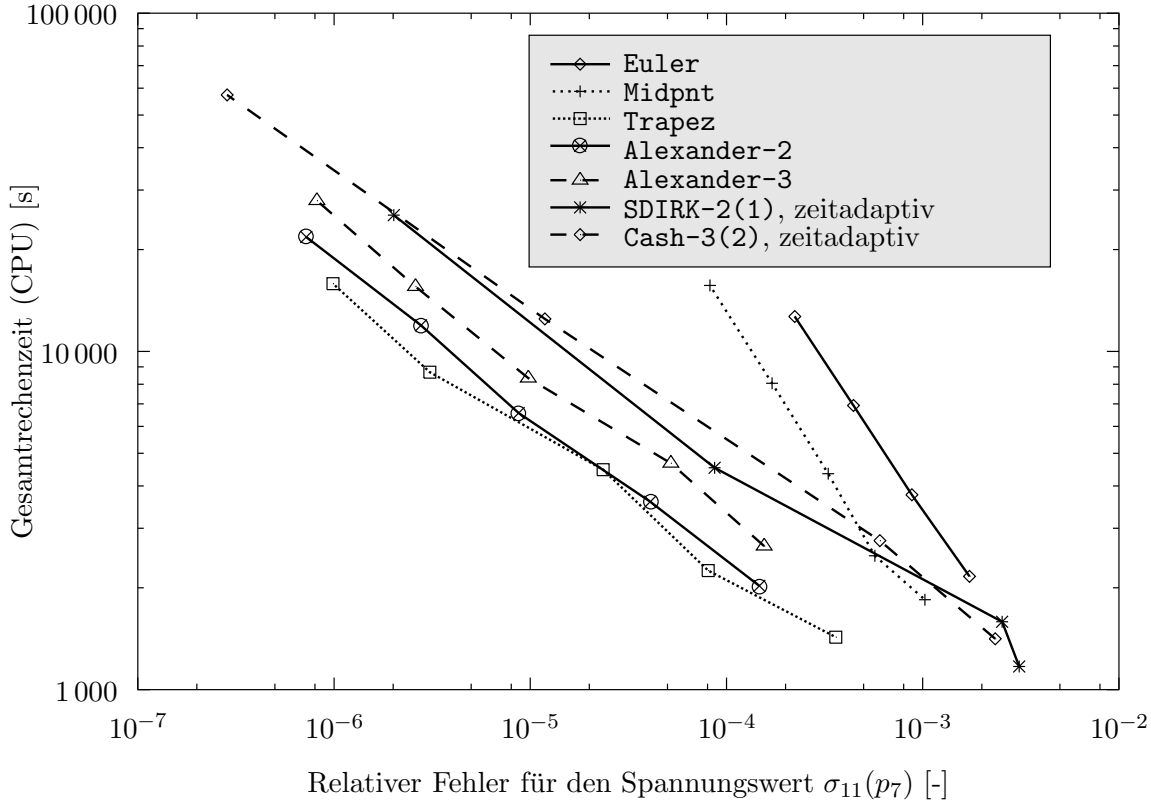


Abbildung 5.10: *DFG-Benchmark: Aufwand über erreichter Genauigkeit*

schritt, d. h. ein um 50% erhöhter Aufwand), der sich wegen der Ordnungsreduktion nicht in einer höheren Genauigkeit niederschlägt.

- Die Ordnungsergebnisse aus dem letzten Abschnitt übertragen sich auf die zeitadaptive Rechnung, d. h. das `SDIRK-2(1)`-Verfahren auf Basis des `Alexander-2`-Verfahrens schneidet besser ab als das `Cash-3(2)`-Verfahren auf Basis des `Alexander-3`-Verfahrens, bei dem wieder eine Ordnungsreduktion stattfindet.

- Bei den beiden zeitadaptiven Verfahren können die vorgegebenen Toleranzen nicht direkt mit der erreichten Genauigkeit verglichen werden, da sich die Toleranzen auf Normen beziehen (2-Norm bzw. Maximumnorm, vgl. Abschnitt 3.4.4), während hier der Fehler in den Spannungen am Punkt p_7 betrachtet wird. Für die punktweise Konvergenz der Spannungen innerhalb der plastischen Zone existieren jedoch keine analytischen Aussagen (*Wieners* [122]). Es kann aber festgestellt werden, daß die Vorgabe einer kleineren Toleranz systematisch zu kleineren Fehlern führt, so daß der Anwender die erforderliche Genauigkeit über die Toleranzen steuern kann.

- Die „Ausreißer“ beim `SDIRK-2(1)`-Verfahren für die Toleranzen 10^{-2} und 10^{-3} (die beiden Kurvenpunkte ganz rechts) können dadurch erklärt werden, daß hier die absolute Toleranz ϵ_a zu groß gewählt ist. Dadurch wird wesentliche Information ignoriert, da die plastischen Verzerrungen bei einsetzender Plastizität gerade in diesem Bereich liegen. Dies führt dazu, daß der relative Fehler nicht mehr kontrolliert

werden kann. Man sollte also lieber eine zu kleine als eine zu große absolute Toleranz wählen und zur Fehlerkontrolle eine geeignete relative Toleranz vorgeben.

- Im Vergleich mit den Lösungen der hand-optimierten Zeitschrittreihe „kostet“ die Zeitadaptivität in etwa einen Faktor 2 an Rechenzeit. Dafür bekommt man automatisch eine Lösung mit vorgegebener Genauigkeit, ohne daß die genaue Kenntnis des Lösungsverlaufs erforderlich ist.

Im Bereich der numerischen Simulation von Plastizitätsproblemen (engl.: computational plasticity) wird bisher weitgehend das implizite *Euler*-Verfahren eingesetzt. Aufgrund obiger Ergebnisse kann daher als neue Erkenntnis auf diesem Gebiet festgehalten werden, daß man auch beim Problem der idealen Elastoplastizität einen Nutzen aus Verfahren höherer Ordnung ziehen kann. Da die ideale Plastizität in bezug auf die numerische Problematik einen Grenzfall darstellt, können ähnliche Ergebnisse auch bei praxisnäheren Materialmodellen erwartet werden, die etwa durch Hinzunahme einer isotropen und/oder kinematischen Verfestigung entstehen.

Vergleicht man bei gleichem Aufwand das Ergebnis des *Euler*-Verfahrens mit 57 Schritten und das Ergebnis des `Alexander-2`-Verfahrens mit 29 Schritten, so liefert die Verwendung des Verfahrens zweiter Ordnung bereits einen Genauigkeitsgewinn von etwa Faktor 12. Wegen der höheren Ordnung nimmt dieser Abstand mit zunehmendem Aufwand dramatisch zu, wie in Abbildung 5.10 gut zu erkennen ist. Bei 449 (bzw. 225) Zeitschritten ist der Verbesserungsfaktor bereits 80.

Im Vergleich des zeitadaptiven Verfahrens `SDIRK-2(1)` mit dem impliziten *Euler*-Verfahren zeigt sich zudem der enorme Vorteil einer Schrittweitensteuerung. Man erhält mit wesentlich geringerem Aufwand und automatischer Fehlerkontrolle eine wesentlich genauere Lösung.

5.2 Verifikation Ortsadaptivität

5.2.1 Elastische Scheibe mit Loch

Problemstellung

Zur Verifikation der Algorithmen der Ortsadaptivität wird ein Randwertproblem der linearen Elastostatik betrachtet, für das eine analytische Lösung vorliegt. Eine unendlich ausgedehnte Scheibe mit kreisförmigem Loch wird im Unendlichen mit einer Spannung σ_∞ in x_1-Richtung auf Zug belastet (Abbildung 5.11). Dabei wird der ebene Verzerrungszustand der linearen Elastizitätstheorie zugrundegelegt. Für dieses Randwertproblem ist

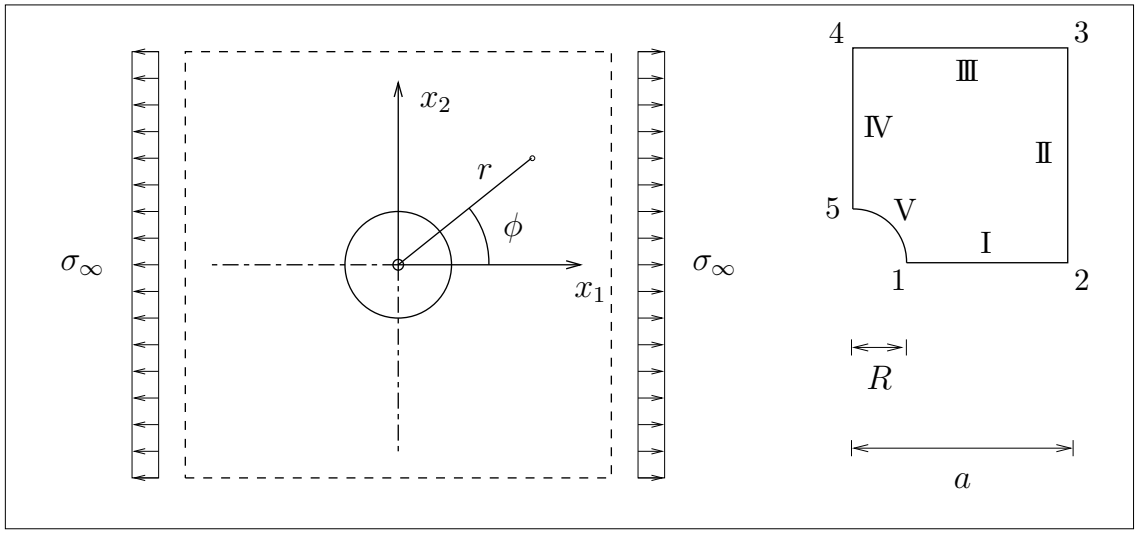


Abbildung 5.11: *Unendlich ausgedehnte Scheibe mit kreisförmigem Loch; links: Skizze des Randwertproblems, rechts: Makroknoten und Makrokanten für die Definition des Randwertproblems zur FE-Diskretisierung*

die Herleitung einer analytischen Lösung möglich (*Girkmann* [61, §59]):

$$\begin{aligned}
\sigma_{11}(r,\phi) &= \sigma_\infty \left[1 - \frac{R^2}{r^2}\left(\frac{3}{2}\cos 2\phi + \cos 4\phi\right) + \frac{3}{2}\frac{R^4}{r^4}\cos 4\phi\right], \\
\sigma_{22}(r,\phi) &= \sigma_\infty \left[- \frac{R^2}{r^2}\left(\frac{1}{2}\cos 2\phi - \cos 4\phi\right) - \frac{3}{2}\frac{R^4}{r^4}\cos 4\phi\right], \\
\sigma_{12}(r,\phi) &= \sigma_\infty \left[- \frac{R^2}{r^2}\left(\frac{1}{2}\sin 2\phi + \sin 4\phi\right) + \frac{3}{2}\frac{R^4}{r^4}\sin 4\phi\right].
\end{aligned} \tag{5.9}$$

Darin sind R der Lochradius und (r,ϕ) die ebenen Polarkoordinaten. Die analytische Lösung ist unabhängig von den gewählten Materialparametern der linearen Elastizitätstheorie, da sie auf einer *Airy*schen Spannungsfunktion als Lösung der biharmonischen Differentialgleichung der Scheibentheorie beruht.

Aus Symmetriegründen muß nur ein Viertel des Problems mit finiten Elementen diskretisiert werden (Abbildung 5.11 rechts). Außerdem wird nur ein endlicher Ausschnitt

der Länge a betrachtet, wobei an den Schnittkanten als Randbedingungen die Werte der analytischen Lösung vorgegeben werden (Freischnitt). Dadurch kann die FE-Lösung im Nahfeld direkt mit der analytischen Lösung der unendlich ausgedehnten Scheibe verglichen werden. Die *Neumann*-Randbedingungen an den Schnittkanten (Makrokanten II und III in Abbildung 5.11) lauten demnach

$$\begin{aligned} \text{Makrokante II}: \quad \bar{\mathbf{t}} &= \mathbf{T}\,\mathbf{n} = \sigma_{11}(r,\phi)\,\mathbf{e}_1 + \sigma_{12}(r,\phi)\,\mathbf{e}_2\,, \\ \text{Makrokante III}: \quad \bar{\mathbf{t}} &= \mathbf{T}\,\mathbf{n} = \sigma_{12}(r,\phi)\,\mathbf{e}_1 + \sigma_{22}(r,\phi)\,\mathbf{e}_2\,, \end{aligned} \tag{5.10}$$

wobei über die Zusammenhänge $r = \sqrt{x_1^2 + x_2^2}$, $\phi = \arctan(x_2/x_1)$ die ebenen Polarkoordinaten (r, ϕ) aus den kartesischen Koordinaten (x_1, x_2) entlang der betrachteten Kante berechnet werden. Aus Symmetriegründen wird entlang der Makrokante I die x_2-Verschiebung festgehalten und entlang der Makrokante IV die x_1-Verschiebung. Für die numerischen Rechnungen werden die in Tabelle 5.5 angegebenen Parameter gewählt.

Parameter	Symbol	Wert
Elastizitätsmodul	E	1000,0 MPa
Querkontraktionszahl	ν	0,3
Lochradius	R	1 mm
Kantenlänge des FE-Rechengebiets	a	10 mm
Spannung im Unendlichen	σ_∞	1 MPa

Tabelle 5.5: *Scheibe mit Loch: Materialparameter, Geometriedaten und Belastung*

Überprüfung der Konvergenzrate

Die *Konvergenzrate* ist ein wichtiges Maß zur Beurteilung der Konvergenzgeschwindigkeit der FEM. In der *A-priori*-Fehlerabschätzung (*Szabó & Babuška* [112])

$$|||\mathbf{u} - \mathbf{u}^h||| \leq \frac{C}{N^\beta} \tag{5.11}$$

bezeichnet C eine problemabhängige Konstante, N die Anzahl der Freiheitsgrade und β die Konvergenzrate (vgl. hierzu auch (4.40)). Bei Problemen mit hinreichend glatter Lösung wie der hier betrachteten Scheibe mit Loch erhält man in zwei Raumdimensionen bei bestmöglicher Netzverfeinerung asymptotisch die Konvergenzrate $\beta = \frac{1}{2}p$, wobei p den Polynomgrad der Ansatzfunktionen bezeichnet.

In Abbildung 5.12 sind die Ergebnisse einiger durchgeführter Rechnungen dargestellt, wobei die folgenden Parameter variiert wurden:

- linearer Ansatz (P1) $\longleftrightarrow$ quadratischer Ansatz (P2)
- uniforme Verfeinerung $\longleftrightarrow$ adaptive Verfeinerung

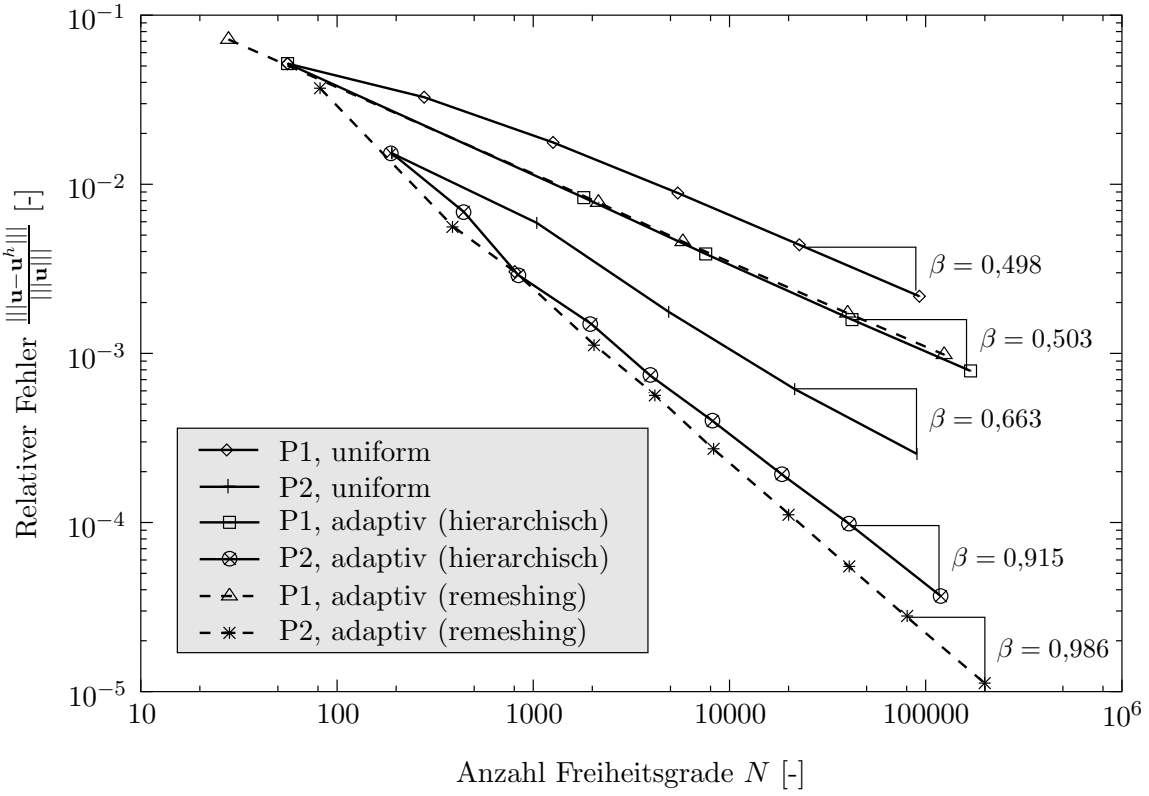


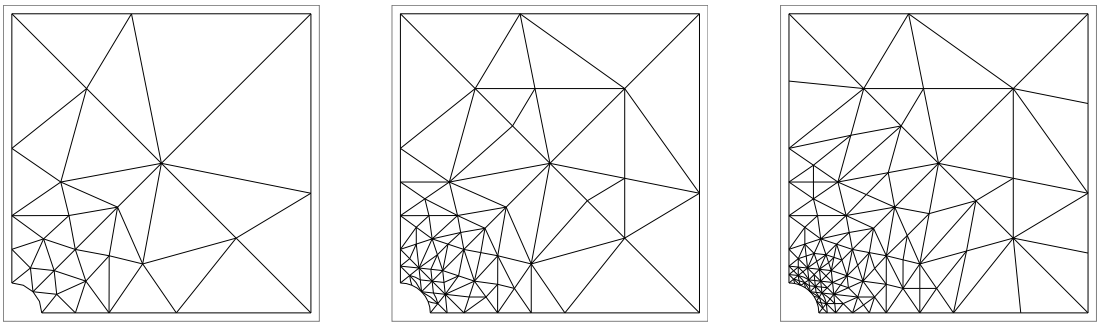
Abbildung 5.12: *Scheibe mit Loch: Konvergenzordnung; P1: Dreiecke mit linearem Ansatz, P2: Dreiecke mit quadratischem Ansatz, jeweils reguläre Verfeinerung gegenüber adaptiver hierarchischer Netzanpassung bzw. Wiedervernetzung (remeshing)*

- hierarchische Netzanpassung $\longleftrightarrow$ Wiedervernetzung

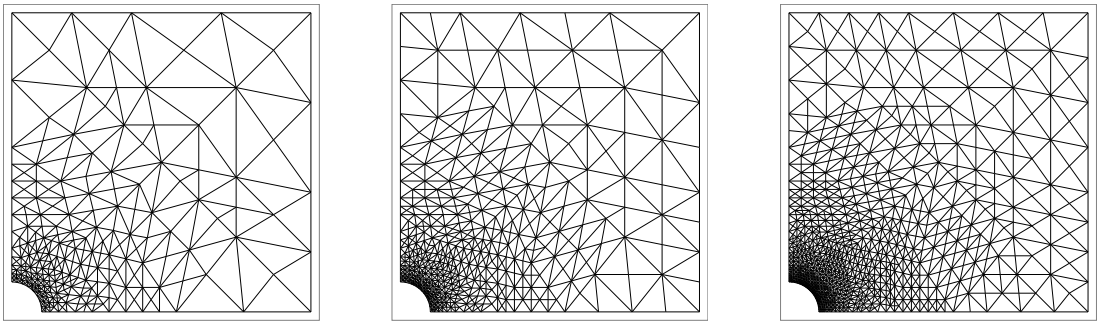
Beim linearen Ansatz wird die theoretisch vorhergesagte Konvergenzrate von $\beta = 0{,}5$ bereits durch die uniforme Verfeinerung erreicht, allerdings führt die adaptive Verfeinerung zu einer höheren Genauigkeit bei der gleichen Anzahl von Freiheitsgraden. Beim quadratischen Ansatz erhält man bei uniformer Verfeinerung eine höhere Konvergenzrate als beim linearen Ansatz, verbunden mit einer wesentlich höheren Genauigkeit. Die adaptive Verfeinerung führt dann auf die theoretisch erreichbare Konvergenzrate von $\beta = 1$. Dabei führt die Wiedervernetzungsstrategie zu einer etwas höheren Genauigkeit bei der gleichen Anzahl von Freiheitsgraden als die hierarchische Netzanpassung. Dies ist darauf zurückzuführen, daß bei einer Wiedervernetzung das Netz genauer an die gegebene Dichtefunktion angepaßt werden kann.

In Abbildung 5.13 sind einige Netze zu verschiedenen Genauigkeitsanforderungen dargestellt, die bei hierarchischer Netzanpassung entstehen. Dabei wurde von einem Startnetz mit einer Forderung von 33° für den Minimalwinkel[3] in den Dreiecken ausgegangen. Die relative Toleranz wurde Schritt für Schritt verkleinert, so daß jeweils das akzeptierte Netz zur vorherigen Toleranz als Startnetz für die nächstkleinere Toleranz diente. In Abbildung 5.14 ist für die gleichen Genauigkeitsanforderungen die Netzfolge bei Wieder-

[3]Der verwendete Netzgenerator Triangle [105] konvergiert garantiert bis zur Vorgabe eines Minimalwinkels von 33°.



(a) Startnetz: $N = 190$ (b) $\epsilon_r = 1\%$: $N = 442$ (c) $\epsilon_r = 0{,}5\%$: $N = 838$



(d) $\epsilon_r = 0{,}2\%$: $N = 1\,978$ (e) $\epsilon_r = 0{,}1\%$: $N = 3\,950$ (f) $\epsilon_r = 0{,}05\%$: $N = 8\,200$

Abbildung 5.13: *Scheibe mit Loch: Adaptiv verfeinerte Netze bei hierarchischer Netzanpassung und quadratischem Ansatz im Dreieck (P2); Variation der relativen Toleranz*

vernetzung dargestellt, wobei hier nur ein Minimalwinkel von 16,5° gefordert wurde, da keine Teilung der Dreiecke mehr erfolgt. Man erkennt, daß beide Strategien in etwa zur gleichen Anzahl von Freiheitsgraden führen. Für die adaptiven Rechnungen wurde der ZZ-Fehlerindikator verwendet und zur Berechnung der Dichtefunktion jeweils die Strategie der Minimierung der Elementanzahl im neuen Netz (Gleichung (4.46)).

Überprüfung des Effektivitätsindex

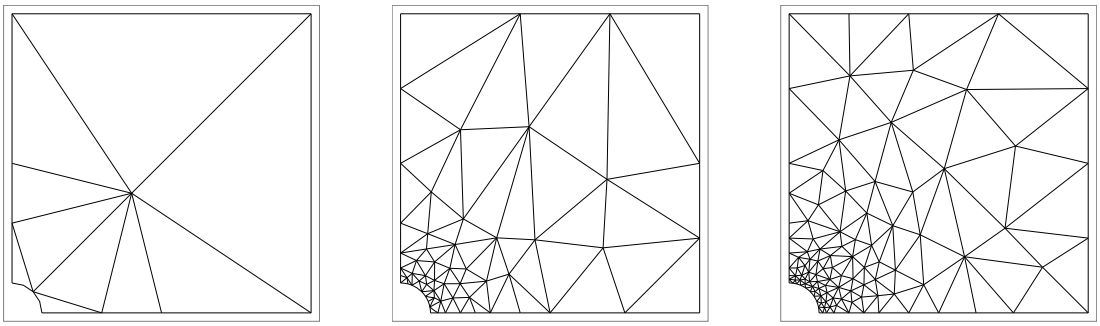
Bei der Überprüfung von *A-posteriori*-Fehlerschätzern stellt der *Effektivitätsindex*

$$\theta \;:=\; \frac{\eta}{|||\mathbf{e}|||} \tag{5.12}$$

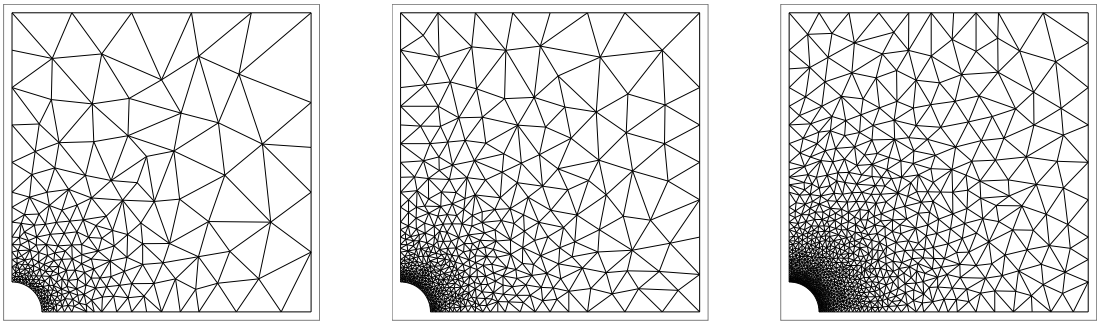
eine wichtiges Maß dar (*Szabó & Babuška* [112]). Er vergleicht den geschätzten Fehler η mit dem wahren Fehler $|||\mathbf{e}||| = |||\mathbf{u} - \mathbf{u}^h|||$ (hier in der Energienorm) und erlaubt so die Beurteilung der Güte eines Fehlerschätzers (vgl. Kapitel 4). Im Idealfall gilt $\theta = 1$, für $\theta < 1$ wird der tatsächliche Fehler unterschätzt, ansonsten überschätzt. Ein Fehlerschätzer wird als *asymptotisch exakt* bezeichnet, wenn gilt:

$$\lim_{h \to 0} \theta \;=\; 1\,, \tag{5.13}$$

d. h. für beliebig verfeinerte Netze wird der wahre Fehler erhalten.



(a) Startnetz: $N = 62$ (b) $\epsilon_r = 1\%$: $N = 408$ (c) $\epsilon_r = 0{,}5\%$: $N = 806$



(d) $\epsilon_r = 0{,}2\%$: $N = 2\,058$ (e) $\epsilon_r = 0{,}1\%$: $N = 4\,110$ (f) $\epsilon_r = 0{,}05\%$: $N = 8\,068$

Abbildung 5.14: *Scheibe mit Loch: Adaptiv verfeinerte Netze bei Wiedervernetzung und quadratischem Ansatz im Dreieck (P2); Variation der relativen Toleranz*

Nr.	ϵ_r	N	$\|\|\|\mathbf{u}^* - \mathbf{u}^h\|\|\|$	$\frac{\|\|\|\mathbf{u}^* - \mathbf{u}^h\|\|\|}{\|\|\|\mathbf{u}^h\|\|\|}$	$\|\|\|\mathbf{u} - \mathbf{u}^h\|\|\|$	$\frac{\|\|\|\mathbf{u} - \mathbf{u}^h\|\|\|}{\|\|\|\mathbf{u}\|\|\|}$	θ
1	$1 \cdot 10^{-2}$	442	$2{,}83 \cdot 10^{-3}$	$9{,}33 \cdot 10^{-3}$	$2{,}08 \cdot 10^{-3}$	$6{,}87 \cdot 10^{-3}$	1,36
2	$5 \cdot 10^{-3}$	838	$1{,}09 \cdot 10^{-3}$	$3{,}59 \cdot 10^{-3}$	$8{,}80 \cdot 10^{-4}$	$2{,}90 \cdot 10^{-3}$	1,24
3	$2 \cdot 10^{-3}$	1978	$5{,}14 \cdot 10^{-4}$	$1{,}70 \cdot 10^{-3}$	$4{,}53 \cdot 10^{-4}$	$1{,}49 \cdot 10^{-3}$	1,14
4	$1 \cdot 10^{-3}$	3950	$2{,}45 \cdot 10^{-4}$	$8{,}07 \cdot 10^{-4}$	$2{,}24 \cdot 10^{-4}$	$7{,}39 \cdot 10^{-4}$	1,09
5	$5 \cdot 10^{-4}$	8200	$1{,}26 \cdot 10^{-4}$	$4{,}15 \cdot 10^{-4}$	$1{,}21 \cdot 10^{-4}$	$3{,}99 \cdot 10^{-4}$	1,04
6	$2 \cdot 10^{-4}$	18448	$5{,}98 \cdot 10^{-5}$	$1{,}97 \cdot 10^{-4}$	$5{,}85 \cdot 10^{-5}$	$1{,}93 \cdot 10^{-4}$	1,02
7	$1 \cdot 10^{-4}$	40786	$2{,}98 \cdot 10^{-5}$	$9{,}82 \cdot 10^{-5}$	$2{,}98 \cdot 10^{-5}$	$9{,}81 \cdot 10^{-5}$	1,00
8	$5 \cdot 10^{-5}$	119568	$1{,}11 \cdot 10^{-5}$	$3{,}66 \cdot 10^{-5}$	$1{,}11 \cdot 10^{-5}$	$3{,}67 \cdot 10^{-5}$	1,00

Tabelle 5.6: *Scheibe mit Loch: Ergebnisse bei adaptiver hierarchischer Netzanpassung und quadratischem Ansatz im Dreieck (P2); geforderte Genauigkeit ϵ_r, Anzahl Freiheitsgrade N, absolute und relative geschätzte und wahre Fehler, Effektivitätsindex θ*

Bei der oben durchgeführten Rechnung mit adaptiver Netzanpassung wurden die in Tabelle 5.6 angegebenen Fehler und Effektivitätsindizes erreicht. Darin bezeichnet $\mathbf{u}$ die exakte Lösung, $\mathbf{u}^h$ die FE-Lösung und $\mathbf{u}^*$ die mit dem SPR-Verfahren geglättete Lösung des ZZ-Fehlerindikators. Man erkennt, daß der Effektivitätsindex von oben gegen $\theta = 1$

konvergiert, so daß der ZZ-Fehlerindikator den Fehler bei zu grober Diskretisierung eher überschätzt. Dies ist im Hinblick auf die Zuverlässigkeit des adaptiven Verfahrens eine sehr wünschenswerte Eigenschaft.

Insgesamt konnte mit den Ergebnissen dieses Rechenbeispiels gezeigt werden, daß die entwickelten ortsadaptiven Verfahren robust und zuverlässig arbeiten, und daß die theoretischen Konvergenzraten erreicht werden. Der Effektivitätsindex des ZZ-Fehlerindikators strebt von oben gegen Eins, so daß eine vom Benutzer vorgegebene Genauigkeit stets eingehalten wird.

Es zeigte sich außerdem, daß die Ergebnisse bei Verwendung der hierarchischen Netzanpassung und der Wiedervernetzung qualitativ gleichwertig sind. Der Aufwand zur Erstellung des neuen Netzes und dem bei zeitabhängigen Problemen notwendigen Datentransfer ist jedoch im Fall der Wiedervernetzungsmethode größer als bei der hierarchischen Netzanpassung. Daher wird in den folgenden Beispielen stets die Methode der hierarchischen Netzanpassung verwendet.

5.3 Verifikation Zeit- und Ortsadaptivität

5.3.1 Biaxialversuch

Anhand eines Biaxialversuchs wird die Wirksamkeit des entwickelten zeit- und ortsadaptiven Verfahrens demonstriert. Beim Biaxialversuch wird eine als starr angenommene Platte verschiebungsgesteuert abgesenkt, bei gleichzeitiger Stabilisierung der Probe durch Aufbringen einer konstanten Seitenspannung (Abbildung 5.15). Die Materialparameter

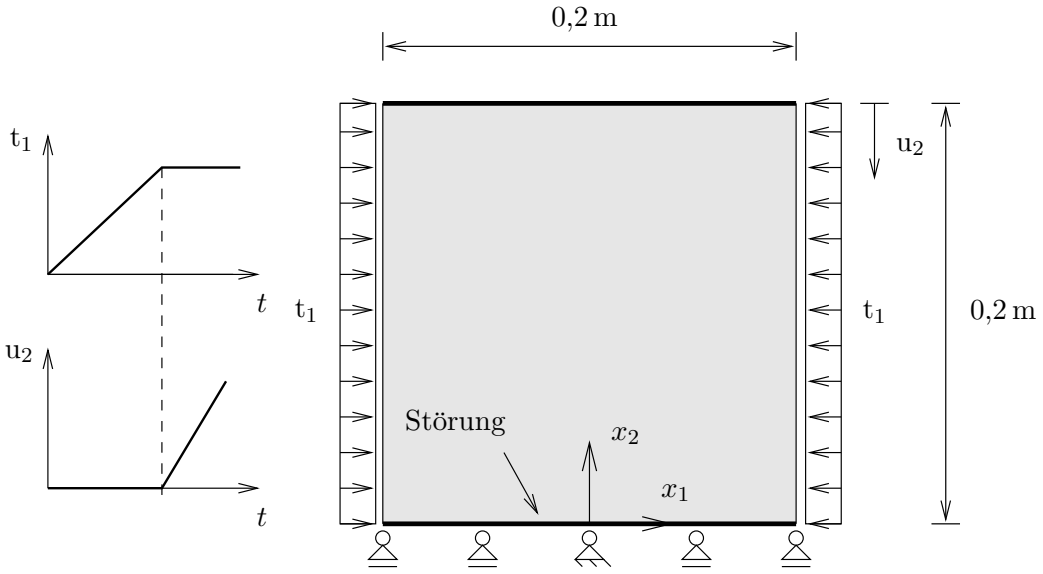


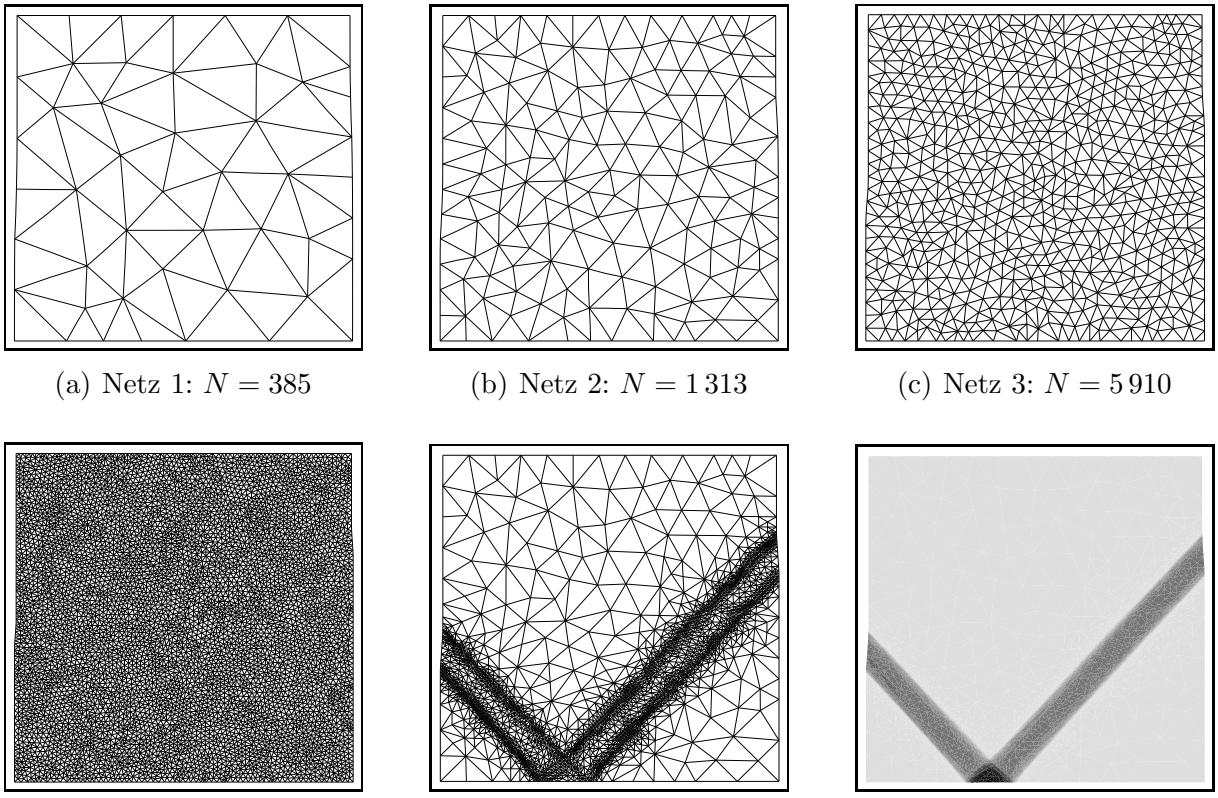
Abbildung 5.15: *Biaxialversuch: Anfangs-Randwertproblem*

des modellierten Tons können Anhang B entnommen werden; hier wird jedoch davon abweichend ein leeres Skelett mit der Relaxationszeit $\eta = 4\,\mathrm{s}$ im elastisch-viskoplastischen Modell betrachtet. Zur Initiierung eines Scherbandes wird in der linken Hälfte des unteren Randes eine Störung der Materialparameter aufgebracht (Imperfektion). Diese Störung muß in Form einer glatten Funktion vorgegeben werden, da ansonsten bei der adaptiven Rechnung die Gebietsgrenze zwischen gestörtem und ungestörtem Materialverhalten beliebig fein aufgelöst würde (Sprung in den Koeffizienten der Differentialgleichung). Hier werden die elastischen Materialparameter μ^S und λ^S mit der ortsabhängigen Funktion

$$f(x_1, x_2;\ x_1^{(0)}, x_2^{(0)}, \delta, s) = 1 + \delta\ \exp\left[-s\left((x_1 - x_1^{(0)})^2 + (x_2 - x_2^{(0)})^2\right)\right] \tag{5.14}$$

multipliziert, so daß die Störung ihr Maximum im Punkt $(x_1^{(0)}; x_2^{(0)})$ annimmt und in alle Richtungen exponentiell abklingt. Im vorliegenden Beispiel wurden die Störungsparameter $\delta = 0{,}05$, $s = 20000$ und $(x_1^{(0)}; x_2^{(0)}) = (-0{,}025; 0)$ gewählt.

Alle Rechnungen wurden mit einer Seitenspannung $\mathrm{t}_1 = 100\,\mathrm{kN/m^2}$ und einer Belastungsgeschwindigkeit $\dot{\mathrm{u}}_2 = 0{,}0002\,\mathrm{m/s}$ bis zu einer Gesamtverschiebung $\mathrm{u}_2 = 0{,}0032\,\mathrm{m}$ durchgeführt ($t = 16\,\mathrm{s}$). Dies entspricht einer durchschnittlichen Verzerrung in vertikaler Richtung von 1,6%, wonach eine geometrisch lineare Rechnung gerechtfertigt erscheint.



(a) Netz 1: $N = 385$ (b) Netz 2: $N = 1\,313$ (c) Netz 3: $N = 5\,910$

(d) Netz 4: $N = 57\,682$ (e) Netz adaptiv (f) Ergebnis adaptiv

Abbildung 5.16: *Biaxialversuch: (a)-(d): gleichmäßig verfeinerte Netze; (e): adaptiv verfeinertes Netz mit $N = 25\,061$ Freiheitsgraden (Startnetz: Netz 2), (f): Scherband-Ergebnis: kleine (hell) bis große (dunkel) plastische Verzerrungen*

Mit dem Startnetz in Abbildung 5.16b ergibt sich bei adaptiver Rechnung ($\epsilon_r = 1\%$, $\epsilon_{a,i} = 0$, $\alpha_1 = 0{,}5$, $\alpha_2 = 0{,}5$, $\alpha_3 = 0$, keine Beschränkung der Elementanzahl) zum Schluß das Netz in Abbildung 5.16e. Abbildung 5.16f zeigt das entstandene Scherband. Man erkennt sehr gut, daß insbesondere die Ränder des Scherbandes verfeinert werden, da hier die größten Gradienten in den Feldgrößen auftreten.

Mit den Netzen aus Abbildung 5.16 wurden Last-Verschiebungskurven aufgenommen (Abbildung 5.17). Man erkennt die Konvergenz der FEM bei gleichmäßiger Verfeinerung (Netze 1-4), was aufgrund der Regularisierung durch die Viskoplastizität zu erwarten war. Die Lösung der adaptiven Rechnung zeigt eine sehr gute Übereinstimmung mit der Lösung auf Netz 4. Hierzu sei noch erwähnt, daß die Elemente in Netz 4 im Bereich des Scherbands immer noch wesentlich größer sind als die Elemente im adaptiv verfeinerten Netz, was die verbleibenden Unterschiede in den Last-Verschiebungskurven erklärt.

Der Aufwand für die Rechnung mit Netz 4 betrug auf einer Silicon Graphics Power Challenge R10000/195 insgesamt 177 h (CPU-Zeit), der für die adaptive Berechnung nur 21 h. Man erhält also durch Verwendung des zeit- und ortsadaptiven Verfahrens in diesem Fall eine Ersparnis von 88% der Rechenzeit.

Die Datenpunkte auf der Last-Verschiebungskurve der adaptiven Rechnung verdeutli-

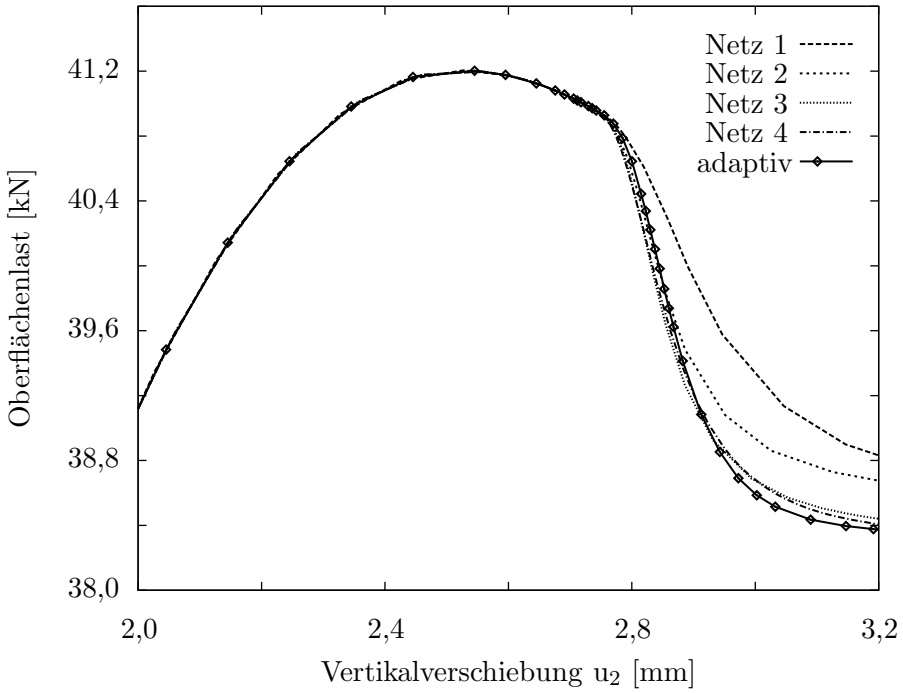


Abbildung 5.17: *Biaxialversuch: Last-Verschiebungskurven*

chen zusätzlich die Wirksamkeit der Zeitadaptivität: Im elastischen Bereich (links) wird mit der erlaubten Maximalschrittweite gerechnet; bei einsetzender Lokalisierung (Maximum der Kurve) wird die Schrittweite automatisch drastisch verkleinert (ohne Schrittweitensteuerung werden u. U. völlig falsche Lösungen berechnet, vgl. *Diebels, Ellsiepen & Ehlers* [41, 42]); im post-kritischen Bereich (Abfall der Kurve) wird die Schrittweite durch die automatische Schrittweitensteuerung so gewählt, daß die Abstände auf der Last-Verschiebungskurve in etwa gleich bleiben (Erfassung der Nichtlinearitäten des Problems).

5.4 Anwendungsbeispiele

5.4.1 Böschungsbruch

Aus der klassischen Literatur (z. B. *Terzaghi & Jelinek* [113]) ist bekannt, daß jede Böschung in Abhängigkeit von Böschungshöhe und Böschungswinkel bereits durch das Eigengewicht des Bodens versagen kann (Abbildung 5.18). Dieses Phänomen tritt beispielsweise bei einem Aushebungsprozeß auf. Die zeitabhängigen Randbedingungen der Böschung werden in der Simulation so gewählt, daß zu Beginn das komplette Gewicht des Aushebungsgebiets auf die Böschung wirkt (kein Aushub) und nach und nach bis zur völligen Entlastung der Böschung abgebaut wird (vollständiger Aushub). Mit Erreichen einer kritischen Höhe setzt der Versagensprozeß ausgehend vom Fuß der Böschung ein. Die Ausbreitungsgeschwindigkeit der Versagenszone wird durch die Permeabilität des Bodens bestimmt.

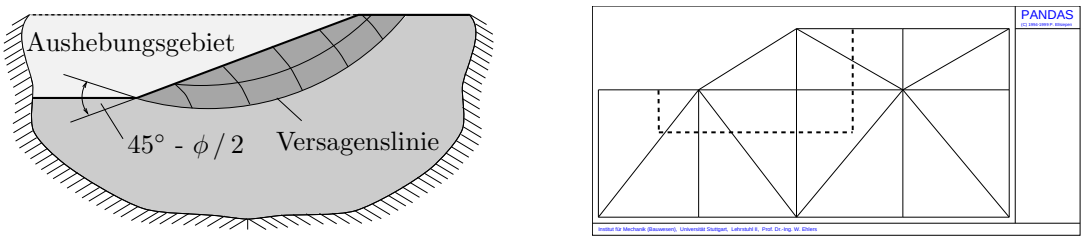


Abbildung 5.18: *Böschungsbruch infolge eines Aushebungsprozesses: links: Problemstellung mit theoretischer Gleitlinie abhängig vom Winkel der inneren Reibung ϕ; rechts: Startnetz der adaptiven Berechnung, Ausschnittsmarkierung*

In Abbildung 5.19 ist der Versagenszustand (Scherbandbildung, vgl. *Ehlers & Volk* [52]) 100 Tage nach Erreichen der kritischen Höhe bei einer Permeabilität von $k^F = 1{,}2{\cdot}10^{-7}\,m/s$ abgebildet (Ausschnitt). Es ist deutlich erkennbar, daß das Scherband mit Hilfe der zeit- und ortsadaptiven Strategien in guter Übereinstimmung zur klassischen Theorie prognostiziert wird.

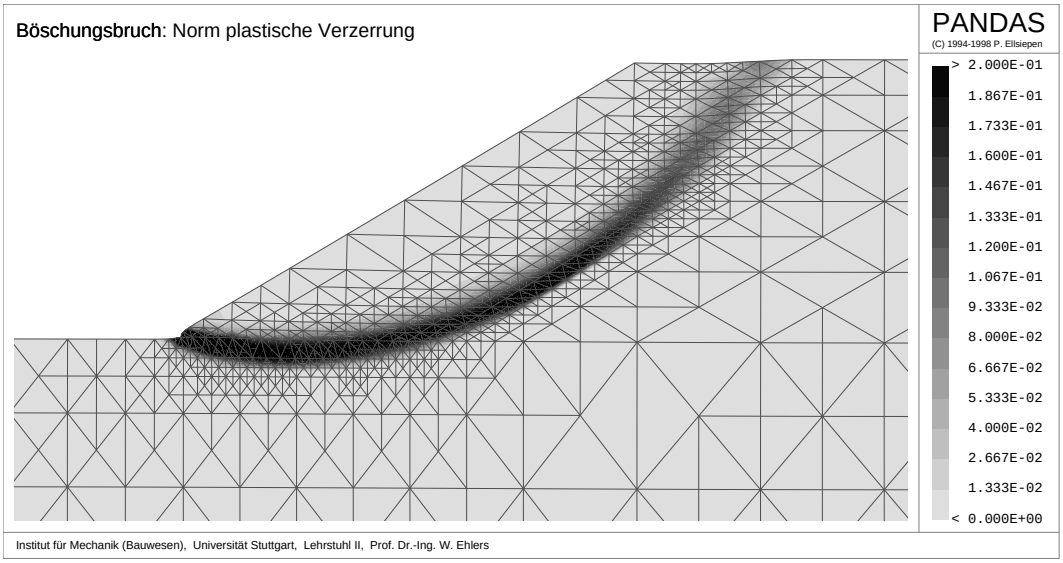


Abbildung 5.19: *Versagenszustand (Ausschnitt; 4 163 Knoten, 2 040 Elemente, 9 388 Freiheitsgrade, 30 600 interne Variablen)*

5.4.2 Grundbruch

Als zweites Anwendungsbeispiel wird das Grundbruchproblem vorgestellt, das für die Geotechnik von großem Interesse ist. Dabei wird eine starre Platte mit konstanter Geschwindigkeit in einen Halbraum aus schluffigem Boden gedrückt. In Abbildung 5.20 ist links die klassische Lösung basierend auf der *Rankine*schen Gleitlinientheorie abgebildet (vgl. *Terzaghi & Jelinek* [113]). In der Simulation wird die Symmetrie des Problems ausgenutzt.

Abbildung 5.21 zeigt bereits, daß die Scherbänder am Rand der Platte beginnen, der durch

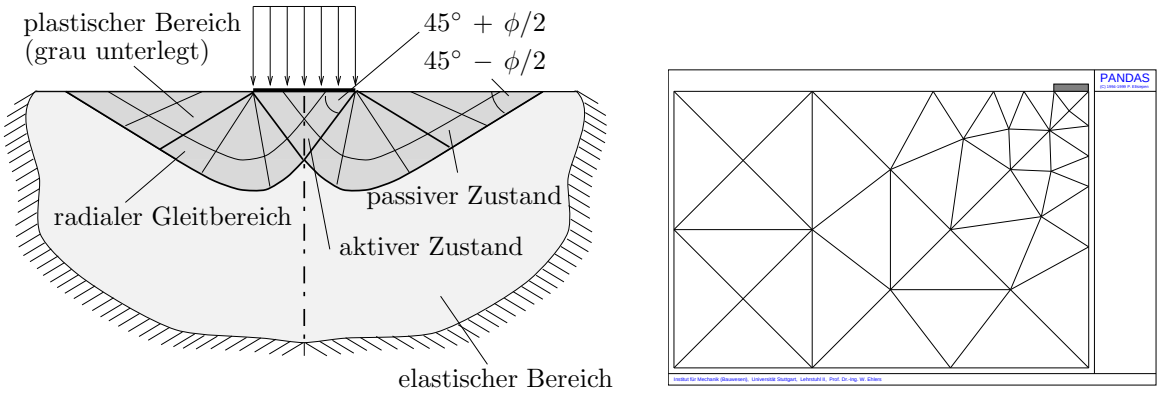


Abbildung 5.20: *Das klassische Grundbruchproblem: links: Problemstellung mit Gleitlinienfeld abhängig vom Winkel der inneren Reibung ϕ sowie Bereichseinteilung; rechts: Startnetz der adaptiven Berechnung (symmetrisches Problem)*

eine Unstetigkeit in den Randbedingungen gekennzeichnet ist. Eine weitere Absenkung der Platte führt zur Ausbildung eines dominanten Scherbands, dessen Verlauf qualitativ mit dem Ergebnis aus der klassischen Gleitlinientheorie übereinstimmt (Abbildung 5.22). Das adaptiv verfeinerte Netz dieses Zustands ist in Bild Abbildung 5.23 dargestellt.

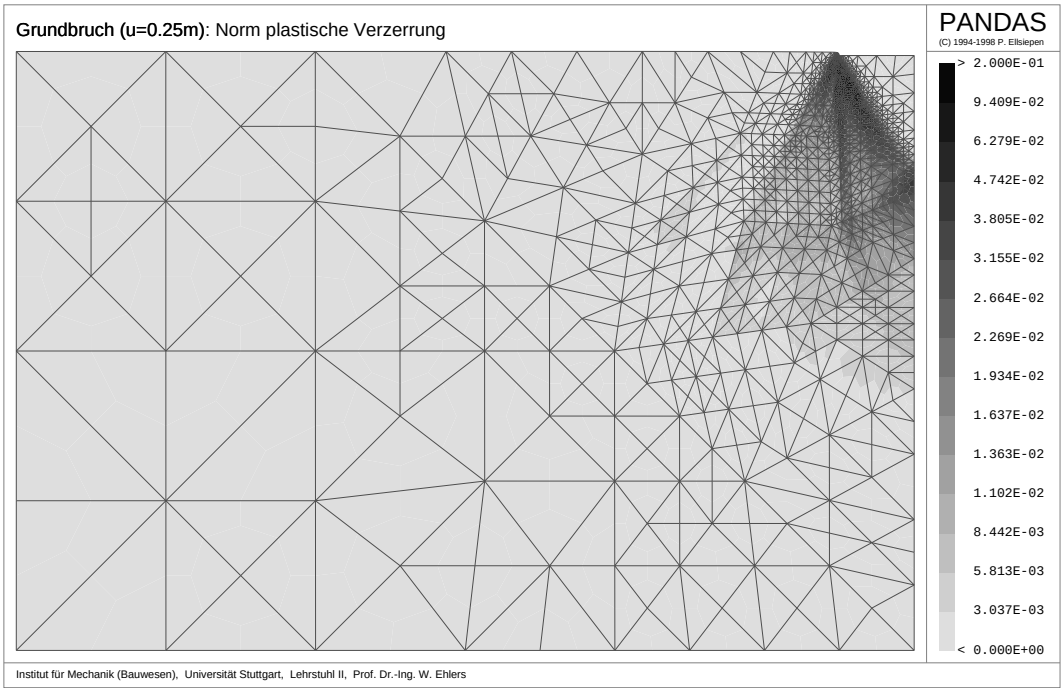


Abbildung 5.21: *Einsetzende Scherbandbildung nach Absenkung der Platte um 0,25 m (3 540 Knoten, 1 735 Elemente, 7 983 Freiheitsgrade, 26 025 interne Variablen)*

An dieser Stelle sei auf eine Problematik bei der Berechnung des Grundbruchproblems hingewiesen. Aufgrund der Unstetigkeit in den Randbedingungen am Plattenrand erhält

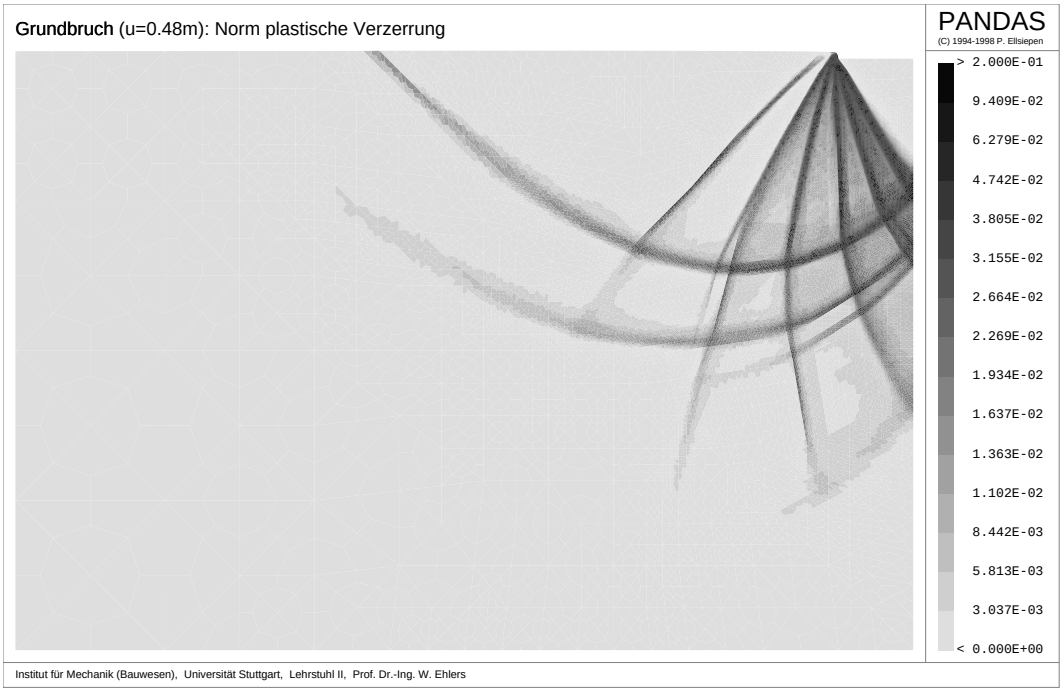


Abbildung 5.22: *Scherbänder nach Absenkung der Platte um 0,48 m*

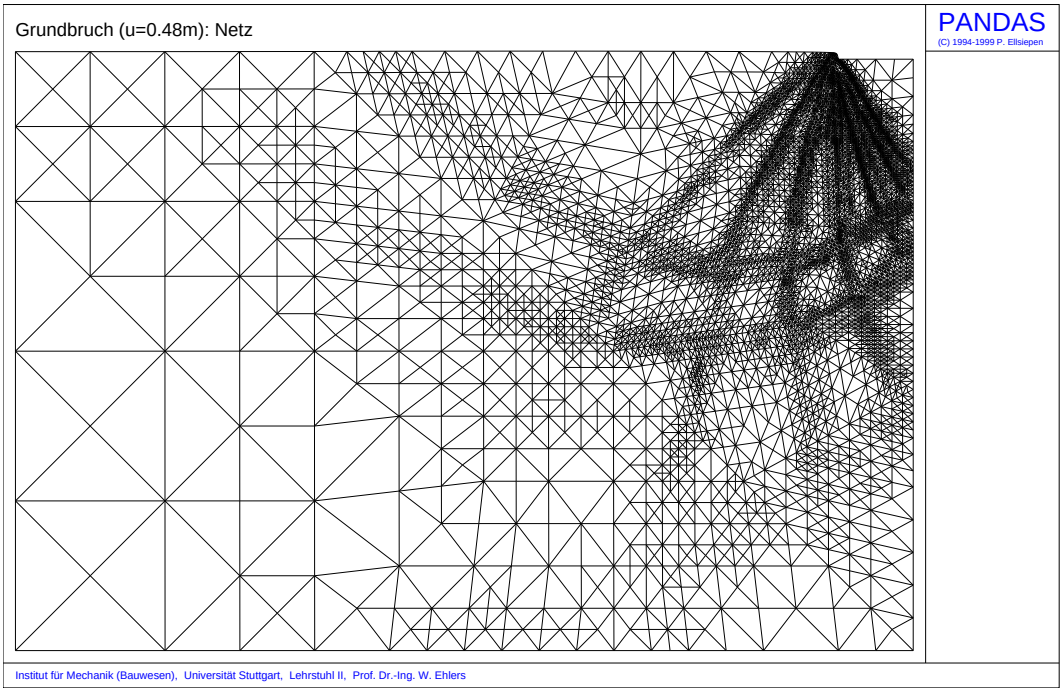


Abbildung 5.23: *Netz nach Absenkung der Platte um 0,48 m (30 643 Knoten, 15 254 Elemente, 68 981 Freiheitsgrade, 228 810 interne Variablen)*

man eine Singularität im Verschiebungsfeld, die von der ortsadaptiven Strategie ohne Begrenzung der Elementzahl beliebig fein aufgelöst würde. Bei der vorliegenden numerischen

Simulation wurde daher eine minimale Elementgröße vorgegeben, so daß eine numerische Lösung mit vertretbarem Aufwand berechnet werden konnte.

Das Grundbruchproblem stellt aus mathematischer Sicht ein schwieriges Anfangs-Randwertproblem dar, bei dem sehr komplexe Phänomene auftreten, wie in diesem Fall die von der Singularität ausgehenden Scherbänder. Es stehen daher i. a. keine theoretischen Aussagen über die Existenz und Eindeutigkeit von Lösungen zur Verfügung. Die entwickelten numerischen Methoden erlauben jedoch die Simulation derartiger Anfangs-Randwertprobleme ohne vorherige Kenntnis theoretischer Aussagen, wobei die Anwendung der Methoden durch die guten Ergebnisse bei den Verifikationsbeispielen der vorangegangenen Abschnitte motiviert ist. Eine Verifikation der berechneten numerischen Lösungen kann bei praxisnahen Anwendungsbeispielen i. a. nur durch Vergleich mit Experimenten oder durch Vergleich mit klassischen Theorien, wie der o. g. *Rankine*schen Gleitlinientheorie im Fall des Grundbruchproblems, erfolgen.

Als Schlußfolgerung aus den Ergebnissen der dargestellten Anwendungsbeispiele kann daher festgehalten werden, daß die entwickelten zeit- und ortsadaptiven Verfahren die numerische Simulation komplexer, praxisnaher Anfangs-Randwertprobleme erlauben, bei denen häufig keine theoretischen Aussagen über die Existenz und Eindeutigkeit von Lösungen existieren. Wie üblich bei praxisnahen Aufgabenstellungen sind in diesem Fall jedoch Vergleiche mit Experimenten oder Ergebnissen klassischer Theorien zur Verifikation der erhaltenen Ergebnisse unerläßlich.

Zusammenfassung und Ausblick

Zusammenfassung

Im Rahmen der vorliegenden Arbeit wurde ein zeit- und ortsadaptives Verfahren entwickelt und auf Mehrphasenprobleme poröser Medien angewandt. Die dabei erarbeiteten Konzepte fanden Einzug in das vom Verfasser konzipierte und implementierte FE-Programmsystem PANDAS, das am Lehrstuhl II des Instituts für Mechanik (Bauwesen) der Universität Stuttgart als Grundlage für Weiterentwicklungen auf diesem Gebiet dient. Darüber hinaus erlaubt die entwickelte Methodik die Einbeziehung von Kontinuumsmodellen aus der Elastizitäts- und Plastizitätstheorie einphasiger Materialien sowie von Fragestellungen aus der Fluidmechanik.

Die Grundlage der Arbeit bildete die Darstellung der *Theorie Poröser Medien* (TPM), die zur Modellierung von Mehrphasenmaterialien verwendet werden kann. Das daraus abgeleitete inkompressible, viskoplastische Zweiphasenmodell in einer geometrisch linearen Formulierung diente im weiteren als Beispiel zur Darstellung der prinzipiellen Vorgehensweise bei der Entwicklung des zeit- und ortsadaptiven Gesamtverfahrens.

Eine konsequente mathematische Notation der auftretenden Anfangs-Randwertprobleme, sowohl in starker als auch in schwacher Formulierung, und die darauf aufbauende abstrakte Darstellung der Semidiskretisierung im Ort mit der Methode der finiten Elemente (FEM) lieferte ein System differential-algebraischer Gleichungen (DAE) in der Zeit. Die so erhaltene Formulierung der ortsdiskreten Gleichungen eröffnete den Weg zur Anwendung moderner Zeitintegrationsverfahren. Da die Struktur der entstehenden DAE-Systeme unabhängig vom gewählten Kontinuums- und Materialmodell ist (ein- bzw. mehrphasig, quasi-statisch bzw. dynamisch, elastisch, viskoelastisch, viskoplastisch bzw. elastoplastisch, geometrisch linear bzw. nichtlinear), wurde auf diese Weise eine breite Basis für weitere Anwendungen gelegt, vgl. *Eipper* [53], *Mahnkopf* [81], *Markert* [82], *Volk* [121].

Aufgrund theoretisch und praktisch motivierter Kriterien konnte im folgenden herausgearbeitet werden, daß sich diagonal-implizite *Runge-Kutta*-Verfahren (DIRK) besonders für die behandelte Problemklasse eignen. Die Diskussion verschiedener Stabilitätsbegriffe machte deutlich, daß aus theoretischer Sicht weitere Forderungen an die Verfahren gestellt werden müssen. Hier sei etwa auf die Bedeutung steif genauer Verfahren für DAE und die Forderung der L-Stabilität bei Plastizitätsproblemen verwiesen, deren Wichtigkeit sich im Rahmen der numerischen Beispielrechnungen bestätigte.

Das implizite *Euler*-Verfahren als Standard-Verfahren bei der numerischen Simulation von Plastizitätsproblemen fällt ebenfalls in die Klasse der DIRK-Verfahren, so daß ein direkter Vergleich mit DIRK-Verfahren höherer Ordnung durchgeführt werden konnte. Es zeigte sich, daß der Einsatz von Verfahren höherer Ordnung trotz des größeren Aufwands je Zeitschritt einen enormen Effizienzgewinn zur Folge hat; so konnte das Erreichen der höheren Ordnung in der Praxis selbst im Fall eines Benchmark-Problems auf Basis der idealen *Prandtl-Reuß*-Plastizität nachgewiesen werden, die als Grenzfall komplexerer

Plastizitätsmodelle einen für die numerische Lösung schwierigen Testfall darstellt.

Darüber hinaus bietet die höhere Ordnung die Möglichkeit zur Einbettung eines Verfahrens niedrigerer Ordnung, das zu einer sehr effizienten Schätzung des lokalen Zeitfehlers herangezogen werden kann. Durch Steuerung der Schrittweite auf Basis dieser Schätzung (Zeitadaptivität) kann auch bei komplexeren Problemen ohne vorherige Kenntnis des Lösungsverlaufs eine vorgegebene Genauigkeit erzielt werden. Diese neue Qualität bei der numerischen Simulation von Plastizitätsproblemen stellt damit neben der Erhöhung der Genauigkeit durch Verwendung von Verfahren höherer Ordnung den entscheidenden Vorteil der zeitadaptiven Verfahren dar. In dieser Hinsicht ist insbesondere das neu konstruierte eingebettete SDIRK-2(1)-Verfahren auf Basis des Verfahrens zweiter Ordnung von *Alexander* [5] zu nennen, das dem impliziten *Euler*-Verfahren bei geringem zusätzlichem Speicherbedarf und Programmieraufwand weit überlegen ist.

In bezug auf die Effizienz von mehrstufigen *Runge-Kutta*-Verfahren kommt der Lösung der nichtlinearen Gleichungssysteme eine wesentliche Bedeutung zu. Hier lieferte die Übertragung von Ideen aus dem Bereich der numerischen Mechanik (algorithmisch konsistente Linearisierung) auf die betrachteten DIRK-Verfahren einen stabilen und schnellen Lösungsalgorithmus, der die Verwendung von üblichen linearen Gleichungslösern für schwach besetzte Systeme erlaubt.

Neben der adaptiven Zeitdiskretisierung ist die adaptive Ortsdiskretisierung insbesondere bei Lokalisierungs-Phänomenen – etwa beim Scherbandproblem in Böden – von entscheidender Bedeutung für die Effizienz des Gesamtverfahrens. In dieser Hinsicht wurde zunächst durch eine vergleichende Gegenüberstellung verschiedener Techniken zur Gewinnung von Fehlerschätzern und Fehlerindikatoren herausgearbeitet, daß gradienten-basierte Fehlerindikatoren im Grenzfall einfacher Modellprobleme gleichwertige Ergebnisse liefern wie mathematisch fundierte Fehlerschätzer; darüber hinaus zeichnen sich diese Fehlerindikatoren dadurch aus, daß sie sehr flexibel auf unterschiedliche Problemklassen angewandt werden können. Diese Ergebnisse stellten die Grundlage zur anschließenden Konstruktion eines neuen gradienten-basierten Fehlerindikators dar, der alle treibenden Größen des behandelten Mehrphasenproblems erfaßt (Festkörper-Elastizität und -Plastizität sowie Fluidströmung) und damit eine durch vorgegebene Toleranzen und Parameter gesteuerte adaptive Ortsdiskretisierung ermöglicht. Anhand eines linear elastischen Modellproblems konnte gezeigt werden, daß der Effektivitätsindex des Fehlerindikators von oben gegen Eins strebt. Dies bedeutet, daß der wahre Fehler bei zu grober Diskretisierung eher überschätzt wird, was im Hinblick auf die Zuverlässigkeit des adaptiven Verfahrens eine sehr wünschenswerte Eigenschaft darstellt.

Zur Umsetzung der Information aus dem Fehlerindikator in ein neues Netz wurden einerseits verschiedene Dichtefunktionen diskutiert und bewertet sowie andererseits zwei Verfahren der Netzgenerierung einander gegenübergestellt. Im Rahmen einer vom Verfasser betreuten Diplomarbeit (*Ammann* [6]) zeigte sich, daß die Dichtefunktion mit dem Ziel der Minimierung der Elementanzahl im neuen Netz zur geringsten Anzahl von Freiheitsgraden bei gleichzeitiger Einhaltung der geforderten Toleranzen führt. Bei der Netzgenerierung lieferten sowohl das hierarchische Verfahren mit Verfeinerung und Vergröberung als auch die Wiedervernetzung qualitativ gleichwertige Ergebnisse.

Zur Verifikation des gekoppelten zeit- und ortsadaptiven Verfahrens diente ein Biaxialversuch, bei dem durch gezielte Störung der Materialparameter ein Scherband initiiert wurde. Es zeigte sich, daß die auftretenden Scherbänder durch den neuen Fehlerindikator zuverlässig lokalisiert werden und daß die gezielte Netzverfeinerung an den Scherbandrändern zu einer effizienten adaptiven Ortsdiskretisierung führt. Durch Vergleich von Last-Verschiebungskurven mit einer numerisch berechneten Referenzlösung konnte zudem nachgewiesen werden, daß das gekoppelte zeit- und ortsadaptive Gesamtverfahren zuverlässige Ergebnisse liefert.

Den Abschluß der Arbeit bildete die numerische Simulation praxisnaher Anwendungsbeispiele aus der Bodenmechanik (Böschungsbruch- und Grundbruchproblem). Hierbei zeigte sich eine qualitativ gute Übereinstimmung der berechneten numerischen Lösungen mit klassischen Theorien der Bodenmechanik wie der *Rankine*schen Gleitlinientheorie.

Rückblickend läßt sich feststellen, daß die Kombination von Ideen aus dem Bereich der numerischen Mathematik – etwa den Konzepten zur Schrittweitensteuerung im Zeitbereich – sowie dem Bereich der numerischen Mechanik – etwa bei der Lösung der großen nichtlinearen Gleichungssysteme – wesentlich zur erfolgreichen Konstruktion eines effizienten Gesamtverfahrens beigetragen hat.

Ausblick

In Erweiterung der im Rahmen dieser Arbeit behandelten Mehrphasenprobleme eröffnet die entwickelte Methodik bei der Darstellung der ortsdiskreten Systeme die Möglichkeit, bestehende Formulierungen nahezu beliebiger Plastizitätsmodelle mit geringem Umstellungsaufwand den entwickelten zeitadaptiven Methoden zugänglich zu machen. Dank der entwickelten Systematik ist der hierbei anfallende Programmieraufwand gering, da lediglich die Bilanzgleichungen in schwacher Formulierung und die Materialgesetze nebst deren Linearisierungen implementiert werden müssen. Diese Vorgehensweise wurde in der vorliegenden Arbeit bereits für die *Prandtl-Reuß*-Plastizität angewandt, deren Implementierung die Grundlage für die numerische Berechnung des DFG-Benchmark-Problems darstellte, das als Verifikationsbeispiel für die zeitadaptiven Verfahren diente.

Die Übertragung der ortsadaptiven Verfahren auf verwandte Probleme erfordert die Bereitstellung eines geeigneten Fehlerschätzers bzw. Fehlerindikators. Bei komplexen Problemen kann dies beispielsweise – wie im Fall des behandelten Mehrphasenproblems – durch die physikalisch motivierte Auswahl treibender Größen geschehen.

Aus mathematischer Sicht stellt die genauere Untersuchung der vorgestellten Verfahren im Hinblick auf Probleme mit Lösungs-Singularitäten (z. B. Grundbruch) ein wünschenswertes Ziel dar. Das bessere Verständnis von ausgewählten Anfangs-Randwertproblemen könnte in dieser Hinsicht Aufschluß über Möglichkeiten zur Verbesserung der Verfahren geben, so daß zukünftig auch bei komplexeren Fragestellungen gesicherte Aussagen über die Güte der numerisch erzielten Ergebnisse ermöglicht würden.

Für die numerische Praxis ist die Einbindung moderner, schneller Gleichungslöser (z. B. Mehrgitterverfahren) das vorrangige Ziel, da der Aufwand des Gesamtverfahrens von der Effizienz des linearen Gleichungslösers dominiert wird. Hierbei muß jedoch zunächst überprüft werden, ob diese Verfahren bei den behandelten Problemen anwendbar sind. Eine mögliche Stoßrichtung wäre die Untersuchung algebraischer Mehrgitterverfahren, die bereits bei Problemen auf Basis der *Prandtl-Reuß*-Plastizität erfolgreich eingesetzt wurden (*Wieners* [122]). Im Rahmen dieser Arbeit wurden sowohl direkte Löser für schwach besetzte Matrizen (z. B. Profil-Löser, *Hoit & Wilson* [72], *Zienkiewicz & Taylor* [127, §16]) als auch iterative Löser eingesetzt (z. B. GMRES[4] mit ILU[5]-Vorkonditionierung, *Saad* [102]). Damit ist jedoch die in einer akzeptablen Rechenzeit behandelbare Problemgröße bei heutigen Workstations auf rund 100 000 Freiheitsgrade beschränkt. Insbesondere bei Erweiterung der Aufgabenstellung auf den dreidimensionalen Fall wird daher die Effizienz des linearen Gleichungslösers zum alles entscheidenden Faktor.

[4] iterative *Krylov*-Unterraum-Methode (engl.: Generalized Minimal RESidual)
[5] unvollständige Dreiecks-Zerlegung (engl.: Incomplete LU)

Anhang A: Differentialgeometrische Notation und Basisdarstellung

Die Notation der klassischen Tensorrechnung, wie sie im Rahmen dieser Arbeit verwendet wurde, wird im folgenden mit einer differentialgeometrisch motivierten Notation verglichen, wie man sie in mathematisch orientierten Arbeiten zur Tensorrechnung und Kontinuumsmechanik findet, z. B. *Abraham, Marsden & Ratiu* [1], *Marsden & Hughes* [83], *Fritzen* [57].

In der Mathematik werden Tensoren häufig als lineare Abbildungen zwischen linearen Vektorräumen aufgefaßt[1]. Daher müssen zunächst einige Begriffe aus der Differentialgeometrie und der linearen Algebra eingeführt werden, was an dieser Stelle jedoch nur informell geschieht. Für eine mathematisch exakte Definition sei auf die o. g. Arbeiten verwiesen.

Def. A.1: Der *Tangentialraum* $T_{\mathbf{x}}\Omega$ einer Menge Ω in einem Punkt $\mathbf{x} \in \Omega$ ist der Vektorraum aller Tangentenvektoren an die Menge Ω in diesem Punkt. Die Elemente des Tangentialraums heißen *Vektoren*.

Der Dualraum eines Tangentialraums heißt *Kotangentialraum* und wird mit $T^*_{\mathbf{x}}\Omega$ bezeichnet[2]. Die Elemente des Kotangentialraums heißen *Kovektoren*, *Einsformen* oder *Linearformen*.

Das *Tangentialbündel* $T\Omega$ einer Menge Ω ist die Menge aller Paare $(\mathbf{x}, T_{\mathbf{x}}\Omega)$ aus Punkten mit ihren jeweiligen Tangentialräumen. Entsprechend ist das *Kotangentialbündel* $T^*\Omega$ einer Menge Ω die Menge aller Paare $(\mathbf{x}, T^*_{\mathbf{x}}\Omega)$.

Ein *Vektorfeld* auf einer Menge Ω ist eine Abbildung $\mathbf{w} : \Omega \longrightarrow T\Omega$, die jedem Punkt $\mathbf{x} \in \Omega$ das Punkt-Vektor-Paar $(\mathbf{x}, \mathbf{w}(\mathbf{x}))$ aus dem Tangentialbündel $T\Omega$ zuordnet[3].

Ein Vektorfeld kann auch in Koordinaten einer Konfiguration parametrisiert sein, aber in das Tangentialbündel einer anderen Konfiguration abbilden. Ist ein Vektorfeld z. B. in Koordinaten der Referenzkonfiguration parametrisiert und bildet in das Tangentialbündel der Momentankonfiguration ab[4], $\mathbf{w} : \Omega_0 \longrightarrow T\Omega_t$, so wird dem Punkt $\mathbf{X} \in \Omega_0$ der Referenzkonfiguration das Paar $(\boldsymbol{\chi}_t(\mathbf{X}), \mathbf{w}(\mathbf{X})) = (\mathbf{x}, \mathbf{w}(\mathbf{X}))$ aus dem Tangentialbündel der Momentankonfiguration zugeordnet.

Ein zweistufiger *Tensor* ist eine lineare Abbildung zwischen Tangential- bzw. Kotangentialräumen, ein zweistufiges *Tensorfeld* eine lineare Abbildung zwischen Tangential- bzw.

[1] Allgemein ist ein Tensor eine multilineare reellwertige Abbildung. Mit diesem Begriff des Tensors kann ein Skalar als ein Tensor 0-ter Stufe, ein Vektor als ein Tensor 1-ter Stufe usw. aufgefaßt werden. Alle Rechenregeln gelten dann für Tensoren beliebiger Stufe. Im Falle zweistufiger Tensoren kann man der Bilinearform eineindeutig eine lineare Abbildung zwischen den zugrundeliegenden Vektorräumen zuordnen.

[2] Allgemein wird mit V^* der Dualraum eines Vektorraums V bezeichnet.

[3] Der Einfachheit halber werden das Vektorfeld und sein Wert im Tangentialraum mit jeweils demselben Buchstaben bezeichnet.

[4] Ein Beispiel hierfür ist die in Definition 1.10 eingeführte materielle Geschwindigkeit.

Kotangentialbündeln.

Ein zweistufiger *Zweipunkttensor* ist eine lineare Abbildung zwischen Tangential- bzw. Kotangentialräumen verschiedener Konfigurationen, z. B. zwischen dem Tangentialraum $T_{\mathbf{X}}\Omega_0$ der Referenzkonfiguration und dem Tangentialraum $T_{\mathbf{x}}\Omega_t$ der Momentankonfiguration. Entsprechend ist ein zweistufiges *Zweipunkttensorfeld* eine lineare Abbildung zwischen Tangential- bzw. Kotangentialbündeln verschiedener Konfigurationen. Dem Punkt $\mathbf{X}$ eines Paares aus dem Tangential- bzw. Kotangentialbündel wird dabei jeweils mit der zugehörigen Bewegungsfunktion der Punkt $\mathbf{x} = \boldsymbol{\chi}_t(\mathbf{X})$ zugeordnet. ■

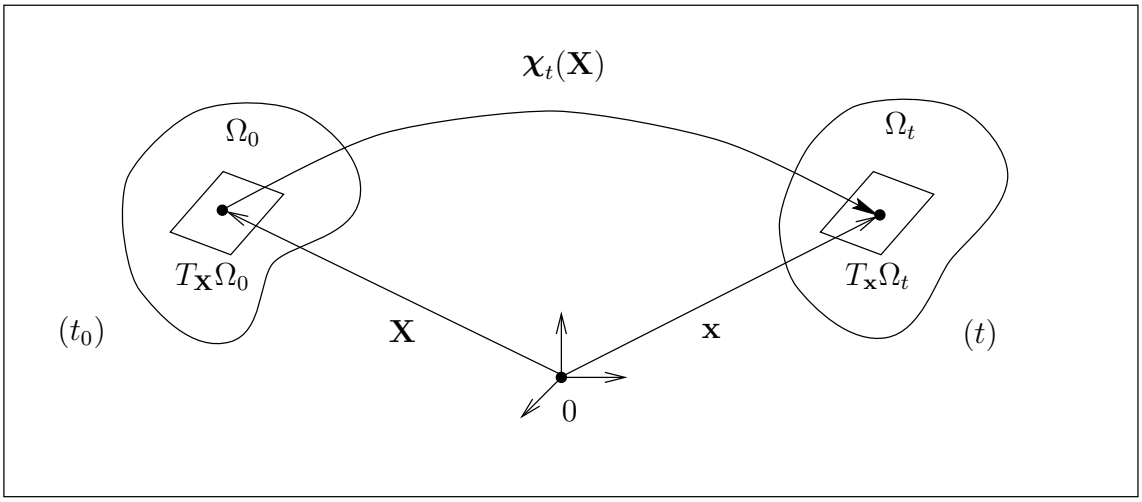


Abbildung A.1: *Tangentialräume*

Alle Definitionen und Begriffe in diesem Abschnitt sind sinngemäß auch für Mischungen zu verstehen, indem Punkte der Referenzkonfiguration wie üblich mit $\mathbf{X}_\alpha$ und die Plazierungsfunktion der Phase φ^α zum Zeitpunkt t mit $\overset{\alpha}{\boldsymbol{\chi}}_t$ bezeichnet werden.

Die erste Komponente eines Punkt-Vektor-Paares aus einem Tangentialbündel wird im folgenden nicht mehr explizit notiert; vielmehr wird stets vorausgesetzt, daß im Falle von Zweipunkt-Operationen zwischen Referenz- und Momentankonfiguration der Punkt $\mathbf{X}$ immer mit der zugehörigen Bewegungsfunktion auf den Punkt $\mathbf{x} = \boldsymbol{\chi}_t(\mathbf{X})$ abgebildet wird.

Der Tangentialraum in einem Punkt des *Euklid*schen Raumes $\mathbb{R}^3$ ist der Vektorraum $\mathbb{R}^3$ selbst[5]. Da man den Dualraum des $\mathbb{R}^n$ über die kanonische Einbettung mit dem $\mathbb{R}^n$ identifizieren kann, ist der Kotangentialraum des $\mathbb{R}^3$ wieder der $\mathbb{R}^3$, allerdings ausgestattet mit der dualen Basis. In der vorliegenden Arbeit ist jede Plazierung eines Körpers eine offene Teilmenge des $\mathbb{R}^3$, so daß die oben eingeführten Vektorräume alle gleich dem $\mathbb{R}^3$ sind.

Def. A.2: Ein *Koordinatensystem* der Referenzkonfiguration ist gegeben durch $\{0, \boldsymbol{\Theta}\}$ mit der Koordinatenabbildung $\boldsymbol{\Theta} : \Omega_0 \longrightarrow \mathbb{R}^3$. Entsprechend ist $\{0, \boldsymbol{\theta}\}$ mit der Koordi-

[5]Wie bereits in Kapitel 1 erwähnt, kann man den *Euklid*schen Anschauungsraum $\mathbb{E}^3$ mit dem Vektorraum $\mathbb{R}^3$ aller reellen Zahlentripel identifizieren. Im folgenden werden diese beiden Bezeichnungen daher nicht unterschieden.

natenabbildung $\boldsymbol{\theta} : \Omega_t \longrightarrow \mathbb{R}^3$ ein Koordinatensystem auf der Momentankonfiguration. Die Koordinatenabbildungen seien *Diffeomorphismen*, d. h. invertierbare C^1-stetige Abbildungen mit C^1-stetigen Inversen. Sie ordnen einem Vektor $\mathbf{X}$ bzw. $\mathbf{x}$ die Koordinaten $(\Theta^1(\mathbf{X}), \Theta^2(\mathbf{X}), \Theta^3(\mathbf{X}))$ bzw. $(\theta^1(\mathbf{x}), \theta^2(\mathbf{x}), \theta^3(\mathbf{x}))$ zu. Durch partielle Ableitung der Ortsvektoren nach den Koordinatenfunktionen erhält man die *Basisvektoren* der Tangentialräume der Referenz- und Momentankonfiguration:

$$\mathbf{G}_i = \frac{\partial \mathbf{X}}{\partial \Theta^i}, \qquad \mathbf{g}_i = \frac{\partial \mathbf{x}}{\partial \theta^i}. \tag{A.1}$$

Die Basisvektoren der Kotangentialräume erhält man aus den *Dualitätsbeziehungen*[6]

$$\mathbf{G}^i \cdot \mathbf{G}_j = \delta^i_j, \qquad \mathbf{g}^i \cdot \mathbf{g}_j = \delta^i_j, \tag{A.2}$$

bzw. durch Ableitung der Koordinatenabbildungen nach den Ortsvektoren:

$$\mathbf{G}^i = \frac{\partial \Theta^i}{\partial \mathbf{X}}, \qquad \mathbf{g}^i = \frac{\partial \theta^i}{\partial \mathbf{x}}. \tag{A.3}$$

■

In Abbildung A.2 sind die wichtigsten kinematischen Abbildungen sowie die zugehörigen Tangentialräume dargestellt.

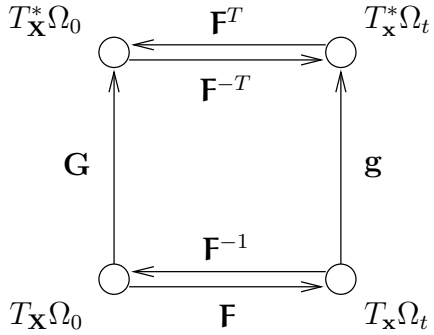


Abbildung A.2: *Kinematische Abbildungen*

Die Abbildungen $\mathbf{G}$ und $\mathbf{g}$ bezeichnen dabei die *Metriktensoren* der Referenz- und der Momentankonfiguration, die in den zugehörigen Basen die folgende Darstellung[7] besitzen:

$$\begin{aligned} \mathbf{G} &: T_{\mathbf{X}}\Omega_0 \longrightarrow T^*_{\mathbf{X}}\Omega_0, \qquad \mathbf{G} = G_{ij}\ \mathbf{G}^i \otimes \mathbf{G}^j \qquad \text{mit} \quad G_{ij} = \mathbf{G}_i \cdot \mathbf{G}_j, \\ \mathbf{g} &: T_{\mathbf{x}}\Omega_t \longrightarrow T^*_{\mathbf{x}}\Omega_t, \qquad \mathbf{g} = g_{ij}\ \mathbf{g}^i \otimes \mathbf{g}^j \qquad \text{mit} \quad g_{ij} = \mathbf{g}_i \cdot \mathbf{g}_j. \end{aligned} \tag{A.4}$$

Der Metriktensor ist also diejenige Abbildung, die einem Vektor $\mathbf{v} = \mathrm{v}^k\, \mathbf{g}_k$ den zugehörigen Kovektor zuordnet (Basiswechsel), z. B. für den Metriktensor der Momentankonfiguration:

$$\mathbf{g}\,(\mathrm{v}^k\, \mathbf{g}_k) = (g_{ij}\ \mathbf{g}^i \otimes \mathbf{g}^j)\,(\mathrm{v}^k\, \mathbf{g}_k) = g_{ij}\, \delta^j_k\, \mathrm{v}^k\, \mathbf{g}^i = g_{ik}\, \mathrm{v}^k\, \mathbf{g}^i =: \mathrm{v}_i\, \mathbf{g}^i. \tag{A.5}$$

[6] Das *Kronecker-Symbol* $\delta^i_j = \delta_{ij} = \delta^{ij}$ nimmt für $i = j$ den Wert 1 und sonst den Wert 0 an.

[7] Gemäß der *Einstein*schen Summenkonvention ist über gleiche gegenständige Indizes zu summieren.

Gemäß der obigen Definition sind die Koordinatenabbildungen auf der Referenz- und Momentankonfiguration beliebig wählbar. Für das Arbeiten mit Tensoren innerhalb der Materialtheorie ist die spezielle Wahl *konvektiver Koordinaten* ein adäquates Hilfsmittel. Dabei werden die Koordinatenlinien der Referenzkonfiguration als materiell betrachtet, also den materiellen Punkten des Körpers „angeheftet“. In der Momentankonfiguration wird dann zu jedem Zeitpunkt ein anderes Koordinatensystem gewählt, nämlich das durch die Bewegung $\boldsymbol{\chi}_t(\mathbf{X})$ deformierte Koordinatensystem der Referenzkonfiguration: $\theta(\mathbf{x}) = \theta(\boldsymbol{\chi}_t(\mathbf{X})) = \Theta(\mathbf{X})$. In Übereinstimmung mit *de Boer* [20], *Ehlers* [44] werden die konvektiven Basisvektoren (*natürliche Basisvektoren*) in der Referenzkonfiguration mit $\mathbf{h}_i$, in der Momentankonfiguration mit $\mathbf{a}_i$ bezeichnet. Damit ergeben sich sehr einfache Darstellungen für den Deformationsgradienten und seine Transponierte bzw. Inverse, da die Information über die Deformation bereits in der Basis steckt, siehe auch Abbildung A.2:

$$\begin{aligned}
\mathsf{F} &: T_{\mathbf{X}}\Omega_0 \longrightarrow T_{\mathbf{x}}\Omega_t, & \mathsf{F} &= \mathbf{a}_i \otimes \mathbf{h}^i, \\
\mathsf{F}^{-1} &: T_{\mathbf{x}}\Omega_t \longrightarrow T_{\mathbf{X}}\Omega_0, & \mathsf{F}^{-1} &= \mathbf{h}_i \otimes \mathbf{a}^i, \\
\mathsf{F}^{T} &: T^*_{\mathbf{x}}\Omega_t \longrightarrow T^*_{\mathbf{X}}\Omega_0, & \mathsf{F}^{T} &= \mathbf{h}^i \otimes \mathbf{a}_i, \\
\mathsf{F}^{-T} &: T^*_{\mathbf{X}}\Omega_0 \longrightarrow T^*_{\mathbf{x}}\Omega_t, & \mathsf{F}^{-T} &= \mathbf{a}^i \otimes \mathbf{h}_i.
\end{aligned} \tag{A.6}$$

In numerischen Algorithmen kommt das Konzept der konvektiven Koordinaten allerdings selten zur Anwendung. Hier wählt man im allgemeinen eine gemeinsame feste Basis für Referenz- und Momentankonfiguration und stellt alle Vektoren und Tensoren in dieser Basis dar. In der Praxis bedeutet dies, daß nur noch das Koeffizientenschema gespeichert und berechnet wird. Stellt man alle Größen in einer kartesischen Basis dar, so können Tensoren wie Matrizen behandelt werden, ansonsten muß man bei allen Operationen die gewählte Basis und die zugehörige Metrik berücksichtigen.

Anhang B: Materialparameter

Die in den Rechnungen in Kapitel 5 im Zusammenhang mit dem inkompressiblen Zweiphasenmodell verwendeten Materialparameter beschreiben einen bindigen Boden (wassergesättigter, schluffiger Ton, siehe *Thamm* [114]) und haben exemplarischen Charakter.

Elastizität

Parameter	Symbol	Wert	Einheit
Lamé-Konstanten	μ^S	$5{,}5833 \cdot 10^6$	$\mathrm{N/m^2}$
	λ^S	$8{,}3750 \cdot 10^6$	$\mathrm{N/m^2}$

Plastizität

Parameter	Symbol	Wert	Einheit
Fließbedingung	α	$1{,}0740 \cdot 10^{-2}$	–
	β	0,11946	–
	γ	1,5550	–
	δ	$1{,}3775 \cdot 10^{-7}$	$\mathrm{m^2/N}$
	ε	$4{,}3303 \cdot 10^{-9}$	$\mathrm{m^2/N}$
	κ	$1{,}0269 \cdot 10^4$	$\mathrm{N/m^2}$
	m	0,5935	–
Viskoplastizität	η	je nach ARWP	s
	σ_0	$1{,}0269 \cdot 10^4$	$\mathrm{N/m^2}$
	r	1,0	–

Bodenmechanische Kenngrößen

Parameter	Symbol	Wert	Einheit
Effektive Dichten	ρ^{SR} (ρ_S)	$2{,}720 \cdot 10^3$	$\mathrm{kg/m^3}$
	ρ^{FR} (ρ_w)	$1{,}000 \cdot 10^3$	$\mathrm{kg/m^3}$
Volumenanteile	n^S_{0S} $(1-n)$	0,540	–
	n^F_{0S} (n)	0,460	–
Effektive Fluidwichte	γ^{FR} (γ_w)	$1{,}000 \cdot 10^4$	$\mathrm{N/m^3}$
Darcy-Parameter	k^F (k_f)	$1{,}0 \cdot 10^{-7}$	m/s

Eine genaue Anpassung an Versuchsergebnisse wurde im Rahmen dieser Arbeit nicht vorgenommen. Bei den bodenmechanischen Kenngrößen ist in Klammern jeweils die in der Bodenmechanik übliche Schreibweise angegeben.

Alle Materialparameter sind hier in SI-Einheiten (MKS-System) notiert. In den numerischen Rechnungen werden jedoch skalierte Parameter verwendet, bei denen als Basiseinheit für die Masse $10^6\,\mathrm{kg}$ zugrundegelegt wird. Dadurch liegen alle Parameter nahezu in der gleichen Größenordnung, was für die Skalierung der linearen Gleichungssysteme von Vorteil ist. Außerdem erhält man die in Ingenieuranwendungen übliche Krafteinheit Mega-*Pascal*: $1\,\mathrm{MPa} = 1\,\mathrm{N/mm^2} = 1\,\mathrm{MN/m^2}$.

Literaturverzeichnis

[1] *Abraham, R.; Marsden, J. E.; Ratiu, T.*: *Manifolds, Tensor Analysis and Applications.* Addison-Wesley, Reading 1982.

[2] *Ainsworth, M.; Craig, A.*: A posteriori error estimators in the finite element method. *Numer. Math.* **60** (1992), 429–463.

[3] *Ainsworth, M.; Senior, B.*: Aspects of an adaptive *hp*-finite element method: Adaptive strategy, conforming approximation and efficient solvers. *Comput. Methods Appl. Mech. Engrg.* **150** (1997), 65–87.

[4] *Ainsworth, M.; Zhu, J. Z.; Craig, A. W.; Zienkiewicz, O. C.*: Analysis of the Zienkiewicz-Zhu a-posteriori error estimator in the finite element method. *Int. J. Numer. Methods Eng.* **28** (1989), 2161–2174.

[5] *Alexander, R.*: Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s. *SIAM J. Numer. Anal.* **14** (1977), 1006–1021.

[6] *Ammann, M.*: Wiedervernetzungsstrategien und Datentransfer in der adaptiven FEM. Diplomarbeit, Institut für Mechanik (Bauwesen), Universität Stuttgart 1998.

[7] *Babuška, I.; Miller, A.*: A feedback finite element method with a posteriori error estimation: Part I. The finite element method and some basic properties of the a posteriori error estimator. *Comput. Methods Appl. Mech. Engrg.* **61** (1987), 1–40.

[8] *Babuška, I.; Rheinboldt, W. C.*: *A-posteriori* error estimates for the finite element method. *Int. J. Numer. Methods Eng.* **12** (1978), 1597–1615.

[9] *Babuška, I.; Rheinboldt, W. C.*: Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.* **15** (1978), 736–754.

[10] *Babuška, I.; Rodriguez, R.*: The problem of the selection of an *a posteriori* error estimator based on smoothening techniques. *Int. J. Numer. Methods Eng.* **36** (1993), 539–567.

[11] *Babuška, I.; Strouboulis, T.; Upadhyay, C. S.*: A model study of the quality of a posteriori error estimators for linear elliptic problems. Error estimation in the interior of patchwise uniform grids of triangles. *Comput. Methods Appl. Mech. Engrg.* **114** (1994), 307–378.

[12] *Babuška, I.; Strouboulis, T.; Upadhyay, C. S.; Gangaraj, S. K.; Copps, K.*: An objective criterion for assessing the reliability of *a-posteriori* error estimators in finite element computations. *IACM Proceedings* **9** (1994), 27–38.

[13] *Babuška, I.; Strouboulis, T.; Upadhyay, C. S.; Gangaraj, S. K.; Copps, K.*: Validation of *a posteriori* error estimators by numerical approach. *Int. J. Numer. Methods Eng.* **37** (1994), 1073–1123.

[14] *Bank, R. E.*: *PLTMG: A Software Package for Solving Elliptic Partial Differential Equations, User's Guide 7.0*, Bd. 15 aus *Frontiers in Applied Mathematics*. SIAM, Philadelphia 1994.

[15] *Bank, R. E.; Smith, R. K.*: A posteriori error estimates based on hierarchical bases. *SIAM J. Numer. Anal.* **30** (1993), 921–935.

[16] *Bank, R. E.; Weiser, A.*: Some a posteriori error estimators for elliptic partial differential equations. *Math. Comput.* **44** (1985), 283–301.

[17] *Barthold, F.-J.; Schmidt, M.; Stein, E.*: Error indicators and mesh refinements for finite-element-computations of elastoplastic deformations. *Comp. Mech.* **22** (1998), 225–238.

[18] *Bathe, K.-J.*: *Finite Element Procedures in Engineering Analysis*. Prentice-Hall, Englewood Cliffs, N. J. 1982.

[19] *Bey, J.*: *Finite-Volumen- und Mehrgitter-Verfahren für elliptische Randwertprobleme*. B. G. Teubner, Stuttgart, Leipzig 1998.

[20] *de Boer, R.*: *Vektor- und Tensorrechnung für Ingenieure*. Springer-Verlag, Berlin 1982.

[21] *de Boer, R.; Ehlers, W.*: *Theorie der Mehrkomponentenkontinua mit Anwendung auf bodenmechanische Probleme*, Bd. 40 aus *Forschungsberichte aus dem Fachbereich Bauwesen*. Universität-GH-Essen, Essen 1986.

[22] *Bowen, R. M.*: Toward a thermodynamics and mechanics of mixtures. *Arch. Rational Mech. Anal.* **24** (1967), 370–403.

[23] *Bowen, R. M.*: Incompressible porous media models by use of the theory of mixtures. *Int. J. Engng. Sci.* **18** (1980), 1129–1148.

[24] *Bowen, R. M.*: Compressible porous media models by use of the theory of mixtures. *Int. J. Engng. Sci.* **20** (1982), 697–735.

[25] *Braess, D.*: *Finite Elemente*. Springer-Verlag, Berlin 1997.

[26] *Brenan, K. E.; Campbell, S. L.; Petzold, L. R.*: *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. North-Holland, New York 1989.

[27] *Brenner, S. C.; Scott, L. R.*: *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, New York 1994.

[28] *Brezzi, F.; Fortin, M.*: *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, New York 1991.

[29] *Campbell, S. L.; Gear, C. W.*: The index of general nonlinear DAEs. *Numer. Math.* **72** (1995), 173–196.

[30] *Cash, J. R.*: Diagonally implicit Runge-Kutta formulae with error estimates. *J. Inst. Maths Applics* **24** (1979), 293–301.

[31] *Ciarlet, P. G.*: *The Finite Element Method for Elliptic Problems.* North-Holland, Amsterdam 1978.

[32] *Ciarlet, P. G.*: *Mathematical Elasticity. Volume I: Three-Dimensional Elasticity.* North-Holland, Amsterdam 1988.

[33] *Clément, P.*: Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.* **2** (1975), 77–84.

[34] *Crawford, R. H.; Anderson, D. C.; Waggenspack, W. N.*: Mesh rezoning of 2d isoparametric elements by inversion. *Int. J. Numer. Methods Eng.* **28** (1989), 523–531.

[35] *Cuitiño, A. M.; Ortiz, M.*: State updates and state-transfer operators in computational plasticity. In *Besdo, D.; Stein, E.* (Hrsg.), *Finite Inelastic Deformations – Theory and Applications*, IUTAM Symposium Hannover/Germany 1991. Springer-Verlag, Berlin 1992, S. 239–258.

[36] *Deb, A.; Prevost, J. H.; Loret, B.*: Adaptive meshing for dynamic strain localization. *Comput. Methods Appl. Mech. Engrg.* **137** (1996), 285–306.

[37] *Demkowicz, L.; Oden, J. T.; Rachowicz, W.; Hardy, O.*: Toward a universal *h-p* adaptive finite element strategy, Part 1. Constrained approximation and data structure. *Comput. Methods Appl. Mech. Engrg.* **77** (1989), 79–112.

[38] *Demkowicz, L.; Rachowicz, W.; Westermann, T. A.*: Toward a universal *h-p* adaptive finite element strategy, Part 2. A posteriori error estimation. *Comput. Methods Appl. Mech. Engrg.* **77** (1989), 113–180.

[39] *Diebels, S.*: Grundgleichungen für Mischungen mit mikropolaren Konstituierenden. *Z. angew. Math. Mech.* **77** (1997), S 73 – S 74.

[40] *Diebels, S.; Ellsiepen, P.; Ehlers, W.*: *A Two-Phase Model for Viscoplastic Geomaterials.* Berichte aus dem Institut für Mechanik (Bauwesen), Nr. 96-II-6. Universität Stuttgart 1996.

[41] *Diebels, S.; Ellsiepen, P.; Ehlers, W.*: Zeitintegration viskoplastischer Zweiphasenmodelle. In *Bruhns, O. T.* (Hrsg.), *Große plastische Formänderungen, Bad Honnef 1997*, Bd. 114 aus *Mitteilungen aus dem Institut für Mechanik.* Ruhr-Universität Bochum 1998, S. 145–148.

[42] *Diebels, S.; Ellsiepen, P.; Ehlers, W.*: Error-controlled Runge-Kutta time integration of a viscoplastic hybrid two-phase model. *Technische Mechanik* **19** (1999), 19–27.

[43] *Dormand, J. R.; Prince, P. J.*: A family of embedded Runge-Kutta formulae. *J. Comp. Appl. Math.* **6** (1980), 19–26.

[44] *Ehlers, W.*: *Poröse Medien – ein kontinuumsmechanisches Modell auf der Basis der Mischungstheorie*, Bd. 47 aus *Forschungsberichte aus dem Fachbereich Bauwesen.* Universität-GH-Essen, Essen 1989.

[45] *Ehlers, W.*: Constitutive equations for granular materials in geomechanical context. In *Hutter, K.* (Hrsg.), *Continuum Mechanics in Environmental Sciences and Geophysics*, Bd. 337 aus *CISM Courses and Lectures.* Springer-Verlag, Wien 1993.

[46] *Ehlers, W.*: A single-surface yield function for geomaterials. *Arch. Appl. Mech.* **65** (1995), 246–259.

[47] *Ehlers, W.*: Grundlegende Konzepte in der Theorie Poröser Medien. *Technische Mechanik* **16** (1996), 63–76.

[48] *Ehlers, W.; Eipper, G.*: Finite elastic deformations in liquid-saturated and empty porous solids. *Transport in Porous Media* **34** (1999), 179–191.

[49] *Ehlers, W.; Ellsiepen, P.*: Adaptive Zeitintegrations-Verfahren für ein elastisch-viskoplastisches Zweiphasenmodell. *Z. angew. Math. Mech.* **78** (1998), S 361–S 362.

[50] *Ehlers, W.; Ellsiepen, P.*: PANDAS: Ein FE-System zur Simulation von Sonderproblemen der Bodenmechanik. In *Wriggers, P.; Meißner, U.; Stein, E.; Wunderlich, W.* (Hrsg.), *Finite Elemente in der Baupraxis: Modellierung, Berechnung und Konstruktion*, Beiträge zur Tagung FEM '98 an der TU Darmstadt am 5. und 6. März 1998. Ernst & Sohn, Berlin 1998, S. 391–400.

[51] *Ehlers, W.; Ellsiepen, P.; Blome, P.; Mahnkopf, D.; Markert, B.*: *Theoretische und numerische Studien zur Lösung von Rand- und Anfangswertproblemen in der Theorie Poröser Medien.* Abschlußbericht zum DFG-Forschungsvorhaben Eh 107/6-2. Berichte aus dem Institut für Mechanik (Bauwesen), Nr. 99-II-1. Universität Stuttgart 1999.

[52] *Ehlers, W.; Volk, W.*: On theoretical and numerical methods in the theory of porous media based on polar and non-polar elasto-plastic porous solid materials. *Int. J. Solids and Structures* **35** (1998), 4597–4617.

[53] *Eipper, G.*: *Theorie und Numerik finiter elastischer Deformationen in fluidgesättigten porösen Medien.* Dissertationen aus dem Institut für Mechanik im Bauwesen, Nr. II-1. Universität Stuttgart 1998.

[54] *Eriksson, K.; Estep, D.; Hansbo, P.; Johnson, C.*: *Computational Differential Equations.* Cambridge University Press, Cambridge 1996.

[55] *Fillunger, P.*: *Erdbaumechanik?.* Selbstverlag des Verfassers, Wien 1936.

[56] *Franz, U.*: *A posteriori* error estimation and adaptivity for finite element approximations of second-order hyperbolic equations: Analysis and numerical implementation for the equation for vibrations of a membrane. Diplomarbeit, Fachbereich Mathematik, Technische Hochschule Darmstadt 1995.

[57] *Fritzen, P.*: *Numerische Behandlung nichtlinearer Probleme der Elastizitäts- und Plastizitätstheorie*. Dissertation, Fachbereich Mathematik, Technische Hochschule Darmstadt 1997.

[58] *Fuenmayor, F. J.; Oliver, J. L.*: Criteria to achieve nearly optimal meshes in the h-adaptive finite element method. *Int. J. Numer. Methods Eng.* **39** (1996), 4039–4061.

[59] *Gallimard, L.; Ladevèze, P.; Pelle, J. P.*: A posteriori errror estimator for non-linear f. e. computations: Application to elastoplasticity. In *Owen, D. R. J.; Oñate, E.; Hinton, E.* (Hrsg.), *Computational Plasticity: Fundamentals and Applications*. Pineridge Press, Swansea 1995, S. 373–382.

[60] *Gallimard, L.; Ladevèze, P.; Pelle, J. P.*: Error estimation and adaptivity in elastoplasticity. *Int. J. Numer. Methods Eng.* **39** (1996), 189–217.

[61] *Girkmann, K.*: *Flächentragwerke*. Springer-Verlag, Wien 1963.

[62] *Hairer, E.; Lubich, C.; Roche, M.*: *The Numerical Solution of Differential-Algebraic Equations by Runge-Kutta Methods*. Springer-Verlag, Berlin 1989.

[63] *Hairer, E.; Nørsett, S. P.; Wanner, G.*: *Solving Ordinary Differential Equations – Nonstiff Problems*, Bd. 1. Springer-Verlag, Berlin 1987.

[64] *Hairer, E.; Wanner, G.*: *Solving Ordinary Differential Equations – Stiff and Differential-Algebraic Problems*, Bd. 2. Springer-Verlag, Berlin 1991.

[65] *Hartmann, S.*: Zur Berechnung inelastischer Festkörper mit der Methode der finiten Elemente. In *Hartmann, S.; Haupt, P.; Ulbricht, V.* (Hrsg.), *Modellierung und Identifikation*. Gesamthochschul-Bibliothek Verlag, Kassel 1998, S. 119–130.

[66] *Hartmann, S.; Lührs, G.; Haupt, P.*: An efficient stress algorithm with applications in viscoplasticity and plasticity. *Int. J. Numer. Methods Eng.* **40** (1997), 991–1013.

[67] *Haupt, P.*: Foundation of continuum mechanics. In *Hutter, K.* (Hrsg.), *Continuum Mechanics in Environmental Sciences and Geophysics*, Bd. 337 aus *CISM Courses and Lectures*. Springer-Verlag, Wien 1993.

[68] *Heinrich, G.; Desoyer, K.*: Hydromechanische Grundlagen für die Behandlung von stationären und instationären Grundwasserströmungen, II. Mitteilung. *Ing.-Archiv* **24** (1956), 81–84.

[69] *Heinrich, G.; Desoyer, K.*: Praktische Methoden zur Lösung von Problemen der stationären und instationären Grundwasserströmungen. *Ing.-Archiv* **26** (1958), 30–42.

[70] *Heinrich, G.; Desoyer, K.*: Theorie dreidimensionaler Setzungsvorgänge in Tonschichten. *Ing.-Archiv* **30** (1961), 225–253.

[71] *Hinton, E.; Campbell, J. S.*: Local and global smoothing of discontinuous finite element functions using a least squares method. *Int. J. Numer. Methods Eng.* **8** (1974), 461–480.

[72] *Hoit, M.; Wilson, E.*: An equation numbering algorithm based on a minimum front criteria. *Comp. Struct.* **16** (1983), 225–239.

[73] *Hughes, T. J. R.*: *The Finite Element Method*. Prentice-Hall, London 1987.

[74] *Johnson, C.; Hansbo, P.*: Adaptive finite element methods in computational mechanics. *Comput. Methods Appl. Mech. Engrg.* **101** (1992), 143–181.

[75] *Kirchner, E.; Simeon, B.*: *A higher order time integration method for viscoplasticity*. Preprint-Nr. 1947. Fachbereich Mathematik, Technische Universität Darmstadt 1997.

[76] *Kossaczký, I.*: A recursive approach to local mesh refinement in two and three dimensions. *J. Comp. Appl. Math.* **55** (1994), 275–288.

[77] *Krause, R.; Rank, E.*: A fast algorithm for point-location in a finite element mesh. *Computing* **57** (1996), 49–62.

[78] *Ladevèze, P.; Pelle, J. P.; Rougeot, P.*: Error estimation and mesh optimization for classical finite elements. *Eng. Comp.* **8** (1991), 69–80.

[79] *Li, L.-Y.; Bettess, P.*: Notes on mesh optimal criteria in adaptive finite element computations. *Commun. Numer. Methods Eng.* **11** (1995), 911–915.

[80] *Lubliner, J.*: *Plasticity Theory*. Macmillan Publishing Company, New York 1990.

[81] *Mahnkopf, D.*: *Lokalisierung fluidgesättigter poröser Medien bei finiten elastoplastischen Deformationen*. Dissertationen aus dem Institut für Mechanik im Bauwesen, in Vorbereitung.

[82] *Markert, B.*: Ein viskoelastisches Modell der Kontinuumsmechanik mit Anwendung in der Theorie Poröser Medien. Diplomarbeit, Institut für Mechanik (Bauwesen), Universität Stuttgart 1998.

[83] *Marsden, J. E.; Hughes, T. J. R.*: *Mathematical Foundations of Elasticity*. Prentice-Hall, Englewood Cliffs, N. J. 1983.

[84] *Mitchell, W. F.*: Adaptive refinement for arbitrary finite-element spaces with hierarchical bases. *J. Comp. Appl. Math.* **36** (1991), 65–78.

[85] *Mücke, R.; Whiteman, J. R.*: *A posteriori* error estimates and adaptivity for finite element solutions in finite elasticity. *Int. J. Numer. Methods Eng.* **38** (1995), 775–795.

[86] *Oñate, E.; Bugeda, G.*: A study of mesh optimality criteria in adaptive finite element analysis. *Eng. Comp.* **10** (1993), 307–321.

[87] *Oden, J.; Brauchli, H. J.*: On the calculation of consistent stress distributions in finite element approximations. *Int. J. Numer. Methods Eng.* **3** (1971), 317–325.

[88] *Oden, J. T.*: *Finite Elements of Nonlinear Continua*. McGraw-Hill, New York 1972.

[89] *Oden, J. T.; Reddy, J. N.*: *An Introduction to the Mathematical Theory of Finite Elements*. John Wiley & Sons, New York 1976.

[90] *Oden, J. T.; Reddy, J. N.*: *Variational Methods in Theoretical Mechanics*. Springer-Verlag, Berlin 1976.

[91] *Ortiz, M.; Quigley, J. J.*: Adaptive mesh refinement in strain localization problems. *Comput. Methods Appl. Mech. Engrg.* **90** (1991), 781–804.

[92] *Pastor, M.; Peraire, J.; Zienkiewicz, O. C.*: Adaptive remeshing for shear band localization problems. *Arch. Appl. Mech.* **61** (1991), 30–39.

[93] *Perić, D.; Hochard, C.; Dutko, M.; Owen, D. R. J.*: Transfer operators for evolving meshes in small strain elasto-plasticity. *Comput. Methods Appl. Mech. Engrg.* **137** (1996), 331–344.

[94] *Perić, D.; Yu, J.; Owen, D. R. J.*: On error estimates and adaptivity in elastoplastic solids: Applications to the numerical simulation of strain localization in classical and Cosserat continua. *Int. J. Numer. Methods Eng.* **37** (1994), 1351–1379.

[95] *Perzyna, P.*: Fundamental problems in viscoplasticity. *Adv. Appl. Mech.* **9** (1966), 243–377.

[96] *Prothero, A.; Robinson, A.*: On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Math. Comput.* **28** (1974), 145–162.

[97] *Rabbat, N. B. G.; Sangiovanni-Vincentelli, A. L.; Hsieh, H. Y.*: A multilevel Newton algorithm with macromodeling and latency for the analysis of large-scale nonlinear circuits in the time domain. *IEEE Trans. Circuits Syst.* **26** (1979), 733–741.

[98] *Rachowicz, W.; Oden, J. T.; Demkowicz, L.*: Toward a universal *h-p* adaptive finite element strategy, Part 3. Design of *h-p* meshes. *Comput. Methods Appl. Mech. Engrg.* **77** (1989), 181–212.

[99] *Rannacher, R.; Suttmeier, F.-T.*: A posteriori error control in finite element methods via duality techniques: Application to perfect plasticity. *Comp. Mech.* **21** (1998), 123–133.

[100] *Rivara, M.-C.*: Algorithms for refining triangular grids suitable for adaptive and multi-grid techniques. *Int. J. Numer. Methods Eng.* **20** (1984), 745–756.

[101] *Rivara, M.-C.*: Mesh refinement processes based on the generalized bisection of simplices. *SIAM J. Numer. Anal.* **21** (1984), 604–613.

[102] *Saad, Y.*: *Iterative Methods for Sparse Linear Systems*. PWS Pubishing Company, Boston 1996.

[103] *Samet, H.*: *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading 1994.

[104] *Schwarz, H. R.*: *Methode der finiten Elemente*. Teubner, Stuttgart 1991.

[105] *Shewchuk, J. R.*: *Triangle: A Two-Dimensional Quality Mesh Generator and Delaunay Triangulator*. School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 1996. `http://www.cs.cmu.edu/~quake/triangle.html`.

[106] *Simo, J. C.; Hughes, T. J. R.*: *Computational Inelasticity*. Springer-Verlag, New York 1998.

[107] *Simo, J. C.; Taylor, R. L.*: Consistent tangent operators for rate-independent elastoplasticity. *Comput. Methods Appl. Mech. Engrg.* **48** (1985), 101–118.

[108] *Stein, E.; Ohnimus, S.*: Coupled model-adaptivity and solution-adaptivity in the finite-element method. *Comput. Methods Appl. Mech. Engrg.* **150** (1997), 327–350.

[109] *Strang, G.; Fix, G. J.*: *An Analysis of the Finite Element Method*. Prentice-Hall, Englewood Cliffs, N. J. 1973.

[110] *Strehmel, K.; Weiner, R.*: *Numerik gewöhnlicher Differentialgleichungen*. Teubner, Stuttgart 1995.

[111] *Suttmeier, F. T.*: *Adaptive Finite Element Approximation of Problems in Elasto-Plasticity Theory*. Dissertation, Institut für Angewandte Mathematik, Universität Heidelberg 1996.

[112] *Szabó, B.; Babuška, I.*: *Finite Element Analysis*. John Wiley & Sons, New York 1991.

[113] *Terzaghi, K.; Jelinek, R.*: *Theoretische Bodenmechanik*. Springer-Verlag, Berlin 1954.

[114] *Thamm, B. R.*: *Berechnung der Anfangssetzungen und der Anfangsporenwasserüberdrücke eines wassergesättigten normalverdichteten Tones*. Mitteilung Nr. 1. Baugrundinstitut Stuttgart 1974.

[115] *Törnig, W.; Spellucci, P.*: *Numerische Mathematik für Ingenieure und Physiker – Numerische Methoden der Analysis*, Bd. 2. Springer-Verlag, Berlin 1990.

[116] *Truesdell, C.*: Sulle basi della termomeccanica. *Accademia Nazionale dei Lincei, Rendiconti della Classe di Scienze Fisiche, Matematiche e Naturali (8)* **22** (1957), 33–38, 158–166.

[117] *Truesdell, C.*: *Rational Thermodynamics*. Springer-Verlag, New York 1984.

[118] *Truesdell, C.; Toupin, R. A.*: The classical field theories. In *Flügge, S.* (Hrsg.), *Handbuch der Physik*, Bd. III/1. Springer-Verlag, Berlin 1960.

[119] *Verfürth, R.*: A posteriori error estimates for nonlinear problems, finite element discretizations of elliptic equations. *Math. Comput.* **62** (1994), 445–475.

[120] *Verfürth, R.*: *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley-Teubner, Chichester, Stuttgart 1996.

[121] *Volk, W.*: *Untersuchung des Lokalisierungsverhaltens mikropolarer poröser Medien mit Hilfe der Cosserat-Theorie*. Dissertationen aus dem Institut für Mechanik im Bauwesen, Nr. II-2. Universität Stuttgart 1999.

[122] *Wieners, C.*: Multigrid methods for Prandtl-Reuß plasticity. Preprint, eingereicht bei *Numerical Linear Algebra with Applications* 1998.

[123] *Wieners, C.*: Diskussion mit Dr. C. Wieners (ICA1, Stuttgart) über die Problematik der Index-Bestimmung des DAE-Systems der Elastoplastizität. März 1999.

[124] *Wittekindt, J.*: *Die numerische Lösung von Anfangs-Randwertproblemen zur Beschreibung inelastischen Werkstoffverhaltens*. Dissertation, Fachbereich Mathematik, Technische Hochschule Darmstadt 1991.

[125] *Wriggers, P.; Scherf, O.*: Adaptive finite element methods for contact problems in plasticity. In *Owen, D. R. J.; Oñate, E.; Hinton, E.* (Hrsg.), *Computational Plasticity: Fundamentals and Applications*. Pineridge Press, Swansea 1995, S. 787–807.

[126] *Zienkiewicz, O. C.; Taylor, R. L.*: *The Finite Element Method – Basic Formulation and Linear Problems*, Bd. 1. McGraw-Hill, London 1989.

[127] *Zienkiewicz, O. C.; Taylor, R. L.*: *The Finite Element Method – Solid and Fluid Mechanics, Dynamics and Non-Linearity*, Bd. 2. McGraw-Hill, London 1991.

[128] *Zienkiewicz, O. C.; Zhu, J. Z.*: A simple error estimator and adaptive procedure for practical engineering analysis. *Int. J. Numer. Methods Eng.* **24** (1987), 337–357.

[129] *Zienkiewicz, O. C.; Zhu, J. Z.*: The superconvergent patch recovery and a posteriori error estimates. Part 1: The recovery technique. *Int. J. Numer. Methods Eng.* **33** (1992), 1331–1364.

[130] *Zienkiewicz, O. C.; Zhu, J. Z.*: The superconvergent patch recovery and a posteriori error estimates. Part 2: Error estimates and adaptivity. *Int. J. Numer. Methods Eng.* **33** (1992), 1365–1382.

Lebenslauf

Peter Ellsiepen

28. April 1967	geboren in Konstanz
1973 – 1977	Grundschule in Dettingen
1977 – 1986	Alexander-von-Humboldt-Gymnasium Konstanz Abschluß: Abitur
10/1986 – 12/1987	Grundwehrdient in Pfullendorf
1/1988 – 9/1988	Tätigkeit im Ingenieurbüro ib datentechnik GmbH, Konstanz
10/1988 – 9/1990	Grundstudium Mathematik und Informatik an der Technischen Hochschule Darmstadt
9/1990	Vordiplom im Studiengang Mathematik Vordiplom im Studiengang Informatik
10/1990 – 4/1994	Hauptstudium Mathematik mit Schwerpunkt Informatik (MSI) an der Technischen Hochschule Darmstadt
9/1993 – 3/1994	Auslandsaufenthalt an der University of Leeds, UK, dabei Anfertigung der Diplomarbeit
4/1994	Diplom im Studiengang Mathematik mit Schwerpunkt Informatik (MSI) Abschluß: Dipl.-Math.
4/1994 – 9/1994	wissenschaftliche Hilfskraft mit Abschluß am Fachbereich Mathematik der Technischen Hochschule Darmstadt bei Prof. Dr. Karl Graf Finck von Finckenstein
10/1994 – 4/1995	wissenschaftliche Hilfskraft mit Abschluß am Fachbereich Mechanik der Technischen Hochschule Darmstadt bei Prof. Dr.-Ing. Wolfgang Ehlers
seit 5/1995	wissenschaftlicher Mitarbeiter am Institut für Mechanik (Bauwesen) der Universität Stuttgart bei Prof. Dr.-Ing. Wolfgang Ehlers

Bisher in dieser Reihe erschienen:

II-1 G. Eipper: Theorie und Numerik finiter elastischer Deformationen in fluidgesättigten porösen Festkörpern, Juni 1998.

II-2 W. Volk: Untersuchung des Lokalisierungsverhaltens mikropolarer poröser Medien mit Hilfe der *Cosserat*-Theorie, Mai 1999.

II-3 P. Ellsiepen: Zeit- und ortsadaptive Verfahren angewandt auf Mehrphasenprobleme poröser Medien, Juli 1999.