Multilevelverfahren für nichtlineare Finite-Element Ausgleichsprobleme

Vom Fachbereich Mathematik und Informatik der Universität Hannover

zur Erlangung des Grades

Doktor der Naturwissenschaften

Dr. rer. nat.

genehmigte Dissertation

von

Dipl. math. techn.

Johannes Rudolf Korsawe

geboren am 22. August 1972 in Hannover

2001

Referent : Prof. Dr. Gerhard Starke

Korreferent : PD Dr. Cornelis W. Oosterlee

Tag der Promotion : 7. November 2001

Datum der Veröffentlichung : Dezember 2001

Danksagung

Die vorliegende Arbeit ist ein Ergebnis meiner Arbeit als wissenschaftlicher Mitarbeiter an den Universitäten Essen und Hannover in den Jahren 1998 bis 2001.

Vor allem danke ich meinem Doktorvater *Herrn Prof. Dr. Gerhard Starke* für seine geduldige und unermüdliche Betreuung beim Zustandekommen dieser Arbeit. Besonders dankbar bin ich für die in vielen detaillierten Diskussionen entstandene Hinführung zum selbständigen wissenschaftlichen Arbeiten.

Für die Unterstützung meiner Arbeit durch die *Arbeitsgruppe Ingenieurmathematik an der Universität Essen (GH)* und das *Institut für Angewandte Mathematik an der Universität Hannover* möchte ich mich an dieser Stelle herzlich bedanken.
Insbesondere danke ich dabei für die hervorragende Zusammenarbeit und die technischen Hilfestellungen bei der Umsetzung der Arbeit am Rechner *Frau Dr. Heike Haschke* von der Universität Essen (GH) und *Herrn Dr. Matthias Maischak* von der Universität Hannover.

Für die Förderung meiner Forschungstätigkeit durch die *Deutsche Forschungsgemeinschaft (DFG)* unter dem Aktenzeichen STA 402/4 bedanke ich mich ebenfalls herzlich.

Weiterhin bedanke ich mich bei *Herrn PD Dr. Cornelis W. Oosterlee* für die freundliche Übernahme des Korreferats.

Meiner *Familie* und meiner Verlobten *Bianca Behrens* danke ich herzlichst für die Unterstützung in allen Bereichen ausserhalb der Wissenschaft.

Beim Entstehen der Arbeit durfte ich deutlich Gottes Hilfe und Leitung spüren. All seiner Liebe, Gnade und Geduld gilt daher mein innigster Dank.

Dezember 2001,

Johannes Korsawe.

Zusammenfassung

Diese Arbeit behandelt die Lösung einer Ausgleichsaufgabe über einem System nichtlinearer partieller Differentialgleichungen erster Ordnung. Für dieses Problem werden zwei Standardansätze besonders im Hinblick auf die Kontrolle des algebraischen und des Diskretisierungsfehlers verglichen: Zum einen die Diskretisierung des Lösungsraums H und die Verwendung eines linearen Multilevel-Verfahrens zur Lösung der linearen Systeme, die sich aus einem globalen Linearisierungsansatz ergeben und zum anderen die Verwendung von nichtlinearen Multilevel-Verfahren zur Lösung des diskretisierten nichtlinearen Problems.
In dieser Arbeit wird eine neue Theorie zu inexakten Newton-Verfahren in unendlichdimensionalen Räumen entwickelt und ergibt in der Anwendung auf das nichtlineare Ausgleichsproblem eine neue Möglichkeit, Exaktheitsbedingungen bei der Lösung der Formulierungen anzugeben. Im Detail ersetzen diese Exaktheitsbedingungen die bisherigen Heuristiken in den Fragen nach der Exaktheit der Lösung der linearen Systeme und der Frage, wann der diskrete Lösungsraum verfeinert werden sollte. Innerhalb der neuen Theorie tragen dann Diskretisierungsfehler und algebraischer Fehler zu gleichen Teilen zur Inexaktheit des Verfahrens bei.
Um optimale Konvergenzraten bei der Anwendung linearer Multilevel-Verfahren in $H(\mathrm{div}, \Omega)$ zu erzielen, werden zwei bekannte Glättungsverfahren für solche Probleme verglichen und auf den Fall wechselnder Randbedingungen und nicht einfach zusammenhängender Gebiete Ω erweitert. Eine realistische Anwendungsaufgabe aus dem Bereich der Hydromechanik wird schliesslich als Vergleichsproblem zum numerischen Vergleich der erhaltenen Lösungsverfahren verwendet.

Stichworte: Multilevel-Verfahren, Finite-Element-Ausgleichsformulierung, inexakte Newton-Verfahren

Abstract

This thesis is about the solution of a first-order system least-squares ansatz for nonlinear partial differential equations of second order. Two standard ways for solving such a formulation are studied with respect to the control of discretization and algebraical error: First, the discretization of the solution space H and the application of linear multilevel methods to the linear systems which arise from global Newton-like approaches for the discretized nonlinear problems and second, the application of nonlinear multilevel techniques to the discretized nonlinear problem.
A new way to deduce exactness bounds in the solution of the finite-dimensional systems as well as refinement strategies for the discretization in order to ensure convergence at optimal cost is described. To this end,a new overall convergence theory for the minimization of the nonlinear least-squares functional in H is applied from the extension of inexact Newton methods in $\mathbb{R}^n$ to the infinite-dimensional case. In effect, this ansatz replaces the heuristics in choosing how exact to solve the linear systems and when to refine the discrete solution space, if nonlinear multilevel methods are not used. Within this theory, the errors from discretization and the approximate solution of the linear systems both contribute to the inexactness of the method.
In order to achieve optimal multilevel convergence rates for problems in $H(\mathrm{div}, \Omega)$, two smoothing schemes for such problems are studied and extended to the case of changing boundary conditions and domains Ω which are not simply connected.
A realistic water infiltration problem serves as a benchmark problem for the numerical comparison of the competetiveness of the respective approaches.

Keywords: multilevel, least-squares finite-element methods, inexact Newton methods

Inhaltsverzeichnis

Kapitel 0

Einleitung

Um die Mechanismen innerhalb der Natur zu begreifen und in der Technik nutzen zu können, werden unter Anwendung mathematischer Abstraktion in den Naturwissenschaften Methoden zur Beschreibung dieser alltäglichsten Abläufe entwickelt. Die Lösung der solchen Modellen zu Grunde liegenden Gleichungen ist die wichtigste Aufgabe der angewandten Mathematik.
Zur Behandlung der bei der Modellierung häufig auftretenden partiellen Differentialgleichungen ist etwa seit der Mitte des vorigen Jahrhunderts die Methode der finiten Elemente (FE-Methode) als Hilfsmittel entwickelt worden (für eine eingehende Beschreibung dieser Entwicklung siehe [53]). Der entscheidende Schritt bei der Anwendung dieser Methode ist die Umformulierung und Abschwächung der Differentialgleichung in eine Variationsaufgabe und die Darstellung dieser Aufgabe als endlichdimensionales Gleichungssystem durch den Ritz-Galerkin-Ansatz. Diese Gleichungssysteme können in der FE-Methode in vielen Fällen durch Multilevelverfahren schnell gelöst werden. Dies und die rasante Entwicklung der Leistungsfähigkeit von Computern trug schließlich zu der weiten Verbreitung der FE-Methode bei.

In dieser Arbeit werden partielle Differentialgleichungen zweiter Ordnung behandelt. Solche Differentialgleichungen können häufig durch Einführung einer Hilfsvariablen in ein System von Gleichungen erster Ordnung überführt werden, wie zum Beispiel die Gleichung

$$-\triangle p \;=\; f$$

in das äquivalente System

$$\begin{aligned} \operatorname{div} u \;&=\; f \\ u + \nabla p \;&=\; 0 \end{aligned} \tag{1}$$

umgeformt werden kann. Dabei besitzt die Hilfsvariable u häufig auch eine physikalische Bedeutung und kann so von ganz eigenem Interesse sein, wie zum Beispiel in der Bodenmechanik der sich aus dem hydraulischen Potential p einstellende Fluss u eines Fluids.
Der klassische Ansatz für die Umformulierung des Systems (1) in eine Variationsaufgabe ist die sogenannte gemischte Methode, bei der Variationsprinzipien direkt auf das System angewendet werden. Seit einiger Zeit werden jedoch auch Ausgleichsfunktionale (Least-Squares-Verfahren) zur Herleitung einer Variationsaufgabe aus (1) verwendet (zur Übersicht siehe [7]). Das Least-Squares-Verfahren bietet in mancher Hinsicht Vorteile gegenüber der gemischten Methode: Bei beliebiger Wahl der Ansatzräume für die Lösungsfunktionen sind die entstehenden algebraischen Systeme positiv definit. Die Wahl der Ansatzräume des gemischten Ansatzes ist dagegen wegen der zu erfüllenden inf-sup-Bedingung ([11]) eingeschränkt und die linearen Gleichungsysteme sind als Repräsentierung eines Sattelpunktproblems indefinit.

Zur Verbesserung einer mit Hilfe dieser Methoden erzielten Näherung an die Lösung kann der Ansatzraum mit Hilfe von a-posteriori Fehlerschätzern gezielt vergrößert werden. Diese Fehlerschätzer benötigen häufig zusätzlichen Berechnungsaufwand und erfordern teilweise zusätzliche Voraussetzungen, wie zum Beispiel die Saturierungsbedingung beim hierarchischen Ansatz. In der Least-Squares-Methode kann jedoch oft direkt das Ausgleichsfunktional, das ohne weiteren Aufwand zur Verfügung steht, als Fehlerschätzer verwendet werden, wie für einige Fälle in [6], [13] und [38] gezeigt wird.

In dieser Arbeit werden nun auch nichtlineare Probleme in den Kontext von Least-Squares-Verfahren gestellt, die als zusätzlichen Lösungsschritt die Anwendung eines Newton-artigen Verfahrens zur Linearisierung der Gleichungen erfordern. Die Finite-Element-Methode gibt dabei zwei gangbare Wege vor: Die Linearisierung nur über dem feinsten Raum und die Lösung des LGS mit einem linearen Multilevelverfahren (siehe z.B. [27]) oder die Anwendung eines nichtlinearen Multilevelverfahrens, wodurch auf jedem Level ein nichtlineares Problem linearisiert und gelöst werden muss (z.B. [9], "FAS").
Da durch die sukzessiven Linearisierungen die nichtlinearen Probleme nur näherungsweise gelöst werden, müssen die verbleibenden Ungenauigkeiten innerhalb der Konvergenzanalyse genau kontrolliert werden (inexakte Verfahren), wie etwa in [4], [30] und [44] für verschiedene spezielle Multilevelverfahren bzw. Voraussetzungen an das Problem angegeben wird. Häufig sind dabei Kenntnisse über die zweiten Ableitungen der Nichtlinearitäten notwendig oder die theoretischen Ergebnisse resultieren nicht in praktisch anwendbaren Exaktheitsbedingungen an die nichtlineare Iteration. Dies zeigt die Schwierigkeit der analytischen Behandlung der Kopplung verschiedener Konvergenzprozesse, wo die Diskretisierung des Lösungsraums, die Inexaktheit des nichtlinearen Verfahrens und die Glättungseigenschaften des Multilevelverfahrens gegeneinander abgewogen werden müssen, insbesondere, wenn nicht mehr als Informationen erster Ordnung über die Nichtlinearitäten zur Verfügung stehen.
Entscheidend für die Konvergenztheorie ist nun die Wahl des (äußeren) Lösungsverfahrens. Wählt man das nichtlineare Multilevelverfahren als Löser, verwendet die bisher bekannte Theorie wie z.B. in [27] oder [28] exakte Lösungen der entstehenden nichtlinearen Aufgaben. Eine Übertragung auf den Fall inexakter nichtlinearer Verfahren kann dort nicht unmittelbar abgeleitet werden. Damit die Gesamtkonvergenz des Verfahrens innerhalb einer Theorie inexakter Verfahren etabliert werden kann, muss daher auf ein inexaktes Newton-Verfahren als äußeren Löser zurückgegriffen werden.
Eine äußerst flexible Definition inexakter Newton-Verfahren im $\mathbb{R}^n$ wurde von Eisenstat und Walker in [19] und [20] entwickelt (in ähnlicher Form auch schon in [16]). Die Konvergenz des Verfahrens wird nachgewiesen, solange die Reduktion der Norm der Funktion durch eine gewisse Abstiegsrichtung mindestens so groß ist wie ein bestimmter Bruchteil der Reduktion der Norm des linearen Modells. Für die Berechnung der Abstiegsrichtung ist dabei jedes Verfahren zugelassen, insbesondere also auch das Gauß-Newton-Verfahren, das ja nur Informationen erster Ordnung von den Nichtlinearitäten benötigt und stets zu einem positiv definiten linearen Gleichungssystem führt. Von hier zur Anwendung dieser Theorie auf das Ausgleichsfunktional des Least-Squares-Verfahrens ist es dann nur noch ein kleiner, aber bedeutsamer Schritt.
Dieser Schritt wird jedoch verhindert, wenn die Diskretisierung der Ansatzräume bereits vorweggenommen wurde, denn die exakte Lösung ist im Allgemeinen nicht in einem endlichdimensionalen Lösungsraum enthalten. Dadurch wird wird die vorher kompatible Minimierungsaufgabe inkompatibel. Auf ein solches Problem ist die Theorie inexakter Newton-Verfahren aus [19] nicht mehr anwendbar. An diese Stelle kann zwar eine Theorie inexakter Gauss-Newton-Verfahren treten, die jedoch bei weitem nicht so deutliche und praxisnahe Ergebnisse liefert.
Eine neue Sichtweise hilft über diese Schwierigkeit hinweg, indem die Diskretisierung erst nach der Linearisierung ausgeführt wird und man die Theorie inexakter Newton-Verfahren auf

das Ausgleichsproblem im unendlichdimensionalen Lösungsraum anwendet. Die Kombination von Least-Squares-Ansatz, Gauß-Newton-Verfahren und Theorie inexakter Newton-Verfahren ergänzt sich dann gegenseitig durch

- Kontrolle sowohl des Diskretisierungsfehlers, als auch des Fehlers in der Lösung des linearen Gleichungssystems durch berechenbare Fehlerschranken,
- Abstiegsbedingungen, deren Verletzung anzeigt, dass der aktuelle Lösungsraum wieder verfeinert werden sollte und dadurch
- Sicherstellung der Konvergenz des gesamten Verfahrens.

Mit diesem Verfahren, in Kapitel 3 als Gauß-Newton-Multilevelverfahren bezeichnet, wird der Verlauf der nichtlinearen Iteration dann gerade so gesteuert, dass auf jedem Level mit optimaler Komplexität bis zum Diskretisierungsfehler gelöst wird.
Damit steht für den Least-Squares-Ansatz eine komplette nichtlineare Lösungstheorie mit berechenbaren Genauigkeitsschranken zur Verfügung.

In der vorliegenden Arbeit werden die genannten Aspekte von nichtlinearen Lösungsverfahren und der Anwendung auf Multilevelverfahren für nichtlineare Ausgleichsprobleme ausführlich behandelt:
Die in [19] enthaltene Theorie inexakter Newton-Verfahren wird auf Nullstellenprobleme $F(x) = 0$ mit einer Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$ erweitert, indem mit Hilfe des Gauß-Newton-Verfahrens Abstiegsrichtungen für das Funktional $\|F(x)\|_2^2$ konstruiert werden. Unter Berücksichtigung von Abstiegsbedingungen wird dabei eine obere Schranke für den größten zugelassenen Fehler bei der Berechnung der Gauß-Newton-Korrektur hergeleitet. Damit ist eine Theorie inexakter Gauß-Newton-Verfahren für kompatible Probleme verfügbar. Daneben wird auch eine Konvergenztheorie zu inexakten Gauß-Newton-Verfahren für inkompatible Probleme (Minimierungsaufgaben, die nicht gleichzeitig Nullstellenprobleme sind) entwickelt.
Die so hergeleiteten Verfahren und Ergebnisse können dann auf Minimierungs- bzw. Nullstellenprobleme in Hilberträumen übertragen werden und führen in der Anwendung auf Finite-Element-Ausgleichsprobleme zu Fehlerschranken für die durch Iterationsverfahren, wie zum Beispiel Multilevelverfahren in der FE-Methode, verursachten Diskretisierungs- und Abbruchfehler. Auf diese Weise kann dann die Anzahl der zur Konvergenz notwendigen linearen Multilevelzyklen zur Lösung des Gauß-Newton-Systems auf dem feinsten Gitter kontrolliert werden. Neu ist ebenfalls die aus dem Ansatz folgende Gleichgewichtung von Diskretisierungsfehler und algebraischem Fehler, so dass zur Konvergenz sowohl eine ausreichende Exaktheit der Lösung der diskreten linearen Systeme, als auch eine sich dem Konvergenzverlauf anpassende Verfeinerungsstrategie notwendig ist. Kontrollmechanismen für beide Aspekte lassen sich direkt nach der hergeleiteten Theorie inexakter Gauß-Newton-Verfahren angeben.

Für die Lösung der linearen Gleichungssysteme, die aus Diskretisierungen des Lösungsraums $H^1(\Omega) \times H(\mathrm{div}, \Omega)$ hervorgehen, werden die in [3] und [29] angegebenen Glättungsverfahren für $\Omega \subseteq \mathbb{R}^2$ verwendet. Dabei wird das Potentialraumverfahren aus [29] auf nicht einfach zusammenhängende Gebiete mit wechselnden Randbedingungen erweitert. Im Rahmen eines Beispielproblems werden die Ansätze numerisch verglichen, um den Einfluss des Glätters beim Vergleich der Multilevelverfahren möglichst gering zu halten.

Weiterhin werden zwei verschiedene Varianten von Korrekturgleichungen innerhalb eines nichtlinearen Multilevelverfahrens ([9], "FAS") für den Least-Squares-Ansatz hergeleitet. Diese Varianten wurden bereits in [35] beschrieben. Hier werden diese Verfahren unter Verwendung

unterschiedlicher Glätter und Abbruchbedingungen weiter getestet und in den Kontext von inexakten Gauß-Newton-Verfahren in Hilberträumen gestellt.
Auch die Konvergenztheorie des neuen Gauß-Newton-Multilevelverfahrens wird unter Verwendung der Exaktheitsvoraussetzungen an das Gauß-Newton-Verfahren ausführlich untersucht. Insbesondere werden dabei die für Multilevelverfahren wichtigen Fragen nach der Verfeinerungsstrategie und nach dem Abbruch der linearen Iteration berücksichtigt.

Diese Ergebnisse werden in folgender Reihenfolge dargestellt:
Im ersten Kapitel werden Grundlagen nichtlinearer Lösungsverfahren wiederholt und einige Erweiterungen dazu hergeleitet. Auch finden sich dort eine Darstellung der benötigten Bestandteile der FE-Methode. In Kapitel 2 wird die Theorie für das nichtlineare Multilevelverfahren und das Verfahren der Linearisierung auf dem feinsten Level dargestellt und die sich ergebenden Algorithmen angegeben. Besonderes Augenmerk liegt dabei auf den Abbruchkriterien der verwendeten iterativen Verfahren. In Kapitel 3 werden die im ersten Kapitel erhaltenen Ergebnisse über nichtlineare Lösungsverfahren zur Herleitung eines kontrollierbaren Multilevelverfahrens für das zu minimierende Ausgleichsfunktional verwendet und ein implementierbarer Algorithmus des daraus resultierenden Gauß-Newton-Multilevelverfahrens angegeben. Das vierte Kapitel enthält die Darstellung eines anspruchsvollen Testproblems aus der Bodenmechanik und numerische Untersuchungen der in den vorhergehenden Kapiteln behandelten Gebiete. Dabei werden sowohl die Varianten des nichtlinearen Multilevelverfahrens, als auch die verschiedenen Glätter der $H(\mathrm{div}, \Omega)$-Probleme verglichen. Die Eigenschaften des Gauß-Newton-Verfahrens werden numerisch nachgewiesen und ein Vergleich der drei entwickelten Multilevelverfahren schließt das Kapitel ab.

Kapitel 1

Grundlagen

Die numerische Lösung nichtlinearer parabolischer Differentialgleichungen lässt sich in mehrere Teilprobleme untergliedern: Die Behandlung der zeitlichen Abhängigkeit, die Ortsdiskretisierung, die Behandlung der Nichtlinearität, und die Lösung der entstehenden linearen Gleichungssysteme.
In den folgenden Abschnitten wird ein Überblick zu diesen Aufgabenfeldern gegeben. Dabei werden für spätere Kapitel wichtige Ergebnisse zur Lösung dieser Teilaufgaben hergeleitet.

1.1 Behandlung nichtlinearer Gleichungen

Werden physikalische Vorgänge, die man durch mathematische Gleichungen beschreiben will, hinreichend vereinfacht bzw. idealisiert, so reichen *lineare* Gleichungen zur Beschreibung aus, wie z.B. bei der Verteilung der Kräfte in einem idealen Stabwerk oder der Beschreibung einer aufgehängten idealen Feder, die nach unten ausgelenkt wird. Werden jedoch realistischere Annahmen getroffen, etwa durch Einbeziehung von Materialeigenschaften oder auch äußeren Einflüssen, so reichen die linearen Gleichungen nicht mehr zur Beschreibung aus. Als Beispiel mag ein Gummiband anstelle der aufgehängten Feder dienen, das nicht mehr linear auf Auslenkungen reagiert, sobald diese groß genug sind.
Die Lösung von nichtlinearen Gleichungen ist daher ein wichtiges anwendungsbezogenes Problemfeld der numerischen Mathematik.

1.1.1 Newton-Verfahren

Gegeben sei das Problem

$$F(x_*) = 0 \tag{1.1}$$

mit $F : \mathbb{R}^n \to \mathbb{R}^n$ differenzierbar.
Das Newton-Verfahren besteht nun aus der Berechnung einer Folge $\{x_k\}$ nach folgendem Algorithmus:

1. Wähle $x_0 \in \mathbb{R}^n$. Setze $k = 0$.
2. Löse die lineare Gleichung
$$F'(x_k)\,\delta_k \;=\; -F(x_k). \tag{1.2}$$
3. Setze $x_{k+1} = x_k + \delta_k$, $k = k + 1$.
4. Abbruchkriterium erfüllt? Sonst gehe zu 2.

Algorithmus 1.1 : Newton-Verfahren

Dabei wird die Berechnung im Allgemeinen abgebrochen, sobald die gewünschte Genauigkeit erreicht ist, also $\|F(x_k)\|_2 \leq \epsilon$ gilt mit einem $\epsilon > 0$.
Das Newton-Verfahren ist, gewisse Lipschitzbedingungen an die Ableitung F' vorausgesetzt, lokal quadratisch konvergent (siehe auch [18] oder [42]). Der Bereich der Startnäherungen, für die das Newton-Verfahren konvergent ist, kann dabei durch Dämpfungsstrategien (z.B. linesearch, Trust-Region-Techniken) erheblich vergrößert werden.

1.1.2 Gauß-Newton-Verfahren

Das Newton-Verfahren ist nur für Probleme geeignet, die sich in der Form (1.1) schreiben lassen und für die $F'(x)$ für alle x invertierbar ist (was gerade im Fall $F : \mathbb{R}^n \to \mathbb{R}^m,\ m \neq n$ nicht möglich ist). Ist das Problem in der Form

$$g(x_*) \stackrel{!}{=} \min_{x \in \mathbb{R}^n} g(x) \tag{1.3}$$

gestellt, so ist das Newton-Verfahren daher nicht unmittelbar anwendbar. Ist das Minimum eindeutig, so ergibt jedoch die notwendige Bedingung für die Minimalstelle

$$g'(x_*) := \frac{\partial g}{\partial x}|_{x=x_*} \stackrel{!}{=} 0\,, \tag{1.4}$$

ein Problem der Form (1.1), das wiederum mit einem Newton-Verfahren lösbar ist. Dazu müssen lineare Operatorgleichungen mit g'' gelöst werden. Diese zweite Ableitung von g ist aber im Allgemeinen nicht vollständig verfügbar bzw. aufwendig zu berechnen.
Eine besondere Klasse von Minimierungsproblemen sind die nichtlinearen Ausgleichsprobleme. Dabei hat g aus (1.3) die Gestalt

$$g(x) \;=\; \|F(x)\|_2^2 \tag{1.5}$$

mit einer zweimal differenzierbaren Funktion $F : \mathbb{R}^n \to \mathbb{R}^m, m \geq n$. Auch hier versucht man, ausgehend von einer Startnäherung x_0, eine Folge von Iterierten $\{x_k\}$ zu berechnen, die gegen die Lösung x_* konvergiert.
Bei Anwendung des Gauß-Newton-Verfahrens wird dazu in jedem Schritt die Funktion F um die aktuelle Näherung x_k linearisiert und statt der eigentlichen Funktion g die sich daraus ergebende quadratische Approximation minimiert. Diese Minimalstelle wird dann gleich x_{k+1} gesetzt, um die verbesserte Näherung zu erhalten. Das nichtlineare Ausgleichsproblem (1.3),(1.5) wird also durch eine Folge von linearen Ausgleichsproblemen ersetzt:

1. Wähle $x_0 \in \mathbb{R}^n$. Setze $k = 0$.
2. Löse das Minimierungsproblem
$$\|F(x_k) + F'(x_k)\,\delta_k\|_2^2 \stackrel{!}{=} \min_{\delta_k} \tag{1.6}$$
3. Setze $x_{k+1} = x_k + \delta_k$, $k = k + 1$.
4. Abbruchkriterium erfüllt? Sonst gehe zu 2.

Algorithmus 1.2 : Gauß-Newton-Verfahren

Zum Vergleich von Newton-Verfahren und Gauß-Newton-Verfahren bildet man die Normalengleichungen zu (1.6):
$$F'(x_k)^T F'(x_k)\,\delta_k = -F'(x_k)^T F(x_k) \tag{1.7}$$
und die zu lösende Newton-Gleichung zu (1.4):
$$(F'(x_k)^T F'(x_k) + \underbrace{F''(x_k)F(x_k)}_{=:S(x_k)})\,\delta_k = -F'(x_k)^T F(x_k). \tag{1.8}$$

Beim Vergleich von (1.7) und (1.8) wird deutlich, dass das Gauß-Newton-Verfahren somit auch als *inexaktes* Newton-Verfahren angesehen werden kann, d.h. als Newton-Verfahren, in dem die Lösung der eigentlichen Newton-Gleichung (1.8) durch die Lösung von (1.7) angenähert wird. Motiviert wird dieses Vorgehen in dem Fall $g(x_*) = 0$ (*kompatibles* Problem) dadurch, dass $S(x_k)$ relativ klein ist, wenn x_k bereits nah an x_* ist und damit auch $F(x_k)$ sich dem minimalen Wert 0 annähert.

Andererseits ist es auch nicht sinnvoll, die Gauß-Newton-Gleichung (1.7) exakt zu lösen, wenn x_k noch weit von der Lösung x_* entfernt ist, da dort das quadratische Modell weniger mit der tatsächlichen Funktion g übereinstimmt. Man spricht dann von *inexakten* Gauß-Newton-Verfahren. In diesem Fall (und im Fall $g(x_*) > 0$) reduziert sich die schnelle lokal-quadratische Konvergenz des exakten Newton-Verfahrens auf nur noch lokal-lineare Konvergenz des inexakten Gauß-Newton-Verfahrens.

1.1.3 Konvergenz des Gauß-Newton Verfahrens

Unter geeigneten Voraussetzungen an das Funktional g konvergiert das exakte Gauß-Newton-Verfahren lokal quadratisch für kompatible ($g(x_*) = 0$) und lokal linear für inkompatible Probleme. In [17, Satz 10.2.1] findet man dazu folgendes Ergebnis :

Satz 1.1 *Sei $F : \mathbb{R}^n \to \mathbb{R}^m, m \geq n$ und $g(x) = \frac{1}{2}F(x)^T F(x)$ zweimal stetig differenzierbar in der offenen, konvexen Menge $D \subset \mathbb{R}^n$. Ist die Jakobi-Matrix $F'(x)$ in D Lipschitz-stetig mit Konstante γ und $\|F'(x)\|_2 \leq \mu$ für alle $x \in D$ und gibt es ein $x_* \in D$ und $0 \leq \sigma < \lambda$ so, dass $F'(x_*)^T F(x_*) = 0$, λ kleinster Eigenwert von $F'(x_*)^T F'(x_*)$ ist und*
$$\|(F'(x) - F'(x_*))^T F(x_*)\|_2 \leq \sigma \|x - x_*\|_2$$

für alle $x \in D$ gilt, dann existiert eine ϵ-Umgebung um x_*, so dass alle Iterierten x_k des Gauß-Newton-Verfahrens darin liegen und gegen x_* konvergieren und es gilt :

$$\|x_{k+1} - x_*\|_2 \leq \frac{c\sigma}{\lambda}\|x_k - x_*\|_2 + \frac{c\mu\gamma}{2\lambda}\|x_k - x_*\|_2^2 \tag{1.9}$$

mit $c \in (1, \lambda/\sigma)$.

Dieses Resultat wird bewiesen, indem induktiv die Differenz $x_{k+1} - x_*$ bestimmt und folgende Abschätzung des Fehlers bei linearer Approximation der Nichtlinearität benutzt wird (vgl. [17, Lemma 4.1.12]):

Lemma 1.2 *Sei $F : \mathbb{R}^n \to \mathbb{R}^m, m \geq n$, in der offenen, konvexen Menge $D \subset \mathbb{R}^n$ Lipschitzstetig differenzierbar mit Konstante γ. Dann ist für jedes $x + p \in D$*

$$\|F(x+p) - F(x) - F'(x)\,p\|_2 \leq \frac{\gamma}{2}\|p\|_2^2. \tag{1.10}$$

Wie oben angeführt ist eine exakte Lösung des Gauß-Newton-Systems weder durchführbar (aufgrund der nur iterativ lösbaren großen linearen Gleichungssysteme) noch wünschenswert, denn das Modell stellt weit entfernt von der Lösung x_* keine gute Approximation an die nichtlineare Funktion g nach (1.5) dar, wie sofort an Lemma 1.2 abgelesen werden kann. Stattdessen wird dazu übergegangen, die linearen Teilprobleme in jedem Gauß-Newton-Schritt nur näherungsweise zu lösen.

1.1.4 Inexakte Newton-Verfahren

Werden die Gleichungen (1.2) oder (1.7) nicht exakt gelöst, so spricht man von inexakten Verfahren. Natürlich muss die Größe des Fehlers, der in Kauf genommen wird, dabei unter Kontrolle gehalten werden. In diesem Abschnitt wird für das Newton-Verfahren von einer Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$ ausgegangen und jeweils mit δ_k die exakte Korrektur des Newton-Verfahrens, sowie mit $\tilde{\delta}_k = \delta_k + \epsilon_k$ die mit einem Fehler ϵ_k behaftete inexakte Korrektur bezeichnet. Die auf x_k folgende Iterierte des inexakten Verfahrens ergibt sich dann durch $x_{k+1} = x_k + \tilde{\delta}_k$.
Es sollen zwei unterschiedliche Ansätze untersucht werden: die Angabe von Abstiegsrichtungen für das Funktional g aus (1.5) und die Überprüfung der Übereinstimmung von g und seinem quadratischen Modell. Für den ersten Ansatz ist aus der Theorie bekannt, dass ein Verfahren der Verwendung von Abstiegsrichtungen $\tilde{\delta}_k$ bei der Berechnung der Iterierten x_k unter bestimmten weiteren Voraussetzungen an g konvergiert (Abstiegsverfahren, siehe [26]). Da sowohl Newton- als auch Gauß-Newton-Verfahren mit dem quadratischen Funktional (1.6) anstelle des ursprünglichen Funktionals arbeiten, ist die Überprüfung der Übereinstimmung dieser beiden Modelle ebenfalls ein Hinweis auf akzeptierbare Korrekturen: Stimmen sie gut überein, so wird ein größerer Fehler zugelassen werden können.
In diesem Abschnitt soll das Newton-Verfahren untersucht werden, es gelte also für die Korrekturrichtung δ_k:

$$\delta_k = -F'(x_k)^{-1}\,F(x_k).$$

Wegen

$$\nabla g(x_k) = 2F'(x_k)^T F(x_k)$$

ist dann

$$\nabla g(x_k)^T \delta_k = -2F(x_k)^T F'(x_k)\,F'(x_k)^{-1}\,F(x_k) = -2\|F(x_k)\|_2^2 = c_{k,N}.$$

Damit schliesst δ_k mit dem Gradienten von g an der Stelle x_k einen Winkel von kleiner als 90° ein, ist also eine Abstiegsrichtung.
Tatsächlich kann δ_k jedoch nur bis auf den Fehler ϵ_k berechnet werden. Die inexakte Korrektur $\tilde{\delta}_k = \delta_k + \epsilon_k$ ist daher genau dann eine Abstiegsrichtung, wenn

$$-c_{k,N} > |\nabla g(x_k)^T \epsilon_k| = \underbrace{|\cos(\angle(\nabla g(x_k), \epsilon_k))|}_{\leq 1} \|\nabla g(x_k)\|_2 \|\epsilon_k\|_2$$

gilt. Umgestellt erhält man als Bedingung

$$\|\epsilon_k\|_2 \leq \eta_k \frac{\|F(x_k)\|_2^2}{\|F'(x_k)^T F(x_k)\|_2} \tag{1.11}$$

mit $\eta_k \in [0, 1)$. Damit die Richtungen $\tilde{\delta}_k$ jedoch auch asymptotisch für $k \to \infty$ Abstiegsrichtungen bleiben, muss man noch $\eta_k \leq \eta < 1$ fordern. Diese Sichtweise führt dann ohne weiteres auf die Anwendung der Theorie für Abstiegsrichtungen.

Eine andere, sehr allgemeine Darstellung von inexakten Newton-Verfahren findet man in [16] und [19]. Der Grundgedanke besteht darin, jede Korrektur δ eine inexakte Newton-Korrektur zu nennen, für die

$$\|F(x_k) + F'(x_k)\delta\|_2 \ \leq \ \eta_k \|F(x_k)\|_2$$

mit einer Kontrollfolge $\{\eta_k\}$, $\eta_k \leq \eta < 1$, gilt. In dieser Gleichung wird gefordert, dass der Wert des quadratischen Modells des Funktionals verringert wird. Wird dies zu einer Abschätzung für ϵ_k umgeformt, ergibt sich

$$\|F'(x_k)\epsilon_k\|_2 \leq \eta_k \|F(x_k)\|_2 \tag{1.12}$$

als Bedingung für die Verringerung des quadratischen Modells.
Bei dem Vergleich dieser beiden Ansätze ergibt sich nach folgendem Lemma (Sonderfall $m = n$ wegen Newton-Verfahren), dass sie für festes x äquivalent sind.

Lemma 1.3 *Sei $F(x) : \mathbb{R}^n \to \mathbb{R}^m, m \geq n$, differenzierbar, $F'(x)$ habe vollen Spaltenrang und $g(x) = \|F(x)\|_2^2$. Sei $x \in \mathbb{R}^n$ beliebig. Dann ist die Richtung $\delta \in \mathbb{R}^n$ (bis auf eine Skalierung ω) genau dann eine Abstiegsrichtung des Funktionals g an der Stelle x, wenn*

$$\|F(x) + F'(x)\delta\|_2 \ \leq \ \eta \|F(x)\|_2 \tag{1.13}$$

mit einem $\eta < 1$ gilt.

Beweis: Sei zunächst δ eine Abstiegsrichtung des Funktionals g an der Stelle x.
Es gilt also

$$\nabla g(x)^T \delta = 2F(x)^T F'(x)\, \delta < 0. \tag{1.14}$$

Sei ρ der Spektralradius von $F'(x)^T F'(x)$. Dann gilt

$$\|F'(x)\, y\|_2^2 \leq \rho \|y\|_2^2 \ \ \forall\, y \in \mathbb{R}^n. \tag{1.15}$$

Setze $\hat{\delta} = \underbrace{-\frac{F(x)^T F'(x)\, \delta}{\rho \|\delta\|_2^2}}_{>0}\, \delta = \omega\, \delta$. Dann ist $\hat{\delta}$ immer noch eine Abstiegsrichtung und (1.15) gilt auch für $y = \hat{\delta}$.

Es gilt

$$\begin{aligned}
\|F(x) + F'(x)\,\hat{\delta}\|_2^2 &= \|F(x)\|_2^2 + 2F(x)^T F'(x)\,\hat{\delta} + \|F'(x)\,\hat{\delta}\|_2^2 \\
&\leq \|F(x)\|_2^2 + 2F(x)^T F'(x)\,\hat{\delta} + \rho\|\hat{\delta}\|_2^2 \\
&= \|F(x)\|_2^2 + (-2\frac{(F(x)^T F'(x)\,\delta)^2}{\rho\|\delta\|_2^2} + \frac{(F(x)^T F'(x)\,\delta)^2}{\rho\|\delta\|_2^2}) \\
&= \|F(x)\|_2^2 - \frac{(F(x)^T F'(x)\,\delta)^2}{\rho\|\delta\|_2^2} \\
&= \tilde{\eta}\|F(x)\|_2^2
\end{aligned}$$

mit

$$\tilde{\eta} = 1 - \frac{(F(x)^T F'(x)\,\delta)^2}{\|F(x)\|_2^2 \rho\|\delta\|_2^2} < 1.$$

Offensichtlich ist $\tilde{\eta} \geq 0$. Damit gilt (1.13) mit $\eta = \sqrt{\tilde{\eta}}$ und der skalierten Richtung $\hat{\delta} = \omega\,\delta$.

Gelte nun Ungleichung (1.13) für ein $\delta \in \mathbb{R}^n$ und $\eta < 1$.
Dann ist wie oben

$$\|F(x) + F'(x)\,\delta\|_2^2 = \|F(x)\|_2^2 + 2F(x)^T F'(x)\,\delta + \|F'(x)\,\delta\|_2^2 \leq \eta^2\|F(x)\|_2^2.$$

Wegen (1.14) ist dann also

$$\nabla g(x)^T\delta \;\leq\; \underbrace{(\eta^2 - 1)}_{<0}\|F(x)\|_2^2 \;-\; \|F'(x)\|_2^2 \;<\; 0,$$

und damit ist δ eine Abstiegsrichtung. □

Um die Konvergenz von Verfahren, die die Fehlerschranken nach (1.12) oder (1.11) einhalten, nachweisen zu können, ist jedoch zusätzlich zu zeigen, dass auch eine gewisse Reduktion im ursprünglichen Funktional stattfindet. Dazu kann eine der folgenden Abschätzungen gefordert werden:

$$\begin{aligned}
&\|F(x+\delta)\|_2 \;\leq\; \|F(x)\|_2 \;+\; 2tF(x)^T F'(x)\delta \;\;, 0 < t < 1, \\
&\|F(x+\delta)\|_2 \;\leq\; (1 - t(1-\eta))\|F(x)\|_2 \;\;, 0 < t < 1 \\
\text{oder} \qquad &\|F(x+\tilde{\omega}\delta)\|_2 \;\stackrel{!}{=}\; \min_{\omega\in\mathbf{R}}\|F(x+\omega\delta)\|_2
\end{aligned} \tag{1.16}$$

Diese Bedingungen können zum Beispiel mit der Anwendung von Line-Search-Verfahren erfüllt werden. Zu Details siehe [17], [19] und [26].
Dass geeignete Suchrichtungen δ existieren, wird in [19] unter allgemeinen Voraussetzungen gezeigt. Weiterhin findet man dort einen Konvergenzbeweis, wenn bei der Fehlerkontrolle bestimmte Forderungen an die Kontrollfolge $\{\eta_k\}$ erfüllt werden. Dieses Ergebnis wird im Folgenden auch zur Unterstützung der zu entwickelnden Resultate für das Gauß-Newton-Verfahren von Bedeutung sein.
Nähere Ausführungen und Details zur Kontrolle der Fehler bei der Berechnung von inexakten Newton-Korrekturen werden auch in [16] und [30] angegeben.

1.1.5 Inexakte Gauß-Newton-Verfahren für kompatible Probleme

Das Minimierungsproblem (1.3) mit g nach (1.5), das mit dem Gauß-Newton-Verfahren behandelt werden kann, stellt hinsichtlich der Kontrolle der Ungenauigkeit ähnliche Probleme wie das

Newton-Verfahren. Steht die zweite Ableitung $\nabla^2 g(x)$ zur Verfügung, so kann dieses Minimierungsproblem auch mit dem Newton-Verfahren, angewendet auf $\nabla g(x) = 0$, gelöst werden. Dies ist jedoch im Allgemeinen nicht der Fall, weswegen im Folgenden auf Informationen über F'' verzichtet werden soll.
Zunächst werden die Ideen aus dem inexakten Newton-Verfahren auf die Übertragbarkeit auf den Gauß-Newton-Fall überprüft.
Für die exakte Gauß-Newton-Korrektur gilt

$$\delta_k = -[F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T F(x_k)$$

und damit auch

$$\nabla g(x_k)^T \delta_k = -2[F'(x_k)^T F(x_k)]^T \left[F'(x_k)^T F'(x_k)\right]^{-1} \left[F'(x_k)^T F(x_k)\right] \;=\; c_{k,GN}.$$

Da bei vollem Rang von $F'(x_k)$ die Matrix $(F'(x_k)^T F'(x_k))^{-1}$ positiv definit ist, ist $\nabla g(x_k)^T \delta_k$ kleiner als 0 und damit der Winkel zwischen der exakten Gauß-Newton-Richtung δ_k und $\nabla g(x_k)$ größer als 90 Grad. Also ist die exakte Gauß-Newton-Korrektur eine Abstiegsrichtung für das Problem

$$g(x) \;=\; \|F(x)\|_2^2 \stackrel{!}{=} \min$$

an der Stelle x_k. Wird die lineare Gleichung jedoch wieder nur bis auf einen Fehler ϵ_k genau gelöst, so ist $\tilde{\delta}_k = \delta_k + \epsilon_k$ genau dann eine Abstiegsrichtung, wenn

$$-c_{k,GN} > |\nabla g(x_k)^T \epsilon_k| = |2F(x_k)^T F'(x_k)\epsilon_k|$$

gilt. Der Ausdruck $|2F(x_k)^T F'(x_k)\epsilon_k|$ kann durch $\|2F(x_k)\|_2 \|F'(x_k)\epsilon_k\|$ abgeschätzt werden. Dies führt zu der Bedingung

$$\|F'(x_k)\epsilon_k\|_2 \;\leq\; \eta_k \, \frac{|c_{k,GN}|}{\|2F(x_k)\|_2}. \tag{1.17}$$

Zu fordern ist dabei $0 < \eta_k \leq \eta < 1$, damit die inexakte Richtung $\tilde{\delta}_k$ auch für $k \to \infty$ eine Abstiegsrichtung an g bleibt. Damit wird bei Verwendung dieser Richtung nach Lemma 1.3 das quadratische Modell des Funktionals reduziert. Im Falle eines kompatiblen Problems ($g(x_*) = 0$) kann dann an diesem Punkt die Theorie aus [19] eingesetzt werden.
Im allgemeinen, nichtkompatiblen Fall hingegen kann das Gauß-Newton-Verfahren natürlich nicht als inexaktes Newton-Verfahren aufgefasst werden, da kein Nullstellenproblem zu Grunde liegt.
Um auch in diesem Fall eine Bedingung an die Genauigkeit zu erhalten, mit der das lineare System (1.7) gelöst werden muss, wird auf das Nullstellenproblem (1.4) zurückgegriffen. Für sowohl kompatible wie nichtkompatible Aufgabenstellungen ist das Minimierungsproblem (1.3) eindeutig mit dem Nullstellenproblem (1.4) verbunden. Im Sinne von Ausgleichsformulierungen liegt dann die Umformulierung des Problems zu

$$\text{Finde } x_* \text{ mit } \|F'(x_*)^T F(x_*)\|_2 \stackrel{!}{=} \min_{x \in \mathbf{R}^n} \|F'(x)^T F(x)\|_2 \tag{1.18}$$

nahe. Eine zu (1.13) analoge Gleichung könnte dann folgendermaßen lauten:

$$\|F'(x_k)^T F(x_k) + F'(x_k)^T F'(x_k)\tilde{\delta}_k\|_2 \leq \eta_k \|F'(x_k)^T F(x_k)\|_2. \tag{1.19}$$

Damit ergibt sich unmittelbar als Genauigkeitsbedingung

$$\|F'(x_k)^T F'(x_k)\epsilon_k\|_2 \;\leq\; \eta_k \|F'(x_k)^T F(x_k)\|_2 \tag{1.20}$$

sowie eine der zu (1.16) analogen Beziehungen über die tatsächliche Reduktion von $\|F'(x)^T F(x)\|_2$.
Die Schwierigkeit des Vergleichs dieser beiden Ansätze für das Gauß-Newton-Verfahren besteht vor allem in der bereits in der Grundgleichung des Gauß-Newton-Verfahrens vorformulierten Inexaktheit durch die ungenaue zweite Ableitung von g. Daher sind die beiden angegebenen Genauigkeitsansätze (1.17) und (1.20) für das Gauß-Newton-Verfahren auch nicht äquivalent wie für das Newton-Verfahren (Lemma 1.3).
Nach den vorhergehenden Überlegungen stehen nun also zwei mögliche theoretische Ansätze für die Formulierung von Konvergenzergebnissen für inexakte Gauß-Newton-Verfahren mit den folgenden Voraussetzungen zur Verfügung. Der erste dieser Ansätze ist so eng mit inexakten Newton-Verfahren verbunden, dass es sich um ein kompatibles Problem handeln muss, damit die Theorie anwendbar ist. Bei inkompatiblen Problemen sind zusätzliche Kenntnisse z.B. über die zweite Ableitung von F notwendig, um die beim Newton-Ansatz für (1.18) vernachlässigten Terme abschätzen zu können.

Konvergenz des inexakten Gauß-Newton-Verfahrens im Rahmen der Theorie inexakter Newton-Verfahren

Voraussetzung 1.4 *Sei $F : \mathbb{R}^n \to \mathbb{R}^m, m \geq n$, differenzierbar und sei $x_* \in \mathbb{R}^n$ die eindeutige Nullstelle von $F : F(x_*) = 0$. Für alle $x \in \mathbb{R}^n$ habe $F'(x)$ vollen Spaltenrang und sei Lipschitzstetig mit Konstante γ.*
Sei zu $x \in \mathbb{R}^n$ ein Vektor $\delta \in \mathbb{R}^n$ durch die exakte Lösung des Gauß-Newton-Systems

$$F'(x)^T F'(x)\,\delta \;=\; -F'(x)^T F(x) \tag{1.21}$$

bestimmt und gelte für einen weiteren Vektor ϵ (später : Fehler in der Lösung des GN-Systems)

$$\|F'(x)\epsilon\|_2 \;<\; \frac{(F'(x)^T F(x))^T (F'(x)^T F'(x))^{-1} (F'(x)^T F(x))}{\|F(x)\|_2}. \tag{1.22}$$

Sei Voraussetzung 1.4 erfüllt. Ziel ist die Konstruktion einer gegen x_* konvergenten Folge $\{x_k\}_{k=0}^{\infty}$ mit Hilfe der Ergebnisse aus [19].
Der Grundgedanke ist dort die Angabe von Bedingungen für eine ausreichende Reduktion sowohl der Funktion als auch des quadratischen Modells der Funktion. Es wird also ausschließlich mit einem Newton-Ansatz gearbeitet.
Eine Erweiterung dieses Ansatzes wird in diesem Abschnitt gegeben, da hier die Ergebnisse aus [19] auf die Suche nach Nullstellen von Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m, m > n$ erweitert werden, denn das reine Newton-Verfahren kann in diesem Fall natürlich nicht mehr angewendet werden ($F'(x) \in \mathbb{R}^{m\times n}$ nicht invertierbar). Hingegen ist wegen der angenommenen Voraussetzung des vollen Spaltenrangs von F' die Matrix $F'(x)^T F'(x)$ stets invertierbar, so dass das zugehörige Gauß-Newton-System (1.21) eindeutig lösbar ist.
Sei $x \in \mathbb{R}^n$ beliebig. Definiert man das Funktional $g : \mathbb{R}^n \to \mathbb{R}$ nach (1.5), so erhält man, wie oben angegeben, bei Einhaltung der Bedingung (1.22) eine Abstiegsrichtung $(\delta + \epsilon)$ für g an der Stelle x in dem Sinne, dass

$$\nabla g(x)^T(\delta + \epsilon) = 2F(x)^T F'(x)(\delta + \epsilon) < 0$$

ist. Nach Lemma 1.3 ergibt sich dann, dass es eine Richtung $\hat{\delta}$ gibt, für die ein $\hat{\eta} < 1$ existiert mit

$$\|F(x) + F'(x)\hat{\delta}\|_2 \;\leq\; \hat{\eta}\,\|F(x)\|_2.$$

Nach [19, Lemma 3.1] existiert dann zu beliebigem $0 < t < 1$ ein $\tilde{\delta} \in \mathbb{R}^n$ und ein $\eta < 1$, so dass die beiden Ungleichungen

$$\begin{aligned} \|F(x) + F'(x)\tilde{\delta}\|_2 &\leq \eta\|F(x)\|_2 \\ \|F(x+\tilde{\delta})\|_2 &\leq [1 - t(1-\eta)]\|F(x)\|_2 \end{aligned}$$

gelten. $\tilde{\delta}$ und η hängen dabei von der Stelle x ab: $\tilde{\delta} = \tilde{\delta}(x), \eta = \eta(x)$.
Nun wird für beliebiges $x_0 \in \mathbb{R}^n$ und beliebiges festes $t \in (0,1)$ die Folge $\{x_k\}_{k=0}^{\infty}$ konstruiert: Setze $x_{k+1} = x_k + \tilde{\delta}(x_k)$ und $\eta_k = \eta(x_k)$. Dann erfüllen die x_k für alle k die Bedingungen

$$\|F(x_k) + F'(x_k)(x_{k+1} - x_k)\|_2 \leq \eta_k\|F(x_k)\|_2, \tag{1.23}$$

$$\|F(x_{k+1})\|_2 \leq [1 - t(1-\eta_k)]\|F(x_k)\|_2. \tag{1.24}$$

Dies sind genau die entscheidenden Voraussetzungen aus [19], wo ein solches Verfahren der Konstruktion der Folge $\{x_k\}$ globales inexaktes Newton-Verfahren mit Kontrollfolge η_k genannt wird. Nach [19, Theoreme 3.3, 3.4] konvergiert diese Folge genau dann gegen das eindeutige Minimum x_* von g, wenn

$$\sum_{k=0}^{\infty}(1 - \eta_k) \tag{1.25}$$

divergiert.
In der Praxis wird man versuchen, die Bedingungen (1.23), (1.24) direkt mit der Definition $x_{k+1} = x_k + \delta_k + \epsilon_k$ zu erfüllen, wobei δ_k und $\epsilon_k = \epsilon$ analog zu den Definitionen in Voraussetzung 1.4 definiert werden. Sind die Bedingungen (insbesondere (1.24)) auf diese Weise nicht erfüllbar, ist nach der Berechnung der Korrektur $(\delta_k + \epsilon_k)$ ein geeignetes Dämpfungsverfahren anzuschließen denn dies ist das Vorgehen im konstruktiven Beweis des oben angegebenen Lemmas [19, Lemma 3.1]. Definiert man dabei

$$\begin{aligned} x_{k+1} &= x_k + \omega_k(\delta_k + \epsilon_k) \\ \eta_k &= \frac{\|F(x_k) + \omega_k F'(x_k)(\delta_k + \epsilon_k)\|}{\|F(x_k)\|_2} \end{aligned}$$

(mit einem geeigneten ω_k aus dem Dämpfungsverfahren so, dass x_k, x_{k+1} die Bedingungen (1.23) und (1.24) erfüllen), bleibt dann noch die Divergenzbedingung in (1.25) nachzuweisen: Gilt $\eta_k \leq \eta < 1$ so ist die Divergenz der Reihe über $(1-\eta_k)$ offensichtlich. Der ungünstigste anzunehmende Fall ist somit $\eta_k \to 1$. Für diesen Fall definiert man als Hilfsgröße

$$c_{k+1} := (\eta_{k+1} - \eta_k)(k+1)k\,.$$

Gilt nun $c_k = O(\frac{1}{k})$, folgt daraus die Divergenzbedingung (1.25), was später in Satz 3.4 für den Hilbert-Raum-Fall gezeigt wird. Die Bedingung an das Verhalten der c_k wird in diesem Fall mit der Erfüllung der Voraussetzungen (1.23) und (1.24) in Konkurrenz treten, die jedoch dabei Vorrang besitzen sollten.
Im Allgemeinen ist die Wahl der Startnäherung x_0 beliebig, solange nur die Konvergenzbedingungen (1.23) und (1.24) für alle k erfüllt werden.

Die Anwendung der Idee einer zusätzlichen Korrektur der inexakten Gauß-Newton-Iterierten über einem von allen oder einem Teil der Abstiegsrichtungen $\tilde{\delta}_l$ aufgespannten Unterraum führt zu einem beschleunigten inexakten (Gauß-)Newton-Verfahren nach [22]. Wie dort vorgeschlagen, kann das Funktional $\|F(x)\|_2^2$ direkt für die Erfüllung einer "Minimal Residual"-Bedingung

verwendet werden, indem nach Berechnung der Abstiegsrichtung $\tilde{\delta}_k$ die nächste Iterierte entsprechend

$$\begin{aligned} x_{k+1} &= x_k + \hat{\tilde{\delta}}_k \quad \text{mit} \\ \|F(x_k + \hat{\tilde{\delta}}_k)\|_2^2 &\stackrel{!}{=} \min_{y \in \mathbf{R}^{k-m}} \|F(x_k + \sum_{l=m}^{k} y_k \, \tilde{\delta}_k)\|_2^2 \end{aligned} \tag{1.26}$$

gesetzt wird (mit einem $m \geq 0$).
Die numerischen Ergebnisse in [22] zeigen eine Beschleunigung und Stabilisierung des nichtlinearen Konvergenzprozesses, auch wenn keine ausführliche Theorie über Ausmaß und Abschätzbarkeit dieser Effekte vorhanden ist. Zu betonen ist, dass bei Einhaltung der Genauigkeitsbedingungen (1.23) und (1.24) für ein mit (1.26) beschleunigtes Verfahren die oben angegebene Konvergenztheorie bestehen bleibt.
Auch bei der Übertragung des Verfahrens auf den Fall der Minimierung in Hilberträumen kann eine Modifikation der Korrektur entsprechend (1.26) vorgenommen werden. Die Untersuchungen in [50] zeigen für anspruchsvolle nichtlineare Probleme auch in diesem Fall eine Stabilisierung und Beschleunigung des nichtlinearen Lösungsverfahrens.

Konvergenz des inexakten Gauß-Newton-Verfahrens in angepassten Normen

Die Theorie des vorigen Abschnitts lehnt sich stark an die Theorie für inexakte Newton-Verfahren an. In diesem Abschnitt wird nun eine Konvergenzaussage ohne Rückgriff auf das inexakte Newton-Verfahren hergeleitet.

Voraussetzung 1.5 *Sei $F : \mathbf{R}^n \to \mathbf{R}^m, m > n$, differenzierbar und sei $x_* \in \mathbf{R}^n$ die eindeutige Nullstelle von $F : F(x_*) = 0$. Für alle $x \in \mathbf{R}^n$ habe $F'(x)$ vollen Spaltenrang und sei Lipschitzstetig mit Konstante γ. Weiterhin gebe es eine Konstante $\mu > 0$, so dass*

$$\frac{1}{\mu} \leq \|F'(x)\|_2 \leq \mu \tag{1.27}$$

gleichmäßig in x gilt.
Sei zu $x \in \mathbf{R}^n$ ein Vektor $\delta \in \mathbf{R}^n$ durch die exakte Lösung des Gauß-Newton-Systems

$$F'(x)^T F'(x)\, \delta = -F'(x)^T F(x) \tag{1.28}$$

bestimmt und gelte für einen weiteren Vektor ϵ

$$\|F'(x_*)^T F'(x_*)\epsilon\|_2 \leq \eta(x)\, \|F'(x)^T F(x)\|_2 \tag{1.29}$$

mit $\eta(x) < 1$.

Zunächst können aus der Voraussetzung 1.5 drei Hilfsresultate hergeleitet werden:

Lemma 1.6 (Lipschitz-Stetigkeit von $F'(x)^T F'(x)$) *Sei F wie in Voraussetzung 1.5. Es gilt also*

$$\begin{aligned} \|F'(x)\|_2 &\leq \mu \quad \forall x \in \mathbf{R}^n \\ \|F'(x) - F'(y)\|_2 &\leq \gamma \|x - y\|_2 \quad \forall x, y \in \mathbf{R}^n. \end{aligned}$$

Dann gilt

$$\|F'(x)^T F'(x) - F'(y)^T F'(y)\|_2 \leq c \|x - y\|_2 \tag{1.30}$$

mit $c = 2\mu\gamma$.

Beweis :

$$\begin{aligned}
&\|F'(x)^T F'(x)\, z - F'(y)^T F'(y)\, z\|_2 \\
&\qquad = \|F'(x)^T [F'(x) - F'(y)]\, z \;-\; [F'(y)^T - F'(x)^T] F'(y)\, z\|_2 \\
&\qquad \le \|F'(x)\|_2 \|[F'(x) - F'(y)]\, z\|_2 \;+\; \|F'(y)^T - F'(x)^T\|_2 \|F'(y)\, z\|_2 \\
&\qquad \le 2\mu\gamma \|x - y\|_2 \|z\|_2.
\end{aligned}$$

□

In den folgenden zwei Resultaten werden, im Ergebnis ähnlich zu Lemma 1.2, für die lineare Approximation von $F'(x+s)^T F(x+s)$ Abschätzungen mit Hilfe einer unvollständigen ersten Ableitung hergeleitet. Da die vollständige erste Ableitung von $F'(x)^T F(x)$ die zweite Ableitung von F beinhaltet, die im Gauß-Newton-Verfahren nicht berücksichtigt wird, soll dieser Teil auch in der Abschätzung nicht benutzt werden, was für kompatible Probleme in einer Umgebung der Lösung möglich ist, da F' dort lokal Lipschitz-stetig ist. Im Folgenden bezeichne U stets eine solche, genügend kleine Umgebung.

Lemma 1.7 (Lineare Approximation, erste Version) *Sei F wie in Voraussetzung 1.5. Es gilt also*

$$\begin{aligned}
\|F'(x)\|_2 &\le \mu \\
\|F'(x) - F'(y)\|_2 &\le \gamma \|x - y\|_2.
\end{aligned}$$

Dann gilt

$$\|F'(x_*)^T F(x_*) - F'(x)^T F(x) - F'(x)^T F'(x)(x_* - x)\|_2 \le c\|x - x_*\|_2^2 \tag{1.31}$$

mit $c = \frac{\mu\gamma}{2}$ für alle $x \in U$.

Beweis :
Nach Lemma 1.2 gilt wegen der Lipschitz-Stetigkeit von F'

$$\|F(x) - F(y) - F'(y)^T (x - y)\|_2 \;\le\; \frac{\gamma}{2}\|x - y\|_2^2. \tag{1.32}$$

Da $F(x_*) = 0$ ist, gilt $F'(x_*)^T F(x_*) = F'(x)^T F(x_*)$ und damit ist

$$\begin{aligned}
& F'(x_*)^T F(x_*) - F'(x)^T F(x) - F'(x)^T F'(x)(x_* - x) \\
= \;& F'(x)^T F(x_*) - F'(x)^T F(x) - F'(x)^T F'(x)(x_* - x) \\
= \;& F'(x)^T [F(x_*) - F(x) - F'(x)(x_* - x)].
\end{aligned}$$

Wendet man die Norm auf beiden Seiten an und benutzt (1.32) und die Beschränktheit von F', erhält man

$$\begin{aligned}
& \|F'(x_*)^T F(x_*) - F'(x)^T F(x) - F'(x)^T F'(x)(x_* - x)\|_2 \\
= \;& \|F'(x)^T [F(x_*) - F(x) - F'(x)(x_* - x)]\|_2 \\
\le \;& \|F'(x)^T\|_2 \, \|F(x_*) - F(x) - F'(x)(x_* - x)\|_2 \\
\le \;& \frac{\mu\gamma}{2}\|x - x_*\|_2^2
\end{aligned}$$

und damit die Behauptung. □

Damit lässt sich das folgende Resultat aus den beiden vorhergehenden Lemmata ableiten:

Lemma 1.8 (Lineare Approximation, zweite Version) *Sei F wie in Voraussetzung 1.5. Es gilt also*

$$\begin{aligned} \|F'(x)\|_2 &\leq \mu \\ \|F'(x) - F'(y)\|_2 &\leq \gamma\|x-y\|_2. \end{aligned}$$

Dann gilt

$$\|F'(x)^T F(x) - F'(x_*)^T F(x_*) - F'(x_*)^T F'(x_*)(x-x_*)\|_2 \leq c\|x-x_*\|_2^2 \tag{1.33}$$

mit $c = \frac{5\mu\gamma}{2}$ für alle $x \in U$.

Beweis :
Nach Lemma 1.6 ist $F'(x)^T F'(x)$ Lipschitz-stetig mit Konstante $2\mu\gamma$.
Es ist

$$\begin{aligned} & F'(x)^T F(x) - F'(x_*)^T F(x_*) - F'(x_*)^T F'(x_*)(x-x_*) \\ = \; - \; & [F'(x_*)^T F(x_*) - F'(x)^T F(x) - F'(x)^T F'(x)(x_*-x)] \\ + \; & [F'(x)^T F'(x) - F'(x_*)^T F'(x_*)](x-x_*). \end{aligned}$$

Wegen der Lipschitz-Stetigkeit von $F'(x)^T F'(x)$ und Lemma 1.7 erhält man

$$\begin{aligned} & \|F'(x)^T F(x) - F'(x_*)^T F(x_*) - F'(x_*)^T F'(x_*)(x-x_*)\|_2 \\ \leq \; & \frac{\mu\gamma}{2}\|x-x_*\|_2^2 + 2\mu\gamma\|x-x_*\|_2^2 \\ = \; & \frac{5\mu\gamma}{2}\|x-x_*\|_2^2 \end{aligned}$$

und damit die Behauptung. □

Die Konvergenzaussage dieses Abschnitts wird in Abhängigkeit von dem Problem angepassten Normen hergeleitet. In gewissem Sinne tritt diese Norm schon in Ungleichung (1.29) auf. Für das Newton-Verfahren findet man angepasste Normen in [16].
Hier werden nun die folgenden Normen eingeführt:

$$\|y\|_i := \|F'(x_i)^T F'(x_i) y\|_2 \tag{1.34}$$

$$\|y\|_* := \|F'(x_*)^T F'(x_*) y\|_2 \tag{1.35}$$

Es ergibt sich sofort aus Voraussetzung 1.5, dass für $\iota \in \{*, 1, 2, \ldots\}$ die Normen $\|\cdot\|_\iota$ gleichmäßig äquivalent sind, denn es folgt unmittelbar für alle Eigenwerte λ von $F'(x)^T F'(x)$: $\mu^{-2} \leq \lambda \leq \mu^2$ für x beliebig und damit gilt

$$\frac{1}{\mu^2}\,\|x\|_2 \;\leq\; \|x\|_\iota \;\leq\; \mu^2\,\|x\|_2 \;\; \forall\, \iota \in \{*, 1, 2, \ldots\}. \tag{1.36}$$

Daher ist die Konvergenz in einer dieser Normen äquivalent zur Konvergenz in der Euklid-Norm. Im Verlauf der Gauß-Newton-Iteration sind die $\|\cdot\|_i$-Normen berechenbar, die $\|\cdot\|_*$-Norm jedoch nicht. Daher wird in der Theorie die $\|\cdot\|_*$-Norm verwendet, während bei praktischen Verfahren die $\|\cdot\|_i$-Norm benutzt werden muss.

Das Hauptergebnis dieses Unterabschnitts ist ein lokales Konvergenzresultat in der $\|\cdot\|_*$-Norm. Dazu werden die Lemmata 1.7 und 1.8 sowie die Genauigkeits-Bedingung (1.29) kombiniert.

Satz 1.9 (Lokale Konvergenz des inexakten Gauß-Newton-Verfahrens) *Gelte Voraussetzung 1.5.*
Dann existieren $0 < t < 1$ und $L > 0$, so dass mit $x_0 \in U_L(x_*) = \{x \in H : \|x - x_*\|_* \leq L\} \cap U$ für die durch die Vorschrift

$$x_{k+1} = x_k + \delta + \epsilon$$

mit δ und ϵ wie in Voraussetzung 1.5 und $\eta(x_k) \leq \eta < 1$ erzeugten Iterierten eines inexakten Gauß-Newton-Verfahrens gilt :

$$\|x_{k+1} - x_*\|_* \leq t \|x_k - x_*\|_* \ \forall\, k = 0, 1, 2, \ldots \tag{1.37}$$

und damit $x_k \in U_L(x_*)$ und $x_k \xrightarrow{\|\cdot\|_*} x_*$ mindestens linear.

Beweis : Gezeigt wird lediglich der Induktionsschritt für den Beweis der Aussage über $k \in \mathbf{N}$. Gelte also $\|x_k - x_*\|_* \leq L$.

$$\begin{aligned} x_{k+1} - x_* &= x_k - x_* + \delta + \epsilon \\ &= x_k - x_* + (F'(x_k)^T F'(x_k))^{-1} F'(x_k)^T F(x_k) + \epsilon \\ &= (F'(x_k)^T F'(x_k))^{-1}[F'(x_k)^T F(x_k) + F'(x_k)^T F'(x_k)(x_k - x_*)] + \epsilon \\ &= (F'(x_k)^T F'(x_k))^{-1}[F'(x_k)^T F(x_k) - F'(x_*)^T F(x_*) - F'(x_k)^T F'(x_k)(x_* - x_k)] \\ &\quad + \epsilon \end{aligned} \tag{1.38}$$

Wegen der Voraussetzung des vollen Spaltenrangs von $F'(x_k)^T F'(x_k)$ und (1.27) ist die Norm von $(F'(x_k)^T F'(x_k))^{-1}$ für alle x_k beschränkt durch μ^2.
Wird die $\|\cdot\|_*$-Norm auf beiden Seiten der Gleichung (1.38) angewendet, erhält man zusammen mit der Abschätzung (1.29) und der Normäquivalenz (1.36)

$$\begin{aligned} \|x_{k+1} - x_*\|_* &\leq \|(F'(x_k)^T F'(x_k))^{-1}\|_* \|F'(x_k)^T F(x_k) - F'(x_*)^T F(x_*) - F'(x_k)^T F'(x_k)(x_* - x_k)\|_* \\ &\quad + \|\epsilon\|_* \\ &\leq \mu^2 \underbrace{\|F'(x_k)^T F(x_k) - F'(x_*)^T F(x_*) - F'(x_k)^T F'(x_k)(x_* - x_k)\|_*}_{\leq \frac{\mu^4 \gamma}{2} \|x_k - x_*\|_*^2 \text{ wg. Lemma 1.7}} \\ &\quad + \eta(x_k) \|F'(x_k)^T F(x_k)\|_2 \\ &\leq \frac{\mu^6 \gamma}{2} \|x_k - x_*\|_*^2 + \eta \|F'(x_k)^T F(x_k)\|_2 \,. \end{aligned} \tag{1.39}$$

Zur Umformung von $\|F'(x_k)^T F(x_k)\|_2$ beachtet man

$$\begin{aligned} F'(x_k)^T F(x_k) &= F'(x_k)^T F(x_k) - F'(x_*)^T F(x_*) - F'(x_*)^T F'(x_*)(x_k - x_*) \\ &\quad + F'(x_*)^T F'(x_*)(x_k - x_*) \end{aligned}$$

und deswegen gilt mit Lemma 1.8

$$\begin{aligned} \|F'(x_k)^T F(x_k)\|_2 &\leq \|F'(x_k)^T F(x_k) - F'(x_*)^T F(x_*) - F'(x_*)^T F'(x_*)(x_k - x_*)\|_2 \\ &\quad + \|F'(x_*)^T F'(x_*)(x_k - x_*)\|_2 \\ &\leq \frac{5\mu\gamma}{2} \|x_k - x_*\|_2^2 + \|x_k - x_*\|_* \\ &\leq \frac{5\mu^3 \gamma}{2} \|x_k - x_*\|_*^2 + \|x_k - x_*\|_* \,. \end{aligned}$$

Setzt man dies in Ungleichung (1.39) ein, erhält man

$$\begin{aligned}
\|x_{k+1} - x_*\|_* &\leq \frac{\mu^6\gamma}{2}\|x_k - x_*\|_*^2 + \eta\,(\frac{5\mu^3\gamma}{2}\|x_k - x_*\|_*^2 + \|x_k - x_*\|_*) \\
&= (\frac{\mu^6\gamma}{2} + \eta\frac{5\mu^3\gamma}{2})\underbrace{\|x_k - x_*\|_*^2}_{\leq L\|x_k - x_*\|_*} + \eta\|x_k - x_*\|_* \\
&\leq \underbrace{(\frac{L\mu^3\gamma(\mu^3 + 5\eta)}{2} + \eta)}_{=:t}\|x_k - x_*\|_* \\
&= t\,\|x_k - x_*\|_*\,. \qquad (1.40)
\end{aligned}$$

Wählt man $0 < L < \frac{2(1-\eta)}{\mu^3\gamma(\mu^3+5\eta)}$, so folgt $0 < t < 1$ und damit die Behauptung. □

Damit sind für die Betrachtung des Gauß-Newton-Verfahrens für kompatible Probleme zwei Voraussetzungen für jeweils zumindest lokale Konvergenz der Verfahren hergeleitet worden.

1.1.6 Inexakte Gauß-Newton-Verfahren für inkompatible Probleme

In vielen Fragestellungen sind die lokalen Minima der Funktion $g(x) = \|F(x)\|_2^2$ nicht gleichzeitig die Nullstellen von g. Auch für diese Probleme ist das Gauß-Newton-Verfahren anwendbar, jedoch kann in manchen Fällen sogar für das exakte Verfahren keine Konvergenz garantiert werden. Dies ist zum Beispiel der Fall, wenn in Satz 1.1 der Fall $\lambda \leq \sigma$ eintritt. Natürlich kann dann auch ein inexaktes Verfahren keine Konvergenz garantieren.
Für inkompatible Probleme muss bei der Betrachtung des Gauß-Newton-Verfahrens die Untersuchung der Bedingung $\nabla g(x) = 0$ an die Stelle von $g(x) = 0$ treten. Das Gauß-Newton-Verfahren ist jedoch kein exaktes Newton-Verfahren zum Finden einer Nullstelle des Gradienten von g. Dadurch kann die Theorie über Newton-Verfahren nur bis zu dem Punkt verwendet werden, bis der Term $F''(x)F(x)$ benutzt werden müsste: bei der Anwendung von Lemma 1.2 für die Funktion $\nabla g(x)$. An die Stelle dieses Lemmas muss dann eine andere Aussage wie Lemma 1.8 treten. Ein solches Lemma ist jedoch für den nicht-kompatiblen Fall neu herzuleiten. Dadurch benötigt man folgende Voraussetzungen:

Voraussetzung 1.10 *Sei $F : \mathbb{R}^n \to \mathbb{R}^m, m > n$, stetig differenzierbar und sei $x_* \in \mathbb{R}^n$ das eindeutige Minimum von $g(x) = \|F(x_*)\|_2^2$. Sei $F'(x)$ Lipschitz-stetig mit Konstante γ und habe für alle $x \in \mathbb{R}^n$ die Matrix $F'(x)$ vollen Spaltenrang. Weiterhin gebe es eine Konstante $\mu > 0$, so dass*

$$\frac{1}{\mu} \leq \|F'(x)\|_2 \leq \mu \qquad (1.41)$$

gleichmässig in x gilt.
Sei zu $x \in \mathbb{R}^n$ ein Vektor $\delta \in \mathbb{R}^n$ durch die exakte Lösung des Gauß-Newton-Systems

$$F'(x)^T F'(x)\,\delta = -F'(x)^T F(x) \qquad (1.42)$$

bestimmt und gelte für einen weiteren Vektor ϵ

$$\|F'(x)^T F'(x)\epsilon\|_2 \leq \eta(x)\,\|F'(x)^T F(x)\|_2 \qquad (1.43)$$

mit $\eta(x) < 1$.

Es ist möglich, die Bedingungen an F' auf eine Umgebung $D \subseteq \mathbb{R}^n$ von x_* einzuschränken. Dementsprechend würden dann die folgenden Aussagen durch lokale Versionen ersetzt werden. Allerdings muss in diesem Fall sichergestellt werden, dass die inexakten Gauss-Newton-Iterierten x_k diese Umgebung D nicht verlassen.
Eine unmittelbare Folgerung aus den Forderungen in Voraussetzung 1.10 ist die folgende Kompatibilitätsbedingung wie sie auch in Satz 1.1 auftritt: Es existiert ein $\sigma > 0$ mit

$$\|(F'(x)^T - F'(x_*)^T)F(x_*)\|_2 \leq \sigma \|x - x_*\|_2 \tag{1.44}$$

($\sigma \leq \gamma\|F(x_*)\|_2$).
Damit ist auch der Fehler abschätzbar, der im Beweis von Lemma 1.7 im Fall $F(x_*) \neq 0$ gemacht würde, und man kann das folgende Lemma analog zu Lemma 1.8 herleiten:

Lemma 1.11 (Lineare Approximation, dritte Version) *Sei F wie in Voraussetzung 1.10. Es gilt also*

$$\begin{aligned} \|F'(x)\|_2 &\leq \mu \\ \|F'(x) - F'(y)\|_2 &\leq \gamma\|x - y\|_2 \\ \|(F'(x)^T - F'(x_*)^T)F(x_*)\|_2 &\leq \sigma \|x - x_*\|_2. \end{aligned}$$

Dann gilt

$$\|F'(x)^T F(x) - F'(x_*)^T F(x_*) - F'(x_*)^T F'(x_*)(x - x_*)\|_2 \leq (c\|x - x_*\|_2 + \sigma)\|x - x_*\|_2 \tag{1.45}$$

mit $c = \frac{5\mu\gamma}{2}$ für alle $x \in U$.

Beweis :
Wie in den Lemmata 1.7 und 1.8 erhält man

$$\begin{aligned} & F'(x)^T F(x) - F'(x_*)^T F(x_*) - F'(x_*)^T F'(x_*)(x - x_*) \\ = \; & -[F'(x_*)^T F(x_*) - F'(x)^T F(x) - F'(x)^T F'(x)(x_* - x)] \\ & +(F'(x)^T F'(x) - F'(x_*)^T F'(x_*))(x - x_*) \\ = \; & -[F'(x)^T F(x_*) - F'(x)^T F(x) - F'(x)^T F'(x)(x_* - x)] \\ & +(F'(x)^T F(x_*) - F'(x_*)^T F(x_*)) + (F'(x)^T F'(x) - F'(x_*)^T F'(x_*))(x - x_*) \\ = \; & -F'(x)^T[F(x_*) - F(x) - F'(x)(x_* - x)] + (F'(x)^T - F'(x_*)^T)F(x_*) \\ & +(F'(x)^T F'(x) - F'(x_*)^T F'(x_*))(x - x_*). \end{aligned}$$

Wendet man nun die Norm auf beiden Seiten an und benutzt die Dreiecks-Ungleichung und die Lemmata 1.2 und 1.6 erhält man die Behauptung. □
Der Beweis für die Konvergenz von inexakten Verfahren basiert hier auf der Übertragung der Beweisidee für das Newton-Verfahren aus [19] auf den Gauß-Newton-Fall und das Finden einer Nullstelle von $F'(x)^T F(x)$. Voraussetzung ist dafür, dass außer der Genauigkeits-Bedingung (1.43) auch tatsächlich eine Reduktion der Norm von $F'(x)^T F(x)$ stattfindet. In diesem Fall hat man das folgende Konvergenzergebnis:

Satz 1.12 *Sei Voraussetzung 1.10 erfüllt und sei $x_0 \in \mathbb{R}^n$ und die Folge $\{x_k\}_{k=0}^{\infty}$ konstruiert nach der Vorschrift $x_{k+1} = x_k + \delta + \epsilon$ mit δ und ϵ passend zu x_k aus (1.42) und (1.43) mit $\eta(x_k) = \eta_k \leq \eta < 1$.*
Gelte weiterhin $F'(x_k)^T F(x_k) \to 0$ und sei für alle k

$$\|F'(x_{k+1})^T F(x_{k+1})\|_2 \leq \|F'(x_k)^T F(x_k)\|_2.$$

Ist x^* ein Häufungspunkt der Folge $\{x_k\}_{k=0}^{\infty}$ derart, dass $F'(x^*)^T F'(x^*)$ invertierbar ist und gilt $\sigma < \frac{1}{2\|(F'(x^*)^T F'(x^*))^{-1}\|_2}$, dann gilt $F'(x^*)^T F(x^*) = 0$ und $x_k \to x^*$.

Beweis:
Es ist klar, dass $F'(x^*)^T F(x^*) = 0$ gilt und damit $x^* = x_*$. Setze $s_k = x_{k+1} - x_k$ und $K = \|(F'(x^*)^T F'(x^*))^{-1}\|_2$.
Sei $\rho > 0$ so klein, dass $(F'(y)^T F'(y))^{-1}$ existiert und

$$\|(F'(y)^T F'(y))^{-1}\|_2 \leq 2K \tag{1.46}$$

$$\|F'(y)^T F(y) - F'(x_*)^T F(x_*) - F'(x_*)^T F'(x_*)(y - x_*)\|_2 \leq \frac{1}{2K}\|y - x_*\|_2 \tag{1.47}$$

gilt für alle y mit $\|y - x_*\|_2 \leq \rho$. Ein solches ρ existiert wegen der Aussage von Lemma 1.11.
Sei y mit $\|y - x_*\|_2 \leq \rho$ gegeben. Dann ist

$$\begin{aligned}
& \|F'(y)^T F(y)\|_2 \\
\geq\ & \|F'(x_*)^T F'(x_*)(y - x_*)\|_2 - \|F'(y)^T F(y) - F'(x_*)^T F(x_*) - F'(x_*)^T F'(x_*)(y - x_*)\|_2 \\
\overset{(1.47)}{\geq}\ & \frac{1}{\|(F'(x^*)^T F'(x^*))^{-1}\|}\|y - x_*\|_2 - \frac{1}{2K}\|y - x_*\|_2 \\
\geq\ & \frac{1}{2K}\|y - x_*\|_2.
\end{aligned}$$

Damit ist dann für $\|y - x_*\|_2 \leq \rho$

$$\|y - x_*\|_2 \leq 2K\|F'(y)^T F(y)\|_2. \tag{1.48}$$

Sei $0 < \alpha < \rho/4$ beliebig. Da x_* ein Häufungspunkt von $\{x_k\}_{k=0}^{\infty}$ ist und $F'(x_*)^T F(x_*) = 0$ gilt, gibt es ein genügend großes k, so dass

$$x_k \in S_\alpha = \{y : \|y - x_*\|_2 \leq \frac{\rho}{2} \text{ und } \|F'(y)^T F(y)\|_2 < \frac{\alpha}{K(1+\eta)}\}.$$

Dann ist mit (1.42),(1.43) und (1.46)

$$\begin{aligned}
\|s_k\|_2 &\leq \|(F'(x_k)^T F'(x_k))^{-1}[-F'(x_k)^T F(x_k) + (F'(x_k)^T F(x_k) + F'(x_k)^T F'(x_k)s_k)]\|_2 \\
&\leq \|(F'(x_k)^T F'(x_k))^{-1}\|_2(\|F'(x_k)^T F(x_k)\|_2 + \|F'(x_k)^T F(x_k) + F'(x_k)^T F'(x_k)s_k\|_2) \\
&\leq 2K(1+\eta)\|F'(x_k)^T F(x_k)\|_2 \\
&< 2\alpha \\
&< \frac{\rho}{2}
\end{aligned}$$

und so gilt

$$\|x_{k+1} - x_*\|_2 \leq \|x_k - x_*\|_2 + \|s_k\|_2 \leq \rho\,.$$

Es ist

$$\|F'(x_{k+1})^T F(x_{k+1})\|_2 \leq \|F'(x_k)^T F(x_k)\|_2 < \frac{\alpha}{K(1+\eta)}$$

und zusammen mit (1.48) gilt dann

$$\|x_{k+1} - x_*\|_2 \leq 2K\|F'(x_{k+1})^T F(x_{k+1})\|_2 < 2K\frac{\alpha}{K(1+\eta)} < \frac{\rho}{2}.$$

Daraus folgt also $x_{k+1} \in S_\alpha$. Dann gilt jedoch für alle genügend großen k $x_k \in S_\alpha$ und damit $\|x_k - x_*\|_2 \leq \rho$. Wegen $\|F'(x_k)^T F(x_k)\|_2 \to 0$ und (1.48) gilt dann $x_k \to x_*$. □
Mit Hilfe dieses Satzes bleibt für die Konvergenz inexakter Gauß-Newton-Verfahren für inkompatible Probleme neben der Genauigkeitsbedingung (1.43) eine Abstiegsbedingung der Art (1.24) für das Funktional $\|F'(x_k)^T F(x_k)\|_2$ nachzuweisen: Gibt es ein t mit $0 < t < 1$ und

$$\|F'(x_{k+1})^T F(x_{k+1})\|_2 \leq [1 - t(1 - \eta_k)]\|F'(x_k)^T F(x_k)\|_2, \tag{1.49}$$

so ist diese Abstiegsbedingung, die auch für Satz 1.12 benötigt wird, erfüllt.
Im Beweis des Satzes 1.12 lässt sich die Voraussetzung $\sigma < (2K)^{-1}$ abschwächen zu $\sigma < (cK)^{-1}$ mit einem $1 \leq c \leq 2$. Dann entspricht diese Bedingung auch der Voraussetzung $\sigma < \lambda$ im Satz 1.1 zur Konvergenz exakter Gauß-Newton-Verfahren angewandt auf inkompatible Probleme. Ist dann also (1.49) bei gleichzeitiger Divergenz der Reihe über $(1 - \eta_k)$ erfüllt, so lässt sich Satz 1.12 anwenden und es gilt $x_k \to x_*$.
Damit stehen auch für inkompatible Probleme die bestmöglichen Genauigkeitsschranken zur Verfügung.

1.2 Diskretisierung parabolischer Differentialgleichungen

Betrachtet wird die folgende partielle Differentialgleichung

$$\begin{array}{rcll} \frac{\partial u(x,t)}{\partial t} + Au(x,t) & = 0 & \text{in} & \Omega, \text{ für } 0 < t \leq T, \\ u(x,t) & = 0 & \text{auf} & \partial\Omega, \text{ für } 0 < t \leq T, \\ u(x,0) & = v(x) & \text{in} & \Omega, \end{array} \tag{1.50}$$

wobei $\Omega \in \mathbb{R}^n$ ein beschränktes Gebiet mit glattem Rand $\partial\Omega$ und A ein linearer Differentialoperator (z.B. $A = -\Delta$) ist. Gesucht ist eine Funktion $u \in L^2([0,T]; H)$, die die Gleichungen (1.50) erfüllt (H ein passender Hilbertraum; zur Existenz von Lösungen siehe [21]).
Diese in Ort (Variable x) und Zeit (Variable t) kontinuierliche Anfangsrandwertaufgabe soll nun mittels geeigneter Diskretisierungstechniken in endlichdimensionale Probleme überführt werden. In diesem Abschnitt wird kurz auf die Diskretisierung bzgl. der Zeit ($[0,T] \to t_0, t_1, t_2, \ldots$) eingegangen. Die Diskretisierung bzgl. des Raumes ($H \to H_h$) wird dann im folgenden Abschnitt 1.3 behandelt.

Die Lösung von (1.50) ergibt sich in Operatorschreibweise zu

$$u(x,t) \;=\; e^{-tA}\, v(x).$$

Ein Einschrittverfahren hat also die Form

$$u_{k+1}(x) \;=\; \Psi^{t_k+\tau,t_k} u_k(x), \;\; u_0(x) = v(x),$$

mit $t_k = k\tau$. Dabei wird $u(x,t_k)$ durch $u_k(x)$ approximiert. Durch Vergleichen der beiden letzten Gleichungen erhält man sofort, dass dieses Verfahren Ordnung $p \in \mathbb{N}$ besitzt, wenn die diskrete Evolution $\Psi^{t_k+\tau,t_k}$ eine Näherung an $e^{-\tau A}$ der Art darstellt, dass

$$\Psi^{t_k+\tau,t_k} \;=\; e^{-\tau A} \;+\; O(\tau^{p+1})$$

gilt.
Eines der einfachsten Verfahren ist das implizite Euler-Verfahren, bei dem die Zeitableitung $\partial u/\partial t$ in (1.50) durch den Differenzenquotienten $\frac{u_{k+1}-u_k}{\tau}$ ersetzt wird. Grob gesprochen wird bei diesem Verfahren also stückweise konstant in der Zeit diskretisiert.
Dies führt auf folgende, im $k+1$-ten Zeitschritt zu lösende Gleichung:

$$u_{k+1}(x) + \tau A u_{k+1}(x) \;=\; u_k(x) \;\text{ in } \Omega. \tag{1.51}$$

Betrachtet man die zugehörige Evolution Ψ, erhält man

$$\begin{array}{rcl} \Psi^{t_k+\tau,t_k} \;=\; (I+\tau A)^{-1} & = & I - \tau A(I - \tau A(I+\tau A)^{-1}) \\ & = & I - \tau A + O(\tau^2) \end{array}$$

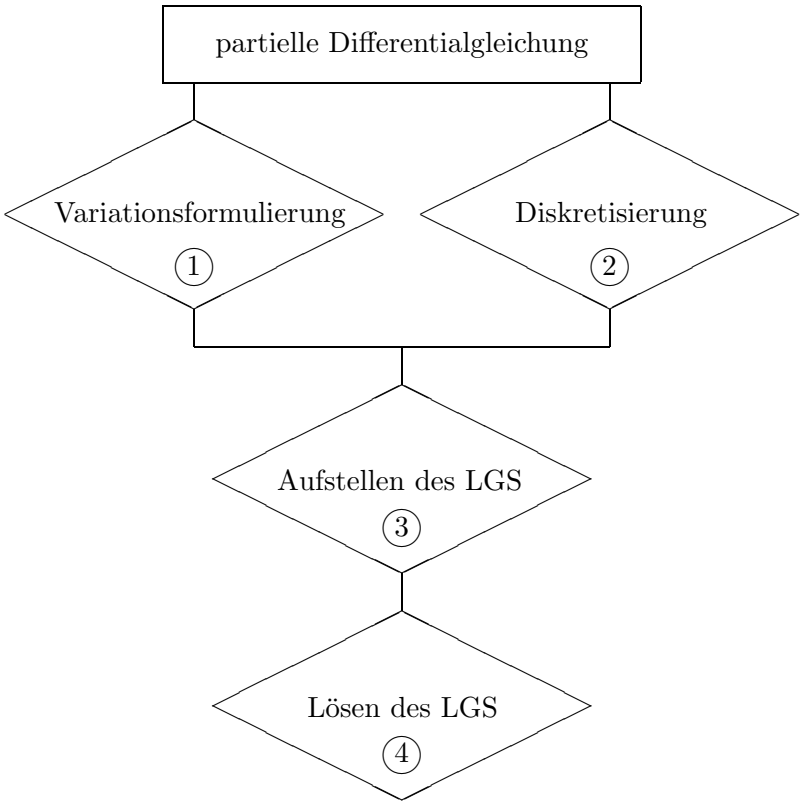
und wegen

$$e^{-\tau A} = I - \tau A + \frac{(\tau A)^2}{2!} + \dots$$

ist dies ein Verfahren erster Ordnung.
Weitere Diskretisierungsschemata (z.B. Crank-Nicholson) sowie Ergebnisse zu Fehlerabschätzungen und Stabilität findet man z.B. in [43].

1.3 Die Methode der Finiten Elemente (FEM)

Der Raum, in dem die Lösungen partieller Differentialgleichungen zu suchen sind, ist im Allgemeinen unendlichdimensional. Wie in Abschnitt 1.2 bezüglich der Zeit diskretisiert wurde, muss daher auch eine Diskretisierung bzgl. des Ortes erfolgen, um das Problem mit dem Rechner behandelbar zu machen. Dafür ist die FEM ist eine der am längsten untersuchten und flexibelsten Methoden. Schematisch dargestellt sind durch die FEM vier Schritte zur approximativen Lösung der Differentialgleichung notwendig, die in den folgenden Unterabschnitten kurz erklärt werden:



Im folgenden Abschnitt wird auf die Aufstellung der Variationsformulierung eingegangen, in Abschnitt 1.3.2 auf die Diskretisierung von Ω und der dadurch induzierten Diskretisierung der Funktionenräume und im letzten Teil wird kurz die Aufstellung der linearen Gleichungssysteme dargestellt. In Abschnitt 1.4 werden schließlich Methoden zur Lösung der linearen Gleichungssysteme angegeben.

1.3.1 Variationsformulierung

Zur besseren Übersicht bezeichne $(\cdot)_{x_i}$ die partielle Ableitung von $(\cdot)$ nach der Variablen x_i. Gegeben sei folgendes Randwertproblem:

$$\begin{aligned} -\sum_{i,j=1}^{n} \left(a_{ij}(x)p_{x_i}(x)\right)_{x_j} + \sum_{i=1}^{n} b_i(x)p_{x_i}(x) + c(x)p(x) &= f \quad \text{in } \Omega \\ p(x) &= g \quad \text{auf } \Gamma_D \subseteq \partial\Omega \\ \vec{n} \cdot \left(\sum_{i,j=1}^{n} \left(a_{ij}(x)p_{x_i}(x)\right)_{x_j}\right) &= h \quad \text{auf } \Gamma_N \subseteq \partial\Omega, \end{aligned} \tag{1.52}$$

mit $\vec{n}$ als äußerer (Einheits-)Normalen an den Rand $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \neq \emptyset$. Werden die Abkürzungen

$$\begin{aligned} A = A(x) &= [a_{ij}(x)]_{i,j=1}^{n} \\ B = B(x) &= (b_1(x), b_2(x), \ldots, b_n(x)) \\ c &= c(x) \end{aligned}$$

eingeführt, so lässt sich dies schreiben als

$$\begin{aligned} -\text{div}\,(A\nabla p) + B\nabla p + cp &= f \quad \text{in } \Omega \\ p &= g \quad \text{auf } \Gamma_D \subseteq \partial\Omega \\ \vec{n} \cdot A\nabla p &= h \quad \text{auf } \Gamma_N \subseteq \partial\Omega. \end{aligned} \tag{1.53}$$

Im Weiteren sei A symmetrisch und uniform positiv definit in Ω. Für dieses Randwertproblem fordert man nun die eindeutige Lösbarkeit mit $p \in H^1(\Omega)$. Mit dieser Forderung schwächt man auch die in (1.53) auftretenden Ableitungen zu schwachen Ableitungen und auch den klassischen Lösungsbegriff zur schwachen Lösung eines Variationsproblems ab.
Eine Variationsformulierung zur Lösung der Aufgabe (1.52) kann durch mehrere Verfahren hergeleitet werden: Integration der Gleichungen (vgl. [21, Kapitel 6]), Euler-Lagrange-Gleichungen (vgl. [21, Kapitel 8]) und den folgenden FOSLS-Ansatz (**F**irst-**O**rder-**S**ystem-**L**east-**S**quares): Man definiert dazu

$$u(x) = \begin{pmatrix} u_1(x) \\ u_2(x) \\ \vdots \\ u_n(x) \end{pmatrix}, \quad u_i(x) = \sum_{j=1}^{n} a_{ij}(x)p_{x_i}, \quad i = 1, \ldots, n$$

und führt damit die Gleichung zweiter Ordnung (1.53) in ein Gleichungssystem erster Ordnung für die unbekannten Funktionen u und p über:

$$\begin{aligned} -\text{div}\; u + B\nabla p + cp &= f \quad \text{in } \Omega \\ u - A\nabla p &= 0 \quad \text{in } \Omega \\ p &= 0 \quad \text{auf } \Gamma_D \subseteq \partial\Omega \\ \vec{n} \cdot u &= 0 \quad \text{auf } \Gamma_N \subseteq \partial\Omega. \end{aligned} \tag{1.54}$$

Durch diese Formulierung sind nun beide Randbedingungen vom Dirichlet-Typ und werden in die Lösungsräume eingebaut. Man sucht also Lösungen $(p, u) \in H^1(\Omega) \times H(\text{div}, \Omega)$ (im schwachen Sinn) die sich als $(p, u) = (p_D + \hat{p}, u_N + \hat{u})$ schreiben lassen. Während (p_D, u_N) einen beliebige Funktion in $H^1(\Omega) \times H(\text{div}, \Omega)$ ist, die die Randbedingung erfüllt, bleibt noch die Funktion $(\hat{p}, \hat{u}) \in V \times W$ mit

$$\begin{aligned} V &= \{q \in H^1(\Omega) : q = 0 \text{ auf } \Gamma_D\} \\ W &= \{v \in H(\text{div}, \Omega) : \vec{n} \cdot v = 0 \text{ auf } \Gamma_N\} \end{aligned} \tag{1.55}$$

zu suchen. Definiert man das Funktional $\mathcal{F} : V \times W \to \mathbb{R}$ mittels

$$\begin{aligned}\mathcal{F}(\hat{p},\hat{u}) =& \| - \text{div } (u_N + \hat{u}) + B\nabla(p_D + \hat{p}) + c(p_D + \hat{p}) - f \|^2_{0,\Omega} \\ &+ \| (u_N + \hat{u}) - A\nabla(p_D + \hat{p}) \|^2_{0,\Omega},\end{aligned} \tag{1.56}$$

ist offensichtlich, dass die eindeutige Lösung p von (1.53) zusammen mit $u = A\nabla p$ das Funktional $\mathcal{F}$ minimiert mit minimalem Wert $\mathcal{F}(p - p_D, u - u_N) = 0$. Die Differentialgleichung (1.52) wird somit auch gelöst durch die Bestimmung von $(\hat{p}, \hat{u}) \in V \times W$ mit

$$\mathcal{F}(\hat{p},\hat{u}) \stackrel{!}{=} \min_{(\hat{q},\hat{v})\in V\times W} \mathcal{F}(\hat{q},\hat{v}). \tag{1.57}$$

Zur Vereinfachung der Schreibweise werden für den Rest des Abschnitts nun homogene Randbedingungen angenommen (d.h. $u_N = 0, p_D = 0, \hat{u} = u, \hat{p} = p$).
Die notwendige Bedingung für ein Minimum von (1.57) ist das Verschwinden der ersten Ableitung (Variationsrechnung). Dies führt auf das Variationsproblem:

Suche $(p, u) \in V \times W$ mit

$$\left(\begin{pmatrix} -\text{div } u + B\nabla p + cp \\ u - A\nabla p \end{pmatrix}, \begin{pmatrix} -\text{div } v + B\nabla q + cq \\ v - A\nabla q \end{pmatrix}\right)_{0,\Omega} = \left(\begin{pmatrix} f \\ 0 \end{pmatrix}, \begin{pmatrix} -\text{div } v + B\nabla q + cq \\ v - A\nabla q \end{pmatrix}\right)_{0,\Omega}$$

für alle $(v, q) \in V \times W$.

Mit den Bezeichnungen $\text{div } u = \nabla \cdot u$ und $X = B\nabla + cI$ wird dies zu

Suche $(p, u) \in V \times W$ mit

$$\left(\begin{pmatrix} -\nabla\cdot & X \\ I & -A\nabla \end{pmatrix}\begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} -\nabla\cdot & X \\ I & -A\nabla \end{pmatrix}\begin{pmatrix} v \\ q \end{pmatrix}\right)_{0,\Omega} = \left(\begin{pmatrix} f \\ 0 \end{pmatrix}, \begin{pmatrix} -\nabla\cdot & X \\ I & -A\nabla \end{pmatrix}\begin{pmatrix} v \\ q \end{pmatrix}\right)_{0,\Omega} \tag{1.58}$$

für alle $(q, v) \in V \times W$.

Die Elliptizität des dadurch definierten Operators wird in [13] nachgewiesen (vgl. (1.60)). Daher ist (1.58) eindeutig lösbar und die Lösung (p, u) ist das eindeutige Minimum von $\mathcal{F}$.

Zur Lösung des Systems (1.54) kann auch die sog. gemischte Methode verwendet werden. Dabei führt die Aufgabe auf ein Sattelpunktproblem (vgl. [12]), dessen eindeutige Lösbarkeit in den (unendlichdimensionalen) Räumen $H^1(\Omega) \times H(\text{div}, \Omega)$ nur bei Erfüllen der inf-sup-Bedingung (auch LBB-Bedingung genannt) gezeigt werden kann. Diese Einschränkung wirkt sich auch entscheidend auf die Auswahl der Diskretisierungen der Lösungsräume (→ Abschnitt 1.3.2) aus, für die ein diskretes Äquivalent der inf-sup-Bedingung gefordert werden muss, um zu stabilen Lösungen zu gelangen (siehe auch [10]). Insbesondere bei dem Wunsch nach nicht-konformer Wahl des Ansatzraumes für die neu eingeführte Variable u führt dies zu wesentlichen Einschränkungen.
Diese Einschränkungen sind bei dem obigen FOSLS-Ansatz nicht zu beachten, da hierbei kein Sattelpunktproblem sondern ein quadratisches Minimierungsproblem zu lösen ist. Dadurch bleiben auch Approximationen in Räumen, in denen die inf-sup-Bedingung nicht erfüllt ist, wertvoll, denn die Erfüllung der Nebenbedingung des Sattelpunktproblems wird durch Minimierung des Funktionals kontrolliert.

Aus dem Hauptergebnis von [13] ergibt sich ein weiterer Vorteil des FOSLS-Ansatzes. Setze dazu die Bilinearform

$$\mathcal{J}^*\mathcal{J}(p,u;q,v) = \left(\begin{pmatrix} -\nabla\cdot & X \\ I & -A\nabla \end{pmatrix}\begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} -\nabla\cdot & X \\ I & -A\nabla \end{pmatrix}\begin{pmatrix} v \\ q \end{pmatrix}\right)_{0,\Omega}$$

entsprechend Gleichung (1.58). In [13] wird gezeigt, dass mit $H = H^1(\Omega) \times H(\mathrm{div}, \Omega)$ und $\|(p,u)\|_H^2 = \|p\|_{1,\Omega}^2 + \|u\|_{\mathrm{div},\Omega}^2$ Konstanten $\underline{\alpha}, \overline{\alpha}$ existieren mit

$$\begin{aligned} \mathcal{J}^*\mathcal{J}(p,u;q,v) &\leq \overline{\alpha}\|(p,u)\|_H \, \|(q,v)\|_H && (1.59)\\ \underline{\alpha}\|(q,v)\|_H^2 &\leq \mathcal{J}^*\mathcal{J}(q,v;q,v) && (1.60) \end{aligned}$$

für alle $(p,u),(q,v) \in V \times W$.
Sei $(p,u) \in V \times W$ die eindeutige Lösung des Systems (1.54). Fügt man diese Gleichungen in das Funktional $\mathcal{F}(q,v)$ nach folgender Weise ein, so verändert sich der Wert nicht:

$$\begin{aligned} \mathcal{F}(q,v) &= \| - \mathrm{div}\ v + B\nabla q + cq - f \|_{0,\Omega}^2 + \| v - A\nabla q \|_{0,\Omega}^2 \\ &= \| - \mathrm{div}\ (v-u) + B\nabla(q-p) + c(q-p) - f + f \|_{0,\Omega}^2 \\ &\quad + \| (v-u) - A\nabla(q-p) \|_{0,\Omega}^2 \qquad (1.61)\\ &= \mathcal{J}^*\mathcal{J}(\underbrace{p-q}_{\in V}, \underbrace{u-v}_{\in W}; p-q, u-v) \overset{(1.59),(1.60)}{\cong} \|(p-q,u-v)\|_H^2 \end{aligned}$$

(Dies gilt offensichtlich auch im Fall $u_N \neq 0$, bzw. $p_D \neq 0$).
Betrachtet man nun (q,v) als eine Näherung an die Lösung, so ist nach (1.61) der Wert des Funktionals $\mathcal{F}(q,v)$ äquivalent zur H-Norm des Fehlers. Da die Approximationen in H durchgeführt werden, ist damit die Größe des Funktionals proportional zur Größe des Fehlers in H. Diese Eigenschaft der FOSLS-Vorgehensweise wird auch im nächsten Abschnitt verwendet und weiter interpretiert werden.

1.3.2 Diskretisierung

Zur Vereinfachung werden auch in diesem Abschnitt homogene Randbedingungen angenommen. Zur numerischen Approximation der Lösung $(p,u) \in V \times W$ des Variationsproblems (1.58) muss zu endlichdimensionalen Teilräumen von V und W übergegangen werden.
Seien $V_h \subseteq V$ und $W_h \subseteq W$ zwei endlichdimensionale Räume. Der Ritz-Galerkin-Ansatz zur Approximation von (p,u) besteht darin, die Variationsformulierung in $V_h \times W_h$ aufzustellen:

Suche $(p_h, u_h) \in V_h \times W_h$ mit

$$\left(\begin{pmatrix} -\nabla\cdot & X \\ I & -A\nabla \end{pmatrix} \begin{pmatrix} u_h \\ p_h \end{pmatrix}, \begin{pmatrix} -\nabla\cdot & X \\ I & -A\nabla \end{pmatrix} \begin{pmatrix} v_h \\ q_h \end{pmatrix} \right)_{0,\Omega} = \left(\begin{pmatrix} f \\ 0 \end{pmatrix}, \begin{pmatrix} -\nabla\cdot & X \\ I & -A\nabla \end{pmatrix} \begin{pmatrix} v_h \\ q_h \end{pmatrix} \right)_{0,\Omega}$$

für alle $(q_h, v_h) \in V_h \times W_h$. (1.62)

Mit der Abkürzung

$$l_f(q,v) = \left(\begin{pmatrix} f \\ 0 \end{pmatrix}, \begin{pmatrix} -\nabla\cdot & X \\ I & -A\nabla \end{pmatrix} \begin{pmatrix} v \\ q \end{pmatrix} \right)_{0,\Omega}$$

wird eine Linearform auf $V \times W$ definiert und zusammen mit der Definition von $\mathcal{J}^*\mathcal{J}$ schreibt sich das Variationsproblem kurz

Suche $(p_h, u_h) \in V_h \times W_h$ mit

$$\mathcal{J}^*\mathcal{J}(p_h, u_h; q_h, v_h) = l_f(q_h, v_h) \qquad (1.63)$$

für alle $(q_h, v_h) \in V_h \times W_h$.

Offensichtlich hängt die Qualität dieser Approximation von den verwendeten Unterräumen V_h und W_h ab. Da $\mathcal{J}^*\mathcal{J}$ stetig, symmetrisch und elliptisch ist, kann das Céa-Lemma angewendet werden (vgl. [8]):

Lemma 1.13 (Céa-Lemma) *Seien $(p,u) \in V \times W$ bzw. $(p_h,u_h) \in V_h \times W_h$ die Lösungen der Variationsprobleme*

$$\begin{aligned} \mathcal{J}^*\mathcal{J}(p,u;q,v) &= l_f(q,v) \;\; \forall\,(q,v) \in V \times W \\ \mathcal{J}^*\mathcal{J}(p_h,u_h;q_h,v_h) &= l_f(q_h,v_h) \;\; \forall\,(q_h,v_h) \in V_h \times W_h \end{aligned}$$

und erfülle $\mathcal{J}^*\mathcal{J}$ (1.59) und (1.60). Dann ist

$$\|(p,u)-(p_h,u_h)\|_H \;\leq\; \underbrace{\frac{\overline{\alpha}}{\underline{\alpha}}}_{=:C} \min_{(q_h,v_h)\in V_h\times W_h} \|(p,u)-(q_h,v_h)\|_H \,. \tag{1.64}$$

Ist $\mathcal{J}^*\mathcal{J}$ auch symmetrisch, so gilt die Behauptung sogar mit $C = \sqrt{\frac{\overline{\alpha}}{\underline{\alpha}}}$.

Je besser die endlichdimensionalen Räume die Lösung überhaupt approximieren können, desto kleiner wird also auch der Fehler bei Verwendung der berechneten Variationslösung (p_h,u_h) sein. Für die Wahl der Teilräume der Galerkin-Approximation fordert man also vor allem, dass sich die Lösung in ihnen gut approximieren lässt. Darauf wird im Zusammenhang mit der Verwendung des Funktionals $\mathcal{F}$ als Fehlerschätzer noch näher eingegangen.
Als weitere Überlegung bei der Wahl der Approximationsräume kommt bei der Formulierung der partiellen Differentialgleichung als Sattelpunktproblem die Erfüllung der diskreten inf-sup-Bedingung (siehe vorhergehender Abschnitt) hinzu.

Ein entscheidender Bestandteil der FEM ist die Konstruktion der verwendeten endlichdimensionalen Räume. Wegen der Approximationseigenschaften von Polynomen wird bei den meisten Methoden der Finiten Elemente eine Basis des endlichdimensionalen Raumes aus Polynomen eines bestimmten vorgegebenen Grades konstruiert, die einen kompakten Träger besitzen. Der Träger der Basisfunktionen hat dabei zumeist Dreiecks- oder Viereckform (im zweidimensionalen Fall) bzw. Tetraeder- oder Quaderform (im dreidimensionalen Fall). Es wird also zuerst eine Entscheidung über die Zerlegung $\mathcal{T} = \{T_i : i = 1,\ldots,n_T\}$ des Gebiets Ω, in dem die PDGL gelöst werden soll, getroffen. Dadurch wird gleichzeitig die gewünschte Feinheit der Zerlegung festgelegt, denn die Feinheit h ist als Maximum über alle Durchmesser von Elementen T_i der Zerlegung $\mathcal{T}$ definiert. Danach wird der maximale Polynomgrad k festgelegt und eine Basis für den gewünschten Teilraum von $\mathbb{P}_k(T_i)$ aufgestellt. ($\mathbb{P}_k(T)$ sei die übliche Bezeichnung für den Raum aller Polynome bis einschließlich Grad k über dem Gebiet T.) Setzt man alle Basisfunktionen über T_i durch Null auf Ω fort und vereinigt die Teilbasen zu allen Elementen T_i, so erhält man eine Basis von $V_h \subseteq V$ bzw. $W_h \subseteq W$.
Wie man sieht, ist die FEM eine äußerst flexible Methode, die sehr viele Gestaltungsspielräume offenlässt, um gegebene Charakteristiken der PDGL möglichst gut zu berücksichtigen.
Im Einzelnen bestehen unter anderem folgende Auswahlmöglichkeiten:

- Art der Zerlegung und Feinheit h
  - Dreieckselemente, Viereckselemente, Bedingung für Innenwinkel der Elemente (z.B. keine zu spitzen Winkel),...
  - Größenordnung des Durchmessers h_i von T_i im ganzen Gebiet Ω etwa gleich oder lokal kleinere h_i zugelassen (Adaptivität)

- maximaler Polynomgrad auf jedem Element der Zerlegung
- Entscheidung über weitere (Stetigkeits-)Bedingungen zwischen den Elementen der Zerlegung:
  - konforme Elemente : $W_h \subseteq W$ (bzw. $V_h \subseteq V$)
  - nichtkonforme Elemente : $W_h \subsetneq W$ (bzw. $V_h \subsetneq V$)

Im Folgenden werden nur konforme Elemente verwendet, weswegen insbesondere auch eine Einschränkung auf Polygongebiete Ω erfolgt. (Weitere natürliche Einschränkungen um sog. zulässige Zerlegungen zu erhalten, findet man in [8].)
Für Diskretisierungen des Raums $V \subseteq H^1(\Omega)$ ergeben sich dann die Stetigkeitsbedingungen aus folgendem Satz (vgl. [8]):

Satz 1.14 *Sei Ω beschränkt und $k \geq 1$. Ist $v : \bar{\Omega} \to \mathbf{R}$ stückweise beliebig oft differenzierbar so gilt*

$$v \in H^k(\Omega) \Leftrightarrow v \in C^{k-1}(\bar{\Omega}).$$

Konforme Funktionen in $V_h \subseteq V \subseteq H^1(\Omega)$ müssen also stetig auf $\bar{\Omega}$ sein, insbesondere muss damit die Funktion beim Übergang zwischen zwei Elementen der Zerlegung stetig sein (z.B. auf Kanten). Ein Beispiel sind die stückweise linearen Funktionen bei Triangulierungen von Gebieten $\Omega \subseteq \mathbf{R}^2$, deren Freiheitsgrade (Basisfunktionen) mit den Knoten (Eckpunkten) der Zerlegung assoziiert sind (sog. nodale Basis). Natürlich müssen auch die homogenen Randbedingungen auf Γ_D für Funktionen in V beachtet werden.
Für die Diskretisierung $W_h \subseteq W \subseteq H(\mathrm{div}, \Omega)$ gilt eine andere Stetigkeitsbedingung, die durch Anwendung des Gaußschen Integralsatzes hergeleitet wird:
Für stückweise polynomiale Ansatzfunktionen ψ muss gefordert werden, dass die Normalkomponente auf Kurven in Ω stetig ist. Es muss also insbesondere die Normalkomponente $n \cdot \psi$ bei dem Übergang zwischen zwei Elementen T_i, T_j stetig sein (und natürlich müssen auch die homogenen Randbedingungen auf Γ_N erfüllt sein). Die Freiheitsgrade sind bei Diskretisierungen von W also mit den Seiten der Triangulierung (im Zweidimensionalen: Kanten) assoziiert. Diese Bedingung ist zum Beispiel für die Räume nach Raviart und Thomas erfüllt (vgl. [12]).

Grundsätzlich kann die Genauigkeit der Approximation sowohl durch Erhöhen des Polynomgrads, als auch durch Wählen feinerer Triangulierungen gesteigert werden. Beides führt dann zu einer höheren Anzahl von Freiheitsgraden, also zu einem größeren linearen Gleichungssystem (→ Abschnitt 1.3.3). Gerade die uniforme, d.h. gleichmässige Verfeinerung der bestehenden Zerlegung erhöht jedoch die Anzahl der Freiheitsgrade häufig unnötig stark.
Eine wirkungsvolle Möglichkeit, die Erhöhung der Anzahl der Freiheitsgrade auf das nötigste zu beschränken, ist die sogenannte adaptive Verfeinerung. Mit Hilfe eines Fehlerschätzers η^2 werden dabei diejenigen Bereiche im Gebiet Ω bestimmt, in denen der Fehler besonders groß ist und dann gezielt dort verfeinert, um die Lösung besser annähern zu können:
Gilt

$$\begin{aligned} \eta^2 &= \sum_{i=1}^{n_T} \eta_{T_i}^2 \\ \underline{a}\,\eta^2 &\leq \|(p,u) - (p_h,u_h)\|_H^2 \leq \overline{a}\,\eta^2 \end{aligned}$$

dann setzt sich der Fehlerschätzer aus lokalen Anteilen η_T^2 zusammen. Eine Verfeinerungsstrategie könnte nun darin bestehen, diejenigen Elemente i zu verfeinern, für die

$$\eta_{T_i}^2 \geq \frac{\eta^2}{n_T}$$

gilt.
Die Konstruktion von Fehlerschätzern für beliebige PDGLen ist ein eigenes Feld innerhalb der FEM (siehe auch [48]) und zieht häufig einen erheblichen Implementierungsaufwand nach sich. Dies ist bei dem im vorhergehenden Abschnitt vorgestellten FOSLS-Ansatz nicht der Fall, denn ein Vorteil dieses Ansatzes besteht in der direkten Verwendung des Funktionals $\mathcal{F}$ als Fehlerschätzer.
Einerseits gilt wegen (1.59) und (1.60)

$$\underline{\alpha}\,\|(p,u)-(p_h,u_h)\|_H^2 \;\leq\; \mathcal{F}(p_h,u_h) \;\leq\; \overline{\alpha}\,\|(p,u)-(p_h,u_h)\|_H^2$$

und andererseits gilt

$$\begin{aligned}
\mathcal{F}(p_h u p_h) &= \| -\operatorname{div} u_h + B\nabla p_h + c p_h - f \|_{0,\Omega}^2 + \| u_h - A\nabla p_h \|_{0,\Omega}^2 \\
&= \sum_{i=1}^{n_T} \| -\operatorname{div} u_h + B\nabla p_h + c p_h - f \|_{0,T_i}^2 + \| u_h - A\nabla p_h \|_{0,T_i}^2 \\
&= \sum_{i=1}^{n_T} \mathcal{F}_i(p_h,u_h)
\end{aligned}$$

und somit sind die Voraussetzungen für den Einsatz von $\mathcal{F}$ als Fehlerschätzer erfüllt.

1.3.3 Aufstellen der linearen Gleichungssysteme

Im Gegensatz zum Verfahren der Finiten Differenzen, bei dem die Werte der Lösung an diskreten (Gitter-)Punkten berechnet werden, ist das Ergebnis der Methode der Finiten Elemente eine Funktion, die eine Approximation an die Lösung nicht nur an einzelnen Punkten, sondern über dem ganzen Gebiet Ω liefert. Diese Näherungslösung ist eine Linearkombination der Basis von $V_h \times W_h$. Sei $\{\Phi_i\}_{i=1}^{n_h}$ eine solche Basis, dann lässt sich die Lösung $(p_h,u_h) \in V_h \times W_h$ schreiben als

$$(p_h,u_h) \;=\; \sum_{i=1}^{n_h} z_i \Phi_i.$$

Eingesetzt in (1.63) ergibt sich damit die Aufgabe:

Suche $z = (z_1, z_2, \ldots, z_{n_h})^T \in \mathbb{R}^{n_h}$ mit

$$\mathcal{J}^*\mathcal{J}\Big(\sum_{i=1}^{n_h} z_i\Phi_i; q_h, v_h\Big) \;=\; l_f(q_h,v_h)$$

für alle $(q_h,v_h) \in V_h \times W_h$.

Wegen der Linearität von $\mathcal{J}^*\mathcal{J}$ und l_f reicht es in dieser Formulierung aus, nur gegen die Basisfunktionen Φ_j zu testen, d.h. das Problem

Suche $z = (z_1, z_2, \ldots, z_{n_h})^T \in \mathbb{R}^{n_h}$ mit

$$\mathcal{J}^*\mathcal{J}\Big(\sum_{i=1}^{n_h} z_i\Phi_i; \Phi_j\Big) \;=\; l_f(\Phi_j) \tag{1.65}$$

für $j = 1, \ldots, n_h$

zu lösen. Da $\mathcal{J}^*\mathcal{J}(\sum_{i=1}^{n_h} z_i\Phi_i;\Phi_j) = \sum_{i=1}^{n_h} z_i\mathcal{J}^*\mathcal{J}(\Phi_i;\Phi_j)$ gilt, ist (1.65) äquivalent zur Lösung des linearen Gleichungssystems

$$A_h\, z \;=\; b_h \tag{1.66}$$

mit

$$\begin{aligned} A_h &= [\mathcal{J}^*\mathcal{J}(\Phi_i;\Phi_j)]_{i,j=1}^{n_h}, \\ b_h &= [l_f(\Phi_j)]_{j=1}^{n_h}. \end{aligned}$$

Bei der Berechnung der Einträge in A_h müssen Integrale über Ω ausgewertet werden. Wegen der Linearität des Integrals reicht es aus, einzeln über den Elementen $T_i \in \mathcal{T}$ zu integrieren (üblicherweise mittels Quadraturformeln von genügend hoher Ordnung). So sind auch z.B. Integrale über den Gradienten von stückweise linearen Funktionen berechenbar (Kanten bilden eine Nullmenge). Wegen der im Allgemeinen nur lokalen Träger der Basisfunktionen Φ_i verschwinden viele dieser Integrale und damit die Einträge in A_h. Die Matrix ist also dünn besetzt, was die numerische Lösung des linearen Gleichungssystems vereinfacht.
Da $\{\Phi_i\}_{i=1}^{n_h}$ eine Basis für den Produktraum $V_h \times W_h$ bildet, wird sie durch Vereinigung der jeweiligen einzelnen Basen $\{\Phi_j^V\}_{j=1}^{n_{V_h}}$ von V_h und $\{\Phi_k^W\}_{k=1}^{n_{W_h}}$ von W_h folgendermaßen konstruiert:

$$\Phi_i \;=\; \begin{cases} (\Phi_i^V, 0) & \text{für} \quad i = 1, \ldots, n_{V_h} \\ (0, \Phi_{i-n_V}^W) & \text{für} \quad i = n_{V_h}+1, \ldots, n_{V_h}+n_{W_h} = n_h. \end{cases}$$

Dadurch besitzt die Matrix A_h eine natürliche Blockstruktur:

$$A_h \;=\; \begin{pmatrix} A_{pp} & A_{up} \\ A_{pu} & A_{uu} \end{pmatrix}.$$

Die Bilinearformen, die jeweils bei der Berechnung der einzelnen Blöcke ausgewertet werden, lassen sich leicht aus der Variationsformulierung (1.62) ablesen. Bei der numerischen Behandlung des LGS kann die Blockstruktur durch Konstruktion geeigneter Vorkonditionierer für die einzelnen Formen ausgenutzt werden.

1.4 FEM : Lösung der Gleichungssysteme

Für mehr Einfachheit in der Darstellung werden in diesem Abschnitt nur Gebiete $\Omega \in \mathbb{R}^2$ betrachtet, die in Triangulierungen zerlegt wurden.
Die linearen Gleichungssysteme, die aus dem Galerkin-Verfahren zur Lösung des Variationsproblems entstehen, sind im Allgemeinen zu groß, um sie direkt mit vertretbarem Aufwand lösen zu können. Allerdings sind diese Gleichungssysteme meistens dünn besetzt und unterliegen bei einer gemischten Formulierung einer natürlichen Blockstruktur.
Die numerische Lösung solcher Gleichungssysteme ist zum einen mittels bekannter numerischer Verfahren möglich, wie dem CG-Verfahren (A_h symmetrisch und positiv definit) oder anderen Krylovraum-Verfahren, ohne zu berücksichtigen, was für ein Problem dem linearen Gleichungssystem zu Grunde liegt. Für das Konvergenzverhalten ist dabei die Konditionszahl $\kappa(A_h)$ der Steifigkeitsmatrix A_h entscheidend, die allerdings häufig wesentlich von der Feinheit der Zerlegung h abhängt. Für Diskretisierungen von $-\triangle$ gilt zum Beispiel

$$\kappa(A_{-\triangle,h}) \;=\; O(\frac{1}{h^2}).$$

Zur stabilen und schnellen Lösung solcher Aufgaben ist daher der Einsatz von Vorkonditionierern notwendig, die die Kondition der Matrix A_h verbessern und, wenn möglich, unabhängig

von der Feinheit h nach oben beschränken.
Die (häufig schlechte) Konditionierung der Steifigkeitsmatrizen A_h wird durch die Variationsformulierung, das Gebiet Ω, die Randbedingungen sowie die benutzten Galerkin-Ansatzräume W_h, V_h beeinflusst. Die Konstruktion von effektiven Vorkonditionierern und Lösern sollte diese Eigenschaften berücksichtigen und gehört deshalb zur Aufgabe der FEM. Da die Berechnungen durch Computer ausgeführt werden, gehört auch eine Abschätzung der Anzahl der zur Lösung benötigten Rechenoperationen zu diesen Aufgaben. Dabei gilt ein Aufwand, der proportional zur Anzahl der Unbekannten n ist, als optimal.

Vorkonditionierer werden häufig auch als Glätter bezeichnet. Diese Ausdrucksweise erklärt sich wie folgt: Sei z_0 eine Anfangsnäherung an die Lösung des LGS $Az = b$. In Gedanken führe man eine Orthonormalbasis aus Eigenvektoren e_i zu den geordneten Eigenwerten $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ ein und zerlege das Anfangsresiduum $r_0 = Az_0 - b$ in seine Darstellung bzgl. dieser Basis :

$$r_0 \;=\; \sum_{i=1}^{n} r_0^{(i)} e_i.$$

Sei z_1 die Näherung nach Anwenden eines Schrittes eines vorkonditionierenden Verfahrens und $r_1 = Az_1 - b$ das neue Residuum mit der Darstellung

$$r_1 \;=\; \sum_{i=1}^{n} r_1^{(i)} e_i.$$

Bei den meisten gebräuchlichen Vorkonditionierern (wie z.B. Krylovraumverfahren) ist nun der Fehleranteil in r_1 zu großen Eigenwerten $\lambda_n, \lambda_{n-1}, \ldots, \lambda_g$, gemessen in $\sum_{i=g}^{n} (r_1^{(i)})^2$, stärker zurückgegangen als derjenige zu kleinen Eigenwerten. (Anschaulich ist klar, wie im Extremfall eines Eigenwerts gleich Null keine Fehleranteile in dieser Richtung gedämpft werden können, das Problem wäre ja auch nicht mehr eindeutig lösbar.) Diesen Effekt nennt man auch die Glättung des Fehlers (weitere Details im nächsten Abschnitt 1.4.1).
Eine effektive Kombination von Glätter und Löser muss also die Fehleranteile nicht nur der großen, sondern auch der kleinen Eigenwerte wirksam verringern. Kleine Eigenwerte entstehen zum einen durch feiner werdende Triangulierungen, zum andern jedoch auch durch die besondere Struktur mancher Lösungsräume (vgl. Abschnitt 1.4.2).

1.4.1 Lineare Multilevelverfahren

Der Zusammenhang zwischen der Größe der Eigenwerte und der Feinheit der Triangulierung lässt sich an folgendem Beispiel verdeutlichen:
Sei A_h die Steifigkeitsmatrix zur Variationsformulierung der Gleichung

$$-\triangle u \;=\; f$$

auf dem Einheitsquadrat $(0,1)^2 \subseteq \mathbb{R}^2$. Die Eigenwerte dieser Matrix lassen sich für eine gleichförmige Zerlegung mit Feinheit $h = 2^{-l}$ ($l \in \mathbb{N}$) explizit angeben:

$$\lambda_{i,j} \;=\; 4(\sin^2 \frac{ih\pi}{2} + \sin^2 \frac{jh\pi}{2}) \tag{1.67}$$

mit $1 \leq i, j \leq 2^l - 1$. Die Zahl l nennt man auch das Level der Zerlegung.
Bei Approximationen auf hohem Level (also feiner Zerlegung) sieht man an (1.67), dass die Anzahl von "kleinen" Eigenwerten (kleiner z.B. als ein festes $\epsilon > 0$) und damit der Anteil von schwerer zu glättenden Fehlerkomponenten stark zunimmt. Die zugehörigen Eigenfunktionen oszillieren nur wenig. Daher lassen sich solche Fehleranteile auch gut auf gröberen Triangulierungen darstellen (niederfrequente Fehleranteile). Auf niedrigeren Leveln gehören diese Fehleranteile dann

(für den Laplace-Operator) zu größeren Eigenwerten und können dort durch Anwenden eines Glätters reduziert werden. Diese Idee der Lösung des linearen Systems (1.66) mittels Glättung und Grobgitterkorrektur liegt jedem Multilevelverfahren (MLV) zu Grunde. Daher basiert die Konvergenztheorie von MLV auf der Untersuchung des Glättungseffekts (Glättungseigenschaft) und dem Einfluss des Übergangs von einem Level auf das nächste (Approximationseigenschaft).

Grundlage für die Darstellung des Multilevelalgorithmus sei eine durch sukzessive Verfeinerung der zu Grunde liegenden Triangulierung entstandene Hierarchie von diskreten Ansatzräumen:

$$V_0 \times W_0 \subseteq V_1 \times W_1 \subseteq \ldots \subseteq V_L \times W_L.$$

Der Subindex l steht dabei für die Zahl des Levels mit h_l als Feinheit der zugehörigen Triangulierung. Zwischen den Räumen bestehe eine Restriktionsabbildung

$$\mathcal{I}_l^{l-1} : V_l \times W_l \rightarrow V_{l-1} \times W_{l-1},$$

für die

$$\mathcal{J}^*\mathcal{J}(p_l, u_l; q_{l-1}, v_{l-1}) = \mathcal{J}^*\mathcal{J}(\mathcal{I}_l^{l-1}(p_l, u_l); q_{l-1}, v_{l-1})$$

für alle $(q_{l-1}, v_{l-1}) \in V_{l-1} \times W_{l-1}$ gilt.
Eine solche Abbildung existiert: Definiert man die Einbettungsabbildung (Prolongation) mit

$$\begin{aligned} \mathcal{I}_{l-1}^{l} : V_{l-1} \times W_{l-1} &\rightarrow V_l \times W_l, \\ (p_{l-1}, u_{l-1}) &\mapsto (p_{l-1}, u_{l-1}), \end{aligned}$$

so bildet die bezüglich $\mathcal{J}^*\mathcal{J}$ gebildete Adjungierte von $\mathcal{I}_{l-1}^{l}$ die sogenannte kanonische Restriktion $\mathcal{I}_l^{l-1}$.
Bezeichne A_l die auf Level l aufgestellte Steifigkeitsmatrix sowie b_l die zugehörige rechte Seite. Sei S_l die Matrixdarstellung des Glättungsverfahrens (z.B. $S_l = (I - \mathrm{diag}(A_l)^{-1}A_l)$ für das Jacobi-Verfahren). Die wesentlichen Bestandteile des MLV lassen sich bereits im Fall von zwei beteiligten Leveln (Zweigitterverfahren) erkennen:
Ist $z_l^{(0)}$ eine Startnäherung an die Lösung z_l von $A_l z_l = b_l$, so gilt für den Fehler nach Anwendung eines Zweigitterschritts

$$z_l - z_l^{(1)} = \underbrace{S_l^{\nu_2}}_{III} \underbrace{(I - I_{l-1}^{l} A_{l-1}^{-1} I_l^{l-1})}_{II} \underbrace{S_l^{\nu_1}}_{I} (z_l - z_l^{(0)}). \tag{1.68}$$

Man erkennt die beiden Aspekte des Verfahrens: Glättung durch Anwendung von ν_1 Vor- und ν_2 Nachglättungsschritten (I und III) und Grobgitterkorrektur (II). Anstelle der Matrix A_{l-1} kann auch die Galerkin-Approximation $\tilde{A}_{l-1} = (I_{l-1}^{l})^T A_l I_{l-1}^{l}$ für die Grobgitterkorrektur verwendet werden, wobei I_{l-1}^{l} die Matrix-Darstellung der Prolongation als linearer Abbildung zwischen zwei endlichdimensionalen Räumen bezeichnet. Im Fall der kanonischen Restriktion und Prolongation sind die Matrizen A_{l-1} und $\tilde{A}_{l-1}$ offensichtlich identisch.
Beim Zweigitterverfahren wird auf dem gröberen Gitter $l-1$ exakt gelöst (siehe (1.68)). Um dieses Verfahren zum MLV zu erweitern, wird das exakte Lösen auf Level $l-1$ durch eine Anzahl $p\,(p \in \mathbb{N})$ weiterer Zweigitterschritte zwischen den Leveln $l-1$ und $l-2$ ersetzt. Ist Level 0 erreicht, so kann das System im Allgemeinen wegen seiner geringen Größe exakt gelöst werden. Durch die Rückführung auf den Zweigitter-Fall kann das MLV rekursiv implementiert werden. Dabei spricht man von einem V-Zyklus für $p = 1$ und einem W-Zyklus für $p = 2$.
"Full Multigrid"-Varianten sind eine weitere Implementierungsmöglichkeit des V-Zyklus. Innerhalb einer geschachtelten Iteration wird dabei die Prolongation einer Näherung auf Level $l-1$, die dort durch einen oder mehrere V-Zyklen bestimmt wird, als Startnäherung für Level l benutzt.

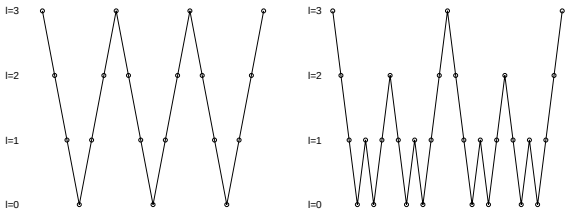


Abbildung 1.1: Schematische Darstellung eines V-Zyklus und eines W-Zyklus

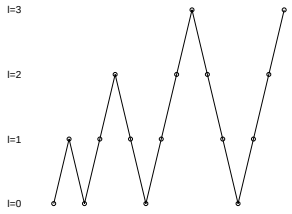


Abbildung 1.2: Schematische Darstellung eines Full Multigrid Zyklus

Als Glättungsiterationen bieten sich leicht zu implementierende Methoden wie Jacobi-, Gauß-Seidel-, CG- oder andere Krylovraum-Verfahren an. Es lässt sich zeigen, dass bei geeigneter Wahl des Glätters S (Glättungseigenschaft) und geeigneten Interpolationsoperatoren (Approximationseigenschaft) der Aufwand zur Lösung des linearen Gleichungssystems nicht mehr von der Feinheit der Zerlegung h_l abhängt, sondern die Kondition des durch ein MLV gelösten Systems sich für $h_l \to 0$ wie $O(1)$ verhält und damit der Aufwand zur Lösung proportional zur Anzahl der Unbekannten ist. Details zu Konvergenzaussagen und Konvergenzvoraussetzungen können z.B. in dem Lehrbuch von W. Hackbusch [27] gefunden werden.

Die Glätter und Interpolationsoperatoren für den Produktraum $V_l \times W_l$ werden über die einzelnen Räume V_l bzw. W_l konstruiert:
V_l ist ein Unterraum von $H^1(\Omega)$ und wird $\mathcal{J}^*\mathcal{J}$ auf $V_l \times V_l$ eingeschränkt, gilt

$$\mathcal{J}^*\mathcal{J}|_{V_l \times V_l} \;=\; (X\cdot, X\cdot)_{0,\Omega} + (A\nabla\cdot, A\nabla\cdot)_{0,\Omega}$$

mit $X = B\nabla + cI$. Da A gleichmäßig positiv definit ist, gilt also

$$\mathcal{J}^*\mathcal{J}|_{V_l \times V_l} \;\cong\; \|\cdot\|^2_{1,\Omega}$$

(siehe auch (1.59) und (1.60)) mit von h_l unabhängigen Konstanten.
Für dieses gleichmäßig elliptische Problem in $H^1(\Omega)$ sind Standard-Glätter wie z.B. das Gauß-Seidel-Verfahren sehr gut geeignet, da hochfrequente Fehleranteile auf allen Gittern zu großen Eigenwerten gehören (siehe dazu oben das Beispiel der Diskretisierung des Laplace-Operators). Zur Anwendung der Interpolationsoperatoren benötigt man nun nur noch eine Matrix-Darstellung des Prolongationsoperators, denn es soll der kanonische Restriktionsoperator mit der Darstellung als Transponierter dieser Matrix verwendet werden. Der Prolongationsoperator für die in Abschnitt 1.3.2 dargestellte nodale Basis von V_l ist einfach zu finden, da ja nur die Darstellung einer stückweise linearen Funktion des groben Gitters über der feineren Zerlegung gesucht wird (siehe z.B. [8]).
Bei der Diskretisierung von W werden die Raviart-Thomas-Räume (RT-Räume) niedrigster Ordnung verwendet, deren Funktionen im zweidimensionalen Fall durch die Vorschrift

$$\varphi(x,y) = \begin{pmatrix} a + cx \\ b + cy \end{pmatrix} \tag{1.69}$$

auf jedem Dreieck $T \in \mathcal{T}$ mit $a, b, c \in \mathbb{R}$ definiert sind. Als Freiheitsgrade dienen (siehe auch Abschnitt 1.3.2) die Normalkomponenten $\vec{n} \cdot \varphi$ auf den Kanten von T.
Wird die Bilinearform auf W_l eingeschränkt, erhält man

$$\begin{aligned} \mathcal{J}^*\mathcal{J}|_{W_l \times W_l} &= (\cdot,\cdot)_{0,\Omega} + (\operatorname{div} \cdot, \operatorname{div} \cdot)_{0,\Omega} \\ &= \|\cdot\|^2_{H(\operatorname{div},\Omega)}. \end{aligned}$$

Es existieren daher hochfrequente Fehleranteile, die auf allen (!) Gittern zu kleinen Eigenwerten gehören. Um dennoch ein optimales MLV zu konstruieren, muss der Glätter auch diese Fehleranteile eliminieren können. Darauf wird im nächsten Abschnitt eingegangen.
Bei der Konstruktion der Prolongation zwischen zwei Räumen W_{l-1} und W_l muss darauf geachtet werden, dass für beliebige $\vec{v} \in \mathbb{R}^2$ und RT-Funktion φ, die auf einem $T \in \mathcal{T}$ definiert sind, auf jeder Kante e_T des Elements T

$$\vec{v} \cdot \varphi(x, y) = c_{\vec{v}} \in \mathbb{R},\ (x, y) \in e_T$$

gilt. Wird ein Dreieck verfeinert, so entstehen in dem Dreieck zusätzliche Kanten, also zusätzliche Freihheitsgrade e_l, auf denen für die Prolongation auf diese Kanten die Stetigkeit der Normalkomponente mit dem entsprechenden Wert $c_{\vec{n}_l}$ gefordert werden muss ($\vec{n}_l$ Normale auf e_l).
Die Randkanten des Dreiecks e_l^r werden bei der Verfeinerung halbiert. Dadurch ändert sich jedoch der Wert der Normalkomponente $c_{\vec{n}_l^r}$ nicht und wird auf die halbierten Kanten "vererbt". Bei den innenliegenden Kanten wird ebenfalls der Wert der Normalkomponente ermittelt und für die Prolongation verwendet. Auf diese Weise wird die Prolongationsmatrix auf jedem Element definiert und zusammengefasst. Dadurch stellt dieses Verfahren die Einbettung des gröberen Raums in den feineren Raum dar.
Die Gesamtprolongationsmatrix für den Raum $V_l \times W_l$ wird schließlich aus den beiden als Blöcke in die Hauptdiagonale geschriebenen einzelnen Prolongationsmatrizen gebildet. Wird für beide Räume die beschriebene Einbettung als Prolongation und die kanonische Restriktion verwendet, so erfüllt dies trivialerweise die Approximationseigenschaft (vgl. [27]), falls die Räume W_l und V_l diese jeweils erfüllen.

1.4.2 Glättungsiterationen für $W_l \subseteq H(\mathbf{div}, \Omega)$

Ein wirksamer Glätter, der innerhalb eines Multilevelverfahrens verwendet werden kann, muss die hochfrequenten Fehleranteile der Approximation dämpfen, während die übrigen Fehleranteile innerhalb der Grobgitterkorrektur eliminiert werden können. Bei der Verwendung von Standard-Glättern für elliptische Probleme wie Gauß-Seidel oder Jacobi-Glättung ist dafür jedoch notwendig, dass die hochfrequenten Fehleranteile - d.h. Eigenfunktionen, die nur auf dem feineren Gitter darstellbar sind - auch zu großen Eigenwerten gehören. Diese Voraussetzung ist jedoch nicht mehr gegeben, wenn die Bilinearform $\mathcal{J}^*\mathcal{J}$ auf dem Raum $W_l \times W_l$ betrachtet wird. In diesem Fall existieren hochfrequente Fehleranteile, die zu kleinen Eigenwerten gehören. Dabei sind diese kleinen Eigenwerte für wachsendes l (also feiner werdende Triangulierung) nicht von der Null weg beschränkt. Diese Beobachtung kann mit Hilfe der diskreten Helmholtz-Zerlegung des Raums W_l erklärt werden.
Sei W_l der oben definierte Raviart-Thomas Raum niedrigster Ordnung über einer Triangulierung $\mathcal{T}$ eines einfach zusammenhängenden Polygongebiets Ω. Zur Vereinfachung sei hier zusätzlich $\Gamma_N = \emptyset$. Definiere die diskreten Räume

$$\begin{aligned} L_l &= \{s \in H^1(\Omega) : s|_T \text{ linear}, T \in \mathcal{T}\} \\ C_l &= \{s \in L^2(\Omega) : s|_T \text{ konstant}, T \in \mathcal{T}\} \end{aligned}$$

und den diskreten Gradient-Operator

$$\begin{aligned} &\nabla_h : C_l \to W_l \\ &(\nabla_h s, u)_{0,\Omega} = -(s, \operatorname{div} u) \ \forall u \in W_l. \end{aligned} \tag{1.70}$$

Unter diesen Voraussetzungen erhält man bei Restriktion auf einzelne Elemente der Zerlegung

Lemma 1.15 (Diskrete Helmholtz-Zerlegung) *Sei W_l, L_l, C_l und ∇_h definiert wie zuvor. Dann gilt (lokal) die Zerlegung*

$$W_l = \nabla_h C_l \oplus \nabla^\perp L_l. \tag{1.71}$$

Diese Zerlegung ist orthogonal bzgl. $\|\cdot\|_{0,\Omega}$ und $\|\cdot\|_{H(div,\Omega)}$.

Zum Beweis siehe [3] oder [33]. Die globale Zerlegung, die auch die Randbedingungen berücksichtigt, wird später in allgemeinerem Zusammenhang angegeben.
Auf $\nabla_h C_l$ ist $\mathcal{J}^*\mathcal{J}$ ein Operator, der $I - \triangle$ entspricht, also ein Operator zweiter Ordnung. Von solchen elliptischen Operatoren ist bekannt, dass hochfrequente Eigenfunktionen auch zu großen Eigenwerten gehören. Auf diesem Teil der Zerlegung (1.71) ist also der Einsatz eines Standard-Glätters wie dem Gauß-Seidel-Verfahren ausreichend.
Die Einschränkung von $\mathcal{J}^*\mathcal{J}$ auf $\nabla^\perp L_l$ ist jedoch wegen $\operatorname{div}\nabla^\perp = 0$ lediglich das L^2-Innenprodukt. Es fehlt die Verstärkung hochfrequenter Funktionen durch diesen Operator, so dass für zu hochfrequenten Eigenfunktionen dieses Teils der Zerlegung nur kleine Eigenwerte gehören. Für Fehlerkomponenten in diesem Raum sind also solche Verfahren, die nur Fehleranteile in großen Eigenwerten reduzieren, als Glätter im Multilevelprozess nicht geeignet.
In gewissem Sinne ist fehlende (gleichmässige) Elliptizität des zu glättenden Operators auf diesem Teilraum die Ursache dieser Schwierigkeit. Daher wird in [29] folgende Umformulierung des Problems vorgeschlagen: Sei $u \in \nabla^\perp L_l$. Dann gibt es eine stückweise lineare Funktion l_u mit $u = \nabla^\perp l_u$. Dann ist jedoch

$$\begin{aligned} \mathcal{J}^*\mathcal{J}(u,0;u,0) &= (u,u)_{0,\Omega} + (\operatorname{div} u, \operatorname{div} u)_{0,\Omega} \\ &= (\nabla^\perp l_u, \nabla^\perp l_u)_{0,\Omega} + (\operatorname{div} \nabla^\perp l_u, \operatorname{div} \nabla^\perp l_u)_{0,\Omega} \\ &= (\nabla^\perp l_u, \nabla^\perp l_u)_{0,\Omega} \\ &= (\nabla l_u, \nabla l_u)_{0,\Omega}. \end{aligned}$$

(NB: Die letzte Umformung gilt nur in Gebieten $\Omega \subseteq \mathbb{R}^2$. In Gebieten $\Omega \subseteq \mathbb{R}^3$ ist in [29] die sich in diesem Fall ergebende diskrete Helmholtz-Zerlegung und weitere Vorgehensweise angegeben.)
Es gilt in gewissem Sinne also

$$\mathcal{J}^*\mathcal{J}|_{\nabla^\perp L_l \times \nabla^\perp L_l} = -\triangle|_{L_l \times L_l}.$$

Damit steht aber auf dem eigentlich schwierig zu vorkonditionierenden (weil divergenzfreien) Raum $\nabla^\perp L_l$ durch Übergang zum sogenannten Potentialraum L_l ein elliptischer Operator zur Verfügung, zu dessen Glättung wieder Standard-Vorkonditionierer eingesetzt werden können (Potentialraum-Verfahren).

Das Problem der divergenzfreien Unterräume bei Diskretisierungen von $H(\operatorname{div}, \Omega)$ kann auch durch Einsatz des in [3] vorgestellten Block-Schwarz-Vorkonditionierers gelöst werden, was auch der Multilevel-Projection-Method in [36] entspricht. Dieses Verfahren ist jedoch nach Untersuchungen in [34] aufwendiger als das Potentialraum-Verfahren und führt zu schlechteren Konvergenzraten. Die Ursache dafür liegt vor allem an der Tatsache, dass das Potentialraum-Verfahren als nodales Gauß-Seidel-Verfahren implementiert werden kann, während bei Block-Verfahren die

Invertierung mehrdimensionaler Probleme notwendig ist.
Bei Anwendung des Potentialraum-Verfahrens auf kompliziertere Gebiete (nur zusammenhängend, nicht einfach zusammenhängend) und Randbedingungen ($\Gamma_N \neq \emptyset$) muss ebenfalls eine Zerlegung des zugehörigen diskreten Raums W_l in eine direkte Summe aus Gradienten und Rotationsanteil wie in (1.71) durchgeführt werden, um Fehleranteile in beiden Teilen glätten zu können. Im Falle nicht einfach zusammenhängender Gebiete muss die direkte Zerlegung dabei um einen weiteren Raum G_0 erweitert werden, der ein Unterraum von W_0, dem Ansatzraum bezüglich des gröbsten Gitters ist:

$$W_l \;=\; \nabla_h \tilde{C}_l \;\oplus\; \nabla^\perp \tilde{L}_l \;\oplus\; G_0 \tag{1.72}$$

mit

$$\begin{aligned} \tilde{C}_l \;&\subseteq\; C_l\,,\; \tilde{L}_l \subseteq L_l \\ \forall g_o \in G_0 \;&:\; \text{div } g_0 = 0 \;\text{ und }\; \nexists\, w_0 \in L_0 : \nabla^\perp w_0 = g_0. \end{aligned}$$

Bezeichne N_p, N_e und N_t die Anzahl der Punkte, Kanten und Dreiecke der Triangulierung $\mathcal{T}_l$ des Gebiets Ω, N_H die Anzahl der Löcher des Gebiets Ω und N_p^Γ bzw. N_e^Γ die Anzahl der zu Γ gehörigen Punkte bzw. Kanten. Dann ist

$$N_e \;=\; N_p \;+\; N_t \;+\; N_H \;-\; 1.$$

Weiterhin sei $N_H = N_{H,N} + N_{H,D} + N_{H,ND}$, wobei der zweite Index die Art der Randbedingung des jeweiligen Lochs des Gebiets angibt ($D \sim$ nur Dirichlet, $N \sim$ nur Neumann, $ND \sim$ sowohl Dirichlet als auch Neumann-Bedingungen auf dem Rand vorgeschrieben). Schließlich sei $N_{N,nc}$ die Anzahl der disjunkten Teile des Randes Γ_N, die für sich keine geschlossene Kurve bilden. Dann gilt (siehe [34]):

$$\begin{aligned} \dim(W_l) \;&=\; N_e - N_e^{\Gamma_N} \\ \dim(\tilde{C}_l) \;&=\; N_t - 0^{N_p^{\Gamma_D}} \\ \dim(\tilde{L}_l) + \dim(G_0) \;&=\; \underbrace{N_p - N_p^{\Gamma_N}}_{I} + \underbrace{N_{N,nc} + N_{H,N}}_{II} + \underbrace{N_{H,D} + N_{H,ND}}_{III} - 1 + 0^{N_p^{\Gamma_D}}. \end{aligned} \tag{1.73}$$

Teil I der letzten Summe entspricht der Darstellung von lokal divergenzfreien Funktionen als $\nabla^\perp f_k$ mit einer nodalen Basisfunktion $f_k \in L_l$. Der zweite Teil entspricht der Darstellung $\nabla^\perp \sum_{k\in\mathcal{I}} f_k$, also einer Summe von Basisfunktionen des Potentialraums L_l. Teil III schließlich entspricht divergenzfreien Flüssen in W_l, die zwei disjunkte Teile des Dirichlet-Randes verbinden, die also offensichtlich nicht als $\nabla^\perp \sum_{k\in\mathcal{I}} f_k,\ f_k \in L_0$ auf dem gröbsten Gitter $\mathcal{T}_0$ dargestellt werden können.
Mit der Beschreibung dieser Komponenten der Zerlegung in divergenzfreien und divergenzbehafteten Teil ist die Konstruktion des zugehörigen Glätters nach dem oben beschriebenen Schema eine einfache Folgerung. Dazu wird eine Basisdarstellung von $\nabla^\perp \tilde{L}_l \oplus G_0$ benötigt, die einfach über die drei Teile I, II, III der Zerlegung (1.73) konstruierbar ist.
Als Glätter werden nun zwei aufeinanderfolgende Operationen benötigt: Als erstes ein nodaler Gauß-Seidel-Schritt über alle Freiheitsgrade in W_l, um die hochfrequenten Anteile in $\nabla_h \tilde{C}_l$ zu dämpfen. Danach wird mittels obiger Basisdarstellung das Residuum in den Potentialraum projiziert und ein nodaler Gauß-Seidel-Schritt auf den Freiheitsgraden des Potentialraums durchgeführt, um die hochfrequenten Anteile in $\nabla^\perp \tilde{L}_l \oplus G_0$ zu dämpfen. Dadurch wird ausgenutzt, dass diese Anteile im Potentialraum wegen Elliptizität des darauf eingeschränkten Operators zu großen Eigenwerten gehören, während diese Komponenten im ursprünglichen Raum zu kleinen Eigenwerten gehörten. Damit steht auch für W_l ein effizienter Glätter zur Verfügung.

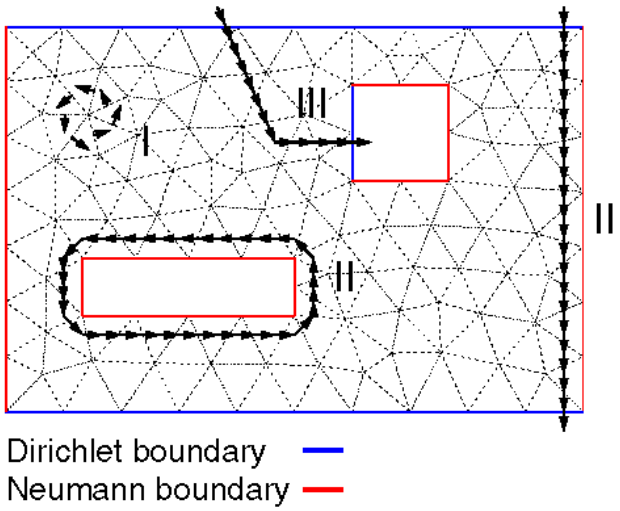


Abbildung 1.3: Schematische Darstellung der divergenzfreien Funktionen

1.4.3 Multilevelverfahren für nichtlineare PDGL

Multilevelverfahren für lineare Probleme basieren auf dem Zweigitteroperator, wie er in (1.68) ablesbar ist. Die Grobgitterkorrektur darin kann auch als Lösung einer Defektgleichung verstanden werden: Sei $A_l z_l = b_l$ das zu lösende Problem auf Level l und $z_L^{(1)}$ eine Näherung an die Lösung auf dem aktuellen Level L. Dann ist

$$\begin{aligned} A_L z_L &= b_L \\ A_L z_L - A_L z_L^{(1)} &= b_L - A_L z_L^{(1)} \\ A_L(\underbrace{z_L - z_L^{(1)}}_{=:d_L}) &= \underbrace{b_L - A_L z_L^{(1)}}_{=:r_L} \\ A_L d_L &= r_L \end{aligned} \tag{1.74}$$

die noch zu lösende Defektgleichung. Ist $d_L^{(1)}$ eine Näherung an d_L, dann ist $z_L^{(2)} = z_L^{(1)} + d_L^{(1)}$ eine verbesserte Näherung an die Lösung z_L des ursprünglichen Systems. Im Zweigitterverfahren wird die letzte Gleichung in (1.74) nach dem Transfer auf dem gröberen Gitter $L-1$ gelöst (kanonische Restriktion):

$$\begin{aligned} I_L^{L-1} A_L I_{L-1}^L d_{L-1} &= I_L^{L-1} r_L \\ \Leftrightarrow A_{L-1} d_{L-1} &= I_L^{L-1} r_L \\ d_L^{(1)} &= I_{L-1}^L d_{L-1}\,. \end{aligned} \tag{1.75}$$

Im nichtlinearen Fall ist dieses Vorgehen nicht mehr möglich, da beim Schritt von der zweiten zur dritten Zeile in (1.74) die Linearität von A_L benutzt wird. Daher wird die Defektgleichung (1.74) im nichtlinearen Fall anders aufgestellt: Zu lösen sei die nichtlineare Gleichung

$$A_L(z_L) = b_L. \tag{1.76}$$

Dann wird (1.74) zu

$$\begin{aligned} A_L(z_L) - A_L(z_L^{(1)}) &= b_L - A_L(z_L^{(1)}) \\ A_{L-1}(\tilde{I}_L^{L-1} z_L^{(1)} + z_{L-1}) - A_{L-1}(\tilde{I}_L^{L-1} z_L^{(1)}) &= I_L^{L-1} r_L \\ A_{L-1}(\tilde{I}_L^{L-1} z_L^{(1)} + z_{L-1}) &= I_L^{L-1} r_L + A_{L-1}(\tilde{I}_L^{L-1} z_L^{(1)}), \end{aligned} \tag{1.77}$$

wobei z_{L-1} gesucht ist. Dadurch ist auch auf Level $L-1$ ein nichtlineares Problem zu lösen. Dabei ist zu beachten, dass im Gegensatz zum linearen Fall im Allgemeinen $A_{L-1} \neq I_L^{L-1} A_L I_{L-1}^L$ gilt. Als Grobgitterkorrektur wird dann

$$d_L^{(1)} = I_{L-1}^L (z_{L-1}^{(1)} - \tilde{I}_L^{L-1} z_L^{(1)})$$

gesetzt, wenn $z_{L-1}^{(1)}$ Gleichung (1.77) näherungsweise löst.

Der Ausdruck $\tilde{I}_L^{L-1} z_L^{(1)}$ ist in diesen Gleichungen also eine Startnäherung für die Lösung auf dem gröberen Level $L-1$. Gleichzeitig wird auf diese Weise der Korrekturcharakter dieser Gleichung deutlich, denn das nichtlineare Problem auf dem gröberen Gitter muss an einer bestimmten, der bisher schon erreichten Näherung entsprechenden Stelle aufgestellt werden. Von dieser Stelle aus wird dann die Korrektur berechnet.
Bei der Wahl der Restriktion $\tilde{I}_L^{L-1}$ stehen zwei Möglichkeiten zur Auswahl:
Zur Reduzierung von Rechenaufwand kann man mit $\tilde{I}_L^{L-1} z_L^{(1)} = z_{L-1}^{(0)}$ für alle $z_L^{(1)}$ stets die gleiche Näherung auf dem gröberen Level wählen. Dies entspricht der Nonlinear-Multilevel-Method von W. Hackbusch ([27]). Ist jedoch $\|I_L^{L-1} A_L(z_L^{(1)}) - A_{L-1}(z_{L-1}^{(0)})\|$ zu groß, so dominiert dieser Anteil die rechte Seite von (1.77), die Grobgitterkorrektur $d_L^{(1)}$ wird sehr klein und damit nahezu nutzlos. Es ist also notwendig, die Werte $z_l^{(0)}$ in der Nähe der Lösung z_L von (1.76) zu wählen (zu Dämpfungsstrategien und der Wahl der $z_l^{(0)}$ siehe [28]).
In [9] schlägt A. Brandt die Verwendung einer Restriktion $\tilde{I}_L^{L-1}$ der Art vor, dass $\tilde{I}_L^{L-1} z_L^{(1)}$ eine Repräsentierung der aktuellen Feingitternäherung $z_L^{(1)}$ auf Level $L-1$ darstellt (FAS-Scheme = Full Approximation Storage-Scheme). Nimmt man diesen zusätzlichen Aufwand in Kauf, ist die Startnäherung für das grobe Gitter automatisch so nahe an der Lösung von (1.76) wie man es eben in diesem Iterationsschritt erreichen kann. Dadurch fallen die Probleme der Nonlinear-Multilevel-Method in Hinsicht auf die Wahl von $z_l^{(0)}$ weg.

Konvergenzaussagen für nichtlineare Multilevelverfahren sind weitaus schwieriger zu beweisen als im linearen Fall. Daher kann zur Lösung eines nichtlinearen Problems auch folgende Methode verwendet werden:
Es wird der Operator bereits auf dem feinsten Gitter linearisiert (siehe Abschnitt 1.1), wonach man das nun lineare Problem mit dem linearen Multilevelverfahren lösen kann. Damit wird auf gröberen Gittern nur noch eine lineare Korrekturgleichung gelöst.
Wird andererseits der nichtlineare Operator auf jedem Gitter (auch für die Grobgitterkorrektur) verwendet, wird die Nichtlinearität des Problems in gewisser Weise besser aufgelöst.
Diese beiden Ansätze werden in den folgenden Kapiteln bezüglich eines Ausgleichsproblems zur Lösung einer nichtlinearen partiellen Differentialgleichung verglichen. Dabei wird auch auf Bedingungen zur Konvergenz der Verfahren eingegangen.

Kapitel 2

Multilevelverfahren für nichtlineare Ausgleichsformulierungen

Ausgleichsformulierungen zur Lösung nichtlinearer partieller Differentialgleichungen sind seit einiger Zeit Gegenstand intensiver Analysen geworden (siehe z.B. [7]). Insbesondere bei der Umformulierung von Gleichungen zweiter Ordnung in ein System von Gleichungen erster Ordnung (Abschnitt 1.3.1) sind Ausgleichsformulierungen ein bewährtes Werkzeug, da bei diesen Problemen die Differenzierbarkeitsvoraussetzungen an die Lösung durch die Ausgleichsformulierung nicht verschärft werden.

Bei der effektiven Lösung von nichtlinearen partiellen Differentialgleichungen mit Hilfe der Methode der Finiten Elemente treffen nun zwei der wichtigsten numerischen Themengebiete aufeinander: Das traditionsreiche Gebiet der Lösung nichtlinearer Gleichungen (Stichwort Newton-Verfahren) trifft auf das junge, aber nicht weniger erfolgreiche Gebiet der Entwicklung von Multilevelverfahren für Variationsformulierungen. Je nach der Reihenfolge, in der diese beiden Aspekte bei der Lösung einer nichtlinearen Variationsformulierung verwendet werden, können so nach der Diskretisierung der Gleichung zwei verschiedene Verfahren verwendet werden:
Das **DLL**-Verfahren: Zuerst die Linearisierung der Systeme und Lösung des linearisierten Systems mit Hilfe eines linearen Mulitilevelverfahrens (**D**iskretisierung, **L**inearisierung, **L**ineares Multilevel). Das **DLL**-Verfahren setzt also nach der Diskretisierung als Löser ein Newton-artiges Verfahren ein, dessen Korrekturen mittels eines linearen Multilevelverfahrens bestimmt werden. Ein zweites Verfahren (**DNL**-Verfahren) besteht in der Verwendung der Möglichkeit, das nichtlineare Variationsproblem durch ein nichtlineares Multilevelverfahren zu lösen. Die Linearisierung der Probleme findet dann erst bei der Lösung der nichtlinearen Korrekturgleichungen auf jedem Level statt (**D**iskretisierung, **N**ichtlineares Multilevel, **L**inearisierung). Das **DNL**-Verfahren setzt als Löser also ein nichtlineares Multilevelverfahren ein, dessen nichtlineare Unterprobleme mit Newton-artigen Verfahren gelöst werden.

Setzt man die Diskretisierung als ersten Lösungsschritt fest, so kann es nur diese beiden Verfahren geben. Aus Gründen der Konvergenztheorie macht es jedoch auch Sinn, die Linearisierung als ersten Lösungsschritt zu setzen und die Diskretisierung der Variationsformulierung und lineares Multilevelverfahren zur Lösung der linearen Systeme folgen zu lassen (**LDL**-Verfahren). Darauf wird in Kapitel 3 eingegangen.

In diesem Kapitel 2 wird die Anwendung der beiden ersten Verfahren auf nichtlineare Ausgleichsformulierungen beschrieben. Das Hauptaugenmerk liegt dabei auf der algorithmischen

Beschreibung, an zweiter Stelle folgen Überlegungen zu Konvergenzaussagen für diese Verfahren. Numerische Ergebnisse zu Beispielproblemen mit allen drei angegebenen Verfahren werden dann im Kapitel 4 behandelt.

2.1 Definitionen

Im Folgenden bezeichnet $\Omega \subseteq \mathbb{R}^2$ das Gebiet, in dem eine partielle Differentialgleichung zu lösen ist. $H \subseteq (L^2(\Omega))^2$ sei ein Hilbertraum mit zugehöriger Norm $\|\cdot\|_{\mathrm{H}} \geq \|\cdot\|_{0,\Omega}$.
Weiterhin sei ein nichtlinearer, Fréchet-differenzierbarer Differentialoperator $\mathcal{R} : H \to (L^2(\Omega))^2$ gegeben, der die Taylorentwicklung

$$\mathcal{R}(x+h) \approx \mathcal{R}(x) + \mathcal{J}(x)h,$$

d.h. mit $\mathcal{J}(x) : H \to (L^2(\Omega))^2$ als Fréchet-Ableitung an der Stelle x, besitzt.
Die Adjungierte $\mathcal{J}^*(x) : (L^2(\Omega))^2 \to H'$ von $\mathcal{J}(x)$ wird dann durch

$$(y_1, \mathcal{J}(x)y_2)_{0,\Omega} = (\mathcal{J}^*(x)y_1, y_2)_{0,\Omega} \;\; \forall\, y_1 \in (L^2(\Omega))^2,\, y_2 \in H$$

definiert.
Weiterhin wird der Operator $\mathcal{J}^*(x)\mathcal{J}(x) : H \to H'$ durch folgende Gleichung definiert :

$$(\mathcal{J}^*(x)\mathcal{J}(x)y_1, y_2)_{0,\Omega} \;=\; (\mathcal{J}(x)y_1, \mathcal{J}(x)y_2)_{0,\Omega}\,. \;\; \forall y_1, y_2 \in H\,. \tag{2.1}$$

Werden nun die Abschätzungen

$$\begin{aligned} \|\mathcal{J}(x)\, y\|_{0,\Omega} &\leq \overline{\alpha}\|y\|_{\mathrm{H}} \quad \forall\, y \in H,\, x \in D \text{ und} && (2.2)\\ \|\mathcal{J}(x)y\|_{0,\Omega} &\geq \underline{\alpha}\|y\|_{\mathrm{H}} \quad \forall\, y \in H,\, x \in D && (2.3) \end{aligned}$$

an $\mathcal{J}(x)$ vorausgesetzt, ist leicht einzusehen, dass die Norm von $\mathcal{J}^*(x)\mathcal{J}(x)$ in D gleichmäßig nach oben und unten beschränkt ist:

$$\begin{aligned} \underline{\alpha}^2(y,y)_{\mathrm{H}} \leq (\mathcal{J}^*(x)\mathcal{J}(x)\, y, y)_{0,\Omega} \leq \overline{\alpha}^2(y,y)_{\mathrm{H}} \quad \forall y \in H,\, x \in D \\ \underline{\alpha}^2 \leq \|\mathcal{J}^*(x)\mathcal{J}(x)\| \leq \overline{\alpha}^2 \quad \forall\, x \in D. \end{aligned} \tag{2.4}$$

Aus der Operatoranalysis ist bekannt, dass dann die Inverse von $\mathcal{J}^*(x)\mathcal{J}(x)$ für $x \in D$ in $H \subseteq H'$ existiert und es gilt

$$\begin{aligned} \frac{1}{\overline{\alpha}^2}(z,z)_{\mathrm{H}} \leq ((\mathcal{J}^*(x)\mathcal{J}(x))^{-1}\, z, z)_{0,\Omega} \leq \frac{1}{\underline{\alpha}^2}\,(z,z)_{\mathrm{H}} \quad \forall z \in H,\, x \in D \\ \frac{1}{\overline{\alpha}^2} \leq \|(\mathcal{J}^*(x)\mathcal{J}(x))^{-1}\| \leq \frac{1}{\underline{\alpha}^2} \quad \forall\, x \in D. \end{aligned}$$

Eine Funktion $w \in H$ mit einem $\lambda \in \mathbb{R}$, für die

$$(\mathcal{J}^*(x)\mathcal{J}(x)w, v)_{\mathrm{H}} \;=\; \lambda(w,v)_{\mathrm{H}} \;\; \forall\, v \in H$$

gilt, heißt Eigenfunktion w zum Eigenwert λ von $\mathcal{J}^*(x)\mathcal{J}(x)$. Wegen (2.4) gilt für alle Eigenwerte λ des symmetrischen Operators $\mathcal{J}^*(x)\mathcal{J}(x)$

$$\underline{\alpha}^2 \;\leq\; \lambda \;\leq\; \overline{\alpha}^2\,. \tag{2.5}$$

Mit diesen Bezeichnungen sei

$$\mathcal{R}(x) = 0 \tag{2.6}$$

die zu lösende Differentialgleichung für $x \in H$ mit Lösung x_*. Gleichung (2.6) ist äquivalent mit der Formulierung

$$(\mathcal{R}(x), y)_{0,\Omega} = 0 \ \ \forall\, y \in (L^2(\Omega))^2\,. \tag{2.7}$$

Wird das Ausgleichsfunktional $\mathcal{F}$ durch

$$\mathcal{F}(x) = \|\mathcal{R}(x)\|^2_{0,\Omega}$$

definiert, so folgt aus (2.7) insbesondere, dass

$$\mathcal{F}(x_*) := \|\mathcal{R}(x_*)\|^2_{0,\Omega} = (\mathcal{R}(x_*), \mathcal{R}(x_*))_{0,\Omega} = 0 \tag{2.8}$$

gelten muss. Wegen $\mathcal{F}(x) \geq 0, \forall\, x \in H$ ist daher Problem (2.6) äquivalent mit dem Minimierungsproblem

$$\begin{aligned} &\text{Suche } x_* \in H \text{ mit} \\ &\mathcal{F}(x_*) \stackrel{!}{=} \min_{x \in H} \mathcal{F}(x)\,. \end{aligned} \tag{2.9}$$

Dieses Minimierungsproblem im unendlichdimensionalen Raum wird diskretisiert, indem nur über einem endlichdimensionalen Raum $H_l \subseteq H$ minimiert wird:

$$\begin{aligned} &\text{Suche } x_{l,*} \in H_l \text{ mit} \\ &\mathcal{F}(x_{l,*}) \stackrel{!}{=} \min_{x_l \in H_l} \mathcal{F}(x_l)\,. \end{aligned} \tag{2.10}$$

Eine solche Optimierungsaufgabe über dem Raum H_l wird mit Hilfe der Variationsrechnung gelöst. Die notwendige Bedingung für ein Minimum lautet dabei

$$\mathcal{F}'(x_{l,*}) \stackrel{!}{=} 0 \ \text{ auf } H_l \tag{2.11}$$

für ein $x_{l,*} \in H_l$.
Wird vorausgesetzt, dass die Differentialgleichung (2.6) eindeutig lösbar und das Funktional $\mathcal{F}$ in einer Kugel mit Durchmesser ρ um die Lösung x_* gleichmäßig elliptisch ist, so besitzt dieses Minimerungsproblem (2.11) dann mindestens ein lokales Minimum $x_{l,*}$, wenn der Abstand $d_l = \text{dist}(x_*, H_l)$ kleiner als ρ ist. Da bei den üblichen Finite-Element-Räumen der Abstand d durch sukzessive Verfeinerungen gegen Null geht, folgt daraus die Existenz einer lokalen Lösung von (2.11) allen passend gewählten Räumen H_l.

2.2 Ein lineares Multilevelverfahren (DLL)

Als Nullstellenproblem könnte die Bedingung (2.11) mit Hilfe des Newton-Verfahrens behandelt werden. Dann wäre jedoch die zweite Ableitung von $\mathcal{R}$ zu berechnen, was allerdings die Differenzierbarkeitsvoraussetzungen an die Nichtlinearitäten von $\mathcal{R}$ verschärft und auch auf Grund höheren Aufwands zu vermeiden ist. Im Gauß-Newton Verfahren wird daher der Term der zweiten Ableitung vernachlässigt. Die Anwendung des Algorithmus aus Abschnitt 1.1 führt dann auf eine Folge von linearen Ausgleichsproblemen, wo im k-ten Schritt die Aufgabe

$$\begin{aligned} &\text{Bestimme } \delta_{l,k} \in H_l \text{ mit} \\ &\|\mathcal{R}(x_{l,k}) + \mathcal{J}(x_{l,k})\delta_{l,k}\|^2_{0,\Omega} \stackrel{!}{=} \min_{\delta_l \in H_l} \|\mathcal{R}(x_{l,k}) + \mathcal{J}(x_{l,k})\delta_l\|^2_{0,\Omega} \end{aligned} \tag{2.12}$$

zu lösen ist (analog zu Gleichung (2.10) mit dem nichtlinearen Funktional $\mathcal{F}$). Dies geschieht durch Lösen der zugehörigen (linearen) Normalengleichung

$$\begin{aligned} &\text{Bestimme } \delta_{l,k} \in H_l \text{ mit} \\ &(\mathcal{J}(x_{l,k})\delta_{l,k}, \mathcal{J}(x_{l,k})x_l)_{0,\Omega} \;=\; -(\mathcal{R}(x_{l,k}), \mathcal{J}(x_{l,k})x_l)_{0,\Omega} \\ &\text{für alle } x_l \in H_l. \end{aligned} \tag{2.13}$$

Insgesamt ergibt sich damit (in Operatorschreibweise) folgender Algorithmus:

1. Wähle $x_{l,0} \in H_l$. Setze $k = 0$.
2. Bestimme $\delta_{l,k} = -(\mathcal{J}(x_{l,k})^*\mathcal{J}(x_{l,k}))^{-1}\mathcal{J}(x_{l,k})^*\mathcal{R}(x_{l,k})$ durch Lösen von Gleichung (2.13).
3. Setze $x_{l,k+1} = x_{l,k} + \delta_{l,k}$, $k = k + 1$.
4. Abbruchkriterium erfüllt? Sonst gehe zu 2.

Algorithmus 2.1 : Gauß-Newton-Verfahren

Die lineare Variationsformulierung (2.13) kann mit einem linearen Multilevelverfahren näherungsweise gelöst werden. Dadurch wird das Gauß-Newton-Verfahren aus Algorithmus 2.1 zu einem inexakten Gauß-Newton Verfahren, dessen Korrekturen $\tilde{\delta}_{l,k}$ mit einem Fehler $\epsilon_{l,k}$ behaftet sind: $\tilde{\delta}_{l,k} = \delta_{l,k} + \epsilon_{l,k}$. Im Folgenden sollen solche Genauigkeitsschranken für dieses Verfahren zur Minimierung in H_l hergeleitet werden, so dass die Konvergenz des Lösungsverfahrens dennoch bestehen bleibt.
Im Allgemeinen wird die dem Minimierungsproblem (2.9) zu Grunde liegende Differentialgleichung, die als in H eindeutig lösbar angenommen wurde, diese Lösung nicht im endlichdimensionalen Unterraum H_l besitzen. Dadurch gilt für den minimalen Wert in diesem Raum $\mathcal{F}(x_{l,*}) > 0$. Beim **DLL**-Ansatz muss daher die Theorie für inexakte Gauß-Newton-Verfahren für nicht kompatible Minimierungsprobleme (Abschnitt 1.1.6) herangezogen werden, um den Fehler $\epsilon_{l,k}$ so zu kontrollieren, dass das inexakte Verfahren dennoch Abstiegsrichtungen für das Funktional $\mathcal{F}$ liefert. Dafür müssen jedoch diese Bedingungen auf den Fall der Minimierung in Funktionenräumen übertragen werden.
An die Stelle von Voraussetzung 1.10 tritt damit

Voraussetzung 2.1 *Seien $\mathcal{R}$ und $\mathcal{J}$ definiert wie in Abschnitt 2.1 mit $x_{l,*} \in H$ als eindeutigem Minimum von $\mathcal{F}(x) = \|\mathcal{R}(x)\|^2_{0,\Omega}$ im Unterraum $H_l \subseteq H$. Sei $\mathcal{J}$ in H_l Lipschitz-stetig mit Konstante γ:*

$$\|\mathcal{J}(x_l)\, z_l \;-\; \mathcal{J}(y_l)\, z_l\|_{0,\Omega} \;\leq\; \gamma\, \|x_l - y_l\|_{\mathrm{H}} \|z_l\|_{\mathrm{H}} \;\forall\, x_l, y_l, z_l \in H_l. \tag{2.14}$$

Sei zu $x_l \in H_l$ ein $\delta_l \in H_l$ bestimmt durch die exakte Lösung der Aufgabe

$$(\mathcal{J}(x_l)\, \delta_l, \mathcal{J}(x_l) v_l)_{0,\Omega} \;=\; -(\mathcal{R}(x_l), \mathcal{J}(x_l) v_l)_{0,\Omega} \;\; \forall\, v_l \in H_l \tag{2.15}$$

und gelte für ein $\epsilon_l \in H$

$$\|\mathcal{J}^*(x_l)\, \mathcal{J}(x_l)\, \epsilon_l\|_{0,\Omega} \;\leq\; \eta(x_l)\, \|\mathcal{J}^*(x_l)\, \mathcal{R}(x_l)\|_{0,\Omega} \tag{2.16}$$

mit einem $\eta(x_l) < 1$.

(Beachte die Bemerkung bei Voraussetzung 1.10 bezüglich der Einschränkung der Voraussetzungen auf eine Umgebung von $x_{l,*}$.) Wegen der Vererbung der Lipschitz-Stetigkeit von $\mathcal{J}(x_l)$ auf $\mathcal{J}^*(x_l)$ gilt folgende Kompatibilitätsbedingung:

$$\|(\mathcal{J}^*(x_l) - \mathcal{J}^*(x_{l,*}))\,\mathcal{R}(x_{l,*})\|_{0,\Omega} \leq \gamma\,\|\mathcal{R}(x_{l,*})\|_{0,\Omega}\,\|x_l - x_{l,*}\|_{\mathrm{H}}\,. \tag{2.17}$$

Über Umformungen, die ähnlich zu den in Lemma 1.11 und Satz 1.12 angegebenen Umformungen ablaufen, gilt damit folgender Satz:

Satz 2.2 *Sei Voraussetzung 2.1 erfüllt, sei $x_{l,0} \in H_l$ und die Folge $\{x_{l,k}\}_{k=0}^{\infty}$ konstruiert nach der Vorschrift $x_{l,k+1} = x_{l,k} + \delta_{l,k} + \epsilon_{l,k}$ mit $\delta_{l,k}$ und $\epsilon_{l,k}$ passend zu $x_{l,k}$ aus (2.15) und (2.16) mit $\eta(x_{l,k}) \leq \eta < 1$.*
Gelte weiterhin $\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k}) \to 0$ für $k \to \infty$ und sei für alle k

$$\|\mathcal{J}^*(x_{l,k+1})\mathcal{R}(x_{l,k+1})\|_{0,\Omega} \leq \|\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})\|_{0,\Omega}.$$

Ist $x_{l,*}$ ein Häufungspunkt der Folge $\{x_{l,k}\}_{k=0}^{\infty}$ und gilt $\gamma\,\|\mathcal{R}(x_{l,*})\|_{0,\Omega} < \frac{1}{\|(\mathcal{J}^*(x_{l,*})\mathcal{J}(x_{l,*}))^{-1}\|}$, dann gilt $\mathcal{J}^*(x_{l,*})\mathcal{R}(x_{l,*}) = 0$ und $x_{l,k} \to x_{l,*}$ für $k \to \infty$.

Mit Hilfe dieses Satzes kann nun also theoretisch die Konvergenz einer durch ein inexaktes Gauß-Newton-Verfahren konstruierten Folge $\{x_{l,k}\}_{k=0}^{\infty}$ gegen eine Funktion $x_{l,*}$ nachgewiesen werden, für die die notwendige Bedingung (2.11) erfüllt ist.
Im Folgenden werden nun zwei Aspekte dieses Verfahrens weitergehend untersucht: Der Einfluss des Raums H_l (bzw. der Einfluss des Diskretisierungsparameters h_l) auf die Erfüllbarkeit der Konvergenzbedingungen und die Kontrolle des algebraischen Fehlers bei der approximativen Lösung von (2.13) entsprechend der Bedingung (2.16). Dabei wird auch kurz auf die Konvergenz des gesamten Verfahrens, dem inexakten Gauß-Newton-Verfahren, in immer feineren Räumen H_l ($l \to \infty$) gegen die Lösung x_* eingegangen.
Zur Übersicht: Das gesamte DLL-Verfahren ist in drei Stufen geschachtelt:

- Folge von Räumen H_l, $l \sim$ Verfeinerungsstufe
  - Folge von Gauß-Newton-Schritten, $k \sim$ wievielter Gauß-Newton-Schritt
    - Folge von linearen Multilevelschritten zur approximativen Lösung der Gauß-Newton-Probleme

2.2.1 Der Diskretisierungsfehler beim DLL-Verfahren

Die Größe des Diskretisierungsfehlers, bzw. die Wahl von H_l, wirkt sich an drei Stellen des Verfahrens aus: Konvergenz der Folge $\{x_{l,k}\}_{k=0}^{\infty}$ gegen $x_{l,*}$, Konvergenz der Folge $\{x_{l,*}\}_{l=0}^{\infty}$ gegen x_*, und Abbruchbedingung an die nichtlineare Iteration (Anzahl der Gauß-Newton-Schritte).

Schon im exakten Gauß-Newton-Verfahren für nicht kompatible Probleme kann die Konvergenz des Verfahrens nur unter gewissen Mindestbedingungen nachgewiesen werden. Diese können hier erfüllt werden, wenn das erreichbare Residuum $\mathcal{F}(x_{l,*}) = \|\mathcal{R}(x_{l,*})\|_{0,\Omega}^2$ klein genug ist. Das Funktional $\mathcal{F}$ ist monoton fallend bezüglich der Dimension des Unterraums H_l in dem Sinne, dass

$$\min_{x_{l+1} \in H_{l+1}} \mathcal{F}(x_{l+1}) \leq \min_{x_l \in H_l} \mathcal{F}(x_l)$$

gilt, solange $H_l \subseteq H_{l+1}$. Da das Minimum von $\mathcal{F}$ über H gleich Null ist, folgt daraus, dass es zu jedem $\epsilon > 0$ einen genügend großen Raum $H_{l,\epsilon}$ gibt, so dass $\min_{x_l \in H_{l,\epsilon}} \mathcal{F}(x_l) \leq \epsilon$ ist. Damit

ist die Bedingung an die Maximalgröße des Residuums $\mathcal{F}(x_{l,*})$ durch eine genügend große und geschickte Wahl von H_l prinzipiell erfüllbar. Konstruiert man dann nach der Vorschrift in Satz 2.2 die Gauß-Newton-Folge $\{x_{l,k}\}_{k=0}^{\infty}$ und überprüft, dass $\|\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})\|_{0,\Omega} \to 0$ monoton gilt, so konvergiert zumindest eine Teilfolge von $\{x_{l,k}\}$ gegen ein $\hat{x}_{l,*}$ mit $\mathcal{J}^*(\hat{x}_{l,*})\mathcal{R}(\hat{x}_{l,*}) = 0$. Die Kontrolle des Fehlers $\epsilon_{l,k}$ in jedem Gauß-Newton-Schritt entspricht dabei der Kontrolle des algebraischen Fehlers bei der Lösung des linearen Systems in jedem Schritt (siehe nächster Abschnitt).

Neben dieser Bedingung an die Wahl von H_l steht die obige Bedingung nach einem genügend kleinen Abstand $d_l = \operatorname{dist}(x_*, H_l)$ und $\|x_{l,0} - x_*\| \leq \rho$, damit die ermittelte Näherung $\hat{x}_{l,*}$ auch tatsächlich $\min_{x_l \in H_l} \mathcal{F}(x_l) = \mathcal{F}(\hat{x}_{l,*})$ erfüllt und damit $\hat{x}_{l,*} = x_{l,*}$ gilt. ρ ist dabei wie in Abschnitt 2.1 die Größe der Umgebung um x_*, in der $\mathcal{F}$ gleichmäßig elliptisch ist.

Damit ist der Diskretisierungsfehler bzw. die Wahl des Raums H_l entscheidend sowohl für die Konvergenz des inexakten Gauß-Newton-Verfahrens gegen eine Näherung $\hat{x}_{l,*}$ als auch - zusammen mit einer genügend guten Wahl der Startnäherung $x_{l,0}$ - für die Folgerung $\hat{x}_{l,*} = x_{l,*}$. Zur Überprüfung dieser Bedingungen ist nach Satz 2.2 zum einen die Kenntnis von $\|\mathcal{R}(x_{l,*})\|_{0,\Omega}$ und der Norm $\|(\mathcal{J}^*(x_{l,*})\mathcal{J}(x_{l,*}))^{-1}\|$ notwendig. Der Wert von $\|\mathcal{R}(x_{l,*})\|_{0,\Omega}$ kann wegen

$$\begin{aligned}
\|\mathcal{R}(x_{l,*})\|_{0,\Omega} &= \|\mathcal{R}(x_{l,*}) - \underbrace{\mathcal{R}(x_*)}_{=0}\|_{0,\Omega} \\
&= \|\mathcal{J}(x_*)(x_* - x_{l,*})\|_{0,\Omega} + O(\|x_* - x_{l,*}\|_{\mathrm{H}}^2) \\
&\leq \overline{\alpha}\|x_* - x_{l,*}\|_{\mathrm{H}} + O(\|x_* - x_{l,*}\|_{\mathrm{H}}^2)
\end{aligned}$$

durch Standard-FE-Ungleichungen (Céa-Lemma) abgeschätzt werden, während nach (2.4)

$$\|(\mathcal{J}^*(x_{l,*})\mathcal{J}(x_{l,*}))^{-1}\| \leq \frac{1}{\underline{\alpha}^2}$$

gilt.
Alles weitere ist im wesentlichen auf eine genügend gute Wahl von $x_{l,0}$ reduzierbar. Gilt $\|x_* - x_{l,0}\|_{\mathrm{H}} \leq \rho$, so folgt $d_l \leq \rho$. Wurde nun in einem Raum H_{l-1} so genau gelöst (zum Beispiel durch exakte Lösung der Gleichungssysteme), dass die Näherung $\tilde{x}_{l-1}$ an $x_{l-1,*}$ in der ρ-Umgebung von x_* liegt, so kann zur Lösung in $H_l \supset H_{l-1}$ die Startnäherung $x_{l,0} = \tilde{x}_{l-1}$ verwendet werden. Dies geschieht automatisch, wenn das Multilevelverfahren in Form eines Full-Multigrid-Zyklus implementiert wird. Dadruch überträgt sich die Qualitätsbedingung an alle Räume auf eine einzelne Bedingung an den gröbsten Raum H_0. Dort kann diese allerdings auch im Allgemeinen nicht praktisch nachgeprüft werden, da die Lösung x_* nicht bekannt ist. Hier ist eine sorgfältige Beobachtung des Konvergenzverlaufs bzw. der berechneten Folge $\{\mathcal{F}(\hat{x}_{l,*})\}_{l=0}^{\infty}$ notwendig.

Nun werde eine Hierarchie von Räumen $H_0 \subseteq H_1 \subseteq \ldots \subseteq H_l \subseteq \ldots$ betrachtet, die den eben angegebenen Bedingungen genügt. Dann konvergiert die Folge der $\{x_{l,*}\}_{l=0}^{\infty}$ für $d_l \to 0$ gegen x_*:
Es gilt (analog zu oben)

$$\begin{aligned}
\|\mathcal{R}(z_l)\|_{0,\Omega} &\leq \overline{\alpha}\|x_* - z_l\|_{\mathrm{H}} + O(\|x_* - z_l\|_{\mathrm{H}}^2), \\
\|\mathcal{R}(z_l)\|_{0,\Omega} &\geq \underline{\alpha}\|x_* - z_l\|_{\mathrm{H}} + O(\|x_* - z_l\|_{\mathrm{H}}^2)
\end{aligned}$$

und wegen

$$\begin{aligned}
d_l &= \min_{z_l \in H_l} \|x_* - z_l\|_{\mathrm{H}} \\
&\geq c_1 \min_{z_l \in H_l} \|\mathcal{R}(z_l)\|_{0,\Omega} \\
&= c_1 \|\mathcal{R}(x_{l,*})\|_{0,\Omega} \\
&\geq c_2 \|x_* - x_{l,*}\|_{\mathrm{H}}
\end{aligned}$$

folgt aus $d_l \to 0$ auch $x_{l,*} \to x_*$.

Dieses Ergebnis gilt jedoch nur für die exakte Lösung $x_{l,*}$ in jedem Raum H_l, die ja im Allgemeinen durch das nichtlineare Lösungsverfahren nicht erreicht wird. Als drittes wirkt sich der Diskretisierungsfehler also als Forderung einer Mindestreduktion in $\mathcal{F}$ aus, wenn das Gauß-Newton-Verfahren abgebrochen wird:
Sei $M > 1$ beliebig, aber fest vorgegeben. Das inexakte Gauß-Newton-Verfahren darf dann abgebrochen werden, wenn für die erhaltene Näherung $\tilde{x}_{l,*}$ an die tatsächliche Lösung $x_{l,*}$ beim Abbruch

$$\|\mathcal{R}(\tilde{x}_{l,*})\|_{0,\Omega} \;\leq\; M\,\|\mathcal{R}(x_{l,*})\|_{0,\Omega}$$

gilt. Wie man sofort sieht, folgt dann wie oben für $d_l \to 0$ auch $\tilde{x}_{l,*} \to x_*$.

Die meisten der Bedingungen an den Diskretisierungsfehler für die Konvergenz des DLL-Verfahrens sind a-priori nicht überprüfbar. Andererseits wird durch diese Bedingungen deutlich, dass der niedrigste Raum $H - 0$ auf keinen Fall zu grob gewählt werden darf. Konvergiert das Verfahren in der Praxis nicht, so kann dies also an einer zu groben Wahl des untersten Raums H_0 liegen., der dann feiner angesetzt werden sollte.

2.2.2 Die Kontrolle der nichtlinearen Iteration beim DLL-Verfahren

Im Hinblick auf die Implementierung des Verfahrens müssen die mehr theoretischen Vorgaben aus Voraussetzung 2.1, Satz 2.2 und dem vorigen Abschnitt in berechenbare Ausdrücke übersetzt werden. Insbesondere stellt sich dabei die Frage nach den Abbruchbedingungen für die beiden inneren Iterationen, d.h. die Anzahl der linearen Multilevelschritte und die Anzahl der inexakten Gauß-Newton-Schritte, da die maximale Anzahl der Verfeinerungsstufen l durch Kapazitätsbegrenzungen im Rechner ohnehin vorgegeben ist.

Zunächst ist die Bedingung für den Fehler nach (2.16) zu betrachten. Von einem festen $x_{l,k} \in H_l$ ausgehend (also für festen Newton-Schritt k und festes Level l) sei $\epsilon_{l,k} \in H_l$ die Differenz zwischen tatsächlicher Lösung des endlichdimensionalen Systems $\delta_{l,k} \in H_l$ (nach (2.15)) und der Lösung $\tilde{\delta}_{l,k} \in H_l$ nach einer bestimmten Anzahl von Multileveliterationen. Für die entsprechenden Vektor-Darstellungen in Bezug auf eine Basis von H_l gilt dann mit $n_l = \dim(H_l)$

$$\begin{aligned}
(2.15) \;&\simeq\; A_{l,k}\, y_{l,k} \;=\; b_{l,k},\\
y_{l,k} \in \mathbb{R}^{n_l} \;&\simeq\; \delta_{l,k} \in H_l,\\
\tilde{y}_{l,k} \in \mathbb{R}^{n_l} \;&\simeq\; \tilde{\delta}_{l,k} \in H_l,\\
\tilde{y}_{l,k} - y_{l,k} \;&\simeq\; \epsilon_{l,k} \in H_l\,.
\end{aligned}$$

Es ergibt sich sofort durch elementare Umformungen die Beziehung

$$\begin{aligned}
\|\mathcal{J}^*(x^*_{l,k})\mathcal{J}(x^*_{l,k})\,\epsilon_{l,k}\|^2_{0,\Omega} \;&=\; (\tilde{y}_{l,k} - y_{l,k})^T\, A^T_{l,k}\, A_{l,k}\,(\tilde{y}_{l,k} - y_{l,k})\\
&=\; r^T_{l,k}\, r_{l,k} = \|r_{l,k}\|^2_2
\end{aligned}$$

mit $r_{l,k} = A_{l,k}\tilde{y}_{l,k} - b_{l,k}$ als Residuum der zu lösenden Gleichung $A_{l,k}\, y_{l,k} = b_{l,k}$. Damit ist die linke Seite der zu beachtenden Genauigkeitsbedingung (2.16) exakt berechenbar. Offensichtlich kann $\|r_{l,k}\|_2$ beliebig klein erreicht werden, wenn ein lineares Multilevelverfahren zur Lösung des LGS angewendet wird.
Die rechte Seite ist nach unten abzuschätzen, da die Norm eines Operators im Dualraum von H nur unter hohem Aufwand genau berechnet werden kann. Die Abschätzung (2.16) muss jedoch in jedem Gauß-Newton-Schritt durchgeführt werden, also ist hier ein einfach zu berechnender

Ausdruck zu verwenden. Die Norm des Operators $\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})$ lässt sich durch die Bildung eines Supremums berechnen und der Wert des Operators für eine Richtung $v \in H_l$ entspricht dem Wert der Richtungsableitung von $\mathcal{F}$ in dieser Richtung v. Das Supremum über alle diese Werte von Richtungsableitungen wird durch die Ableitung in Richtung des (negativen) Gradienten erreicht. Das Gauß-Newton-Verfahren konstruiert näherungsweise eine Abstiegsrichtung $\delta_{l,k}$ bzw. $y_{l,k}$. Insgesamt ist daher eine Ableitung in Richtung dieses Abstiegs eine vernünftige Abschätzung für die rechte Seite von (2.16):

$$\begin{aligned} \|\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})\|_{0,\Omega} &= \sup_{v\in H} \frac{(\mathcal{R}(x_{l,k}), \mathcal{J}(x_{l,k})\, v)_{0,\Omega}}{\|v\|_{\mathrm{H}}} \\ &\geq \frac{|(\mathcal{R}(x_{l,k}), \mathcal{J}(x_{l,k})\, \tilde{\delta}_{l,k}))_{0,\Omega}|}{\|\tilde{\delta}_{l,k}\|_{\mathrm{H}}} \\ &= \frac{|b_{l,k}^T\, \tilde{y}_{l,k}|}{(\tilde{y}_{l,k}^T\, M_l\, \tilde{y}_{l,k})^{1/2}}\,. \end{aligned}$$

Hierbei ist M_l die Massenmatrix für die Darstellung der Norm $\|\cdot\|_{\mathrm{H}}$ in H_l. Diese Abschätzung für die rechte Seite bleibt bei Konvergenz von $\tilde{y}_{l,k}$ gegen $y_{l,k}$, von der Null weg beschränkt. Es ergibt sich insgesamt als Abschätzung zum Abbruch des linearen Multilevelverfahrens die Vorschrift

$$\begin{aligned} &\text{Breche mit } \tilde{y}_{l,k} \text{ als Näherung an } A_{l,k}^{-1}\, b_{l,k} \text{ ab, sobald} \\ &\|r_{l,k}\|_2 = \|A_{l,k}\tilde{y}_{l,k} - b_{l,k}\|_2 \;\leq\; \frac{|b_{l,k}^T\, \tilde{y}_{l,k}|}{(\tilde{y}_{l,k}^T\, M_l\, \tilde{y}_{l,k})^{1/2}}. \end{aligned} \tag{2.18}$$

Damit ist eine implementierbare Vorschrift für den Abbruch der innersten Iteration (Lösung des LGS nach (2.15) mit linearem Multilevelverfahren) angegeben.

In Satz 2.2 ist als notwendige Bedingung für die Konvergenz des Gauß-Newton-Verfahrens die Bedingung $\|\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})\|_{0,\Omega} \to 0\,(k \to \infty)$ vorgegeben. Diese Bedingung betrifft damit die mittlere Iteration, die Lösung des nichtlinearen Minimierungsproblems mittels einer Folge von Gauß-Newton-Schritten. Diese Bedingung an die erste Ableitung des Funktionals $\mathcal{F}$ steht im Zusammenhang mit der notwendigen Bedingung für ein Minimum von $\mathcal{F}$ in H_l nach (2.11). Die Bedingung zweiter Ordnung für ein solches Minimum ist erfüllt, wenn, wie im vorigen Abschnitt 2.2.1 behandelt, die Startnäherung bereits im gleichmäßig elliptischen Bereich um x_* gewählt wird (was wiederum vom Diskretisierungsfehler abhängt). Als Abbruchbedingung der mittleren Iterationsvorschrift kann dann eine "genügend starke" Reduktion von $\|\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})\|_{0,\Omega}$ verwendet werden, d.h.

$$\|\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})\|_{0,\Omega} \leq \rho^{(l)} \tag{2.19}$$

für ein vorgegebenes $\rho^{(l)} > 0$ mit $\rho^{(l+1)} \leq \rho^{(l)}$.
Für die Approximation von $\|\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})\|_{0,\Omega}$ kann zum einen die eben angewendete Approximation des Wertes der Richtungsableitung von $\mathcal{F}$ an der Stelle $x_{l,k}$ in Richtung $\tilde{\delta}_{l,k} = x_{l,k+1} - x_{l,k}$ verwendet werden. Da jedoch letzten Endes das Minimum des Funktionals $\mathcal{F}$ gefunden werden

soll, kann auch die folgende Approximation sinnvoll sein:

$$\begin{aligned}
\|\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})\|_{0,\Omega} &= \sup_{v\in H} \frac{|(\mathcal{R}(x_{l,k}), \mathcal{J}(x_{l,k})\, v)_{0,\Omega}|}{\|v\|_{\mathrm{H}}} \\
&= \sup_{v\in H} \lim_{t\to 0} \frac{1}{2} \frac{|\|\mathcal{R}(x_{l,k}+tv)\|_{0,\Omega}^2 - \|\mathcal{R}(x_{l,k})\|_{0,\Omega}^2|}{t\|v\|_{\mathrm{H}}} \\
&\overset{v=\tilde{\delta}_{l,k}}{\approx} \lim_{t\to 0} \frac{1}{2} \frac{\|\mathcal{R}(x_{l,k})\|_{0,\Omega}^2 - \|\mathcal{R}(x_{l,k}+t\tilde{\delta}_{l,k})\|_{0,\Omega}^2}{t\|\tilde{\delta}_{l,k}\|_{\mathrm{H}}} \\
&\overset{t=1}{\approx} \frac{1}{2} \frac{\|\mathcal{R}(x_{l,k})\|_{0,\Omega}^2 - \|\mathcal{R}(x_{l,k+1})\|_{0,\Omega}^2}{\|\tilde{\delta}_{l,k}\|_{\mathrm{H}}} \\
&= \frac{\|\mathcal{R}(x_{l,k})\|_{0,\Omega}^2 - \|\mathcal{R}(x_{l,k+1})\|_{0,\Omega}^2}{2\,(\tilde{y}_{l,k}^T M_l\, \tilde{y}_{l,k})^{1/2}}\,.
\end{aligned} \tag{2.20}$$

Eingesetzt in (2.19) wird dann also die nichtlineare Iteration abgebrochen (und zum nächsten Raum H_{l+1} übergegangen), sobald die Reduktion des Funktionals eine gewisse Mindestreduktion $2\rho^{(l)}(\tilde{y}_{l,k}^T M_l\, \tilde{y}_{l,k})^{1/2}$ unterschreitet. Ein Abbruchkriterium dieser Art findet man auch z.B. in [35].

2.2.3 Der DLL-Algorithmus

Zur besseren Übersicht werden hier die Abbruchbedingungen und Voraussetzungen der vorigen Teile dieses Abschnitts noch einmal im resultierenden Algorithmus kurz zusammengefasst angegeben. Dabei gibt das Zeichen $\simeq$ die Beziehung zwischen dem Vektor der Koeffizienten der Basisdarstellung eines Elements $v_j \in H_j$ und dem Element selbst an:

$$\begin{aligned} y_j &\simeq R_j v_j, \\ (R_j)^{-1} y_j &\simeq v_j. \end{aligned}$$

Dabei ist

$$R_j : H_j \to \mathbb{R}^{N_j},$$

wenn N_j =dim(H_j) gilt.

1. Gegeben: Hierarchie von Räumen $H_0 \subseteq H_1 \subseteq \ldots \subseteq H_l$ und Parametern $\rho^{(l)} > 0$.
2. Sei $x_{l,0} \in H_l$ Startnäherung. Setze $k = 0$. Berechne M_l Massenmatrix über H_l.
3. Stelle das Gleichungssystem
$$A_{l,k}\, y_{l,k} \;=\; b_{l,k} \tag{2.21}$$
entsprechend der Variationsformulierung
$$\begin{aligned} &\text{Berechne } \delta_{l,k} \in H_l \text{ mit} \\ &(\mathcal{J}(x_{l,k})\, \delta_{l,k}, \mathcal{J}(x_{l,k})\, v_l)_{0,\Omega} \;=\; -(\mathcal{R}(x_{l,k}), \mathcal{J}(x_{l,k})\, v_l)_{0,\Omega} \\ &\text{für alle } v_l \in H_l \end{aligned} \tag{2.22}$$
auf.
4. Wähle $\tilde{y}_{l,k}^{(0)} \in \mathbb{R}^{\dim(A_{l,k})=n_l}$ als Startnäherung für die Lösung von (2.21), $j = 0$.
   (a) Wende einen Multilevelschritt auf (2.21) mit Startnäherung $\tilde{y}_{l,k}^{(j)}$ an $\Rightarrow$ $\tilde{y}_{l,k}^{(j+1)}$ als Näherung an $y_{l,k}$ in (2.21). Setze $j = j + 1$.
   (b) Gilt
   $$\|A_{l,k}\tilde{y}_{l,k}^{(j)} - b_{l,k}\|_2 \;>\; \frac{|b_{l,k}^T\, \tilde{y}_{l,k}^{(j)}|}{((\tilde{y}_{l,k}^{(j)})^T\, M_l\, \tilde{y}_{l,k}^{(j)})^{1/2}}$$
   dann gehe zu (a).
5. Setze $x_{l,k+1} \;=\; x_{l,k} + \tilde{\delta}_{l,k}^{(j)}$ mit $\tilde{\delta}_{l,k}^{(j)} \in H_l \simeq \tilde{y}_{l,k}^{(j)} \in \mathbb{R}^{n_l}$.
6. Gilt
   $$\|\mathcal{R}(x_{l,k+1})\|_{0,\Omega}^2 - \|\mathcal{R}(x_{l,k})\|_{0,\Omega}^2 \;>\; 2\rho^{(l)}((\tilde{y}_{l,k}^{(j)})^T\, M_l\, \tilde{y}_{l,k}^{(j)})^{1/2},$$
   dann setze $k = k + 1$ und gehe zu 3.
7. Wähle $H_{l+1} \supset H_l$. Setze $x_{l+1,0} = x_{l,k}$. Gehe zu 2. (oder Abbruch, wenn maximale Verfeinerungsstufe erreicht)

Algorithmus 2.2 : DLL-Verfahren

2.3 Ein nichtlineares Multilevelverfahren (DNL)

Im vorigen Abschnitt wurde das nichtlineare Problem (2.10) durch Anwendung des Gauß-Newton-Verfahrens über dem Raum H_l linearisiert und das entstandene lineare Problem mit Hilfe eines linearen Multilevelverfahrens im Algorithmus 2.2 gelöst. In diesem Abschnitt wird eine alternative Herangehensweise angegeben, die das nichtlineare Problem (2.11) mit Hilfe eines nichtlinearen Multilevelverfahrens löst und die Teilprobleme erst auf den einzelnen Unterräumen linearisiert.

In Abschnitt 1.4.3 wurde das Grundprinzip von nichtlinearen Multilevelverfahren (FAS) angegeben. Dies soll jetzt auf das Problem (2.11) bzw. die Aufgabe

$$\begin{aligned} &\text{Finde } x_{l,*} \text{ in } H_l \text{ mit} \\ &(\mathcal{R}(x_{l,*}), \mathcal{J}(x_{l,*})\, v_l)_{0,\Omega} \;=\; 0 \\ &\text{für alle } v_l \in H_l \end{aligned} \tag{2.23}$$

angewendet werden. Gleichung (2.23) ist eine nichtlineare Gleichung im Dualraum von H_l. Setzt man $A_l(x_l)$ für den Operator $\mathcal{J}^*(x_l)\mathcal{R}(x_l)|_{H_l'}$, so lässt sich diese Gleichung auch vereinfacht schreiben als

$$\begin{aligned} &\text{Finde } x_{l,*} \text{ in } H_l \text{ mit} \\ &A_l(x_{l,*})[v_l] \;=\; 0 \\ &\text{für alle } v_l \in H_l. \end{aligned} \tag{2.24}$$

Für die Lösung von (2.24) über ein nichtlineares Multilevelverfahren muss nun für eine gegebene Näherung x_l im Raum H_l eine nichtlineare Korrekturgleichung im Raum H_{l-1} aufgestellt werden, um x_l durch die Lösung dieses niedrigerdimensionalen Problems zu verbessern. Im folgenden Abschnitt werden zwei Möglichkeiten angegeben, diese Korrekturgleichung aufzustellen. Anschließend wird auf die Lösung der nichtlinearen Gleichungen in einem festen Raum H_l eingegangen und abschließend der sich ergebende Algorithmus angegeben.
Zur Übersicht: Das gesamte DNL-Verfahren ist in drei Stufen geschachtelt:

- Folge von Räumen H_l, $l \sim$ Verfeinerungsstufe
  - Folge von nichtlinearen Multilevelschritten, k wievielte Multileveliteration
    - iterative Lösung der nichtlinearen Probleme auf festem Level l, Iterationsindex m

2.3.1 Herleitung der nichtlinearen Korrekturgleichungen

Gegeben sei die Hierarchie von Räumen $H_0 \subseteq H_1 \subseteq \ldots \subseteq H_l$. Zwischen zwei aufeinanderfolgenden Räumen H_j und H_{j+1} existiert die kanonische Einbettungsabbildung $\mathcal{I}_j^{j+1}$, deren Matrixrepräsentation mit I_j^{j+1} bezeichnet wird:

$$I_j^{j+1} \;=\; R_{j+1} \circ \mathcal{I}_j^{j+1} \circ R_j^{-1}$$

(R_j wie oben bei Algorithmus 2.2). Der entsprechende Restriktionsoperator $\mathcal{I}_{j+1}^{j}$ wird mit Hilfe von

$$(\mathcal{I}_{j+1}^{j} u_{j+1}, v_j)_{0,\Omega} \;=\; (u_{j+1}, v_j)_{0,\Omega} \quad \forall\, v_j \in H_j$$

definiert. In Abschnitt 1.4.3 wurde angegeben, dass für nichtlineare Multilevelverfahren auch eine solche Restriktion $\tilde{\mathcal{I}}_{j+1}^{j}$ benötigt wird, so dass für $x_{j+1} \in H_{j+1}$ das Bild $\tilde{\mathcal{I}}_{j+1}^{j} x_{j+1}$ eine Repräsentation von x_{j+1} im Raum H_j darstellt. Diese Abbildung $\tilde{\mathcal{I}}_{j+1}^{j}$ hängt wesentlich von dem jeweiligen, dem Problem zu Grunde liegenden Raum H ab. Daher wird diese Abbildung im Detail erst bei der Behandlung des Beispielproblems in Kapitel 4 angegeben.

Für die Herleitung von Grobraumkorrekturgleichungen für das FAS-Schema gibt es bei Ausgleichsformulierungen wie (2.9) zwei verschiedene Ansätze. Zum einen ist es möglich, die in (2.24) definierte nichtlineare Operatorgleichung zum Ausgangspunkt des FAS-Schemas zu machen. Zum anderen ist es jedoch auch denkbar, direkt von der ursprünglichen nichtlinearen Gleichung (2.6) auszugehen. Im ersteren Fall ist es der nichtlineare Operator A_l, im zweiten der ursprüngliche nichtlineare Differentialoperator $\mathcal{R}$, auf den das FAS-Schema angewendet wird.

Zunächst soll die Aufgabe (2.24) zum Ausgangspunkt gemacht werden. Die im Abschnitt über das FAS-Schema hergeleitete Grobraumkorrekturgleichung ausgehend von einer Näherung $x_l \in H_l$ lautet dann

$$\begin{aligned} &\text{Finde } x_{l-1} \text{ in } H_{l-1} \text{ mit} \\ &A_{l-1}(\tilde{\mathcal{I}}_l^{l-1} x_l + x_{l-1})[v_{l-1}] \;=\; \mathcal{I}_l^{l-1} r_l[v_{l-1}] \;+\; A_{l-1}(\tilde{\mathcal{I}}_l^{l-1} x_l)[v_{l-1}] \\ &\text{für alle } v_{l-1} \in H_{l-1}. \end{aligned} \tag{2.25}$$

Dabei ist $r_l = b_l - A_l(x_l) = -A_l(x_l)$ das Residuum der Gleichung (2.24) im "reicheren" Raum H_l (da hier l der oberste Raum ist, gilt $b_l \equiv 0$). Es ist zu beachten, dass x_{l-1} keine Näherung an die Lösung des Problems (2.24) im Raum H_{l-1} ist, sondern eine Korrektur an die aktuelle Näherung x_l im Raum H_l.
Setzt man in (2.25) die Definition des Operators ein, erhält man als Grobraumkorrekturgleichung:

$$\begin{aligned} &\text{Finde } x_{l-1} \text{ in } H_{l-1} \text{ mit} \\ &(\mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l + x_{l-1}), \mathcal{J}(\tilde{\mathcal{I}}_l^{l-1} x_l + x_{l-1})\, v_{l-1})_{0,\Omega} \;=\; \underbrace{-(\mathcal{R}(x_l), \mathcal{J}(x_l)\, v_{l-1})_{0,\Omega}}_{=\mathcal{I}_l^{l-1}(-A_l(x_l)[v_{l-1}])} \\ &\qquad\qquad + \underbrace{(\mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l), \mathcal{J}(\tilde{\mathcal{I}}_l^{l-1} x_l)\, v_{l-1})_{0,\Omega}}_{=A_{l-1}(\tilde{\mathcal{I}}_l^{l-1} x_l)[v_{l-1}]} \\ &\text{für alle } v_{l-1} \in H_{l-1}. \end{aligned} \tag{2.26}$$

Dies ist eine nichtlineare Gleichung im Raum H_{l-1}. Hat man eine Näherung an die Lösung x_{l-1} des Problems (2.26) gefunden, so kann diese wiederum durch eine Grobraumkorrektur verbessert werden. Dadurch werden rekursive Aufrufe notwendig. Die Grobraumkorrekturgleichung für einen Raum H_j mit $j \leq l$ beliebig und einer Näherung $x_j \in H_j$ lässt sich dann allgemein schreiben als

$$\begin{aligned} &\text{Finde } x_{j-1} \text{ in } H_{j-1} \text{ mit} \\ &(\mathcal{R}(\tilde{\mathcal{I}}_j^{j-1} x_j + x_{j-1}), \mathcal{J}(\tilde{\mathcal{I}}_j^{j-1} x_j + x_{j-1})\, v_{j-1})_{0,\Omega} \;=\; f_{j-1}[v_{j-1}] \\ &\text{für alle } v_{j-1} \in H_{j-1} \end{aligned} \tag{2.27}$$

mit einem linearen Funktional f_{j-1} auf H_{j-1}.
Von einem Raum H_l ausgehend lassen sich dann die rechten Seiten der zu lösenden Systeme aus (2.27) folgendermaßen angeben:

$$\begin{aligned} &f_l[v_l] \;=\; (\mathcal{R}(x_l), \mathcal{J}(x_l)\, v_l)_{0,\Omega} \\ &\text{und rekursiv von } j = l, l-1, \ldots, 1 \\ &f_{j-1}[v_{j-1}] \;=\; (\mathcal{R}(\tilde{\mathcal{I}}_j^{j-1} x_j), \mathcal{J}(\tilde{\mathcal{I}}_j^{j-1} x_j)\, v_{j-1})_{0,\Omega} - f_j[v_{j-1}]\,. \end{aligned} \tag{2.28}$$

Dadurch wird ein FAS-Multilevelschritt für die Aufgabe (2.23) formal beschrieben. Auf die Lösung solcher nichtlinearen Gleichungen wird im Abschnitt 2.3.2 eingegangen.

Ein zweiter Weg zum Einsatz des FAS-Schemas für die zu lösende nichtlineare Differentialgleichung besteht in der Anwendung des Schemas auf die Differentialgleichung (2.6) selbst. Offensichtlich ist die entsprechende Grobraumkorrekturgleichung für den Operator $\mathcal{R}$ an der Stelle x_l im Raum H_{l-1} gegeben durch

$$\mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l + x_{l-1}) \;=\; -\mathcal{R}(x_l) + \mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l). \tag{2.29}$$

Diese Gleichung ist für das unbekannte $x_{l-1} \in H_{l-1}$ zu lösen. Dieses Problem wird näherungsweise durch eine Ausgleichsformulierung für (2.29) gelöst. Die Aufgabe lautet dann

$$\begin{aligned} &\|\mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l + x_{l-1}) - \mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l) + \mathcal{R}(x_l)\|_{0,\Omega}^2 \\ &\qquad \stackrel{!}{=} \min_{v_{l-1} \in H_{l-1}} \|\mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l + v_{l-1}) - \mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l) + \mathcal{R}(x_l)\|_{0,\Omega}^2. \end{aligned} \tag{2.30}$$

Eine notwendige Bedingung für die Lösung dieser Aufgabe ist das Finden einer Nullstelle der Ableitung dieses Funktionals in H_{l-1}. Mit Hilfe der Variationsrechnung erhält man dafür das nichtlineare Variationsproblem

$$\begin{aligned} &\text{Finde } x_{l-1} \text{ in } H_{l-1} \text{ mit} \\ &(\mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l + x_{l-1}), \mathcal{J}(\tilde{\mathcal{I}}_l^{l-1} x_l + x_{l-1})\, v_{l-1})_{0,\Omega} \;=\; -(\mathcal{R}(x_l), \mathcal{J}(\tilde{\mathcal{I}}_l^{l-1} x_l + x_{l-1})\, v_{l-1})_{0,\Omega} \\ &\qquad\qquad + (\mathcal{R}(\tilde{\mathcal{I}}_l^{l-1} x_l), \mathcal{J}(\tilde{\mathcal{I}}_l^{l-1} x_l + x_{l-1})\, v_{l-1})_{0,\Omega} \\ &\text{für alle } v_{l-1} \in H_{l-1}. \end{aligned} \tag{2.31}$$

Ein wesentlicher Unterschied zu (2.26) liegt darin, dass hier die rechte Seite von der gesuchten Lösung x_{l-1} abhängt. Dadurch lässt sich dieser Ansatz auch nicht mehr rekursiv wie in (2.27) beschreiben. Hinzu kommt, dass die rechte Seite jeder weiteren Korrekturgleichung für $j = l-2, l-3, \ldots$ im Multilevelprozess an der jeweils aktuellen Näherung im zu behandelnden Raum neu ausgewertet werden muss. Dies bedeutet einerseits einen höheren Aufwand als bei dem Ansatz (2.26). Zum anderen muss die rechte Seite also erst in den höheren Räumen ausgewertet werden und kann erst danach im aktuellen Raum aufgestellt werden.
Insgesamt bedeutet diese Formulierung also einen erheblichen Mehraufwand bei der Implementierung. Andererseits wird dadurch ein nichtlineares Problem zur Grobraumkorrektur aufgestellt, das die Nichtlinearität insofern besser widerspiegelt, als stets eine aktuelle Näherung bei der Aufstellung des Problems verwendet wird. Im vorherigen Ansatz wird demgegenüber eine Näherung des obersten Raums H_l nicht mehr verändert, bis die Multileveliteration beendet ist.

Diese beiden Ansätze (und ein weiterer, hybrider Ansatz) wurden bereits in [35] hinsichtlich Schnelligkeit und Aufwand verglichen. Ein Ergebnis aus diesen Untersuchungen zeigte deutlich, dass der erhöhte Aufwand beim hier als zweites dargestellten Ansatz (FAS für $\mathcal{R}$) nicht eine entsprechend verbesserte Konvergenzrate zur Folge hat. Damit lohnt sich dieser Mehraufwand also im Vergleich mit dem hier als erstes dargestellten Ansatz (FAS für $A_l(\cdot)[\cdot]$ aus (2.24)) nicht. Daher wird in den folgenden Abschnitten nur noch der erste Ansatz berücksichtigt.

2.3.2 Lösung der nichtlinearen Gleichungen

Über jedem Raum H_j müssen nichtlineare Gleichungen nach (2.27) und (2.28) gelöst werden. Dazu wird der Ansatz der iterativen Linearisierung verwendet. Ausgehend von der Startnäherung

$x_{j-1,0} = \tilde{\mathcal{I}}_j^{j-1} x_j$ wird die Gleichung (2.27) linearisiert: Die Aufgabe

$$\begin{aligned} &\text{Finde } x_{j-1} \text{ in } H_{j-1} \text{ mit} \\ &(\mathcal{R}(x_{j-1,0} + x_{j-1}), \mathcal{J}(x_{j-1,0} + x_{j-1})\, v_{j-1})_{0,\Omega} \;=\; f_{j-1}[v_{j-1}] \\ &\text{für alle } v_{j-1} \in H_{j-1} \end{aligned} \tag{2.32}$$

wird dadurch zu

$$\begin{aligned} &\text{Finde } \delta_{j-1,m} \text{ in } H_{j-1} \text{ mit} \\ &(\mathcal{R}(x_{j-1,m}) + \mathcal{J}(x_{j-1,m})\, \delta_{j-1,m}, \mathcal{J}(x_{j-1,m})\, v_{j-1})_{0,\Omega} \;=\; f_{j-1}[v_{j-1}] \\ &\text{für alle } v_{j-1} \in H_{j-1} \\ &\text{und setze } x_{j-1,m+1} \;=\; x_{j-1,m} \;+\; \delta_{j-1,m} \end{aligned} \tag{2.33}$$

für $m = 0, 1, 2, \ldots$
Die folgende Beobachtung stützt diesen Ansatz: Durch Einsetzen der Definition von f_l nach (2.28) sieht man durch überprüfung der entsprechenden Normalengleichungen, dass die Iteration nach (2.33) einer Folge von Gauß-Newton-Schritten für die Aufgabe

$$\begin{aligned} &\text{Suche } x_{l-1,*} \in H_{l-1} \text{ mit} \\ &\qquad \mathcal{F}(x_l + x_{l-1,*}) \stackrel{!}{=} \min_{x_{l-1} \in H_{l-1}} \mathcal{F}(x_l + x_{l-1}) \end{aligned}$$

entspricht, solange nur die Räume H_l und H_{l-1} einbezogen sind. Sobald jedoch die nichtlineare Korrekturgleichung für den Raum H_{l-1} über dem Raum H_{l-2} aufgestellt wird, gibt es natürlich kein entsprechendes Minimierungsproblem mehr, da im Raum H_{l-2} nun eine solche Korrekturgleichung an die Korrekturgleichung im Raum H_{l-1} aufgestellt wird, zu der die ursprünglich zu lösende Gleichung nur noch das Residuum auf der rechten Seite beiträgt.

Im Zusammenhang mit der Konvergenztheorie für nichtlineare Multilevelverfahren nach [27] wäre es notwendig, das Lösungsverfahren (Glättungsverfahren) für die nichtlinearen Korrekturgleichungen mit Hilfe sogenannter dividierter Differenzen darzustellen. Dies ist bei dem obigen Lösungsansatz (sukzessive Linearisierungen) nicht möglich. Um zu überprüfen, ob Konvergenz vorliegt, müssen daher andere Kontrollverfahren angewendet werden. Ein der Theorie konformer Lösungsansatz für die nichtlinearen Gleichungen wäre ein nichtlineares Jacobi- oder Gauß-Seidel-Verfahren, innerhalb dessen die Newton-Iterationen zur Lösung der eindimensionalen Probleme so exakt durchgeführt werden müssen, dass die notwendige Glättungseigenschaft ("smoothing property") nicht verlorengeht. Dabei müsste auch die Ableitung von $\mathcal{J}$ verwendet werden. Eine Implementierung dieses Ansatzes wäre also sehr aufwendig, da hier genausoviele eindimensionale nichtlineare Gleichungen mit Hilfe eines Newton-Verfahrens nacheinander gelöst werden müssen, wie die Dimension von H_j angibt.
Im Zusammenhang mit der Multilevel-Philosophie ist dabei zu bedenken, dass eine exakte Lösung der Grobraumkorrekturgleichungen wie im linearen Fall jede weitere Grobraumkorrektur überflüssig machen würde, da ja $H_j \supset H_{j-1} \supset \ldots$ gilt. Es ist also lediglich eine näherungsweise Lösung der Grobraumkorrekturgleichung notwendig. Aus diesem Grund soll der in (2.33) angegebene Ansatz für die praktische Implementierung zur Lösung von (2.27) und (2.28) verwendet werden.

Damit entschieden werden kann, bei welchem Wert von m die iterative Lösung der nichtlinearen Gleichung über dem Raum H_j abgebrochen werden soll, muss der Verlauf der Konvergenz bei Anwendung der sukzessiven Linearisierungen kontrolliert werden. Dies ist beispielsweise möglich durch eine Kontrolle des nichtlinearen Residuums der Gleichung (2.32): Abzuschätzen ist

$$R_{\mathrm{nl},m} := \sup_{v_{j-1} \in H_{j-1}} \frac{|(\mathcal{R}(x_{j-1,m}), \mathcal{J}(x_{j-1,m})\, v_{j-1})_{0,\Omega} - f_{j-1}[v_{j-1}]|}{\|v_{j-1}\|_{H_{j-1}}}. \tag{2.34}$$

Sei M_{j-1} eine Darstellung der Norm $\|\cdot\|_{H_{j-1}}$ über einer N_{j-1}-elementigen Basis des H_{j-1} (vergleiche Algorithmus 2.2, Punkt 2). Durchläuft man mit den v_{j-1} aus (2.34) diese Basis und stellt die sich ergebenden Zähler aus (2.34) in einen Vektor b, so ist das Supremum $R_{\mathrm{nl},m}$ zu berechnen über

$$R_{\mathrm{nl},m}^2 \;=\; \max_{c\in\mathbf{R}^{N_{j-1}}} \frac{(c^T b)^2}{c^T M_{j-1} c} \;=\; b^T M_{j-1}^{-1} b. \tag{2.35}$$

Dieser Ausdruck kann dann mit Hilfe der Anwendung eines linearen Multilevelschritts auf die Gleichung

$$M_{j-1}\, z \;=\; b$$

(M_{j-1} ist offensichtlich positiv definit und symmetrisch) durch

$$R_{\mathrm{nl},m}^2 \;\approx\; b^T z$$

angenähert werden.
Da eine näherungsweise Lösung der Gleichung (2.32) ausreicht, kann der Iterationsprozess der sukzessiven Linearisierung zum Beispiel dann abgebrochen werden, wenn für ein ρ_1 mit $0 < \rho_1 < 1$

$$R_{\mathrm{nl},m} \;\le\; \rho_1\, R_{\mathrm{nl},0} \tag{2.36}$$

gilt.
Nachdem auf diese Weise ein sinnvolles Abbruchkriterium für die nichtlineare Iteration innerhalb eines festen Raums H_{j-1} vorhanden ist, stellt sich nun die Frage, wie genau die linearen Gleichungssysteme gemäß (2.33) zu lösen sind. Neben der Möglichkeit, das lineare Residuum stets bis auf ein gewisses $0 < \rho_2 < 1$ genau zu lösen, besteht auch die Möglichkeit, in jedem Schritt der sukzessiven Linearisierung sicherzustellen, dass das nichtlineare Residuum verringert wird:

$$R_{\mathrm{nl},m+1} \;<\; \rho_3\, R_{\mathrm{nl},m},\quad 0 < \rho_3 < 1.$$

Da jedoch für ein nichtlineares Problem normalerweise keine monotone Konvergenz vorliegt, ist dieses Kriterium im Allgemeinen nicht praktikabel. Daher soll hier lediglich eine gewisse Mindestgenauigkeit für das Residuum der linearen Systeme aus (2.33) durch ein $0 < \rho_2 < 1$ vorgeschrieben werden.

Damit ist ein nichtlinearer FAS-Multilevelschritt auch in Hinsicht auf Abbruchbedingungen für die iterativen Lösungsansätze beschrieben worden. Sei $\{x_{l,k}\}_{k=0,1,2,\ldots} \subset H_l$ die durch wiederholte Anwendung des FAS-Verfahrens erhaltene Folge von Iterierten. Damit bleibt noch die Frage, wieviele nichtlineare Multilevelschritte verwendet werden sollen, um die Gleichung (2.23) hinreichend genau zu lösen. Die Betrachtungen am Schluss des Abschnitts 2.2.2 treffen hier im Hinblick auf die Bedingung $\|\mathcal{J}^*(x_{l,k})\mathcal{R}(x_{l,k})\|_{0,\Omega} \to 0\,(k\to\infty)$ ebenso zu. Es kann daher dafür die in Algorithmus 2.2, Punkt 5 und 6 angegebene Abbruchbedingung verwendet werden.

2.3.3 Der DNL-Algorithmus

Wie in Abschnitt 2.2.3 werden hier die Ergebnisse der vorigen Teile angegeben. Dabei entspricht das Zeichen $\simeq$ wie in Abschnitt 2.2.3 die Beziehung zwischen dem Vektor der Koeffizienten der Basisdarstellung eines Elements $v_j \in H_j$ und dem Element selbst.

1. Gegeben: Hierarchie von Räumen $H_0 \subseteq H_1 \subseteq \ldots \subseteq H_l$, $\rho^{(l)}, \rho_1, \rho_2 > 0$.

2. Wähle Startnäherung $x_{l,0} \in H_l$, setze $k = 0, m = 0$ und berechne $M_j, j = 0, \ldots, l$ Massenmatrizen über H_j.

3. Vorglättung: Löse die nichtlineare Aufgabe

$$\begin{aligned} &\text{Finde } \delta_{l,k} \in H_l \text{ mit:} \\ &(\mathcal{R}(x_{l,k} + \delta_{l,k}), \mathcal{J}(x_{l,k} + \delta_{l,k})\, v_l)_{0,\Omega} \;=\; 0 \\ &\text{für alle } v_l \;\in H_l \end{aligned}$$

   näherungsweise: Setze $x_{l,k,m} = x_{l,k}$, berechne $R_{\mathrm{nl},0}$ nach (2.35).

   (a) Stelle das lineare Gleichungssystem $A_{l,k,m}\, y_{l,k,m} \;=\; b_{l,k,m}$ analog zur linearen Aufgabe

$$\begin{aligned} &\text{Finde } \delta_{l,k,m} \in H_l \text{ mit:} \\ &(\mathcal{R}(x_{l,k,m}) + \mathcal{J}(x_{l,k,m})\delta_{l,k,m}), \mathcal{J}(x_{l,k,m})\, v_l)_{0,\Omega} \;=\; 0 \\ &\text{für alle } v_l \;\in H_l \end{aligned}$$

   auf und löse näherungsweise
   $\Longrightarrow \tilde{y}_{l,k,m}$ mit $\|A_{l,k,m}\, \tilde{y}_{l,k,m} - b_{l,k,m}\|_2 \leq \rho_2$.

   (b) Setze $x_{l,k,m+1} = x_{l,k,m} + \tilde{\delta}_{l,k,m}$ mit $\tilde{\delta}_{l,k,m} \simeq \tilde{y}_{l,k,m}$ und $m = m + 1$.
   Berechne $R_{\mathrm{nl},m}$. Gilt $R_{\mathrm{nl},m} > \rho_1 R_{\mathrm{nl},0}$ dann gehe zu (a).

4. Setze $x_{l,k}^{(1)} = x_{l,k,m}$.

5. FAS:
   Wende einen nichtlinearen Multilevelschritt an. Löse die dabei entstehenden nichtlinearen Korrekturgleichungen wie in 3.(a) und (b) $\Rightarrow \tilde{\delta}_{l,k}$ als nichtlineare Multilevelkorrektur.

6. Setze $x_{l,k}^{(2)} = x_{l,k}^{(1)} + \tilde{\delta}_{l,k}$.

7. Nachglättung:
   Berechne wie oben bei Punkt 3 Näherung an Lösung von

$$\begin{aligned} &\text{Finde } \delta_{l,k} \in H_l \text{ mit:} \\ &(\mathcal{R}(x_{l,k}^{(2)} + \delta_{l,k}), \mathcal{J}(x_{l,k}^{(2)} + \delta_{l,k})\, v_l)_{0,\Omega} \;=\; 0 \\ &\text{für alle } v_l \;\in H_l. \end{aligned}$$

   mit Hilfe von sukzessiven Linearisierungen 3. (a) und (b) $\Rightarrow \tilde{\delta}_{l,k}$ Näherung an $\delta_{l,k}$
   $\Rightarrow$ Neue Näherung $x_{l,k}^{(3)} = x_{l,k}^{(2)} + \tilde{\delta}_{l,k}$.

8. Gilt mit $\tilde{y}_{l,k} \simeq x_{l,k}^{(3)} - x_{l,k}$

$$\|\mathcal{R}(x_{l,k})\|_{0,\Omega} - \|\mathcal{R}(x_{l,k}^{(3)})\|_{0,\Omega} \;>\; 2\rho^{(l)}((\tilde{y}_{l,k})^T\, M_l\, \tilde{y}_{l,k})^{1/2}\,,$$

   dann setze $x_{l,k+1} = x_{l,k}^{(3)}$, $k = k + 1$ und gehe zu 3.

9. Wähle $H_{l+1} \supset H_l$. Setze $x_{l+1,0} = x_{l,k}^{(3)}$. Gehe zu 2.

Algorithmus 2.3 : DNL-Verfahren

Kapitel 3

Gauß-Newton Multilevelverfahren (LDL)

Die Lösungsverfahren aus Kapitel 2 nehmen zuerst eine Diskretisierung des Problems (2.9) vor und bearbeiten dann das neue, diskretisierte Problem (2.10) (**D**···-Verfahren). Dadurch wird jedoch das ursprünglich kompatible Problem inkompatibel, da das neue Minimum des Funktionals $\mathcal{F}$ im endlichdimensionalen Teilraum im Allgemeinen nicht Null ist. Für inkompatible Probleme ist allerdings die Konvergenztheorie des Gauß-Newton-Verfahrens wesentlich eingeschränkter als für kompatible Probleme (siehe Kapitel 1). Weiterhin ist die Konvergenzgeschwindigkeit auch bei exakter Lösung der Gleichungssysteme maximal linear, während für kompatible Probleme theoretisch quadratische Konvergenzgeschwindigkeit erreichbar ist.
Daher liegt es nahe, das Problem (2.9) zunächst nicht zu diskretisieren, sondern als kompatibles Minimierungsproblem im unendlichdimensionalen Hilbertraum H aufzufassen und so ein Gauß-Newton-Verfahren in H und nicht in einem endlichdimensionalen Raum H_l zu verwenden. Der Diskretisierungsfehler, der zwangsläufig durch die Implementierung des Verfahrens auf einem Rechner entsteht, wird dann als Exaktheitsfehler in einem inexakten Gauß-Newton-Verfahren Dadurch kann der Diskretisierungsfehler genau wie der durch die iterative Lösung bedingte algebraische Fehler behandelt werden. Mit Hilfe der Genauigkeitsschranken aus Abschnitt 1.1.5 ist der Diskretisierungsfehler in dem Sinne kontrollierbar, dass auch für die Diskretisierung Schranken so vorgegeben werden können, dass das inexakte Gauß-Newton-Verfahren konvergiert. Im Gegensatz dazu musste im vorherigen Kapitel der Diskretisierungsfehler als gegeben hingenommen werden, so dass die gegebenen Genauigkeitsschranken nur auf den algebraischen Fehler angewendet werden konnten.
Das **LDL**-Verfahren besteht also in der Aufstellung des kompatiblen Minimierungsproblems in H, Anwenden des Gauß-Newton-Verfahrens in H und Berechnen einer Näherungslösung $\tilde{\delta}_l \in H_l$ an die exakte Gauß-Newton-Korrektur $\delta \in H$ mittels eines linearen Multilevelverfahrens.
Im Folgenden bezeichne l den Index eines endlichdimensionalen Unterraums von H und k den Index des linearen Iterationsverfahrens für festes l. Das bedeutet, dass zum einen durch $l \to \infty$ der Raum H immer besser ausgeschöpft wird, als auch auf jedem Level für immer größeres k die zu dem Unterraum gehörige Korrektur immer exakter berechnet wird. Formal wird nun allerdings bei jedem Gauß-Newton-Schritt auch zum nächsten Level übergegangen (auch wenn sich der Raum gar nicht ändert, also $H_{l+1} = H_l$ gilt). Damit steht l also auch für den Laufindex des nichtlinearen Lösungsverfahrens (Gauß-Newton) und k weiterhin für den Laufindex der iterativen Lösung der dazugehörigen linearen Probleme.

3.1 Formulierung des inexakten Gauss-Newton-Verfahrens in H

Das Ziel ist die Berechnung des eindeutigen Minimums $x_* \in H$ von

$$\mathcal{F}(x_*) \;=\; \|\mathcal{R}(x_*)\|_{0,\Omega}^2 \;=\; 0 \tag{3.1}$$

mit Hilfe des Gauß-Newton-Verfahrens. Die eindeutige Lösbarkeit des Problems $\mathcal{R}(x) = 0$ wird dabei vorausgesetzt. Wie in Abschnitt 1.1.2 angegeben, wird der Operator $\mathcal{R}$ um eine aktuelle Näherung linearisiert und das entstehende lineare Ausgleichsproblem mittels der Normalengleichungen gelöst:
Sei x_l die Näherung an x_* im l-ten Iterationsschritt. Dann gilt für die Gauß-Newton-Korrektur δ_l

$$\|\mathcal{R}(x_l+\delta_l)\|_{0,\Omega}^2 \;\approx\; \|\mathcal{R}(x_l)+\mathcal{J}(x_l)\,\delta_l\|_{0,\Omega}^2 \;\stackrel{!}{=}\; \min_{\delta\in H}\|\mathcal{R}(x_l)+\mathcal{J}(x_l)\,\delta\|_{0,\Omega}^2$$

und diese Korrektur δ_l wird durch Lösung der folgenden Gleichung bestimmt:

$$(\mathcal{J}(x_l)\,\delta_l, \mathcal{J}(x_l)v)_{0,\Omega} \;=\; -(\mathcal{R}(x_l), \mathcal{J}(x_l)v)_{0,\Omega} \;\; \forall\, v \,\in H\,. \tag{3.2}$$

Bei geschickter Wahl der Startnäherung x_0 ist dieses Verfahren auch in H (lokal) quadratisch konvergent, denn Satz 1.1 lässt sich sinngemäß auf einen Operator $\mathcal{R} : H \to (L^2(\Omega))^2$ und die jeweiligen Normen übertragen:

Satz 3.1 *Sei $\mathcal{R} : H \to (L^2(\Omega))^2$ zweimal stetig differenzierbar und die Fréchet-Ableitung $\mathcal{J}(x)$ wie in Abschnitt 2.1. Sei zusätzlich $\mathcal{J}(x)$ Lipschitz-stetig in H mit Konstante γ*

$$\|\mathcal{J}(x)\,z \;-\; \mathcal{J}(y)\,z\|_{0,\Omega} \;\le\; \gamma\,\|x-y\|_{\mathrm{H}}\|z\|_{\mathrm{H}} \;\forall\, x,y,z \,\in H.$$

Dann existiert ein $\epsilon > 0$, so dass für eine Startnäherung $x_0 \in H$ mit $\|x_* - x_0\|_{\mathrm{H}} < \epsilon$ die Folge der x_l, deren Korrekturen durch das Gauß-Newton-Verfahren nach (3.2) definiert werden, quadratisch gegen x_* konvergiert.

Die Gleichungen (3.2) können allerdings nicht exakt gelöst werden, sondern werden durch Lösungen in einem Raum $H_l \subseteq H$ nur näherungsweise gelöst. Der dabei entstehende Fehler kann wie folgt beschrieben werden: Sei $x_{l,k} \in H_l$ eine durch k Teilraumiterationen entstandene Näherung an x_*. Mit $x_k = x_{l,k}$ sei $\delta_l \in H$ die exakte Lösung der Gleichung (3.2). Weiterhin bezeichne $\delta_{l,k}$ die exakte Lösung der endlichdimensionalen Aufgabe

$$(\mathcal{J}(x_{l,k})\,\delta_{l,k}, \mathcal{J}(x_{l,k})v_l)_{0,\Omega} \;=\; -(\mathcal{R}(x_{l,k}), \mathcal{J}(x_{l,k})v_l)_{0,\Omega} \;\; \forall\, v_l \,\in H_l \tag{3.3}$$

und schließlich sei $\tilde{\delta}_{l,k} = \delta_{l,k} + \epsilon_{l,k} \in H_l$ eine Näherung an die Lösung von (3.3).
Damit ist der Fehler in der Berechnung der (kontinuierlichen) Gauß-Newton-Korrektur δ_l durch zwei Komponenten abschätzbar. Eine sinnvolle Norm zur Abschätzung des Fehlers ist dabei die Energienorm $\|\cdot\|_{\mathcal{J},y} = \|\mathcal{J}(y)\cdot\|_{0,\Omega}$ für $y \in H$. Damit wird

$$\|\delta_l - \tilde{\delta}_{l,k}\|_{\mathcal{J},x_{l,k}} \;\le\; \underbrace{\|\delta_l - \delta_{l,k}\|_{\mathcal{J},x_{l,k}}}_{\text{Diskretisierungsfehler}} \;+\; \underbrace{\|\epsilon_{l,k}\|_{\mathcal{J},x_{l,k}}}_{\text{algebraischerFehler}}\,. \tag{3.4}$$

Wegen der Voraussetzung (2.4) ist diese Norm für beliebiges $x_{l,k}$ beliebig äquivalent zur Norm des Hilbertraums H.
Nach den Ergebnissen von Kapitel 1 verbleibt nun noch, diesen Fehler so klein zu halten, dass die Konvergenz des Gauß-Newton-Verfahrens in H (!) gewährleistet bleibt.

In Teil 1.1 werden solche fehlerkontrollierenden Genauigkeitsbedingungen sowohl für inkompatible als auch für kompatible Probleme in endlichdimensionalen Räumen angegeben. Der wesentliche Vorteil der Aufstellung des Gauß-Newton-Systems im (ganzen) Hilbertraum H ist nun, dass es sich hier um ein kompatibles Problem handelt, da das Problem als in H eindeutig lösbar angenommen wurde. Damit können die dementsprechenden, besseren Konvergenzergebnisse verwendet werden, nachdem sie für den unendlichdimensionalen Fall übertragen worden sind. Dies steht vor allem im Gegensatz zum **DLL**-Verfahren aus Abschnitt 2.2, wo es sich wegen der Aufstellung des Minimierungsproblems in einem endlichdimensionalen Teilraum H_l des Lösungsraums H nur um ein inkompatibles Problem handelt.
Es stehen damit zwei mögliche Kontrollabschätzungen aus dem Abschnitt 1.1.5 zur Verfügung. Wegen der einfacheren Handhabbarkeit wird hier die erste dieser beiden Abschätzungen verwendet, die sich nach einer sinngemäßen Anpassung der Voraussetzung 1.4 ergibt: An die Stelle dieser Voraussetzungen an die endlichdimensionale Abbildung F tritt nun

Voraussetzung 3.2 *Seien $\mathcal{R}$ und $\mathcal{J}$ definiert wie in Abschnitt 2.1 mit $x_* \in H$ als eindeutiger Lösung von $\mathcal{R}(x) = 0$. Sei $\mathcal{J}$ Lipschitz-stetig mit Konstante γ :*

$$\|\mathcal{J}(x)\, z \,-\, \mathcal{J}(y)\, z\|_{0,\Omega} \,\leq\, \gamma\, \|x - y\|_{\mathrm{H}}\, \|z\|_{\mathrm{H}}\ \forall\, x, y, z \,\in H. \tag{3.5}$$

Sei zu $x_l \in H$ ein $\delta \in H$ bestimmt durch die exakte Lösung der Aufgabe

$$(\mathcal{J}(x_l)\, \delta, \mathcal{J}(x_l) v)_{0,\Omega} \,=\, -(\mathcal{R}(x_l), \mathcal{J}(x_l) v)_{0,\Omega}\ \ \forall\, v \,\in H \tag{3.6}$$

und gelte für ein $\epsilon_l \in H$

$$\|\epsilon_l\|_{\mathcal{J},x_l} \,=\, \|\mathcal{J}(x_l)\, \epsilon_l\|_{0,\Omega} \,<\, -\frac{(\mathcal{R}(x_l), \mathcal{J}(x_l)\, \delta)_{0,\Omega}}{\|\mathcal{R}(x_l)\|_{0,\Omega}} \,=\, e_l\,. \tag{3.7}$$

Wieder ist eine Einschränkung dieser Voraussetzungen auf eine Umgebung D von x_* möglich, wodurch die folgenden Ergebnisse dann nur in dieser Umgebung gültig sind. Da jedoch im Folgenden Abstiegsverfahren bzgl. des Funktionals $\mathcal{F}(x) = \|\mathcal{R}(x)\|^2_{0,\Omega}$ konstruiert werden, und dieses Funktional ein Fehlerschätzer für $\|x - x_*\|_{\mathrm{H}}$ ist, kann eine solche Umgebung von einem x_0 ausgehend angegeben werden:

$$D = \{y \in H : \|\mathcal{R}(y)\|_{0,\Omega} \,\leq\, \|\mathcal{R}(x_0)\|_{0,\Omega}\}.$$

Dies ist dann sinnvoll, wenn für $\mathcal{J}$ nur lokale Abschätzungen oder Eigenschaften, wie zum Beispiel lokale Lipschitz-Stetigkeit, zur Verfügung stehen.
Voraussetzung 3.2 stellt sicher, dass $\delta_l + \epsilon_l$ eine Abstiegsrichtung an das Fehlerfunktional $\mathcal{F}(x)$ an der Stelle x_l darstellt. Die Ableitung von $\mathcal{F}$ an der Stelle x in Richtung v ergibt sich nach der Definition der Fréchet-Ableitung $\mathcal{J}$ zu

$$\mathcal{F}'(x)\, v \,=\, 2(\mathcal{R}(x), \mathcal{J}(x)\, v)_{0,\Omega}\ \ \forall\, x, v \in H\,. \tag{3.8}$$

Die exakte Gauß-Newton-Richtung ist wegen (2.4) eine Abstiegsrichtung und damit ist die rechte Seite der Ungleichung (3.7), also der Wert e_l positiv.
Daher ist unter den Bedingungen von Voraussetzung 3.2 die Richtung $\delta_l + \epsilon_l$ eine Abstiegsrichtung an $\mathcal{F}$ an der Stelle x_l, denn

$$\begin{aligned}
(\mathcal{R}(x_l), \mathcal{J}(x_l)\, (\delta_l + \epsilon_l))_{0,\Omega} \quad &= \quad (\mathcal{R}(x_l), \mathcal{J}(x_l)\, \delta_l)_{0,\Omega} \,+\, (\mathcal{R}(x_l), \mathcal{J}(x_l)\, \epsilon_l)_{0,\Omega} \\
&\leq \quad (\mathcal{R}(x_l), \mathcal{J}(x_l)\, \delta_l)_{0,\Omega} \,+\, \|\mathcal{R}(x_l)\|_{0,\Omega}\, \|\epsilon_l\|_{\mathcal{J},x_l} \\
&\overset{(3.7)}{<} \quad (\mathcal{R}(x_l), \mathcal{J}(x_l)\, \delta_l)_{0,\Omega} \,-\, (\mathcal{R}(x_l), \mathcal{J}(x_l)\, \delta_l)_{0,\Omega} \\
&= \quad 0\,.
\end{aligned}$$

In einer Aussage analog zu Lemma 1.3 erhält man daher, dass ein ω_1 mit $0 < \omega_1 \leq 1$ und ein $\tilde{\eta}_l < 1$ existiert mit

$$\|\mathcal{R}(x_l) + \mathcal{J}(x_l)\,\omega_1(\delta_l + \epsilon_l)\|_{0,\Omega} \leq \tilde{\eta}_l \|\mathcal{R}(x_l)\|_{0,\Omega} \tag{3.9}$$

(verwende dabei die Voraussetzung (2.2) für den Schritt (1.15)).
Um mit dieser Richtung auch eine Reduktion in $\mathcal{F}$ selbst und nicht nur in einer Linearisierung davon zu erhalten, muss ω_1 unter Umständen weiter reduziert werden, was dann $\tilde{\eta}_l$ vergrößert. In welchem Sinne dies zu verstehen ist, ergibt sich aus dem folgenden Satz, der genau wie im endlichdimensionalen Fall in [19, Lemma 3.1] bewiesen werden kann.

Satz 3.3 *Sei $x_l \in H$ und $t \in (0,1)$ gegeben. Sei $\bar{\delta}$ so, dass*

$$\|\mathcal{R}(x_l) + \mathcal{J}(x_l)\,\bar{\delta}\|_{0,\Omega} < \|\mathcal{R}(x_l)\|_{0,\Omega}$$

gilt. Dann existiert ein $\eta_{\min} \in [0,1)$, so dass für alle $\eta \in [\eta_{\min}, 1)$ ein $\omega_2 < 1$ existiert mit

$$\|\mathcal{R}(x_l) + \mathcal{J}(x_l)\,\omega_2\bar{\delta}\|_{0,\Omega} \leq \eta \|\mathcal{R}(x_l)\|_{0,\Omega}\,, \tag{3.10}$$

$$\|\mathcal{R}(x_l + \omega_2\bar{\delta})\|_{0,\Omega} \leq [1 - t(1-\eta)]\|\mathcal{R}(x_l)\|_{0,\Omega}\,. \tag{3.11}$$

Aus Satz 3.3 folgt, dass zu δ_l, ϵ_l und ω_1, wie oben definiert, ein ω_2 existiert, so dass für die Korrektur $\hat{\delta}_l = \omega_1\omega_2(\delta_l + \epsilon_l)$ und $\hat{\eta}_l = \eta_{\min} < 1$ die beiden Abschätzungen

$$\|\mathcal{R}(x_l) + \mathcal{J}(x_l)\,\hat{\delta}_l\|_{0,\Omega} \leq \hat{\eta}_l \|\mathcal{R}(x_l)\|_{0,\Omega}\,, \tag{3.12}$$

$$\|\mathcal{R}(x_l + \hat{\delta}_l)\|_{0,\Omega} \leq [1 - t(1-\hat{\eta}_l)]\|\mathcal{R}(x_l)\|_{0,\Omega} \tag{3.13}$$

gelten.
Ein Konvergenzresultat für die Folge $\{x_l\}_{l=0}^{\infty}$ bei Vorliegen von Reduktion nach (3.13) auf jedem Level l folgt ebenfalls analog zu [19, Satz 3.3]. Nach den Herleitungen aus dem endlichdimensionalen Fall in Abschnitt 1.1.5 ergibt sich folgendes Konvergenzergebnis als einfache Folgerung:

Satz 3.4 *Sei Voraussetzung 3.2 erfüllt und t fest mit $0 < t < 1$. Dann können in jedem Schritt (durch Dämpfung) für ein $\hat{\eta}_l$ und $\hat{\delta}_l = \omega(\delta_l + \epsilon_l)$ die Bedingungen (3.12) und (3.13) erfüllt werden. Setzt man $H_{l+1} \ni x_{l+1} = \mathcal{I}_l^{l+1}(x_l + \hat{\delta}_l)$ und definiert*

$$\hat{c}_{l+1} = (\hat{\eta}_{l+1} - \hat{\eta}_l)(l+1)l \tag{3.14}$$

eine Hilfsgröße mit $\hat{c}_l = O(\frac{1}{l})$, so gilt

$$x_l \to x_*.$$

Dabei ist $\mathcal{I}_l^{l+1}$ die Einbettungsabbildung für die geschachtelten Räume $H_l \subseteq H_{l+1}$ (siehe auch Abschnitt 3.3).

Beweis:
Die Erfüllbarkeit der Bedingungen (3.12) und (3.13) folgt aus dem Vorhergehenden. Mit der auf den Hilbertraum-Fall übertragenen Theorie aus [19] lässt sich damit die Konvergenz von x_l gegen die Lösung x_* aus der Divergenz der Reihe

$$\sum_{l \geq 0}(1 - \hat{\eta}_l)$$

folgern (siehe (1.25)).
Sei $\hat{c}_{l+1}$ nach (3.14) für alle $l \geq 0$ definiert und gelte $\hat{c}_l = O(\frac{1}{l})$.
Dann ist

$$\begin{aligned}(\hat{\eta}_{l+1} - \hat{\eta}_l) &= [(1-\hat{\eta}_l) - (1-\hat{\eta}_{l+1})] \\ &\overset{(3.14)}{=} \frac{\hat{c}_{l+1}}{l(l+1)}\,.\end{aligned}$$

Daraus folgt unmittelbar

$$(1-\hat{\eta}_{l+1}) \;=\; (1-\hat{\eta}_l) - \frac{C}{l(l+1)}$$

und damit für $n \geq 3$

$$\begin{aligned}\sum_{l=0}^{n}(1-\hat{\eta}_l) &\geq \sum_{l=0}^{n-1}(1-\hat{\eta}_l) + 1\cdot(1-\hat{\eta}_{n-1}) - \frac{\hat{c}_n}{(n-1)n} \\ &\geq \sum_{l=0}^{n-2}(1-\hat{\eta}_l) + 2\cdot(1-\hat{\eta}_{n-2}) - \frac{\hat{c}_n}{(n-1)n} - \frac{2\hat{c}_{n-1}}{(n-2)(n-1)} \\ &\vdots \\ &\geq \underbrace{n\,(1-\hat{\eta}_0)}_{\to\infty (n\to\infty)} - \underbrace{\sum_{j=1}^{n-1}\frac{\hat{c}_{n+1}(n-j)}{j(j+1)}}_{\text{beschränkt}},\end{aligned}$$

da der Zähler innerhalb der Summe beschränkt ist und die Summe damit konvergiert. Es ist also (1.25) erfüllt. Nun folgt wie in der Theorie in [19, Satz 3.4] die Konvergenz $x_l \to x_*$. □

In den folgenden Abschnitten werden nun Methoden angegeben, mit Hilfe derer die Bedingung (3.7) im Zusammenhang mit (3.4) erfüllt werden kann. Sind dann auch die übrigen Forderungen in Voraussetzung 3.2 erfüllt, so können durch Dämpfung die Bedingungen (3.12) und (3.13) erfüllt werden, so dass dann mit Hilfe von (3.14) die Konvergenz der Folge $\{x_l\}_{l=0}^{\infty}$ kontrolliert werden kann.

3.2 Kontrolle des Diskretisierungsfehlers

Die maximale Ungenauigkeit e_l gemäß (3.7) kann auf den Diskretisierungsfehler und den algebraischen Fehler wegen (3.4) zum Beispiel zu gleichen Teilen aufgeteilt werden. Für ein festes (Level) l ist auch e_l eine feste Größe, auf deren (näherungsweise) Berechnung in Abschnitt 3.3 eingegangen wird. Für diesen Abschnitt sei e_l also gegeben und fest.
In Bezug auf den Diskretisierungsfehler ist dann nur noch von Bedeutung, ob die Genauigkeitsbedingung für eine bestimmte Wahl des diskreten Raums H_l überhaupt erfüllt werden kann. Um zu sehen, wie typische FE-Abschätzungen angewendet werden können, führt man für festes $x = x_l$ die Bilinearform $a(\cdot,\cdot)$ und die Linearform g ein: (vergleiche (1.63)):

$$a(u,v) \;=\; (\mathcal{J}(x)\,u, \mathcal{J}(x)\,v)_{0,\Omega} \quad u,v \in H\,, \tag{3.15}$$

$$g(v) \;=\; -(\mathcal{R}(x), \mathcal{J}(x)\,v) \quad v \in H\,. \tag{3.16}$$

Wegen der Voraussetzungen in Abschnitt 2.1 gilt

$$\begin{aligned}a(u,v) &= a(v,u) \quad \forall\, u,v \in H\,, \\ |a(u,v)| &\leq \bar{\alpha}^2 \|u\|_{\mathrm{H}}\|v\|_{\mathrm{H}} \quad \forall\, u,v \in H\,, \\ a(u,u) &\geq \underline{\alpha}^2 \|u\|_{\mathrm{H}}^2 \quad \forall\, u \in H.\end{aligned}$$

Seien $\delta \in H$ und $\delta_l \in H_l$ die jeweiligen Lösungen der Variationsaufgaben

$$\begin{aligned} a(\delta, v) &= g(v) \quad \forall\, v \in H\,, \\ a(\delta_l, v_l) &= g(v_l) \quad \forall\, v_l \in H_l, \end{aligned}$$

was offensichtlich der Lösung der Normalengleichungen in den jeweiligen Räumen H und H_l entspricht.
Nach dem Céa-Lemma (vgl. Gleichung 1.64) gilt dann

$$\|\delta - \delta_l\|_{\mathrm{H}} \;\leq\; \frac{\bar{\alpha}}{\underline{\alpha}} \inf_{v_l \in H_l} \|\delta - v_l\|_{\mathrm{H}}. \tag{3.17}$$

Unter genügenden Regularitätsvoraussetzungen an die Lösung δ und geeigneten FE-Räumen H_l gilt eine Interpolationsabschätzung der Art

$$\|\delta - v_l\|_{\mathrm{H}} \;\leq\; C\, h_l^j\, \|\delta\|_j \tag{3.18}$$

(siehe z.B. [25] und Anhang B), wobei mit $j \in (0, 1]$ die Norm $\|\cdot\|_j$ von der vorgegebenen Differentialgleichung und der Regularität der Lösung δ abhängt. h_l bezeichnet die Feinheit der Zerlegung $\mathcal{T}_l$ von Ω, über der der Raum H_l stückweise linear aufgestellt wird (siehe Abschnitt 1.3.2). Wegen $\|\cdot\|_{\mathrm{H}} \geq (1/\bar{\alpha})\|\cdot\|_{\mathcal{J},x}$ in H gilt dann zusammen mit (3.17) folgende Abschätzung des Diskretisierungsfehlers

$$\|\delta - \delta_l\|_{\mathcal{J},x} \;\leq\; h_l^j\, \frac{C\bar{\alpha}^2}{\underline{\alpha}}\, \|\delta\|_j. \tag{3.19}$$

Wird nun durch (3.4) und (3.7) ein maximaler Diskretisierungsfehler vorgeschrieben, so ist danach die Feinheit h_l in (3.19) klein genug zu wählen.
Bedingung (3.19) ist für x_l in der Nähe der Lösung eine Forderung, die sich auch direkt bei der Betrachtung des nichtlinearen Minimierungsproblems ergibt. Es stelle x_* die Lösung des Minimierungsproblems in H und $x_{l,*}$ die Lösung des Minimierungsproblems in H_l dar, d.h.

$$\begin{aligned} \|\mathcal{R}(x_*)\|_{0,\Omega}^2 &= \min_{x \in H} \|\mathcal{R}(x)\|_{0,\Omega}^2\,, \\ \|\mathcal{R}(x_{l,*})\|_{0,\Omega}^2 &= \min_{x_l \in H_l} \|\mathcal{R}(x_l)\|_{0,\Omega}^2\,. \end{aligned}$$

Für eine Näherung $x_l \in H_l$, die nahe genug an der Lösung x_* liegt, ist dann

$$\begin{aligned} x_* &\approx x_l + \delta, \\ x_{l,*} &\approx x_l + \delta_l, \end{aligned}$$

wobei δ, δ_l die Lösungen der Normalengleichungen darstellen. Damit ist jedoch

$$\|x_* - x_{l,*}\|_{\mathrm{H}} \;\approx\; \|\delta - \delta_l\|_{\mathrm{H}}.$$

Für x_l in der Nähe der Lösung ist die Bedingung (3.19) also von gleicher Ordnung wie die üblichen Finite-Element-Interpolationsbedingungen für die Approximierbarkeit von x_* durch $x_{l,*}$ (siehe auch [24, Kapitel IV]).
Bei Konvergenz des Gauß-Newton-Verfahrens in H folgt, dass sowohl die Gauß-Newton-Korrektur δ_l, als auch der störende Fehler ϵ_l gegen Null gehen muss. Natürlich ist dies für den Anteil des Diskretisierungsfehlers nur eine theoretische Forderung, die auf Grund der Beschränkung der Rechnerkapazität nicht erfüllt werden kann. Ab einer gewissen Gauß-Newton-Iterationsanzahl wird die vom Programm zur Verfügung gestellte Schachtelungstiefe der Verfeinerungen nicht mehr ausreichen. Dadurch besitzt der Diskretisierungsfehler also eine natürliche, von Implementierung und Hardware abhängige untere Schranke. Wenn nun die Bedingungen

(3.12) und (3.13) wegen des Diskretisierungsfehlers auf einem Level l nicht mehr erfüllt werden können, ist dies sofort ein Abbruchkriterium an die Iteration auf diesem Level l, da Abstiegsrichtungen bei weiterer Berechnung nicht mehr garantiert werden können. Ist die Rechnergrenze noch nicht erreicht, so kann der Diskretisierungsfehler durch Übergang zu einem feineren Raum H_{l+1} so verkleinert werden, dass die Konvergenzbedingungen (3.12) und (3.13) wieder erfüllbar sind. Kann nicht mehr weiter verfeinert werden, so ist, wie eben gezeigt, nahe der Lösung x_* der Diskretisierungsfehler in x_* von der gleichen Größenordnung wie der Diskretisierungsfehler in der Berechnung der Gauß-Newton-Korrektur δ. Dies ist jedoch das Ziel jedes nichtlinearen Lösungsverfahrens. Für ähnliche Beobachtungen eines kaskadischen Multilevelverfahrens für semilineare Probleme siehe [44].

3.3 Kontrolle des algebraischen Fehlers

Während der Diskretisierungsfehler natürlich nur durch eine entsprechende Feinheit h_l der zu Grunde liegenden Zerlegung kontrolliert werden kann, ist für die Kontrolle des algebraischen Fehlers das Abbruchkriterium bei der iterativen Lösung des linearen Problems entscheidend.
Zu lösen ist für gegebenes $x_l \in H_l$ die lineare Aufgabe

$$\begin{aligned} &(\mathcal{J}(x_l)\,\delta_l, \mathcal{J}(x_l)\,v_l)_{0,\Omega} \;=\; -(\mathcal{R}(x_l), \mathcal{J}(x_l)\,v_l)_{0,\Omega} \\ &\text{oder auch} \\ &a(\delta_l, v_l) \;=\; l(v_l) \end{aligned} \tag{3.20}$$

für alle $v_l \in H_l$ mit den Definitionen von (3.15) und (3.16).
Wie schon im vorhergehenden Abschnitt festgestellt, ist dies ein lineares, H-koerzives Variationsproblem. Für solche Probleme ist das Prinzip von linearen Multilevelverfahren bereits in Abschnitt 1.4.1 beschrieben worden.
Am Ende des Abschnitts 1.3.2 wurde auch auf den Einsatz von Fehlerschätzern beim Aufbau geeigneter Diskretisierungen von Gebieten, auf denen lokal stark nichtlineare Probleme gelöst werden sollen, eingegangen. Wegen der Kontrolle des Diskretisierungsfehlers wie im vorhergehenden Abschnitt ähnelt nun das Gauß-Newton-Multilevelverfahren dem Full-Multigrid-Zyklus insofern, als bei Erreichen eines Residuums in der Größenordnung des Diskretisierungsfehlers der Raum H_l vergrößert werden muss, um weiterhin Konvergenz zu ermöglichen, bis die Grenzen der Hard- oder Software erreicht sind.
Für die folgenden Überlegungen zur Kontrolle des algebraischen Fehlers sei ein Level $l > 1$ fest gewählt. Diesem entspricht dann der zugehörige endlichdimensionale Raum H_l. Die Startnäherung sei unter Berücksichtigung der Genauigkeitsbedingung bereits auf Level $l-1$ berechnet worden und stehe durch Prolongation auf Level l zur Verfügung. Zu untersuchen ist nun, wann der V-Zyklus, der zur Lösung des linearen Problems (3.20) auf Level l verwendet wird, abgebrochen werden kann. Anders ausgedrückt stellt sich die Frage, wann der algebraische Fehler bei der Lösung von (3.20) so klein ist, dass die Genauigkeitsbedingung (3.7) für den algebraischen Fehler erfüllt ist. Dabei wird angenommen, dass der Anteil des Diskretisierungsfehlers nach (3.4) diese Bedingung bereits erfüllt.

Zum Aufstellen der Gleichungssysteme, die aus der Aufgabe (3.20) hervorgehen, siehe Abschnitt 1.3.3. Dabei werden im Folgenden die Besonderheiten bei der Einbeziehung von inhomogenen Randbedingungen außer Acht gelassen. Für die formale Beschreibung des in diesem Abschnitt zu entwickelnden Abbruchkriteriums werden folgende Bezeichnungen eingeführt:
Sei

$$H_0 \subseteq H_1 \subseteq \ldots \subseteq H_l$$

eine Familie von endlichdimensionalen Unterräumen von H, die durch sukzessive Verfeinerung entstanden ist. In jedem dieser Räume lässt sich eine Basis einführen, mit Hilfe derer eine Funktion $u_j \in H_j$ als Vektor $x_j \in \mathbb{R}^{N_j}$ darstellbar ist, wobei $N_j = \dim(H_j)$ gilt. Beschrieben werde dies durch die bijektive Abbildung

$$R_j : H_j \rightarrow \mathbb{R}^{N_j} \ , \ u_j \mapsto y_j.$$

Zwischen zwei aufeinanderfolgenden Räumen H_j und H_{j+1} existiert die kanonische Einbettungsabbildung $\mathcal{I}_j^{j+1}$, deren Matrixrepräsentation mit I_j^{j+1} bezeichnet wird :

$$I_j^{j+1} \ = \ R_{j+1} \circ \mathcal{I}_j^{j+1} \circ R_j^{-1}.$$

Dadurch lässt sich zwischen H_i und H_j, mit $i > j$ bel., ein Prolongationsoperator definieren:

$$I_j^i \ = \ I_{i-1}^i \circ \ldots \circ I_{j+1}^{j+2} \circ I_j^{j+1} \, .$$

Wird der Restriktionsoperator $\mathcal{I}_{j+1}^j : H_{j+1} \rightarrow H_j$ mit Hilfe von

$$(\mathcal{I}_{j+1}^j u_{j+1}, v_j)_{0,\Omega} \ = \ (u_{j+1}, v_j)_{0,\Omega} \quad \forall \, v_j \in H_j$$

definiert, so kann man schnell feststellen, dass sich die entsprechende Matrixdarstellung durch Transponieren der Prolongation erhalten lässt:

$$I_{j+1}^j \ = \ (I_j^{j+1})^T,$$

(siehe z.B. [10]).
Die lineare Aufgabe (3.20) werde dann für ein beliebiges Level j beschrieben durch das lineare Gleichungssystem

$$A_j \, y_{j,*} \ = \ b_j.$$

Dabei können (für die oben definierten Interpolations- und Restriktionsabbildungen) die Matrizen auf niedrigeren Leveln durch die sogenannte Galerkin-Approximation der nächsthöheren Matrix gebildet werden:

$$A_{j-1} \ = \ I_j^{j-1} \, A_j \, I_{j-1}^j.$$

Die Systemmatrizen dieser Gleichungssysteme sind wegen der Eigenschaften der Bilinearform a positiv definit und symmetrisch.
Die Iterierten zur Näherung an y_l werden durch einen weiteren unteren Index bezeichnet, also zum Beispiel die Startnäherung auf Level l mit $y_{l,0}$. Das Residuum zum k-ten Iterationsschritt auf Level l wird mit

$$r_{l,k} \ = \ b_l \ - \ A_l \, y_{l,k} \ = \ A_l \, (y_{l,*} - y_{l,k})$$

bezeichnet.
Wesentlich ist im Folgenden auch das Verhältnis zwischen den Normen in H_j und $\mathbb{R}^{N_j}$. In der Genauigkeitsabschätzung taucht die Energienorm $\| \cdot \|_{\mathcal{J},x}$ auf. Wegen (3.15) wird diese Norm für festes x durch die Bilinearform a induziert, durch die wiederum die Matrix A_j aufgestellt wird. Dadurch besteht der folgende Zusammenhang:

$$\|u\|_{\mathcal{J},x}^2 \ = \ a(u,u) \ = \ (R_j u)^T \, A_j \, (R_j u) \ = \ y_j^T \, A_j \, y_j \quad \forall \, u \in H_j \, . \tag{3.21}$$

Als Glättungsiteration innerhalb des V-Zyklus wird auf allen Leveln ein nodaler Gauß-Seidel-Glätter mit je einem Vor-und einem Nachglättungsschritt angewendet. Beschreibt dann C_l^{-1} die (symmetrische) Iterationsmatrix eines V-Zyklus-Schrittes, d.h.

$$y_{l,k+1} \ = \ y_{l,k} \ + \ \underbrace{C_l^{-1} \, r_{l,k}}_{\text{Multilevelkorrektur}} \quad , \tag{3.22}$$

und damit

$$r_{l,k+1} \;=\; (I - A_l\, C_l^{-1})\, r_{l,k}\,, \tag{3.23}$$

so ist aus der Theorie von Multilevelverfahren für elliptische Probleme (siehe z.B. [27]) bekannt, dass für die Eigenwerte λ der Fehlerfortpflanzungsmatrix $E_l \;=\; (I - A_l\, C_l^{-1})$ gilt : $0 < \lambda \leq \bar{\lambda} < 1$. Offensichtlich ist daher der V-Zyklus eines Multilevelverfahrens ein effektiver Vorkonditionierer für die Matrix A_l: Es ist $\rho(E_l) < 1$ für den Spektralradius von E_l, und damit gilt für die Konditionszahl von $A_l\, C_l^{-1}$: $\kappa(A_l\, C_l^{-1}) \;=\; \frac{1+\rho(E_l)}{1-\rho(E_l)}$. Wegen (3.23) kann $\rho(E_l)$ durch

$$\rho(E_l) \;\approx\; \frac{\|r_{l,k+1}\|_2}{\|r_{l,k}\|_2} \tag{3.24}$$

angenähert werden.
Damit lässt sich auch die Konditionszahl der durch den V-Zyklus vorkonditionierten Matrix A_l schätzen und ergibt

$$\kappa(A_l\, C_l^{-1}) \;\approx\; \frac{\|r_{l,k}\|_2 + \|r_{l,k+1}\|_2}{\|r_{l,k}\|_2 - \|r_{l,k+1}\|_2}. \tag{3.25}$$

Mit Hilfe der vorhergehenden Definitionen und Ergebnisse sind jetzt die Bedingungen zur Einhaltung der Genauigkeitsbedingung (3.7) einfach herzuleiten. Zunächst gilt für die linke Seite der Ungleichung

$$\begin{aligned}
\|\epsilon_{l,k}\|^2_{\mathcal{J},x_l} \;&=\; \|\mathcal{J}(x_l)\,\epsilon_{l,k}\|^2_{0,\Omega} \\
&=\; \|\mathcal{J}(x_l)\,(\delta - \tilde{\delta}_{l,k})\|^2_{0,\Omega} \\
&=\; \underbrace{\|\mathcal{J}(x_l)\,(\delta - \delta_l)\|^2_{0,\Omega}}_{=err^2_{\mathrm{dis}}} + \|\mathcal{J}(x_l)\,(\delta_l - \tilde{\delta}_{l,k})\|^2_{0,\Omega} \\
&\quad + \underbrace{(\mathcal{J}(x_l)(\delta - \delta_l), \mathcal{J}(x_l)(\delta_l - \tilde{\delta}_{l,k}))_{0,\Omega}}_{=0\ (\text{Galerkin Projektion}))} \\
&=\; err^2_{\mathrm{dis}} + (y_{l,*} - y_{l,k})^T\, A_l\, (y_{l,*} - y_{l,k}) \\
&=\; err^2_{\mathrm{dis}} + r_{l,k}^T\, A_l^{-1}\, r_{l,k} \\
&\leq\; err^2_{\mathrm{dis}} + \kappa(A_l\, C_l^{-1})\, r_{l,k}^T\, C_l^{-1}\, r_{l,k} \\
&=\; err^2_{\mathrm{dis}} + \kappa(A_l\, C_l^{-1})\, r_{l,k}^T\, d_{l,k} \\
&\approx\; err^2_{\mathrm{dis}} + \frac{\|r_{l,k}\|_2 + \|r_{l,k+1}\|_2}{\|r_{l,k}\|_2 - \|r_{l,k+1}\|_2}\, r_{l,k}^T\, d_{l,k}\,.
\end{aligned} \tag{3.26}$$

Dabei bezeichnet $d_{l,k}$ die Multilevelkorrektur an die aktuelle Näherung $y_{l,k}$ nach (3.22). Damit kann dieser Teil der zu prüfenden Bedingung (3.7) durch einen berechenbaren Ausdruck abgeschätzt werden.
Nun ist noch e_l auf der rechten Seite der Bedingung (3.7) abzuschätzen, denn die exakte Berechnung von e_l würde die exakte Gauß-Newton-Korrektur δ voraussetzen, die ja offensichtlich nicht zur Verfügung steht.
Zunächst ist $F_l \;=\; \|\mathcal{R}(x_l)\|_{0,\Omega}$ mit Quadraturformeln sehr genau approximierbar. Da F_l ohnehin in der Bedingung (3.13) benötigt wird, muss dieser Wert ohnehin innerhalb des Verfahrens

berechnet werden. Nun gilt zum anderen

$$\begin{aligned} e_l = -\frac{(\mathcal{R}(x_l), \mathcal{J}(x_l)\,\delta)_{0,\Omega}}{\|\mathcal{R}(x_l)\|_{0,\Omega}} &= \frac{(\mathcal{J}(x_l)\,\delta, \mathcal{J}(x_l)\,\delta)_{0,\Omega}}{F_l} \\ &= \frac{err_{\text{dis}}^2 + y_{l,*}^T\, A_l\, y_{l,*}}{F_l} \\ &\geq \frac{b_l^T\, A_l^{-1}\, b_l}{F_l} \\ &\geq \frac{1}{\kappa(A_l\, C_l^{-1})}\, \frac{b_l^T\, C_l^{-1}\, b_l}{F_l} \\ &= \frac{1}{\kappa(A_l\, C_l^{-1})}\, \frac{b_l^T\, d_{l,-1}}{F_l} \\ &\approx \frac{\|r_{l,k}\|_2 - \|r_{l,k+1}\|_2}{\|r_{l,k}\|_2 + \|r_{l,k+1}\|_2}\, \frac{b_l^T\, d_{l,-1}}{F_l}. \end{aligned} \tag{3.27}$$

Hier muss noch der Vektor $d_{l,-1}$ als Ergebnis eines V-Zyklus mit rechter Seite b_l und Startnäherung $x_{l,0} = 0$ berechnet werden. Damit stehen alle für die Kontrolle des algebraischen Fehlers notwendigen Abschätzungen als berechenbare Ausdrücke zur Verfügung. Zum Schluss dieses Abschnitts wird der damit vollständige Algorithmus zur Kontrolle des algebraischen Fehlers mit Abbruchkriterien zur Einhaltung der Genauigkeitsbedingung (3.7) angegeben.

- Gegeben: Startnäherung $y_{l,0} \in \mathbb{R}^{N_l}$, A_l, b_l, F_l Setze $k = 0$.
- Führe einen V-Zyklus mit 0 als Startnäherung durch. Ergebnis $d_{l,-1}$.
- Setze $r_{l,k} = b_l - A_l\, y_{l,k}$.
- 1. Berechne $d_{l,k} = C_l^{-1}\, r_{l,k}$ (V-Zyklus).
  2. Setze $y_{l,k+1} = y_{l,k} + d_{l,k}$ und $r_{l,k+1} = b_l - A_l\, y_{l,k+1}$.
  3. Berechne $err_k^2 = \frac{\|r_{l,k}\|_2 + \|r_{l,k+1}\|_2}{\|r_{l,k}\|_2 - \|r_{l,k+1}\|_2}\, r_{l,k}^T\, d_{l,k}$.
  4. Berechne $con_k = \frac{\|r_{l,k}\|_2 - \|r_{l,k+1}\|_2}{\|r_{l,k}\|_2 + \|r_{l,k+1}\|_2}\, \frac{b_l^T\, d_{l,-1}}{F_l}$.
  5. Gilt $err_k \geq con_k$, setze $k = k + 1$ und gehe zu 1.
- Ergebnis $y_{l,k+1}$.

Algorithmus 3.1 : V-Zyklus im Gauß-Newton-Verfahren

Da der Fehlerfortsetzungsoperator E_l einen Spektralradius kleiner 1 besitzt, gilt

$$y_{l,k} \overset{k\to\infty}{\longrightarrow} y_{l,*} = A_l^{-1}\, b_l\,,$$

und damit gilt

$$r_{l,k} \overset{k\to\infty}{\longrightarrow} 0.$$

Dann konvergiert jedoch auch err_k aus Algorithmus 3.1 wegen (3.26) gegen 0. Die rechte Seite hingegen bleibt echt größer Null. Daher muss dieser Algorithmus nach einer gewissen Anzahl von Schritten k von selbst abbrechen.

Es ist auch denkbar, $con_k \equiv con_0$ zu setzen und auf die jeweils aktuelle Näherung an die Konditionszahl $\kappa(A_l\, C_l^{-1})$ an dieser Stelle zu verzichten.
Bei der Anwendung des V-Zyklus zur Lösung linearer Probleme sind Konvergenzraten von $\frac{1}{10}$ nicht ungewöhnlich. Dann liegt jedoch die Konditionszahl $\kappa(A_l\, C_l^{-1})$ nahe bei 1 und es geht bei den Abschätzungen in (3.26) und (3.27) nicht viel Genauigkeit verloren.
Wenn der Diskretisierungsfehler mit in die Abschätzung einbezogen werden soll, muss in Punkt 5 des obigen Algorithmus das Abbruchkriterium $err_k \geq con_k$ durch $err_k \geq C \cdot con_k$ mit einem $C < 1$ ersetzt werden, um einen Anteil von $(1 - C)con_k$ für den Diskretisierungsfehler verwenden zu können. Insbesondere folgt aus der Wahl von $C = 0.5$ bei Nichterfüllbarkeit der Konvergenzbedingungen (3.12) und (3.13) die Abschätzung $err_{\text{dis}} \geq err_k$, was der Reduktion des algebraischen Fehlers auf die Größe des Diskretisierungsfehlers entspricht.
Schließlich ist die berechnete Korrektur $\tilde{\delta}_{l,k} = R_l^{-1} y_{l,k}$ noch so weit zu dämpfen, bis die Bedingungen (3.12) und (3.13) erfüllt sind.

3.3.1 Der LDL-Algorithmus

1. Gegeben : Hierarchie von Räumen $H_0 \subseteq H_1 \subseteq \ldots \subseteq H_l$, $k_{\max} \in \mathbb{N}$, $0 < t < 1$, $l = 0$.
2. Sei $x_l \in H_l$ die aktuelle Näherung an die Lösung der Differentialgleichung und
$$A_l y_l = b_l \tag{3.28}$$
das zum nächsten Gauß-Newton-Schritt gehörige Gleichungssystem.
3. Wende den V-Zyklus aus Algorithmus 3.1 auf (3.28) an, zähle die Anzahl der V-Zyklen mit k, breche dabei spätestens nach $k = k_{\max}$ Schritten ab
$\Rightarrow$ Näherung $y_{l,k}$ an Lösung des LGS (3.28).
   (a) Ist $k = k_{\max}$ (d.h. war die Erfüllung der Bedingung aus Punkt 5 in Algorithmus 3.1 für $k \leq k_{\max}$ nicht möglich) dann gehe zu 4.
   (b) (Kontrolle der Konvergenzbedingungen (3.12) und (3.13))
      - Suche $\omega = 1, 1/2, 1/4, 1/8, \ldots$ so, dass mit $\hat{\delta}_l \simeq \omega\, y_{l,k}$ die Bedingungen (3.12) und (3.13) erfüllt sind $\Rightarrow \hat{\eta}_l$
      - Wenn $l \geq 2$ setze $\hat{c}_l = (\hat{\eta}_l - \hat{\eta}_{l-1})(l+1)l$. Überprüfe $\hat{c}_l \leq C$, $l = 1, 2, \ldots$ für ein $C > 0$ zur Kontrolle der Konvergenz des Gesamtverfahrens.
      - Setze $x_l = x_l + \omega\, \delta_l$, mit $\delta_l \simeq y_{l,k}$. Stelle das neue Gleichungssystem an der Stelle x_l auf und berechne den nächsten Gauß-Newton-Schritt bei 2.
4. Wähle $H_{l+1} \supset H_l$. Setze $x_{l+1,0} = x_{l,k_{\max}}$ und $l = l + 1$. Gehe zu 2. oder breche das Verfahren ab, wenn die maximale Verfeinerungsstufe erreicht war.

Algorithmus 3.2 : LDL-Verfahren

Wie in den Abschnitten 2.2.3 und 2.3.3 wurde hier ein kompletter Algorithmus angegeben, der allerdings im Wesentlichen auf Algorithmus 3.1 zurückgreift. Das Zeichen $\simeq$ gibt wie in den anderen Algorithmen die Beziehung zwischen dem Vektor der Koeffizienten der Basisdarstellung

eines Elements $v_j \in H_j$ und dem Element selbst an:

$$\begin{aligned} y_j &\simeq R_j v_j \,, \\ (R_j)^{-1} y_j &\simeq v_j \,. \end{aligned}$$

Kapitel 4

Anwendungsbeispiel und numerischer Vergleich

In den vorhergehenden Kapiteln wurden verschiedene Methoden zur Lösung eines Systems von nichtlinearen partiellen Differentialgleichungen erster Ordnung durch Multilevelverfahren für den Ausgleichsformulierungsansatz entwickelt. Am Beispiel eines realistischen Anwendungsproblems aus der Strömungsmechanik werden in diesem Kapitel die verschiedenen Methoden angewendet und in ihrem Konvergenzverhalten verglichen.

4.1 Das Beispielproblem

4.1.1 Physikalischer Hintergrund und mathematische Formulierung

Zur Beschreibung von Strömungen einer flüssigen Phase (z.B. Wasser) in einem als starr angenommenen porösen Festkörper sind in der Literatur verschiedene Ansätze zu finden, siehe z.B. [5], [31] und [39], die üblicherweise ein Materialgesetz mit dem Prinzip der Massenerhaltung kombinieren.
Sei ein Gebiet $\Omega \subseteq \mathbf{R}^2$ gegeben und es werde die Strömung von Wasser innerhalb dieses durch ein poröses Material ausgefüllten Gebietes betrachtet. Aus der Beobachtung der Massenbilanz in kleinen Gebieten erhält man das Gesetz von der Erhaltung der Masse, auch Kontinuitätsgleichung genannt:

$$\frac{\partial \Theta(p)}{\partial t} + \operatorname{div}(u) = S. \tag{4.1}$$

In dieser Gleichung bezeichnet u den unbekannten Fluss bzw. das Strömungsfeld und S eine bekannte Quelle oder Senke innerhalb des Gebiets. Es gilt allgemein

$$\begin{aligned} u &: \Omega \times [0,T] \to \mathbf{R}^2 \\ S &: \Omega \times [0,T] \to \mathbf{R}. \end{aligned}$$

Mit Hilfe der Funktion Θ wird die zeitliche Veränderung der Menge des Wassers innerhalb des Gebietes beschrieben, indem

$$\Theta : \mathbf{R} \to \mathbf{R}$$

den Wassergehalt in Abhängigkeit der Druckvariable p angibt. Diese für poröses Material typische Beziehung modelliert den Zusammenhang zwischen Druck, Gravitation und äußeren

(Kapillar-)kräften und dem resultierenden Wassergehalt im Equilibrium: In den ungesättigten Bereichen des Gebiets erzeugen die kapillaren Kräfte innerhalb des porösen Mediums eine gewisse "Saugspannung", die umso höher ist, je geringer der Wassergehalt ist. Sind jedoch die Hohlräume des Mediums mit Wasser gefüllt, besteht ein prinzipiell messbarer hydrostatischer Druck und der Wassergehalt verändert sich nicht. Nun wird die durch die kapillaren Kräfte erzeugte Saugspannung (in der ungesättigten Zone) mit dem hydrostatischen Druck in der gesättigten Zone und mit dem überall wirkenden Gravitationspotential in der Variable p zusammengefasst. Dann kann die nichtlineare Beziehung zwischen diesem Potential p und dem Wassergehalt Θ durch empirisch ermittelte Daten modelliert werden.

Weiterhin besteht zwischen den beiden unbekannten Größen, dem Potential p und dem Fluss u, eine Beziehung, die in der Theorie poröser Medien durch das Darcy-Buckingham-Gesetz beschrieben wird:

$$u + K(p)\nabla p \;=\; 0. \tag{4.2}$$

Danach ist also der sich einstellende Fluss u proportional zum Gradienten des Druckpotentials p. Die Proportionalität hängt offensichtlich von weiteren Materialeigenschaften ab und wird durch eine nichtlineare Funktion K, die hydraulische Leitfähigkeit oder Permeabilität, dargestellt. Diese ebenfalls empirisch zu bestimmende Funktion hängt von Viskosität und Dichte der Flüssigkeit und der örtlichen Beschaffenheit des porösen Mediums ab (für Details siehe z.B. [49]). In [37] und [46] werden Parametrisierungen dieser Funktionen angegeben, deren typische Verläufe für Sand und Lehm in Abbildung 4.1 dargestellt sind ($z \leq 0$ entspricht der Bodentiefe, $z = 0$ der Oberfläche). Für die genauen Parametrisierungen siehe Anhang A.

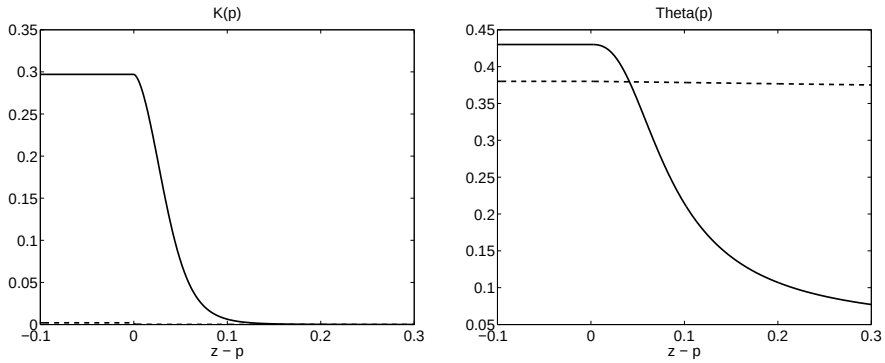


Abbildung 4.1: Mualem/van Genuchten Parametrisierungen für Sand (durchgezogen) und Lehm (gestrichelt)

Wird die Beziehung (4.2) in die Massenerhaltung (4.1) eingesetzt, erhält man die Richards-Gleichung

$$\frac{\partial \Theta(p)}{\partial t} \;-\; \mathrm{div}(K(p)\nabla p) \;=\; S, \tag{4.3}$$

eine quasi-lineare, parabolische Differentialgleichung.
Bei der formalen Untersuchung dieser Gleichung, wie z.B. in [52], werden die drei möglichen Sättigungsgrade innerhalb des Gebiets berücksichtigt: voll gesättigt, teilweise gesättigt, voll ungesättigt (voll gesättigt mit Luft). In [1] wird die Existenz und Eindeutigkeit einer Lösung von (4.3) unter Bedingungen an die nichtlinearen Funktionen Θ und K gezeigt. Hinreichende Bedingungen sind zum Beispiel $\Theta, K \in L^\infty(\mathbb{R}), \Theta \geq \overline{\Theta} > 0, K \geq 0$, die bei Verwendung der oben angegebenen Mualem/van Genuchten-Parametrisierungen erfüllt sind. Die Zeitabhängigkeit der Gleichung fordert dabei das Vorschreiben von Anfangsbedingungen für die Druckvariable p zum Zeitpunkt $t = 0$.

Zur numerischen Approximation der Zeitableitung wird die implizite Euler-Diskretisierung aus Abschnitt 1.2 verwendet: Ausgehend von einer Zerlegung $\{t_i\}_{i=1}^{n_t}$ des Zeitintervalls $[0,T]$ und der gegebenen Näherung $p(\cdot,t_i)=p_i$ wird p zum Zeitpunkt t_{i+1} berechnet. Aus diesem Verfahren erhält man als zeitdiskrete Gleichung

$$\begin{aligned}
\frac{\Theta(p_{i+1})-\Theta(p_i)}{t_{i+1}-t_i} - \operatorname{div}(K(p_{i+1})\nabla p_{i+1}) &= S \\
\Leftrightarrow \Theta(p_{i+1}) - \underbrace{(t_{i+1}-t_i)}_{=:\tau_i}\operatorname{div}(K(p_{i+1})\nabla p_{i+1}) &= \underbrace{(t_{i+1}-t_i)S+\Theta(p_i)}_{=:f_i} \\
\Leftrightarrow \Theta(p_{i+1}) - \tau_i \operatorname{div}(K(p_{i+1})\nabla p_{i+1}) &= f_i. \qquad (4.4)
\end{aligned}$$

Alternative Zeitdiskretisierungen können durch Integration bezüglich der Zeit wie in [2] erhalten werden.
Die zu lösende Formulierung (4.4) ist wegen der oben angegebenen Voraussetzungen für die nichtlinearen Funktionen stets eine quasi-lineare elliptische Gleichung zweiter Ordnung. Damit diese Gleichung eine eindeutige Lösung besitzt, müssen auf zumindest einem Teil des Randes Γ des Gebiets Ω Dirichlet-Randbedingungen vorgeschrieben werden. Lässt sich der Rand gemäß

$$\Gamma = \Gamma_D \cup \Gamma_N$$

in zwei disjunkte Teile zerlegen und gilt $\Gamma_D \neq \emptyset$, so ist die Aufgabe

Sei $p_0(\cdot)=p(\cdot,0)$ gegeben.
Für $i=0,1,2,\ldots,n_t-1$ berechne p_{i+1} aus:

$$\left\{\begin{array}{rcll}
\Theta(p_{i+1}) - \tau_i \operatorname{div}(K(p_{i+1})\nabla p_{i+1}) &=& f_i & \text{in } \Omega \\
p_{i+1}(\cdot) &=& g(\cdot,t_{i+1}) & \text{auf } \Gamma_D \\
n\cdot(K(p_{i+1})\nabla p_{i+1}) &=& h(\cdot,t_{i+1}) & \text{auf } \Gamma_N
\end{array}\right. \qquad (4.5)$$

eindeutig lösbar.
Häufig ist der sich einstellende Fluss des Fluids von größerem Interesse als die aus mehreren Anteilen zusammengesetzte Mischvariable p. Dieser Fluss u könnte dabei aus der Lösung von (4.5) gemäß (4.2) näherungsweise berechnet werden. Um den dadurch entstehenden Fehler zu vermeiden, bietet sich die Möglichkeit an, mit Hilfe eines gemischten Ansatzes beide Unbekannte u und p gleichzeitig zu berechnen. Dazu wird (4.5) in ein System von Gleichungen erster Ordnung umgeschrieben und man erhält folgende Aufgabe zur gleichzeitigen Berechnung von u und p:

Sei $p_0(\cdot)=p(\cdot,0)$ gegeben.
Für $i=0,1,2,\ldots,n_t-1$ berechne p_{i+1} und u_{i+1} aus:

$$\left\{\begin{array}{rcll}
\Theta(p_{i+1}) + \tau_i \operatorname{div}(u_{i+1}) &=& f_i & \text{in } \Omega \\
u_{i+1} + K(p_{i+1})\nabla p_{i+1} &=& 0 & \text{in } \Omega \\
p_{i+1}(\cdot) &=& g(\cdot,t_{i+1}) & \text{auf } \Gamma_D \\
n\cdot u_{i+1} &=& -h(\cdot,t_{i+1}) & \text{auf } \Gamma_N.
\end{array}\right. \qquad (4.6)$$

Offensichtlich hat das Vorschreiben eines Normalenflusses auf einem Teil des Randes des Gebiets auch eine physikalische Bedeutung und wird in der Aufgabe durch eine wesentliche Randbedingung an u modelliert. Wegen der eindeutigen Lösbarkeit von (4.5) ist auch (4.6) eindeutig lösbar.

4.1.2 Numerische Behandlung der Aufgabe

Das hergeleitete System partieller Differentialgleichungen (4.6) soll numerisch gelöst werden. Dazu wird eine Variationsformulierung wie in 1.3.1 unter Verwendung des Ausgleichsfunktionals

hergeleitet. Dabei wird in diesem Abschnitt von einem festen Zeitpunkt t_i ausgegangen und wegen Vereinfachung der Schreibweise $\tau_i = \tau$ und $f_i = f$ gesetzt.
Für die obige Aufgabe (4.6) ergibt sich aus Differenzierbarkeitsgründen $H^1(\Omega)$ als Lösungsraum für p und $H(\mathrm{div}, \Omega)$ als Lösungsraum für u, sowie $f \in L^2(\Omega)$. Für die Berücksichtigung der Randbedingungen definiere $p^D \in H^1(\Omega)$ und $u^N \in H(\mathrm{div}, \Omega)$ so, dass

$$\begin{aligned} p^D(\cdot) &= g(\cdot, t_{i+1}) \text{ auf } \Gamma_D \\ n \cdot u^N &= -h(\cdot, t_{i+1}) \text{ auf } \Gamma_N \end{aligned}$$

gilt. Werden die Räume

$$Q := \{q \in H^1(\Omega) : q = 0 \text{ auf } \Gamma_D\} \tag{4.7}$$
$$V := \{v \in H(\mathrm{div}, \Omega) : n \cdot v = 0 \text{ auf } \Gamma_N\} \tag{4.8}$$

eingeführt, so kann die Lösung in der Form $(p_{i+1}, u_{i+1}) = (p^D + \hat{p}, u^N + \hat{u})$ mit $\hat{p} \in Q, \hat{u} \in V$ geschrieben werden.
Analog zu der Bezeichnungsweise in den vorhergehenden Kapiteln wird $H := H^1(\Omega) \times H(\mathrm{div}, \Omega)$ gesetzt und der Operator $\mathcal{R}$ definiert über

$$\begin{aligned} \mathcal{R} \,:\, H &\to (L^2(\Omega))^3 \\ (p, u) &\mapsto \begin{pmatrix} \tau \,\mathrm{div}\, u + \Theta(p) - f \\ u + K(p)\nabla p \end{pmatrix}, \end{aligned} \tag{4.9}$$

womit die Aufgabe (4.6) äquivalent damit ist, eine Nullstelle von $\mathcal{R}$ unter Berücksichtigung der Randbedingungen zu finden.
Das Ausgleichsfunktional wird definiert durch

$$\begin{aligned} \mathcal{F}(p^D + \hat{p}, u^N + \hat{u}) &= \|\mathcal{R}(p^D + \hat{p}, u^N + \hat{u})\|^2_{0,\Omega} \\ &= \|\Theta(p^D + \hat{p}) + \tau \,\mathrm{div}(u^N + \hat{u}) - f\|^2_{0,\Omega} \\ &\quad + \|(u^N + \hat{u}) + K(p^D + \hat{p})\nabla(p^D + \hat{p})\|^2_{0,\Omega} \geq 0. \end{aligned} \tag{4.10}$$

Wegen der eindeutigen Lösbarkeit der ursprünglichen Aufgabe gilt

$$\begin{aligned} 0 &= \min_{(\hat{p},\hat{u}) \in Q \times V} \mathcal{F}(p^D + \hat{p}, u^N + \hat{u}) \\ &= \mathcal{F}(p_*, u_*)\,. \end{aligned}$$

Es bezeichne also (p_*, u_*) die eindeutige Lösung. Die neue nichtlineare Minimierungsaufgabe lautet dann

$$\mathcal{F}(p^D + \hat{p}, u^N + \hat{u}) \to \min_{(\hat{p},\hat{u}) \in Q \times V} !\,. \tag{4.11}$$

Dieses Minimierungsproblem wird nun mit Hilfe der drei in den Kapiteln 2 und 3 dargestellten Multilevelverfahren gelöst. Dafür wird zunächst die Fréchet-Ableitung von $\mathcal{R}$ bereitgestellt. Man findet

$$\mathcal{J}(p, u) \begin{pmatrix} \hat{q} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} \tau \,\mathrm{div}\, \hat{v} + \Theta'(p)\hat{q} \\ \hat{v} + K'(p)\nabla p\, \hat{q} + K(p)\nabla \hat{q} \end{pmatrix} \tag{4.12}$$

und es gilt die Entwicklung

$$\mathcal{R}(p + \hat{\delta}_p, u + \hat{\delta}_u) = \mathcal{R}(p, u) + \mathcal{J}(p, u) \begin{pmatrix} \hat{\delta}_p \\ \hat{\delta}_u \end{pmatrix} + O(\|(\hat{\delta}_p, \hat{\delta}_u)\|^2_{\mathrm{H}})$$

mit der Norm $\|(p, u)\|^2_{\mathrm{H}} = \|p\|^2_{1,\Omega} + \|u\|^2_{\mathrm{div},\Omega}$.
Für die Verfahren aus Kapitel 2 werden zunächst die Räume Q und V geeignet diskretisiert.

Wegen der Ausgleichsformulierung braucht dabei keine besondere Bedingung wie die inf-sup-Bedingung für gemischte FE-Formulierungen erfüllt werden. Daher können die einfachsten konformen FE-Räume als endlichdimensionale Unterräume für Q und V ausgewählt werden.
Für $Q \subseteq H^1(\Omega)$ sind dies die üblichen stückweise linearen stetigen "Hut"-Funktionen über einer Triangulierung

$$\mathcal{T}_l \;=\; \{T_i : i = 1, \ldots, n_T, T_i \subseteq \Omega \text{ ist ein Dreieck}\}$$

mit $\bigcup_{i=1}^{n_T} T_i \;=\; \overline{\Omega}$.
Als einfachste Diskretisierung für $H(\mathrm{div}, \Omega)$ werden die Raviart-Thomas-Elemente niedrigster Ordnung verwendet (siehe [40]), deren Freiheitsgrade mit den Kanten der Triangulierung $\mathcal{T}$ assoziiert sind. Bezeichne also

$$Q \supset Q_l \;=\; \{q \in Q : q|_{T_i} \text{ linear für alle } T_i \in \mathcal{T}_l\} \tag{4.13}$$

$$V \supset V_l \;=\; \{v \in V : v|_{T_i} = \begin{pmatrix} \alpha + \gamma x \\ \beta + \gamma y \end{pmatrix}, \alpha, \beta, \gamma \in \mathbb{R}, T_i \in \mathcal{T}_l\} \tag{4.14}$$

die Unterräume der Lösungsräume Q und V mit l als Level-Parameter.

Die Variationsformulierung für das DLL-Verfahren

Für die Anwendung von linearen Multilevelverfahren auf das diskrete nichtlineare Minimierungsproblem

$$\mathcal{F}(p^D + p_l, u^N + u_l) \;\rightarrow\; \min_{(p_l, u_l) \in Q_l \times V_l} !$$

wird zunächst innerhalb des Funktionals linearisiert (siehe (2.12)). Nach dem Aufstellen der Normalengleichungen (siehe (2.13)) für diese lineare Minimierungsaufgabe erhält man die in Gleichung (2.22) zu lösende lineare Variationsformulierung

$$\begin{aligned} &\text{Berechne } (\hat{\delta}_p, \hat{\delta}_u) \in Q_l \times V_l \text{ mit} \\ &(\mathcal{J}(p,u) \begin{pmatrix} \hat{\delta}_p \\ \hat{\delta}_u \end{pmatrix}, \mathcal{J}(p,u) \begin{pmatrix} \hat{q}_l \\ \hat{v}_l \end{pmatrix})_{0,\Omega} \;=\; -(\mathcal{R}(p,u), \mathcal{J}(p,u) \begin{pmatrix} \hat{q}_l \\ \hat{v}_l \end{pmatrix})_{0,\Omega} \\ &\text{für alle } (\hat{q}_l, \hat{v}_l) \in Q_l \times V_l. \end{aligned} \tag{4.15}$$

Hier steht (p, u) für die aktuelle Näherung und $(\hat{\delta}_p, \hat{\delta}_u)$ für die zu berechnende Korrektur an die aktuelle Näherung (wobei die Randbedingungen bereits in (p, u) enthalten sein sollen).
Die weiteren Schritte innerhalb von Algorithmus 2 folgen analog.

Die Variationsformulierung für das DNL-Verfahren

Für die Verwendung von nichtlinearen Multilevelverfahren ist das nichtlineare Variationsproblem

$$\begin{aligned} &\text{Berechne } (p_{l,*}, u_{l,*}) \text{ in } Q_l \times V_l \text{ mit} \\ &(\mathcal{R}(p^D + p_{l,*}, u^N + u_{l,*}), \mathcal{J}(p^D + p_{l,*}, u^N + u_{l,*}) \begin{pmatrix} \hat{q}_l \\ \hat{v}_l \end{pmatrix})_{0,\Omega} \;=\; 0 \\ &\text{für alle } (\hat{q}_l, \hat{v}_l) \in Q_l \times V_l \end{aligned}$$

der Ausgangspunkt (siehe (2.23)). Entsprechend den Herleitungen in Abschnitt 2.3 erhält man zwei verschiedene nichtlineare Grobgitterkorrekturgleichungen durch Anwendung des FAS-Schemas auf die nichtlinearen Normalengleichungen bzw. auf die ursprüngliche Differentialgleichung. Der Unterschied besteht dabei darin, dass bei dem letzteren Verfahren die rechte Seite der

Grobgitterkorrekturgleichung auch nichtlinear von der gesuchten Grobgitterkorrektur abhängt, während dies bei Verwendung von FAS für die Normalengleichungen nicht der Fall ist. Für die Details der Gleichungen siehe Abschnitt 2.3 bzw. [35].
Die Restriktionsabbildung $\mathcal{I}_{j+1}^{j}$ wird über den einzelnen Räumen zusammengesetzt. Für die Freiheitsgrade in Q_{j+1} wird die $L(\Omega)^2$-Adjungierte der Standard-Prolongation verwendet. Für die kantenbezogenen Freiheitsgrade im Raum der Raviart-Thomas-Elemente niedrigster Ordnung wird zur Interpolation der Wert der Normalenableitung der feineren Funktion auf den groben Kanten gemittelt.
In Abschnitt 4.2.1 werden numerische Ergebnisse zum Vergleich der beiden FAS-Ansätze angegeben.

Die Variationsformulierung für das LDL-Verfahren

Die zu lösende Variationsformulierung hat für das LDL-Verfahren genau die gleiche Form wie oben (4.15). Lediglich die Abbruchkriterien und der Dämpfungsalgorithmus werden zusätzlich wie in den Algorithmen 3.1 und 3.2 angegeben implementiert.

Damit stehen die notwendigen Variationsformulierungen zur Verfügung und die drei Algorithmen können auf das Problem angewendet werden. Als Beispielaufgabe dient ein Bewässerungsversuch nach [47]:

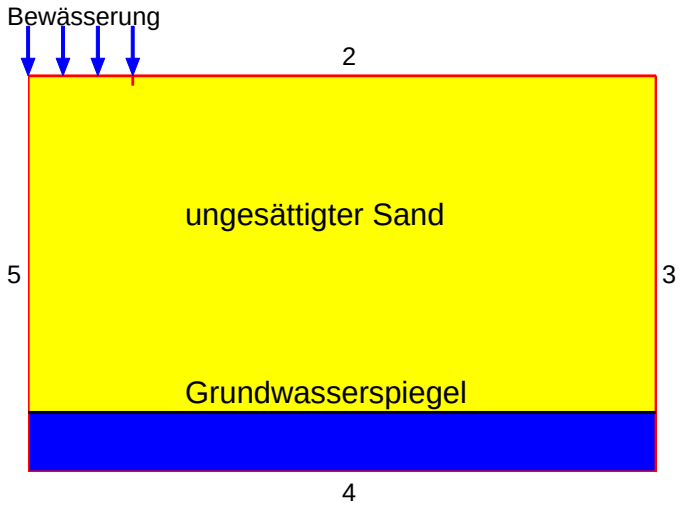


Abbildung 4.2: Beispielversuch

Das Simulationsgebiet ist ein zweidimensionaler Bodenausschnitt der Breite 3 m und Tiefe 2 m. Zu Beginn der Simulation ($t = 0$) ist das Gebiet oberhalb des Grundwasserspiegels bei -1.7 m nahezu völlig ungesättigt. Dies wird durch die Anfangsbedingung für p

$$p_0(x) \equiv -1.7$$

modelliert. Der Rand des Gebiets teilt sich in $\Gamma_D \stackrel{\wedge}{=}$ Rand 3 und $\Gamma_N = \partial\Omega \setminus \Gamma_D$. Am rechten Rand des Gebietes (Rand 3) sollen für alle Zeiten konstante Dirichlet-Bedingungen für p vorgeschrieben sein, die den Anfangsbedingungen entsprechen ($g \equiv -1.7$).
Auf dem Rest des Randes werden Neumann-Randbedingungen für p, d.h. wesentliche Randbedingungen für u vorgeschrieben: An den Rändern 2,4 und 5 wird der Normalenfluss auf 0 gesetzt (d.h. über diese Ränder findet kein Zu-oder Abfluss von Wasser statt.) Am oberen linken Rand

des Gebiets wird ein konstanter Fluss (Bewässerung) von 0.148 m/h für u vorgeschrieben. Für die Zeitdiskretisierung wird eine äquidistante Zerlegung mit $\tau_i = \tau = 0.05$ Stunden verwendet.

Durch die Bewässerung bildet sich am oberen linken Rand eine nach unten durchdringende Wasserfront aus. Im Bereich des Übergangs zwischen (fast völlig) gesättigtem und (noch) ungesättigtem Gebiet verändern sich die Koeffizienten der Differentialgleichung entsprechend ihren nichtlinearen Abhängigkeiten relativ stark. Um dieses Übergangsgebiet gut aufzulösen und eine gute Approximation der Lösung in diesem Bereich zu erzielen, wird das Funktional $\mathcal{F}$ als Fehlerschätzer für die erzielte Näherung verwendet, um die gröbste Ausgangstriangulierung $\mathcal{T}_0$ und die darauf folgenden Triangulierungen im Bereich der Wasserfront adaptiv zu verfeinern. Im Anschluss an die Berechnung einer Näherung nach der 5. Verfeinerung, also über der Triangulierung $\mathcal{T}_5$, wird zum nächsten Zeitschritt wieder auf Level 0 übergegangen.
Wie in Abschnitt 2.2.1 vorgeschlagen und in den Algorithmen der drei Multilevelverfahren zu sehen, wird in den Verfahren ein Full-Multigrid-Zyklus verwendet, um geeignete Startnäherungen für die feineren Triangulierungen bereitzustellen.

4.1.3 Nachweis der Voraussetzungen aus Kapitel 3

Um die theoretischen Ergebnisse des Gauß-Newton-Multilevelverfahrens (LDL) auch im Beispielfall verwenden zu können, müssen für das Funktional $\mathcal{F}$ bzw. die zu Grunde liegende Differentialgleichung $\mathcal{R}(p,u) = 0$ die in Abschnitt 2.1 und in den Sätzen in Kapitel 3 verwendeten Voraussetzungen nachgewiesen werden.

Zunächst sind die Abschätzungen (2.2) und (2.3) für $\mathcal{J}$ zu zeigen. Wird die vereinfachende Umformung

$$\mathcal{J}(p,u)\begin{pmatrix}\hat{q}\\ \hat{v}\end{pmatrix} = \begin{pmatrix}\tau \operatorname{div} \hat{v} + \Theta'(p)\hat{q}\\ \hat{v} + K'(p)\nabla p\,\hat{q} + K(p)\nabla\hat{q}\end{pmatrix} = \begin{pmatrix}\tau \operatorname{div} \hat{v} + \Theta'(p)\hat{q}\\ \hat{v} + \nabla(K(p)\hat{q})\end{pmatrix} \tag{4.16}$$

verwendet, so ergeben sich diese Abschätzungen als unmittelbare Folgerung aus dem Hauptergebnis von [13]. Um die dazu notwendigen Voraussetzungen zu erfüllen, werden die Parametrisierungen von $\Theta(p)$ und $K(p)$ leicht abgeändert (siehe [23] und [41] für Details).

Für die Anwendbarkeit des Gauß-Newton-Verfahrens reicht die lokale Lipschitz-Stetigkeit von $\mathcal{J}$ aus. Da $K(p)$ und $\Theta(p)$ jeweils beliebig oft stetig differenzierbar und beschränkte Funktionen sind, ist $\mathcal{J}$ auf ganz H stetig differenzierbar und damit lokal Lipschitz-stetig. Da im Verfahren Abstiegsrichtungen konstruiert werden, wird die durch die Startnäherung x_0 bestimmte Umgebung

$$D = \{y \in H : \mathcal{F}(y) \leq \mathcal{F}(x_0)\}$$

um die Lösung x_* ($\mathcal{F}(x_*) = 0$!) von den konstruierten Iterierten x_l nicht verlassen, so dass die Einschränkung von Voraussetzung 3.2 auf diese Umgebung D für die Anwendbarkeit des Verfahrens ausreicht.

Die Abhängigkeit des Diskretisierungsfehlers von der Feinheit h_l der Zerlegung $\mathcal{T}_l$ nach (3.18) ist bei dem Beispielproblem für die verwendeten FE-Elemente (siehe (4.13) und (4.14)) im Zusammenhang mit der Differentialgleichung (4.6) zu untersuchen.
In [25] findet man, dass für ein $\alpha \in (0,1]$ $p \in H^{1+\alpha}(\Omega)$ liegt. Daraus folgt mit der zweiten Gleichung in (4.6), dass $u \in (H^\alpha(\Omega))^2$ liegt. Mit Hilfe der ersten Gleichung gilt mit $f \in H^\alpha(\Omega)$ auch $\operatorname{div} u \in H^\alpha(\Omega)$. Unter diesen Voraussetzungen existieren die Projektionen $I_l : H^1(\Omega) \to Q_l$ und $R_l : H(\operatorname{div},\Omega) \to V_l$ für die oben angegebenen FE-Räume. Für diese Projektionen wird in Anhang B die Abschätzung

$$\|(p - I_l\,p, u - R_l\,u)\|_{\mathrm{H}} \leq C\,h_l^\alpha\,\|(p,u)\|_{\mathrm{H},\alpha}$$

gezeigt, wobei

$$\|(p,u)\|^2_{\mathrm{H},\alpha} \;=\; \|p\|^2_{1+\alpha,\Omega} \;+\; \|u\|^2_{\alpha,\Omega} \;+\; \|\mathrm{div}\; u\|^2_{\alpha,\Omega}$$

gilt. Zusammen mit dem Céa-Lemma (1.64) und der Stetigkeit von $\mathcal{J}$ (2.2) ist der Diskretisierungsfehler dann bei genügend kleiner Wahl von h_l beliebig klein.

Schließlich ist für die Eindeutigkeit des Minimums von $\mathcal{F}$ im Raum H_l notwendig, dass das Funktional zumindest in einer Umgebung D der Lösung (p_*, u_*) gleichmäßig elliptisch ist. Dazu wird die zweite Ableitung von $\mathcal{F}$ untersucht. Es gilt

$$\begin{aligned}\frac{\partial^2\mathcal{F}(p,u)}{\partial(q_1,v_1)\partial(q_2,v_2)} &= (\mathcal{R}(p,u), \mathcal{J}'(p,u)\Big[\begin{pmatrix}q_1\\v_1\end{pmatrix}, \begin{pmatrix}q_2\\v_2\end{pmatrix}\Big])_{0,\Omega}\\ &\quad +(\mathcal{J}(p,u)\begin{pmatrix}q_1\\v_1\end{pmatrix}, \mathcal{J}(p,u)\begin{pmatrix}q_2\\v_2\end{pmatrix})_{0,\Omega}\end{aligned} \tag{4.17}$$

mit

$$\mathcal{J}'(p,u)\Big[\begin{pmatrix}q_1\\v_1\end{pmatrix}, \begin{pmatrix}q_2\\v_2\end{pmatrix}\Big] \;=\; \begin{pmatrix}\Theta''(p)\, q_1\, q_2\\ K'(p)(\nabla q_1\cdot q_2 + \nabla q_2\cdot q_1) + K''\nabla p\cdot q_1\, q_2\end{pmatrix}. \tag{4.18}$$

Die gleichmäßige Elliptizität in D folgt nach [15] aus

$$\frac{\partial^2\mathcal{F}(p,u)}{\partial(q,v)^2} \;\geq\; C\,\|(q,v)\|^2_{\mathrm{H}} \tag{4.19}$$

mit $C>0$ für alle $(p,u)\in D$ und $(q,v)\in H$.
Wegen (2.3) gilt bereits

$$(\mathcal{J}(p,u)\begin{pmatrix}q\\v\end{pmatrix}, \mathcal{J}(p,u)\begin{pmatrix}q\\v\end{pmatrix})_{0,\Omega} \;\geq\; \underline{\alpha}^2\|(q,v)\|^2_{\mathrm{H}}\,.$$

Sei $D_1 = \{(p,u)\in H \,:\, \|(p-p_*, u-u_*)\|_{\mathrm{H}} \leq \rho\}$ eine beliebige beschränkte Umgebung um (p_*,u_*). Die verwendeten Parametrisierungen von Θ und K sind über ihrem gesamten Definitionsbereich unendlich oft differenzierbar und ihre Ableitungen sind beschränkt (z.B. durch M). Damit erhält man für die Abschätzung des ersten Teils in (4.17) für $(p,u)\in D_1$

$$\begin{aligned}&(\mathcal{R}(p,u), \mathcal{J}'(p,u)\Big[\begin{pmatrix}q\\v\end{pmatrix}, \begin{pmatrix}q\\v\end{pmatrix}\Big])_{0,\Omega}\\
&\qquad\geq -\|\mathcal{R}(p,u)\|_{0,\Omega}\|\mathcal{J}'(p,u)\begin{pmatrix}q\\v\end{pmatrix}\begin{pmatrix}q\\v\end{pmatrix}\|_{0,\Omega}\\
&\qquad\geq -\|\mathcal{R}(p,u)\|_{0,\Omega}\big(\|\underbrace{\Theta''(p)}_{\leq M} q^2\|_{0,\Omega} + 2\|\underbrace{K'(p)}_{\leq M}\nabla q\cdot q\|_{0,\Omega} + \|\underbrace{K''(p)}_{\leq M}\nabla p\, q^2\|_{0,\Omega}\big)\\
&\qquad\geq -M\|\mathcal{R}(p,u)\|_{0,\Omega}\big(\underbrace{\|q^2\|_{0,\Omega}}_{\leq\|q\|^2_{0,\Omega}} + 2\;\underbrace{\|\nabla q\cdot q\|_{0,\Omega}}_{\leq\|\nabla q\|_{0,\Omega}\|q\|_{0,\Omega}} \;+\; \underbrace{\|\nabla p\, q^2\|_{0,\Omega}}_{\leq\|\nabla p\|_{0,\Omega}\|q\|^2_{0,\Omega}}\big)\\
&\qquad\geq -M\|\mathcal{R}(p,u)\|_{0,\Omega}\big(\|q\|^2_{0,\Omega} + \underbrace{2\|\nabla q\|_{0,\Omega}\|q\|_{0,\Omega}}_{\leq\|q\|^2_{0,\Omega}+\|\nabla q\|^2_{0,\Omega}} + \underbrace{\|\nabla p\|_{0,\Omega}}_{\text{beschränkt}}\|q\|^2_{0,\Omega}\big)\\
&\qquad\geq -C\|\mathcal{R}(p,u)\|_{0,\Omega}\,\|q\|^2_{1,\Omega}\\
&\qquad\geq -C\|\mathcal{R}(p,u)\|_{0,\Omega}\,\|(q,v)\|^2_{\mathrm{H}}\,.\end{aligned}$$

Dabei wurden alle auftretenden Konstanten in C zusammengefasst.
Wird diese Abschätzung mit der Vorhergehenden verknüpft, erhält man

$$\frac{\partial^2\mathcal{F}(p,u)}{\partial(q,v)^2} \;\geq\; (\underline{\alpha}^2 - C\|\mathcal{R}(p,u)\|_{0,\Omega})\,\|(q,v)\|^2_{\mathrm{H}}\,. \tag{4.20}$$

Da $\|\mathcal{R}(p_*, u_*)\|_{0,\Omega} = 0$ und $\mathcal{R}$ stetig ist, gibt es eine echte Umgebung $D = \{(p,u) \in D_1 : \|\mathcal{R}(p,u)\|_{0,\Omega} < \frac{\alpha^2}{C}\}$ von (p_*, u_*), in der die Bedingung (4.19) erfüllt ist und in der $\mathcal{F}$ damit gleichmäßig elliptisch ist.

4.2 Numerische Ergebnisse

Im Verlauf dieser Arbeit ist an mehreren Teilaspekten der Finite-Element-Lösung der Aufgabe (4.11) gearbeitet worden. Zunächst wurden nichtlineare Multilevelvarianten ausgearbeitet, dann zwei verschiedene Glätter für den $H(\mathrm{div}, \Omega)$-Anteil getestet. Schließlich folgte ein Vergleich der verschiedenen Abbruchkriterien. Die Darstellung der numerischen Ergebnisse dieser Arbeit soll nun dem gleichen Verlauf folgen. Es wird also zunächst aus den verschiedenen nichtlinearen Multilevelvarianten die wettbewerbsfähigste Methode ermittelt. Unter Verwendung beider Multilevelverfahren DNL und DLL werden dann die beiden Glätter für die $H(\mathrm{div}, \Omega)$-Probleme in Hinsicht auf das günstigste Aufwand/(Nutzen = Konvergenzrate)-Verhältnis verglichen.
Für den Vergleich der Abbruchkriterien wird dann die Performance des in Kapitel 3 entwickelten LDL-Verfahrens untersucht und schließlich mit den zwei anderen Methoden verglichen.

4.2.1 Varianten des nichtlinearen DNL-Verfahrens

Die innerhalb des nichtlinearen Multilevelverfahrens zu lösende lineare Variationsgleichung wurde in Abschnitt 2.3 hergeleitet. Durch die Anwendung des FAS-Schemas ([9]) auf verschiedene nichtlineare Ausgangsprobleme ergaben sich zwei Varianten (*FAS für die Normalengleichungen* und *FAS für den Operator* $\mathcal{R}$), die sich in der rechten Seite der Grobgitterkorrektur-Gleichung unterschieden. Diese nichtlinearen Multilevelverfahren werden nun auf das Beispielproblem angewendet und hinsichtlich Konvergenzrate und Aufwand verglichen.
Um auch den Effekt der Grobgitterkorrektur berücksichtigen zu können, wird der Vergleich mit einem Ein-Gitter-Verfahren (d.h. nur Glättung auf dem feinsten Gitter ohne Grobgitterkorrektur) vorgenommen.

Das Minimum von $\mathcal{F}$ ist über jedem endlich dimensionalen Teilraum H_l von H echt größer als Null. Bei der Konvergenz gegen die minimierende Funktion $(p_{l,*}, u_{l,*}) \in H_l$ ist somit eine Stagnation in der Reduktion des Funktionals von Schritt zu Schritt gemäß Abbildung 4.3 zu erwarten und auch tatsächlich zu beobachten. Da diese Stagnation stets bei Näherung an die

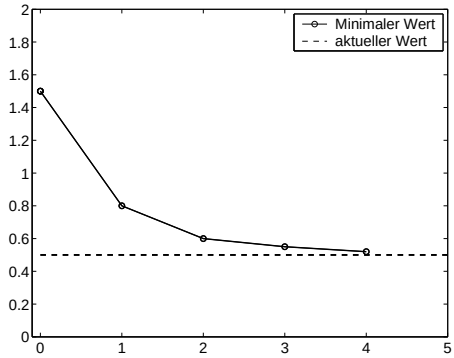


Abbildung 4.3: Reduktion des Funktionals über H_l

Lösung $(p_{l,*}, u_{l,*})$ zu beobachten ist, dient dies als erstes Abbruchkriterium. Ist $(p_{l,k}, u_{l,k})$ die

Näherung im k-ten Iterationsschritt so wird die äußerste Iteration dann abgebrochen, wenn

$$\frac{\mathcal{F}(p_{l,k}, u_{l,k})}{\mathcal{F}(p_{l,k-2}, u_{l,k-2})} \geq \rho_1 \tag{4.21}$$

mit einem $\rho_1 > 0$ gilt, z.B. mit $\rho_1 = 1 - 10^{-4}$. In (4.21) wird also die Reduktionsrate des Funktionals in zwei aufeinanderfolgenden Schritten gemessen und die Iteration abgebrochen, wenn die Reduktion des Funktionals zu gering ist. Ein ähnliches Abbruchkriterium wurde auch zum Ende von Abschnitt 2.2.2 vorgeschlagen.

Als Glättung der linearen Probleme wird ein nodaler Gauß-Seidel-Schritt für den Fehler in $H^1(\Omega)$ und ein Schritt des Block-Schwarz-Verfahrens wie in [3] für den Fehler in $H(\mathrm{div}, \Omega)$ eingesetzt. Unter Verwendung des Funktionals als Fehlerschätzer wird maximal fünfmal adaptiv verfeinert ($l_{\max}$=5).
Für $t = 3$ Stunden sind in Tabelle 4.1 die Anzahl der Freiheitsgrade auf jedem Level und der durch das Ein-Gitter-Verfahren erreichte Wert des Funktionals angegeben.

	$\dim V_l$	$\dim Q_l$	$\|\mathcal{R}(p^D + \hat{p}_l, u^N + \hat{u}_l)\|^2_{0,\Omega}$
$l = 0$	346	128	$2.26 * 10^{-4}$
$l = 1$	583	210	$1.01 * 10^{-4}$
$l = 2$	1105	388	$3.86 * 10^{-5}$
$l = 3$	2254	774	$1.39 * 10^{-5}$
$l = 4$	3936	1338	$5.46 * 10^{-6}$
$l = 5$	6053	2047	$2.87 * 10^{-6}$

Tabelle 4.1: Anzahl der Freiheitsgrade und Funktional nach drei Stunden

In Abbildung 4.4 ist das der Diskretisierung auf Level $l = 5$ zum Zeitpunkt $t = 3$ zu Grunde liegende Gitter zu sehen. Die einsickernde Wasserfront kann leicht an der oberen linken Ecke des Gebiets erkannt werden.

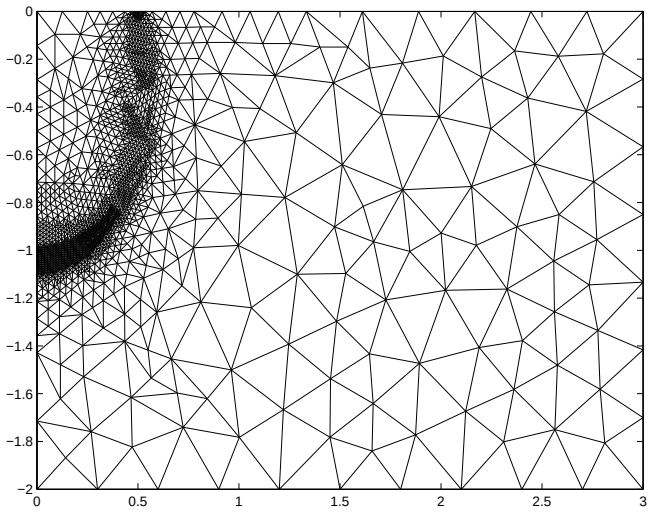


Abbildung 4.4: Adaptiv verfeinerte Triangulierung

Für den Vergleich der drei Verfahren ("Ein-Gitter" ohne Grobgitterkorrektur, "FAS-Normalgl." und "FAS-Operator") wird in Tabelle 4.2 die bis zum Abbruch notwendige Anzahl von Iterationsschritten angegeben.
Während die Anzahl der benötigten Ein-Gitter-Iterationen mit der Verfeinerung wächst, scheint

	Ein-Gitter	FAS-Normalgl.	FAS-Operator
$l = 0$	16	-	-
$l = 1$	7	6	5
$l = 2$	14	7	7
$l = 3$	35	11	11
$l = 4$	65	14	13
$l = 5$	66	13	12

Tabelle 4.2: Anzahl von Iterationsschritten

die benötigte Iterationsanzahl der Multilevelverfahren beschränkt zu sein.
Da das Abbruchkriterium auf Stagnation in der Reduktion von $\mathcal{F}$ beruht, zeigt Tabelle 4.2, dass das Minimum bei der Verwendung der Multilevelverfahren schnell erreicht wird, während das Ein-Gitter-Verfahren für die Dämpfung von Fehleranteilen, die zu gröberen Gittern gehören, mehr Iterationen benötigt.
Zu bedenken ist bei diesem Abbruchkriterium allerdings auch, dass ein Verfahren auch dann abgebrochen wird, wenn es nicht im Stande ist, weitere Reduktion zu liefern, obwohl vielleicht noch weitere Reduktion möglich wäre. Daher wird nun in Tabelle 4.3 der erreichte Funktionalwert für jede Methode in Prozent des erreichten Ein-Gitter-Funktionalwerts angegeben.

	Ein-Gitter	FAS-Normalgl.	FAS-Operator
$l = 1$	100	100.15	99.99
$l = 2$	100	100.19	100.20
$l = 3$	100	99.72	99.71
$l = 4$	100	98.79	98.80
$l = 5$	100	97.50	97.50

Tabelle 4.3: Anteil des erreichten Ein-Gitter-Funktionalwerts

Aus den Werten von Tabelle 4.3 wird deutlich, dass die Multilevelverfahren auf höheren Gittern bei Abbruch des Verfahrens tatsächlich ein geringeres Funktional erreichen. Entsprechend schlecht, nämlich bestenfalls gleich groß wie $\rho_1 = 0.999$, ist die Dämpfungsrate des Ein-Gitter-Verfahrens für die Grobgitter-Fehleranteile.
Eine weitere Folgerung aus Tabelle 4.3 ist, dass das Abbruchkriterium modifiziert werden sollte, denn es führt offensichtlich nicht zu einer für alle Verfahren gleichmäßig hohen Reduktion des Fehlerfunktionals. Die erhaltenen numerischen Ergebnisse für dieses Abbruchkriterium zeigen deutlich, dass auf den gröberen Gittern Fehleranteile nicht gleichmäßig gut durch alle Verfahren geglättet werden. Dies legt als Modifikation des Abbruchkriteriums nahe, gerade diese übrigbleibenden Fehleranteile, zusammengefasst in der Norm der rechten Seite der linearisierten Korrekturgleichung (siehe Abschnitt 2.3.2), zu messen und eine Mindestreduktion von $\rho_2 = 1e - 4$ dieser Norm für den Abbruch des Verfahrens zu verlangen. Dieses Abbruchkriterium führt zu den Iterationszahlen in Tabelle 4.4. In dieser Tabelle ist das minimal erreichte Funktional auf Level $l = 5$ mit angegeben.
Mit diesem Abbruchkriterium wird ein gleichmäßig kleines Funktional $\mathcal{F}$ erreicht (vgl. Tabelle 4.1) und damit ist dieses Abbruchkriterium tatsächlich für das Minimierungsproblem (4.11) geeignet. Weiterhin bemerkt man den deutlichen Anstieg in der Anzahl benötigter Iterationsschritte für das Ein-Gitter-Verfahren, während die Multilevelverfahren tatsächlich eine unabhängig vom Gitter beschränkte Zahl von Multilevelschritten für die Reduktion benötigen. Dieses Ergebnis ist auch deutlich in Abbildung 4.5 zu sehen, wo das erreichte Residuum gegen die Anzahl von Iterationen für jeweils zwei Varianten dargestellt ist. Die Unstetigkeitsstelle markiert jeweils den Start der Iteration auf dem nächstfeineren Gitter.

	Ein-Gitter	FAS-Normalgl.	FAS-Operator
$l = 0$	7	-	-
$l = 1$	18	13	13
$l = 2$	37	10	10
$l = 3$	81	11	11
$l = 4$	122	11	12
$l = 5$	158	10	11
$\mathcal{F}_{min} \cdot 10^6$	2.81	2.80	2.80

Tabelle 4.4: Anzahl von Iterationsschritten, zweites Abbruchkriterium

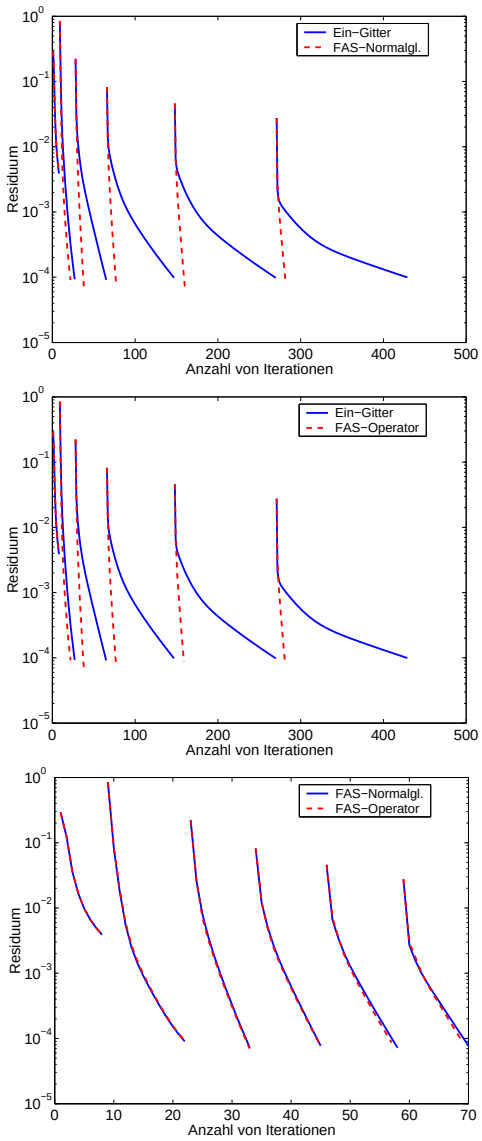


Abbildung 4.5: Residuum gegen Iterationen

Aus den Ergebnissen in den Tabellen 4.2 bis 4.4 bzw. Abbildung 4.5 läßt sich ablesen, dass die in Abschnitt 2.3.1 hergeleiteten Grobgitterkorrekturgleichungen für beide Varianten zu funktionierenden Multilevelverfahren mit gitterunabhängiger Iterationszahl führen. Nun kann im verbleibenden Teil dieses Abschnitts der Aufwand beider FAS-Varianten verglichen werden, um

das schnellere Verfahren herauszufinden.

Bei die diesbezüglichen Überlegungen in Abschnitt 2.3.1 ist schon deutlich geworden, dass die FAS-Variante für den Operator $\mathcal{R}$ wegen der aufwendiger aufzustellenden rechten Seite der Grobgittergleichung wahrscheinlich mehr Aufwand nach sich zieht. Wird die durchschnittliche Zeit je Multileveliteration für beide Verfahren ermittelt, ergibt sich Tabelle 4.5.

	FAS-Normalgl.	FAS-Operator
$l = 1$	7.867	11.654
$l = 2$	17.240	27.100
$l = 3$	36.382	64.038
$l = 4$	69.656	131.855
$l = 5$	127.05	256.194

Tabelle 4.5: Sekunden je Multileveliteration

Ähnlich zu Abbildung 4.5 wird in Abbildung 4.6 für beide FAS-Varianten das erreichte Residuum durch alle Level gegen die Zeit in Sekunden dargestellt.

Durch die in Tabelle 4.5 und Abbildungen 4.5 und 4.6 angegebenen numerischen Ergebnis-

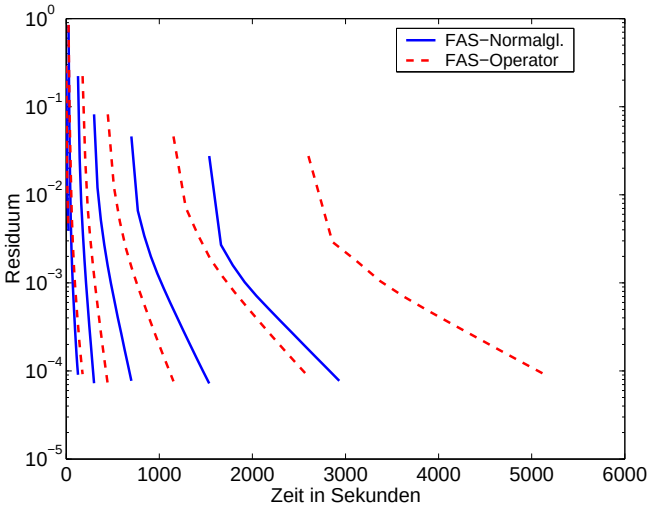


Abbildung 4.6: Residuum gegen Zeit in Sekunden

se wird deutlich, dass der leichte Vorteil, den das FAS-Operator-Verfahren in der Reduktion des Residuums je Iteration erzielt, durch den ungleich höheren Aufwand je Iteration nicht gerechtfertigt ist. Im Gesamtvergleich erzielt so das FAS-Normalgl.-Verfahren wesentlich bessere Konvergenzraten pro Sekunde. Im Vergleich in Abschnitt 4.2.4 wird daher diese FAS-Variante für das DNL-Verfahren verwendet.

4.2.2 Vergleich der Glättungsverfahren aus Abschnitt 1.4.2

Für Testfunktionen aus dem Raum V_l besitzt die Steifigkeitsmatrix zum linearen Variationsproblem (4.15) mit immer feiner werdender Triangulierung immer näher an Null liegende Eigenwerte. Verursacht wird dies durch die divergenzfreien Anteile einer diskreten Helmholtz-Zerlegung von V_l. Zur Behandlung dieser Fehleranteile muss ein geeigneter Glätter eingesetzt werden, wenn dennoch levelunabhängige Konvergenzraten erzielt werden sollen.

Aus der Literatur sind zwei Verfahren bekannt: Ein Block-Gauß-Seidel-Verfahren (siehe [3]) entsprechend der additiven oder multiplikativen Schwarz-Methode ("Projektions-Verfahren"(proj.verf.)) oder die Glättung der divergenzfreien Anteile in deren Potentialraum nach [29] ("Potentialraum-Verfahren"(pot.verf.)), was ausführlich in Abschnitt 1.4.2 vorgestellt wurde.
Für Block-Verfahren ist bekannt, dass die Konvergenzrate umso günstiger ist, je größer die Überschneidungen der gewählten Blöcke sind, während ein nodales Verfahren wie bei der Potentialraumkorrektur weniger Aufwand je Iteration verwendet. Es ist also ein Vergleich dieser Verfahren notwendig, um das effektiv schnellere Verfahren herauszufinden und dieses beim abschließenden Vergleich aller Multilevelverfahren verwenden zu können.
Zuerst wird dieser Vergleich für ein echt lineares Problem in $H(\mathrm{div}, \Omega)$ durchgeführt. Danach werden die Glättungsverfahren auch für das DLL-und das DNL-Verfahren zur Lösung des nichtlinearen Beispielproblems verglichen.

Als lineares Problem wird die Aufgabe

$$u - \triangle u = f$$

durch ein lineares Multilevelverfahren in $H(\mathrm{div}, \Omega)$ gelöst. In Abbildung 4.7 ist die Norm des Residuums gegen die Iterationsanzahl auf dem dritten, vierten und fünften Gitter (von links nach rechts) für diese Verfahren angegeben. Dabei entsprechen die gestrichelten Kurven mit "1", "2" und "3" dem Potentialraum-Verfahren mit ein, zwei oder drei nodalen Gauß-Seidel-Schritten je Glättungsiteration, während die durchgezogenen Linien die erzielten Werte des Projektions-Verfahrens angeben.

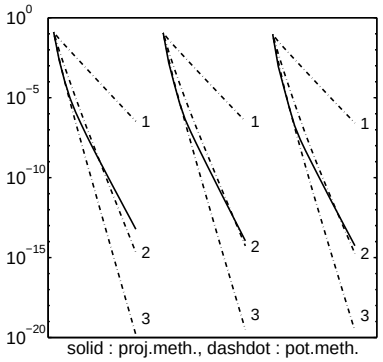


Abbildung 4.7: Konvergenz der Verfahren für das lineare Beispielproblem

Es ist deutlich zu sehen, dass die Konvergenzrate für alle Verfahren unabhängig von der Anzahl der Unbekannten ist.
Um den Aufwand der Verfahren zu vergleichen, wird die Kennzahl $k_m = r^{\frac{m \cdot 1e9}{f}}$ eingeführt, die mit r = Konvergenz je Iteration und f = Anzahl der Flops (= Gleitkommaoperationen) je Iteration ein Maß für die Konvergenzrate je m Milliarden Flops angibt. In Tabelle 4.6 wird die Maßzahl k_{30} für die verschiedenen Verfahren angegeben.
In der Zeile "# DOF" ist die Anzahl der Freiheitsgrade auf den jeweiligen Gittern abzulesen. Die 1, 2 oder 3 hinter "pot.verf." bezeichnet die Anzahl der verwendeten nodalen Gauß-Seidel-Iterationen im Potentialraum.
An Tabelle 4.6 kann deutlich erkannt werden, dass für das lineare Problem das Potentialraum-Verfahren deutlich schneller ist, wobei zwei oder drei nodale Gauß-Seidel-Iterationen je Glättungsschritt verwendet werden sollten.

Für das nichtlineare Beispielproblem wird nun das DLL-Verfahren angewendet und die Konvergenzrate je Iteration verglichen. Zu beachten ist dabei, dass diese Konvergenzrate sich auf die

Level	3	4	5
# DOF	1095	1716	3128
proj.verf.	0.118	0.358	0.685
pot.verf.1	0.101	0.325	0.645
pot.verf.2	0.028	0.203	0.552
pot.verf.3	0.028	0.206	0.573

Tabelle 4.6: k_{30} beim Einsatz der Glättungsverfahren für das lineare Problem

Konvergenz des nichtlinearen Verfahrens bezieht. Daher werden die mit dem Multilevelverfahren erzielten Konvergenzraten gegen diejenigen Konvergenzraten verglichen, die man bei exakter Lösung der linearen Gleichungssysteme für die nichtlineare Iteration erhalten hätte. Der Aufwandsvergleich liefert dabei für das "proj.verf." etwa dreimal mehr Flops wie für das "pot.verf.2" und zweimal mehr Flops wie für das "pot.verf.3".

Level	3	4	5
# DOF	1276 + 450	2115 + 737	3986 + 1374
exakte Lösung	0.517	0.465	0.371
proj.verf.	0.720	0.680	0.637
pot.verf.2	0.759	0.713	0.693
pot.verf.3	0.704	0.640	0.607

Tabelle 4.7: Konvergenzraten für das nichtlineare Problem (DLL)

Auch aus den in Tabelle 4.7 angegebenen Ergebnissen lässt sich sofort erkennen, dass das Potentialraum-Verfahren das schnellere der beiden Verfahren ist, denn sogar mit etwa nur der Hälfte des Aufwands werden mit dem "pot.verf.3" bessere Konvergenzraten je Iteration als beim "proj.verf." erzielt.

Als letzter Vergleich wird nun das nichtlineare Multilevelverfahren DNL auf das Beispielproblem angewendet. Die in Tabelle 4.8 angegebenen Konvergenzraten je Iteration beziehen sich dabei wieder auf die Konvergenz des nichtlinearen Verfahrens, dessen lineare Gleichungssysteme mit dem jeweiligen Glättungsverfahren vorkonditioniert werden.

level	3	4	5
# DOF	1032 + 367	1629 + 573	3041 + 1056
proj.verf.	0.872	0.947	0.958
pot.verf.2	0.726	0.807	0.853
pot.verf.3	0.570	0.690	0.693
pot.verf.4	0.483	0.610	0.627

Tabelle 4.8: Konvergenzraten für das nichtlineare Problem (DNL)

Bei der Auswertung der Tabelle 4.8 ist insbesondere die relativ levelunabhängige Konvergenzrate der Potentialraum-Korrektur-Verfahren zu beachten. Diese Verfahren sind also sehr viel besser zur Glättung der beim nichtlinearen Multilevelverfahren DNL entstehenden linearen Gleichungssysteme geeignet.
Bezieht man den Aufwand der jeweiligen Glättungsverfahren in den Vergleich mit ein, so ergibt das die in Tabelle 4.9 angegebenen Werte. Dabei ist eindeutig der Vorteil der Anwendung des Potentialraum-Verfahrens gegenüber dem Projektions-Verfahren zu erkennen.

Die numerischen Ergebnisse dieses Abschnitts zeigen deutlich, dass das Potentialraum-

level	3	4	5
proj.verf.	0.683	0.930	0.978
pot.verf.2	0.232	0.589	0.836
pot.verf.3	0.143	0.504	0.741
pot.verf.4	0.138	0.489	0.743

Tabelle 4.9: k_{200} beim Einsatz der Glättungsverfahren für das nichtlineare Problem

Verfahren als Glättungsverfahren in $H(\mathrm{div}, \Omega)$ das schnellere der beiden möglichen Verfahren darstellt. Darüberhinaus zeigen die Ergebnisse beim Einsatz der Glättungsvarianten für das DLL-und DNL-Multilevelverfahren, dass die Verwendung des Projektions-Verfahrens im späteren Vergleich der Multilevelvarianten (Abschnitt 4.2.4) das DNL-Verfahren unangemessen benachteiligen würde, während das Potentialraum-Verfahren auch hier vergleichbare Konvergenzraten für DLL und DNL liefert. Daher wird in den verbleibenden Vergleichen nur noch das Potentialraum-Verfahren mit zwei nodalen Gauß-Seidel-Iterationen als Glättungsverfahren in $H(\mathrm{div}, \Omega)$ verwendet werden.

4.2.3 Performance des LDL-Algorithmus

An den Ergebnissen in Abschnitt 4.2.1 ist deutlich geworden, dass das Abbruchkriterium für die nichtlineare Iterationsvorschrift großen Einfluss sowohl (trivialerweise) auf die Anzahl der Iterationsschritte auf jedem Level, als auch auf den minimal erhaltenen Wert des Funktionals $\mathcal{F}$ hat.
In der Literatur (z.B. [32] oder [45]) findet man für lineare wie nichtlineare Multilevelverfahren keine exakt hergeleiteten Genauigkeitsbedingungen, die die Anzahl von Iterationsschritten oder das Abbruchkriterium vorgeben könnten. Für den linearen Fall (DLL) wird zum Beispiel vorgeschlagen, die Anzahl der V-Zyklen von Newton-Schritt zu Newton-Schritt zu verdoppeln, um die quadratische Konvergenz des Newton-Verfahrens nicht zu zerstören.
In Kapitel 3 dieser Arbeit wurden ebenfalls Genauigkeitsbedingungen für die Lösung des linearen Systems angegeben, die aus der Einhaltung von Abstiegsbedingungen hergeleitet wurden. Der daraus resultierende Algorithmus 3.2 wird nun in diesem Abschnitt weiter untersucht.

In Voraussetzung 3.2 ist die Genauigkeitsbedingung (3.7) für das inexakte Gauß-Newton-Verfahren angegeben, damit die berechnete Suchrichtung eine Abstiegsrichtung an das Funktional ist. Obwohl sich die obere Genauigkeitsschranke e_l ohne die exakte Lösung der linearen Systeme nicht genau berechnen lässt, wird zunächst von einer exakten Berechnung dieser Genauigkeitsschranken ausgegangen, um den prinzipiellen Ablauf des LDL-Verfahrens innerhalb eines Zeitschritts illustrieren zu können. Dazu ist die Genauigkeitsbedingung in Abbildung 4.8 für das LDL-Verfahren in einem Zeitschritt mit fünfmaliger Verfeinerung dargestellt.
Entsprechend der Idee des LDL-Verfahrens durchläuft die x-Achse der Abbildung 4.8 die Anzahl der Gauß-Newton-Schritte, die bis zum Abbruch auf Level 5 nötig waren, ohne beim Übergang zum nächstfeineren Gitter wieder zu beginnen neu zu zählen. Die obere, blaue Linie gibt den Wert von e_l nach (3.7) für die Näherung x_l in jedem l-ten Gauß-Newton-Schritt an. Dies ist damit der maximale Gesamtfehler bzw. die maximale Ungenauigkeit, die bei der Berechnung der Suchrichtung im l-ten Schritt zugelassen werden darf, damit das inexakte Gauß-Newton-Verfahren eine Abstiegsrichtung berechnet.
Jeweils vor dem steilen Anstieg dieser Kurve findet ein Übergang zum nächstfeineren Gitter statt. Dieser Sprung bedeutet, dass auf einem Level j zu Beginn der Minimierung ein relativ großer Fehler zugelassen werden kann. Dies ist dadurch zu erklären, dass die aktuelle Näherung

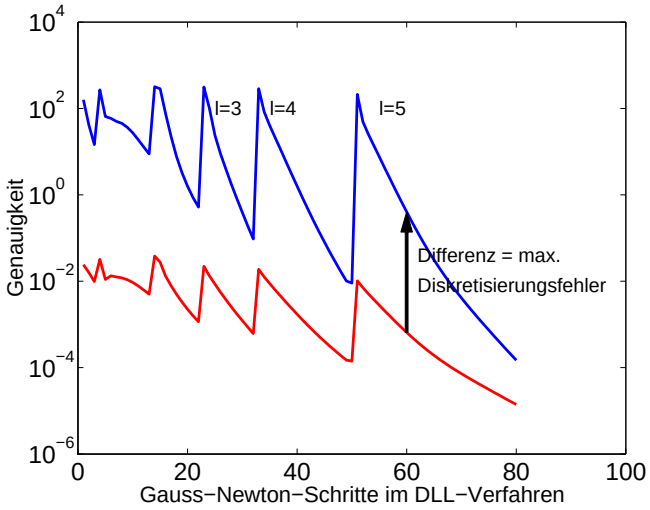


Abbildung 4.8: Exakte Genauigkeitsbedingung (3.7)

x_l noch relativ weit von der Minimalstelle $x_{j,*}$ auf diesem Level entfernt ist. Im Verlauf der Gauß-Newton-Iteration auf diesem Level nähert sich x_l jedoch immer mehr dieser Minimalstelle und daher nimmt auch die zugelassene Ungenauigkeit immer stärker ab. Dies erklärt das Abfallen der Kurve nach dem steilen Anstieg.
Werden V-Zyklen nach Algorithmus 3.1 durchgeführt und danach der algebraische Fehler $\|(\tilde{\delta}_{l,k} - \delta_l)\|_{\mathcal{J},x_l}$ exakt berechnet, erhält man die untere, rote Kurve in Abbildung 4.8. Dies ist genau die linke Seite der Genauigkeitsbedingung (3.7) nach Durchführung von Algorithmus 3. Liegt die rote Kurve unter der blauen Kurve, so ist bei diesem Gauß-Newton-Schritt damit die Genauigkeitsbedingung für den algebraischen Fehler erfüllt. Damit also der gesamte Fehler, der ja gemäß (3.4) aus algebraischem Fehler plus Diskretisierungsfehler besteht, die Bedingung (3.7) erfüllt, darf der Anteil des übriggebliebenen Diskretisierungsfehlers in diesem Schritt nicht größer sein als die Differenz der blauen und der roten Kurve (siehe Pfeil in Abbildung 4.8). Sobald diese Bedingung jedoch verletzt ist, geht der LDL-Algorithmus nach adaptiver Verfeinerung entsprechend dem Fehlerschätzer $\|\mathcal{R}(x_{l,k})\|_{0,\Omega}$ zum nächsthöheren Gitter über. Dabei ist der Diskretisierungsfehler in Abbildung 4.8 in der von Level zu Level immer geringer werdenden Differenz zwischen blauer und roter Kurve unmittelbar vor dem Sprung zum nächsthöheren Level zu erkennen, da ja unmittelbar vor dem Sprung der Diskretisierungsfehler die Abstiegsbedingung verletzt.
Es ist damit im LDL-Algorithmus gerade so angelegt, dass genau so lange auf einem Gitter iteriert wird, bis der Diskretisierungsfehler größer als e_l und damit größer als der erreichte algebraische Fehler ist. Daher steht durch diesen Algorithmus beim Übergang zum nächsthöheren Level eine optimale Näherungslösung des niedrigeren Gitters als Startnäherung zur Verfügung. Auf diese Weise erreicht man für dieses nichtlineare Verfahren die optimale Komplexität des linearen Full-Multigrid-Zyklus.

Da die exakte Lösung der linearen Systeme natürlich nicht zur Verfügung steht, wurde die Genauigkeitsbedingung nach (3.7) in den Gleichungen (3.26) und (3.27) durch berechenbare Ausdrücke abgeschätzt. Dies führt zu den in Abbildung 4.9 eingezeichneten gestrichelten Abschätzungen der (durchgezogen eingezeichneten) exakten Terme wie in Abbildung 4.8.
In dieser Abbildung ist deutlich zu sehen, dass die Abschätzung überall exakt ist, d.h. überall bei Erfüllen der Bedingung "rote gestrichelte Kurve unter blauer gestrichelter Kurve" auch die Bedingung "rote durchgezogene Kurve unter blauer durchgezogener Kurve" erfüllt ist. Wegen dieser Eigenschaft bricht dieses Verfahren natürlich die Berechnung auf einem Level früher ab,

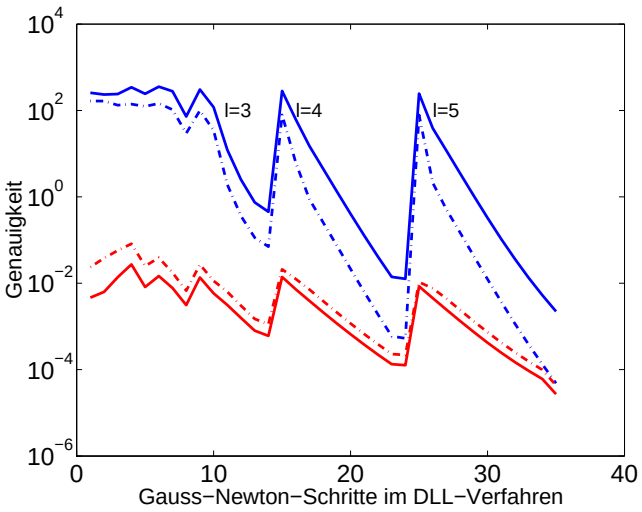


Abbildung 4.9: Abschätzung der Bedingung (3.7)

als wenn die exakten Ausdrücke verwendet werden. Es ist also auszuschliessen, dass zu lange auf einem Gitter gerechnet wird, obwohl der algebraische Fehler schon längst geringer als der Diskrtisierungsfehler geworden ist.
Ob durch das LDL-Verfahren auch das nichtlineare Residuum (wie beim zweiten Abbruchkriterium in Abschnitt 4.2.1) genügend reduziert wird und welcher Wert des Funktionals $\mathcal{F}$ erreicht wird, wird im Vergleich aller drei Multilevelverfahren im letzten Abschnitt 4.2.4 angegeben.

Die Kontrollgröße $\hat{c}_l$ nach (3.14) ergab für die Testläufe ein Verhalten wie $O(l^{-(1+a)})$ mit einem $a > 0$ und erfüllte damit die Voraussetzung aus Satz 3.4 über die Konvergenz des Verfahrens.

4.2.4 Vergleich aller drei Multilevelverfahren

In den Kapiteln 2 und 3 sind drei Multilevelverfahren entwickelt worden, deren wesentliche Eigenschaften in Tabelle 4.10 kurz zusammengefasst werden.

	DLL	DNL	LDL
äußerer Löser	Gauß-Newton (inexakt)	FAS	Newton (inexakt)
Multilevelverfahren	linear	nichtlinear	linear
Abbruchkriterium	$\|\mathcal{J}^*\mathcal{R}\| < \rho$	$\|\mathcal{J}^*\mathcal{R}\| < \rho$	(3.12) und (3.13) verletzt

Tabelle 4.10: Eigenschaften der verschiedenen Multilevelverfahren

Die Verfahren werden zum gleichen Zeitpunkt $t = 0.3$ für das Beispielproblem verglichen. Das Abbruchkriterium des DLL- und DNL-Verfahrens verwendet $\rho = 10^{-6}$ und wird mit Hilfe von (2.20) implementiert. Dies führt zu den Ergebnissen in Tabelle 4.11.

Zunächst wird deutlich, dass durch das DNL-Verfahren ein wesentlich stärker verfeinertes Gitter erzeugt wird und die dadurch erzielten Ergebnisse für das Funktional und den Wert von $\|\mathcal{F}^*\mathcal{R}\|$ daher mit den anderen Verfahren nur eingeschränkt vergleichbar sind. Es ist zu beobachten, dass alle Verfahren ein etwa gleich großes Funktional $\mathcal{F}$ erreichen. Der größte Unterschied zwischen den beiden linearen Verfahren DLL und LDL ist der durch das Gauss-Newton-Multilevel(LDL)-Verfahren erzielte kleinere Wert von $\|\mathcal{F}^*\mathcal{R}\|$. Zwar wurde dazu insgesamt ein etwa zweieinhalbfacher Aufwand benötigt, jedoch ist die dadurch erzielte Lösung näher am Mi-

Level	DLL				DNL			
	# DOF	$\mathcal{F}$	$\|\mathcal{F}^*\mathcal{R}\|$	Zeit	# DOF	$\mathcal{F}$	$\|\mathcal{F}^*\mathcal{R}\|$	Zeit
1	628	$6.58 \cdot 10^{-3}$	$3.48 \cdot 10^{-1}$	5.15	744	$7.66 \cdot 10^{-3}$	$4.48 \cdot 10^{-1}$	12.01
2	872	$3.90 \cdot 10^{-3}$	$1.69 \cdot 10^{-1}$	4.04	1078	$4.59 \cdot 10^{-3}$	$2.40 \cdot 10^{-1}$	12.20
3	1410	$2.42 \cdot 10^{-3}$	$5.85 \cdot 10^{-2}$	10.26	1714	$2.92 \cdot 10^{-3}$	$1.36 \cdot 10^{-1}$	33.10
4	2437	$1.47 \cdot 10^{-3}$	$4.19 \cdot 10^{-2}$	24.62	2883	$1.68 \cdot 10^{-3}$	$4.71 \cdot 10^{-3}$	106.26
5	2670	$1.33 \cdot 10^{-3}$	$3.31 \cdot 10^{-2}$	23.72	3236	$1.29 \cdot 10^{-3}$	$3.79 \cdot 10^{-2}$	100.67

Level	LDL			
	# DOF	$\mathcal{F}$	$\|\mathcal{F}^*\mathcal{R}\|$	Zeit
1	659	$6.73 \cdot 10^{-3}$	$3.56 \cdot 10^{-1}$	7.98
2	903	$3.91 \cdot 10^{-3}$	$1.78 \cdot 10^{-1}$	7.70
3	1425	$2.40 \cdot 10^{-3}$	$5.81 \cdot 10^{-2}$	13.46
4	2444	$1.46 \cdot 10^{-3}$	$1.71 \cdot 10^{-2}$	63.37
5	2661	$1.33 \cdot 10^{-3}$	$6.82 \cdot 10^{-4}$	66.21

Tabelle 4.11: Konvergenzergebnisse der drei Multilevelverfahren

nimierer $(p_{j,*}, q_{j,*})$ im entsprechenden Raum $Q_j \times V_j$, für den ja $\|\mathcal{F}^*(p_{j,*}, q_{j,*})\mathcal{R}(p_{j,*}, q_{j,*})\| = 0$ gilt. Um für die beiden Verfahren DLL und DNL ebenfalls den Wert von $\|\mathcal{F}^*\mathcal{R}\|$ zu verringern, muss das Abbruchkriterium für diese Verfahren verschärft werden. Dazu wird im nächsten Testlauf auf dem höchsten Level $l = 5$ der Wert $\rho = 10^{-8}$ vorgeschrieben. Die Ergebnisse dieses Testlaufs finden sich in Tabelle 4.12.

Level	DLL				DNL			
	# DOF	$\mathcal{F}$	$\|\mathcal{F}^*\mathcal{R}\|$	Zeit	# DOF	$\mathcal{F}$	$\|\mathcal{F}^*\mathcal{R}\|$	Zeit
5	2670	$1.33 \cdot 10^{-3}$	$1.68 \cdot 10^{-4}$	58.27	3236	$1.28 \cdot 10^{-3}$	$6.37 \cdot 10^{-4}$	382.42

Tabelle 4.12: wie Tabelle 4.11 mit $\rho = 10^{-8}$ für $l = 5$

In dieser Tabelle ist zu erkennen, dass offensichtlich die Wahl eines genügend kleinen ρ beim Abbruchkriterium der DLL-und DNL-Verfahren das nichtlineare Residuum $\|\mathcal{F}^*\mathcal{R}\|$ auf die gleiche Größenordnung reduzieren kann wie das Abbruchkriterium des Gauss-Newton-Multilevelverfahrens. Welcher Wert von ρ welche Reduktion von $\|\mathcal{F}^*\mathcal{R}\|$ zur Folge hat, ist von vornherein nicht feststellbar. Ein zu großes ρ lässt, wie in Tabelle 4.11 zu sehen, keine ausreichend große Reduktion von $\|\mathcal{F}^*\mathcal{R}\|$ zu, während das Abbruchkriterium mit einer zu kleinen Wahl von ρ auf dem aktuellen Level die dadurch benötigte Anzahl von Iterationen des Verfahrens sehr stark wachsen lässt. Dementgegen iteriert, wie in Abschnitt 4.2.3 gesehen, das Gauss-Newton-Multilevelverfahren auf jedem Level nicht weiter, als bis der algebraische Fehler gerade kleiner als der Diskretisierungsfehler geworden ist.

Um einen echten Aufwandsvergleich der Verfahren zu erhalten, wird abschließend angenommen, man könnte den Wert von ρ gerade so wählen, dass der mit Hilfe des DLL-und DNL-Verfahrens erreichte Wert von $\|\mathcal{F}^*\mathcal{R}\|$ stets kleiner oder gleich dem vom Gauss-Newton-Multilevelverfahren erreichten Wert ist. Dies wird so implementiert, dass immer zuerst der LDL-Schritt berechnet wird und danach die beiden anderen Verfahren auf der gleichen Triangulierung und mit derselben Startnäherung rechnen, bis der durch sie erzielte Wert von $\|\mathcal{F}^*\mathcal{R}\|$ kleiner gleich dem vom LDL-verfahren berechneten Wert ist. Damit unterscheiden sich das LDL-und das DLL-Verfahren nur noch in der fehlenden Dämpfung beim DLL-Verfahren. Mit Hilfe dieses

Abbruchkriteriums erhält man dann den in Tabelle 4.13 angegebenen Konvergenzverlauf.

Level	DLL		DNL		LDL	
	$\|\mathcal{F}^*\mathcal{R}\|$	Zeit	$\|\mathcal{F}^*\mathcal{R}\|$	Zeit	$\|\mathcal{F}^*\mathcal{R}\|$	Zeit
1	$3.55 \cdot 10^{-1}$	5.08	$3.36 \cdot 10^{-1}$	5.36	$3.57 \cdot 10^{-1}$	7.64
2	$1.69 \cdot 10^{-1}$	2.63	$1.53 \cdot 10^{-1}$	10.33	$1.78 \cdot 10^{-1}$	7.27
3	$5.15 \cdot 10^{-2}$	8.17	$5.71 \cdot 10^{-2}$	27.41	$5.79 \cdot 10^{-2}$	12.91
4	$1.51 \cdot 10^{-2}$	31.57	$1.19 \cdot 10^{-2}$	83.31	$1.68 \cdot 10^{-2}$	60.88
5	$2.30 \cdot 10^{-4}$	51.94	$6.71 \cdot 10^{-4}$	153.11	$7.04 \cdot 10^{-4}$	63.12

Tabelle 4.13: wie Tabelle 4.11, zweites Abbruchkrterium

Der Vergleich zeigt, wie alle Verfahren in der Lage sind, bei geeignetem Abbruchkriterium ein vergleichbar kleines nichtlineares Residuum $\|\mathcal{F}^*\mathcal{R}\|$ zu erzielen. Deutlich wird weiterhin, dass das echt nichtlineare DNL-Verfahren im Zeitvergleich am schlechtesten abschneidet und sich der größere Aufwand der Aufstellung der nichtlinearen FAS-Gleichungen nicht in einer entsprechend besseren Konvergenzrate auszahlt. Für die beiden linearen Verfahren LDL und DLL zeigt sich, dass das DLL-Verfahren (ohne Dämpfung) schneller ist, das vom LDL-Verfahren vorgegebene nichtlineare Residuum zu unterschreiten.

4.2.5 Fazit

Aus den numerischen Ergebnissen lässt sich der Schluss ziehen, dass sich echt nichtlineare Multilevelverfahren wie DNL bei der Anwendung auf das Testbeispiel nicht auszahlen.
Für das DLL-Verfahren stellt sich die Frage nach einem geeigneten Abbruchkriterium für die nichtlineare Iteration auf einem vorgegebenen Level. Ein solches Abbruchkriterium kann aus der Theorie inexakter Newton-Verfahren abgeleitet werden und führt in den Beispielrechnungen zu einer monotonen Reduktion des nichtlinearen Residuums $\|\mathcal{F}^*\mathcal{R}\|$. Gleichzeitig wird durch diese Theorie eine Vorschrift über die Genauigkeit gewonnen, mit der die in jedem Schritt entstehenden linearen Gleichungssysteme mindestens gelöst werden müssen. Zusammen mit dem von der Theorie geforderten Dämpfungsverfahren wird so das DLL-Verfahren zum in Kapitel 3 angegebenen LDL-Verfahren. In der Beispielrechnung wurden die für das LDL-Verfahren theoretisch hergeleiteten Ergebnisse bestätigt.

Damit werden die bisher heuristisch angegebenen Abbruchkriterien für die lineare Iteration durch ein berechenbares Abbruchkriterium ersetzt, das mit Hilfe der Theorie inexakter Newton-Verfahren nach [19] zu einem konvergenten Gesamtverfahren zur Lösung des Minimierungsproblems

$$\mathcal{F}(x_*) = \min_{x \in H} \mathcal{F}(x)$$

führt.

Anhang A

Das Modell von Mualem/van Genuchten

Eine mögliche Parametrisierung der nichtlinearen Funktionen $\Theta(p)$ und $K(p)$ aus den Gleichungen (4.1) und (4.2) findet man in [37] und [46]. Für den Wassergehalt Θ wird dort in Abhängigkeit der Bodentiefe z

$$\Theta(p) = \begin{cases} \Theta_s & \text{für} \quad p \geq z \\ \Theta_r + \frac{\Theta_s - \Theta_r}{(1+\alpha(z-p)^\beta)^{1-1/\beta}} & \text{für} \quad p < z \end{cases} \tag{A.1}$$

gesetzt. Θ_s bzw. Θ_r bezeichnen den gesättigten bzw. residualen Wassergehalt und können durch Probemessungen für jede Bodenart bestimmt werden. Mit Hilfe der Parameter $\alpha > 0$ und $\beta > 1$ wird in (A.1) eine zwischen Θ_r und Θ_s interpolierende Funktion beschrieben, die überall beliebig oft stetig differenzierbar ist. Diese Parameter werden durch Messungen für jede Bodenart empirisch bestimmt.
Die Permeabilität $K(p)$ wird im gleichen Modell wie folgt parametrisiert:

$$K(p) = K_s \Big(\frac{\Theta(p) - \Theta_r}{\Theta_s - \Theta_r} \Big)^{1/2} \left(1 - \left(1 - \Big(\frac{\Theta(p) - \Theta_r}{\Theta_s - \Theta_r} \Big)^{\beta/(\beta-1)} \right)^{1-1/\beta} \right)^2 . \tag{A.2}$$

K_s bezeichnet in (A.2) die Permeabilität eines völlig mit dem Fluid gesättigten Bodens. Im Fall $p \geq z$ ist $\Theta(p) = \Theta_r$ und damit $K \equiv K_s$. Für vier verschiedene Böden sind in Tabelle A.1 Parameter für (A.1) und (A.2) aus [14] angegeben.

Boden	Θ_r [1]	Θ_s [1]	α $[\frac{1}{m}]$	β [1]	K_s $[\frac{m}{s}]$
Sand	0.045	0.43	14.5	2.68	$8.25 \cdot 10^{-5}$
sandiger Lehm	0.065	0.41	7.5	1.89	$1.23 \cdot 10^{-5}$
Lehm	0.078	0.43	3.6	1.56	$2.89 \cdot 10^{-6}$
Ton	0.068	0.38	0.8	1.09	$5.56 \cdot 10^{-7}$

Tabelle A.1: Parameter für das Mualem/van Genuchten-Modell

Anhang B

Fehlerabschätzungen in $H^1(\Omega) \times H(\mathbf{div}, \Omega)$

Für die Abschätzung des Diskretisierungsfehlers müssen Aussagen über die Interpolationsoperatoren zwischen kontinuierlichen und diskreten Räumen verwendet werden. Dadurch hängen diese Fehlerabschätzungen von den eingesetzten Finite-Element-Räumen ab. Für Abschätzungen, in denen Konstanten vorkommen, wird hier $a \lesssim b$ statt $a \leq Cb$ geschrieben.
In dieser Arbeit werden die üblichen stückweise linearen Ansatzfunktionen für $H^1(\Omega)$ zusammen mit den Raviart-Thomas-Funktionen niedrigster Ordnung für $H(\mathrm{div}, \Omega)$ verwendet. In Abhängigkeit von der Geometrie des Gebiets und der Inhomogenität f erhält man für ein Problem der Art (4.6) eine Lösung (p, u) mit

$$p \in H^{1+\alpha}(\Omega), \;\; u \in (H^\alpha(\Omega))^2, \;\; \mathrm{div}\; u \in H^\alpha(\Omega)$$

für ein $\alpha \in (0, 1]$.

Für die Interpolation von $H^1(\Omega)$ auf stückweise lineare Funktionen in Q_l (siehe (4.13)) durch den Operator $I_l : H^1(\Omega) \to Q_l$ findet man in [12]

$$\begin{aligned} \|p - I_l\, p\|_{0,\Omega} &\lesssim h_l^{1+\alpha} \|p\|_{1+\alpha,\Omega} \\ \|\nabla(p - I_l\, p)\|_{0,\Omega} &\lesssim h_l^{\alpha} \|p\|_{1+\alpha,\Omega} \end{aligned}$$

als Abschätzungen und damit

$$\|p - I_l\, p\|_{1,\Omega} \lesssim h_l^{\alpha} \|p\|_{1+\alpha,\Omega}\,. \tag{B.1}$$

Für die Abschätzung in u wird die Norm

$$\|u\|_{\mathrm{div},\alpha,\Omega} := \|u\|_{\alpha,\Omega} + \|\mathrm{div}\; u\|_{\alpha,\Omega} \tag{B.2}$$

definiert. Um die Helmholtz-Zerlegung von $H(\mathrm{div}, \Omega)$ anwenden zu können wird

$$\begin{aligned} H^\circ(\mathrm{div}, \Omega) &:= \{v \in H(\mathrm{div}, \Omega) : \mathrm{div}\; v = 0\} \\ V_l^\circ &:= \{v_l \in V_l : \mathrm{div}\; v_l = 0\} \end{aligned}$$

gesetzt. Unter den gleichen Bedingungen an die Geometrie des Gebiets wie oben gilt dann die Darstellung

$$\begin{aligned} H^\circ(\mathrm{div}, \Omega) &= \{v = \nabla^\perp \psi : \psi \in H^{1+\alpha}(\Omega)\} \\ V_l^\circ &:= \{v_l = \nabla^\perp \psi_l : \psi_l \in Q_l\} \end{aligned}$$

(siehe z.B. [3]). Die Helmholtz-Zerlegung lautet dann

$$H(\mathrm{div}, \Omega) \;=\; H^{\circ}(\mathrm{div}, \Omega) \,\oplus\, H^{g}(\mathrm{div}, \Omega) \tag{B.3}$$

mit dem Raum der schwachen Gradienten $H^g(\mathrm{div}, \Omega)$. Die Zerlegung ist dabei orthogonal zum $L^2(\Omega)$-und $H(\mathrm{div}, \Omega)$-Innenprodukt.
In $H^g(\mathrm{div}, \Omega)$ gilt die Abschätzung (siehe [51, Lemma 3.3])

$$\|\mathrm{div}\; v\|_{0,\Omega} \lesssim \|v\|_{\mathrm{div},\Omega} \lesssim \|\mathrm{div}\; v\|_{0,\Omega} \;\; \forall\, v \in H^{g}(\mathrm{div}, \Omega)\,. \tag{B.4}$$

Sei $Q_l : L^2(\Omega) \to \mathrm{div}(V_l)$ die L^2-Projektion auf $\mathrm{div}(V_l)$. Zwischen den Funktionenräumen bestehen dann folgende kommutative Beziehungen:

$$\begin{array}{ccc} H^1(\Omega) & \xrightarrow{\mathrm{div}} & L^2(\Omega) \\ R_l \downarrow & & Q_l \downarrow \\ V_l & \xrightarrow{\mathrm{div}} & \mathrm{div}(V_l) \end{array} \tag{B.5}$$

und

$$\begin{array}{ccc} H^1(\Omega) & \xrightarrow{\nabla^{\perp}} & H^{\circ}(\mathrm{div}, \Omega) \\ I_l \downarrow & & R_l \downarrow \\ Q_l & \xrightarrow{\nabla^{\perp}} & V_l^{\circ} \end{array} \quad . \tag{B.6}$$

$R_l : H^1(\Omega) \to V_l$ ist die verwendete Projektion. Nun wird zuerst $\|\mathrm{div}(v - R_l v)\|_{0,\Omega}$ abgeschätzt. Setzt man das Diagramm (B.5) ein, so erhält man wie in [12, Proposition 3.9] die Ungleichung

$$\|\mathrm{div}(v - R_l v)\|_{0,\Omega} = \|\mathrm{div}\; v - Q_l \mathrm{div}\; v\|_{0,\Omega} = \|(I - Q_l)\mathrm{div}\; v\|_{0,\Omega} \lesssim h^{\alpha} \|\mathrm{div}\; v\|_{\alpha,\Omega}\,. \tag{B.7}$$

Für die Abschätzung von $\|(v - R_l v)\|_{0,\Omega}$ wird zunächst die Helmholtz-Zerlegung eingesetzt und man erhält

$$\|(v - R_l v)\|_{0,\Omega}^2 \;=\; \|(I - R_l)\nabla^{\perp}\psi\|_{0,\Omega}^2 \;+\; \|(I - R_l)v^g\|_{0,\Omega} \tag{B.8}$$

mit $v^g \in H^g(\mathrm{div}, \Omega)$. Die beiden Teile von Zerlegung (B.8) können getrennt voneinander behandelt werden. Für den zweiten Teil sieht man mit Hilfe von (B.4) und (B.7)

$$\begin{aligned} \|(I - R_l)v^g\|_{0,\Omega} &\lesssim \|\mathrm{div}\; (v^g - R_l v^g)\|_{0,\Omega} \\ &\lesssim h^{\alpha} \|\mathrm{div}\; v^g\|_{\alpha,\Omega} \\ &\leq h^{\alpha} \|v^g\|_{\mathrm{div},\alpha,\Omega} \end{aligned} \tag{B.9}$$

und ist für diesen Teil fertig.
Nun wird das Diagramm (B.6) auf den ersten Teil in (B.8) angewendet und die Identität

$$\|\nabla^{\perp} q\|_{0,\Omega} \;=\; \|\nabla q\|_{0,\Omega}$$

für $\Omega \subseteq \mathbb{R}^2$ verwendet. Sei zunächst $\psi \in H^2(\Omega)$. Dann gilt

$$\begin{aligned} \|(I - R_l)\nabla^{\perp}\psi\|_{0,\Omega} &= \|\nabla^{\perp}(\psi - I_l\psi)\|_{0,\Omega} \\ &= \|\nabla(\psi - I_l\psi)\|_{0,\Omega} \\ &\lesssim h\|\psi\|_{2,\Omega} \end{aligned}$$

(siehe [12]).
Mit Hilfe der Ergebnisse zu Interpolation zwischen Sobolev-Räumen (siehe z.B. [10]) erhält man damit für $\psi \in H^{1+\alpha}(\Omega)$ das Ergebnis

$$\begin{aligned} \|(I - R_l)\nabla^{\perp}\psi\|_{0,\Omega} &\lesssim h^{\alpha}\|\psi\|_{1+\alpha,\Omega} \\ &= h^{\alpha}\|\nabla^{\perp}\psi\|_{\alpha,\Omega} \\ &\leq h^{\alpha}\|\nabla^{\perp}\psi\|_{\mathrm{div},\alpha,\Omega} \end{aligned}$$

und zusammen mit dem vorigen Ergebnis in (B.9) die gewünschte Abschätzung

$$\|(I - R_l)v\|_{0,\Omega} \lesssim h^{\alpha}\|v\|_{\mathrm{div},\alpha,\Omega}\,.$$

Nun kann dieses Ergebnis noch mit (B.7) zusammengesetzt werden und es ergibt sich die gesuchte Fehlerabschätzung

$$\|(I - R_l)v\|_{\mathrm{div},\Omega} \lesssim h^{\alpha}\|v\|_{\mathrm{div},\alpha,\Omega}\,. \tag{B.10}$$

(B.1) und (B.10) können dann zum gesuchten Ergebnis zusammengesetzt werden.

Literaturverzeichnis

[1] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[2] T. Arbogast, M. F. Wheeler, and N.-Y. Zhang, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in porous media*, SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687.

[3] D. N. Arnold, R. S. Falk, and R. Winther, *Preconditioning in $H(div)$ with applications*, Math. Comp., 66 (1997), pp. 957–984.

[4] R. E. Bank and D. J. Rose, *Analysis of a multilevel iterative method for nonlinear finite element equations*, Math. Comp., 39 (1982), pp. 453–465.

[5] J. Bear, *Dynamics of Fluids in Porous Media*, Elsevier, New York, 1972.

[6] M. Berndt, T. A. Manteuffel, and S. F. McCormick, *Local error estimates and adaptive refinement for first-order system least squares*, Electr. Trans. Numer. Anal., 6 (1997), pp. 35–43.

[7] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Review, 40 (1998), pp. 789–837.

[8] D. Braess, *Finite Elements*, Cambridge University Press, Cambridge, 1997.

[9] A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390.

[10] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[11] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers*, RAIRO Anal. Numer., 8 (1974), pp. 129–151.

[12] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.

[13] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part I*, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[14] R. F. Carsel and R. S. Parrish, *Developing joint probability distributions of soil water retention characteristics*, Water Resources Research, 24 (1988), pp. 755–769.

[15] J. W. Daniel, *The Approximate Minimization of Functionals*, Prentice Hall, Englewood Cliffs, 1971.

[16] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[17] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.

[18] P. Deuflhard and A. Hohmann, *Numerische Mathematik I*, De Gruyter, Berlin, 1993.

[19] S. C. Eisenstat and H. F. Walker, *Globally convergent Newton methods*, SIAM J. Optimization, 4 (1994), pp. 393–422.

[20] ——, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[21] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, Rhode Island, 1998.

[22] D. Fokkema, G. Sleijpen, and H. van der Vorst, *Accelerated inexact Newton schemes for large systems of nonlinear equations*, SIAM J. Sci. Comput., 19 (1998), pp. 657–674.

[23] J. Fuhrmann, *Zur Verwendung von Mehrgitterverfahren bei der numerischen Behandlung elliptischer partieller Differentialgleichungen mit variablen Koeffizienten*, PhD thesis, Technische Universität Chemnitz-Zwickau, Aachen, 1995.

[24] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer, New York, 1984.

[25] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[26] C. Grossmann and J. Terno, *Numerik der Optimierung*, Teubner, Stuttgart, 1997.

[27] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer, Berlin, 1985.

[28] W. Hackbusch and A. Reusken, *Analysis of a damped nonlinear multilevel method*, Numer. Math., 55 (1989), pp. 225–246.

[29] R. Hiptmair, *Multigrid method for H(div) in three dimensions*, Electr. Trans. Numer. Anal., 6 (1997), pp. 133–152.

[30] A. Hohmann, *Inexact Gauss Newton methods for parameter dependent nonlinear problems*, Technical Report ZIB, Berlin, 93-13 (1993).

[31] U. Hornung and W. Messing, *Poröse Medien — Methoden und Simulation*, Beiträge zur Hydrologie, Kirchzarten, 1984.

[32] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.

[33] J. Korsawe, *Finite-Element-Ausgleichsprobleme mit den Raviart-Thomas-Räumen : Theorie, Implementierung und Multilevel-Verfahren*, Diplomarbeit, Technische Universität Karlsruhe, 1998.

[34] ——, *Nonlinear first order system least squares finite element multilevel computations : Extended relaxation in H(div)*, in Proceedings of the 6th Copper Mountain Conference on Iterative Methods, T. A. Manteuffel and S. F. McCormick, eds., 2000.

[35] J. Korsawe and G. Starke, *Multilevel projection methods for nonlinear least-squares finite element computations*, Electr. Trans. Numer. Anal., 10 (2000), pp. 56–73.

[36] S. F. McCormick, *Multilevel Projection Methods for Partial Differential Equations*, SIAM, Philadelphia, 1992.

[37] Y. Mualem, *A new model for predicting the hydraulic conductivity of unsaturated porous media*, Water Resources Research, 12 (1976), pp. 513–522.

[38] A. I. Pehlivanov and G. F. Carey, *Error estimates for least-squares mixed finite elements*, Math. Model. Numer. Anal., 28 (1994), pp. 499–516.

[39] E. Perau, *Die Phasen des Bodens und ihre mechanischen Wechselwirkungen*, Mitteilungen aus dem Fachgebiet Grundbau und Bodenmechanik, 2001. In Press.

[40] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method, E. M. I. Galligani, ed., Springer, 1977, pp. 292–315.

[41] G. Starke, *Least-squares mixed finite element solution of variably saturated subsurface flow problems*, SIAM J. Sci. Comput., 21 (2000), pp. 1869–1885.

[42] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer, New York, 2nd ed., 1993.

[43] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer, Berlin, 1997.

[44] G. Timmermann, *A cascadic multigrid algorithm for semilinear elliptic problems*, Numerische Mathematik, 86 (2000), pp. 717–731.

[45] U. Trottenberg, C. W. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, New York, 2000.

[46] M. T. Van Genuchten, *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils*, Soil Sci. Soc. Am. J., 44 (1980), pp. 892–898.

[47] M. Vauclin, D. Khanji, and G. Vachaud, *Experimental and numerical study of a transient, two-dimensional unsaturated-saturated water table recharge problem*, Water Resources Research, 15 (1979), pp. 1089–1101.

[48] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester-Stuttgart, 1996.

[49] C. Wagner, G. Wittum, R. Fritsche, and H.-P. Haar, *Diffusions-Reaktionsprobleme in ungesättigten porösen Medien*, in Mathematik – Schlüsseltechnologie für die Zukunft, K.-H. Hoffmann, W. Jäger, T. Lohmann, and H. Schunck, eds., Springer, 1997.

[50] T. Washio and C. W. Oosterlee, *Krylov subspace acceleration for nonlinear multigrid schemes*, ETNA, 6 (1997), pp. 271–290.

[51] B. I. Wohlmuth and R. H. W. Hoppe, *A comparison of a posteriori error estimators for mixed finite element discretizations by Raviart-Thomas elements*, Math. Comp., 68 (1999), pp. 1347–1378.

[52] C. S. Woodward and C. N. Dawson, *Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media*, SIAM J. Numer. Anal., 37 (2000), pp. 701–724.

[53] O. Zienkiewicz, *The finite element method – From intuition to generality*, Appl. Mech. Rev., 23 (1970), pp. 249–256.

Lebenslauf

Johannes Rudolf Korsawe

22. August 1972	Geboren in Hannover Eltern: Rudolf & Marianne Korsawe, geb. Stielau
	Ausbildung
1979 - 1983	Grundschule Am Lindener Marktplatz, Hannover
1983 - 1985	Orientierungsstufe Ludwig-Windthorst-Schule, Hannover
1985 - 1992	Gymnasium St.-Ursula-Schule, Hannover
5/1992	Abitur
7/1992 - 9/1993	Zivildienst bei der Lebenshilfe für Behinderte, Hannover
1993 - 1998	Studium Technomathematik, Nebenfach Elektrotechnik, Zusatzfach Betriebspädagogik, Universität Karlsruhe (TH)
9/1998	Diplom in Technomathematik, Universität Karlsruhe (TH)
	Wissenschaftliche Tätigkeit
10/1995 - 9/1998	Wissenschaftliche Hilfskraft am Mathematischen Institut I der Universität Karlsruhe (TH)
10/1998 - 12/1998	Wissenschaftlicher Mitarbeiter am Institut für Grundbau und Bodenmechanik der Universität Essen
1/1999 - 9/2000	Wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Ingenieurmathematik der Universität Essen
1/1999 - 12/2001	Bearbeitung des DFG-Projekts STA 402/4 "Iterative Verfahren für nichtlineare Finite-Element-Ausgleichsprobleme"
10/2000 - 12/2001	Wissenschaftlicher Mitarbeiter am Institut für Angewandte Mathematik der Universität Hannover