Globalisierte Newton-Verfahren mit Anwendung auf das Navier-Stokes-Problem

Diplomarbeit

vorgelegt von
Wiebke Lemster
aus
Hamburg

angefertigt im
Institut für Numerische und Angewandte Mathematik
der Georg-August-Universität Göttingen
2008

Inhaltsverzeichnis

Kapitel 1

Einleitung

Fluide, d.h. Gase und Flüssigkeiten, beeinflussen fast jeden Bereich des alltäglichen Lebens und der Umwelt. So ist zum Beispiel das Strömungsverhalten von Fluiden in Rohren sowohl für das Transportieren von Öl und Gas durch Pipelines als auch für den Bluttransport in unseren Adern von entscheidender Bedeutung. Die Bewegung von Newtonschen Fluiden wird durch die Navier-Stokes-Gleichungen

$$\rho\frac{\partial u}{\partial t} + \rho\,(u\cdot\nabla)\,u + \nabla\tilde{p} - \eta\Delta u - (\eta+\lambda)\nabla(\nabla\cdot u) = \tilde{f}$$
$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho u) = 0$$

beschrieben. Hierbei ist u bzw. $\tilde{p}$ die Geschwindigkeit bzw. der Druck des Fluides im betrachteten Gebiet. Mit ρ wird die Dichte des Fluides bezeichnet und λ und η bezeichnen Viskositäten. $\tilde{f}$ ist die von außen wirkende Kraft. Die erste Gleichung zählt zu den partiellen Differentialgleichungen 2.Ordnung und ist durch den Term $\rho\,(u\cdot\nabla)\,u$ nichtlinear. Da man die Gleichung im Allgemeinen nicht analytisch lösen kann, ist man auf numerische Verfahren angewiesen. Für die Diskretisierung des Navier-Stokes-Problems wird die Finite-Elemente-Methode verwendet.

In den hier betrachteten Zusammenhängen gibt es zwei grundlegende Verfahren zur Lösung nichtlinearer Gleichungen, die Picard- und die Newton-Iteration. Beide Verfahren linearisieren das Gleichungssystem. Für die so entstandenen linearen Systeme gibt es zwei Arten von Lösern, direkte und iterative. Direkte Verfahren lösen bis zur Rechnergenauigkeit. Bei iterativen Methoden kann man die Genauigkeit des Residuums angeben, bei der das Verfahren abbricht. Bei moderaten Problemgrößen sind die direkten Löser den iterativen überlegen, während sie bei sehr großen Problemen aufgrund ihres hohen Speicherbedarfs versagen. Daher ist man dort auf iterative Verfahren angewiesen.

Die iterativen Verfahren müssen vorkonditioniert werden, damit sie effizient sind. Das System wird so umgeformt, dass ein äquivalentes, leichter zu lösendes System entsteht. Verwendet man innerhalb der Picard- oder der Newton-Iteration einen iterativen Löser, um die linearen Gleichungen zu lösen, und regelt man die Genauigkeit, mit der das System berechnet wird, so spricht man von einem inexakten Verfahren. Im anderen Fall wird die Methode exakt genannt.

Das Picard-Verfahren ist in dem hier betrachteten Zusammenhang die gebräuchlichste Methode nichtlineare Gleichungen zu lösen. Es ist robust und bietet den Vorteil der a-priori Fehlerabschätzung. Die Methode konvergiert aber nur linear.

Der größte Teil dieser Arbeit beschäftigt sich mit der Newton-Iteration. Für ein nichtlineares Problem $F(x) = 0$ hat der Basisalgorithmus folgende Form.

Zu gegebener Startlösung x_0 berechnet man iterativ s_k durch Lösen des Systems

$$F'(x_k)s_k = -F(x_k).$$

Als neuen Iterationswert erhält man dann $x_{k+1} = x_k + s_k$.
Das Newton-Verfahren ist im Gegensatz zur Picard-Iteration quadratisch konvergent. Man könnte sich nun fragen, warum man nicht immer das Newton-Verfahren benutzt. Der Nachteil ist, dass man eine sehr gute Approximation der gesuchten Lösung als Startwert benötigt, denn das Verfahren konvergiert nur lokal.

Des Weiteren besitzt das inexakte Newton-Verfahren eine etwas schlechtere Konvergenz als das exakte. Es konvergiert immer noch superlinear bei geeigneter Wahl der Forcing-Terme. Die Forcing-Terme beschreiben die Genauigkeit, mit der das lineare System gelöst wird.

Um den Konvergenzradius des Newton-Verfahrens zu vergrößern, verwendet man sogenannte Globalisierungsstrategien. Diese Verfahren verändern den sogenannten Newton-Schritt, das Ergebnis der linearen Gleichung. Man hofft, dass dieser modifizierte Schritt eher zur gesuchten Lösung führt. Allgemein kann man zeigen, dass das globalisierte inexakte Newton-Verfahren unter bestimmten Voraussetzungen an die Forcing-Terme und das System konvergiert.

In Kapitel 2 werden die allgemeine und die relaxierte Picard-Iteration eingeführt. Das exakte und das inexakte Newton-Verfahren werden vorgestellt. Neben einer Konvergenzbetrachtung wird auch die Wahl der Forcing-Terme untersucht.

In Kapitel 3 wird die Globalisierung des inexakten Newton-Verfahrens dargestellt. Die Globalisierungsstrategien teilen sich in zwei Untergruppen, Backtracking- und Trust-Region-Verfahren. Aus beiden Gruppen wird ein Vertreter vorgestellt und sein Einfluß auf die Konvergenz untersucht.

Neben der starken Formulierung des Navier-Stokes-Problems wird die schwache Formulierung derselben in Kapitel 4 eingeführt. Einige funktionalanalytische Grundlagen und die bekannte Lösungstheorie der Navier-Stokes-Gleichung werden zitiert. Sowohl die relaxierte Picard-Iteration als auch das inexakte Newton-Verfahren mit Backtracking werden an das Navier-Stokes-Problem angepasst.

Numerische Tests zu den in Kapitel 4 vorgestellten Algorithmen werden an zwei Testbeispielen durchgeführt. Eines der beiden Beispiele, das Lid-Driven-Cavity-Problem, ist eines der am besten untersuchten Probleme in der numerischen Behandlung der Navier-Stokes-Gleichungen.

Ein kurzes Fazit rundet die Arbeit in Kapitel 6 ab. Die Ergebnisse werden zusammengefasst und kritisch beurteilt.

Im Anhang A findet sich eine Herleitung der Q- und R-Konvergenzordnung und in Anhang B werden die Ergebnisse der Rechnungen nachgetragen.

Kapitel 2

Linearisierungsverfahren

Die beiden Linearisierungsmethoden Picard- und Newton-Verfahren werden vorgestellt. Neben den jeweiligen Konvergenzresultaten wird das inexakte Newton-Verfahren motiviert und eingeführt.

2.1 Picard-Iteration

In der Natur und somit auch in der Physik und der Mathematik gibt es viele nichtlineare Probleme. Dafür benötigt man iterative Verfahren. Ein weit verbreitetes Verfahren ist die Picard-Iteration.

2.1.1 Der Banachsche Fixpunktsatz

Sie beruht auf dem wichtigen Satz von Banach (vgl. [Lub06a] Theorem 2.12 und 2.13).

Satz 2.1.1
Sei g ein Kontraktionsoperator eines vollständigen metrischen Raumes (X, d) in sich, d.h. es gibt ein $q \in [0, 1)$ so, dass

$$d(g(x), g(\tilde{x})) \leq q\, d(x, \tilde{x}) \qquad \forall x, \tilde{x} \in X.$$

Dann konvergiert das Verfahren der sukzessiven Approximation

$$x_{n+1} := g(x_n), \qquad n = 0, 1, \ldots$$

für beliebigen Startwert x_0 gegen den eindeutig bestimmten Fixpunkt x^* von g.

Die sukzessive Approximation wird auch sukzessive Iteration oder Picard-Iteration genannt. Ein großer Vorteil dieses Verfahrens ist die Möglichkeit einer Fehlerabschätzung.

Korollar 2.1.2
Seien die Voraussetzungen von Satz 2.1.1 gegeben. Dann erhält man für beliebige Zahlen $n \in \mathbb{N}$ die a-priori Fehlerabschätzung

$$d(x^*, x_n) \leq \frac{q^n}{1-q}\, d(x_0, x_1)$$

sowie für $n \geq 1$ die a-posteriori Fehlerabschätzung

$$d(x^*, x_n) \leq \frac{q}{1-q}\, d(x_n, x_{n-1}).$$

2.1.2 Dämpfung

Leider sind die Bedingungen aus Satz 2.1.1 in der Praxis nicht immer einhaltbar. Daher bedient man sich des Tricks der Relaxation. Man berechnet die Iterationswerte für gegebenes x_0 folgendermaßen:

$$\begin{aligned} y_{n+1} &:= \omega\,(g(x_n) - x_n)\,, \qquad n = 0, 1, \ldots \\ x_{n+1} &:= x_n + y_{n+1}, \qquad n = 0, 1, \ldots, \end{aligned}$$

wobei $\omega \in (0, 1]$ der Relaxationsparameter ist. Diese Methode ist in vielen Fällen ausschlaggebend für die Konvergenz des Verfahrens.

Korollar 2.1.3
Sei g ein Kontraktionsoperator eines vollständigen metrischen Raumes (X, d) in sich, d.h. es gibt ein $q \in [0, 1)$ so, dass

$$d(g(x), g(\tilde{x})) \leq q\, d(x, \tilde{x}) \qquad \forall x, \tilde{x} \in X.$$

Sei weiter $\omega \in [0, 1)$. Dann konvergiert das Verfahren der relaxierten sukzessiven Approximation

$$\begin{aligned} y_{n+1} &:= \omega\,(g(x_n) - x_n)\,, \qquad n = 0, 1, \ldots \\ x_{n+1} &:= x_n + y_{n+1}, \qquad n = 0, 1, \ldots \end{aligned}$$

für beliebigen Startwert x_0 gegen den eindeutig bestimmten Fixpunkt x^* von g.

Beweis:
Sei g Kontraktionsoperator. Man schreibe die Iterationswerte x_{n+1} so um, dass diese nur noch von x_n abhängen, d.h.

$$x_{n+1} = x_n + y_{n+1} = x_n + \omega\,(g(x_n) - x_n) = (1 - \omega)x_n + \omega g(x_n) =: g_\omega(x_n), \qquad n = 0, 1, \ldots.$$

Betrachtet man nun den Operator g_ω, so gilt

$$\begin{aligned} d(g_\omega(x), g_\omega(\tilde{x})) &= d((1-\omega)x + \omega g(x), (1-\omega)\tilde{x} + g(\tilde{x})) \\ &\leq (1-\omega)d(x, \tilde{x}) + \omega d(g(x), g(\tilde{x})) \\ &\leq (1-\omega)d(x, \tilde{x}) + \omega q\, d(x, \tilde{x}) \\ &= (1 - (1-q)\omega)d(x, \tilde{x}). \end{aligned}$$

Also ist g_ω eine Kontraktion und Satz 2.1.1 kann angewendet werden. □

2.2 Newton-Verfahren

Die im vorherigen Abschnitt vorgestellte Picard-Iteration ist die am häufigsten verwendete numerische Methode zur Lösung der Navier-Stokes-Gleichungen. Da sie aber höchstens linear konvergiert, stellt sich die Frage nach einem besseren Verfahren. Ein Ansatz ist das Newton-Verfahren. Es hat den Vorteil der quadratischen Konvergenz. Wie man aber sehen wird, gilt dies nur lokal (vgl. Satz 2.2.1). Neben diesem wichtigen Satz wird die Konvergenz des inexakten Verfahrens in den Sätzen 2.2.6 und 2.2.10 bewiesen. Für den eindimensionalen Fall kann man sogar kubische Konvergenz für einige modifizierte Newton-Verfahren zeigen (vgl. [KLW06]). Dies wird hier nicht weiter untersucht.

2.2.1 Exaktes Newton-Verfahren

Das Newton-Verfahren ist ein iteratives Verfahren zur Nullstellenbestimmung einer Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$, d.h. man sucht ein $x \in \mathbb{R}^n$ mit

$$F(x) = 0.$$

F ist im Allgemeinen nicht (affin-)linear, ansonsten könnte man ein Verfahren für lineare Gleichungssysteme verwenden. Um eine Motivation des Verfahrens zu erhalten, nehme man an, dass F stetig differenzierbar sei. Für den Fall nicht-glatter, aber immer noch Lipschitz-stetiger Funktionen, stellen Pu und Tian in [PT02] ein Verfahren vor, dass unter bestimmten Voraussetzungen global superlinear konvergiert. Wenn man nun aber davon ausgeht, was im Folgenden vorrausgesetzt wird, dass F differenzierbar ist, gilt nach dem Satz von Taylor

$$F(x+s) = F(x) + F'(x)s + \text{weitere Terme.}$$

Beim Newton-Verfahren linearisiert man die Gleichung, d.h. man vernachlässigt die "weiteren" Terme und berechnet die Nullstelle dieser linearen Gleichung. Die berechnete Stelle ist der neue Iterationswert, d.h. man erhält x_{k+1} aus x_k durch Lösung des Systems

$$0 = F(x_{k+1}) = F(x_k) + F'(x_k)s_k,$$

wobei $s_k = x_{k+1} - x_k$ ist.

Somit ergibt sich

ALGORITHMUS 2.1 N: Newton-Verfahren

Sei x_0 gegeben.
for $k = 0, 1, \ldots$ until Konvergenz **do**
  Löse $F'(x_k)s_k = -F(x_k)$.
  Setze $x_{k+1} := x_k + s_k$.
end for

Über die Konvergenz des Verfahrens kann man das Folgende sagen (vgl. [Pla00]):

Satz 2.2.1
Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ auf einer offenen konvexen Menge D stetig differenzierbar und $x^* \in D$, wobei $F(x^*) = 0$, $F'(x^*)$ invertierbar und $\left\|F'(x^*)^{-1}\right\| \leq \beta < \infty$ ist.
Sei weiter F' Lipschitz-stetig auf D mit Lipschitz-Konstante L und $r > 0$ so, dass $B_r(x^*) \subset D$.
Dann gilt für $x_0 \in B_\delta(x^*)$ mit $\delta = \min\left\{r, \frac{1}{2\beta L}\right\}$, dass das Newton-Verfahren wohldefiniert ist und lokal quadratisch konvergiert, d.h.

$$\|x_{k+1} - x^*\| \leq L\beta \|x_k - x^*\|^2 .$$

Auf zwei Dinge sollte man bei diesem Satz hinweisen. Auf der einen Seite liegt die Konvergenz nur lokal vor, d.h. die ungefähre Lage der Nullstelle (Existenz wird vorausgesetzt!) muss bekannt sein. Auf der anderen Seite konvergiert das Verfahren quadratisch.

Der Beweis des Satzes beruht auf relativ elementaren Eigenschaften der Funktion F, die im Folgenden zitiert bzw. bewiesen werden. Für die Wohldefiniertheit von x_{k+1} benötigt man die Invertierbarkeit von $F'(x_k)$. Dies beruht auf Satz 10.5 aus [Lub06a].

Bemerkung 2.2.2
Sei $B : \mathbb{R}^n \to \mathbb{R}^n$ ein linearer beschränkter Operator mit $\|B\| < 1$. Dann ist $(I - B)$ invertierbar und es gilt

$$\left\|(I-B)^{-1}\right\| \leq \frac{1}{1 - \|B\|},$$

wobei I der Einheitsoperator ist.

Denn diese impliziert die folgende Abschätzung (vgl. [Pla00]), die in Lemma 2.2.4 für die Abschätzung von $\left\|F'(x_k)^{-1}\right\|$ gebraucht wird.

Lemma 2.2.3
Sei $A : \mathbb{R}^n \to \mathbb{R}^n$ regulär. Für jede lineare Abbildung $(B - A) : \mathbb{R}^n \to \mathbb{R}^n$ mit $\left\|A^{-1}\right\| \|B - A\| < 1$ ist $B : \mathbb{R}^n \to \mathbb{R}^n$ regulär und es gilt

$$\left\|B^{-1}\right\| \leq \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}\| \|B - A\|}.$$

Beweis:
Da $\left\|A^{-1}(B - A)\right\| \leq \left\|A^{-1}\right\| \|B - A\| < 1$, folgt aus Bemerkung 2.2.2, dass

$$I + A^{-1}(B - A) \text{ und damit } B = A + (B - A) = A\left[I + A^{-1}(B - A)\right]$$

regulär ist.
Es gilt zum einen $B^{-1} = \left(I + A^{-1}(B - A)\right)^{-1} A^{-1}$ und zum anderen mit Bemerkung 2.2.2

$$\begin{aligned} \left\|B^{-1}\right\| &\leq \left\|\left(I + A^{-1}(B - A)\right)^{-1}\right\| \left\|A^{-1}\right\| \\ &\leq \frac{1}{1 - \|A^{-1}(B - A)\|} \left\|A^{-1}\right\| \\ &\leq \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}\| \|B - A\|}. \end{aligned}$$

□

F' ist invertierbar in x^* und stetig in einer Umgebung des selben Punktes. Somit kann man die Invertierbarkeit von F' in diese Umgebung folgern.

Lemma 2.2.4
Falls die Voraussetzungen von Satz 2.2.1 erfüllt sind und $\|y - x^*\| < \delta$ gilt für $y \in D$, so ist F' invertierbar in D und es gilt $\left\|F'(y)^{-1}\right\| \leq 2\beta$.

Beweis:
Mit der vorausgesetzten Abschätzung für die Norm von $F'(x^*)$ und der Lipschitz-Stetigkeit von F' gilt

$$\left\|F'(x^*)^{-1}\right\| \|F'(y) - F'(x^*)\| \leq \beta L \|y - x^*\|.$$

Da $y \in B_\delta(x^*)$ gilt und $\delta \leq \frac{1}{2\beta L}$ nach Definition ist, folgt

$$\left\|F'(x^*)^{-1}\right\| \|F'(y) - F'(x^*)\| \leq \beta L \delta \leq \frac{1}{2}.$$

Damit sind die Bedingungen von Lemma 2.2.3 erfüllt. Folglich ist $F'(y)$ invertierbar und es gilt

$$\left\|F'(y)^{-1}\right\| \leq \frac{\left\|F'(x^*)^{-1}\right\|}{1 - \|F'(x^*)^{-1}\| \|F'(y) - F'(x^*)\|} \leq \frac{\beta}{\frac{1}{2}} = 2\beta.$$

□

Für den Abstand des neuen Iterationswertes zur gesuchten Lösung benötigt man die folgende Abschätzung (vgl. [Pla00]):

Lemma 2.2.5
Sei F' stetig und Lipschitz-stetig in x^* mit der Lipschitz-Konstante L und der Lipschitz-Umgebung U. Dann gilt für $y \in U$, dass

$$\|F(y) - F(x^*) - F'(y)(y - x^*)\| \leq \frac{L}{2} \|y - x^*\|^2.$$

Beweis:
Der Hauptsatz der Integral- und Differentialrechnung ergibt

$$\|F(y) - F(x^*) - F'(y)(y-x^*)\| = \left\| \int_{x^*}^{y} F'(t)dt - \int_0^1 F'(y)(y-x^*)dt \right\|.$$

Durch Umparametrisierung erhält man

$$\begin{aligned} \|F(y) - F(x^*) - F'(y)(y-x^*)\| &= \left\| \int_0^1 \left[F'(x^* + t(y-x^*)) - F'(y)\right](y-x^*)dt \right\| \\ &\leq \int_0^1 \|F'(x^* + t(y-x^*)) - F'(y)\| \, \|y-x^*\| \, dt. \end{aligned}$$

Zusammen mit der Lipschitz-Stetigkeit von F' ergibt das

$$\begin{aligned} \|F(y) - F(x^*) - F'(y)(y-x^*)\| &\leq \int_0^1 L \, \|x^* + t(y-x^*) - y\| \, \|y-x^*\| \, dt \\ &= L \, \|y-x^*\|^2 \int_0^1 (1-t)dt. \end{aligned}$$

Führt man die Integration aus, so folgt die gewünschte Abschätzung. □

Diese Ergebnisse kommen im Beweis der lokalen Konvergenz des Newton-Verfahrens zum Tragen.

Beweis von Satz 2.2.1:
Um die Wohldefiniertheit zu zeigen, beweist man durch vollständige Induktion, dass die Iterationswerte in einer bestimmten Umgebung von x^* bleiben, d.h. $x_k \in B_\delta(x^*)$ und die Abschätzung $\|x_{k+1} - x^*\| \leq \beta L \, \|x_k - x^*\|^2$ für $k \in N$.

Nach Voraussetzung ist $x_0 \in B_\delta(x^*)$.
Sei also $x_k \in B_\delta(x^*) \subset D$. x_{k+1} ist wohldefiniert, da $F'(x_k)$ nach Lemma 2.2.4 invertierbar ist. Es gilt

$$\begin{aligned} x_{k+1} &= x_k - F'(x_k)^{-1} F(x_k) \\ &= x_k - F'(x_k)^{-1} \left[F(x_k) - \underbrace{F(x^*)}_{=0} \right] \end{aligned}$$

bzw. für den Abstand des neuen Iterationswertes x_{k+1} zu x^*

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - F'(x_k)^{-1} \left[F(x_k) - F(x^*)\right] \\ &= -F'(x_k)^{-1} \left[F(x_k) - F(x^*) - F'(x_k)(x_k - x^*)\right]. \end{aligned}$$

Betrachtet man die Normen, so ergibt sich

$$\begin{aligned} \|x_{k+1} - x^*\| &\leq \left\|F'(x_k)^{-1}\right\| \, \|F(x_k) - F(x^*) - F'(x_k)(x_k - x^*)\| \\ &\overset{\text{Lemmata 2.2.4,2.2.5}}{\leq} 2\beta \frac{L}{2} \, \|x_k - x^*\|^2. \end{aligned}$$

Damit ist die gesuchte Abschätzung bewiesen.

Zu zeigen bleibt noch, dass alle Iterationswerte in $B_\delta(x^*)$ liegen. Das folgt aus der Voraussetzung $\|x_k - x^*\| \leq \delta \leq \frac{1}{(2L\beta)}$ wie folgt:

$$\|x_{k+1} - x^*\| \leq \frac{1}{2}\delta < \delta.$$

□

2.2.2 Das Inexakte Newton-Verfahren

Das exakte Newton-Verfahren hat den Vorteil der quadratischen Konvergenz, ist aber für die Praxis von großen Systemen nicht geeignet. In jedem Schritt muss ein lineares Gleichungssystem gelöst werden. Verwendet man direkte Verfahren, um eine exakte Lösung zu erhalten, so ist der Rechenaufwand $O(n^3)$ für ein dicht besetztes $n \times n$-Problem. Selbst, wenn die Matrix dünn besetzt ist, kann ein ebenso großer Rechenaufwand erforderlich sein ("fill in"). Bei iterativen Lösern benötigt man n Operationen pro Iteration (ein Matrix-Vektor-Produkt) bei dünn besetzten Matrizen. D.h. man erhält einen Rechenaufwand $O(n)$, falls die Iterationszahl nicht von n abhängt. Dies versucht man durch gute Vorkonditionierungsstrategien zu erreichen.

Da die zu lösenden linearen Gleichungssysteme bei nichtlinearen Funktionen nur eine Approximation an die eigentliche Gleichung sind, ist es eventuell möglich, iterative Verfahren zu nutzen, um die Gleichung approximativ zu lösen.
Damit erhält man den

ALGORITHMUS 2.2 IN: Inexaktes Newton-Verfahren

Sei x_0 gegeben.
for $k = 0, 1, \ldots$ until Konvergenz **do**
  Wähle η_k und s_k so, dass
    $\|F(x_k) + F'(x_k)s_k\| \leq \eta_k \|F(x_k)\|$.
  Setze $x_{k+1} := x_k + s_k$.
end for

Beim exakten Verfahren gilt $\eta_k = 0$. Je größer man die *Forcing-Terme* η_k wählt, desto ungenauer wird die Newton-Gleichung erfüllt. Damit regeln die η_k auch die Konvergenz-Geschwindigkeit.

Wahl des Forcing-Terms

Ohne genauere Angaben zu den Forcing-Termen η_k kann man die folgende allgemeine Aussage beweisen (vgl. [DES82]).

Satz 2.2.6
Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar in einer Umgebung von $x^* \in \mathbb{R}^n$, wobei $F(x^*) = 0$ und $F'(x^*)$ nicht singulär ist. Des Weiteren seien $0 \leq \eta_k \leq \eta_{max} \leq t < 1$ gegeben.
Dann existiert ein $\delta > 0$, so dass für $x_0 \in B_\delta(x^*)$ die Folge des Inexakten Newton-Verfahrens $\{x_k\}$ gegen x^* konvergiert und

$$\|x_{k+1} - x^*\|_* \leq t \|x_k - x^*\|_*$$

mit $\|y\|_* := \|F'(x^*)y\|$ gilt.

Die einzige Bedingung an die η_k ist eine obere Schranke $\eta_{max} < 1$. Man erhält lineare Konvergenz, die durch geeignete Wahl der Forcing-Terme verbessert werden kann (vgl. Satz 2.2.10).

Auch für diesen Beweis benötigt man man einige Vorbereitungen. Um die Invertierbarkeit des Operators $F'(x_k)$ in einer Umgebung von x^* zu gewährleisten, benötigt man das folgende Resultat aus der Analysis.

Bemerkung 2.2.7
Sei $A : \mathbb{R}^n \to \mathbb{R}^n$ stetig im Punkt x^* und $A(x^*)$ sei invertierbar. Dann ist A in einer Umgebung von x^* stetig und invertierbar.

Desweiteren kann man folgende Abschätzungen für F' machen (vgl. [OR70]).

Lemma 2.2.8
Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar in einer Umgebung von x^* und $F'(x^*)$ sei nicht singulär.

1. *Für alle $\epsilon > 0$ gibt es $\delta > 0$, so dass $F'(y)$ nicht singulär ist und*
$$\left\|F'(y)^{-1} - F'(x^*)^{-1}\right\| < \epsilon \qquad \text{für alle } y \text{ mit } \|y - x^*\| < \delta.$$
2. *Für alle $\epsilon > 0$ gibt es $\delta > 0$, so dass*
$$\|F(y) - F(x^*) - F'(x^*)(y - x^*)\| < \epsilon \|y - x^*\| \qquad \text{für alle } y \text{ mit } \|y - x^*\| < \delta.$$

Beweis:
1. Nach Bemerkung 2.2.7 ist F' stetig.
2. Nach Voraussetzung ist F differenzierbar und damit auch Frechet-differenzierbar. □

Der Konvergenzsatz 2.2.6 ist für die Norm $\|\cdot\|_*$ gegeben. Einen Zusammenhang zu $\|\cdot\|$ liefert

Lemma 2.2.9
Mit $\mu := \max\left\{\|F'(x^*)\|, \left\|F'(x^*)^{-1}\right\|\right\}$ gilt
$$\frac{1}{\mu}\|y\| \le \|y\|_* \le \mu \|y\|.$$

Beweis:
Nach Definition von $\|\cdot\|_*$ und von μ gilt
$$\|y\|_* = \|F'(x^*)y\| \le \|F'(x^*)\| \|y\| \le \mu \|y\|.$$
Die zweite Ungleichung erhält man aus
$$\|y\| = \left\|F'(x^*)^{-1}F'(x^*)y\right\| \le \left\|F'(x^*)^{-1}\right\| \|y\|_* \le \mu \|y\|_*.$$
□

Nun kann man die Konvergenz des inexakten Newton-Verfahrens beweisen.

Beweis von Satz 2.2.6:
Der Beweis wird mittels Induktion durchgeführt. Dazu benötigt man die folgende Abschätzung für t.

Da $\eta_{max} < t$, gibt es ein genügend kleines $\gamma > 0$ mit
$$(1 + \mu\gamma)\left[\eta_{max}(1 + \mu\gamma) + 2\mu\gamma\right] \le t.$$
Denn $f(\gamma) := (1 + \mu\gamma)\left[\eta_{max}(1 + \mu\gamma) + 2\mu\gamma\right]$ ist stetig mit $f(0) = \eta_{max} < t$.
D.h. es gibt eine Umgebung von 0 so, dass $f(\gamma) \le t$.

Man wähle $\delta > 0$ so klein, dass für $\|y - x^*\| \leq \mu^2\delta$ gilt

(a) $\|F'(y) - F'(x^*)\| \leq \gamma$,

(b) $\left\|F'(y)^{-1} - F'(x^*)^{-1}\right\| \leq \gamma$ und

(c) $\|F(y) - F(x^*) - F'(x^*)(y - x^*)\| \leq \gamma \|y - x^*\|$.

So ein δ existiert, weil F' stetig ist und Bemerkung 2.2.7 und Lemma 2.2.8 gelten.

Dann nehme man an, dass $\|x_0 - x^*\| < \delta$.

Die Behauptung

$$\|x_{k+1} - x^*\|_* \leq t \|x_k - x^*\|_*$$

folgt durch vollständige Induktion nach k:
Damit der Beweis ein wenig übersichtlicher wird, teilt man ihn in drei Teile.

(1) Als Erstes zeigt man, dass die Iterationswerte in der Kugel $B_{\mu^2\delta}(x^*)$ liegen, denn dann kann man die oben beschriebenen Abschätzungen verwenden.

$$\begin{aligned}
\|x_k - x^*\| &\overset{\text{Lemma 2.2.9}}{\leq} \mu \|x_k - x^*\|_* \\
&\overset{I.V.}{\leq} \mu t^k \|x_0 - x^*\|_* \\
&\overset{\text{Lemma 2.2.9}}{\leq} \mu^2 t^k \|x_0 - x^*\| \\
&\overset{t<1}{\leq} \mu^2 \|x_0 - x^*\| \leq \mu^2\delta.
\end{aligned}$$

(2) Als Zweites benötigt man eine Abschätzung für die Norm von $F(x_k)$. Man macht eine Nullergänzung und ordnet die Terme in folgender Weise:

$$\begin{aligned}
F(x_k) &= F(x_k) - \overbrace{F(x^*)}^{=0} + F'(x^*)(x_k - x^*) - F'(x^*)(x_k - x^*) \\
&= [F'(x^*)(x_k - x^*)] + [F(x_k) - F(x^*) - F'(x^*)(x_k - x^*)] .
\end{aligned}$$

Mittels Dreiecksungleichung ergibt sich für $\|F(x_k)\|$

$$\begin{aligned}
\|F(x_k)\| &\leq \|F'(x^*)(x_k - x^*)\| + \|F(x_k) - F(x^*) - F'(x^*)(x_k - x^*)\| && (2.1) \\
&\leq \|x_k - x^*\|_* + \gamma \|x_k - x^*\| . && (2.2)
\end{aligned}$$

(3) Als Drittes werden (1) und (2) verwendet, um die gesuchte Ungleichung zu erhalten. Dazu sei $r_k := F'(x_k)s_k + F(x_k)$. Dann gilt

$$\begin{aligned}
F'(x^*)(x_{k+1} - x^*) &= F'(x^*)\left(s_k + x_k - x^* + F'(x_k)^{-1}F(x^*)\right) \\
&\overset{\text{Def. } r_k}{=} F'(x^*)\left[F'(x_k)^{-1}(r_k - F(x_k)) + x_k - x^* + F'(x_k)^{-1}F(x^*)\right] \\
&= F'(x^*)F'(x_k)^{-1}r_k + F'(x^*)(x_k - x^*) + F'(x^*)F'(x_k)^{-1}F(x^*) \\
&\quad -F'(x^*)F'(x_k)^{-1}F(x_k) + F'(x^*)F'(x_k)^{-1}F'(x^*)(x_k - x^*) \\
&\quad -F'(x^*)F'(x_k)^{-1}F'(x^*)(x_k - x^*).
\end{aligned}$$

Umordnen und mehrfache Addition mit Null ergibt

$$\begin{aligned}
F'(x^*)(x_{k+1}-x^*) &= F'(x^*)F'(x_k)^{-1}r_k + F'(x^*)F'(x_k)^{-1}\left[F'(x_k)-F'(x^*)\right](x_k-x^*)\\
&\quad -F'(x^*)F'(x_k)^{-1}\left[F(x_k)-F(x^*)-F'(x^*)(x_k-x^*)\right]+r_k-r_k\\
&\quad +\left[F'(x_k)-F'(x^*)\right](x_k-x^*)-\left[F'(x_k)-F'(x^*)\right](x_k-x^*)+F(x^*)\\
&\quad -F(x^*)+F(x_k)-F(x_k)+F'(x^*)(x_k-x^*)-F'(x^*)(x_k-x^*)\\
&= r_k+F'(x^*)\left[F'(x_k)^{-1}-F'(x^*)^{-1}\right]r_k\\
&\quad +\left[F'(x_k)-F'(x^*)\right](x_k-x^*)\\
&\quad +F'(x^*)\left[F'(x_k)^{-1}-F'(x^*)^{-1}\right]\left[F'(x_k)-F'(x^*)\right](x_k-x^*)\\
&\quad -F(x_k)+F(x^*)+F'(x^*)(x_k-x^*)\\
&\quad -F'(x^*)\left[F'(x_k)^{-1}-F'(x^*)^{-1}\right]\left[F(x_k)-F(x^*)-F'(x^*)(x_k-x^*)\right].
\end{aligned}$$

Betrachtet man die letzten Terme, so sieht man, dass dreimal der Faktor $\left(I+F'(x^*)\left[F'(x_k)^{-1}-F'(x^*)^{-1}\right]\right)$ ausgeklammert werden kann. Man erhält

$$\begin{aligned}
&F'(x^*)(x_{k+1}-x^*)\\
=\ &\left(I+F'(x^*)\left[F'(x_k)^{-1}-F'(x^*)^{-1}\right]\right)r_k\\
&+\left(I+F'(x^*)\left[F'(x_k)^{-1}-F'(x^*)^{-1}\right]\right)\left[F'(x_k)-F'(x^*)\right](x_k-x^*)\\
&+\left(I+F'(x^*)\left[F'(x_k)^{-1}-F'(x^*)^{-1}\right]\right)\left[-F(x_k)+F(x^*)+F'(x^*)(x_k-x^*)\right]\\
=\ &\left(I+F'(x^*)\left[F'(x_k)^{-1}-F'(x^*)^{-1}\right]\right)\\
&\cdot\left(r_k+\left[F'(x_k)-F'(x^*)\right](x_k-x^*)-F(x_k)+F(x^*)+F'(x^*)(x_k-x^*)\right).
\end{aligned}$$

Betrachtet man die Normen, so folgt aus der Definition von $\|\cdot\|_*$ und der obigen Rechnung

$$\begin{aligned}
\|x_{k+1}-x^*\|_* &= \|F'(x^*)(x_{k+1}-x^*)\|\\
&\le \left\|I+F'(x^*)\left[F'(x_k)^{-1}-F'(x^*)^{-1}\right]\right\|\\
&\quad \cdot\|r_k+\left[F'(x_k)-F'(x^*)\right](x_k-x^*)\\
&\qquad -F(x_k)+F(x^*)+F'(x^*)(x_k-x^*)\|.
\end{aligned}$$

Mittels Dreiecksungleichung und $\|I\|=1$ ergibt sich

$$\begin{aligned}
\|x_{k+1}-x^*\|_* &\le \left(1+\|F'(x^*)\|\left\|F'(x_k)^{-1}-F'(x^*)^{-1}\right\|\right)\\
&\quad \cdot(\|r_k\|+\|F'(x_k)-F'(x^*)\|\,\|x_k-x^*\|\\
&\qquad +\|F(x_k)-F(x^*)-F'(x^*)(x_k-x^*)\|).
\end{aligned}$$

Benutzt man die Definition von μ und die Ungleichungen (a), (b) und (c) so gilt

$$\begin{aligned}
\|x_{k+1}-x^*\|_* &\le (1+\mu\gamma)\\
&\quad \cdot\left(\|r_k\|+\gamma\,\|x_k-x^*\|+\gamma\,\|x_k-x^*\|\right).
\end{aligned}$$

Da für die Forcing-Terme $\eta_k \le \eta_{max}$ gilt, folgt

$$\begin{aligned}
\|x_{k+1}-x^*\|_* &\le (1+\mu\gamma)(\eta_{max}\,\|F(x_k)\|+2\gamma\,\|x_k-x^*\|)\\
&\overset{\text{Gleichung (2.2)}}{\le} (1+\mu\gamma)(\eta_{max}\left(\|x_k-x^*\|_*+\gamma\,\|x_k-x^*\|\right)\\
&\qquad +2\gamma\,\|x_k-x^*\|).
\end{aligned}$$

Mit Hilfe von Lemma 2.2.9 kann man folgende Ungleichung folgern

$$\|x_{k+1}-x^*\|_* \le (1+\mu\gamma)(\eta_{max}(1+\mu\gamma)+2\mu\gamma)\,\|x_k-x^*\|_*.$$

Aus der Vorüberlegung folgt dann die gewünschte Abschätzung

$$\|x_{k+1}-x^*\|_* \le t\,\|x_k-x^*\|_*.$$

□

Der Satz 2.2.6 besagt etwas über die Konvergenz in der $\|\cdot\|_*$-Norm. Für $\|\cdot\|$ erhält man

$$\|x_{k+1} - x^*\| \leq \mu \|x_{k+1} - x^*\|_* \leq \mu\ t \|x_k - x^*\|_* \leq \mu^2 t \|x_k - x^*\| .$$

Falls $\|F'(x^*)\|$ oder $\left\|F'(x^*)^{-1}\right\|$ sehr groß ist, erhält man eine sehr langsame Konvergenz.

Wählt man η_k konstant, so erhält man lineare Konvergenz. Dabei kann t, je nach Wahl der η_k, sehr klein werden bzw. nahe bei 1 liegen. D.h. je kleiner man seine η_k wählt, desto besser ist die Konvergenz des Verfahrens. Wählt man sie allerdings zu klein, kann die Laufzeit erheblich zunehmen, da das lineare System zu genau gelöst werden muss. In [TWS02] wird dies an Hand des Backtracking-Verfahrens (vgl. Abschnitt 3.4) analytisch und numerisch untersucht.

In vielen Fällen ist es sinnvoller, die η_k dem Problem angepasst zu wählen. Eine Möglichkeit, die sich bei vielen Problemen bewährt hat, ist die Einbeziehung des vorherigen Newton-Schrittes (vgl. [EW96]):

Man wählt ein $\eta_0 \in [0, 1)$ und betrachtet das Verhältnis von erwarteter Reduktion $\|F(x_k)\| - \|F(x_{k-1}) + F'(x_{k-1})s_{k-1}\|$ und $\|F(x_{k-1})\|$ und wählt $\tilde{\eta_k}$ als den Betrag dieses Verhältnisses. $\tilde{\eta_k}$ wird dann wie folgt berechnet.

Sei $\eta_0 \in [0, 1)$ gegeben. Man definiert

$$\tilde{\eta_k} := \frac{|\ \|F(x_k)\| - \|F(x_{k-1}) + F'(x_{k-1})s_{k-1}\|\ |}{\|F(x_{k-1})\|} \qquad k = 1, 2, \ldots .$$

Damit die $\tilde{\eta_k}$ in der Praxis nicht zu schnell abnehmen und damit das Finden eines geeigneten Newton-Schrittes verhindern, setzt man als neue Schranke $\tilde{\eta_k} = \max\left\{\tilde{\eta_k}, \tilde{\eta}_{k-1}^{\frac{1+\sqrt{5}}{2}}\right\}$, falls $\tilde{\eta}_{k-1}^{\frac{1+\sqrt{5}}{2}}$ größer als eine Konstante c ist. In der Praxis wird oft 0.1 als diese Konstante gewählt.
Der Exponent berücksichtigt die R-Ordnung des Inexakten Newton-Verfahrens. Diese ist $\frac{1+\sqrt{5}}{2}$, wie man der Motivation im Anhang (vgl. Abschnitt A) entnehmen kann.

Es kann vorkommen, dass die $\tilde{\eta_k}$ größer als 1 werden. Da die Norm von $F(x_k)$ in diesem Fall größer als die Norm von $F(x_{k-1})$ werden könnte, zwingt man die $\tilde{\eta_k}$ unter eine gegebene Grenze $\eta_{max} < 1$. Insgesamt berechnet man die η_k mit Hilfe der Sicherheit

$$\eta_k := \min\left\{\eta_{max}, \max\left\{\tilde{\eta_k}, \tilde{\eta}_{k-1}^{\frac{1+\sqrt{5}}{2}}\right\}\right\}, \qquad \text{falls } \tilde{\eta}_{k-1}^{\frac{1+\sqrt{5}}{2}} > c. \tag{2.3}$$

Neben dieser Wahl gibt es natürlich noch viele andere Möglichkeiten, die Forcing-Terme zu wählen. In [EW96] werden neben den hier gewählten Forcing-Termen die Werte η_k auf folgende Weise berechnet: Sei $\gamma \in [0, 1]$, $\alpha \in (1, 2]$ und $\eta_0 \in [0, 1)$ gegeben

$$\eta_k := \gamma \left(\frac{\|F(x_k)\|}{\|F(x_{k-1})\|}\right)^{\alpha} \qquad k = 1, 2, \ldots .$$

Für diese Forcing-Terme kann man ebenfalls superlineare Konvergenz zeigen, falls $\gamma \leq 1$. Numerische Tests und weitere mögliche Forcing-Terme werden in [AML07] dargestellt.

Konvergenz des Inexakten Newton-Verfahrens

Für die zuvor beschriebene Wahl der Forcing-Terme kann man 2-Schritt quadratische Konvergenz zeigen (vgl. [EW96]).

Satz 2.2.10
Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar in einer Umgebung von $x^* \in \mathbb{R}^n$, so dass $F(x^*) = 0$ und

$F'(x^*)$ nicht singulär ist. Desweiteren sei F Lipschitz-stetig differenzierbar in x^* und die η_k wie in (2.3) beschrieben gewählt.
Dann konvergiert die Folge $\{x_k\}$ des Inexakten Newton-Verfahrens gegen x^* für hinreichend kleine $\|x_0 - x^*\|$ und es gibt $t > 0$ mit

$$\begin{aligned} \|x_{k+1} - x^*\| &\leq t \|x_k - x^*\| \|x_{k-1} - x^*\|, \text{ bzw.} \\ \|x_{k+1} - x^*\| &\leq t \|x_{k-1} - x^*\|^2 \\ \Rightarrow \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} &\overset{k\to\infty}{\longrightarrow} 0. \end{aligned}$$

Bevor man diesen Satz beweisen kann, fehlen noch einige Abschätzungen. Als Erstes geht es um eine Abschätzung der Norm von $F(y)$ gegenüber dem Abstand von y zu x^* (vgl. [DES82]).

Lemma 2.2.11
Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar in einer Umgebung von x^*, wobei $x^* \in \mathbb{R}^n$ eine Nullstelle der Funktion sei. Desweiteren sei $F'(x^*)$ nicht singulär.
Setzt man $\alpha := \max\left\{\|F'(x^*)\| + \frac{1}{2\|F'(x^*)^{-1}\|}, 2\left\|F'(x^*)^{-1}\right\|\right\}$, so gilt

$$\frac{1}{\alpha}\|y - x^*\| \leq \|F(y)\| \leq \alpha \|y - x^*\|,$$

falls $\|y - x^*\|$ hinreichend klein ist.

Beweis:
Setze $\beta := \left\|F'(x^*)^{-1}\right\|$. Nach Lemma 2.2.8 gibt es ein hinreichend kleines $\delta > 0$, so dass

$$\|F(y) - F(x^*) - F'(x^*)(y - x^*)\| \leq \frac{1}{2\beta}\|y - x^*\| \text{ für } \|y - x^*\| < \delta. \tag{2.4}$$

Es gilt (vgl. Beweis von Satz 2.2.6)

$$F(y) = [F'(x^*)(y - x^*)] + [F(y) - F(x^*) - F'(x^*)(y - x^*)].$$

Betrachtet man die Normen, so ergibt sich

$$\begin{aligned} \|F(y)\| &\leq \|F'(x^*)(y - x^*)\| + \|F(y) - F(x^*) - F'(x^*)(y - x^*)\| \\ &\leq \|F'(x^*)\| \|y - x^*\| + \frac{1}{2\beta}\|y - x^*\| \\ &= \left[\|F'(x^*)\| + \frac{1}{2\beta}\right]\|y - x^*\| \leq \alpha \|y - x^*\| \text{ für } \|y - x^*\| < \delta. \end{aligned}$$

Damit ist die erste Ungleichung gezeigt.
Für die zweite Ungleichung betrachtet man

$$\begin{aligned} \|y - x^*\| &= \left\|F'(x^*)^{-1}F'(x^*)(y - x^*)\right\| \\ &\leq \left\|F'(x^*)^{-1}\right\| \|F'(x^*)(y - x^*)\| \end{aligned} \tag{2.5}$$

und damit für y mit $\|y - x^*\| < \delta$

$$\begin{aligned} \|F(y)\| &\geq \|F'(x^*)(y - x^*)\| - \|F(y) - F(x^*) - F'(x^*)(y - x^*)\| \\ &\overset{(2.4),(2.5)}{\geq} \left\|F'(x^*)^{-1}\right\|^{-1}\|y - x^*\| - \frac{1}{2\beta}\|y - x^*\|. \end{aligned}$$

Einfache Umformungen und die Definition von β ergibt

$$\|F(y)\| \geq \frac{1}{2\beta}\|y - x^*\|.$$

Da nach Definition von α die Ungleichung $\alpha \geq 2\beta$ gilt, folgt $\frac{1}{2\beta} \geq \frac{1}{\alpha}$ und damit

$$\|F(y)\| \quad \geq \quad \frac{1}{\alpha}\|y - x^*\| \text{ für } \|y - x^*\| < \delta.$$

□

Für die folgenden Lemmata mache man die (vgl. [EW96])

Annahme 2.2.12 .

(a) Sei $\mu := \max\left\{\|F'(x^*)\|, \left\|F'(x^*)^{-1}\right\|\right\}$.

(b) Sei $\epsilon > 0$ *so klein, dass*

(1) F *stetig differenzierbar und* F' *nicht-singulär in* $B_\epsilon(x^*)$*ist,*

(2) $\left\|F'(y)^{-1}\right\| \leq 2\mu$ *für* $y \in B_\epsilon(x^*)$,

(3) $\|F'(y) - F'(x^*)\| \leq L\|y - x^*\|$ *für* $y \in B_\epsilon(x^*)$,

(4) $\epsilon < \frac{2}{L\alpha}$ *für* $\alpha := \max\left\{\|F'(x^*)\| + \frac{1}{2\|F'(x^*)^{-1}\|}, 2\left\|F'(x^*)^{-1}\right\|\right\}$.

Als Zweites sucht man eine Abschätzung für den Nenner der Forcing-Terme (vgl. [EW96]).

Lemma 2.2.13
Sei Annahme 2.2.12 erfüllt und $y \in B_\epsilon(x^*)$ *und* s *so gewählt, dass* $\tilde{y} := y + s \in B_\epsilon(x^*)$. *Dann gilt*

$$|\,\|F(\tilde{y})\| - \|F(y) + F'(y)s\|\,| \leq \|F(\tilde{y}) - F(y) - F'(y)s\| \leq L\left(2\|y - x^*\| + \frac{\|s\|}{2}\right)\|s\|.$$

Beweis:
Die erste Ungleichung ergibt sich wie folgt:

$$\begin{aligned} |\,\|F(\tilde{y})\| - \|F(y) + F'(y)s\|\,| &\leq \|F(\tilde{y}) - (F(y) + F'(y)s)\| \\ &= \|F(\tilde{y}) - F(y) - F'(y)s\|. \end{aligned}$$

Nach dem Hauptsatz der Integral- und Differentialrechnung und einer Umparametrisierung ergibt sich (vgl. Beweis von Lemma 2.2.5)

$$\begin{aligned} \|F(\tilde{y}) - F(y) - F'(y)s\| &= \left\|\int_0^1 F'(y + t(\tilde{y} - y))(\tilde{y} - y)dt - F'(y)s\right\| \\ &= \left\|\int_0^1 \left[F'(y + t(\tilde{y} - y)) - F'(x^*)\right](\tilde{y} - y)dt - \left[F'(y) - F'(x^*)\right]s\right\| \\ &\leq \left[\int_0^1 \|F'(y + t(\tilde{y} - y)) - F'(x^*)\|\,dt + \|F'(y) - F'(x^*)\|\right]\|s\|. \end{aligned}$$

Nutzt man die Lipschitz-Stetigkeit von F' aus, so folgt

$$\begin{aligned} |\,\|F(\tilde{y})\| - \|F(y) - F'(y)s\|\,| &\leq \left[\int_0^1 L\|y + t(\tilde{y} - y) - x^*\|\,dt + L\|y - x^*\|\right]\|s\| \\ &\leq \left[L\|y - x^*\| + L\|s\|\frac{1}{2} + L\|y - x^*\|\right]\|s\| \end{aligned}$$

und damit die gesuchte Gleichung. □

Als nächstes braucht man eine Abschätzung für die Norm von s (vgl. [EW96]).

Lemma 2.2.14
Sei Annahme 2.2.12 erfüllt und $y \in B_\epsilon(x^*)$ und $\|F(y) + F'(y)s\| \leq \eta \|F(y)\|$ für ein s und ein $\eta \in [0,1)$. Dann gilt

$$\|s\| \leq 4\mu \|F(y)\| .$$

Beweis:
Es gilt

$$\begin{aligned} \|s\| &= \left\|F'(y)^{-1}F'(y)s\right\| \\ &\leq \left\|F'(y)^{-1}\right\| \|F'(y)s\| . \end{aligned}$$

Verwendet man die Abschätzung für $\left\|F'(y)^{-1}\right\|$ aus Annahme 2.2.12 und addiert mit $0 = F(y) - F(y)$, so ergibt sich

$$\begin{aligned} \|s\| &\leq 2\mu(\|F(y)\| + \|F(y) + F'(y)s\|) \\ &\overset{Vor.}{\leq} 2\mu(1+\eta) \|F(y)\| \\ &\leq 4\mu \|F(y)\| . \end{aligned}$$

□

Als Letztes fehlt noch eine Abschätzung für die Norm des neuen Iterationswertes (vgl. [EW96]).

Lemma 2.2.15
Sei Annahme 2.2.12 erfüllt und für $y \in B_\epsilon(x^*)$ existiere s und $\eta \in [0,1)$, so dass $\|F(y) + F'(y)s\| \leq \|F(y)\|$ und $\tilde{y} := y + s \in B_\epsilon(x^*)$. Dann gibt es ein $B > 0$ mit

$$\|F(\tilde{y})\| \leq (\eta + B \|F(y)\|) \|F(y)\| .$$

Beweis:
Bei Addition mit Null erhält man

$$\begin{aligned} \|F(\tilde{y})\| &= \|F(\tilde{y}) - F(y) - F'(y)s + F(y) + F'(y)s\| \\ &\leq \|F(\tilde{y}) - F(y) - F'(y)s\| + \|F(y) + F'(y)s\| \\ &\overset{\text{Lemma 2.2.13}}{\leq} \eta \|F(y)\| + L(2\|y - x^*\| + \frac{\|s\|}{2}) \|s\| . \end{aligned}$$

Da durch Annahme 2.2.12 die Voraussetzungen von Lemma 2.2.11 erfüllt sind, ergibt sich

$$\begin{aligned} \|F(\tilde{y})\| &\leq \eta \|F(y)\| + L(2\alpha \|F(y)\| + \frac{\|s\|}{2}) \|s\| \\ &\overset{\text{Lemma 2.2.14}}{\leq} \eta \|F(y)\| + L(2\alpha \|F(y)\| + 2\mu \|F(y)\|)4\mu \|F(y)\| \\ &= (\eta + \underbrace{\left[8\alpha L\mu + 8\mu^2 L\right]}_{=:B} \|F(y)\|) \|F(y)\| . \end{aligned}$$

□

Damit sind die Grundlagen für den Beweis von Satz 2.2.10 geschaffen.

Beweis von Satz 2.2.10:
Sei $\tilde{\epsilon}$ mit $0 < \tilde{\epsilon} \leq \epsilon \frac{1}{1+4\mu\alpha}$ gegeben. Für $y \in B_{\tilde{\epsilon}}(x^*)$ mit $\|F(y) + F'(y)s\| \leq \eta \|F(y)\|$ und geeignet gewähltem $s \in \mathbb{R}^n$, $\eta \in [0, 1)$ gilt

$$y + s \in B_\epsilon(x^*),$$

denn

$$\begin{aligned} \|y + s - x^*\| &\leq \|y - x^*\| + \|s\| \\ &\overset{\text{Lemma 2.2.14}}{\leq} \|y - x^*\| + 4\mu \|F(y)\|. \end{aligned}$$

Aufgrund von Lemma 2.2.11 und der Tatsache, dass $\|y - x^*\| \leq \tilde{\epsilon}$, folgt

$$\|y + s - x^*\| \leq (1 + 4\mu\alpha)\tilde{\epsilon} \leq \epsilon.$$

Sei $\eta_0 \in [0, 1)$ gegeben und τ so gewählt, dass $\eta_0 < \tau < 1$. Wähle $\delta > 0$ so klein, dass

(1) $\eta_0 + B\delta \leq \tau$,

(2) $[8L\mu(\alpha + \mu) + B]\,\delta \leq \tau$ und

(3) $\delta < \frac{\tilde{\epsilon}}{\alpha}$.

Dann gilt für $y \in B_\epsilon(x^*)$ und $\|F(y)\| \leq \delta$

$$\|y - x^*\| \overset{\text{Lemma 2.2.11}}{\leq} \alpha \|F(y)\| \leq \alpha\delta \overset{(3)}{<} \tilde{\epsilon}.$$

Also gilt $y \in B_{\tilde{\epsilon}}(x^*)$.

Für $\|F(x_0)\| \leq \delta$ zeigt man $x_k \in B_{\tilde{\epsilon}}(x^*)\ \forall k$ und $\|F(x_k)\| \leq \delta$ durch vollständige Induktion. Daraus ergibt sich mit dem eben Gezeigtem, dass alle neuen Iterationswerte in $B_{\tilde{\epsilon}}(x^*)$ liegen.

Induktionsanfang:

Sei $x_0 \in B_\epsilon(x^*)$ genügend nahe bei x^*, so dass $\|F(x_0)\| \leq \delta$.
Da $x_0 \in B_{\tilde{\epsilon}}(x^*)$, gilt $x_1 \in B_\epsilon(x^*)$. Nach Lemma 2.2.15 gilt somit

$$\|F(x_1)\| \leq (\eta + B \|F(x_0)\|) \|F(x_0)\|.$$

Aufgrund der Abschätzung für die Norm von $F(x_0)$ und (1) gilt

$$\|F(x_1)\| \leq (\eta + B\delta) \|F(x_0)\| \leq \tau \|F(x_0)\|$$

und damit

$$\|F(x_1)\| \leq \|F(x_0)\| \leq \delta.$$

D.h. es folgt $x_1 \in B_{\tilde{\epsilon}}(x^*)$.

Induktionsschritt:

Nach Voraussetzung gilt $x_{k-1}, x_k \in B_{\tilde{\epsilon}}(x^*)$ und $\|x_i\| \leq \delta$ für $i = k - 1, k$.
D.h. es gilt $x_{k+1} \in B_\epsilon(x^*)$. Für die Forcing-Terme gilt

$$\begin{aligned} \eta_k &= \frac{|\,\|F(x_k)\| - \|F(x_{k-1}) + F'(x_{k-1})s_{k-1}\|\,|}{\|F(x_{k-1})\|} \\ &\overset{\text{Lemma 2.2.13}}{\leq} \frac{L(2\|x_{k-1} - x^*\| + \frac{\|s_{k-1}\|}{2})\|s_{k-1}\|}{\|F(x_{k-1})\|}. \end{aligned}$$

Mit Lemma 2.2.11 und Lemma 2.2.14 erhält man

$$\eta_k \leq L8\mu(\alpha+\mu)\left\|F(x_{k-1})\right\|.$$

Damit ergibt sich

$$\begin{aligned}
\left\|F(x_{k+1})\right\| &\overset{\text{Lemma 2.2.15}}{\leq} (\eta_k + B\left\|F(x_k)\right\|)\left\|F(x_k)\right\| \\
&\leq (8L\mu(\alpha+\mu)\left\|F(x_{k-1})\right\| + B\left\|F(x_k)\right\|)\left\|F(x_k)\right\| \\
&\overset{IV}{\leq} ([8L\mu(\alpha+\mu)+B]\,\delta)\left\|F(x_k)\right\| \\
&\overset{(2)}{\leq} \tau\left\|F(x_k)\right\| \leq \delta,
\end{aligned}$$

d.h. es gilt $x_{k+1} \in B_{\bar{\epsilon}}(x^*)$ nach dem oben Gezeigtem.

Daraus kann man

$$\{x_k\} \subset B_{\bar{\epsilon}}(x^*) \subset B_\epsilon(x^*)$$

entnehmen. Desweiteren sieht man, dass $\left\|F(x_{k+1})\right\| \leq \tau\left\|F(x_k)\right\|$ für alle $k \geq 0$ gilt.
Also folgt $\left\|F(x_k)\right\| \to 0$ und damit $F(x_k) \to 0$. Aus Lemma 2.2.11 folgt

$$\left\|x_k - x^*\right\| \leq \alpha\left\|F(x_k)\right\| \to 0$$

und damit $x_k \to x^*$.
Aus dem Induktionsbeweis ersieht man, dass, für $k \geq 1$, $\left\|F(x_k)\right\| \leq \left\|F(x_{k-1})\right\|$ und damit gilt:

$$\begin{aligned}
\left\|F(x_{k+1})\right\| &\leq (8L\mu(\alpha+\mu)\left\|F(x_{k-1})\right\| + B\left\|F(x_k)\right\|)\left\|F(x_k)\right\| \\
&\leq (8L\mu(\alpha+\mu)+B)\left\|F(x_{k-1})\right\|\left\|F(x_k)\right\|.
\end{aligned}$$

Mit Lemma 2.2.11 folgt

$$\begin{aligned}
\left\|x_{k+1} - x^*\right\| \leq \alpha\left\|F(x_{k+1})\right\| &\leq \alpha(8L\mu(\alpha+\mu)+B)\left\|F(x_{k-1})\right\|\left\|F(x_k)\right\| \\
&\leq \underbrace{\alpha^3(8L\mu(\alpha+\mu)+B)}_{=:t}\left\|x_{k-1}-x^*\right\|\left\|x_k-x^*\right\|.
\end{aligned}$$

Betrachtet man erneut die vorletzte Zeile, so folgt die zweite Ungleichung für $\left\|x_{k-1} - x^*\right\|$.

$$\begin{aligned}
\left\|x_{k+1} - x^*\right\| &\leq \alpha(8L\mu(\alpha+\mu)+B)\left\|F(x_{k-1})\right\|\left\|F(x_k)\right\| \\
&\leq \alpha(8L\mu(\alpha+\mu)+B)\left\|F(x_{k-1})\right\|^2 \\
&\leq t\left\|x_{k-1}-x^*\right\|^2.
\end{aligned}$$

□

Man beachte, dass es sich bei den wichtigen Sätzen 2.2.1 und 2.2.10 dieses Abschnittes nur um lokale Aussagen handelt. Der Rechenaufwand wird durch den inexakten Algorithmus wesentlich verringert, allerdings erhält man keine quadratische Konvergenz. Durch geeignete Wahl der Forcing-Terme kann man superlineare Konvergenz garantieren, die den inexakten Algorithmus zu einer attraktiven Alternative der Picard-Iteration macht. Die Picard-Iteration hat allerdings einen größeren Konvergenzradius. Daher muss man den Newton-Algorithmus globalisieren.

Kapitel 3

Globalisiertes Newton-Verfahren

Inhalt dieses Kapitels ist es, das inexakte Newton-Verfahren zu globalisieren, d.h. das Verfahren so zu erweitern, dass man mit einem beliebigem Startwert beginnen kann. Dabei soll die Konvergenzordnung möglichst erhalten bleiben. Wichtigstes theoretisches Ergebnis dieses Kapitels ist Satz 3.1.3, in dem die Konvergenz des globalisierten Vefahrens bewiesen wird. Als Anwendung werden zwei konkrete globalisierte Verfahren vorgestellt.

3.1 Bedingungen für globale Konvergenz

Die folgenden Sätze und Beweise sind zum größten Teil [EW94] entnommen. Eine weitere gute Übersicht bieten Brown und Saad in [BS89].

Als Erstes braucht man die Definition eines stationären Punktes.

Definition 3.1.1
$x \in \mathbb{R}^n$ *heißt* stationärer Punkt *von* $\|F\|$, *falls*

$$\|F(x)\| \leq \|F(x) + F'(x)s\| \qquad \forall s \in \mathbb{R}^n.$$

Aus Abschnitt 2.2.2 kann man folgende Aussagen ziehen:

Bemerkung 3.1.2

(1) Ein inexakter Newton-Schritt s_k existiert für beliebiges $\eta_k \in [0,1)$ genau dann, wenn $F(x_k) \in R(F'(x_k))$, da sonst die Bedingung $\|F(x_k) + F'(x_k)s_k\| \leq \eta_k \|F(x_k)\|$ im Allgemeinen nicht erfüllbar ist.

(2) Es gibt $\eta_k \in [0,1)$, so dass ein inexakter Newton-Schritt s_k genau dann existiert, wenn $F(x_k) = 0$ oder x_k kein stationärer Punkt von $\|F\|$ ist.

Damit erhält man den folgenden Algorithmus, dessen globale Konvergenz im Anschluß bewiesen wird.

ALGORITHMUS 3.1 GIN: Globales inexaktes Newton-Verfahren

Sei x_0 und $t \in (0,1)$ gegeben.
for $k = 0, 1, \ldots$ until Konvergenz **do**
  Wähle $\eta_k \in (0,1)$ und s_k, so dass
    $\|F(x_k) + F'(x_k)s_k\| \leq \eta_k \|F(x_k)\|$ und
    $\|F(x_k + s_k)\| \leq [1 - t(1 - \eta_k)] \|F(x_k)\|$.
  Setze $x_{k+1} := x_k + s_k$.
end for

Satz 3.1.3
Man nehme an, dass der Algorithmus GIN nicht abbricht, d.h. für jedes η_k ein s_k berechnet werden kann. Sei weiter $\sum_{k=0}^{\infty}(1-\eta_k)$ divergent.

(1) Dann gilt $F(x_k) \to 0$.

(2) Falls x^* ein Limes-Punkt von $\{x_k\}$ ist, für den $F'(x^*)$ invertierbar ist, so gilt $F(x^*) = 0$ und $x_k \to x^*$.

Die Bedingung, dass $\sum_{k=0}^{\infty}(1-\eta_k)$ divergiert, impliziert $(1-\eta_k) \nrightarrow 0$ bzw. $(1-t(1-\eta_k)) \nrightarrow 1$. D.h. falls $\sum_{k=0}^{\infty}(1-\eta_k)$ divergiert, erhält man eine sich verstärkende Reduktion der Norm von F. Damit wird $F(x_k) \to 0$ gewährleistet. Wie diese Reduktion im k-ten Schritt genau aussieht, wird im nächsten Lemma gezeigt.

Lemma 3.1.4
Für t, $\eta_j \in [0,1)$ gilt

$$\prod_{j=0}^{k-1}[1-t(1-\eta_j)] \leq \exp\left(-t\sum_{j=0}^{k-1}(1-\eta_j)\right) \forall k \geq 1.$$

Beweis:
Es gilt $e^x \geq 1 + x + \frac{x^2}{2} + \cdots + \frac{x^{2n-1}}{(2n-1)!}$ für $n \geq 1$ (vgl. [Heu03], S.361). Damit gilt für $x := -t(1-\eta_j)$ und $n = 1$

$$e^{-t(1-\eta_j)} \geq 1 + (-t(1-\eta_j)) = 1 - t(1-\eta_j).$$

Dann ergibt sich

$$\begin{aligned}\prod_{j=0}^{k-1}[1-t(1-\eta_j)] \leq \prod_{j=0}^{k-1}\exp\left(-t(1-\eta_j)\right) &= \exp\left(\sum_{j=0}^{k-1}(-t(1-\eta_j))\right) \\ &= \exp\left(-t\sum_{j=0}^{k-1}(1-\eta_j)\right).\end{aligned}$$

□

Damit kann man den Beweis von Satz 3.1.3 führen.

Beweis von Satz 3.1.3:
(1) Aus der 2. Bedingung von Algorithmus GIN folgt

$$\begin{aligned}\|F(x_k)\| &\leq [1-t(1-\eta_{k-1})]\,\|F(x_{k-1})\| \\ &\leq \prod_{j=0}^{k-1}[1-t(1-\eta_j)]\,\|F(x_0)\| \\ &\overset{\text{Lemma 3.1.4}}{\leq} \exp\left(-t\sum_{j=0}^{k-1}(1-\eta_j)\right)\|F(x_0)\|.\end{aligned}$$

Da $\|F(x_0)\|$ konstant ist, $\sum_{j=0}^{k-1}(1-\eta_j)$ divergiert und $t > 0$ gilt, konvergiert $\exp\left(-t\sum_{j=0}^{k-1}(1-\eta_j)\right)$ gegen $\exp(-\infty) = 0$. Es folgt

$$\|F(x_k)\| \to 0 \text{ und damit } F(x_k) \to 0.$$

(2) Da x^* ein Häufungspunkt von $\{x_k\}$ ist, ist $F(x^*)$ aufgrund der Stetigkeit von F Häufungspunkt von $\{F(x_k)\}$. Da $F(x_k)$ gegen Null konvergiert, ist 0 einziger Häufungspunkt von $\{F(x_k)\}$, d.h. $F(x^*) = 0$.

Setze $K := \left\|F'(x^*)^{-1}\right\|$. Wähle $\delta > 0$ so klein, dass für alle $y \in B_\delta(x^*)$ gilt:

(a) $F'(y)$ existiert,

(b) $\left\|F'(y)^{-1}\right\| \leq 2K$ und

(c) $\|F(y) - F(x^*) - F'(x^*)(y - x^*)\| \leq \frac{1}{2K} \|y - x^*\|$.

Dieses δ existiert nach Bemerkung 2.2.7 und Lemma 2.2.8. Für $y \in B_\delta(x^*)$ gilt

$$\begin{aligned} \|F(y)\| &= \left\| F(y) - \overbrace{F(x^*)}^{=0} - F'(x^*)(y - x^*) + F'(x^*)(y - x^*) \right\| \\ &\geq \|F'(x^*)(y - x^*)\| - \|F(y) - F(x^*) - F'(x^*)(y - x^*)\| . \end{aligned}$$

Nach Erweiterung mit 1 und mit (b) ergibt sich

$$\begin{aligned} \|F(y)\| &\overset{(b)}{\geq} \frac{\left\|F'(x^*)^{-1}\right\|}{\|F'(x^*)^{-1}\|} \|F'(x^*)(y - x^*)\| - \frac{1}{2K} \|y - x^*\| \\ &\geq \left\|F'(x^*)^{-1}\right\|^{-1} \|y - x^*\| - \frac{1}{2K} \|y - x^*\| \\ &= \left((\underbrace{\left\|F'(x^*)^{-1}\right\|}_{=K})^{-1} - \frac{1}{2K} \right) \|y - x^*\| \\ &= \frac{1}{2K} \|y - x^*\| . \end{aligned}$$

D.h. $\|y - x^*\| \leq 2K \|F(y)\|$ für $y \in B_\delta(x^*)$.
Sei nun $\epsilon \in \left(0, \frac{\delta}{4}\right)$ gegeben. Setze $S_\epsilon := \left\{y \middle| \|y - x^*\| < \frac{\delta}{2} \text{ und } \|F(y)\| < \frac{\epsilon}{2K}\right\}$.
$F(x^*)$ ist Grenzwert von $\{F(x_k)\}$ und x^* Limes-Punkt von $\{x_k\}$. D.h. es gibt ein $\tilde{k}$ (genügend groß) so, dass $x_{\tilde{k}} \in S_\epsilon$.

Durch vollständige Induktion nach k zeigt man $x_k \in S_\epsilon$ für alle $k \geq \tilde{k}$

Sei also $x_k \in S_\epsilon$. D.h. $\|x_k - x^*\| < \frac{\delta}{2}$ und $\|F(x_k)\| < \frac{\epsilon}{2K}$. Dann gilt

$$\begin{aligned} \|s_k\| &= \left\|F'(x_k)^{-1} \left(-F(x_k) + [F(x_k) + F'(x_k)s_k]\right)\right\| \\ &\leq \left\|F'(x_k)^{-1}\right\| \left(\|F(x_k)\| + \|F(x_k) + F'(x_k)s_k\|\right) \\ &\overset{\text{Vor.}}{\leq} 2K \left(\|F(x_k)\| + \|F(x_k)\|\right) \\ &\overset{\text{Vor.}}{\leq} 2K \left(2\frac{\epsilon}{2K}\right) = 2\epsilon. \end{aligned}$$

Da $\epsilon \in \left(0, \frac{\delta}{4}\right)$ ist, erhält man

$$\|s_k\| \leq \frac{\delta}{2}.$$

Aus der Induktionsvoraussetzung und der obigen Abschätzung folgt

$$\|x_{k+1} - x^*\| = \|x_k + s_k - x^*\| \leq \|x_k - x^*\| + \|s_k\| < \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Nach Voraussetzung gilt

$$\|F(x_{k+1})\| \leq \|F(x_k)\| \leq \frac{\epsilon}{2K}.$$

Es folgt dann

$$\begin{aligned} \|x_{k+1} - x^*\| &\leq 2K \|F(x_{k+1})\| \\ &\leq 2K \frac{\epsilon}{2K} \\ &= \epsilon < \frac{\delta}{4}. \end{aligned}$$

Es gilt $x_k \in S_\epsilon \subset B_\delta(x^*)$ für $k \geq \tilde{k}$ und $\|F(x_k)\| \to 0$. Damit folgt

$$\|x_k - x^*\| \leq 2K \|F(x_k)\| \to 0,$$

also $x_k \to x^*$. □

Die Bedingung, dass $\sum_{k=0}^{\infty}(1 - \eta_k)$ divergiert, wird von den betrachteten Verfahren automatisch erfüllt, d.h. man muss sie nicht explizit fordern.

Eine wichtige Bedingung ist, dass die Norm von F durch den Newton-Schritt reduziert wird. Globalisierte Verfahren prüfen dies und entscheiden dann, ob der Schritt akzeptiert wird oder ob der Schritt modifiziert werden muss.

Um die Notation etwas zu erleichtern, werden die Begriffe der aktuellen und der zu erwartenden Reduktion eingeführt.

1. $ared_k$ ist die aktuelle Reduktion. Es berechnet sich durch

$$ared_k := \|F(x_k)\| - \|F(x_k + s_k)\| .$$

2. $pred_k$ ist die zu erwartende Reduktion. Es berechnet sich durch

$$pred_k := \|F(x_k)\| - \|F(x_k) + F'(x_k)s_k\| .$$

Dazu noch einige Bemerkungen

1. Die Bedingung

$$ared_k \geq t(1 - \eta_k) \|F(x_k)\|$$

ist äquivalent zu der Bedingung

$$\|F(x_k + s_k)\| \leq [1 - t\,(1 - \eta_k)] \|F(x_k)\| .$$

2. Gilt nun $ared_k \geq t\, pred_k$ und $\|F(x_k) + F'(x_k)s_k\| \leq \eta_k \|F(x_k)\|$, so folgt aus den Definitionen von $ared_k$ und $pred_k$, dass

$$\begin{aligned} \|F(x_k)\| - \|F(x_k + s_k)\| &\geq t\,(\|F(x_k)\| - \|F(x_k) + F'(x_k)s_k\|) \\ \Rightarrow (1 - t) \|F(x_k)\| + t \|F(x_k) + F'(x_k)s_k\| &\geq \|F(x_k + s_k)\| \\ \Rightarrow (1 - t) \|F(x_k)\| + t\eta_k \|F(x_k)\| &\geq \|F(x_k + s_k)\| \\ \Rightarrow \|F(x_k + s_k)\| &\leq [1 - t\,(1 - \eta_k)] \|F(x_k)\| . \end{aligned}$$

D.h. die 2. Bedingung aus Algorithmus GIN kann durch $ared_k(s_k) \geq tpred_k(s_k)$ ersetzt werden, sofern die 1. Bedingung erfüllt ist.

3.2 Einteilung der globalisierten Verfahren

Bei den globalisierten Verfahren unterscheidet man grundsätzlich zwischen zwei Arten:

1. Die *Trust-Region-Verfahren* verwenden den Newton-Schritt s_k, variieren Länge und Richtung, bis eine starke Abstiegsrichtung des Residuums gefunden ist.

2. Die *Backtracking-Verfahren* verwenden ebenfalls s_k. Sie variieren allerdings nur dessen Länge so, dass er die Bedingung für globale Konvergenz erfüllt.

Beide Verfahren haben ihre Vor- und Nachteile:

1. Die Trust-Region-Verfahren verändern auch die Richtung und können so bessere Abstiegsrichtungen erhalten. Dafür sind sie schwer zu implementieren.

2. Die Backtracking-Verfahren sind leicht zu implementieren. Allerdings sind sie an die Richtung des Schrittes s_k gebunden.

In [PSSW06] findet man einen kurzen Überblick über diese Verfahren und deren numerische Vergleiche. Neben dem Backward-Facing-Step und dem thermalen Konvektions-Problem wird auch das Lid-Driven-Cavity-Problem untersucht. In [STW97] von Shadid, Tuminaro und Walker findet man eine etwas genauere Analyse des Backtracking-Verfahrens auf die gleichen Probleme. Mögliche Fehler des Backtracking-Verfahrens werden in [TWS02] genauer untersucht.

Im Folgenden werden zunächst allgemeine Ergebnisse über diese beiden Verfahrens-Arten vorgestellt und jeweils ein konkreter Algorithmus genauer untersucht.

3.3 Trust-Region-Verfahren

Die Ausführungen dieses Abschnittes beruhen zum größten Teil auf dem Buch [DS83].

3.3.1 Motivation und Einführung

Die Herleitung des Verfahrens erfolgt für das exakte Newton-Verfahren. Die Änderungen und Schwierigkeiten, die beim inexakten Newton-Verfahren auftauchen, werden im Anschluß erläutert.
Bei den Trust-Region-Verfahren versucht man die Norm des lokalen linearen Modells
$\|F(x) + F'(x)s\|$ in einer Kugel um den aktuellen Wert zu berechnen. Den Radius δ dieser Kugel nennt man *Trust-Region-Radius*.

Als Norm verwendet man die l_2-Norm, so dass $\|F\|_2^2 = F(x)^T F(x)$ gilt. Im Folgenden werden die Indizes der Norm weggelassen, da es nicht zu Verwechslungen kommen kann.

Um eine Abstiegsrichtung der Norm zu finden, definiert man sich zunächst die Funktion $f(x) := \frac{1}{2}F(x)^T F(x)$. Der Faktor $\frac{1}{2}$ kommt nur dazu, um die Rechnungen zu vereinfachen. Zusammenhänge zwischen dem Problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{3.1}$$

und dem ursprünglichen Problem, eine Nullstelle von F zu finden, sind in der folgenden Bemerkung aufgeführt.

Bemerkung 3.3.1

(1) Eine Nullstelle von F löst auch das Problem (3.1). Die Umkehrung ist nicht notwendigerweise richtig.

(2) Man könnte versuchen, (3.1) an Stelle des Nullstellenproblems zu lösen. Allerdings wird die Struktur des Problems nicht genutzt. D.h. man modifiziert (3.1).

Für eine Abstiegsrichtung p von (3.1) gilt

$$\nabla f(x)^T p < 0.$$

Dabei ist

$$\begin{aligned} \nabla f(x) &= \frac{d}{dx} \sum_{i=1}^{n} \frac{1}{2} \left(F_i(x)\right)^2 = \sum_{i=1}^{n} F_i(x) \frac{d}{dx} Fi(x) \\ &= F'(x)^T F(x). \end{aligned}$$

Damit ist die stärkste Abstiegsrichtung durch $-F'(x)^T F(x)$ gegeben.

Um die Struktur des Problems in das Problem einfließen zu lassen, definiert man

1. $M(x,s) := F(x) + F'(x)s$ und damit
2. $m(x,s) := \frac{1}{2} M(x,s)^T M(x,s) = \frac{1}{2} \|M(x,s)\|^2$.

D.h. $m(x,s)$ ist $\frac{1}{2} \|F(x) + F'(x)s\|^2$, also gerade der Wert, den man bei Trust-Region-Verfahren minimieren möchte. Es ergibt sich

$$\begin{aligned} m(x,s) &= \frac{1}{2} \left(F(x) + F'(x)s\right)^T \left(F(x) + F'(x)s\right) \\ &= \frac{1}{2} \left(F(x)^T F(x) + F(x)^T F'(x)s + s^T F'(x)^T F(x) + s^T F'(x)^T F'(x)s\right) \\ &= \frac{1}{2} F(x)^T F(x) + \left(F'(x)^T F(x)\right)^T s + \frac{1}{2} s^T F'(x)^T F'(x)s. \end{aligned}$$

Um die Wohldefiniertheit dieser Definition zu zeigen, betrachte man den Gradienten bei $s = 0$.
Es gilt

$$\nabla m(x,0) = \nabla \left(\frac{1}{2} F(x)^T F(x)\right) = \nabla f(x).$$

D.h. die Abstiegsrichtungen von f und m sind für $s = 0$ identisch. Damit kann man die folgende Aussage über den Newton-Schritt s_k machen.

Lemma 3.3.2
Der exakte Newton-Schritt s_k löst das Problem

$$\min_{s \in \mathbb{R}^n} m(x_k, s). \tag{3.2}$$

Beweis:
Nach Definition von M gilt $M(x_k, s) = F(x_k) + F'(x_k)s$. Da s_k als Nullstelle von $F(x_k) + F'(x_k)s$ berechnet wird, verschwindet M bei (x_k, s_k) und damit gilt

$$m(x_k, s_k) = \frac{1}{2} M(x_k, s_k)^T M(x_k, s_k) = 0.$$

Für alle anderen s gilt $m(x_k, s) = \frac{1}{2} \|M(x_k, s)\|^2 \geq 0$. D.h. s_k minimiert m. □

D.h. eine Lösung von (3.2) minimiert die Norm von F. Für gegebenes $x \in \mathbb{R}^n$ erhält man folgendes zu lösendes Problem:

$$\min_{s \in B_\delta(x)} m(x,s) = \min_{s \in B_\delta(x)} \left(f(x) + \nabla f(x)^T s + \frac{1}{2} s^T \left(F'(x)^T F'(x)\right) s\right),$$

wobei δ der Trust-Region-Radius ist.

3.3.2 Dogleg-Verfahren

Bei Trust-Region-Methoden sucht man ein geeignetes x_{k+1} in $B_{\delta_k}(x_k)$. Im Newton-Algorithmus berechnet man x_{k+1} als Summe von x_k und dem Newton-Schritt s_k. D.h. es genügt das äquivalente Problem zu betrachten, ein geeignetes s_k in $B_{\delta_k}(0)$ zu finden.

Beim Dogleg-Verfahren definiert man die *Dogleg-Kurve* Γ_k. Γ_k verbindet stückweise linear 0, den Cauchy-Schritt s_k^{CP} und den Newton-Schritt $s_k^N := -F'(x_k)F(x_k)$. Der Cauchy-Schritt ist als Minimum des lokalen linearen Modells in Richtung des stärksten Abstiegs der Norm von F definiert. Entweder wird der Newton-Schritt s_k^N, falls Γ_k in $B_{\delta_k}(0)$ enthalten ist, oder der Schnittpunkt von Γ_k mit $B_{\delta_k}(0)$ als neuer Schritt gewählt. x_{k+1} erhält man durch Addition dieses neuen Schrittes mit x_k. Die drei möglichen Wahlen werden in Abbildung 3.1 graphisch dargestellt.

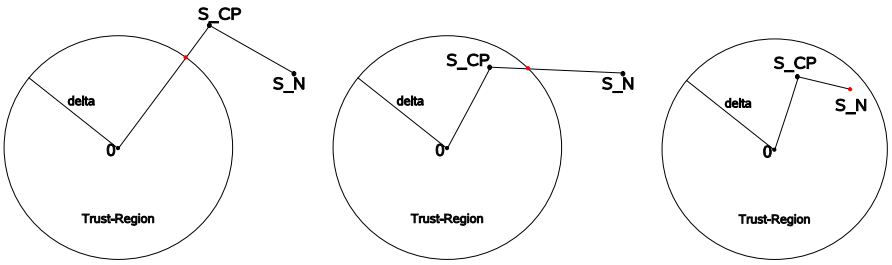

Abbildung 3.1: Die drei möglichen Dogleg-Kurven. Der rote Punkt zeigt den neuen Schritt an.

Es ist gewährleistet, dass es höchstens einen Schnittpunkt gibt, denn $\|s\|$ nimmt auf dem Weg von 0 zu s_k^N monoton zu. Außerdem nimmt die Norm des lokalen linearen Modells auf dem gleichen Weg monoton ab. D.h. der gewählte Punkt ist auch das Minimum desselben auf $\Gamma_k \cap B_{\delta_k}(0)$. Darüber hinaus erhält man, wenn $\delta_k \leq \left\|s_k^{CP}\right\|$ ist, einen Schritt in die stärkste Abstiegsrichtung.
Das sogenannte *Double-Dogleg-Verfahren* wird in [DS83] vorgestellt. Dabei wird ein zusätzlicher Schritt zwischen Cauchy-Schritt und Newton-Schritt auf Γ_k durchlaufen.

Um das Ganze formal zu zeigen, benötigt man die folgenden Definitionen.

Definition 3.3.3

(1) Der Cauchy-Schritt s_k^{CP} *ist definiert durch*

$$\begin{aligned} s_k^{CP} &:= \arg\min_{\lambda\in\mathbb{R}} \|F(x_k) + F'(x_k)(-\lambda\nabla f(x_k))\| \\ &= \arg\min_{\lambda\in\mathbb{R}} \|F(x_k) + F'(x_k)(-\lambda\nabla m(x_k,0))\|\,. \end{aligned}$$

(2) Der Cauchy-Punkt x_k^{CP} *berechnet sich wie folgt:* $x_k^{CP} := x_k + s_k^{CP}$.

(3) Das Sehnen-Polygon Γ_k *ist die stückweise lineare Verbindung von 0,* s_k^{CP} *und* s_k^N.

(4) Man sagt, eine Funktion $g : \mathbb{R}^n \to \mathbb{R}$ *nimmt auf* Γ_k *monoton zu bzw. ab, falls* $g(s) \leq g(s')$ *bzw.* $g(s) \geq g(s')$ *und* s *auf dem Weg von 0 zu* s_k^N *auf* Γ_k *vor* s' *erreicht wird.*

Wie man konkret s_k^{CP} berechnen kann, zeigt die folgende Bemerkung.

Bemerkung 3.3.4
Mit

$$\lambda^* = \frac{\|\nabla f(x_k)\|^2}{\nabla f(x_k)^T \left(F'(x_k)^T F'(x_k)\right) \nabla f(x_k)}$$

gilt

(1) $x_k^{CP} = x_k - \lambda^* \nabla f(x_k)$ *und*

(2) $s_k^{CP} = -\lambda^* \nabla f(x_k)$.

Beweis:
$-\nabla f(x_k)$ ist die stärkste Abstiegsrichtung von f und von $m(x_k, \cdot)$. D.h. es wird λ^* so gesucht, dass

$$m(x_k, -\lambda^* \nabla f(x_k)) = \min_{\lambda \in \mathbb{R}} m\left(x_k, -\lambda \nabla f(x_k)\right).$$

Dieses Problem besitzt die eindeutige Lösung

$$\lambda^* = \frac{\|\nabla f(x_k)\|^2}{\nabla f(x_k)^T \left(F'(x_k)^T F'(x_k)\right) \nabla f(x_k)}.$$

□

Das nächste Lemma zeigt, dass es höchstens einen Punkt auf Γ_k mit Norm δ gibt, der gleichzeitig das lokale lineare Modell auf Γ_k in $B_{\delta_k}(0)$ minimiert.

Lemma 3.3.5
$\|s\|$ *nimmt auf* Γ_k *monoton zu, während die Funktion* $m(x_k, s)$ *monoton fällt.*

Um den Beweis nicht zu lang werden zu lassen, zeigt man

Lemma 3.3.6
Seien A eine positiv definite Matrix und $\{\phi_i\}$ *ein vollständiges Orthonormalsystem aus Eigenvektoren von A. Die Eigenwerte seien mit* λ_i *bezeichnet.*
Dann gilt

$$\frac{\|x\|^4}{[x^T A x]\,[x^T A^{-1} x]} \leq 1.$$

Beweis:
Man kann x auf folgende Weise schreiben: $x = \sum_i \alpha_i \phi_i$. Für die quadratische Form $Q_A(x) := x^T A x$ gilt mit dieser Schreibweise $Q_A(x) = \sum_i \alpha_i^2 \lambda_i$ bzw. $Q_{A^{-1}}(x) = \sum_i \alpha_i^2 \lambda_i^{-1}$. Für die Norm von x erhält man $\|x\|^4 = \left(\sum_i \alpha_i^2\right)^2$.
Nach Cauchy-Schwarz gilt $\left(\sum_i \alpha_i^2\right)^2 = \left(\sum_i \alpha_i \sqrt{\lambda_i} \frac{\alpha_i}{\sqrt{\lambda_i}}\right)^2 \leq \left(\sum_i \alpha_i^2 \lambda_i\right) \left(\sum_i \frac{\alpha_i^2}{\lambda_i}\right)$ und damit die Behauptung. □

Nun kann auch der Beweis von 3.3.5 geführt werden.

Beweis von 3.3.5:
(1) Dass s auf Γ_k zwischen 0 und s_k^{CP} monoton zunimmt, ist offensichtlich. D.h. man muss noch zeigen, dass

$$\left\|s_k^N\right\| \geq \left\|s_k^{CP}\right\|.$$

Setze $H := F'(x_k)^T F'(x_k)$. Mit Lemma 3.3.6 gilt

$$\gamma := \frac{\|\nabla f(x_k)\|^4}{[\nabla f(x_k)^T H \nabla f(x_k)]\,[\nabla f(x_k)^T H^{-1} \nabla f(x_k)]} \leq 1$$

und damit

$$\begin{aligned}
\left\|s_k^{CP}\right\| &= \frac{\|\nabla f(x_k)\|^3}{\nabla f(x_k)^T H \nabla f(x_k)} \\
&= \frac{\|\nabla f(x_k)\|^3}{\nabla f(x_k)^T H \nabla f(x_k)} \frac{\left\|\nabla f(x_k)^T H^{-1} \nabla f(x_k)\right\|}{\nabla f(x_k)^T H^{-1} \nabla f(x_k)} \\
&\leq \frac{\|\nabla f(x_k)\|^3}{\nabla f(x_k)^T H \nabla f(x_k)} \frac{\|\nabla f(x_k)\|}{\nabla f(x_k)^T H^{-1} \nabla f(x_k)} \left\|H^{-1} \nabla f(x_k)\right\| \\
&= \frac{\|\nabla f(x_k)\|^4}{\left[\nabla f(x_k)^T H \nabla f(x_k)\right] \left[\nabla f(x_k)^T H^{-1} \nabla f(x_k)\right]} \cdot \\
&\quad \cdot \underbrace{\left\|F'(x_k)^{-1} \left(F'(x_k)^T\right)^{-1} F'(x_k)^T F(x_k)\right\|}_{=\|F'(x_k)^{-1} F(x_k)\| = \left\|s_k^N\right\|} \\
&= \gamma \left\|s_k^N\right\| \leq \left\|s_k^N\right\|.
\end{aligned}$$

(2) Man definiere

$$\begin{aligned}
x_k^{neu}(\lambda) &:= x_k^{CP} + \lambda \left(x_{k+1} - x_k^{CP}\right) \\
&= x_k + s_k^{CP} + \lambda \left(s_k^N - s_k^{CP}\right).
\end{aligned}$$

Damit setze man

$$\tilde{x}_k^{neu}(\lambda) := \left(x_k, s_k^{CP} + \lambda \left(s_k^N - s_k^{CP}\right)\right)$$

und H wie in (1). Dann ist zu zeigen, dass

$$\nabla m(\tilde{x}_k^{neu}(\lambda))^T \left(s_k^N - s_k^{CP}\right) < 0 \text{ für alle } \lambda \in [0, 1).$$

(a)

$$\begin{aligned}
m(\tilde{x}_k^{neu}(\lambda))^T \left(s_k^N - s_k^{CP}\right) &= \left[\nabla f(x_k) + H\left(s_k^{CP} + \lambda(s_k^N - s_k^{CP})\right)\right]^T \cdot \\
&\quad \cdot \left(s_k^N - s_k^{CP}\right) \\
&= \underbrace{\left[\nabla f(x_k) + H s_k^{CP}\right]^T \left(s_k^N - s_k^{CP}\right)}_{\text{unabhängig von } \lambda} \\
&\quad + \lambda \underbrace{\left(s_k^N - s_k^{CP}\right)^T H \left(s_k^N - s_k^{CP}\right)}_{\geq 0},
\end{aligned}$$

d.h. $\nabla m(\tilde{x}_k^{neu})^T \left(s_k^N - s_k^{CP}\right)$ nimmt monoton mit λ zu. Wenn man nun zeigt, dass die Funktion Null bei $\lambda = 1$ annimmt, so weiß man, dass sie vorher negativ war und die Behauptung somit wahr ist.

(b)

$$\begin{aligned}
\nabla m(\tilde{x}_k^{neu}(\lambda))^T \left(s_k^N - s_k^{CP}\right) &= \left[\nabla f(x_k) + H s_k^N\right]^T \left(s_k^N - s_k^{CP}\right) \\
&\overset{\text{Def. } H}{=} \left[\nabla f(x_k) + F'(x_k)^T F'(x_k) s_k^N\right]^T \left(s_k^N - s_k^{CP}\right) \\
&\overset{\text{Def. } s_k^N}{=} \left[\nabla f(x_k) + F'(x_k)^T (-F(x_k))\right]^T \left(s_k^N - s_k^{CP}\right) \\
&= \left[\nabla f(x_k) - \nabla f(x_k)\right]^T \left(s_k^N - s_k^{CP}\right) = 0.
\end{aligned}$$

□

Die theoretischen Grundlagen sind geschaffen. Es bleibt die Frage der Wahl des neuen Schrittes s_k. Dabei geht man folgendermaßen vor.

1. Man prüft, ob s_k^N in $B_{\delta_k}(0)$ liegt.
   Wenn ja, wählt man s_k^N als den neuen Schritt.
   Wenn nein, folgt 2. .
2. Man prüft, ob $\delta_k \leq \left\|s_k^{CP}\right\|$.
   Wenn ja, wählt man $-\delta_k \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$ als den neuen Schritt.
   Wenn nein, folgt 3. .
3. Man wählt den neuen Schritt $s = s_k^{CP} + \lambda(s_k^N - s_k^{CP})$, wobei λ die Gleichung $\delta_k = \left\|s_k^{CP} + \lambda(s_k^N - s_k^{CP})\right\|$ erfüllt.

Eine alternative Strategie der Wahl des Newton-Schrittes, der insbesondere für das inexakte Newton-Verfahren konstruiert wurde, wird in [PSWS05] beschrieben. Ein Vorteil dieser Wahl ist, dass der inexakte Newton-Schritt eventuell überhaupt nicht berechnet werden muss. Dafür wird aber in jedem Fall der Cauchy-Schritt s_k^{CP} berechnet.

3.3.3 Der inexakte Algorithmus

ALGORITHMUS 3.2 INDL: Inexaktes Dogleg-Newton-Verfahren

Seien x_0, $\eta_{max} \in [0,1)$, $t \in (0,1)$, $0 < \theta_{min} < \theta_{max} < 1$ und $0 < \delta_{min} \leq \delta$ gegeben.
for $k = 0, 1, \ldots$ until Konvergenz **do**
  Wähle $\eta_k \in [0, \eta_{max}]$ und s_k^{IN} so, dass
    $\left\|F(x_k) + F'(x_k)s_k^{IN}\right\| \leq \eta_k \|F(x_k)\|$.
  Berechne s_k^{CP} und **bestimme** $s_k \in \Gamma_k$.
  while $ared_k < t\ pred_k$ **do**

    if $\delta = \delta_{min}$ **then**
      stop

    else
      wähle $\theta \in [\theta_{min}, \theta_{max}]$.

    Setze $\delta := \max\{\theta\delta, \delta_{min}\}$.

    Bestimme erneut $s_k \in \Gamma_k$.
  end while
  Setze $x_{k+1} := x_k + s_k$.
end for

Betrachtet man Algorithmus INDL und zieht die Vorüberlegungen mit ein, so sieht man, dass falls die Bedingung der While-Schleife nicht mehr erfüllt ist, die Bedingung

$$\|F(x_k + s_k)\| \quad \geq \quad [1 - t\,(1 - \eta_k)]\,\|F(x_k)\|$$

erfüllt ist. Diese ist bereits aus Abschnitt 3.1 bekannt und sichert hier wie dort die globale Konvergenz.

Verwendet man statt des exakten Newton-Schrittes s_k^N den inexakten Newton-Schritt s_k^{IN}, so treten folgende theoretische Probleme auf.

1. Die monotone Abnahme von $\|F(x_k) + F'(x_k)s_k\|$ auf Γ_k zwischen s_k^{CP} und s_k^{IN} ist für kein $\eta_k \in (0, \eta_{max}]$ gesichert (vgl. linkes Bild in Abbildung 3.2).
2. Die monotone Zunahme von $\|s_k\|$ auf Γ_k zwischen s_k^{CP} und s_k^{IN} ist für kein $\eta_k \in (0, \eta_{max}]$ gesichert (vgl. rechtes Bild in Abbildung 3.2).

Im Fall $\left\|s_k^{CP}\right\| > \delta_k$ kann es bis zu drei Schnittpunkte von Γ_k und $B_{\delta_k}(x_k)$ geben und das heißt wiederum, dass s_k nicht eindeutig bestimmt ist (vgl. rechtes Bild in Abbildung 3.2).

3. Falls η_k so groß gewählt wurde, dass

$$\eta_k \left\|F(x_k)\right\| \geq \left\|F(x_k) + F'(x_k)s_k^{IN}\right\| \geq \left\|F(x_k) + F'(x_k)s_k^{CP}\right\|$$

und

$$\left\|s_k^{IN}\right\| \leq \delta_k \leq \left\|s_k^{CP}\right\|,$$

dann wählt das Verfahren den inexakten Newton-Schritt als neuen Schritt s_k, obwohl $\delta_k \frac{s_k^{CP}}{\left\|s_k^{CP}\right\|}$ eine größere Reduktion von $\|F(x_k) + F'(x_k)s_k\|$ bewirken würde (vgl. rechtes Bild in Abbildung 3.2).

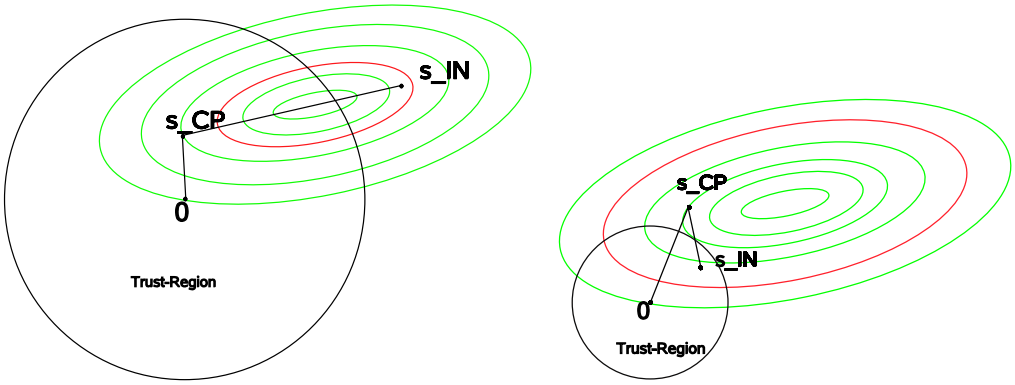

Abbildung 3.2: Illustration einer Dogleg-Kurve. Der schwarze Kreis beschreibt die Trust-Region. Die Ellipsen beschreiben Äquipotential-Linien der Norm des lokalen linearen Modells $F(x_k) + F'(x_k)s$. Die rote Linie repräsentiert $\{s \in \mathbb{R} | \left\|F(x_k) + F'(x_k)s\right\| = \eta_k \left\|F(x_k)\right\|\}$.

3.4 Backtracking-Verfahren

Eine der einfachsten Backtracking-Methoden ist die Folgende:

ALGORITHMUS 3.3 INB: Inexaktes Newton-Verfahren mit Backtracking

Seien x_0, $\eta_{max} \in [0, 1)$, $t \in (0, 1)$ und $0 < \theta_{min} < \theta_{max} < 1$ gegeben.
for $k = 0, 1, \ldots$ until Konvergenz **do**
  Wähle $\eta_k \in [0, \eta_{max}]$ und s_k so, dass
    $\left\|F(x_k) + F'(x_k)s_k\right\| \leq \eta_k \left\|F(x_k)\right\|$.
  while $\left\|F(x_k + s_k)\right\| > [1 - t\,(1 - \eta_k)] \left\|F(x_k)\right\|$ **do**
    Wähle $\theta \in [\theta_{min}, \theta_{max}]$ und
    setze $s_k := \theta s_k$ und
      $\eta_k := 1 - \theta\,(1 - \eta_k)$.
  end while
  Setze $x_{k+1} := x_k + s_k$.
end for

Die while-Bedingung $\left\|F(x_k + s_k)\right\| > [1 - t\,(1 - \eta_k)] \left\|F(x_k)\right\|$ garantiert die globale Konvergenz.

Falls F stetig differenzierbar ist, endet die While-Schleife nach endlich vielen Durchläufen. Die nach der While-Schleife gewählten s_k und η_k erfüllen weiterhin die erste Bedingung. Doch wie ist das θ zu wählen?

Zunächst berechnet man das quadratische Polynom p, dass die folgenden Bedingungen erfüllt

1. $p(0) = \frac{1}{2} \|F(x_k)\|^2$,
2. $p(1) = \frac{1}{2} \|F(x_k + s_k)\|^2$ und
3. $p'(0) = \frac{1}{2} \frac{d}{dt} \left[\|F(x_k + ts_k)\|^2 \right]_{t=0}$.

θ wählt man nun so, dass $p(\theta)$ minimal ist.

Zusammenfassend passiert im $(k-1)$-ten Schritt das Folgende:

> x_k ist aus dem vorherigen Schritt bekannt.
> Man berechnet $F(x_k)$ und minimiert mittels GMRES $\|F(x_k) + F'(x_k)s_k\|$ bis dieses kleiner als die Schranke $\eta_k \|F(x_k)\|$ ist bzw. man berechnet das vorkonditionierte System (vgl. Abschnitt 5.1).
> Mit dem berechneten s_k wird nun die Bedingung aus der While-Schleife überprüft.
>
> > Wenn diese nicht erfüllt ist, wird das θ wie oben beschrieben gewählt und s_k mit Hilfe dieses Faktors verkürzt.
> > Das η_k wird ebenfalls verändert, damit die erste Bedingung, die den Newton-Schritt und damit seine Konvergenz bestimmt, erhalten bleibt.
> > Die Bedingung der While-Schleife wird erneut nachgeprüft und die Schleife eventuell erneut durchlaufen.
>
> Die While-Schleife bricht nach endlich vielen Durchläufen ab. Mit dem so erhaltenen s_k wird der neue Iterationswert x_{k+1} durch $x_k + s_k$ definiert.

Über die Konvergenz des Backtracking-Verfahrens kann man folgende Aussage treffen, die zum großen Teil auf dem Satz 3.1.3 basiert.

Satz 3.4.1
Sei F stetig differenzierbar. Falls der Algorithmus INB die Folge $\{x_k\}$ so erzeugt, dass x^* Limes-Punkt der Folge und $F'(x^*)$ nicht-singulär ist, gilt

$$F(x^*) = 0 \quad \textit{und} \quad x_k \to x^*.$$

Doch bevor man dies beweist, muss das Backtracking-Verfahren etwas genauer betrachtet werden.

Bemerkung 3.4.2
Sei s_k ein von IN erzeugter Newton-Schritt zum Forcing-Term η_k. Dann definiere man

$$\sigma_k(\eta) := \frac{1-\eta}{1-\eta_k} s_k \textit{ für } \eta_k \leq \eta \leq 1$$

als Backtracking-Kurve.
Nun betrachte man das Update des inexakten Newton-Schrittes und des Forcing-Terms innerhalb der While-Schleife des inexakten Backtracking-Verfahrens. Es gilt

$$1 - \theta(1-\eta_k) \leq 1 - (1-\eta_k) \leq \eta_k \textit{ und}$$
$$\sigma_k(1 - \theta(1-\eta_k)) = \frac{\theta(1-\eta_k)}{1-\eta_k} s_k = \theta s_k.$$

D.h. die in INB gewählten Updates erfüllen weiterhin die erste Konvergenzbedingung.

Lemma 3.4.3
Sei im k-ten Schritt des inexakten Backtracking-Verfahrens $F(x_k) \neq 0$ und existiere ein Γ mit

$$\|\tilde{\sigma}_k(\eta)\| \leq \Gamma_k(1-\eta)\|F(x_k)\|, \ \tilde{\eta}_k \leq \eta \leq 1.$$

Sei weiter $\tilde{\eta}_k$ der vom Verfahren berechnete Forcing-Term, dann endet die While-Schleife nach endlich vielen Schritten mit einem Forcing-Term, für den

$$1-\eta_k \geq \min\left\{1-\tilde{\eta}_k, \frac{\theta_{min}\delta}{\Gamma_k\|F(x_k)\|}\right\}$$

gilt. Dabei ist $\delta > 0$ so gewählt, dass

$$\|F(y) - F(x_k) - F'(x_k)(y-x_k)\| \leq \frac{1-t}{\Gamma_k}\|y-x_k\|$$

für $y \in B_\delta(x_k)$ gilt.

Beweis:
(1) Falls $\eta \in [\eta_k, 1]$ so gegeben ist, dass $1-\eta < \frac{\delta}{\Gamma_k\|F(x_k)\|}$. Dann impliziert die erste Bedingung des Lemmas die erste Bedingung des Algorithmus und die zweite Bedingung des Lemmas $\|\tilde{\sigma}_k(\eta)\| < \delta$ und

$$\begin{aligned}\|F(x_k+\tilde{\sigma}_k(\eta))\| &\leq \|F(x_k)+F'(x_k)\tilde{\sigma}_k(\eta)\| + \|F(x_k+\tilde{\sigma}_k(\eta)) - F(x_k) - F'(x_k)\tilde{\sigma}_k(\eta)\| \\ &\leq \eta\|F(x_k)\| + \frac{1-t}{\Gamma_k}\|\tilde{\sigma}_k(\eta)\| \\ &\leq [1-t(1-\eta)]\|F(x_k)\|.\end{aligned}$$

(2) $1-\eta_k$ wird bei jedem Durchlauf durch die While-Schleife um den Faktor $\theta \leq \theta_{min} < 1$ verringert. D.h. die Schleife endet nach endlich vielen Schritten.

(3) Wird die Schleife nicht durchlaufen, d.h. gilt $\eta_k = \tilde{\eta_k}$, so gilt offensichtlich die Bedingung.
Sei also $\eta_k \neq \tilde{\eta}_k$ und η_k^- der vorletzte Schritt. Dann gilt nach (1), dass $1-\eta_k^- \geq \frac{\delta}{\Gamma_k\|F(x_k)\|}$, da sonst die While-Schleife nicht mehr durchlaufen wird. Der neue Schritt ist gegeben durch $1-\eta_k = \theta(1-\eta_k^-)$ für ein $\theta \geq \theta_{min}$. Damit gilt

$$1-\eta_k = \theta(1-\eta_k^-) \geq \theta\frac{\delta}{\Gamma_k\|F(x_k)\|} \geq \frac{\theta_{min}\delta}{\Gamma_k\|F(x_k)\|}.$$

□

Es stellt sich die Frage nach der Existenz einer solchen Kurve $\tilde{\sigma}_k$. Es stellt sich heraus, dass die in Bemerkung 3.4.2 eingeführte Kurve die Bedingungen aus Lemma 3.4.3 erfüllt.

Bemerkung 3.4.4

1. *Im Abschnitt 3.3 wurde gezeigt, dass der Newton-Schritt eine Abstiegsrichtung beschreibt. D.h. für $\sigma_k(\eta) = \frac{1-\eta}{1-\tilde{\eta}_k}\tilde{s_k}$ gilt*

$$\|F(x_k)+F'(x_k)\sigma_k(\eta)\| \leq \eta\|F(x_k)\|, \tilde{\eta}_k \leq \eta \leq 1.$$

2. *Per vollständiger Induktion kann man zeigen, dass $s_k = \sigma_k(\eta_k)$ in der While-Schleife von INB gilt.*

3. Mit $F(x_k) \neq 0$ folgt

$$\begin{aligned}\|\sigma_k(\eta)\| &= \left\|\frac{1-\eta}{1-\eta_k}s_k\right\| \leq \frac{1-\eta}{1-\eta_k}\|s_k\| \\ &= \underbrace{\frac{\|s_k\|}{(1-\eta_k)\|F(x_k)\|}}_{=\Gamma_k}(1-\eta)\|F(x_k)\| \\ &= \Gamma_k(1-\eta)\|F(x_k)\|.\end{aligned}$$

D.h. die erste Bedingung aus Lemma 3.4.3 ist erfüllt für σ_k.

Als Nächstes zeigt man, dass $\sum_{k\geq 0}(1-\eta_k)$ divergiert. Dies war eine Bedingung für die Konvergenz des globalen Verfahrens (vgl. Satz 3.1.3).

Lemma 3.4.5
Man nehme an, dass INB nicht abbricht, weil es einen Schritt nicht berechnen kann. Sei weiter x^* Limes-Punkt der $\{x_k\}$ so, dass es ein von k unabhängiges Γ_k gibt, für das

$$\|\sigma_k(\eta)\| \leq \Gamma_k(1-\eta)\|F(x_k)\|, \ \tilde{\eta}_k \leq \eta \leq 1,$$

erfüllt ist, falls x_k nah genug an x^* liegt. Dann gilt $F(x^*) = 0$ und $x_k \to x^*$. Außerdem gilt $\eta_k = \tilde{\eta}_k$ für große k.

Beweis:
(1) Zunächst beweist man $x_k \to x^*$ durch einen Widerspruchsbeweis. Man nehme also an, dass $x_k \not\to x^*$.

(a) Setze $\delta > 0$ so, dass es unendlich viele k gibt mit $x_k \notin B_\delta(x^*)$, und so, dass $\|s_k\| \leq \Gamma_k(1-\eta_k)\|F(x_k)\|$ gilt, falls $x_k \in B_\delta(x^*)$ und k groß genug ist.

(b) Da x^* Limes-Punkt von $\{x_k\}$ ist, gibt es eine Teilfolge $\{x_{k_j}\}$, die gegen x^* konvergiert und für die $x_{k_j} \in B_{\frac{\delta}{j}}(x^*)$ gilt. Desweiteren gibt es ein $l_j \in \mathbb{R}$, so dass

$$\begin{aligned}&x_{k_j+i} \in B_\delta(x^*) \text{ für } i \in \{0,\ldots,l_j-1\} \text{ und} \\ &x_{k_j+l_j} \notin B_\delta(x^*).\end{aligned}$$

Weiter sei $k_j + l_j < k_{j+1}$.

(c) Für genügend große j gilt

$$\begin{aligned}\frac{\delta}{2} &\leq \left\|x_{k_j+l_j} - x_{k_j}\right\| \\ &= \left\|\sum_{k=k_j}^{k_j+l_j-1} x_{k+1} - x_k\right\| \\ &= \left\|\sum_{k=k_j}^{k_j+l_j-1} s_k\right\| \\ &\leq \sum_{k=k_j}^{k_j+l_j-1}\|s_k\| \\ &\overset{Vor}{\leq} \sum_{k=k_j}^{k_j+l_j-1}\Gamma_k(1-\eta_k)\|F(x_k)\|.\end{aligned}$$

Die Bedingungen

$$\begin{aligned} \|F(x_k + s_k)\| &\leq [1 - t(1 - \eta_k)] \|F(x_k)\| \text{ und} \\ t(1 - \eta_k) \|F(x_k)\| &\leq \|F(x_k)\| - \|F(x_{k+1})\| \end{aligned}$$

sind äquivalent. Damit ergibt sich

$$\begin{aligned} \frac{\delta}{2} &\leq \sum_{k=k_j}^{k_j+l_j-1} \frac{\Gamma_k}{t} \left(\|F(x_k)\| - \|F(x_{k+1})\|\right) \\ &= \frac{\Gamma_k}{t} \left(\left\|F(x_{k_j})\right\| - \left\|F(x_{k_j+l_j})\right\|\right). \end{aligned}$$

Da $k_j + l_j < k_{j+1}$ ist, gilt $\left\|F(x_{k_j+l_j})\right\| \geq \left\|F(x_{k_{j+1}})\right\|$ und damit folgt

$$\frac{\delta}{2} \leq \frac{\Gamma_k}{t} \left(\left\|F(x_{k_j})\right\| - \left\|F(x_{k_{j+1}})\right\|\right).$$

Auf Grund der Stetigkeit von $\|F\|$ konvergiert die rechte Seite gegen 0 für $j \to \infty$. Dies ist ein Widerspruch zur Ungleichung.

(2) Falls $F(x_k)$ für irgendein k verschwindet, gilt auch $F(x_j) = 0$ für alle $j \geq k$. Dann folgt aus der Stetigkeit von F und der Tatsache, dass $x_k \to x^*$, auch $F(x^*) = 0$.

(3) Sei also $F(x_k) \neq 0$ für alle k. Desweiteren sei $\delta > 0$ so klein, dass

(a) $\|\sigma_k(\eta)\| \leq \Gamma_k(1 - \eta) \|F(x_k)\|$, $\tilde{\eta} \leq \eta \leq 1$, falls $x_k \in B_\delta(x^*)$ und

(b) $\|F(y) - F(x) - F'(x)(y - x)\| \leq \frac{1-t}{\Gamma_k} \|y - x\|$, falls x, $y \in B_{2\delta}(x^*)$.

Nach Lemma 3.4.3 bricht die While-Schleife mit

$$1 - \eta_k \geq \min\left\{1 - \tilde{\eta}_k, \frac{\theta_{min}\delta}{\Gamma_k \|F(x_k)\|}\right\} \geq \min\left\{1 - \eta_{max}, \frac{\theta_{min}\delta}{\Gamma_k K}\right\}$$

ab, falls $x_k \in B_\delta(x^*)$ und $K := \sup_{x \in B_\delta(x^*)} \|F(x_k)\|$.

(4) Da $x_k \to x^*$, gibt es unendlich viele $x_k \in B_\delta(x^*)$, d.h. $\sum_{k \geq 0}(1 - \eta_k)$ divergiert. Nach Satz 3.1.3 folgt $F(x^*) = 0$. Damit folgt $F(x_k) \to 0$ und $\frac{\delta\theta_{min}}{\Gamma_k \|F(x_k)\|} \to \infty$, d.h. irgendwann ist $\frac{\delta\theta_{min}}{\Gamma_k \|F(x_k)\|}$ größer als $1 - \tilde{\eta}_k$. Dann gilt $1 - \eta_k \geq 1 - \tilde{\eta}_k$ und $1 - \eta_k = \underbrace{\theta_1 \cdots \theta_m}_{<1}(1 - \tilde{\eta}_k) < 1 - \tilde{\eta}_k$.

Damit wird η_k als $\tilde{\eta}_k$ für große k gewählt. □

Nun kann der Beweis von Satz 3.4.1 geführt werden.

Beweis von Satz 3.4.1:
Setze $K := \left\|F'(x^*)^{-1}\right\|$ und sei $\delta > 0$ so klein, dass

(1) $F'(y)^{-1}$ existiert und

(2) $\left\|F'(y)^{-1}\right\| \leq 2K$, falls $x \in B_\delta(x^*)$.

Sei weiter $x_k \in B_\delta(x^*)$, dann gilt für $\sigma_k(\eta)$ wie im Beweis von Lemma 3.4.3

$$\begin{aligned}
\|\sigma_k(\eta)\| &\leq \left\|F'(x_k)^{-1}\right\| \|F'(x_k)\sigma_k(\eta)\| \\
&= 2K\frac{1-\eta}{1-\eta_k} \|F'(x_k)s_k + F(x_k) - F(x_k)\| \\
&\leq 2K\frac{1-\eta}{1-\eta_k} \left[\underbrace{\|F'(x_k)s_k + F(x_k)\|}_{\leq \eta_k \|F(x_k)\|} + \|F(x_k)\| \right] \\
&= 2K\frac{1+\eta_k}{1-\eta_k}(1-\eta) \|F(x_k)\| \\
&\leq \underbrace{2K\frac{1+\eta_{max}}{1-\eta_{min}}}_{\text{konstant}}(1-\eta) \|F(x_k)\| .
\end{aligned}$$

D.h. die erste Bedingung aus dem Lemma 3.4.3 gilt mit $\Gamma_k = 2K\frac{1+\eta_{max}}{1-\eta_{min}}$ und die Aussage folgt mit Lemma 3.4.5. □

Neben den hier vorgestellten Verfahren gibt es auch hybride Verfahren. Diese zweifach globalisierten Verfahren setzen an den Schwachstellen der vorhandenen Verfahren an. Für das Backtracking-Verfahren werden in [BM01] und [AB07] zwei unterschiedliche Methoden vorgestellt, die eine neue Richtung bestimmen. In [BB07] wird diese Suchrichtung durch eine Trust-Region-Methode berechnet.

Kapitel 4

Navier-Stokes-Problem

Dieses Kapitel behandelt die Navier-Stokes-Gleichungen. Die Variationsformulierung und die Diskretisierung derselben wird nach Einführung der nötigen Hilfsmitteln hergeleitet. Nach einer Existenz-Analyse werden die Picard-Iteration sowie die Verfahren aus Kapitel 3 auf die Navier-Stokes-Gleichungen angewendet.

4.1 Inkompressibles Navier-Stokes-Problem

Die Bewegung von newtonschen Fluiden wie Wasser und Gase werden durch die Navier-Stokes-Gleichungen

$$\rho\frac{\partial u}{\partial t} + \rho\,(u\cdot\nabla)\,u + \nabla\tilde{p} - \eta\Delta u - (\eta+\lambda)\nabla(\nabla\cdot u) = \tilde{f} \tag{4.1}$$

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho u) = 0 \tag{4.2}$$

beschrieben. Hierbei sind u bzw. $\tilde{p}$ die Geschwindigkeit bzw. der Druck des Fluides im betrachteten Gebiet. Mit ρ wird die Dichte des Fluides bezeichnet und λ und η bezeichnen Viskositäten. η wird auch dynamische Viskosität genannt. $\tilde{f}$ ist die von außen wirkende Kraft. Die zweite Gleichung nennt man auch Kontinuitätsgleichung. Eine physikalische Herleitung findet sich in [Wie05].

Für einige Spezialfälle kann man auch analytische Lösungen angeben. Allerdings müssen im Allgemeinen die Gleichungen numerisch gelöst werden. Die Lösung ist stark abhängig von der Reynoldszahl

$$Re = \frac{Lu_\infty\rho}{\eta} = \frac{Lu_\infty}{\nu},$$

wobei L die charakteristische Größe des Problems, z.B. der Durchmesser des Rohres, u_∞ die charakteristische Geschwindigkeit, ρ die Dichte und ν die Viskosität ist. $\nu = \frac{\eta}{\rho}$ heißt auch kinematische Viskosität. Betrachtet man zwei Strömungen mit gleicher Reynoldszahl, so erhält man ein ähnliches Strömungsverhalten, d.h. ist die eine Strömung turbulent, so ist es auch die andere.
Für die inkompressiblen, stationären Gleichungen verschwindet die Zeitableitung. Unter der Annahme, dass die Dichte ρ konstant ist, fällt die Ableitung nach ρ in der Kontinuitätsgleichung (4.2) weg und man kann die Dichte vor die Divergenz ziehen, d.h. (4.2) vereinfacht sich zu

$$\rho\nabla\cdot u = 0 \Leftrightarrow \nabla\cdot u = 0.$$

Durch Substitution $p := \frac{\tilde{p}}{\rho}$ und $f := \frac{\tilde{f}}{\rho}$ und Einsetzen von $\nabla\cdot u = 0$ und $\eta = \frac{\nu}{\rho}$ erhält man die stationären, inkompressiblen Navier-Stokes-Gleichungen:

$$\begin{aligned}(u\cdot\nabla)u + \nabla p - \nu\Delta u &= f \text{ in } \Omega,\\ \nabla\cdot u &= 0 \text{ in } \Omega,\\ u &= 0 \text{ auf } \partial\Omega.\end{aligned}$$

Dabei sind das Geschwindigkeitsfeld u und der Druck p gesucht. Der Einfachheit halber werden homogene Dirichlet-Bedingungen angenommen.

Diskretisiert man die inkompressiblen, instationären Navier-Stokes-Gleichungen implizit in der Zeit, so erhält man einen zusätzlichen Term cu. Dabei ist $c \geq 0$ proportional zur inversen Zeitschrittweite. Man kann diesen Term als zusätzliche Reaktion auffassen. Die Navier-Stokes-Gleichungen mit homogenen Dirichlet-Randbedingungen lauten dann:

$$cu + (u \cdot \nabla)u + \nabla p - \nu \Delta u = f \text{ in } \Omega, \tag{4.3}$$

$$\nabla \cdot u = 0 \text{ in } \Omega, \tag{4.4}$$

$$u = 0 \text{ auf } \partial\Omega. \tag{4.5}$$

4.2 Funktionalanalytische Grundlagen

Im Folgenden werden verschiedene Funktionenräumen benötigt. Die wichtigsten Räume und ihre Eigenschaften werden hier vorgestellt, damit es nicht zu Verwechslungen kommt. Außerdem werden einige funktionanlanalytische Werkzeuge bereitgestellt. Die Ausführungen stammen zum großen Teil aus den Skripten [Lub06a] und [Lub06b].

Als erstes braucht man den Raum der m-fach stetig differenzierbaren Funktionen.

Definition 4.2.1

(1) Sei $\alpha = (\alpha_1, \ldots, \alpha_n)$ ein Vektor mit nichtnegativen ganzen Zahlen α_i. Dann nennt man α Multiindex *mit der* Länge

$$|\alpha| := \sum_{i=1}^{n} \alpha_i.$$

(2) Um die Notation etwas zu vereinfachen, wird die partielle Ableitung der Ordnung α *einer hinreichend oft im Punkt $x \in \Omega \subset \mathbb{R}^n$ differenzierbaren Funktion $u : \Omega \to \mathbb{R}$ in folgender Form geschrieben.*

$$\begin{aligned} D^\alpha u(x) &:= \frac{\partial^{|\alpha|} u}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_n} x_n}(x), \quad \forall\, |\alpha| \geq 1 \text{ und} \\ D^{(0,\ldots,0)} u(x) &:= u(x). \end{aligned}$$

(3) Sei m eine nichtnegative ganze Zahl. Dann ist

$$C^m(\Omega) := \{v : \Omega \to \mathbb{R} | D^\alpha v \in C(\Omega), \text{ für alle } \alpha \text{ mit } |\alpha| \leq m\}$$

die Menge der m-fach auf Ω stetig differenzierbaren Funktionen.

(4) $C^m(\overline{\Omega})$ ist die Menge der Funktionen aus $C^m(\Omega)$ mit stetig auf den Abschluß von Ω, $\overline{\Omega}$, fortsetzbaren Ableitungen bis zur Ordnung m.

Bemerkung 4.2.2
Wenn $\overline{\Omega}$ kompakt ist, bildet $C^m(\overline{\Omega})$ mit der Norm

$$\|u\|_{C^m(\overline{\Omega})} := \max_{|\alpha| \leq m} \sup_{x \in \overline{\Omega}} |D^\alpha u(x)|, \quad u \in C^m(\overline{\Omega})$$

einen Banach-Raum.

Für die folgenden Ausführungen benötigt man außerdem noch die Menge der stetig differenzierbaren Funktionen mit kompaktem Träger.

Definition 4.2.3

(1) Sei die Funktion u auf dem Gebiet Ω definiert. Dann heißt

$$supp(u) := \overline{\{x \in \Omega | u(x) \neq 0\}}$$

Träger von u.

(2) Eine Funktion heißt finit *in Ω, wenn ihr Träger kompakt im Gebiet Ω ist.*

(3) Sei $C^m(\Omega)$ wie in Definition 4.2.1. Dann definiert man

$$C_0^m(\Omega) := \{u \in C^m(\Omega) |\ u \text{ finit}\}.$$

(4) Der Raum $C_0^\infty(\Omega)$ wird auch Raum der Testfunktionen *genannt.*

Diese Räume sind nicht ausreichend, wenn es um die Suche nach verallgemeinerten Lösungen geht. Dazu müssen der Begriff der Lebesgue-Räume und darauf aufbauend die Sobolev-Räume eingeführt werden. Insbesondere sind die unendlich oft differenzierbaren Funktionen mit kompaktem Träger bei der Variationsformulierung wichtig. Sie werden auch Testfunktionen genannt.

Definition 4.2.4

(1) Sei (Ω, B, μ) ein Maß-Raum. Sei v eine nichtnegative und μ-meßbare Funktion. Dann wird durch

$$\int_\Omega v \, d\mu \in [0, \infty]$$

das Lebesgue-Integral *von v auf Ω definiert.*

(2) Eine Teilmenge N von Ω die das Maß 0 hat, heißt Nullmenge.

(3) Sei weiter Y ein Banach-Raum über $\mathbb{R}$ oder $\mathbb{C}$ mit der Norm $\|\cdot\|$, so ist für μ-meßbare Funktionen $u : \Omega \to Y$ auch $\|u\|$ eine μ-meßbare Funktion. Durch

$$\|u\|_{L^p} := \begin{cases} \left(\int_\Omega \|u\|^p \, d\mu\right)^{\frac{1}{p}} & , 1 \leq p < \infty \\ \inf_{\mu(N)=0} \sup_{x \in \Omega/N} \|u(x)\| & , p = \infty \end{cases}$$

*wird die L^p-*Norm *definiert.*

(4) Auf $\widetilde{L}^p(\mu, Y) := \{u : \Omega \to Y |\ u\ \mu - meßbar, \|u\|_{L^p} < \infty\}$ kann man für $1 \leq p \leq \infty$ eine Äquivalenzrelation $\sim$ folgendermaßen definieren. u und v sind genau dann äquivalent, wenn fast überall $u = v$ gilt, d.h nur auf einer Nullmenge gilt $u \neq v$. Die Menge der Äquivalenzklassen auf $\widetilde{L}^p(\mu, Y)$ heißt Lebesgue-Raum L^p.

Bemerkung 4.2.5
L^p bildet mit der Lebesgue-Norm einen Banach-Raum.

Das erste wichtige funktionalanalytische Werkzeug ist die partielle Integration:

Bemerkung 4.2.6
Für $u, v \in C^1(\Omega)$ gilt mit dem äußeren Normaleneinheitsvektor $\nu = (\nu_i)$ auf $\partial\Omega$ für $i = 1, \ldots, n$

$$\int_\Omega \frac{\partial u}{\partial x_i} v \, dx = \int_{\partial\Omega} uv\nu_i \, dx - \int_\Omega u \frac{\partial v}{\partial x_i} \, dx.$$

Um Sobolev-Räume zu definieren, benötigt man den Begriff der verallgemeinerten Ableitung. Dazu braucht man zunächst einen neuen Funktionenraum:

Definition 4.2.7

(1) Man definiert $L^1_{loc}(\Omega)$ wie folgt:

$$L^1_{loc}(\Omega) := \left\{u : \Omega \to \mathbb{R} \text{ oder } \mathbb{C} |\ u \in \mathrm{L}^1(\Omega_0) \text{ für alle kompakten Teilmengen } \Omega_0 \text{ von } \Omega\right\}.$$

(2) $u \in L^1_{loc}(\Omega)$ besitzt die α-te verallgemeinerte Ableitung *$w_\alpha \in L^1_{loc}(\Omega)$, falls*

$$\int_\Omega w_\alpha v \, dx = (-1)^{|\alpha|} \int_\Omega u D^\alpha v \, dx, \qquad \forall v \in C_0^\infty(\Omega)$$

gilt. Man schreibt auch $w_\alpha = D^\alpha u$.

Bemerkung 4.2.8
Dass die verallgemeinerte Ableitung auch wohldefiniert ist, garantiert Lemma 5.4 aus [Dob06]: Die verallgemeinerte Ableitung ist eindeutig, falls sie existiert. Besitzt eine Funktion eine klassische Ableitung, so ist die Funktion auch verallgemeinert differenzierbar und die Ableitungen stimmen überein.

Definition 4.2.9

(1) Sei $1 \le p \le \infty$. Dann heißt die Menge

$$W^{k,p}(\Omega) := \{v \in L^p(\Omega) |\ \exists D^\alpha v \in L^p(\Omega), \forall \alpha \text{ mit } |\alpha| \le k\}$$

Sobolev-Raum *der Funktionen mit verallgemeinerten und zur p-ten Potenz auf Ω integrierbaren Ableitungen bis zur Ordnung k.*

(2) Der Raum $W_0^{k,p}(\Omega)$ ist der Abschluß der Menge $C_0^\infty(\Omega)$ in der Norm

$$\begin{aligned} \|u\|_{W^{k,p}(\Omega)} &:= \left(\sum_{|\alpha|\le k} \|D^\alpha u\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}} \quad \text{für } p < \infty \text{ und} \\ \|u\|_{W^{k,\infty}(\Omega)} &:= \sum_{|\alpha|\le k} \|D^\alpha u\|_{L^\infty(\Omega)} \quad \text{für } p = \infty. \end{aligned}$$

Bemerkung 4.2.10

(1) Sei Ω Gebiet im $\mathbb{R}^n$. Dann sind die Räume $W^{k,p}(\Omega)$ und $W_0^{k,p}(\Omega)$ für $1 \le p < \infty$ mit der Norm $\|\cdot\|_{W^{k,p}(\Omega)}$ Banach-Räume. $W^{k,\infty}(\Omega)$ ist mit $\|\cdot\|_{W^{k,\infty}(\Omega)}$ ebenfalls Banach-Raum.

(2) Sei Ω ein beschränktes, offenes Lipschitz-stetiges Gebiet in $\mathbb{R}^n$. Dann schreibt man für $W^{k,2}$, $k \in \mathbb{N}_{>0}$, auch H^k und für $W_0^{k,2}$ schreibt man H_0^k.

(3) $H_0^1(\Omega)$ bzw. $H^1(\Omega)$ bilden mit dem Skalarprodukt

$$(u, v) = \sum_{|\alpha|} \int_\Omega D^\alpha u(x) D^\alpha v(x) \, dx$$

Hilbert-Räume.

4.3 Variationsformulierung und Diskretisierung

Im Folgenden wird die Variationsformulierung der Navier-Stokes-Gleichungen hergeleitet und eine Lösungstheorie vorgestellt. Anschließend wird das kontinuierliche Problem diskretisiert. Dabei wird insbesondere auf finite Elemente eingegangen.

4.3.1 Variationsformulierung des kontinuierlichen Problems

Man führt die folgenden Räume ein:

1. Geschwindigkeitsraum $V := \left(H_0^1(\Omega)\right)^n$ und
2. Druckraum $Q := L^2_*(\Omega) := \left\{q \in L^2(\Omega) | \int_\Omega q\, dx = 0\right\}$.

Multiplikation der Navier-Stokes-Gleichung (4.3) mit einer Testfunktion $v \in V$ und Integration über Ω liefert

$$\int_\Omega cuv\, dx - \int_\Omega \nu \Delta uv\, dx + \int_\Omega (u \cdot \nabla)uv\, dx + \int_\Omega \nabla pv\, dx = \int_\Omega fv\, dx.$$

Mittels partieller Integration erhält man

$$\underbrace{\int_\Omega cuv\, dx + \nu \int_\Omega \nabla u \nabla v\, dx}_{=:a(u,v)} + \underbrace{\int_\Omega (u \cdot \nabla)uv\, dx}_{=:a_1(u,u,v)} \underbrace{- \int_\Omega p \nabla \cdot v\, dx}_{=:b(p,v)} = \underbrace{\int_\Omega fv\, dx}_{=:\langle \mathbb{F}, v\rangle}$$

mit $a_1(w,u,v) := \int_\Omega (w \cdot \nabla)uv\, dx$. Daraus ergibt sich das Variationsproblem:
Finde die Geschwindigkeit u und der Druck p so, dass

$$\begin{aligned} a(u,v) + a_1(u,u,v) + b(p,v) &= \langle \mathbb{F}, v\rangle \qquad \forall v \in V \\ b(q,u) &= 0 \qquad \forall q \in Q. \end{aligned}$$

Um die Notation etwas zu vereinfachen, wird das Problem umgeschrieben.

(P) *Gesucht ist die Geschwindigkeit u und der Druck p so, dass*

$$\begin{aligned} \tilde{a}(u,u,v) + b(p,v) &= \langle \mathbb{F}, v\rangle \qquad \forall v \in V \\ b(q,u) &= 0 \qquad \forall q \in Q. \end{aligned}$$

Dabei ist $\tilde{a}(w,u,v) = a(u,v) + a_1(w,u,v)$ linear in u und v. Zu den Bilinearformen $\tilde{a}(w,\cdot,\cdot)$ und b definiert man die Operatoren $A(w) : V \to V'$ und $B : V \to Q'$ durch

$$\begin{aligned} \langle A(w)u, v\rangle &= \tilde{a}(w,u,v) \qquad \forall u, v \in V \text{ bzw.} \\ \langle Bv, q\rangle &= b(v,q) \qquad \forall v \in V, \forall q \in Q. \end{aligned}$$

Dann ist das Problem (P) äquivalent zu dem Problem: *Finde $u \in V$ und $p \in Q$ mit*

$$\begin{aligned} A(u)u + B'p &= \mathbb{F} \text{ in } V', \\ Bu &= 0 \text{ in } Q'. \end{aligned}$$

Mit den Definitionen für Z und π aus Definition 4.3.1 kann man das folgende Problem aufstellen.

(Q) *Finde $u \in Z$, so dass*

$$\begin{aligned} \tilde{a}(u,u,v) &= \langle \mathbb{F}, v\rangle \qquad \forall v \in Z \\ \Leftrightarrow \quad \pi A(u)u &= \pi \mathbb{F} \text{ in } Z'. \end{aligned}$$

Falls (u,p) eine Lösung von (P) ist, so löst u auch (Q). Andersherum sichert die Inf-Sup-Bedingung (vgl. Satz 4.3.3) ein eindeutig bestimmtes p für jede Lösung u von (Q), so dass (u,p) (P) löst (vgl. [GR86]).

Um Aussagen über Existenz und Eindeutigkeit für dieses Problem machen zu können, benötigt man die folgenden Definitionen.

Definition 4.3.1

(1) Im Folgenden soll mit Z der Kern von B bezeichnet werden, d.h.

$$Z = \{v |\ v \in V,\ b(v,q) = 0\ \forall q \in Q\}\,.$$

(2) Die orthogonale Projektion von V' auf Z' erhält die Bezeichnung π.

Nun kann man das folgende Existenzresultat zeigen (vgl. [GR86] Theorem 1.2 in Teil IV.).

Satz 4.3.2
Falls für $\tilde{a}$ eine Konstante $\gamma > 0$ existiert mit

$$\tilde{a}(v,v,v) \geq \gamma \|v\|_V^2 \ \ \forall z \in Z,$$

Z separabel und für alle $z \in Z$ die Abbildung

$$u \longmapsto \tilde{a}(u,u,v)$$

schwach folgenstetig auf Z ist, so besitzt das Problem (Q) mindestens eine Lösung $u \in Z$.

Gilt eine der Bedingungen aus Satz 4.3.3, so gibt es zu jeder Lösung u für das Problem (Q) ein eindeutig bestimmtes Paar (u,p), das Problem (P) löst, d.h. unter den Voraussetzungen von Satz 4.3.2 und der Inf-Sup-Bedingung, besitzt das Navier-Stokes-Problem mindestens eine Lösung. Da $\int_\Omega (v\cdot\nabla)vv\ dx = 0$ und $\int_\Omega cv^2\ dx \geq 0$ gelten, folgt

$$\begin{aligned}\tilde{a}(v,v,v) &= \int_\Omega cv^2\ dx + \nu \int_\Omega (\nabla v)^2\ dx + \int_\Omega (v\cdot\nabla)vv\ dx \\ &\geq \nu\,|v|_1^2\,.\end{aligned}$$

Mittels Friedrichs-Ungleichung ergibt sich

$$\tilde{a}(v,v,v) \geq c\,\|v\|_V^2\,.$$

Damit ist die erste Bedingung für die Navier-Stokes-Gleichungen erfüllt. Da die zweite Bedingung offensichtlich erfüllt ist, besitzen die Gleichungen mindestens eine Lösung. Für das allgemeine (nichtlineare) Sattelpunktproblem kann man folgende nützliche Äquivalenz zeigen (vgl. Theorem 0.1 in [Bre74]).

Satz 4.3.3
Seien H_1 und H_2 reelle Hilbert-Räume, $b(\cdot,\cdot)$ eine Bilinearform auf $H_1 \times H_2$, B der durch

$$\left\langle Bh, \tilde{h}\right\rangle = b(h,\tilde{h}) \qquad \forall h \in H_1, \tilde{h} \in H_2$$

definierte und B' der zu B duale Operator.
Dann sind für alle $k > 0$ die folgenden drei Aussagen äquivalent.

$$\begin{aligned}&(1) && \sup_{h\in H_1\setminus\{0\}} \frac{b(h,\tilde{h})}{\|h\|} &&\geq k\left\|\tilde{h}\right\| && \forall \tilde{h} \in H_2,\\ &(2) && \left\|B'\tilde{h}\right\| &&\geq k\left\|\tilde{h}\right\| && \forall \tilde{h} \in H_2,\\ &(3) && \exists S \in \mathcal{L}(H_2', H_1),\ \text{so dass } BS = Id_{H_2'} \text{ und } \|S\| \leq k^{-1}.\end{aligned}$$

Die erste Bedingung wird Inf-Sup- oder auch Babuška-Brezzi-Bedingung genannt. Im Falle der Navier-Stokes-Gleichungen erfüllt b die Bedingung (1), wobei $H_1 = V$ und $H_2 = Q$ (vgl. [GR86] im Beweis von Theorem 5.1 in Teil I.).

Für die Eindeutigkeit braucht man einige weitere Voraussetzungen, die in der folgenden Annahme vereint sind.

Annahme 4.3.4

(a) Die Bilinearform $\tilde{a}(w,\cdot,\cdot)$ sei Z-elliptisch, d.h. es existiert ein $\gamma > 0$, so dass

$$\tilde{a}(w,v,v) \geq \gamma \|v\|_V^2 \qquad \forall v, w \in Z.$$

(b) Die Abbildung $w \longmapsto \pi A(w)$ sei lokal Lipschitz-stetig in Z, d.h. es existiere eine stetige und monoton wachsende Funktion $L : \mathbb{R}^+ \to \mathbb{R}^+$, so dass für alle $\mu > 0$

$$|\tilde{a}(w,u,v) - \tilde{a}(\tilde{w},u,v)| \leq L(\mu) \|u\|_V \|v\|_V \|w - \tilde{w}\|_V$$

für alle u, $v \in Z$ und für alle w, $\tilde{w} \in Z_\mu := \{z \in Z |\ \|z\|_V \leq \mu\}$ gilt.

(c) Es gelte

$$\frac{\|\pi\mathbb{F}\|_{Z'}}{\gamma^2} L\left(\frac{\|\pi\mathbb{F}\|_{Z'}}{\gamma}\right) < 1.$$

Man kann die Abbildung

$$\begin{aligned} T : Z' &\to Z, \\ w &\longmapsto (\pi A(w))^{-1} \end{aligned}$$

definieren, da $\pi A(w)$ unter Annahme 4.3.4 ein Isomorphismus ist. Problem (Q) ist dann äquivalent zu dem Problem: *Finde $u \in Z$ mit*

$$u = T(u)\pi\mathbb{F}.$$

Auf diese Fixpunktgleichung kann man den Satz von Banach 2.1.1 anwenden und erhält eine konvergente Folge.

Satz 4.3.5
Unter Annahme 4.3.4 hat Problem (Q) genau eine Lösung $u \in Z$. Für die Iteration

$$u_{n+1} = T(u_n)\pi\mathbb{F} \tag{4.6}$$

gilt die Fehlerabschätzung

$$\lim_{n\to\infty} \|u - u_n\|_V = 0.$$

Beweis:
(1) $T(w) = (\pi A(w))^{-1}$ ist eine lineare, stetige Funktion von Z' nach Z und es gilt

$$\|T(w)\| = \left\|(\pi A(w))^{-1}\right\| \leq \frac{1}{\|\pi A(w)\|^{-1}} \leq \frac{1}{\gamma}.$$

Man zeigt nun, dass $z \to T(z)\pi\mathbb{F}$ Z in $Z_\mu := \{z \in Z |\ \|z\|_V \leq \mu\}$ abbildet und eine strikte Kontraktion in Z_μ ist, wobei $\mu = \frac{1}{\gamma}\|\pi\mathbb{F}\|_{Z'}$ gilt. Für alle $z \in Z$ erhält man die Abschätzung

$$\|T(z)\pi\mathbb{F}\|_V \leq \|T(z)\| \|\pi\mathbb{F}\|_{Z'} \leq \frac{1}{\gamma}\|\pi\mathbb{F}\|_{Z'} = \mu.$$

D.h. $T(z)\pi\mathbb{F}$ liegt in Z_μ. Um die Kontraktivität zu zeigen, betrachtet man zunächst die Identität

$$\begin{aligned} T(z) - T(\tilde{z}) &= T(z)\underbrace{T(\tilde{z})^{-1}}_{=\pi A(\tilde{z})} T(\tilde{z}) - T(z)\underbrace{T(z)^{-1}}_{=\pi A(z)} T(\tilde{z}) \\ &= T(z)\left[\pi A(\tilde{z}) - \pi A(z)\right] T(\tilde{z}). \end{aligned}$$

Damit erhält man die folgende Abschätzung.

$$\begin{aligned}
\|T(z)\pi\mathbb{F} - T(\tilde{z})\pi\mathbb{F}\|_V &= \|[T(z)\,(\pi A(\tilde{z}) - \pi A(z))\,T(\tilde{z})]\,\pi\mathbb{F}\|_V \\
&\leq \|T(z)\|\,\|\pi A(\tilde{z}) - \pi A(z)\|\,\|T(\tilde{z}\|\,\|\pi\mathbb{F}\|_{Z'} \\
&\overset{\text{An. 4.3.4}}{\leq} \frac{1}{\gamma^2} L\mu\,\|z - \tilde{z}\|_V\,\|\pi\mathbb{F}\|_{Z'} \\
&= \frac{\|\pi\mathbb{F}\|_{Z'}}{\gamma^2} L(\frac{1}{\gamma}\,\|\pi\mathbb{F}\|_{Z'})\,\|z - \tilde{z}\|_V\,.
\end{aligned}$$

Nach Annahme 4.3.4 (c) ist der Faktor vor $\|z - \tilde{z}\|_V$ kleiner als eins und das heißt, dass $z \to T(z)\pi\mathbb{F}$ eine Kontraktion auf Z ist. Die Voraussetzungen von Satz 2.1.1 sind erfüllt und man erhält für beliebige Startwerte $u_0 \in Z$ eine Folge $\{u_{n+1}\}$, deren Folgenglieder durch die Vorschrift 4.6 bestimmt sind und gegen den eindeutig bestimmten Fixpunkt u konvergieren. Dieses u ist die eindeutig bestimmte Lösung von Problem (Q).

(2) Aus Korollar 2.1.2 erhält man die Abschätzung

$$\|u - u_n\|_V \leq \frac{l^n}{1-l}\,\|u_0 - u_1\|_V \overset{n\to\infty}{\to} 0$$

mit $l = \frac{\|\pi\mathbb{F}\|_{Z'}}{\gamma^2} L(\frac{1}{\gamma}\,\|\pi\mathbb{F}\|_{Z'})$. □

Im Falle der Navier-Stokes-Gleichungen kann man für kleine Daten den folgenden Satz beweisen (vgl. [GR86] Theorem 2.2 Teil IV.).

Satz 4.3.6
Sei $n \leq 4$ und Ω ein beschränktes Gebiet mit Lipschitz-stetigem Rand. Gilt zusätzlich

$$\frac{\mathcal{N}}{\nu^2}\,\|\mathbb{F}\| < 1,$$

so besitzt das Problem (Q) genau eine Lösung (u,p) in $Z \times Q$. Dabei ist $\mathcal{N}$ folgendermaßen definiert:

$$\mathcal{N} := \sup_{u,v,w\in Z} \frac{a_1(w,u,v)}{|u|_1\,|v|_1\,|w|_1}.$$

4.3.2 Diskretisierung des kontinuierlichen Problems

Die Diskretisierung wird mittels Finite-Elemente-Methode durchgeführt. Es wird ein kleiner Überblick über die Methode gegeben. Dabei wird sich hier auf Rechteckelemente beschränkt.

Um das Problem zu diskretisieren führt man die Räume V_h und Q_h ein. Dies seien endlichdimensionale Unterräume von V bzw. Q. Dann heißt:
Finde $u_h \in V_h$ und $p_h \in Q_h$ so, dass

$$\begin{aligned}
\tilde{a}(u_h, u_h, v_h) + b(p_h, v_h) &= \langle \mathbb{F}, v_h\rangle \qquad \forall v_h \in V_h \\
b(q_h, u_h) &= 0 \qquad \forall q_h \in Q_h
\end{aligned}$$

Galerkin-Verfahren zur Variationsformulierung (P). Die *Finite-Elemente-Methode* ist eine spezielle Variante dieses Verfahrens.

Zunächst definiert man eine Zerlegung $T_h = \{K_i\}_{i=1}^M$ des beschränkten Gebietes Ω in kompakte Teilgebiete K_i mit folgenden Eigenschaften.

$$\overline{\Omega} = \cup_{j=1}^M \overline{K}_j, \qquad K_i \cap K_j = \emptyset, i \neq j, \qquad h_i := diam(K_i), \qquad h := \max_{i=1,\ldots,M} h_i.$$

K_i ist dabei das Bild der Abbildung eines Referenzelementes $\widehat{K}$ unter der Abbildung F_i. Eine Zerlegung heißt *zulässig*, falls zwei verschiedene abgeschlossene Teilgebiete $\overline{K}_i$ und $\overline{K}_j$ keinen Punkt, genau einen Punkt oder eine gemeinsame Kante/Fläche besitzen.

Ein *finites Element* besteht aus dem Element K, dem Raum der Formfunktionen $\mathcal{P}$ (auf $\widehat{K}$ definierter endlichdimensionaler linearer Funktionenraum der Dimension n) und der Menge der Freiheitsgrade $\sum \subset P'$ (n linear unabhängige Funktionale über $\mathcal{P}$).

Den *Finite-Element-Raum* definiert man lokal auf den einzelnen Elementen $K \in T_h$. Durch die Funktionen F_i kann man sich auf das Referenzelement zurückziehen. Daher definiert man den Raum der Polynome vom Grad $\leq k$ auf $\widehat{K}$:

$$\mathbb{Q}_k(\widehat{K}) := \operatorname{span}\left\{ p : \widehat{K} \to \mathbb{R} |\ p(x_1, \ldots, x_n) = x_1^{\alpha_1} \cdots x_n^{\alpha_n} \text{ mit } 0 \leq \alpha_i \leq k \right\}.$$

Der skalare Finite-Elemente-Raum Q_k auf Ω wird dann folgendermaßen definiert.

$$Q_k := \left\{ v \in L^2(\Omega) |\ v|_{K_i} \circ F_i \in \mathbb{Q}_k(\widehat{K}),\ K_i \in T_h \right\}.$$

Nun kann man die Räume V_h und Q_h einführen. Für das Taylor-Hood-Paar $Q_k - Q_{k-1}$ besitzen sie die Gestalt

$$\begin{aligned} V_h &= [Q_k]^n \cap V \subset V \text{ und} \\ Q_h &= Q_{k-1} \cap Q \cap C(\Omega) \subset Q. \end{aligned}$$

Im Folgenden sollen die Indizes k implizieren, dass es sich um die im k-ten Iterations-Schritt benutzten Größen handeln. Es sei ϕ_i eine Basis von V_h und ψ_i eine von Q_h. Dann ergeben sich folgende Basisdarstellungen:

$$u_k = \sum_{i=1}^{n} u_{ki}\phi_i \quad \text{und} \quad p_k = \sum_{i=1}^{l} p_{ki}\psi_i.$$

Dann definiere man die folgenden Vektoren als Koeffizientenvektoren:

$$u_k := \begin{pmatrix} u_{k1} \\ \vdots \\ u_{kn} \end{pmatrix}, \quad p_k := \begin{pmatrix} p_{k1} \\ \vdots \\ p_{kl} \end{pmatrix} \quad \text{und} \quad f := \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} f(\phi_1) \\ \vdots \\ f(\phi_n) \end{pmatrix}.$$

Nun seien A, $\tilde{A}(u)$, $\widehat{A}(u)$ und B die nach der Ritz-Galerkin-Methode entstandenen Matrizen, d.h.

$$\begin{aligned} A = (A_{ij})_{i,j=1}^{n}, &\qquad \text{wobei } A_{ij} = a(\phi_j, \phi_i), \\ \tilde{A}(u) = \left(\tilde{A}_{ij}(u)\right)_{i,j=1}^{n}, &\qquad \text{wobei } \tilde{A}_{ij}(u) = a_1(u, \phi_j, \phi_i), \\ \widehat{A}(u) = \left(\widehat{A}_{ij}(u)\right)_{i,j=1}^{n}, &\qquad \text{wobei } \widehat{A}_{ij}(u) = a_1(\phi_j, u, \phi_i) \\ \text{und } B = (B_{i,j})_{i,j=1}^{l,n}, &\qquad \text{wobei } B_{ij} = b(\psi_i, \phi_j). \end{aligned}$$

Mit diesen Bezeichnungen erhält man das Systems

$$\begin{pmatrix} A + \tilde{A}(u_k) & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u_k \\ p_k \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}.$$

4.4 Anwendung auf die Picard-Iteration

Betrachtet man die Fixpunktgleichung, die für den Eindeutigkeitsbeweis in Abschnitt 4.3.1 verwendet wurde, so hat man gleich den Ausgangpunkt für die Picard-Iteration (vgl. Kapitel 2). Schreibt man dies in der Formulierung mit der Bilinearform, so erhält man die folgende Iterationsvorschrift.

$$\tilde{a}(u_n, u_{n+1}, v) = \langle \mathbb{F}, v \rangle \qquad \forall v \in V.$$

Erweitert man dieses Schema auf Problem (P), so erhält man für $u_0 \in Z$ das Paar (u_n, p_n) durch Lösen des Systems

$$\begin{aligned} \tilde{a}(u_n, u_{n+1}, v) + b(v, p_{n+1}) &= \langle \mathbb{F}, v \rangle \qquad \forall v \in V, \\ b(u_{n+1}, q) &= 0 \qquad \forall q \in Q. \end{aligned}$$

Äquivalent dazu liefert für beliebiges $u_0 \in V$ mit $\nabla \cdot u_0 = 0$ und $u_0|_{\partial\Omega} = 0$ das Iterationsschema

$$\begin{aligned} c u_{n+1} + (u_n \cdot \nabla) u_{n+1} + \nabla p_{n+1} - \nu \Delta u_{n+1} &= f \qquad \text{in } \Omega \\ \nabla \cdot u_{n+1} &= 0 \qquad \text{in } \Omega \\ u_{n+1} &= 0 \qquad \text{auf } \partial\Omega \end{aligned}$$

eine eindeutige Folge $(u_n, p_n) \in V \times Q$, so dass

$$\lim_{n \to \infty} \left\{ \|u_n - u\|_V + \|p_n - p\|_Q \right\} = 0.$$

Um die in Kapitel 2 eingeführte relaxierte Variante zu erhalten, betrachtet man wieder die Fixpunktgleichung

$$u = T(u)\pi\mathbb{F}.$$

Nun berechnet man u_{n+1} als Summe der alten Iterationswerte u_n und einer Hilfsgröße w_{n+1}, die die Gleichung

$$\begin{aligned} w_{n+1} &= \omega(T(u_n)\pi\mathbb{F} - u_n) \qquad \text{in } Z, \\ \Leftrightarrow A(u_n)\left(\frac{1}{\omega} w_{n+1} + u_n\right) &= \pi\mathbb{F} \qquad \text{in } Z \end{aligned}$$

erfüllt. Analog zu u kann man auch mit p verfahren. D.h. man berechnet (u_{n+1}, p_{n+1}) als Summe von (u_n, p_n) und (w_{n+1}, r_{n+1}), wobei (w_{n+1}, r_{n+1}) die Lösung folgender Variante von Problem (P) ist.

$$\begin{aligned} \tilde{a}(u_n, u_n + \frac{1}{\omega} w_{n+1}, v) + b(p_n + \frac{1}{\omega} r_{n+1}, v) &= \langle \mathbb{F}, v \rangle \qquad \forall v \in V \\ b(q, u_n + \frac{1}{\omega} w_{n+1}) &= 0 \qquad \forall q \in Q. \end{aligned}$$

Da $\tilde{a}$ im zweiten und b im ersten Argument linear sind, kann man die Terme mit den alten Iterationswerten auf die rechte Seite ziehen und erhält das Iterationsschema für u_0 mit $b(q, u_0) = 0$ für alle $q \in Q$.

$$\tilde{a}(u_n, w_{n+1}, v) + b(r_{n+1}, v) = \omega\left(\langle \mathbb{F}, v \rangle - \tilde{a}(u_n, u_n, v) - b(p_n, v)\right) \qquad \forall v \in V \tag{4.7}$$

$$b(q, w_{n+1}) = 0 \qquad \forall q \in Q. \tag{4.8}$$

In der starken Formulierung sieht das ganze folgendermaßen für Startwerte u_0 mit $\nabla \cdot u_0 = 0$ aus.

$$\begin{aligned} c w_{n+1} - \nu \Delta w_{n+1} + (u_n \cdot \nabla) w_{n+1} + \nabla p_{n+1} &= \omega\left(f + c u_n + \nu \Delta u_n - (u_n \cdot \nabla) u_n - \nabla p_n\right) \\ \nabla \cdot w_{n+1} &= 0 \\ u_{n+1} &:= u_n + w_{n+1}. \end{aligned}$$

Um eine Aussage über die Lösbarkeit dieser linearen Probleme machen zu können, betrachtet man das allgemeine lineare Sattelpunktproblem (vgl. [Bre74]): *Finde* (u,p) *in* $V \times Q$ *so, dass*

$$\begin{aligned} d(u,v) + e(v,p) &= \langle f, v\rangle \quad \forall v \in V, \\ e(u,q) &= \langle g, q\rangle \quad \forall q \in Q, \end{aligned}$$

wobei V, Q reelle Hilbert-Räume und $d(\cdot,\cdot)$ und $e(\cdot,\cdot)$ stetige Bilinearformen auf $V \times V$ bzw. $V \times Q$ sind. Die Gleichungen (4.7), (4.8) sind ein Spezialfall dieses Systems. Um die Notation etwas zu vereinfachen, werden folgende Begriffe definiert.

Definition 4.4.1

(1) Zu den Bilinearformen $d(\cdot,\cdot)$ *und* $e(\cdot,\cdot)$ *definiere man die linearen Operatoren* $D : V \to V'$ *und* $E : V \to Q'$ *durch*

$$\langle Du, v\rangle = d(u,v) \quad \forall u, v \in V, \tag{4.9}$$
$$\langle Ev, q\rangle = e(v,q) \quad \forall v \in V,\ \forall q \in Q. \tag{4.10}$$

(2) Eine Bilinearform $d : V \times V \to \mathbb{R}$ *heißt* V-elliptisch, *falls eine Konstante* $\gamma > 0$ *existiert mit*

$$d(v,v) \geq \gamma \|v\|_V^2 \quad \forall v \in V.$$

(3) Definiere $\Lambda : V \times Q \to V' \times Q'$ *folgendermaßen.*

$$\Lambda(v,q) := (Dv + E'q, Ev)\,.$$

Mit den Definitionen 4.3.1 und 4.4.1 kann man die folgenden Resultate zeigen.

Satz 4.4.2
Sei πA *ein Isomorphismus von* Z *nach* Z' *und gebe es ein* $k > 0$ *so, dass* $\|E'q\| \geq k\|q\|$ *für alle* $q \in Q$. *Dann ist* Λ *aus Definition 4.4.1 ein Isomorphismus von* $V \times Q$ *nach* $V' \times Q'$.

Korollar 4.4.3
Sei $d(\cdot,\cdot)$ Z*-elliptisch und gebe es ein* $k > 0$ *so, dass* $\|E'q\| \geq k\|q\|$ *für alle* $q \in Q$.
Dann ist Λ *aus Definition 4.4.1 ein Isomorphismus von* $V \times Q$ *nach* $V' \times Q'$.

Man findet zu jeder vorgegebenen rechten Seite $\mathbb{F}$ ein eindeutig bestimmtes Paar (u,p), d.h. jedes der bei der Picard-Iteration entstandenen linearen Probleme ((4.7), (4.8)) ist eindeutig lösbar.

Um nun ein Resultat für den diskreten Fall zu erhalten, benötigt man einige Definitionen.

Definition 4.4.4

(1) Ein Projektionsoperator von $V' \times Q'$ *nach* $V_h' \times Q_h'$ *werde mit* ρ_h *bezeichnet.*

(2) Dann definiert man $\Lambda_h : V_h \times Q_h \to V_h' \times Q_h'$ *durch*

$$\Lambda_h := \rho_h \Lambda(v_h, q_h) \quad \forall v_h \in V_h, q_h \in Q_h,$$

wobei Λ *der Operator aus Definition 4.4.1 ist.*

(3) Analog zu Z *definiert man* Z_h *durch*

$$Z_h = \{v_h |\ v_h \in V_h,\ e(v_h, q_h) = 0\ \forall q_h \in Q_h\}\,.$$

Für den diskreten Fall kann man folgendes Resultat angeben (vgl. [Bre74] Korollar 2.1.).

Satz 4.4.5
Gibt es eine positive Konstante k_h so, dass

$$\sup_{v_h \in V_h \setminus \{0\}} \frac{e(v_h, q_h)}{\|v_h\|} \geq k_h \|v_h\| \qquad \forall q_h \in Q_h,$$

und eine Konstante $\gamma_h > 0$ mit

$$d(v_h, v_h) \geq \gamma_h \|v_h\|^2 \qquad \forall v_h \in Z_h,$$

so ist Λ_h ein Isomorphismus von $V_h \times Q_h$ nach $V_h' \times Q_h'$. Außerdem gilt für jedes Paar $(f, g) \in V' \times Q'$, falls $(u, p) = \Lambda^{-1}(f, g)$ und $(u_h, q_h) = \Lambda_h^{-1} \rho_h(f, g)$, die Fehlerabschätzung

$$\|u - u_h\| + \|p - p_h\| \leq \sigma_h \left(\inf_{v_h \in V_h} \|u - v_h\| + \inf_{q_h \in Q_h} \|p - q_h\| \right),$$

wobei

$$\sigma_h = \max \left\{ \left(\gamma_h^{-1} + k_h^{-1}(1 + \|D\| \gamma_h^{-1})\right), \left(k_h^{-1} + \|D\| k_h^{-2}\right) \left(1 + \|D\| \gamma_h^{-1}\right) \right\} (\|D\| + \|E\|) + 1.$$

Grad-Div-Stabilisierung
In der Praxis ist es oft sinnvoll, das zu lösende System so zu verändern, dass es günstigere Eigenschaften hat. Diese Variation heißt Stabilisierung. Eine mögliche Variante ist die Grad-Div-Stabilisierung. Dabei wird der skalierte Term $(\nabla \cdot u_h, \nabla \cdot v_h)$ zur Bilinearform $a(u_h, v_h)$ hinzuaddiert. Für Taylor-Hood-Elemente (vgl. Abschnitt 4.3.2)hat sie die folgende Gestalt

$$a(u_h, v_h) = \nu(\nabla u_h, \nabla v_h) + \alpha(\nabla \cdot u_h, \nabla \cdot v_h) + (c u_h, v_h).$$

Man kann $\alpha \in O(1)$ wählen. Die Grad-Div-Stabilisierung ist konsistent und folglich wird die kontinuierliche Lösung nicht verändert. Weitere Ausführungen zur numerischen Analyse finden sich in [OR04, Hei08].

4.5 Anwendung des Newton-Verfahrens

Im Fall der Newton-Iteration setzt man (vgl. Abschnitt 4.3.2 bezüglich der Notation)

$$F \begin{pmatrix} u_k \\ p_k \end{pmatrix} := \underbrace{\begin{pmatrix} A + \tilde{A}(u_k) & B^T \\ B & 0 \end{pmatrix}}_{:= M(u_k)} \begin{pmatrix} u_k \\ p_k \end{pmatrix} - \begin{pmatrix} f \\ 0 \end{pmatrix}$$

und erhält für die Ableitung von F

$$F' \begin{pmatrix} u_k \\ p_k \end{pmatrix} \begin{pmatrix} s_k^1 \\ s_k^2 \end{pmatrix} = \underbrace{\begin{pmatrix} A + \tilde{A}(u_k) & B^* \\ B & 0 \end{pmatrix}}_{= M(u_k)} \begin{pmatrix} s_k^1 \\ s_k^2 \end{pmatrix} + \underbrace{\begin{pmatrix} \widehat{A}(u_k) & 0 \\ 0 & 0 \end{pmatrix}}_{=: M_k} \begin{pmatrix} s_k^1 \\ s_k^2 \end{pmatrix}.$$

Um die Schreibweise etwas zu vereinfachen, führt man die folgende Notation ein:

$$f := \begin{pmatrix} f \\ 0 \end{pmatrix}, \qquad x_k := \begin{pmatrix} u_k \\ p_k \end{pmatrix} \text{ und } \qquad s_k := \begin{pmatrix} s_k^1 \\ s_k^2 \end{pmatrix}.$$

Es gibt mehrere Möglichkeiten, den Startwert x_0 zu wählen. Damit das Newton-Verfahren konvergiert, sollte x_0 nahe der wirklichen Lösung gewählt werden. Dazu kann man z.B. einige Picard-Iterationen durchlaufen oder das zugehörige Stokes-Problem lösen und die Lösung als Anfangswert wählen. Eine Fehleranalyse für den zweiten Fall wird in [SDLS06] gemacht.

Das inexakte Newton-Verfahren mit Dogleg-Verfahren (Algorithmus 3.2) aus Abschnitt 3.3.2 und das inexakte Newton-Verfahren (Algorithmus 3.3) aus Abschnitt 3.4 werden nun auf das diskretisierte Navier-Stokes-Problem angewendet. Dabei werden zur besseren Übersicht einige Teile des Dogleg-Algorithmen ausgelagert. Im Einzelnen sind das die Schrittbestimmung (Algorithmus 4.2), die Veränderung des Schrittes um die Globalitätsbedingung zu erfüllen (Algorithmus 4.3) und das Update des Trust-Region-Radius am Ende eines Newton-Schrittes (Algorithmus 4.4).

Die Bezeichnungen der Variablen sind etwas anders als in den Algorithmen in Kapitel 3. Die Iterationswerte x_k werden mit *solution* bezeichnet. Sie werden nicht gespeichert, sondern in jedem Schritt überschrieben. So benötigt man weniger Speicher. Die Funktion F setzt sich aus der Matrix *matrix* angewendet auf die Stelle x_k, an der F ausgewertet wird und f zusammen (vgl. Abschnitt 4.5). Dabei ist *matrix* die in Abschnitt 4.5 eingeführte Matrix $M(u_k)$, d.h. sie ist von der Stelle x_k abhängig.

4.5.1 Dogleg-Algorithmus

Das Basis-Dogleg-Verfahren besitzt folgende Gestalt.

Algoritmus 4.1 INDL: Inexaktes Newton-Verfahren mit Dogleg (Navier-Stokes)

Seien $\eta_0 \leq \eta_{max} \in [0,1)$, $t \in (0,1)$, $\theta = 0.25$.
Wähle einen Startwert *solution*.
$matrix := M(solution)$;
$y := matrix * solution - f$;
$D := \|y\|$;
for $k = 0, 1, \ldots$ until ($k > 100$ oder $D < 1e-12$) **do**

$\eta := \frac{|\ D-r\ |}{D_{last}}$ für $k \geq 1$.

if $k > 0 \wedge \eta_{last}^{\frac{1+\sqrt{5}}{2}} > 0.1$ **then**

$$\eta := \max\left\{\eta, \eta_{last}^{\frac{1+\sqrt{5}}{2}}\right\};$$

$\eta := \min\{\eta_{max}, \eta\}$;

$funktion := y$;

$D_{last} := D$;

$y := -y$;

$matrix := matrix + M_k$;

$fstrich := matrix$;

Löse bis zu einer Genauigkeit von $D_k\eta_k$ das Gleichungssystem

$$y = matrix * s$$

und erhalte den Vektor s.

$s_{laenge} := \|s\|$;

//Im Folgenden wird das anfängliche δ bestimmt.

if $k = 1$ **then**

if $s_{laenge} < \delta_{min}$ **then**
$\delta := 2\delta_{min}$

else
$\delta := s_{laenge}$

Wende Algorithmus S an. *//Bestimmen der Schrittweite*

Wende Algorithmus GDL an.

$solution := solution + s$;

$r := D_{last} - \|funktion + fstrich * s\|$;

Wende Algorithmus UP an.

end for

In Algorithmus S (Algorithmus 4.2) wird zunächst geprüft, ob der aktuelle Schritt in der Trust-Region liegt. Falls dies nicht der Fall ist, wird ein geeigneter Schritt (vgl. Abschnitt 3.3.2) folgendermaßen gewählt.

Der Cauchy-Schritt wird berechnet und geprüft, ob er in der Trust-Region liegt. Falls dies der Fall ist, wird der Schnittpunkt der Gerade, die den Cauchy-Punkt und den Newton-Schritt verbindet, und der Kugel um 0 mit Radius δ als neuer Schritt gewählt. Im anderen Fall wird ein Schritt in Richtung des Cauchy-Punktes mit Länge δ als neuer Schritt berechnet.

Algoritmus 4.2 S: Bestimmung des Schrittes bei INDL

if $s_{laenge} \geq \delta$ **then**

$s_{CP} = matrix^T funktion;$

$s_{CP} := -\frac{\|s_{CP}\|^2}{\|matrix\ s_{CP}\|^2} s_{CP};$

if $\|s_{CP}\| \geq \delta$ **then**
$s := \frac{\delta}{\|s_{CP}\|} s_{CP};$

else
$\lambda := -\frac{s_{CP}^T(s-s_{CP})}{\|s-s_{CP}\|} + \sqrt{\left(\frac{s_{CP}^T(s-s_{CP})}{\|s-s_{CP}\|}\right)^2 - \frac{\|s_{CP}\|^2-\delta^2}{\|s-s_{CP}\|^2}};$
$s := s_{CP} + \lambda\,(s - s_{CP});$

$s_{laenge} := \|s\|;$

Algorithmus 4.3 berechnet zunächst den Funktionswert des Schrittes $solution + s$, d.h. es wird $F(x_k + s_k)$ ermittelt. Falls $solution + s$ die Globalitätsbedingung $ared < t\,pred$ erfüllt, wird der Schritt nicht weiter verändert. Falls dies nicht der Fall ist, wird der Trust-Region-Radius δ um den Faktor 0.25 reduziert und Algorithmus 4.2 angewendet, bis die Globalitätsbedingung erfüllt ist oder der minimal erlaubte Trust-Region-Radius δ_{min} erreicht wurde.

Algoritmus 4.3 GDL: Veränderung des Schrittes, bis die Globalitätsbedingung erfüllt ist

while true **do**

$matrix := M(solution + s);$

$y := matrix * (solution + s) - f;$

$D := \|y\|;$

if $D_{last} - D < t\ (D_{last} - \|funktion + fstrich * s\|)$ **then**
break;

if $\delta = \delta_{min}$ **then**
stop

$\delta := \max\{0.25\delta, \delta_{min}\}.$

Wende Algorithmus S an.

end while

Während einer Newton-Iteration wird in Algorithmus 4.3 die Trust-Region so lange verkleinert, bis die Globalitätsbedingung erfüllt ist. Da der so entstandene Trust-Region-Radius δ sehr klein sein kann, wird man schnell an den minimalen Trust-Region-Radius δ_{min} gelangen, wenn man δ nicht am Ende des Newton-Schrittes updatet. Insbesondere benötigt man eventuell beim nächsten Schritt einen größeren Radius. Um zu bestimmen, wie mit δ verfahren werden soll, betrachtet man den

Quotienten aus der aktuellen Reduktion *ared* und der zu erwartenden Reduktion *pred* der Norm des linearen Modells der Funktion F.

Ist dieser Quotient kleiner als eine obere Schranke $0 < \rho_s < 1$, so unterscheidet man zwischen zwei Fällen. Liegt der inexakte Newton-Schritt in der Trust-Region, so wird diese auf die Länge des Newton-Schrittes bzw. auf δ_{min} verkürzt, je nachdem was größer ist. Falls ein Punkt auf der Dogleg-Kurve Γ gewählt wurde, so wird die Trust-Region um einen festen Faktor $0 < \beta_s < 1$ reduziert.

Falls der Quotient größer als eine untere Schranke $\rho_s < \rho_e < 1$ ist, d.h. die zu erwartende Reduktion der Norm des linearen Modells der Funktion F kleiner als die aktuelle Reduktion ist, so vergrößert man die Trust-Region um einen Faktor $1 < \beta_e$, falls δ damit nicht größer als der maximal erlaubte Radius δ_{max} wird.

Trifft keiner der beiden Fälle auf, d.h. $\rho_s < \frac{ared}{pred} < \rho_e$, so wird δ nicht verändert.

Algoritmus 4.4 UP: Update von δ

Seien ρ_e, ρ_s, β_e und β_s gegeben.
if *quotient* := $\frac{D_{last}-D}{D_{last}-\|funktion+fstrich*s\|} < \rho_s$ **then**

  if $s_{laenge} < \delta$ **then**
    $\delta := \max\{s_{laenge}, \delta_{min}\}$;

  else
    $\delta := \max\{\beta_s * \delta, \delta_{min}\}$;

else
  if *quotient* $> \rho_e$ und $|s_{laenge} - \delta| < 10^{-9}$ **then**
    $\delta := \min\{\beta_e * \delta, \delta_{max}\}$;

4.5.2 Backtracking-Algorithmus

Im Dogleg-Verfahren muss man die Matrix *matrix* zwischenspeichern, um die zu erwartende Reduktion $pred = \|F(x_k)\| - \|F(x_k) + F'(x_k)s_k\|$, die für die Globalitätsbedingung innerhalb der while-Schleife benötigt wird, zu berechnen. Im Fall des Backtracking-Verfahrens ist dies nicht nötig, da man eine restriktivere Bedingung an die Residuen wählt (vgl. Bemerkung am Ende von Abschnitt 3.1). Dadurch benötigt das Backtracking-Verfahren weniger Speicher als das Dogleg-Verfahren. Zudem gibt es weniger Vektor-Vektor-Produkte.

Beim Dogleg-Verfahren wird die Richtung und die Länge des Newton-Schrittes variiert, beim Backtracking-Verfahren nur die Länge. Daher wird kein Update des Schrittes vor der while-Schleife benötigt. Im eigentlichen Backtracking-Verfahren wird der Reduktionsfaktor Θ adaptiv gewählt. Θ wird als Minimum eines Polynoms zweiten Grades (vgl. Abschnitt 3.4) berechnet.

Algoritmus 4.5 INB: Inexaktes Newton-Verfahren mit Backtracking (Navier-Stokes)

Seien $\eta_0 \leq \eta_{max} \in [0,1)$, $t \in (0,1)$, $0 < \theta_{min} < \theta_{max} < 1$.
Wähle einen Startwert *solution*.
$matrix := M(solution)$;
$y := matrix * solution - f$;
$D := \|y\|$;
for $k = 0, 1, \ldots$ until ($k > 100$ oder $D < 1e-12$) **do**

$\eta := \frac{|\, D-r \,|}{D_{last}}$ für $k \geq 1$.

if $k > 0 \wedge \eta_{last}^{\frac{1+\sqrt{5}}{2}} > 0.1$ **then**

$$\eta := \max\left\{\eta, \eta_{last}^{\frac{1+\sqrt{5}}{2}}\right\};$$

$\eta := \min\{\eta_{max}, \eta\}$;

$temp_1 := y$;

$D_{last} := D$;

$y := -y$;

$matrix := matrix + M_k$;

Löse bis zu einer Genauigkeit von $D_k\eta_k$ das Gleichungssystem

$$y = matrix * s$$

und erhalte den Vektor s.

$tmp_2 := matrix * s$;

//Im Folgenden wird die Schrittweite so verändert, dass sie der Globalitätsbedingung genügt. θ ist der Stauchungsfaktor.

while true **do**

$matrix := M(solution + s)$;

$y := matrix * (solution + s) - f$;

$D := \|y\|$;

if $D \leq [1 - t(1-\eta)]\, D_{last}$ **then**
break;

Berechne $\theta \in [\theta_{min}, \theta_{max}]$ wie in Abschnitt 3.4.

$s := \theta s$;

$\eta := 1 - \theta(1-\eta)$;

end while

$solution := solution + s$;

$r := \|tmp_1 + tmp_2\|$;

end for

Kapitel 5

Numerische Experimente

Die Picard- und die Newton-Iteration werden in diesem Kapitel anhand von zwei Testbeispielen getestet. In den Tests wurde das Krylov-Verfahren GMRES als iterativer Löser innerhalb des Newton-Verfahrens verwendet. Daher wird GMRES am Anfang dieses Kapitels kurz vorgestellt. Darauf folgt eine kurze Darstellung des verwendeten Vorkonditionierers, bevor die Testergebnissen vorgestellt werden.

5.1 Lösung linearer Gleichungssysteme

Nach einer kurzen Einführung in Krylov-Unterraum-Methoden wird das GMRES-Verfahren als ein solches Verfahren genauer untersucht. Die möglichen Vorkonditionierungsstrategien werden vorgestellt, bevor der in den Tests verwendete Vorkonditionierer vorgestellt wird.

5.1.1 Krylov-Verfahren und allgemeine Vorkonditionierung

Eine große Klasse von iterativen Lösungsmethoden eines linearen, algebraischen Systems

$$Ax = b$$

sind die Krylov-Unterraum-Methoden. Sie konstruieren Lösungen mit Hilfe von an die Matrix A angepassten Teilräumen des $\mathbb{R}^n$, den sogenannten Krylov-Unterräumen.

Definition 5.1.1
(1) Seien die Matrix $A \in \mathbb{R}^{n\times n}$ und der Vektor $y \in \mathbb{R}^n \setminus \{0\}$ gegeben. Dann definiert man den zugehörigen Krylov-Unterraum

$$\mathcal{K}_k(A, y) := span\left\{y, Ay, A^2y, \ldots, A^{k-1}y\right\},\ k \in \mathbb{N}.$$

(2) Zum Iterationswert x_k berechnet man das zugehörige Residuum $r_k := b - Ax_k$.

Man unterscheidet zwischen zwei Arten von Krylov-Unterraum-Methoden (vgl. [Lub06b]).

1. Die *Galerkin-Verfahren* konstruieren x_k so, dass das Residuum r_k orthogonal zu $\mathcal{K}(A, r_0)$ oder einem anderen geeigneten Krylov-Raum steht, wobei x_0 der Startwert ist.
2. Die *Minimierungs-Verfahren* minimieren das Residuum r_k in einer passenden Norm auf $\mathcal{K}(A, r_0)$ oder einem anderen geeigneten Krylov-Raum. Ein wichtiger Vertreter dieser Gruppe ist das GMRES-Verfahren.

Hat man seinen Iterationswert erhalten, erhöht man k oder setzt $x_0 := x_k$ und $r_0 := b - Ax_0$. Die zweite Alternative wird auch Restart genannt. Im Folgenden wird sich etwas genauer mit dem GMRES-Verfahren beschäftigt, da dieses Verfahren bei der Implementierung verwendet wurde.

Im Allgemeinen kann die Kondition der Matrix A sehr schlecht sein. Da das GMRES-Verfahren aber stark von dieser abhängt, verändert man das zu lösende System so, dass die Kondition verbessert wird. Es gibt zwei unterschiedliche Vorgehensweisen.

(1) Sei P eine invertierbare Matrix. Bei der Links-Vorkonditionierung wird das System

$$P^{-1}Ax = P^{-1}b$$

anstelle des Originalsystems gelöst, d.h. GMRES minimiert

$$\left\|P^{-1}b - P^{-1}Ax\right\|_2$$

über alle Vektoren aus dem affinen Unterraum

$$x_0 + \mathcal{K}(P^{-1}A, P^{-1}r_0).$$

(2) Sei P eine invertierbare Matrix. Bei der Rechts-Vorkonditionierung wird das System

$$AP^{-1}u = b,\ u = Px$$

anstelle des Originalsystems gelöst, d.h. GMRES minimiert

$$\left\|b - AP^{-1}u\right\|_2$$

über alle Vektoren u aus dem affinen Unterraum

$$u_0 + \mathcal{K}(AP^{-1}, r_0).$$

Für beide vorkonditionierten Verfahren kann man das folgende Resultat zeigen (vgl. [Saa96]).

Bemerkung 5.1.2
Die approximierte Lösung des GMRES-Verfahrens mit Links- oder Rechts-Vorkonditionierung ist von der Form

$$x_m = x_0 + s_{m-1}(P^{-1}A)z_0 = x_0 + P^{-1}s_{m-1}(AP^{-1})r_0,$$

wobei $z_0 = P^{-1}r_0$ *und* s_{m-1} *ein Polynom vom Grad* $m-1$ *ist. Das Polynom* s_{m-1} *minimiert* $\|b - Ax_m\|_2$ *für den Fall der Rechts-Vorkonditionierung und* $\left\|P^{-1}(b - Ax_m)\right\|_2$ *im Fall der Links-Vorkonditionierung.*

Ein GMRES-Algorithmus mit Rechts-Vorkonditionierung und Restart von m sieht folgendermaßen aus (vgl. [Saa96]).

Algoritmus 5.1 GMRES mit Rechts-Vorkonditionierung

$r_0 := b - Ax_0$
$\beta := \|r_0\|_2$
$v_1 := \frac{r_0}{\beta}$
for $j = 1, \ldots, m$ **do**
  $w := AP^{-1}v_j$
  for $i = 1, \ldots, j$ **do**
    $h_{i,j} := (w, v_i)$
    $w := w - h_{i,j}v_i$
  end for
  $h_{j+1,j} := \|w\|_2$
  $v_{j+1} := \frac{w}{h_{j+1,j}}$
  Definiere $V_m := [v_1, \ldots, v_m]$ und $\overline{H}_m := (h_{i,j})_{1 \le i \le j+1; 1 \le j \le m}$.
end for
$y_m :=$argmin$_y \left\|\beta e_1 - \overline{H}_m y\right\|_2$
$x_m := x_0 + P^{-1}V_m y_m$
if Abbruchbedingung erfüllt **then**
  STOPP
else
  $x_0 := x_m$
  Beginne von vorne.

5.1.2 Ein Grad-Div-Vorkonditionierer

Eine spezielle Klasse von linearen, algebraischen Problemen sind Sattelpunktprobleme. Sie haben die folgende Form (vgl. Abschnitt 4.3).

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} x = \mathbb{F}.$$

Für diese Art von Problemen wurde der verwendete Grad-Div-Vorkonditionierer (GD) entwickelt. Er gehört zu den Block-Dreiecks-Vorkonditionierern für Sattelpunktprobleme, die die Form

$$P = \begin{pmatrix} \tilde{A} & B^T \\ 0 & \tilde{S} \end{pmatrix}$$

haben, wobei $\tilde{A}$ eine Approximation an A und $\tilde{S}$ eine Approximation an das Schurkomplement $S = -BA^{-1}B^T$ sind. Die Inverse kann man folgendermaßen berechnen

$$\begin{aligned} P^{-1} &= \begin{pmatrix} \tilde{A}^{-1} & -\tilde{A}^{-1}B^T\tilde{S}^{-1} \\ 0 & \tilde{S}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \tilde{A}^{-1} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & B^T \\ 0 & -I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & -\tilde{S}^{-1} \end{pmatrix}. \end{aligned}$$

Beim GD wird das ursprüngliche Problem Grad-Div-stabilisiert (vgl. Abschnitt 4.4). Man berechnet die Matrix $\tilde{A}^{-1}$ exakt, d.h. man verwendet direkte Verfahren wie "umfpack" oder iterative Verfahren wie GMRES, um $\tilde{A}$ als Inverses von A zu berechnen, und setzt

$$\tilde{S}^{-1} := -(\nu + \gamma)M_p^{-1},$$

wobei M_p die Massematrix, ν die Viskosität und γ der Parameter aus der Grad-Div-Stabilisierung sind. M_p^{-1} kann auf Grund der Struktur von M_p leicht berechnet werden. Üblicherweise wird $\gamma \in O(1)$ gewählt. Der Nachteil dieses Vorkonditionierers ist die Veränderung der diskreten Lösung. Jedoch verringert die Grad-Div-Stabilisierung auch die Oszillationen der Lösung, was erwünscht ist. Nähere Ausführungen finden sich in der Diplomarbeit von Timo J. Heister [Hei08].

5.2 Testbeispiele

Nach den theoretischen Grundlagen geht es nun darum, die beschriebenen Verfahren auf die Navier-Stokes-Gleichungen anzuwenden. Dabei werden ein Testbeispiel mit bekannter Lösung und das Lid-Driven-Cavity-Problem behandelt.

5.2.1 Allgemeines

Für die Implementierung wurde die auf C^{++} basierende Programm-Bibliothek `deal.II` verwendet. `deal.II` unterstützt die numerische Lösung von partiellen Differentialgleichungen mittels adaptiver Finiter Elemente. Für die Testbeispiele wurde das Taylor-Hood-Element Q_2-Q_1 benutzt. Dieses Element erfüllt die diskrete Inf-Sup-Bedingung.

Beim Newton-Verfahren wurde die Grad-Div-Stabilisierung für den Vorkonditionierer eingebaut.
Als Abbruchbedingung wurde bei der Picard-Iteration sowie beim Newton-Verfahren 10^{-12} für die Norm des Residuums gewählt. Als obere Grenze für die Iterationszahlen wurden 100 beim Newton-Verfahren und 500 bei der Picard-Iteration festgesetzt.
Der Startwert war in allen Fällen der Nullvektor.

5.2.2 Testbeispiel 1: Beispiel mit vorgegebener Lösung

Bei diesem zweidimensionalen Beispiel ist die Lösung

$$\begin{aligned} u_1(x_1, x_2) &= \sin(\pi x_1), \\ u_2(x_1, x_2) &= -\pi y \cos(\pi x_1), \\ p(x_1, x_2) &= \sin(\pi x_1)\cos(\pi x_2) \end{aligned}$$

auf $\Omega = (0,1)^2$ vorgegeben. Aus ihr kann man durch Einsetzen in die Navier-Stokes-Gleichungen die rechte Seite f in Abhängigkeit von der Viskosität ν und die Randwerte berechnen.
Es wird auf unregelmäßigen Gittern mit 960 bzw. 3840 Zellen gerechnet (vgl. Abbildung 5.1).

5.2.3 Testbeispiel 2: Lid-Driven-Cavity-Problem

Das 2D-Lid-Driven-Cavity-Problem beschreibt eine Strömung im Einheitsquadrat, die durch eine am oberen Rand angelegte Geschwindigkeit u^0 erzeugt wird. Die Randbedingungen lauten

$$\begin{aligned} u_1 &= 0 \quad \text{für} \quad x_1 = 0,\ x_1 = 1,\ x_2 = 0 \text{ und} \\ u_1 &= u^0 = 1 \quad \text{sonst und} \\ u_2 &= 0. \end{aligned}$$

Durch diese Normierung der Geschwindigkeit auf 1, erhält man die Beziehung $Re = \frac{1}{\nu}$ zwischen Reynoldszahl Re und Viskosität ν. Je größer die Reynoldszahl wird, desto schwieriger ist das Problem. Einfacher ist das instationäre Problem (vgl. Abschnitt 4.1) zu lösen, da der durch die Zeitdiskretisierung auftretende Reaktionsterm das Gewicht zugunsten der linearen Terme verschiebt. Daher wird in diesem Beispiel das Problem mit einer konstanten Reaktion c untersucht.
Es wird auf einem äquidistanten Gitter mit 64x64 bzw. 128x128 Zellen gerechnet.

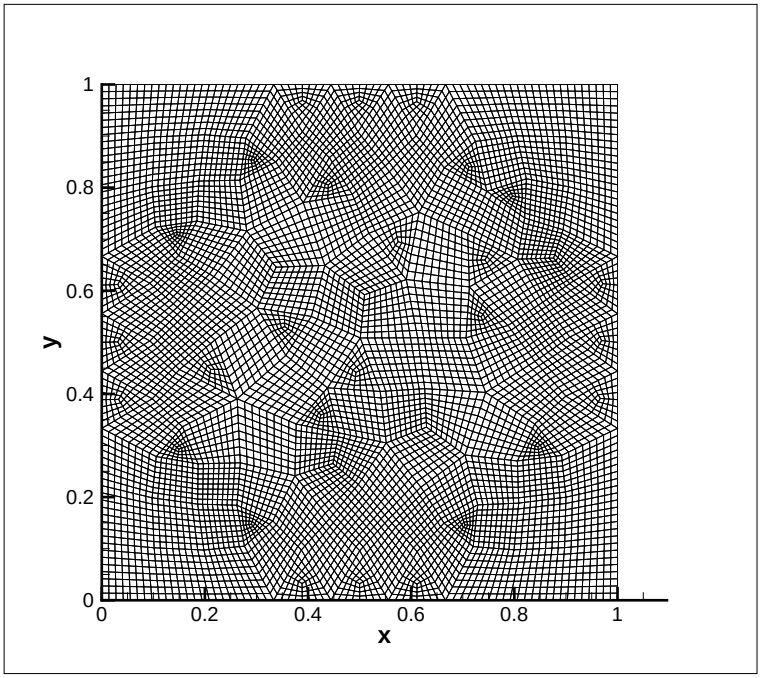

Abbildung 5.1: Grobes Gitter für das erste Testbeispiel

5.3 Ergebnisse

Bei Testbeispiel 1 wurde die Globalisierung des Newton-Verfahrens nur sehr selten benötigt. Daher eignet sich dieses Beispiel sehr gut dazu, dass exakte und inexakte Newton-Verfahren mit der Picard-Iteration zu vergleichen. Testbeispiel 2 war um einiges komplexer und schwieriger zu lösen. Für den Fall ohne künstliche Reaktion benötigt man für große Reynoldszahlen, also kleine Viskositäten, das Backtracking-Verfahren. Die Ergbenissse verbessern sich deutlich bei Einführung eines künstlichen Reaktionsterms, der beim instationären Navier-Stokes-Problem bei Verwendung einer impliziten Zeitdiskretisierung in jedem Zeitschritt entsteht.

5.3.1 Testbeispiel 1

Als Erstes wurde die optimale Dämpfung der Picard-Iteration untersucht. Die Picard-Iteration wurde wie in Abschnitt 4.4 verwendet. Die Iterationszahlen wurden ermittelt und in Abhängigkeit der Viskosität ν und der Dämpfung ω tabellarisiert (vgl. Tabelle 5.1). Dieser Test wurde für das grobe sowie für das feine Gitter durchgeführt. Ein Strich in der Tabelle bedeutet, dass das Verfahren nicht konvergiert. Je mehr man dämpft, desto mehr Iterationsschritte benötigt man. Des Weiteren

ν	ω					
	1	0.9	0.8	0.6	0.4	0.2
10^0	5	13	18	30	51	112
10^{-2}	13	25	30	44	70	146
10^{-4}	25	32	38	54	86	176
10^{-6}	-	-	-	-	-	-

ν	ω					
	1	0.9	0.8	0.6	0.4	0.2
10^0	4	13	18	31	50	109
10^{-2}	10	25	29	42	68	142
10^{-4}	21	31	37	53	83	170
10^{-6}	-	-	-	-	-	-

Tabelle 5.1: Picard-Iterationen in Abhängigkeit des Dämpfungsparameters ω für 960(links) und 3840(rechts) Zellen.

verhindert die Dämpfung das Versagen bei niedrigen Viskositäten nicht, das heißt, man sollte nicht dämpfen. Die Picard-Iteration versagt für $\nu = 10^{-6}$ und kleinere Viskositäten.

Vergleicht man die beiden Tabellen, so sieht man, dass die Feinheit des Gitters keinen großen Einfluß auf die Iterationszahlen hat. Für ein feineres Gitter benötigt man etwas weniger Schritte. Die einzige Ausnahme stellen die Iterationszahlen für $\nu = 1$ und $\omega = 0.6$ dar. Dort vergrößert sich die Iterationszahl um eins bei Verfeinerung des Gitters.

Als Zweites wurde die Effizienz des exakten Newton-Verfahrens untersucht. Auch hier wurden die Iterationszahlen in Abhängigkeit der Viskosität betrachtet (vgl. Tabelle 5.2). Als Löser wurde "umfpack" anstatt GMRES verwendet. Der Newton-Algorithmus versagt auch bei Viskositäten von 10^{-10} im

ν	960 Zellen	3840 Zellen
10^0	2	2
10^{-2}	3	2
10^{-4}	3	3
10^{-6}	4	4
10^{-8}	4	4
10^{-10}	5	5

Tabelle 5.2: Exakte Newton-Iterationen.

Gegensatz zur Picard-Iteration nicht. Des Weiteren sind die Iterationszahlen nahezu invariant bezüglich der Verfeinerung des Gitters. Einzig bei $\nu = 10^{-2}$ unterscheiden sich die Iterationszahlen.
Die Iterationszahlen der Picard-Iteration steigen stark mit sinkendem ν an. Dagegen erhöhen sich die Iterationen beim Newton-Verfahren nur unwesentlich.

Vergleicht man nun die Picard-Iteration mit dem exakten Newton-Verfahren (vgl. Tabelle 5.3), so sieht man, dass die Newton-Iteration der Picard-Iteration überlegen ist. Dieser Effekt verstärkt sich mit sinkender Viskosität ν. Am aufwendigsten war die Wahl der optimalen Parameter für das in-

ν	Picard	Newton
10^0	5	2
10^{-2}	13	3
10^{-4}	25	3
10^{-6}	-	4

ν	Picard	Newton
10^0	4	2
10^{-2}	10	2
10^{-4}	21	3
10^{-6}	-	4

Tabelle 5.3: Vergleich von exakter Newton- und Picard-Iterationen mit $\omega = 1$ für 960(l) und 3840(r) Zellen.

exakte Newton-Verfahren. Das Backtracking-Verfahren besitzt fünf Parameter, die man variieren kann: η_0, η_{max}, θ_{min}, θ_{max} und t.

Zunächst wurden $\theta_{min} = 0.1$, $\theta_{max} = 0.5$ und $t = 10^{-4}$ konstant gehalten und die Newton- bzw. GMRES-Schritte für unterschiedliche η_0 und η_{max} untersucht. Für η_0 wurden die Werte 10^{-1} , 10^{-2}, 10^{-4} und 10^{-6} und für η_{max} die Werte 0.75, 0.5, 0.25, 10^{-1}, 10^{-2} und 10^{-3} untersucht. Die Ergebnisse für $\nu = 1$, $\nu = 10^{-2}$ bis $\nu = 10^{-10}$ finden sich im Anhang B.
Es stellt sich heraus, dass man η_0 und η_{max} möglichst klein wählen muss, d.h. man muss das lineare System sehr genau lösen.

Während der 170 erfolgreichen Durchläufe des Verfahrens, wurde das Backtracking-Verfahren nur 13 mal verwendet. Dies war der Fall für $\eta_0 = 10^{-1}$ und $\eta_0 = 10^{-2}$ bei $\nu = 10^{-2}$. Die maximale Anzahl der Backtracking-Schritte pro Durchlauf betrug 3, während pro Iterationsschritt höchstens einmal die Schrittweite reduziert wurde. Daher spielen die Parameter θ_{min} und θ_{max}, die nur für die Reduktion des Schrittes verwendet werden, keine Rolle bei der Effizienz des Verfahrens.

Für die optimalen Werte $\eta_0 = 10^{-6}$ und $\eta_{max} = 10^{-3}$ wurde der Einfluß von t untersucht. Dabei wurden die Werte 10^{-1}, 10^{-2}, 10^{-3} und 10^{-4} eingesetzt. Für diese verschiedenen t gab es weder

Unterschiede in der Anzahl der Newton-Schritte, noch in der der GMRES-Schritte.

Als Letztes sollte noch der Einfluß des Vorkonditionierers für die optimalen Parameter betrachtet werden. Doch GMRES bricht schon für $\nu = 1$ nach 10.000 Iterationsschritten ab. Da GMRES das System zu genau lösen muss, würde eine andere Wahl von η_0 und η_{max} eventuell ein vorzeitiges Abbrechen verhindern. Dies wurde hier aber nicht weiter untersucht.

5.3.2 Testbeispiel 2

Beim Lid-Driven-Cavity-Problem wurde ebenfalls die optimale Dämpfung der Picard-Iteration untersucht. Die Picard-Iteration wurde wie in Abschnitt 4.4 verwendet. Die Iterationszahlen wurde ermittelt und in Abhängigkeit der Viskosität ν und der Dämpfung ω tabellarisiert (vgl. Tabelle 5.4). Dieser Test wurde für die Gitterweiten $h = \frac{1}{64}$ und $h = \frac{1}{128}$ durchgeführt. Ein Strich in der Tabelle bedeutet, dass das Verfahren nicht konvergiert.

ν	ω					
	1	0.9	0.8	0.6	0.4	0.2
10^0	6	12	16	25	45	96
10^{-1}	9	13	17	27	45	97
10^{-2}	18	21	25	37	59	122
$5 \cdot 10^{-3}$	25	28	33	46	72	146
10^{-3}	41	46	53	72	110	217
$5 \cdot 10^{-4}$	67	49	56	76	115	227
10^{-4}	-	-	-	-	-	-

ν	ω					
	1	0.9	0.8	0.6	0.4	0.2
10^0	6	12	17	26	44	94
10^{-1}	9	12	16	26	44	94
10^{-2}	17	20	24	36	57	118
$5 \cdot 10^{-3}$	24	28	32	45	70	141
10^{-3}	40	45	51	70	106	209
$5 \cdot 10^{-4}$	51	47	54	73	111	218
10^{-4}	-	-	176	148	218	423

Tabelle 5.4: Picard-Iterationen in Abhängigkeit des Dämpfungsparameters ω für $h = \frac{1}{64}$(links) und $h = \frac{1}{128}$(rechts).

Für das grobe Gitter hilft die Dämpfung nicht. Sie erhöht nur die Iterationszahlen. Dagegen kann man beim feinen Gitter mit einer Dämpfung $\omega \leq 0.8$ auch das Problem für die Viskosität $\nu = 10^{-4}$ lösen.

Bei diesem Beispiel kann das exakte Newton-Verfahren die Ergebnisse der Picard-Iteration nicht erreichen, denn es versagt schon für die Viskosität $\nu = 5 \cdot 10^{-4}$. Jedoch kann man mit einem künstlichen Reaktionsterm die Ergebnisse verbessern. Ein derartiger Term entsteht bei impliziter Zeitdiskretisierung des instationären Navier-Stokes-Problems in jedem Zeitschritt (vgl. Abschnitt 4.1). Praktisch sind eventuell hinreichend viele Zeitschritte erforderlich. Durch die Lösung vom vorherigen Zeitschritt hat man aber auch in der Regel einen sehr guten Startwert für das Newton-Verfahren. Die Ergebnisse für zusätzliche Reaktionen von $c = 0.1$ und $c = 1$ werden in Tabelle 5.5 mit den Ergebnissen des stationären Verfahrens verglichen.

Je größer man c wählt, desto robuster wird das Verfahren, bzw. desto einfacher ist die Gleichung zu lösen. Bei $c = 1$ werden keine Backtracking-Schritte benötigt und die Iterationszahlen sind niedriger als die der anderen beiden Varianten. Für $c = 0.1$ kann man das System bis zu einer Viskosität von $\nu = 5 \cdot 10^{-4}$ berechnen. Im Gegensatz zum Testbeispiel 1 werden auch im exakten Verfahren Backtracking-Schritte benötigt. Dabei ist die maximale Anzahl von hintereinander ausgeführten Schritte gleich drei, was auch nur einmal bei $c = 0.1$ und $\nu = 10^{-3}$ auf dem feinen Gitter auftritt. Ansonsten gibt es keinen gravierenden Unterschied zwischen dem groben und dem feinen Gitter.

Nun vergleicht man das exakte Newton-Verfahren ohne Reaktion mit der Picard-Iteration (vgl. Tabelle 5.6).

Das Newton-Verfahren benötigt höchstens halb so viele Schritte wie die Picard-Iteration, um die Lösung zu berechnen. Die sukzessive Iteration versagt dafür bei niedrigen Viskositäten nicht. Auf-

	Reaktion					
	0		0.1		1	
ν	64 x 64	128 x 128	64 x 64	128 x 128	64 x 64	128 x 128
10^0	3	3	3	3	3	3
10^{-1}	4	4	4	4	4	-
10^{-2}	5	5	5	6	4	4
$5 \cdot 10^{-3}$	6	6	6	6	5	5
10^{-3}	15(8)	12(4)	13(4)	24(22)	6	6
$5 \cdot 10^{-4}$	-	-	10(2)	12(1)	7	7
10^{-4}	-	-	-	-	7	10

Tabelle 5.5: Exakte Newton-Iterationen in Abhängigkeit der zusätzlichen Reaktion und des Gitters. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

ν	Picard	Newton
10^0	6	3
10^{-1}	9	4
10^{-2}	18	5
$5 \cdot 10^{-3}$	25	6
10^{-3}	41	15
$5 \cdot 10^{-4}$	67	-

ν	Picard	Newton
10^0	17	3
10^{-1}	16	4
10^{-2}	24	5
$5 \cdot 10^{-3}$	32	6
10^{-3}	51	12
$5 \cdot 10^{-4}$	54	-

Tabelle 5.6: Exakte Newton- und Picard-Iterationen in Abhängigkeit der Viskosität ν für $h = \frac{1}{64}$ mit $\omega = 1$(l) und $h = \frac{1}{128}$ mit $\omega = 0.8$(r).

fällig ist noch, dass sich die Iterationszahlen des exakten Newton-Verfahrens zwischen 10^0, 10^{-1} und 10^{-2} jeweils nur um eins unterscheiden, sich jedoch zwischen 10^{-2} und 10^{-3} verdreifachen bzw. verdoppeln. Der Einfluß der Gitterweite ist vernachlässigbar.

Wie in Testbeispiel 1 wurde die Abhängigkeit der Effektivität des Verfahrens von den Parametern η_0, η_{max}, θ_{min}, θ_{max} und t untersucht.

Zunächst wurden die Parameter $\theta_{min} = 0.1$, $\theta_{max} = 0.5$ und $t = 10^{-4}$ konstant gewählt. Die Ergebnisse dieser Tests mit Reaktion $c = 1$ und ohne zusätzlicher Reaktion werden im Anhang B.2 dargestellt.

Betrachtet man den Fall der zusätzlichen Reaktion, so ist auffällig, dass das Verfahren nie versagt und auch nie Backtracking-Schritte benötigt. Die pro Rechnung benötigte Anzahl an GMRES-Schritten liegt zwischen 15 und 35. Da sich die Newton- sowie die GMRES-Iterationszahlen für $\eta_0 = 0.1$ und $\eta_0 = 0.01$ nicht unterscheiden, sind die optimalen Parameter $\eta_0 = 0.1$ und $\eta_{max} = 0.01$. In diesem Testbeispiel ist es nicht von Vorteil möglichst genau zu lösen, was ja die Angabe von η_{max} beschreibt. Auch die ersten Schritte müssen nicht sonderlich gut sein. Die Anzahl der GMRES-Schritte ist sogar teilweise größer, wenn man am Anfang genauer löst.

Der stationäre Fall ist derjenige ohne Reaktion. Für große Viskositäten ν unterscheidet sich dieser Fall nur unwesentlich von dem Fall mit Reaktion. Erst bei $\nu = 10^{-3}$ treten größere Unterschiede auf. Zum Einen werden in dem Fall ohne Reaktion Backtracking-Schritte benötigt und zum Andern erhöht sich der Aufwand. Die Anzahl der GMRES-Schritte liegt zwischen 37 und 69, das heißt sie verdoppelt sich. Extrem wird dies für die Viskosität $\nu = 5 \cdot 10^{-4}$. Die Verfahren, die nicht versagen, benötigen zwischen 91 und 159 GMRES-Schritte. Für das feine Gitter gab es keine getestete Parameterkombination, die eine Lösung berechnen konnte.

Im Folgenden wird sich auf den Fall $c = 0$ beschränkt. Den geringsten Aufwand über alle getesteten Fälle benötigten $\eta_0 = 10^{-4}$ und $\eta_{max} = 10^{-2}$. Mit diesen Parametern untersucht man den Einfluß von θ_{min} und θ_{max} auf die Effektivität und Robustheit. $t = 10^{-4}$ wird auch hier festgehalten. θ_{min} und θ_{max} beeinflussen lediglich die Reduktion des berechneten inexakten Newton-Schrittes. Das heißt die Rechnungen für $\nu \in \{1, 10^{-1}, 10^{-2}\}$ werden dadurch nicht beeinflusst. Daher wird sich auf den Fall $\nu = 10^{-3}$ beschränkt.

In den folgenden Tabellen 5.7 und 5.8 sind die Ergebnisse zusammengefasst. Die leeren Zellen bedeuten, dass die Rechnung nicht durchgeführt wurde. Die Striche bedeuten einen Abbruch des Verfahrens.

	θ_{max}				
θ_{min}	0.1	0.25	0.5	0.75	0.99
0	-	-	20(12)	20(12)	20(12)
0.1	-	21(15)	17(9)	17(9)	17(9)
0.25		-	15(8)	15(8)	15(8)
0.5			-	-	-

	θ_{max}				
θ_{min}	0.1	0.25	0.5	0.75	0.99
0	-	-	63	63	63
0.1	-	71	54	54	54
0.25		-	48	48	48
0.5			-	-	-

Tabelle 5.7: Newton- und GMRES-Iterationen in Abhängigkeit von θ_{min} und θ_{max} für $h = \frac{1}{64}$, $\eta_0 = 10^{-4}$ und $\eta_{max} = 10^{-2}$. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

	θ_{max}				
θ_{min}	0.1	0.25	0.5	0.75	0.99
0	-	-	12(4)	12(4)	12(4)
0.1	-	24(16)	12(4)	12(4)	12(4)
0.25		-	12(4)	12(4)	12(4)
0.5			-	-	-

	θ_{max}				
θ_{min}	0.1	0.25	0.5	0.75	0.99
0	-	-	40	40	40
0.1	-	87	40	40	40
0.25		-	41	41	41
0.5			-	-	-

Tabelle 5.8: Newton- und GMRES-Iterationen in Abhängigkeit von θ_{min} und θ_{max} für $h = \frac{1}{128}$, $\eta_0 = 10^{-4}$ und $\eta_{max} = 10^{-2}$. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

Man sieht, dass die Iterationszahlen sowohl für das Newton-Verfahren als auch für GMRES im Falle der Konvergenz nur von θ_{min} abhängen. Einzig die Wahl $\theta_{min} = 0.1$ und $\theta_{max} = 0.25$ sticht heraus. Der Aufwand für diese Parameter ist für beide Gitterweiten am größten von allen Parameterwahlen, die erfolgreich waren. Beim feinen Gitter ist die Abhängigkeit der Iterationszahlen geringer als beim groben Gitter. Es ist wichtig, dass θ_{max} größer als θ_{min} ist, denn für konstante θ (vgl. Diagonale über den leeren Kästchen) versagt das Verfahren auf dem groben und auf dem feinen Gitter.

Nachdem die ersten vier Parameter optimiert wurden, wurde noch die Abhängigkeit der Iterationszahlen von t untersucht. Dabei wurden die Werte 10^{-1}, 10^{-2}, 10^{-4} und 10^{-6} für t gewählt. Es zeigte sich jedoch keine Abhängigkeit. Sowohl die Newton- und die Backtracking-Schritte als auch die GMRES-Iterationszahlen waren identisch. Als optimale Parameter werden somit $\eta_0 = 10^{-4}$, $\eta_{max} = 10^{-2}$, $\theta_{min} = 0.25$, $\theta_{max} = 0.5$ und $t = 10^{-4}$ gewählt (vgl. Abschnitt 5.3.3).

Mit diesen Parameterwerten wurde der Einfluß der Reaktion c untersucht. Die Ergebnisse sind in Tabelle 5.9 dargestellt. Das Verfahren ist für Reaktionen $c = 0.001$ in der Lage das Problem für die Viskosität $\nu = 5 \cdot 10^{-4}$ zu lösen. Nur wenn $c \geq 0.4$ ist, kann man auch die Gleichungen für $\nu = 10^{-4}$ berechnen.

Als Letztes wurde der Einfluß des Vorkonditionierers untersucht. Mit den Parameterwerten $\eta_0 = 10^{-4}$, $\eta_{max} = 10^{-2}$, $\theta_{min} = 0.25$, $\theta_{max} = 0.5$ und $t = 10^{-4}$ erhält man folgende Tabelle (vgl. Tabelle 5.10).

	Reaktion									
ν	1	0.8	0.6	0.4	0.2	0.1	0.01	0.001	0.0001	0
$5 \cdot 10^{-4}$	8	8	9	9(1)	18(11)	10(2)	-	19(19)	-	-
10^{-4}	7	8	9	13(3)	-	-	-	-	-	-

	Reaktion									
ν	1	0.8	0.6	0.4	0.2	0.1	0.01	0.001	0.0001	0
$5 \cdot 10^{-4}$	20	19	21	21	46	24	-	62	-	-
10^{-4}	17	18	20	28	-	-	-	-	-	-

Tabelle 5.9: Newton- und GMRES-Iterationen in Abhängigkeit der Reaktion c und der Viskosität ν für $h = \frac{1}{64}$. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

	Vorkonditionierung	
ν	mit	ohne
10^{0}	6	7
10^{-1}	5	6
10^{-2}	6	7
$5 \cdot 10^{-3}$	7	8
10^{-3}	15(8)	-
$5 \cdot 10^{-4}$	-	-

	Vorkonditionierung	
ν	mit	ohne
10^{0}	29	43684
10^{-1}	17	10795
10^{-2}	18	12418
$5 \cdot 10^{-3}$	22	19118
10^{-3}	48	-
$5 \cdot 10^{-4}$	-	-

Tabelle 5.10: Newton- und GMRES-Iterationen mit und ohne Vorkonditionierer für $h = \frac{1}{64}$. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

Das Verfahren ohne Vorkonditionierung benötigt eine Newton-Iteration weniger als das vorkonditionierte System. Das nicht-vorkonditionierte System versagt bei $\nu = 10^{-3}$, weil der Löser mehr als 10.000 GMRES-Iterationen braucht, um das System zu lösen. Ohne Vorkonditionierung benötigt man die 635-1507-fache Anzahl an GMRES-Schritten. Allerdings sollte man dabei beachten, dass die Approximation an die Matrix A (vgl. Abschnitt 5.1) beim Berechnen der Vorkonditionierungsmatrix mit einem direkten oder iterativen Verfahren berechnet werden muss.

5.3.3 Mögliche Verbesserungen

Da sehr viele Parameter zu wählen sind, muss viel gerechnet werden. Die hier gewählten Parameter sind höchstwahrscheinlich nicht ganz optimal, denn es wurden immer drei der Parameter festgehalten, während die übrigen zwei variiert wurden.

Die in Abschnitt 2.2.2 beschriebene Sicherheit für die Forcing-Terme ist in den hier betrachten Fällen irrelevant, denn die maximale Größe der Forcing-Terme η_{max} ist in den meisten Fällen kleiner als die obere Schranke 0.1 gewählt worden. Eine Verbesserung würde eventuell eine Anpassung der Sicherheit an η_{max} bringen. Man könnte die Sicherheit z.B. als $\frac{\eta_{max}}{10}$ wählen.

Kapitel 6

Fazit und Ausblick

Ziel dieser Arbeit war es das Newton-Verfahren vorzustellen. Es wurde untersucht, ob es als Alternative zur Picard-Iteration bei der Berechnung der inkompressiblen Navier-Stokes-Gleichungen von Bedeutung ist.

Die Picard-Iteration wurde vorgestellt. Es wurde bewiesen, dass die relaxierte Variante unter den Bedingungen des Fixpunkt-Satzes von Banach ebenfalls konvergiert.

Es wurden Konvergenzaussagen für das exakte und das inexakte Newton-Verfahren gezeigt. Das exakte Verfahren konvergiert quadratisch, während beim inexakten Verfahren lediglich superlineare Konvergenz bewiesen werden kann, falls die Forcing-Terme adaptiv gewählt wurden. Für konstante Forcing-Terme erhält man lediglich lineare Konvergenz. Alle diese Aussagen sind aber auf einen kleinen Konvergenzbereich eingeschränkt, der im Allgemeinen kleiner als der der Picard-Iteration ist.

Daher wurde das inexakte Newton-Verfahren globalisiert. Es zeigt sich, dass es verschiedene Ansätze gibt, einen geeigneten Schritt zu wählen. Es wurde gezeigt, dass das inexakte globalisierte Verfahren konvergiert, falls in jedem Schritt eine Berechnung des Newton-Schrittes möglich ist und die Forcing-Terme geeignet gewählt wurden. Über die Konvergenzordnung wurden keine Aussagen gemacht. Man kann beweisen, dass das Verfahren nach endlich vielen Schritten keine Globalisierung mehr benötigt. Das heißt, dass für die letzten Schritte die Konvergenzordnung des nichtglobalisierten Verfahrens ausschlaggebend ist.

In Kapitel 4 wurden zwei globalisierte inexakte Newton-Verfahren sowie die relaxierte Picard-Iteration auf die diskretisierte Navier-Stokes-Gleichungen angewendet und der für die numerischen Tests verwendeten INB-Algorithmus dokumentiert.

Die in der Einleitung beschriebenen Vor- bzw. Nachteile des Newton-Verfahrens zeigen sich auch in den hier gemachten Tests. Falls das Newton-Verfahren konvergiert, ist es sehr viel schneller als die Picard-Iteration. Allerdings reagiert es sensibel auf die Wahl der Parameter. Desweiteren ist das Backtracking-Verfahren von der Wahl und der Güte des Vorkonditionierers abhängig.
Nach den vorliegenden Ergebnissen eignet sich das inexakte Newton-Verfahren mit Backtracking nicht als Alternative für die Picard-Iteration beim stationären Navier-Stokes-Problem. Eventuell ist das Newton-Verfahren bei instationären Problemen besser geeignet, da dort zusätzliche Reaktionsterme vorhanden sind.

Die numerischen Ergebnisse zeigen, dass es notwendig ist, das Newton-Verfahren zu globalisieren. Die Tests wurden ausschließlich mit dem Backtracking-Verfahren als inexaktes Newton-Verfahren gemacht. Es ist zu vermuten, dass das Dogleg-Verfahren bessere Ergebnisse liefert. Jedoch sind in diesem Verfahren noch mehr Parameter zu variieren, als beim Backtracking-Verfahren.

Die numerischen Tests konnten die in der Literatur beschriebenen Ergebnisse für große Reynoldszahlen nicht bestätigen. In [STW97, PSSW06] wird das inexakte Newton-Verfahren mit Backtracking auf das Lid-Driven-Cavity-Problem mit Reynoldszahl $Re = 10.000$ angewendet. In den hier durchgeführten Test konnten lediglich Reynoldszahlen ≤ 2000 berechnet werden.

Eine Kombination von Picard- und Newton-Iteration könnte die Vorteile beider Verfahren vereinen. Mit der sukzessiven Iteration nähert man sich linear der Lösung, bis man in den Konvergenzradius des Newton-Verfahrens gelangt. Dank der quadratischen Konvergenz erreicht man von dort schnell die Lösung.

Anhang A

Q- und R-Konvergenzordnung

In Abschnitt 2.2.2 wurde die Wahl der Forcing-Terme beim inexakten Newton-Verfahren beschrieben. Dabei wird eine Sicherheit eingeführt, damit die Terme nicht zu schnell zu klein werden. Die Sicherheit wird angewendet, falls die Potenz $\eta_{k-1}^{\frac{1+\sqrt{5}}{2}}$ unterhalb einer konstanten Schranke liegt, wobei η_{k-1} der Forcing-Term aus dem vorherigen Schritt ist. Um diesen willkürlich erscheinenen Exponenten zu erklären, benötigt man die Ausführungen dieses Anhangs. Es wird sich zeigen, dass die R-Konvergenzordnung des exakten Newton-Verfahrens gerade $\frac{1+\sqrt{5}}{2}$ beträgt.

A.1 Grundlagen

Die Ausführungen dieses Anhangs beruhen auf [OR70]. Das Neumann-Lemma (vgl. Bemerkung 2.2.2) garantiert die Existenz des Inversen von $I - B$, falls B ein linearer stetiger Operator mit $\|B\| < 1$ ist. Lemma 2.2.3 liefert eine Abschätzung für die Norm von B^{-1}. Zunächst werden einige Eigenschaften von Frechet-differenzierbaren Funktionen vorgestellt.

Bemerkung A.1.1
Falls $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ Frechet-differenzierbar in $x \in D$ ist, so ist F stetig an x auf Geraden durch x.

Bemerkung A.1.2
Sei $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ Frechet-differenzierbar auf einem konvexen Gebiet $D_0 \subset D$. Dann gilt für alle x, $y \in D_0$

$$\|Fx - Fy\| \leq \sup_{0 \leq t \leq 1} \|F'(x + t(y - x))\| \, \|x - y\| .$$

Bemerkung A.1.3
Sei $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ Frechet-differenzierbar auf einem konvexen Gebiet $D_0 \subset D$. Dann gilt für alle x, y, $z \in D_0$

$$\|Fy - Fz - F'(x)(y - z)\| \leq \sup_{0 \leq t \leq 1} \|F'(x + t(y - z)) - F'(x)\| \, \|x - y\| .$$

Definition A.1.4
Sei $I(\tau, x^*)$ die Menge aller Folgen mit Limes-Punkt x^*, die durch den iterativen Prozeß τ erzeugt werden. Sei $\{x_k\} \in \mathbb{R}^n$ eine Folge, die gegen x^* konvergiert, dann heißt

$$Q_p(\{x_k\}) := \begin{cases} 0 & \text{, falls } x_k = x^* \text{ für fast alle } k \\ \limsup \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} & \text{, falls } x_k \neq x^* \text{ für fast alle } k \\ \infty & \text{, sonst} \end{cases}$$

Q_p-Faktor *von* $\{x_k\}$.

$$Q_p(\tau, x^*) := \sup_{\{x_k\} \in I(\tau, x^*)} Q_p(\{x_k\})$$

heißt Q_p-Faktor *von* τ *bei* x^*.

Lemma A.1.5
Sei $Q_p(\tau, x^*)$, $p \in [1, \infty)$, *wie in Definition A.1.4 gegeben. Dann gilt genau eine der folgenden Aussagen.*

(a) $Q_p(\tau, x^*) = 0$ *für alle* $p \in [1, \infty)$,

(b) $Q_p(\tau, x^*) = \infty$ *für alle* $p \in [1, \infty)$ *oder*

(c) es gibt $p_0 \in [1, \infty)$, *so dass* $Q_p(\tau, x^*) = 0$ *für alle* $p \in [1, p_0)$ *und* $Q_p(\tau, x^*) = \infty$ *für alle* $p \in (p_0, \infty)$.

Beweis:
Sei $\{x_k\} \in I(\tau, x^*)$ und $\epsilon_k := \|x_k - x^*\|$, $k = 0, 1, \ldots$. Falls $\epsilon_k = 0$ für fast alle k gilt, so folgt $Q_p(\{x_k\}) = 0$ für alle $p \in [1, \infty)$.
Man setze also $\epsilon_k > 0$ für $k \geq k_0$ voraus und nehme weiter an, dass $Q_p(\{x_k\}) < \infty$ für ein $p \in (1, \infty)$. Dann gilt für jedes $q \in [0, p)$

$$\begin{aligned} Q_q(\{x_k\}) &= \limsup_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^q} = \limsup_{k\to\infty} \frac{\epsilon_{k+1}}{\epsilon_k^q} \\ &\leq \limsup_{k\to\infty} \frac{\epsilon_{k+1}}{\epsilon_k^p} \limsup_{k\to\infty} \epsilon_k^{p-q} \\ &= Q_p(\{x_k\}) \underbrace{\limsup_{k\to\infty} \epsilon_k^{p-q}}_{=0} = 0. \end{aligned}$$

D.h. falls $1 \leq q < p$ und $Q_q(\{x_k\}) > 0$, muss $Q_p(\{x_k\}) = \infty$ gelten. Damit ist die Aussage für jede Folge $\{x_k\} \in I(\tau, x^*)$ gezeigt. Jetzt muss die Aussage natürlich noch für die Q_p-Faktoren gezeigt werden.

Man nehme an, dass weder (a) noch (b) gelten und setze $p_0 := \arg\inf_{p\in[0,\infty)} \{Q_p(\tau, x^*) = \infty\}$.
Um die eine Seite zu zeigen, nehme man an, dass es $p > p_0$ gebe, so dass $Q_p(\tau, x^*) < \infty$. D.h. $Q_p(\{x_k\}) < \infty$ für alle $\{x_k\} \in I(\tau, x^*)$. Nach Definition von p_0 existiert ein $p' \in [p_o, p)$ mit $Q_{p'}(\tau, x^*) = \infty$ und damit gilt $Q_{p'}(\{x_k\}) > 0$ für irgendeine Folge $\{x_k\} \in I(\tau, x^*)$. Nach dem oben Gezeigten folgt damit $Q_p(\{x_k\}) = \infty$. Das ist aber ein Widerspruch und damit gilt $Q_p(\tau, x^*) = \infty$ für alle $p \in (p_o, \infty)$.

Falls $p_0 > 1$ und $Q_p(\tau, x^*) \neq 0$ für ein $p \in [1, p_0)$, so folgt mit dem gleichen Argument
$Q_{p'}(\tau, x^*) = \infty$ für alle $p' \in (p, p_0)$. Das ist aber ein Widerspruch zur Definition von p_0.
Und damit gilt $Q_p(\tau, x^*) = 0$ für alle $p \in [1, p_0)$. □

Definition A.1.6
Man definiert als die Q-Ordnung *von* τ *bei* x^*

$$O_Q(\tau, x^*) := \begin{cases} \infty & , \textit{falls } Q_p(\tau, x^*) = 0 \quad \forall p \in [1, \infty) \\ \textit{arg}\inf_{p\in[1,\infty)} \{Q_p(\tau, x^*) = \infty\} & , \textit{sonst.} \end{cases}$$

Aus dem Gezeigten kann man folgende Eigenschaften ablesen.

Bemerkung A.1.7
Sei τ ein iterativer Prozeß mit Limes-Punkt x^*.

1. *Falls $Q_p(\tau, x^*) < \infty$ für ein $p \in [1, \infty)$, so gilt $O_Q(\tau, x^*) \geq p$.*
2. *Falls $Q_q(\tau, x^*) > 0$ für ein $q \in [1, \infty)$, so gilt $O_Q(\tau, x^*) \leq q$.*
3. *Falls $0 < Q_p(\tau, x^*) < \infty$ für ein $p \in [1, \infty)$, so gilt $O_Q(\tau, x^*) = p$.*

Analog zur Definition A.1.4 erhält man den R_p-Faktor durch die folgende Definition.

Definition A.1.8
Sei $\{x_k\} \in \mathbb{R}^n$ eine Folge, die gegen x^* konvergiert, dann heißt

$$R_p(\{x_k\}) := \begin{cases} \limsup_{k\to\infty} \|x_k - x^*\|^{\frac{1}{k}}, & \text{für } p = 1, \\ \limsup_{k\to\infty} \|x_k - x^*\|^{\frac{1}{p^k}}, & \text{für } p > 1, \end{cases}$$

R_p-Faktor *von* $\{x_k\}$.

$$R_p(\tau, x^*) := \sup_{\{x_k\} \in I(\tau, x^*)} R_p(\{x_k\})$$

heißt R_p-Faktor *von τ bei x^*.*

Lemma A.1.9
Sei $R_p(\tau, x^*)$, $p \in [1, \infty)$, wie in Definition A.1.8 gegeben. Dann gilt genau eine der folgenden Aussagen.

(a) $R_p(\tau, x^*) = 0$ für alle $p \in [1, \infty)$,

(b) $R_p(\tau, x^*) = 1$ für alle $p \in [1, \infty)$ oder

(c) es gibt $p_0 \in [1, \infty)$ so, dass $R_p(\tau, x^*) = 0$ für alle $p \in [1, p_0)$ und $R_p(\tau, x^*) = 1$ für alle $p \in (p_0, \infty)$.

Beweis:
Sei $\{x_k\}$ eine beliebige Folge aus $I(\tau, x^*)$. Setze $\gamma_{1k} := \frac{1}{k}$ und $\gamma_{pk} := \frac{1}{p^k}$ für $p > 1$ und $k = 1, 2, \ldots$.
Es gilt

$$\lim_{k\to\infty} \frac{\gamma_{qk}}{\gamma_{pk}} = \lim_{k\to\infty} \frac{p^k}{q^k} = \lim_{k\to\infty} \left(\frac{p}{q}\right)^k = \infty \text{ für } 1 < q < p.$$

Man nehme an, dass $R_p(\{x_k\}) < 1$ für ein $p \in (1, \infty)$ und man wähle $\epsilon > 0$ so, dass für ein festes $R_p(\{x_k\}) < \alpha < 1$ die Gleichung $R_p(\{x_k\}) + \epsilon = \alpha$ erfüllt ist. Man setze $\epsilon_k := \|x_k - x^*\|$ und wähle k_0 so, dass

$$\epsilon_k^{\gamma_{pk}} \leq \alpha \quad \forall k \geq k_0.$$

Damit gilt für alle $q \in [1, p)$

$$\begin{aligned} R_q(\{x_k\}) &= \limsup_{k\to\infty} \|x_k - x^*\|^{\frac{1}{q^k}} = \limsup_{k\to\infty} \epsilon_k^{\gamma_{qk}} \\ &= \limsup_{k\to\infty} \left(\epsilon_k^{\gamma_{pk}}\right)^{\frac{\gamma_{qk}}{\gamma_{pk}}} \\ &\leq \lim_{k\to\infty} \alpha^{\frac{\gamma_{qk}}{\gamma_{pk}}} = 0. \end{aligned}$$

D.h. $R_q(\{x_k\}) = 0$ falls $q < p$ und $R_p(\{x_k\}) < 1$. Gilt nun aber $R_q(\{x_k\}) > 0$ für $q < p$, so muss die Annahme $R_p(\{x_k\}) < 1$ falsch sein und es gilt $R_p(\{x_k\}) = 1$. Damit hat man die Behauptung für jede Folge $\{x_k\} \in I(\tau, x^*)$. Es fehlt nun noch der Beweis für den R_p-Faktor von τ bei x^*.

Als erstes gehe man davon aus, dass weder (a) noch (b) gelten, und setze
$p_0 := \operatorname{arginf}_{p\in[1,\infty)} \{R_p(\tau, x^*) = 1\}$. Des Weiteren nehme man an, dass $p > p_0$ existiert, so dass $R_p(\tau, x^*) < 1$.

Dann gilt $R_p(\{x_k\}) < 1$ für alle Folgen aus $I(\tau, x^*)$ und nach Definition von p_0 gibt es $p' \in [p_0, p)$ mit $R_{p'}(\tau, x^*) = 1$ und damit existiert eine Folge $\{x_k\}$ mit $R_{p'}(\{x_k\}) > 0$. Mit dem oben Gezeigten folgt $R_p(\{x_k\}) = 1$. Dies ist ein Widerspruch und es gilt $R_p(\tau, x^*) = 1$ für $p \in (p_0, \infty)$.
Analog zu Lemma A.1.5 folgt $R_p(\tau, x^*) = 0$ für $p < p_0$. □

Analog zur Q-Ordnung definiert man die R-Ordnung folgendermaßen.

Definition A.1.10
Man definiert als die R*-Ordnung von τ bei x^**

$$O_R(\tau, x^*) := \begin{cases} \infty & \text{, falls } R_p(\tau, x^*) = 1 \quad \forall p \in [1, \infty) \\ \operatorname{arginf}_{p\in[1,\infty)} \{R_p(\tau, x^*) = 1\} & \text{, sonst.} \end{cases}$$

Auch hier erhält man ähnliche Eigenschaften wie bei der Q-Ordnung.

Bemerkung A.1.11
Sei τ ein iterativer Prozeß mit Limes-Punkt x^*.

1. *Falls $R_p(\tau, x^*) < 1$ für ein $p \in [1, \infty)$, so gilt $O_R(\tau, x^*) \geq p$.*
2. *Falls $R_q(\tau, x^*) > 0$ für ein $q \in [1, \infty)$, so gilt $O_R(\tau, x^*) \leq q$.*
3. *Falls $0 < R_p(\tau, x^*) < \infty$ für ein $p \in [1, \infty)$, so gilt $O_R(\tau, x^*) = p$.*

Jetzt folgt eine kleine analytische Überlegung.

Bemerkung A.1.12
Für jede ganze Zahl $m \geq 1$ besitzt das Polynom

$$p_m(t) = t^{m+1} - t^m - 1$$

eine eindeutige positive Nullstelle $\tau_m \in (1, 2]$.

Lemma A.1.13
Sei $I(\tau, x^*)$ gegeben und $\gamma_0, \ldots, \gamma_m$ nichtnegative Konstanten. Falls für eine Folge $\{x_k\} \in I(\tau, x^*)$ ein $k_0 \geq m$ existiert, so dass

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\| \sum_{j=0}^{m} \gamma_j \|x_{k-j} - x^*\| \quad \forall k \geq k_0,$$

so gilt $O_R(\tau, x^*) \geq t$, wobei t die eindeutige positive Wurzel von $\tilde{t}^{m+1} - \tilde{t}^m - 1 = 0$ ist.
Falls zusätzlich ein $\beta > 0$ und eine Folge $\{x_k\} \in I(\tau, x^*)$ existieren, so dass für $k_0 \geq m$

$$\|x_{k+1} - x^*\| \geq \beta \|x_k - x^*\| \|x_{k-m} - x^*\| > 0 \quad \forall k \geq k_0,$$

so gilt $O_R(\tau, x^*) = t$.

Beweis:
(1) Sei $\gamma := \sum_{j=0}^{m} \gamma_j$. Gilt $\gamma = 0$ so ist die erste Aussage trivial und die zweite Bedingung ist nicht erfüllt.
Sei also $\gamma > 0$ und weiter $\{x_k\} \in I(\tau, x^*)$. Zur Vereinfachung der Schreibweise setze man $\epsilon_k := \|x_k - x^*\|$, $\eta_k := \gamma\epsilon_k$ für $k = 0, 1, \ldots$ und $\delta_j := \frac{\gamma_j}{\gamma}$ für $j = 1, 2, \ldots, m$. Dann gilt $\sum_{j=0}^{m} \delta_j = \sum_{j=0}^{m} \frac{\gamma_j}{\gamma} = 1$ und

$$\underbrace{\gamma \|x_{k+1} - x^*\|}_{\eta_{k+1}} \leq \underbrace{\gamma \|x_k - x^*\|}_{\eta_k} \sum_{j=0}^{m} \gamma_j \underbrace{\|x_{k-j} - x^*\|}_{\epsilon_{k-j} = \frac{\eta_{k-j}}{\gamma}}.$$

D.h.

$$\eta_{k+1} \leq \eta_k \sum_{j=0}^{m} \delta_j \eta_{k-j} \qquad \forall k \geq k_0 \geq m.$$

Da $\epsilon_k \to 0$ gibt es $\eta < 1$ und k' mit $\eta_k \leq \eta$ für $k \geq k' \geq k_0$. Damit folgt

$$\eta_{k'+m+1} \leq \eta_{k'+m} \sum_{j=0}^{m} \delta_j \eta_{k'+m-j} \leq \eta \underbrace{\sum_{j=0}^{m} \delta_j}_{=1} \eta = \eta^2$$

und

$$\eta_{k'+m+2} \leq \eta_{k'+m+1} \sum_{j=0}^{m} \delta_j \eta_{k'+m+1-j} \leq \eta^2 \eta = \eta^3.$$

Per vollständigen Induktion kann man zeigen, dass

$$\eta_{k'+i} \leq \eta_{k'+i-1} \sum_{j=0}^{m} \delta_j \eta_{k'+i-1-j} \leq \eta^{\mu_i} \eta^{\mu_{i-m}} = \eta^{\mu_{i+1}},$$

wobei $i = m, m+1, \ldots$, $\mu_0 = \mu_1 = \cdots = \mu_m = 1$ und $\mu_{i-1} := \mu_i + \mu_{i-m}$ gilt. D.h. es gilt

$$\eta_{k'+i} \leq \eta^{\mu_{i+1}}.$$

Als nächsten Schritt zeigt man für $i = 0, 1, \ldots$ die Ungleichung $\mu_i \geq \alpha t^i$ mit $\alpha = t^{-m}$ (t wie oben). Es gilt $\alpha t^i = t^{i-m}$.
Nach Definition ist $\mu_i > 1$ und nach Lemma A.1.12 gilt $t > 1$, d.h. für $i \leq m$ ist $i - m \leq 0$ und damit ist $\mu_i \geq 1 \geq \alpha t^i$. Aufgrund von

$$t^{-1} + t^{-m-1} = \frac{1}{t} + \frac{1}{t^{m+1}} = \frac{t^m + 1}{t^{m+1}} \overset{\text{Vor.}}{=} \frac{t^{m+1}}{t^{m+1}} = 1$$

folgt

$$\mu_{i+1} = \mu_i + \mu_{i-m} \geq \alpha t^i + \alpha t^{i-m} = \alpha t^{i+1} \left(t^{-1} + t^{-m-1}\right) = \alpha t^{i+1}.$$

Desweiteren erhält man folgende Abschätzung:

$$\epsilon_{k'+i} = \frac{\eta_{k'+i}}{\gamma} \leq \frac{\eta^{\mu_i}}{\gamma} \leq \frac{\eta^{\alpha t^i}}{\gamma} \qquad \forall i \geq 0.$$

Somit gilt

$$R_t(\{x_k\}) = \limsup_{i \to \infty} \epsilon_{k'+i}^{\frac{1}{t^{k'+i}}} \leq \limsup_{i \to \infty} \underbrace{\frac{\eta^{\frac{\alpha t^i}{t^{k'+i}}}}{\gamma^{\frac{1}{t^{k'+i}}}}}_{\to 1} = \eta^{\frac{\alpha}{t^{k'}}} < 1.$$

Nach Lemma A.1.9 gilt für jedes $\epsilon > 0$, dass $R_{t-\epsilon}(\{x_k\}) = 0$ für alle Folgen $\{x_k\}$ aus $I(\tau, x^*)$. Daraus folgt, dass

$$R_{t-\epsilon}(\tau, x^*) = 0.$$

Es existiert also $R_{t-\epsilon}(\tau, x^*) < 1$ und da ϵ beliebig gewählt werden kann, gilt nach Bemerkung A.1.11

$$O_R(\tau, x^*) \geq t.$$

(2) Man nehme an, dass die zweite Bedingung erfüllt sei für eine beliebige Folge $\{x_k\}$ aus $I(\tau, x^*)$. Definiere $\eta_k := \beta\epsilon_k = \beta \|x_k - x^*\|$. Dann gilt $\eta_{k+1} \geq \eta_k \eta_{k-m}$ für alle $k \geq k_0$.
Dann sei $k' \geq k_0$ so gegeben, dass für festes $\tilde{\eta} < 1$ die Ungleichung $\eta_k \leq \tilde{\eta}$ für alle $k \geq k'$ gilt. Dann setze man $\eta := \min\{\eta_{k'}, \ldots, \eta_{k'+m}\}$. Analog zu (1) kann man zeigen, dass

$$\eta_{k'+i} \geq \eta^{\mu_i}, i = 0, 1, \ldots,$$

wobei μ_i wie in Teil (1) gewählt wird. Per vollständiger Induktion kann man $\mu_i \leq t^i$ für $i = 0, 1, \ldots$ zeigen, da die Aussage für $i = 0, 1, \ldots, m$ offensichtlich ist und für alle anderen i analog zu Teil (1)

$$\mu_{i+1} = \mu_i + \mu_{i-m} \leq t^i + t^{i-m} = t^{i+1}\left(t^{-1} + t^{-m-1}\right) = t^{i+1}$$

gilt. Zusammen ergibt das $\eta_{k'+i} \geq \eta^{\mu_i} \geq \eta^{t^i}$ für alle $i \geq 0$ bzw.

$$R_t(\{x_k\}) = \limsup_{i\to\infty} \epsilon_{k'+i}^{\frac{1}{t^{k'+i}}} = \limsup_{i\to\infty} \frac{\eta_{k'+i}^{\frac{1}{t^{k'+i}}}}{\underbrace{\beta^{\frac{1}{t^{k'+i}}}}_{\to 1}} \geq \limsup_{i\to\infty} \eta^{\frac{t^i}{t^{k'+i}}} = \eta^{\frac{1}{t^{k'}}} > 0.$$

D.h. $R_t(\tau, x^*) > 0$ und mit Bemerkung A.1.11 folgt

$$O_R(\tau, x^*) \leq t.$$

□

Lemma A.1.14
Sei τ ein iterativer Prozeß mit Limes-Punkt x^*. Dann gilt

$$O_Q(\tau, x^*) \leq O_R(\tau, x^*).$$

Beweis:
Sei $\{x_k\} \in I(\tau, x^*)$. Man muss zeigen, dass aus $Q_p(\{x_k\}) < \infty$ für $p > 1$ folgt, dass $R_p(\{x_k\}) < 1$ gilt.

(1) Sei $\epsilon_k := \|x_k - x^*\|$ für $k = 0, 1, \ldots$ und zu gegebenem $\epsilon > 0$ sei $\gamma := Q_p(\{x_k\}) + \epsilon$.
Dann existiert k_0 so, dass

$$\begin{aligned} \epsilon_{k+1} &\leq \gamma\epsilon_k^p \\ &\leq \gamma\gamma^p\epsilon_{k-1}^{p^2} \\ &\leq \cdots \leq \gamma^{1+p+\cdots+p^{k-k_0}}\epsilon_0^{p^{k-k_0+1}} \qquad \forall k \geq k_0. \end{aligned}$$

Dann gilt mit $\gamma' := \max\left\{1, \gamma^{\frac{1}{1-p}}\right\}$

$$\epsilon_{k+1}^{\frac{1}{p^{k+1}}} \leq \gamma^{\sum_{j=k_0+1}^{k+1}\frac{1}{p^j}}\epsilon_{k_0}^{\frac{1}{p^{k_0}}} = \left(\gamma^{\sum_{j=+1}^{k-k_0+1}\frac{1}{p^j}}\epsilon_{k_0}\right)^{\frac{1}{p^{k_0}}}.$$

Es gilt (geometrische Reihe) für $p \neq 1$

$$\frac{1}{p-1} = \frac{p}{p-1} - \frac{p-1}{p-1} = \frac{p}{p-1} - 1 = \sum_{j=0}^{\infty} \frac{1}{p^j} - 1 = \sum_{j=1}^{\infty} \frac{1}{p^j} \geq \sum_{j=1}^{k-k_0+1} \frac{1}{p^j}.$$

Damit folgt

$$\begin{aligned} \gamma^{\sum_{j=1}^{k-k_0+1} \frac{1}{p^j}} &\leq \gamma^{\frac{1}{p-1}} && \text{für } \gamma > 1, \\ \gamma^{\sum_{j=1}^{k-k_0+1} \frac{1}{p^j}} &= 1 && \text{für } \gamma = 1 \text{ und} \\ \gamma^{\sum_{j=1}^{k-k_0+1} \frac{1}{p^j}} &< 1 && \text{für } \gamma < 1. \end{aligned}$$

D.h. es gilt $\epsilon_{k+1}^{\frac{1}{p^{k+1}}} \leq (\gamma' \epsilon_{k_0})^{\frac{1}{p^{k_0}}}$ für $k \geq k_0$, wobei γ' als Maximum von $\left\{1, \gamma^{\frac{1}{p-1}}\right\}$ definiert ist. Da $\epsilon_k \to 0$ für $k \to \infty$, kann man annehmen, dass k_0 so groß gewählt wurde, dass $\gamma' \epsilon_{k_0} < 1$. Damit gilt

$$R_p(\{x_k\}) = \limsup_{k \to \infty} \epsilon_k^{\frac{1}{p^k}} \leq (\gamma' \epsilon_{k_0})^{\frac{1}{p^{k_0}}} < 1.$$

(2) Man nehme nun an, das $r = O_R(\tau, x^*) < O_Q(\tau, x^*) = q$. Nach Lemma A.1.5 gilt $Q_p(\tau, x^*) = 0$ für $p \in (r, q)$ und nach Lemma A.1.9 $R_p(\tau, x^*) = 1$ für $p \in (r, q)$. D.h. für jede Folge $\{x_k\} \in I(\tau, x^*)$ gilt $Q_{p'}(\{x_k\}) = 0$ für $p' := \frac{q+r}{2}$ und nach dem eben Gezeigten folgt $R_{p'}(\{x_k\}) < 1$.

Dann gilt aber $R_p(\{x_k\}) = 0$ für $p \in (r, p')$. Da das für alle Folgen aus $I(\tau, x^*)$ zutrifft, folgt $R_p(\tau, x^*) = 0$. Das ist ein Widerspruch zur Annahme $O_Q(\tau, x^*) > O_R(\tau, x^*)$ und damit erhält man

$$O_Q(\tau, x^*) \leq O_R(\tau, x^*).$$

□

Lemma A.1.15

(1) Gibt es $p \in [1, \infty)$ und eine Konstante c_2 so, dass für alle Folgen $\{x_k\} \in I(\tau, x^*)$ gilt

$$\|x_{k+1} - x^*\| \leq c_2 \|x_k - x^*\|^p > 0 \quad \forall k \geq k_0 = k_0(\{x_k\}),$$

so folgt $O_R(\tau, x^*) \geq O_Q(\tau, x^*) \geq p$.

(2) Gibt es eine Konstante c_1 und eine Folge $\{x_k\} \in I(\tau, x^*)$ mit

$$\|x_{k+1} - x^*\| \geq c_1 \|x_k - x^*\|^p > 0 \quad \forall k \geq k_0 = k_0(\{x_k\}),$$

so gilt $O_Q(\tau, x^*) \leq O_R(\tau, x^*) \leq p$.

(3) Gelten (1) und (2), so folgt $O_Q(\tau, x^*) = O_R(\tau, x^*) = p$.

Beweis:
(1) Falls die Ungleichung $\|x_{k+1} - x^*\| \leq c_2 \|x_k - x^*\|^p$ erfüllt ist, so gilt ebenfalls $Q_p(\{x_k\}) \leq c_2$ für alle Folgen $\{x_k\} \in I(\tau, x^*)$. Daraus folgt, dass $Q_p(\tau, x^*) \leq c_2 < \infty$, und mit Lemma A.1.7 und Lemma A.1.14 ergibt sich $O_R(\tau, x^*) \geq O_Q(\tau, x^*) \geq p$.

(2) Es gelte $\|x_{k+1} - x^*\| \geq c_1 \|x_k - x^*\|^p$ für irgendeine Folge $\{x_k\} \in I(\tau, x^*)$. Definiere $\epsilon_k := \|x_k - x^*\| > 0$ für $k \geq k_0$. Dann gilt

$$\epsilon_{k+1} \geq c_1 \epsilon_k^p \geq \cdots \geq c_1^{1+p+\cdots+p^{k-k_0}} \epsilon_{k_0}^{p^{k-k_0+1}} \quad \forall k \geq k_0.$$

Damit erhält man für $p = 1$

$$\begin{aligned} R_1(\{x_k\}) &= \limsup_{k\to\infty} \|x_k - x^*\|^{\frac{1}{k}} = \limsup_{k\to\infty} \epsilon_k^{\frac{1}{k}} \\ &\geq \lim_{k\to\infty} \left(c_1^{k-k_0+1}\epsilon_{k_0}\right)^{\frac{1}{k}} = \lim_{k\to\infty} c_1^{1-\frac{k_0-1}{k}} \epsilon_{k_0}^{\frac{1}{k}} = c_1 > 0 \end{aligned}$$

und für $p > 1$

$$\epsilon_{k+1}^{\frac{1}{p^{k+1}}} \geq c_1^{\frac{1}{p^{k+1}}} \epsilon_k^{\frac{1}{p^k}} \geq \cdots \geq c_1^{\sum_{j=k_0}^{k} \frac{1}{p^{j+1}}} \epsilon_{k_0}^{\frac{1}{p^{k_0}}} \geq \min\left\{1, c_1^{\frac{1}{p-1}}\right\} \epsilon_{k_0}^{\frac{1}{p^{k_0}}}.$$

Damit gilt

$$R_p(\{x_k\}) = \limsup_{k\to\infty} \epsilon_k^{\frac{1}{p^k}} \geq \lim_{k\to\infty} \min\left\{1, c_1^{\frac{1}{p-1}}\right\} \epsilon_{k_0}^{\frac{1}{p^{k_0}}} > 0.$$

Somit folgt $R_p(\tau, x^*) > 0$. Nach Bemerkung A.1.11 gilt $O_R(\tau, x^*) \leq p$ und mit Lemma A.1.14 folgt $O_Q(\tau, x^*) \leq O_R(\tau, x^*) \leq p$. □

A.2 Anwendung auf das Newton-Verfahren

Definition A.2.1

(1) Die Menge der stetigen und linearen Funktionen von $\mathbb{R}^n \to \mathbb{R}^n$ wird hier mit $\mathcal{L}(\mathbb{R}^n)$ bezeichnet.

(2) Sei $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ Frechet-differenzierbar auf $D_0 \subset D$ und $J : D_J \times D_h \subset \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{L}(\mathbb{R}^n)$. Dann ist J eine konsistente Approximation *von F' auf $D_0 \subset D_J$, falls $0 \in \overline{D_h}$ und $\lim_{h\to 0; h\in D_h} J(x,h) = F'(x)$ einheitlich für $x \in D_0$.*

(3) Gibt es Konstanten c und $r > 0$ so, dass

$$\|F'(x) - J(x,h)\| \leq c\|h\| \qquad \forall x \in D_0, h \in D_h \cap B_r(0),$$

so heißt J starke konsistente Approximation *von F' auf D_0.*

Lemma A.2.2
Sei $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ Frechet-differenzierbar in einer offenen Umgebung $S_0 \subset D$ eines Punktes $x^* \in D$, für den $F(x^*) = 0$ gilt, $F'(x^*)$ nicht-singulär und F' stetig ist. Sei weiter $J : D_J \times D_h \subset \mathbb{R}^n \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^n)$ eine konsistente Approximation von F' in S_0.
Dann existieren $\delta > 0$, $r > 0$ so, dass die Funktion

$$G(x,h) = x - J(x,h)^{-1}F(x)$$

wohldefiniert ist für alle $x \in S = B_\delta(x^*)$, $h \in D_h' = D_h \cap B_r(0)$. Die Funktion erfüllt

$$\|x^* - G(x,h)\| \leq \omega(x,h)\|x - x^*\| \qquad \forall x \in S, h \in D_h',$$

wobei $\omega(x,h) \to 0$ für $x \to x^*$, $h \to 0$ und $h \in D_h'$.
Ist J eine starke konsistente Approximation von F' in S_0 und gilt

$$\|F'(x) - F'(x^*)\| \leq \gamma\|x - x^*\| \qquad \forall x \in S_0$$

so gibt es Konstanten α_1, α_2 so, dass

$$\|x^* - G(x,h)\| \leq \alpha_1\|x - x^*\|^2 + \alpha_2\|h\|\|x - x^*\| \qquad \forall x \in S, h \in D_h'.$$

Beweis:
Setze $\beta := \left\|F'(x^*)^{-1}\right\|$ und sei $\epsilon \in \left(0, \frac{1}{2\beta}\right)$. Da J konsistente Approximation in S_0 ist, gibt es $r > 0$ so, dass D'_h nicht leer ist und dass

$$\|F'(x) - J(x,h)\| \leq \frac{1}{2}\epsilon \quad \forall x \in s_0, h \in D'_h.$$

Aufgrund der Stetigkeit von F' in x^* gibt es $\delta > 0$, so dass $S := B_\delta(x^*) \subset S_0$ und

$$\|F'(x) - F'(x^*)\| \leq \frac{1}{2}\epsilon \quad \forall x \in S.$$

Dann gilt

$$\begin{aligned} \|F'(x^*) - J(x,h)\| &\leq \|F'(x^*) - F'(x)\| + \|F'(x) - J(x,h)\| \\ &\leq \frac{1}{2}\epsilon + \frac{1}{2}\epsilon = \epsilon \quad \forall x \in S, h \in D'_h \end{aligned}$$

und nach Lemma 2.2.3 ist $J(x,h)$ invertierbar und erfüllt

$$\left\|J(x,h)^{-1}\right\| < \eta = \frac{\beta}{1-\beta\epsilon} \quad \forall x \in S, h \in D'_h.$$

D.h. G ist wohldefiniert auf $S \times D'_h$ und

$$\begin{aligned} \|G(x,h) - x^*\| &= \left\|x - J(x,h)^{-1}f(x) - x^*\right\| \\ &= \left\|J(x,h)^{-1}J(x,h)(x-x^*) - J(x,h)^{-1}F(x)\right\| \\ &= \left\|J(x,h)^{-1}\left[J(x,h)(x-x^*) - F(x)\right]\right\| \\ &\leq \left\|J(x,h)^{-1}\right\| \| J(x,h)(x-x^*) - F'(x)(x-x^*) + F'(x)(x-x^*) \\ &\qquad -F'(x^*)(x-x^*) + F'(x^*)(x-x^*) + F(x^*) - F(x) \| \\ &\leq \eta\left[\|J(x,h) - F'(x)\| + \|F'(x) - F'(x^*)\|\right]\|x-x^*\| \\ &\quad +\eta\|F(x) - F(x^*) - F'(x^*)(x-x^*)\|. \end{aligned}$$

Definiere

$$q(x) := \begin{cases} 0 & , x = x^* \\ \frac{\|F(x) - F(x^*) - F'(x^*)(x-x^*)\|}{\|x-x^*\|} & , \text{sonst} \end{cases}$$

und

$$\omega(x,h) := \eta\left[\|J(x,h) - F'(x)\| + \|F'(x) - F'(x^*)\| + q(x)\right].$$

Für $x = x^*$ gilt

$$\omega(x^*,h) = \eta \underbrace{\|J(x^*,h) - F'(x)\|}_{\to 0 \text{ für } h \to 0}.$$

D.h. $\|x^* - G(x,h)\| \leq \omega(x,h)\|x-x^*\|$ mit $\omega(x,h) \to 0$ für $x \to x^*$ und $h \to 0$.
Man nehme nun an, dass $\|F'(x) - F'(x^*)\| \leq \gamma\|x-x^*\|$ für alle $x \in S_0$ gilt. Dann folgt nach Bemerkung A.1.3

$$\begin{aligned} q(x) = \frac{\|F(x) - F(x^*) - F'(x^*)(x-x^*)\|}{\|x-x^*\|} &\leq \sup\|F'(x^* + t(x-x^*)) - F'(x^*)\| \\ &\leq \gamma\|x-x^*\| \quad \forall x \in S. \end{aligned}$$

Dann gilt

$$\begin{aligned}\|x^* - G(x,h)\| &\leq \eta\left[\|J(x,h) - F'(x)\| + \|F'(x) - F'(x^*)\| + q(x)\right]\|x - x^*\| \\ &\leq \eta\left[\|J(x,h) - F'(x)\| + \gamma\|x - x^*\| + \gamma\|x - x^*\|\right]\|x - x^*\| \\ &\leq \eta\underbrace{\|J(x,h) - F'(x)\|}_{c\|h\|}\|x - x^*\| + 2\gamma\|x - x^*\|\|x - x^*\|.\end{aligned}$$

Mit $\alpha_2 := \eta c$ und $\alpha_1 := 2\gamma$ folgt die gesuchte Ungleichung

$$\|x^* - G(x,h)\| \leq \alpha_2\|h\|\|x - x^*\| + \alpha_1\|x - x^*\|^2.$$

□

Satz A.2.3
Sei $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ Frechet-differenzierbar in einer offenen Umgebung $S_0 \subset D$ eines Punktes $x^* \in D$, für den $F(x^*) = 0$ gilt, $F'(x^*)$ nicht-singulär und F' Lipschitz-stetig ist. Sei weiter $J : D_J \times D_h \subset \mathbb{R}^n \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^n)$ eine starke konsistente Approximation von F' in S_0.
Angenommen für einige $\{h_k\} \subset D_h$ ist die durch die Iterationsvorschrift

$$x_{k+1} := x_k - J(x_k, h_k)^{-1}F(x_k) \qquad \forall k = 0, 1, \ldots$$

definierte Folge $\{x_k\} \in I(\tau, x^*)$ wohldefiniert und konvergieren gegen x^*. Gilt zusätzlich

$$\|h_k\| \leq \beta_1\|F(x_k)\| \tag{A.1}$$

für alle $k \geq k_0$, dann gilt $O_{\mathrm{R}}(\{x_k\}) \geq O_{\mathrm{Q}}(\{x_k\}) \geq 2$. Gilt

$$\|h_k\| \leq \beta_2\|x_k - x_{k-1}\| \tag{A.2}$$

für alle $k \geq k_0$, dann gilt $O_{\mathrm{R}}(\{x_k\}) \geq \frac{1+\sqrt{5}}{2}$.

Beweis:
Nach Lemma A.2.2 existieren $\delta > 0$ und $r > 0$ so, dass für jeder $x \in S := B_\delta(x^*) \subset S_0$ und $h \in D_h' := D_h \cap B_r(0)$ gilt

$$\|x^* - G(x,h)\| \leq \alpha_1\|x - x^*\|^2 + \alpha_2\|h\|\|x - x^*\|.$$

Man nehme nun an, dass (A.1) gilt. Dann folgt aus $x_k \to x^*$ die Konvergenz von $F(x_k)$ gegen $F(x^*) = 0$ für $k \to \infty$. D.h. $x_k \in S$ und $h \in D_h'$ für alle $k \geq k_1 \geq k_0$. Es gilt

$$\begin{aligned}\|x_{k+1} - x^*\| &= \left\|x_k - J(x_k,h_k)^{-1}F(x_k) - x^*\right\| \\ &= \|G(x_k,h_k) - x^*\| \\ &\leq \alpha_1\|x_k - x^*\|^2 + \alpha_2\|h_k\|\|x_k - x^*\| \\ &\leq \alpha_1\|x_k - x^*\|^2 + \alpha_2\beta_1\|F(x_k)\|\|x_k - x^*\|.\end{aligned}$$

Für $F(x_k)$ erhält man die Abschätzung

$$\begin{aligned}\|F(x_k)\| &\leq \|F(x_k) - F(x^*) - F'(x^*)(x_k - x^*)\| + \|F'(x^*)(x_k - x^*)\| \\ &\leq \epsilon_k\|x_k - x^*\| + \|F'(x^*)\|\|x_k - x^*\| \\ &= \left[\epsilon_k + \|F'(x^*)\|\right]\|x_k - x^*\|.\end{aligned}$$

Es gilt $\epsilon_k \to 0$ für $k \to \infty$. Damit folgt

$$\begin{aligned}\|x_{k+1} - x^*\| &\leq \alpha_1\|x_k - x^*\|^2 + \alpha_2\beta_1\left[\epsilon_k + \|F'(x^*)\|\right]\|x_k - x^*\|^2 \\ &\leq c_2\|x_k - x^*\|^2.\end{aligned}$$

Nach Lemma A.1.14 gilt somit $O_R(\{x_k\}) \geq O_Q(\{x_k\}) \geq 2$.
Ist (A.2) erfüllt, so folgt für alle $k \geq k_1$

$$\begin{aligned} \|x_{k+1} - x^*\| &\leq \alpha_1 \|x_k - x^*\|^2 + \alpha_2\beta_2 \|x_k - x_{k-1}\| \|x_k - x^*\| \\ &\leq (\alpha_1 + \alpha_2\beta_2) \|x_k - x^*\|^2 + \alpha_2\beta_2 \|x_{k-1} - x^*\| \|x_k - x^*\| . \end{aligned}$$

Nach Lemma A.1.13 mit m=1 ist $O_R(\{x_k\}) \geq \tau$, wobei $\tau = \frac{1+\sqrt{5}}{2}$ die eindeutige positive Wurzel von $t^2 - t - 1 = 0$ ist. □

Dieses allgemeine Resultat kann man nun auf das Newton-Verfahren anwenden. Für das exakte Verfahren ist bereits in Kapitel 2 die quadratische Konvergenz bewiesen worden. Einen alternativen Beweis findet man hier, indem man $J(x_k, h_k) := F'(x_k)$ definiert. Damit sind die h_k irrelevant und können als 0 gewählt werden. Damit erfüllt man beide Ungleichungen aus Satz A.2.3. Die übrigen Voraussetzungen sind die gleichen wie in Satz 2.2.1. Damit sind diese ebenfalls erfüllt. Die R-Konvergenzordnung des exakten Newton-Verfahrens ist also ≥ 2.

In Abschnitt 2.2.2 wurde eine Sicherheit bei der Wahl der Forcing-Terme eingeführt, die dann anschlägt, wenn $\eta_{k-1}^{\frac{1+\sqrt{5}}{2}} > 0.1$ ist, wobei η_{k-1} der Forcing-Term aus der vorherigen Newton-Iteration ist. Um eine Motivation für den Exponenten zu bekommen, betrachtet man zunächst die Bedingung an den neuen Newton-Schritt s_k:

$$\|F(x_k) + F'(x_k)s_k\| \leq \eta_k \|F(x_k)\| . \tag{A.3}$$

Diese Ungleichung ist nicht eindeutig lösbar. Daher geht man zur Motivation von zwei Vereinfachungen aus. Statt $\leq$ wird $=$ gewählt und man betrachtet die Gleichung ohne die Norm. D.h. man wählt s_k, so dass

$$F(x_k) + F'(x_k)s_k = \eta_k F(x_k)$$

gilt. Dann erfüllt s_k die Gleichung (A.3). Da $s_k = x_{k+1} - x_k$ gilt, folgt

$$\begin{aligned} F'(x_k)s_k &= (\eta_k Id - Id) F(x_k) \\ x_{k+1} &= x_k + F'(x_k)^{-1} (\eta_k Id - Id) F(x_k). \end{aligned}$$

Dann kann man J folgendermaßen definieren:

$$\begin{aligned} J(x_k, \eta_k) &:= -(\eta_k Id - Id) F'(x_k) \\ &= (1 - \eta_k) F'(x_k). \end{aligned}$$

Für $\eta_k \to 0$ gilt $J(x_k, \eta_k) \to F'(x_k)$. Außerdem erhält man die Abschätzung

$$\begin{aligned} \|F'(x_k) - J(x_k, \eta_k)\| &= \|F'(x_k) - (1 - \eta_k)F'(x_k)\| \\ &= \|\eta_k F'(x_k)\| \\ &\leq \|\eta_k\| \underbrace{\|F'(x_k)\|}_{konstant} . \end{aligned}$$

D.h. dieses J ist eine stark konsistente Approximation von F' und die Bedingungen von Satz A.2.3 sind erfüllt. Für diese spezielle Wahl von J besitzt das inexakte Newton-Verfahren die R-Konvergenzordnung $\frac{1+\sqrt{5}}{2}$.

Anhang B

Weitere Testergebnisse

Im Folgenden werden die Testergebnisse, die in Kapitel 5 aus Platzgründen keinen Platz fanden, nachgetragen. Die optimalen Parameter η_0 und η_{max} wurden gesucht. Die übrigen Parameter θ_{min}, θ_{max} und t wurden konstant gehalten (vgl. Kapitel 5).

Die Striche in den folgenden Tabellen bedeuten, dass das Verfahren nicht erfolgreich war. In den hier behandelten Beispielen traten zwei Abbruchmöglichkeiten auf. Entweder konnte der Löser keine Lösung innerhalb seiner 10.000 Iterationen berechnen oder die vorgeschriebene Grenze für das Residuum wurde innerhalb der 100 Iterationen des Newton-Verfahrens nicht unterschritten.
Falls das Verfahren für eine Viskosität ν und die Parameter η_0 und η_{max} versagt hat, wurde aus theoretischen Überlegungen darauf verzichtet, die gleichen Parameter mit einer niedrigeren Viskosität zu testen. In diesen Fällen wurde ebenfalls ein Strich in der Tabelle gemacht.

B.1 Testbeispiel 1

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	23	20	16	11	7	4	4
10^{-2}	23	20	16	11	7	4	4
10^{-4}	22	20	16	10	6	4	4
10^{-6}	21	19	15	10	6	4	4

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	40	39	44	37	30	28	28
10^{-2}	40	39	44	37	30	28	28
10^{-4}	39	41	47	38	29	31	28
10^{-6}	40	41	46	40	31	32	30

Tabelle B.1: Newton- und GMRES-Iterationen für $\nu = 10^0$ in Abhängigkeit von η_0 und η_{max} für 960 Zellen.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	23	21	15	11	7	5	4
10^{-2}	23	21	15	11	7	5	4
10^{-4}	20	18	16	10	6	5	4
10^{-6}	21	19	15	9	6	5	4

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	39	40	38	36	29	27	26
10^{-2}	39	40	38	36	29	27	26
10^{-4}	37	37	46	36	29	29	26
10^{-6}	39	42	46	36	31	32	31

Tabelle B.2: Newton- und GMRES-Iterationen für $\nu = 10^0$ in Abhängigkeit von η_0 und η_{max} für 3840 Zellen.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	10^{-4}
10^{-1}	21(3)	18(3)	16(3)	14(2)	6	4	4
10^{-2}	23	23	12	10	6	4	4
10^{-4}	23	23	13	10	6	4	4
10^{-6}	23	23	13	10	6	4	4

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	10^{-4}
10^{-1}	41	38	38	36	23	23	22
10^{-2}	36	40	28	28	23	23	22
10^{-4}	37	39	25	28	24	25	22
10^{-6}	38	42	28	28	25	26	24

Tabelle B.3: Newton- und GMRES-Iterationen für $\nu = 10^{-2}$ in Abhängigkeit von η_0 und η_{max} für 960 Zellen. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	10^{-4}
10^{-1}	21(2)	20(2)	16(2)	14(3)	10(2)	5	4
10^{-2}	21(2)	20(2)	16(2)	14(3)	10(2)	5	4
10^{-4}	23	23	12	10	6	5	4
10^{-6}	23	22	13	10	6	5	4

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	10^{-4}
10^{-1}	40	39	33	37	34	23	20
10^{-2}	40	39	33	37	34	23	20
10^{-4}	36	39	25	28	23	23	20
10^{-6}	38	41	25	28	24	22	24

Tabelle B.4: Newton- und GMRES-Iterationen für $\nu = 10^{-2}$ in Abhängigkeit von η_0 und η_{max} für 3840 Zellen. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	6	5	5
10^{-2}	-	-	12	10	6	5	5
10^{-4}	-	-	12	10	6	5	5
10^{-6}	-	-	12	10	6	5	4

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	25	24	24
10^{-2}	-	-	30	27	25	24	24
10^{-4}	-	-	30	27	25	24	24
10^{-6}	-	-	28	29	24	24	23

Tabelle B.5: Newton- und GMRES-Iterationen für $\nu = 10^{-4}$ in Abhängigkeit von η_0 und η_{max} für 960 Zellen.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	-	5	4
10^{-2}	-	-	-	-	-	5	4
10^{-4}	-	-	13	11	7	6	4
10^{-6}	-	-	12	10	6	5	4

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	-	24	20
10^{-2}	-	-	-	-	-	24	20
10^{-4}	-	-	30	30	25	24	20
10^{-6}	-	-	28	29	24	20	23

Tabelle B.6: Newton- und GMRES-Iterationen für $\nu = 10^{-4}$ in Abhängigkeit von η_0 und η_{max} für 3840 Zellen.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	7	6	6
10^{-2}	-	-	15	11	7	6	6
10^{-4}	-	-	15	11	7	6	6
10^{-6}	-	-	13	10	6	5	4

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	25	26	30
10^{-2}	-	-	30	30	25	26	30
10^{-4}	-	-	30	30	25	26	30
10^{-6}	-	-	29	29	25	24	22

Tabelle B.7: Newton-und GMRES- Iterationen für $\nu = 10^{-6}$ in Abhängigkeit von η_0 und η_{max} für 960 Zellen.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	-	5	6
10^{-2}	-	-	-	-	-	6	6
10^{-4}	-	-	13	11	7	6	6
10^{-6}	-	-	12	10	6	4	4

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	-	23	28
10^{-2}	-	-	-	-	-	26	28
10^{-4}	-	-	27	30	26	26	28
10^{-6}	-	-	29	27	25	20	24

Tabelle B.8: Newton- und GMRES-Iterationen für $\nu = 10^{-6}$ in Abhängigkeit von η_0 und η_{max} für 3840 Zellen.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	8	8	7
10^{-2}	-	-	-	-	8	8	7
10^{-4}	-	-	-	-	8	8	7
10^{-6}	-	-	12	10	6	5	5

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	27	33	34
10^{-2}	-	-	-	-	27	33	34
10^{-4}	-	-	-	-	27	33	34
10^{-6}	-	-	28	29	25	24	26

Tabelle B.9: Newton- und GMRES-Iterationen für $\nu = 10^{-8}$ in Abhängigkeit von η_0 und η_{max} für 960 Zellen.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	-	7	9
10^{-2}	-	-	-	-	-	7	9
10^{-4}	-	-	-	-	10	9	9
10^{-6}	-	-	12	10	6	5	4

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	-	-	29	43
10^{-2}	-	-	-	-	-	29	43
10^{-4}	-	-	-	-	31	31	43
10^{-6}	-	-	29	27	25	24	21

Tabelle B.10: Newton- und GMRES-Iterationen für $\nu = 10^{-8}$ in Abhängigkeit von η_0 und η_{max} für 3840 Zellen.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-4}	-	-	-	-	-	-	-
10^{-6}	-	-	12	10	6	5	5

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-4}	-	-	-	-	-	-	-
10^{-6}	-	-	28	29	25	23	26

Tabelle B.11: Newton- und GMRES-Iterationen für $\nu = 10^{-10}$ in Abhängigkeit von η_0 und η_{max} für 960 Zellen.

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-4}	-	-	-	-	-	-	-
10^{-6}	-	-	12	10	6	5	5

η_0	η_{max}						
	0.75	0.5	0.25	0.1	0.01	0.001	0.0001
10^{-4}	-	-	-	-	-	-	-
10^{-6}	-	-	29	27	25	24	26

Tabelle B.12: Newton- und GMRES-Iterationen für $\nu = 10^{-10}$ in Abhängigkeit von η_0 und η_{max} für 3840 Zellen.

B.2 Testbeispiel 2

B.2.1 Zusätzliche Reaktion $c = 1$

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	20	9	7	5	4
10^{-2}	20	9	7	5	4
10^{-3}	20	9	6	5	4
10^{-4}	20	9	6	4	4
10^{-5}	18	9	6	5	4

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	32	29	31	26	27
10^{-2}	32	29	31	26	27
10^{-3}	35	33	26	26	27
10^{-4}	34	29	29	24	27
10^{-5}	33	34	30	29	28

Tabelle B.13: Newton- und GMRES-Iterationen für $\nu = 1$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	21	10	7	5	4
10^{-2}	21	10	7	5	4
10^{-3}	21	10	7	5	4
10^{-4}	18	9	6	5	4
10^{-5}	17	9	6	4	4

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	35	35	31	26	27
10^{-2}	35	35	31	26	27
10^{-3}	28	35	31	26	27
10^{-4}	33	29	29	28	27
10^{-5}	30	33	30	24	27

Tabelle B.14: Newton- und GMRES-Iterationen für $\nu = 1$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	20	10	6	5	4
10^{-2}	20	10	6	5	4
10^{-3}	20	8	6	5	4
10^{-4}	17	9	6	5	4
10^{-5}	16	8	6	4	4

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	30	20	19	19	18
10^{-2}	30	20	19	19	18
10^{-3}	30	21	21	19	18
10^{-4}	26	24	21	20	18
10^{-5}	26	23	22	17	19

Tabelle B.15: Newton- und GMRES-Iterationen für $\nu = 10^{-1}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	19	10	6	5	4
10^{-2}	19	10	6	5	4
10^{-3}	19	7	5	5	4
10^{-4}	17	8	6	5	4
10^{-5}	16	8	5	4	4

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	29	20	19	19	19
10^{-2}	29	20	19	19	19
10^{-3}	28	21	17	19	19
10^{-4}	26	22	21	20	19
10^{-5}	26	23	18	17	19

Tabelle B.16: Newton- und GMRES-Iterationen für $\nu = 10^{-1}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	16	9	6	5	5
10^{-2}	16	9	6	5	5
10^{-3}	15	7	6	5	5
10^{-4}	13	8	6	5	5
10^{-5}	13	9	6	5	5

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	23	20	16	17	19
10^{-2}	23	20	16	17	19
10^{-3}	20	18	18	17	19
10^{-4}	20	19	17	17	19
10^{-5}	20	22	18	18	20

Tabelle B.17: Newton- und GMRES-Iterationen für $\nu = 10^{-2}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	16	8	6	5	5
10^{-2}	16	8	6	5	5
10^{-3}	21	7	6	5	5
10^{-4}	12	8	5	5	5
10^{-5}	12	8	5	5	5

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	23	17	16	17	17
10^{-2}	23	17	16	17	17
10^{-3}	27	18	18	17	17
10^{-4}	19	19	15	17	17
10^{-5}	19	19	15	17	17

Tabelle B.18: Newton- und GMRES-Iterationen für $\nu = 10^{-2}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	18	8	7	6	5
10^{-2}	18	8	7	6	5
10^{-3}	16	8	6	6	5
10^{-4}	15	8	6	5	5
10^{-5}	13	8	6	5	5

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	25	16	18	19	19
10^{-2}	25	16	18	19	19
10^{-3}	22	18	17	19	19
10^{-4}	21	17	16	16	19
10^{-5}	20	18	17	17	20

Tabelle B.19: Newton- und GMRES-Iterationen für $\nu = 5 \cdot 10^{-3}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	17	9	6	5	5
10^{-2}	17	9	6	5	5
10^{-3}	16	7	6	5	5
10^{-4}	15	8	6	5	5
10^{-5}	15	8	6	5	5

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	22	18	15	15	18
10^{-2}	22	18	15	15	18
10^{-3}	22	15	17	15	18
10^{-4}	21	16	17	16	18
10^{-5}	21	16	17	16	18

Tabelle B.20: Newton- und GMRES-Iterationen für $\nu = 5 \cdot 10^{-3}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	16	8	6	6	6
10^{-2}	16	8	6	6	6
10^{-3}	17	9	7	6	6
10^{-4}	14	9	6	6	6
10^{-5}	14	9	6	6	6

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	20	19	15	17	20
10^{-2}	20	19	15	17	20
10^{-3}	22	18	18	17	20
10^{-4}	22	19	16	18	20
10^{-5}	22	19	16	18	20

Tabelle B.21: Newton- und GMRES-Iterationen für $\nu = 10^{-3}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	13	8	7	6	6
10^{-2}	13	8	7	6	6
10^{-3}	16	9	7	6	6
10^{-4}	14	8	6	6	6
10^{-5}	14	8	6	6	6

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	18	19	17	17	20
10^{-2}	18	19	17	17	20
10^{-3}	20	18	18	17	20
10^{-4}	22	16	16	17	20
10^{-5}	22	16	16	17	20

Tabelle B.22: Newton- und GMRES-Iterationen für $\nu = 10^{-3}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	18	9	7	7	7
10^{-2}	24	9	7	7	7
10^{-3}	13	9	8	7	7
10^{-4}	13	8	8	7	7
10^{-5}	13	8	8	7	7

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	17	18	17	20	22
10^{-2}	24	18	17	20	22
10^{-3}	19	19	19	20	22
10^{-4}	21	17	20	20	22
10^{-5}	21	17	20	20	22

Tabelle B.23: Newton- und GMRES-Iterationen für $\nu = 5 \cdot 10^{-4}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	15	9	7	7	7
10^{-2}	15	9	7	7	7
10^{-3}	14	9	7	7	7
10^{-4}	14	9	7	7	7
10^{-5}	14	9	7	7	7

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	22	20	16	19	22
10^{-2}	22	20	16	19	22
10^{-3}	21	17	17	19	22
10^{-4}	21	17	18	20	22
10^{-5}	21	17	18	20	22

Tabelle B.24: Newton- und GMRES-Iterationen für $\nu = 5 \cdot 10^{-4}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	17	11	8	7	7
10^{-2}	18	11	8	7	7
10^{-3}	14	9	7	7	7
10^{-4}	15	8	7	7	7
10^{-5}	15	8	7	7	7

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	24	21	17	19	22
10^{-2}	24	21	17	19	22
10^{-3}	21	18	18	19	22
10^{-4}	23	17	17	20	22
10^{-5}	23	17	17	20	22

Tabelle B.25: Newton- und GMRES-Iterationen für $\nu = 10^{-4}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$ und $c = 1$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	15	11	10	11	10
10^{-2}	15	11	10	11	10
10^{-3}	16	12	11	11	10
10^{-4}	17	12	11	10	10
10^{-5}	17	12	11	10	10

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	23	22	22	26	30
10^{-2}	23	22	22	26	30
10^{-3}	24	25	24	26	30
10^{-4}	26	21	24	23	30
10^{-5}	26	21	24	23	30

Tabelle B.26: Newton- und GMRES-Iterationen für $\nu = 10^{-4}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$ und $c = 1$.

B.2.2 Ohne zusätzliche Reaktion

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	20	10	7	5	4
10^{-2}	20	10	7	5	4
10^{-3}	20	10	6	5	4
10^{-4}	21	9	6	4	4
10^{-5}	18	9	6	5	4

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	31	33	31	26	27
10^{-2}	31	33	31	26	27
10^{-3}	35	35	28	26	27
10^{-4}	35	29	29	24	27
10^{-5}	32	35	30	29	27

Tabelle B.27: Newton- und GMRES-Iterationen für $\nu = 1$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	20	10	6	5	4
10^{-2}	20	10	6	5	4
10^{-3}	20	10	6	5	4
10^{-4}	18	9	6	4	4
10^{-5}	22	8	6	4	4

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	34	35	26	27	27
10^{-2}	34	35	26	27	27
10^{-3}	34	35	26	27	27
10^{-4}	31	29	29	22	27
10^{-5}	39	33	30	24	27

Tabelle B.28: Newton- und GMRES-Iterationen für $\nu = 1$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	20	9	6	5	4
10^{-2}	20	9	6	4	4
10^{-3}	19	8	6	5	4
10^{-4}	16	8	5	5	4
10^{-5}	16	8	6	4	4

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	30	19	19	19	18
10^{-2}	30	19	19	19	18
10^{-3}	28	21	21	19	18
10^{-4}	25	22	17	20	18
10^{-5}	25	23	22	17	19

Tabelle B.29: Newton- und GMRES-Iterationen für $\nu = 10^{-1}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	19	10	6	4	4
10^{-2}	19	10	6	4	4
10^{-3}	19	8	6	5	4
10^{-4}	16	8	5	4	4
10^{-5}	16	8	5	4	4

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	29	21	19	15	18
10^{-2}	29	21	19	15	18
10^{-3}	28	20	21	15	18
10^{-4}	25	22	18	16	18
10^{-5}	26	23	18	17	19

Tabelle B.30: Newton- und GMRES-Iterationen für $\nu = 10^{-1}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	17	8	6	6	5
10^{-2}	17	8	6	6	5
10^{-3}	15	8	6	6	5
10^{-4}	15	9	6	6	5
10^{-5}	15	9	6	6	5

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	24	21	17	20	20
10^{-2}	24	21	17	20	20
10^{-3}	22	20	19	20	20
10^{-4}	22	22	18	21	20
10^{-5}	22	22	18	21	20

Tabelle B.31: Newton- und GMRES-Iterationen für $\nu = 10^{-2}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	16	8	6	5	5
10^{-2}	16	8	6	5	5
10^{-3}	14	8	6	5	5
10^{-4}	15	8	6	5	5
10^{-5}	15	8	6	5	5

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	23	21	17	17	20
10^{-2}	23	21	17	17	20
10^{-3}	21	19	19	17	20
10^{-4}	23	21	18	17	20
10^{-5}	23	21	18	17	20

Tabelle B.32: Newton- und GMRES-Iterationen für $\nu = 10^{-2}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	19	9	7	6	6
10^{-2}	19	9	7	6	6
10^{-3}	21	9	7	6	6
10^{-4}	16	9	7	6	6
10^{-5}	16	9	7	6	6

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	27	22	20	20	24
10^{-2}	27	22	20	20	24
10^{-3}	30	21	22	20	24
10^{-4}	23	23	22	20	24
10^{-5}	23	23	22	20	24

Tabelle B.33: Newton- und GMRES-Iterationen für $\nu = 5 \cdot 10^{-3}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	18	9	7	6	6
10^{-2}	18	9	7	6	6
10^{-3}	20	9	7	6	6
10^{-4}	16	9	7	6	6
10^{-5}	16	9	7	6	6

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	25	22	20	20	24
10^{-2}	25	22	20	20	24
10^{-3}	28	21	21	20	24
10^{-4}	25	23	22	20	24
10^{-5}	25	23	22	20	24

Tabelle B.34: Newton- und GMRES-Iterationen für $\nu = 5 \cdot 10^{-3}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	25(6)	19(7)	16(9)	17(10)	15(8)
10^{-2}	25(6)	19(7)	18(9)	17(10)	15(8)
10^{-3}	20(4)	19(10)	18(10)	17(10)	15(8)
10^{-4}	21(3)	20(10)	17(9)	16(8)	15(8)
10^{-5}	21(3)	20(4)	17(9)	16(8)	15(8)

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	49	54	54	65	69
10^{-2}	49	54	54	65	69
10^{-3}	38	51	54	65	69
10^{-4}	37	54	54	65	69
10^{-5}	37	54	54	65	69

Tabelle B.35: Newton- und GMRES-Iterationen für $\nu = 10^{-3}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	24(3)	15(4)	16(7)	12(4)	12(4)
10^{-2}	24(3)	15(4)	16(7)	12(4)	12(4)
10^{-3}	-	14(4)	13(4)	12(4)	12(4)
10^{-4}	-	15(4)	12(4)	12(4)	12(4)
10^{-5}	-	15(4)	12(4)	12(4)	12(4)

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	41	39	54	48	58
10^{-2}	41	39	54	48	58
10^{-3}	-	37	42	48	58
10^{-4}	-	40	40	49	58
10^{-5}	-	40	40	49	58

Tabelle B.36: Newton- und GMRES-Iterationen für $\nu = 10^{-3}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{128}$. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	36(33)	-
10^{-2}	-	-	-	36(33)	-
10^{-3}	-	32(23)	38(40)	36(33)	-
10^{-4}	-	32(24)	-	-	-
10^{-5}	-	31(24)	-	-	-

η_0	η_{max}				
	0.8	0.1	0.01	0.001	0.0001
10^{-1}	-	-	-	159	-
10^{-2}	-	-	-	159	-
10^{-3}	-	98	144	159	-
10^{-4}	-	91	-	-	-
10^{-5}	-	91	-	-	-

Tabelle B.37: Newton- und GMRES-Iterationen für $\nu = 5 \cdot 10^{-4}$ in Abhängigkeit von η_0 und η_{max} für $h = \frac{1}{64}$. Die Zahlen in den Klammern sind die Anzahlen der Backtracking-Schritte, die während der jeweiligen Rechnung gemacht wurden.

Für die Viskosität $\nu = 10^{-4}$ konnte keine Lösung mit dem inexakten Newton-Verfahren ohne zusätzliche Reaktion berechnet werden. Auch für $\nu = 5 \cdot 10^{-4}$ brach das Verfahren für jede Variation der Parameter η_0 und η_{max} auf dem feinen Gitter ab.

Algorithmen

Literaturverzeichnis

[AB07] Heng-Bin An and Zhong-Zhi Bai. A globally convergent newton-gmres method for large sparse systems of nonlinear equations. *Appl. Numer. Math.*, 57(3):235–252, 2007.

[AML07] Heng-Bin An, Ze-Yao Mo, and Xing-Ping Liu. A choice of forcing terms in inexact Newton method. *J. Comput. Appl. Math.*, 200(1):47–60, 2007.

[BB07] Stefania Bellavia and Stefano Berrone. Globalization strategies for newton-Krylov methods for stabilized FEM discretization of Navier-Stokes equations. *J. Comput. Phys.*, 226(2):2317–2340, 2007.

[BM01] Stefania Bellavia and Benedetta Morini. A globally convergent newton-gmres subspace method for systems of nonlinear equations. *SIAM J. Sci. Comput.*, 23(3):940–960, 2001.

[Bre74] Franco Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. 1974.

[BS89] Peter N. Brown and Youcef Saad. Globally convergent techniques in nonlinear newton-krylov algorithms. *Technical Report L-316, Computing and Mathematics Research Division, Lawrence Livermore National Lab*, 1989.

[DES82] Ron S. Dembo, Stanley C. Eisenstat, and Trond Steihaug. Inexact newton methods. *SIAM Journal on Numerical Analysis*, 19(2):400–408, 1982.

[Dob06] Manfred Dobrowolski. *Applied functional analysis. Functional analysis, Sobolev spaces and elliptic differential equations. (Angewandte Funktionalanalysis. Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen.).* Berlin: Springer. xii, 266 p. , 2006.

[DS83] John E.jun. Dennis and Robert B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations.* Prentice-Hall Series in Computational Mathematics. Englewood Cliffs, New Jersey: Prentice-Hall, Inc. XIII, 378 p., 1983.

[EW94] Stanley C. Eisenstat and Homer F. Walker. Globally convergent inexact newton methods. *SIAM Journal on Optimization*, 4(2):393–422, 1994.

[EW96] Stanley C. Eisenstat and Homer F. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM Journal on Scientific Computing*, 17(1):16–32, 1996.

[GR86] Vivette Girault and Pierre-Arnaud Raviart. *Finite element methods for Navier-Stokes equations. Theory and algorithms. (Extended version of the 1979 publ.).* Springer Series in Computational Mathematics, 5. Berlin etc.: Springer- Verlag. X, 374 p. , 1986.

[Hei08] Timo J. Heister. Vorkonditionierungsstrategien für das stabilisierte Oseen-Problem - Theorie und Anwendung. Diplomarbeit, Georg-August-Universität Göttingen, 2008.

[Heu03] Harro Heuser. *Textbook of analysis. Part 1. (Lehrbuch der Analysis. Teil 1.) 15., durchgesehene Aufl.* Mathematische Leitfäden. Stuttgart: Teubner. 643 S. , 2003.

[KLW06] Jisheng Kou, Yitian Li, and Xiuhua Wang. On modified Newton methods with cubic convergence. *Appl. Math. Comput.*, 176(1):123–127, 2006.

[Kon02] Konrad Königsberger. *Analysis 2. 4. überarb. Aufl.* Springer-Lehrbuch. Berlin: Springer. xii, 459 S. , 2002.

[Lub06a] Gert Lube. *Skript zur Vorlesung "Lineare Funktionalanalysis und Anwendung auf partielle DGL".* 2005/2006.

[Lub06b] Gert Lube. *Skript zur Vorlesung "Theorie und Numerik elliptischer Randwertprobleme".* 2006.

[OR70] James M. Ortega and Werner C. Rheinboldt. *Iterative Solutions of Nonlinear Equations in Several Varibles.* Computer Science and Applied Mathematics. Academic Press, New York, 1970.

[OR04] Maxim A. Olshanskii and Arnold Reusken. Grad-div stablilization for Stokes equations. *Math. Comput.*, 73(248):1699–1718, 2004.

[Pla00] Robert Plato. *Numerical mathematics compact. (Numerische Mathematik kompakt. Grundlagenwissen für Studium und Praxis.).* Braunschweig: Vieweg. xiv, 360 S. , 2000.

[PSSW06] Roger P. Pawlowski, John N. Shadid, Joseph P. Simonis, and Homer F. Walker. Globalization techniques for newton-krylov methods and applications to the fully coupled solution of the navier-stokes equations. *SIAM Rev.*, 48(4):700–721, 2006.

[PSWS05] Roger P. Pawlowski, Joseph P. Simonis, Homer F. Walker, and John N. Shadid. Inexact newton dogleg methods. *WPI Math. Sciences Dept. Tech. Rep. MS-5-05-36*, 2005.

[PT02] Dingguo Pu and Weiwen Tian. Globally convergent inexact generalized Newton's methods for nonsmooth equations. *J. Comput. Appl. Math.*, 138(1):37–49, 2002.

[PW98] Michael Pernice and Homer F. Walker. NITSOL: A Newton iterative solver for nonlinear systems. *SIAM J. Sci. Comput.*, 19(1):302–318, 1998.

[Saa96] Yousef Saad. *Iterative methods for sparse linear systems.* The PWS Series in Computer Science. Boston, MA: PWS Publishing Company. xvi, 447 p. , 1996.

[Sch86] Hans-Rudolf Schwarz. *Numerische Mathematik.* Stuttgart: B. G. Teubner. 496 S. , 1986.

[SDLS06] Yong Hun Lee Sang Dong Lee and Byeong Chun Shin. Newton's Method for the Navier-Stokes Equations with Finite-Element Initial Guess of Stokes Equations. *Computers and Mathematics with Applications*, 51:805–816, 2006.

[SK04] Hans-Rudolf Schwarz and Norbert Kockler. *Numerical mathematics. (Numerische Mathematik.) 5th ed.* Stuttgart: Teubner. 573 p. , 2004.

[STW97] John N. Shadid, Ray S. Tuminaro, and Homer F. Walker. An inexact newton method for fully coupled solution of the navier-stokes equations with heat and mass transport. *J. Comput. Phys.*, 137(1):155–185, 1997.

[TWS02] Raymond S. Tuminaro, Homer F. Walker, and John N. Shadid. On backtracking failure in newton-gmres methods with a demonstration for the navier-stokes equations. *J. Comput. Phys.*, 180(2):549–558, 2002.

[Wie05] Karl Wieghardt. *Theoretische Strömungslehre. Nachdruck der 2., überarbeiteten und erweiterten Auflage 1974.* Göttinger Universitätsverlag. Herausgeber der Reihe "Göttinger Klassiker der Strömungsmechanik": Prof. Dr. rer.nat. Dr.-Ing.habil. Andreas Dillmann (Georg-August-Universität Göttingen und Deutsches Zentrum für Luft- und Raumfahrt) , 2005.

Danksagung

An dieser Stelle möchte ich Professor Dr. Gert Lube danken, der es mir ermöglicht hat, diese Arbeit zu schreiben.

Großer Dank gebührt Dr. Gerd Rapin, der mit seiner Unterstützung und Motivation viel zur Entstehung dieser Arbeit beigetragen hat. Er hat sich immer die Zeit genommen, um Probleme mit mir zu diskutieren.

Für ihre Hilfe bei der Programmierung der hier verwendeten Programme danke ich Johannes Löwe und Timo Heister.

Meinen Eltern danke ich für ihre Unterstützung, die mein Studium erst möglich gemacht hat.

Abschließend möchte ich allen danken, die ich nicht genannt habe und die zur Erstellung dieser Arbeit beigetragen haben.