Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

FELIX LEHMANN

Inexaktheit in Newton-Lagrange-Verfahren für Optimierungsprobleme mit Elliptischen PDGL-Nebenbedingungen

ZIB-Report 13-37 (August 2013)

Herausgegeben vom
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Takustraße 7
D-14195 Berlin-Dahlem

Telefon: 030-84185-0
Telefax: 030-84185-125

e-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Diplomarbeit

Inexaktheit in Newton-Lagrange-Verfahren für Optimierungsprobleme mit Elliptischen PDGL-Nebenbedingungen

Felix Lehmann

31. Januar 2013

Betreuer: Dr. Martin Weiser

Zweitkorrektor: Prof. Dr. Ralf Kornhuber

Institut für Mathematik

Freie Universität Berlin

Zusammenfassung

Bei der numerischen Lösung von Optimalsteuerungsproblemen mit elliptischen partiellen Differentialgleichungen als Nebenbedingung treten unvermeidlich Diskretisierungs- und Iterationsfehler auf. Man ist aus Aufwandsgründen daran interessiert die dabei entstehenden Fehler nicht sehr klein wählen zu müssen. In der Folge werden die linearisierten Nebenbedingungen in einem Composite-Step-Verfahren nicht exakt erfüllt. In dieser Arbeit wird der Einfluss dieser Ungenauigkeit auf das Konvergenzverhalten von Newton-Lagrange-Verfahren untersucht. Dabei sollen mehrere einschlägige lokale Konvergenzresultate diskutiert werden. Anschließend wird ein konkretes Composite-Step-Verfahren formuliert, in dem die Genauigkeit der inneren Iterationsverfahren adaptiv gesteuert werden kann. Am Ende der Arbeit wird an zwei Musterproblemen die hohe Übereinstimmung der analytischen Voraussagen und der tatsächlichen Perfomanz der dargestellten Methoden demonstriert.

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation

In zahlreichen naturwissenschaftlichen Anwendungen müssen kontinuierliche Optimierungsaufgaben gelöst werden. Bei Optimalsteuerungsproblemen etwa muss ein Kostenfunktional unter vorgegebenen Nebenbedingungen minimiert werden, welche zum Beispiel physikalische Gesetzmäßigkeiten eines Systems widerspiegeln. Eine praxisnahe Problemstellung ist die therapeutische Hyperthermieplanung. Dabei werden tiefsitzende Tumoren durch elektromagnetische Strahlung erwärmt, um sie zu zerstören oder empfänglicher für weitere Therapien zu machen. Es wird versucht, krankhaftes Gewebe einer Zieltemperatur von 43°C und höher auszusetzen, gleichzeitig jedoch gesunde Zellen in der Umgebung zu schonen und deren Temperatur auf ein nicht schädliches Maß zu beschränken. Vereinfacht lässt sich diese Aufgabe durch die Optimierung eines sogenannten Tracking-Type-Funktionals mit PDGL-Nebenbedingungen modellieren (vgl. [9], [33]). Es bezeichnen Ω_0 das Tumorgebiet und $\Omega \supset \Omega_0$ das Gebiet in der näheren Umgebung sowie y die lokale Temperatur und u die Strahlungsintensität. Dann stellt sich das Optimierungsproblem

$$\min_{y,u} J(y,u) = \int_\Omega (y - y_d)^2 dt + \tfrac{\sigma}{2} \int_\Omega u^2 dt, \qquad \text{wobei } y_d = \begin{cases} 37 & \text{in } \Omega \backslash \Omega_0 \\ 43 & \text{in } \Omega_0 \end{cases}$$

mit dem Regularisierungsterm $\frac{\sigma}{2} \int_\Omega u^2 dt$. Die Nebenbedingung wird durch die Bio-Wärmeaustauschgleichung von Pennes [25] formuliert:

$$-\nabla \cdot (\kappa \nabla y) + f(y) = g(u).$$

Dabei modelliert $f(y)$ den Einfluss der Blutperfusion des Gewebes auf die Temperaturverteilung sowie $g(u)$ die lokale Energieaufnahme abhängig

von der Strahlungsintensität.

Die Analysis und Algorithmen für konvexe Probleme sind dabei gut erforscht, sobald die auftretenden Funktionale jedoch nichtkonvex werden, ist die Theorie ungleich schwieriger und das Vorliegen nichtlinearer Nebenbedingungen erschwert das Problem weiter. Im vorliegenden Fall ist das Funktional J zwar konvex, im Allgemeinen muss dies für das entstehende Lagrange-Funktional aber nicht mehr gelten.

Für die Lösung derartiger Optimalsteuerungsaufgaben müssen die kontinuierlichen Probleme diskretisiert und mit iterativen numerischen Verfahren behandelt werden. Die resultierenden Systeme nehmen mit feiner werdendem Gitter schnell gewaltige Ausmaße an, so dass direkte Löser rasch an ihre Grenzen kommen würden. Man ist dann gezwungen, für jeden Schritt erneut iterative Verfahren einzusetzen. Es wird die Idee verfolgt, die sukzessiv zu lösenden gleichungsbeschränkten Teilprobleme auf leicht gestörten Unterräumen zu lösen. Die Konstruktion dieser Räume erlaubt eine deutlich günstigere Berechnung als es auf den exakten Räumen möglich wäre. In der vorliegenden Arbeit wird die beschriebene Vorgehensweise und deren Auswirkungen auf das Konvergenzverhalten des äußeren Verfahrens analytisch und numerisch untersucht.

1.2 Aufbau

Die Struktur dieser Arbeit soll im Folgenden kurz umrissen werden. Am Ende dieses Einleitungskapitels werden zunächst die theoretischen Grundlagen der unbeschränkten und der gleichungsbeschränkten Optimierung wiederholt. Im Kapitel 2 wird die Vorgehensweise bei Newton-Lagrange-Verfahren beschrieben. Es werden lokale und globale Konvergenzaussagen getroffen und Techniken zur Globalisierung vorgestellt, auch in Hinblick auf nichtkonvexe Probleme. Im darauf folgenden Kapitel wird der Einfluss von inexakten Lösungen des KKT-Systems auf den Tangentialschritt sowie den gesamten Newton-Schritt analysiert und entsprechende Konvergenzresultate präsentiert. Im Zentrum von Kapitel 4 steht die numerische Umsetzung des vorgeschlagenen Ansatzes. Dazu werden zunächst die notwendigen Werkzeuge wiederholt, insbesondere die Konzepte Finite Elemente, CG-Verfahren und Vorkonditionierer und anschließend eine konkrete Implementierung vorgestellt. Im letzten Kapitel 5 werden zwei nichtlineare Optimierungsprobleme gelöst, um die erarbeiteten Ergebnisse zu veranschaulichen.

1.3 Grundlagen der Optimierung

Unbeschränkte Optimierung

Für ausreichend glatte Optimierungsprobleme

$$\min_{x \in X} J(x)$$

ist das Newton-Verfahren für die Lösung von $J'(x) =: F(x) = 0$ von größter Bedeutung, ob seiner hervorragenden Konvergenzeigenschaften. Ein Großteil der existierenden Iterationsverfahren orientiert sich an dessen Grundidee. Zur Lösung einer Gleichung $F(x) = 0$ werden iterativ die Nullstellen x_{k+1} der Linearisierung $F(x_k) + F'(x_k)(x - x_k)$ von F an der Stelle x_k berechnet. Der resultierende Prozess schreibt sich dann

$$F'(x_k)\Delta x_k = -F(x_k), \quad x_{k+1} = x_k + \Delta x_k, \quad k = 1, 2, ... \tag{1.1}$$

Das Verfahren ist äquivalent zur iterativen Bestimmung der stationären Punkte der quadratischen Taylor-Entwicklung $T_J^2(x_k)$ von J an den Stellen x_k. In der Nähe einer Lösung x^* stimmen die beiden Funktionale sehr genau überein, was die hohe Konvergenzgeschwindigkeit erklärt.

Hier wird auch die große Bedeutung der durch den Newton-Schritt definierten Richtung deutlich. Der negative Gradient $-\nabla J$ ist zwar die Richtung des steilsten Abstiegs (in der euklidischen Norm), kann aber bereits für konvexe quadratische Probleme zu extrem langsamer Konvergenz führen. Wenn diese schlecht konditioniert sind, ist diese Richtung annähernd orthogonal zu jener, auf der das eigentliche Optimum liegt und es kommt zum sogenannten „Zick-Zacking“. Der Newton-Schritt Δx dagegen bildet gerade diese optimale Suchrichtung. Für quadratische konvexe Modelle erreicht man mit einem Schritt die exakte Lösung. Man kann den Schritt daher auch als Richtung des steilsten Abstiegs interpretieren, allerdings in der durch $J''(x)$ induzierten Metrik. Im nichtquadratischen aber strikt konvexen Fall wird durch den Newton-Schritt lokal Abstieg garantiert. Dann ist nämlich die Hesse-Matrix positiv definit und es folgt

$$\Delta x^T \nabla J(x) = -\Delta x^T J''(x) \Delta x \leq -\lambda_{min} \|\Delta x\|^2$$

für den kleinsten Eigenwert λ_{min} von $J''(x)$.

Ohne Konvexität geht diese wünschenswerte Eigenschaft leider verloren. Ist die Hesse-Matrix $J''(x)$ indefinit (aber invertierbar), dann ist das Modell nach unten unbeschränkt und der eindeutige stationäre Punkt ein hochdimensionaler Sattelpunkt. Der Newton-Schritt kann in diesem Fall eine Aufstiegsrichtung darstellen und es wäre von geringem Nutzen,

dieses teure System zu berechnen. Lokal bildet das quadratische Modell dennoch eine gute Approximation des Funktionals. Um nicht auf den minderwertigen Gradienten-Abstieg zurückgreifen zu müssen, kann daher eine Modifikation der Hesse-Matrix in Betracht gezogen werden, indem man eine regularisierende Matrix addiert. Eine beliebte Wahl ist ein Vielfaches der Einheitsmatrix. Offensichtlich kann Positiv-Definitheit erzwungen werden für $H - \theta I$ für beliebige $\theta < \lambda_{min}$.

Bei Problemen sehr großer Dimension kann es von Nutzen sein, den Newtonschritt nicht mehr exakt zu berechnen, sondern mittels eines (frühzeitig und adaptiv abzubrechenden) iterativen Verfahrens zu approximieren

$$F'(x_k)\delta x_k^{(i)} = F(x_k) + r_k^{(i)}.$$

Unter bestimmten Anforderungen an den Fehler $\|\delta x_k^{(i)} - \Delta x_k\|$ kann lokal quadratische Konvergenz erhalten werden (siehe Kapitel 3).

Global konvergiert das Newton-Verfahren leider nur für sehr schwach nichtlineare Probleme. Eine intuitive Globalisierung kann durch Dämpfung der Schritte erreicht werden,

$$x_{k+1} = x_k + \tau_k \Delta x_k,$$

mit dem Dämpfungsfaktor τ_k, welcher durch das eindimensionale Optimierungsproblem

$$\min_{\tau_k > 0} J(x_k + \tau_k \Delta x_k)$$

bestimmt werden kann. Für gleichmäßig konvexe Funktionale (Def. A.5) kann globale Konvergenz des beschriebenen Verfahrens bewiesen werden.

Das Verfahren kann auch für nicht konvexe Probleme genutzt werden. Δx sollte dann aber durch eine leicht zu bestimmende Abstiegsrichtung ersetzt werden, so dass mit einer geeigneten Dämpfungsstrategie monotoner Abstieg der Funktionswerte $J(x_k)$ gewährleistet werden kann. Dann kann Konvergenz auch für ausreichend reguläre ($J \in C^1$, J' Lipschitzstetig) aber nichtkonvexe Funktionale gezeigt werden, unter der Forderung des hinreichenden Abstiegs der Funktionswerte:

$$\min_{x \in X} J(x) \leq J(x_k) - c_k \|J'(x_k)\|^2, \qquad \text{für } c_k \geq c > 0$$

(siehe Proposition 2.2.2 in [28]). Einen anderen Ansatz verfolgen Trust-Region-Methoden, in denen ein Modell zweiter oder geringerer Ordnung minimiert wird, in einem adaptiv gewählten Bereich, in dem das Modell das Funktional ausreichend gut approximiert.

$$\begin{aligned} &m_{x_k}(\delta x_k) \approx J(x_k + \delta x_k), \quad \text{für } \|\delta x_k\| \text{ klein}, \\ &\min_{\delta x_k} m_{x_k}(\delta x_k), \quad \text{so dass } \|\delta x_k\| \leq \Delta. \end{aligned}$$

Die Qualität des Modells wird durch einen Vergleich der vorhergesagten und der tatsächlichen Reduktion des Funktionswertes bewertet und die Größe des Bereiches entsprechend angepasst.

$$\rho_k = \frac{J(x_k + \delta x_k) - J(x_k)}{m_{x_k}(0) - m_{x_k}(\delta x_k)},$$
$$\text{falls } \rho_k > c \text{ verkleinere } \Delta$$
$$\text{falls } \rho_k \leq c \text{ vergrößere } \Delta$$

Diese Verfahren zeigen ein robustes Konvergenzverhalten auch für stark nichtkonvexe Probleme, allerdings gilt zu bedenken, dass die Teilprobleme ungleichungsbeschränkt sind und eine approximative Lösung auch bei Modellen geringer Komplexität recht aufwendig werden kann.

Optimierung mit Gleichungsrestriktionen

In dieser Arbeit sollen Probleme behandelt werden, in denen ein Funktional nur noch auf einer Teilmenge des Definitionsbereichs minimiert werden soll. Häufig ist diese Menge implizit durch die Niveaumenge einer Funktion gegeben.

$$\min_{x \in N^0} J(x), \quad N^0 = \{x \in X | c(x) = 0\}. \tag{1.2}$$

Ein gängiger Lösungsansatz ist die Umformulierung in ein unbeschränktes Optimierungsproblem. Bei (affin-) linearen oder nur leicht nichtlinearen Restriktionen sichern manche Verfahren, dass, sobald ein zulässiger Punkt $x_0 \in N^0$ gefunden wurde, die folgenden Iterierten zulässig bleiben. Dies generell zu fordern, wäre jedoch meist sehr kostspielig und daher nicht unbedingt von Vorteil. Eine alternative Verfahrensklasse bilden die Penalty-Methoden. Dabei wird die Addition eines Strafterms benutzt, um die Verletzung der Nebenbedingung direkt in das Funktional zu integrieren. Wichtige Vertreter sind die exakten und die quadratischen Penalty-Verfahren

$$\min_{x \in X} J_E(x) = J(x) + \mu \|c(x)\|, \tag{1.3}$$
$$\min_{x \in X} J_Q(x) = J(x) + \mu \|c(x)\|^2. \tag{1.4}$$

Beide Funktionale sind auf N^0 mit J identisch. Für das exakte Penalty-Verfahren stimmt die (lokale) Lösung mit der eigentlichen Lösung für hinreichend große $\mu < \infty$ überein. Leider ist das entsprechende Funktional im Allgemeinen nicht mehr glatt und es müssten kompliziertere Methoden als für glatte Probleme eingesetzt werden. Im Gegensatz dazu

ist das Funktional in (1.4) differenzierbar. Allerdings stimmen hier die Optima von J_Q und Problem (1.2) im Allgmeinen nur noch im Grenzfall $\mu \to \infty$ überein, denn obwohl die Ableitung des Strafterms $\mu\|c(x)\|^2$ auf N^0 verschwindet, muss das nicht für J in der Nähe des restringierten Minimierers gelten. Mit wachsendem Parameter μ entstehen zudem immer schmalere Täler und zunehmend schlecht konditionierte Teilprobleme. Aus den genannten Gründen wird eher selten versucht, diese Funktionale direkt zu minimieren. Sie leisten jedoch gute Dienste als „merit-function“, mit deren Hilfe man die Qualität eines Schrittes bewerten kann. Bei der beschränkten Optimierung ist das Lagrange-Funktional

$$L(x, \lambda) = J(x) + \langle \lambda, c(x) \rangle$$

von großer Bedeutung. Es stellt sich heraus, das für optimale Punkte x^* in (1.2) eine Lagrange-Multiplikator λ^* existiert, so das die Gleichung

$$L'(x^*, \lambda^*) = 0$$

erfüllt ist. Anschaulich betrachtet kippt der zusätzliche Term $\langle \lambda, c(x) \rangle$ für den optimalen Lagrange-Multiplikator λ^* die Tangentialebene an J derartig, dass die Ableitung von $L(x, \lambda^*)$ bei x^* verschwindet (siehe Abbildung 1.1). Im folgenden Kapitel wird aufgezeigt, wie man sich diese Erkenntnis zu Nutze machen kann.

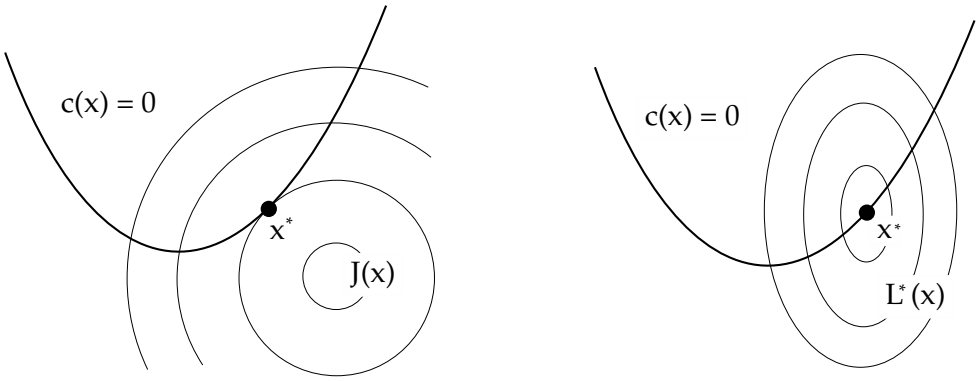

Abbildung 1.1: Niveaulinien von $J(x) = (x_1 - 1)^2 + (x_2 + 1)^2$ und der Lagrangefunktion $L^*(x) = L(x, \lambda^*) = J(x) + \lambda^* c(x)$ unter der Nebenbedingung $c(x) = x_1^2 - x_2$.

Kapitel 2

Exaktes Newton-Lagrange-Verfahren

Für das zu betrachtende Minimierungsproblem

$$\min_{y\in Y, u\in U} J(y,u), \quad \text{so dass } c(y,u)=0, \tag{2.1}$$

seien $J : X = Y \times U \to \mathbb{R}$ und $c : X \to W$ zweimal stetig Fréchet-differenzierbar auf den Hilberträumen Y, U und W. Dann gilt der folgende wohlbekannte Satz.

Satz 2.1. *Ist $x^* = (y^*, u^*) \in X$ eine Lösung von (2.1) mit surjektiver Ableitung $c'(x^*)$, dann existiert ein Lagrange-Multiplikator $\bar{\lambda} \in W^*$, für den die notwendigen KKT-Bedingungen*

$$J'(x^*) + c'(x^*)^*\bar{\lambda} = 0 \tag{2.2}$$
$$c(x^*) = 0 \tag{2.3}$$

erfüllt sind.

Unter Zuhilfenahme des Lagrange-Funktionals

$$L : Z = X \times W^* \to \mathbb{R}, \qquad L(x,\lambda) = J(x) + \langle \lambda, c(x)\rangle$$

mit

$$L'(x,\lambda) = \begin{pmatrix} J_y(x) + c_y(x)^*\lambda \\ J_u(x) + c_u(x)^*\lambda \\ c(x) \end{pmatrix}$$

lässt sich (2.2) und (2.3) kompakter formulieren durch

$$L'(x^*, \bar{\lambda}) = 0.$$

Als Kandidaten für die Minimierung von (2.1) kommen nur stationäre Punkte des Lagrange-Funktionals in Frage und daher gilt es, die Operatorgleichung

$$F(z) := L'(z) = 0,$$

mit $z := (x, \lambda) = (y, u, \lambda)$ zu lösen. Die Lösung dieser gemeinhin nichtlinearen Gleichung kann durch ein Newton-Verfahren berechnet werden, der resultierende Prozess wird als Newton-Lagrange-Verfahren bezeichnet. Es muss dann iterativ das folgende System gelöst werden.

$$\begin{pmatrix} L_{xx}(x,\lambda) & c'(x)^* \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} + \begin{pmatrix} J'(x) + c'(x)^*\lambda \\ c(x) \end{pmatrix} = L''(z)\Delta z + L'(z) = 0.$$

Setzt man voraus, dass L_{xx} positiv definit ist, so existiert ein eindeutiger Minimierer des linear-quadratischen Optimierungsproblems

$$\min_p J'(x)p + \tfrac{1}{2}L_{xx}(x,\lambda)(p,p), \quad \text{so dass } c'(x)p + c(x) = 0 \tag{2.4}$$

und dieser ist identisch mit der primalen Variable Δx im exakten Newton-Lagrange-Schritt. Iteriert man die Lösung des Problems (2.4) an Stelle der direkten Lösung des Newton-Systems, erhält man ein sogenanntes SQP-Verfahren (Sequential Quadratic Programming). Wohlgemerkt stellen bei konvexen Zielfunktionalen beide Ansätze zunächst nur verschiedene Interpretationen ein und desselben Problems dar (bis auf Berechnung der dualen Variable λ).

2.1 Lokale Konvergenzresultate

Die Popularität des Newton-Verfahren beruht stark auf der lokal guten Konvergenzgeschwindigkeit, wobei der Begriff *lokal* sich auf die Annahme bezieht, bereits einen guten Startwert nah an der Lösung parat zu haben. Der folgende Satz garantiert lokal quadratische Konvergenz des SQP-Verfahrens (2.4).

Satz 2.2 (Konvergenz des SQP-Verfahrens, Theorem 15.2.1 in [8]). *Es sei $x^* = (y^*, u^*)$ ein Minimierer für (2.1) mit dem entsprechenden Lagrange-Multiplikator λ^*. In einer Umgebung Q um (x^*, λ^*) seien J und c zweimal Frechét-differenzierbar, deren Ableitung Lipschitz-stetig und das entsprechende KKT-System invertierbar. Dann existiert eine Umgebung $Q^* \subset Q$ um (x^*, λ^*) so dass für jede gegen λ^* konvergente Folge $\{\lambda_k\}$ die durch das SQP-Verfahren (2.4) definierte Folge $\{x^k\}$ superlinear gegen x^* konvergiert für beliebige Startwerte in Q^*. Für $\|\lambda_k - \lambda^*\| = \mathcal{O}(\|x_k - x^*\|)$ ist die Konvergenz quadratisch.*

Die implizierte Wohldefiniertheit des SQP-Schritts (2.4) garantiert somit, dass in der Nähe eines Optimums $L_{xx}(x, \lambda)$ positiv definit auf dem Kern der linearisierten Nebenbedingungen ist. In diesem Fall sind SQP und Newton-Lagrange äquivalent und folglich von gleicher Konvergenzgeschwindigkeit.

Die folgenden Ausführungen bieten Informationen über die Größe der Bereiche, in denen das Newton-Verfahren quadratisch konvergiert. Die Argumentation orientiert sich an der Monografie von Peter Deuflhard [13].

In den vierziger Jahren des letzten Jahrhunderts wurde durch den Russen Leonid Kantorovich quadratische Konvergenz erstmals in allgemeinen Banachräumen gezeigt. Dabei wies er auch Existenz und Eindeutigkeit der Lösung unter minimalen Forderungen an Startwerte und Funktional nach. Zunächst wird eine ähnliche Aussage unter stärkeren Forderungen diskutiert. Der Newton-Algorithmus (1.1) legt nahe, die beschränkte Invertierbarkeit von $F'(z)$ zu fordern, das heißt

$$\|F'(z)^{-1}\| \leq \beta < \infty. \tag{2.5}$$

Weiterhin sind bestimmte Informationen zweiter Ordnung an F nötig. Eine sehr einfache Möglichkeit ist die Beschränkung der zweiten Ableitung,

$$\|F''(z)\| \leq \gamma < \infty. \tag{2.6}$$

Diese restriktiven Forderungen an die Glattheit von F garantieren nun Existenz einer Lösung z^* von $F(z) = 0$ und lokal quadratische Konvergenz der Newton-Iteration gegen diesen Punkt.

Satz 2.3 (Theorem 4.2.4 in Argyros [3]). *Sei $F : D \to \mathbb{R}$ zweimal Fréchet-differenzierbar und $D \subset Z$ offen und konvex. Für die abgeschlossene Kugel $B(z_0, r) \subset D$ um den Startpunkt z_0 gelte (2.5) und (2.6) für alle $z \in B(z_0, r)$. Die Größe $h_0 = \beta\gamma\|\Delta z_0\|$ erfülle $h_0 < \frac{1}{2}$ und definiere den Kugelradius*

$$r = \frac{1 - \sqrt{1 - 2h_0}}{\beta\gamma}. \tag{2.7}$$

Dann existiert eine Lösung $z^* \in B(z_0, r)$, die Newton-Iteration $\{z_k\}$ verbleibt in $B(z_0, r)$ und konvergiert quadratisch gegen z^*.

Beweis. Da aus (2.6) die Lipschitz-Stetigkeit von F' folgt, ist der Satz ein Spezialfall von Theorem 2.1 in Deuflhard [13]. □

Es stellt sich heraus, dass die zweimalige Differenzierbarkeit von F nicht gebraucht wird. Die notwendigen Informationen zweiter Ordnung lassen sich bereits aus einer Lipschitz-Bedingung an die Ableitung

$$\|F'(z) - F'(w)\| \leq \gamma\|z - w\| \tag{2.8}$$

entnehmen, so dass der obere Satz für nur stetig Fréchet-differenzierbare Funktionen gültig bleibt, wenn man (2.6) durch (2.8) ersetzt. Diese Bedingung erlaubt noch eine Vereinfachung der Voraussetzungen. Die punktweise Schranke

$$\|F'(z_0)^{-1}\| \leq \beta_0 < \infty \tag{2.9}$$

liefert unter der Verwendung der oben genannten Lipschitzbedingung und mit Hilfe des Störungslemmas für Operatoren (A.1) bereits eine lokale Schranke für $\|F'(z)^{-1}\|$ abhängig von β_0, γ und $\|z - z_0\|$.

Für die Entwicklung effizienter Algorithmen ist das Konzept der affinen Invarianz von großer Bedeutung. Endlichdimensionale Funktionen $F : \mathbb{R}^n \to \mathbb{R}^n$ können beliebig schlechte Skalierungseigenschaften haben. Betrachtet man die affin transformierte Funktion $G_A(z) := AF(z)$ für eine nichtsinguläre Matrix A, produziert das Newton-Verfahren für F und G_A identische Iterierte, denn

$$\Delta z_G = -G'_A(z)^{-1} G_A(z) = -F'(z)^{-1} A^{-1} AF(z) = -F'(z)^{-1} F(z) = \Delta z_F.$$

Der Konvergenzradius r in klassischen Newton-Konvergenz-Theoremen ist ungefähr proportional zu $(\gamma\beta)^{-1}$ bzw. $(\gamma\beta_0)^{-1}$. Die darin vorkommenden Schranken können abhängig von A beliebig groß werden, so dass die Kugeln garantierter quadratischer Konvergenz fast verschwinden. Um dieses Problem zu beheben, kann man die Bedingungen (2.8) und (2.9) in einer affin-invarianten Lipschitzbedingung koppeln. Für den affin-kovarianten Fall (affin-invariant unter Transformation im Bildraum) kann eine Bedingung der Art

$$\|F'(z_0)^{-1}(F'(z) - F'(w))\| \leq \omega_0 \|z - w\|, \qquad \text{für } z_0, z, w \in D \tag{2.10}$$

formuliert werden, welche die gleichen Resultate wie in Satz 2.3 liefert, wobei im Konvergenzradius r und der Kantorovich-Größe h_0 das Produkt $\beta\gamma$ durch ω_0 ersetzt wird. Offensichtlich ist ω_0 invariant unter der Transformation A. Da ω_0 oft deutlich kleiner als $\beta_0\gamma$ bzw. $\beta\gamma$ ist erhält man größere Konvergenzradien. Leider kommt in (2.10) noch eine Operatornorm vor. Diese ist numerisch schlecht handhabbar und hängt von den Normen im Urbild- und im Bildraum ab und kann bei schlechter Wahl derer stark anwachsen. Es ist aber möglich, ein sogenanntes Newton-Mysovskikh-Theorem mit einer affin-kovarianten Lipschitzbedingung folgender Art zu versehen:

$$\|F'(z)^{-1}(F'(w) - F'(z))(w - z)\| \leq \omega \|w - z\|^2, \qquad \text{für } z, w \in D.$$

Offenbar werden hier nur noch Vektornormen benutzt. Folgendes affin-kovariantes Newton-Mysovskikh-Theorem aus dem Artikel [34] vereint die bisherigen Ausführungen für allgemeine Banachräume.

Satz 2.4. *Es sei $F : D \to \mathbb{R}$ stetig Fréchet-differenzierbar auf einem konvexen Gebiet $D \subset Z$. $F'(z)$ sei für alle $z \in D$ invertierbar. Weiterhin gelte*

$$\|F'(z)^{-1}(F'(w) - F'(y))v\| \leq \omega\|w - y\|\|v\|$$

für kolineare $z, w, y \in D$, versehen mit der Norm $\|\cdot\|$ des Urbildraums Z. Für $z_0 \in D$ sei

$$h_0 = \omega\|\Delta z_0\| < 2$$

und die abgeschlossene Kugel $B(z_0, r) \subset D$ für $r = \frac{\Delta z_0}{1-h_0/2}$. Dann ist die durch das Newton-Verfahren definierte Folge $\{z_k\}$ wohldefiniert, verbleibt in $B(z_0, r)$ und konvergiert gegen die eindeutige Lösung $z^* \in B(z_0, r)$. Die Konvergenzgeschwindigkeit kann abgeschätzt werden gemäß

$$\|z_{k+1} - z_k\| \leq \tfrac{1}{2}\omega\|z_k - z_{k-1}\|^2.$$

2.2 Globalisierung mit Composite-Step-Verfahren

Der in (2.4) definierte SQP-Algorithmus setzt voraus, dass das betrachtete Funktional auf dem Kern der linearisierten Nebenbedingungen $\ker c'(x)$ konvex ist, was in der Nähe des Minimierers eine sinnvolle Annahme ist. Im Allgemeinen, weit weg von der Lösung, muss diese Bedingung natürlich nicht mehr erfüllt sein. In diesem Fall ist die Hesse-Matrix des Funktionals nicht mehr positiv definit auf $\ker c'(x)$ und das entsprechende Teilproblem nach unten unbeschränkt. In diesem und dem folgenden Abschnitt wird beschrieben, wie man diese Schwierigkeit beheben kann, um ein global wohldefiniertes Verfahren zu erhalten.

Eine Möglichkeit zur Globalisierung besteht darin, die Aspekte Optimierung und Zulässigkeit (Einhaltung der Nebenbedingung) zu trennen. Bei sogenannten Composite-Step-Verfahren zerlegt man daher den Newton-Schritt Δx in einen Normalschritt δn und einen Tangentialschritt δt. Der Normalschritt dient dabei der Verbesserung der Zulässigkeit in dem Sinne, dass $\|c(x)\| > \|c(x + \delta n)\| \approx 0$. Bei der Bezeichnung „normal“ denke man an Orthogonalität bezüglich der Niveaumengen $\{x \in X | c(x) = 0\}$. Um Ressourcen zu sparen, soll das entsprechende Teilproblem nur für die Näherung erster Ordnung umgesetzt werden, d.h. δn orthogonal zu $\{v \in X | c(x_k) + c'(x_k)v = 0\}$.
Für die Optimierung betrachte man die Taylor-Entwicklung der Nebenbedingung,

$$c(x + \delta t) = c(x) + c'(x)\delta t + \mathcal{O}(\|\delta t\|^2)\,.$$

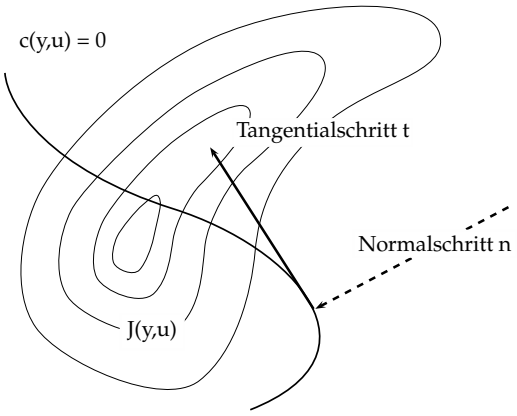

Abbildung 2.1: Normal- und Tangentialschritt

Der Tangentialschritt sollte daher aus dem Kern von $c'(x)$ gewählt werden, so dass $c(x+\delta t) - c(x) \in \mathcal{O}(\|\delta t\|^2)$. Im Funktional dagegen bleibt eine Reduktion des Funktionswertes in der Größenordnung $\mathcal{O}(\|\delta t\|)$ zu erwarten. Für ausreichend kleine Tangentialschritte optimiert man daher im Wesentlichen das Problem, ändert aber wenig an der Nebenbedingung.

Für die Berechnung des Tangentialschritts kann man sich an dem eingangs beschriebenen gewöhnlichen Newton-Lagrange-Verfahren orientieren. Motiviert durch die Lösung von

$$\min_{\delta t} L(x+\delta t, \lambda), \quad \text{so dass } c'(x)\delta t = 0\,,$$

betrachtet man das Modellproblem

$$\min_{\delta t} m(\delta t)\,, \quad \text{so dass } c'(x)\delta t = 0\,, \tag{2.11}$$

mit der quadratischen Approximation

$$\begin{aligned} m(\delta t) &= L(x,\lambda) + L_x(x,\lambda)\delta t + \tfrac{1}{2} L_{xx}(x,\lambda)(\delta t, \delta t) \\ &\approx L(x+\delta t, \lambda)\,. \end{aligned} \tag{2.12}$$

Bei dieser Vorgehensweise können Schwierigkeiten auftreten. Zum Einen ist $m(\delta t)$ nur lokal eine gute Approximation für $L(x+\delta t, \lambda)$. Gegebenenfalls muss daher eine Dämpfung des Tangentialschritts erwogen werden. Schlimmer noch kann das Lagrangefunktional bei x nichtkonvex auf dem Kern der Nebenbedingung sein. In diesem Fall ist (2.11) nicht mehr wohldefiniert. In Abschnitt 2.3 wird eine Methode beschrieben, mit der beide Probleme bewältigt werden können.

Merit-Funktion

Um zu entscheiden, ob die berechneten Schritte in einem iterativen Verfahren akzeptiert oder gegebenenfalls gedämpft werden sollten, muss deren Qualität bewertet werden. Bei der unbeschränkten Optimierung stellt der Wert des Funktionals bereits ein gutes alleinstehendes Kriterium dar. In der beschränkten Optimierung konkurrieren jedoch Abstieg des Funktionals und Einhaltung der Zulässigkeit. In diesem Fall kann eine Merit-Funktion zur Bewertung herangezogen werden, in der beide Aspekte berücksichtigt werden. In dieser Arbeit wird die Merit-Funktion

$$\Phi(x, \mu) = J(x) + \mu \|c(x)\|\,, \tag{2.13}$$

mit dem Penalty-Parameter $\mu > 0$ benutzt. Der Einsatz dieser nicht differenzierbaren Funktion wird besonders dadurch gerechtfertigt, dass für hinreichend große $\mu \geq \mu_0$ der beschränkte Minimierer und der unbeschränkte Minimierer von (2.13) übereinstimmen, was auch die Bezeichnung „exakt“ rechtfertigt (siehe Einleitung). Möglichkeiten zur adaptiven Bestimmung des Parameters μ werden in den Büchern [22] und [8] dargestellt.

Algorithmisches Konzept

Die vorangegangenen Abschnitte motivieren den folgenden Prototypen eines Composite-Step-Algorithmus. Eine detaillierte Umsetzung folgt in Kapitel 4.

Algorithmus 1 Composite-Step Algorithmus (CS)

$(x, \lambda) = CS(x_0, \lambda_0, \mu_0)$
for $k = 0, \ldots, k_{\max}$ **do**
if Abbruchkriterium positiv **then**
Stop.
else
Berechne Normalschritt δn
Minimiere approximativ $\|c(x_k + \delta n)\|$
Berechne Tangentialschritt δt
Minimiere approximativ Modell $m(\delta t)$
$\min_{\delta t} m(\delta t) \approx L(x_k + \delta n + \delta t, \lambda_k)$
so dass $c'(x_k)\delta t = 0$
Finde Dämpfungsfaktor $\tau > 0$
so dass $L(x_k + \delta n + \tau\delta t, \lambda_k) < L(x_k + \delta n, \lambda_k)$
$\delta x = \delta n + \tau\delta t$
Finde Dämpfungsfaktor $\alpha > 0$
so dass $\Phi(x_k + \alpha\delta x, \mu_k) < \Phi(x_k, \mu_k)$
$x_{k+1} = x_k + \alpha\delta x$
Aktualisiere λ_{k+1} und μ_{k+1}
end if
end for

Es wird darauf hingewiesen, dass die vorgestellte Herangehensweise nicht ausreicht, um globale Konvergenz zu sichern. Zum Einen genügt es nicht, nur Abstieg in Funktional und Nebenbedingung zu fordern, dieser muss auch hinreichend stark ausfallen, da der Algorithmus sonst fern eines stationären Punktes zum Stehen kommen kann. Zum Anderen muss eine Strategie entworfen werden dem sogenannten Maratos-Effekt vorzubeugen. Damit wird der Effekt bezeichnet, dass die Merit-Funktion manchmal Schritte abweist, welche eigentlich guten Fortschritt in Richtung Optimum bedeuten, aber lokal keinen Abstieg in Funktional oder Nebenbedingung bringen. In [22] werden effektive Gegenmaßnahmen vorgestellt die in dieser Arbeit jedoch nicht thematisiert werden sollen.

Im nächsten Abschnitt wird ein Werkzeug entwickelt, mit dem der

Dämpfungsfaktor τ für den Tangentialschritt δt in Algorithmus 1 bestimmt werden kann.

2.3 Kubische Regularisierung

Bei globalisierten Optimierungsverfahren kommt der Dämpfung der berechneten Schritte eine gehobene Bedeutung zu. In Trust-Region-Methoden wird eine quadratische Näherung des Funktionals auf einem bereits beschränkten Bereich minimiert und man erhält ein ungleichungsbeschränktes Problem. Mit der Idee der kubischen Regularisierung, welche auf Griewank [15] zurückgeht, kann diese zusätzliche Komplexität vermieden werden. Dabei versucht man das kubisches Modell

$$\psi_{x,\omega}(\delta x) = J(x) + J'(x)\delta x + \frac{1}{2}J''(x)(\delta x, \delta x) + \frac{\omega}{6}\|\delta x\|_M^3 \tag{2.14}$$

zu optimieren. Man erhält ein koerzitives Minimierungsproblem ohne Nebenbedingungen. Diese Methode wird besonders dadurch motiviert, dass das Modell $\psi_{x,\omega}(\delta x)$ für korrekte Wahl des Parameters ω (und alle größeren) eine obere Schranke des Funktionals darstellt. Besitzt J eine Lipschitz-stetige zweite Ableitung

$$\|J''(x) - J''(y)\| \leq \omega\|x - y\|,$$

erfüllt die Lipschitz-Konstante diese Forderung. Wie sich herausstellt, genügt für den Beweis eine schwächere und gleichzeitig leichter zugängliche affin-invariante Lipschitz-Bedingung.

Lemma 2.5 (Lemma 3.2 in [32]). *Sei $J : D \to Y$ zweimal Fréchet-differenzierbar auf einer offenen Teilmenge D des Hilbertraums X. Der Operator $M \in \mathcal{L}(X, X^*)$ sei symmetrisch und positiv-definit und definiere die Normen $\|x\|_M^2 = (x, Mx)$ auf X und $\|y\|_{M^{-1}}^2 = (y, M^{-1}y)$ auf dem Dualraum. Es sei die Lipschitz-Bedingung*

$$\|(J''(x + dx) - J''(x))\delta x\|_{M^{-1}} \leq \omega\|\delta x\|_M^2 \tag{2.15}$$

erfüllt. Dann bildet das Modell (2.14) *eine echte obere Schranke für J.*

Beweis. Mit der Taylor-Formel A.7 erhält man die Abschätzung

$$\begin{aligned}
&|J(x+\delta x) - J(x) - J'(x)\delta x - \tfrac{1}{2}J''(x)(\delta x,\delta x)| \\
&\qquad = \left|\int_0^1 [J'(x+s\delta x) - J'(x)]\delta x - sJ''(x)(\delta x,\delta x)ds\right| \\
&\qquad = \left|\int_0^1\int_0^1 J''(x+ts\delta x)(\delta x, s\delta x)dt - sJ''(x)(\delta x,\delta x)ds\right| \\
&\qquad = \left|\int_0^1\int_0^1 s[J''(x+ts\delta x) - J''(x)](\delta x,\delta x)dt\,ds\right| \\
&\qquad \le \int_0^1\int_0^1 s\big|[J''(x+ts\delta x) - J''(x)](\delta x,\delta x)\big|dt\,ds \\
&\qquad \le \int_0^1\int_0^1 s\big\|[J''(x+ts\delta x) - J''(x)]\delta x\big\|_{M^{-1}}\|\delta x\|_M dt\,ds \\
&\qquad \le \int_0^1\int_0^1 \tfrac{s}{ts}\big\|[J''(x+ts\delta x) - J''(x)]ts\delta x\big\|_{M^{-1}}\|\delta x\|_M dt\,ds \\
&\qquad \le \int_0^1\int_0^1 ts^2 dt\,ds\;\omega\|\delta x\|_M^3 \\
&\qquad \le \tfrac{\omega}{6}\|\delta x\|_M^3.
\end{aligned}$$

□

In der Folge liefert ein Schritt, der den Wert des kubischen Modells reduziert, auch Abstieg im eigentlichen Funktional. Im Allgemeinen ist es sehr schwierig, den Parameter ω exakt zu bestimmen und man ist daher auf Schätzungen angewiesen. Aus der Ungleichung in obigem Beweis gilt

$$\omega_{\text{global}} = \sup_{\delta x} \tfrac{6}{\|\delta x\|_M^3}\big|J(x+\delta x) - J(x) - J'(x)\delta x - \tfrac{1}{2}J''(x)(\delta x,\delta x)\big|.$$

Durch Einsetzen eines bestimmten Vektors $\underline{\delta x}$, erhält man die lokale Schätzung

$$[\omega]_3 := \tfrac{6}{\|\underline{\delta x}\|_M^3}\big|J(x+\underline{\delta x}) - J(x) - J'(x)\underline{\delta x} - \tfrac{1}{2}J''(x)(\underline{\delta x},\underline{\delta x})\big|.$$

Offensichtlich gilt für den Schätzer $[\omega]_3 \le \omega$. Eine weitere Möglichkeit ist

$$[\omega]_2 := \tfrac{2}{\|\underline{\delta x}\|_M^3}\big|J'(x+\underline{\delta x})\underline{\delta x} - J'(x)\underline{\delta x} - J''(x)(\underline{\delta x},\underline{\delta x})\big|.$$

In Weiser [32] wird gezeigt, dass auch hier $[\omega]_2 \le \omega$ gilt. Im selben Artikel werden auch Auswirkungen und Strategien diskutiert, falls $[\omega]$ deutlich zu klein ist.

Kubische Regularisierung für die Tangentialschrittberechnung

Die Berechnung globaler Minima des kubischen Modells wäre ein sehr schwieriges Unterfangen, denn $\psi_{x,\omega}$ besitzt bis zu $2m+1$ stationäre Punkte, wobei m die Anzahl der negativen Eigenwerte von $J''(x)$ beziffert ([15]). Das Verfahren soll daher vornehmlich der Schrittweitenbestimmung einer vorgegebenen (möglicherweise inexakten) Newton-Suchrichtung dienen. Nur wenn bei deren Ermittlung Nichtkonvexität auftaucht, soll das kubische Modell auf einem zweidimensionalen Unterraum minimiert werden. Der Aufwand für dieses Problem ist dann vernachlässigbar. In Abschnitt 4.3.2 wird erläutert, wie dieser Raum definiert wird.

Zunächst soll aber im nächsten Kapitel die Auswirkung inexakter, ungedämpfer Tangentialschritte bei der Optimierung konvexer Probleme untersucht werden.

Kapitel 3

Inexaktes Newton-Lagrange-Verfahren

3.1 Der Inexakte Tangentialschritt

Im vorangegangenen Kapitel wurde die Notwendigkeit für die Berechnung des Tangentialschritts in einem Composite-Step-Verfahren begründet. Es wird nun davon ausgegangen, dass dieser Schritt $\delta t = (y, u)$ Lösung eines konvexen Minimierungsproblems

$$\min_{y \in Y, u \in U} J(y,u) = \langle \tfrac{1}{2} H_{yy} y - b_y, y \rangle + \langle \tfrac{1}{2} H_{uu} u - b_u, u \rangle$$

unter linearen, homogenen Nebenbedingungen

$$Ay + Bu = 0 \tag{3.1}$$

darstellt. Dabei seien Y und U Banachräume, A ein symmetrischer und beschränkt invertierbarer Operator, $B : U \to Y^*$ und $B^* : Y^* \to U$ beschränkt und stetig und $H_{yy} : Y \to Y^*$, $H_{uu} : U \to U^*$ beschränkt, symmetrisch und auf der Lösungsmenge von (3.1) positiv definit. Der Zustand y kann dann eliminiert werden via $y = -A^{-1}Bu$ und man erhält das reduzierte Funktional

$$\begin{aligned} J_R(u) &= \langle \tfrac{1}{2} H_{yy}(-A^{-1}Bu) - b_y, -A^{-1}Bu \rangle + \langle \tfrac{1}{2} H_{uu} u - b_u, u \rangle \\ &= \langle \tfrac{1}{2} B^* A^{-1} H_{yy} A^{-1} Bu + B^* A^{-1} b_y, u \rangle + \langle \tfrac{1}{2} H_{uu} u - b_u, u \rangle \\ &= \langle \tfrac{1}{2} (\underbrace{B^* A^{-1} H_{yy} A^{-1} B + H_{uu}}_{=: H_R}) u - (\underbrace{b_u - B^* A^{-1} b_y}_{=: b}), u \rangle \, . \end{aligned}$$

Der reduzierte Operator H_R ist nun positiv definit, deshalb gilt für die optimale Steuerung und den zugehörigen Zustand

$$u = H_R^{-1} b \, , \quad y = -A^{-1} B u.$$

Bei den Problemen, die hier von Interesse sein sollen, stellt A einen elliptischen Differentialoperator dar. Die Existenz effektiver, spektraläquvialenter und symmetrischer Mehrgittervorkonditionierer für das diskretisierte Problem motiviert die Untersuchung des inexakten Tangentialschritts

$$\tilde{u} = \widetilde{H}_R^{-1}\tilde{b}\,, \quad \tilde{y} = -\widetilde{A}^{-1}B\tilde{u},$$

definiert durch den gestörten Operator $\widetilde{A}^{-1}$ und die entsprechend definierten Größen $\widetilde{H}_R$ und $\tilde{b}$. Wie später noch gezeigt wird, erlaubt die i-fach iterierte Anwendung des besagten Vorkonditionierers, hier mit $\widetilde{A}_i^{-1}$ bezeichnet, die Abschätzung $\|I - \widetilde{A}_i^{-1}A\| \leq \rho^i$, für ein $\rho \leq 1/2$ (siehe Satz 4.2 und Gleichung (4.11), S.38). Das folgende Lemma zeigt, dass der Fehler des inexakten Tangentialschritts für $i \to \infty$ verschwindet.

Lemma 3.1. *Neben den oben genannten Voraussetzungen wird zusätzlich gefordert, dass H_R und $\widetilde{H}_R$ beschränkt invertierbar seien. Für eine Folge von Operatoren $\{\widetilde{A}_i\,, i \in \mathbb{N}\}$, gelte $\|I - \widetilde{A}_i^{-1}A\| \to 0$, für $i \to \infty$. Dann sind $\|\tilde{u}_i - u\|$ und $\|\tilde{y}_i - y\|$ ebenfalls Nullfolgen.*

Beweis. Der Übersichtlichkeit halber werden die Indizes der Operatorfolge weggelassen. Zunächst gilt

$$\|A^{-1} - \widetilde{A}^{-1}\| \leq \|I - \widetilde{A}^{-1}A\|\|A^{-1}\| \to 0,$$

da A^{-1} nach Voraussetzung beschränkt ist. Weiterhin gilt

$$\begin{aligned}\|\tilde{u} - u\| &= \|\widetilde{H}_R^{-1}\tilde{b} - H_R^{-1}b\| \\ &\leq \|\widetilde{H}_R^{-1}\tilde{b} - \widetilde{H}_R^{-1}b\| + \|\widetilde{H}_R^{-1}b - H_R^{-1}b\|.\end{aligned}$$

Die linke Seite lässt sich abschätzen gemäß

$$\begin{aligned}\|\widetilde{H}_R^{-1}\tilde{b} - \widetilde{H}_R^{-1}b\| &\leq \|\widetilde{H}_R^{-1}\|\|\tilde{b} - b\| \\ &\leq \underbrace{\|\widetilde{H}_R^{-1}\|}_{\text{beschränkt}}\ \underbrace{\|B^*\|}_{\text{beschränkt}}\ \underbrace{\|A^{-1} - \widetilde{A}^{-1}\|}_{\longrightarrow 0}\ \underbrace{\|b\|}_{\text{beschränkt}}\end{aligned}$$

und die rechte Seite

$$\|\widetilde{H}_R^{-1}b - H_R^{-1}b\| \leq \underbrace{\|H_R^{-1}\|}_{\text{beschränkt}}\ \|H_R - \widetilde{H}_R\|\ \underbrace{\|\widetilde{H}_R^{-1}\|}_{\text{beschränkt}}\ \underbrace{\|b\|}_{\text{beschränkt}}\,.$$

Für den mittleren Teil erhält man

$$\|H_R - \widetilde{H}_R\| \leq \underbrace{\|B^*\|}_{\text{beschränkt}}\ \|A^{-1}H_{yy}A^{-1} - \widetilde{A}^{-1}H_{yy}\widetilde{A}^{-1}\|\ \underbrace{\|B\|}_{\text{beschränkt}}$$

und schließlich nach einer weiteren Addition von Null

$$\begin{aligned}
&\|A^{-1}H_{yy}A^{-1} - \widetilde{A}^{-1}H_{yy}\widetilde{A}^{-1}\| \\
&\qquad = \|A^{-1}H_{yy}A^{-1} - A^{-1}H_{yy}\widetilde{A}^{-1} + A^{-1}H_{yy}\widetilde{A}^{-1} - \widetilde{A}^{-1}H_{yy}\widetilde{A}^{-1}\| \\
&\qquad = \|A^{-1}H_{yy}A^{-1}(I - A\widetilde{A}^{-1}) + (A^{-1}\widetilde{A} - I)\widetilde{A}^{-1}H_{yy}\widetilde{A}^{-1}\| \\
&\qquad \leq \underbrace{\|A^{-1}H_{yy}A^{-1}\|}_{\text{beschränkt}}\underbrace{\|I - A\widetilde{A}^{-1}\|}_{\longrightarrow 0} + \underbrace{\|A^{-1}\widetilde{A} - I\|}_{\longrightarrow 0}\underbrace{\|\widetilde{A}^{-1}H_{yy}\widetilde{A}^{-1}\|}_{\text{beschränkt}}.
\end{aligned}$$

Analog gilt für den Fehler im Zustand

$$\begin{aligned}
\|y - \tilde{y}\| &= \|A^{-1}Bu - \widetilde{A}^{-1}B\tilde{u}\| \\
&\leq \|A^{-1}Bu - A^{-1}B\tilde{u}\| + \|A^{-1}B\tilde{u} - \widetilde{A}^{-1}B\tilde{u}\| \\
&\leq \underbrace{\|A^{-1}\|}_{\text{beschränkt}}\ \underbrace{\|B\|}_{\text{beschränkt}}\ \underbrace{\|u - \tilde{u}\|}_{\longrightarrow 0} + \underbrace{\|A^{-1} - \widetilde{A}^{-1}\|}_{\longrightarrow 0}\ \underbrace{\|B\|}_{\text{beschränkt}}\ \|\tilde{u}\|
\end{aligned}$$

□

Beispiel 3.2. *Das folgende linear-quadratisches Optimalsteuerungsproblem rechtfertigt die oben geforderten Eigenschaften.*

$$\min_{\substack{y\in H^1(\Omega)\\ u\in L^2(\Omega)}} J(y,u) = \tfrac{1}{2}(y - y_d, y - y_d)_{L^2} + \tfrac{1}{2}(u,u)_{L^2},$$
$$-\Delta y = u \ \text{ in } \Omega\,, \quad n^T\nabla y + y = 0 \ \text{ auf } \partial\Omega.$$

Für dieses Problem sind alle obigen Voraussetzungen erfüllt. Insbesondere sind H_{yy}, H_{uu} und B gleich der Identität auf $L^2(\Omega)$, wenn man $L^2(\Omega)$ mit dessen Dualraum identifiziert. Des Weiteren ist der Operator A definiert durch

$$A : y \mapsto \big(v \mapsto \int_\Omega \nabla y^T \nabla v\, dt + \int_{\partial\Omega} yv\, ds\big).$$

A ist beschränkt invertierbar, denn für dessen kleinsten Eigenwert λ_{min} gilt

$$\lambda_{min} = \min_{v\in H^1(\Omega)} \frac{(Av)(v)}{\|v\|^2_{L^2(\Omega)}} \geq \frac{c(v,v)_{H^1(\Omega)}}{\|v\|^2_{L^2(\Omega)}} \geq c > 0, \tag{3.2}$$

wobei die verallgemeinerte Friedrichs'sche Ungleichung A.3 verwendet wurde. Zudem ergibt sich die beschränkte Invertierbarkeit von H_R automatisch, da $H_R - H_{uu} = B^*A^{-1}H_{yy}A^{-1}B$ zumindest positiv semidefinit ist. Unter Verwendung des Rayleigh-Quotienten lässt sich dann leicht zeigen, dass $\|H_R^{-1}\| \leq \|H_{uu}^{-1}\|$.

3.2 Der Fehler im Newton-Lagrange-Schritt

In der Nähe der Lösung sollte das Composite-Step-Verfahren in ein inexaktes Newton-Verfahren übergehen, um dessen hohe Konvergenzgeschwindigkeit auszuschöpfen. Die folgenden Ausführungen dienen der Vorbereitung entsprechender lokaler Konvergenzresultate, welche im anschließenden Abschnitt bewiesen werden. Die Inexaktheit stammt hier ebenfalls aus der Verwendung des gestörten Operators $\widetilde{A}^{-1}$. Die obigen Bezeichnungen sollen für die Bestandteile der zweiten Ableitung der Lagrange-Funktion, der KKT-Operatormatrix übernommen werden.

$$L'' =: K = \begin{pmatrix} H_{yy} & & A \\ & H_{uu} & B^* \\ A & B & \end{pmatrix}$$

Die gestörte KKT-Matrix $\widetilde{K}$ definiert sich analog mit $\widetilde{A} := (\widetilde{A}^{-1})^{-1}$ an Stelle von A. Der exakte und der inexakte Newton-Lagrange-Schritt sind dann die Lösungen von

$$K\,\Delta z = r \quad \text{bzw.} \quad \widetilde{K}\,\delta z = r.$$

Die Konvergenzaussagen im kommenden Abschnitt werden in Abhängigkeit der Größe

$$\delta := \frac{\|\delta z - \Delta z\|}{\|\delta z\|}$$

formuliert. Es ist nun möglich die Störung $\widetilde{A}^{-1}$ aus δ zu isolieren. Dazu sei zunächst K^{-1} explizit angegeben. Mit der reduzierten Hesse-Matrix $H_R = B^*A^{-1}H_{yy}A^{-1}B + H_{uu}$ und der Hilfsmatrix $F = A^{-1}BH_R^{-1}B^*$ gilt

$$K^{-1} = \begin{pmatrix} FA^{-1} & -A^{-1}BH_R^{-1} & A^{-1} - FA^{-1}H_{yy}A^{-1} \\ -H_R^{-1}B^*A^{-1} & H_R^{-1} & H_R^{-1}B^*A^{-1}H_{yy}A^{-1} \\ A^{-1} - A^{-1}H_{yy}FA^{-1} & A^{-1}H_{yy}A^{-1}BH_R^{-1} & -A^{-1}H_{yy}(A^{-1} - FA^{-1}H_{yy}A^{-1}) \end{pmatrix}.$$

Dann lässt sich die Matrix $K^{-1}\widetilde{K} - I$ faktorisieren durch

$$\begin{aligned} K^{-1}\widetilde{K} - I &= \begin{pmatrix} (FA^{-1}H_{yy} - I)(I - \widetilde{I}) & 0 & -F(I - \widetilde{I}) \\ -H_R^{-1}B^*A^{-1}H_{yy}(I - \widetilde{I}) & 0 & H_R^{-1}B^*(I - \widetilde{I}) \\ A^{-1}H_{yy}(I - FA^{-1}H_{yy})(I - \widetilde{I}) & 0 & (A^{-1}H_{yy}F - I)(I - \widetilde{I}) \end{pmatrix} \\ &= RD_{I-\widetilde{I}} \end{aligned}$$

mit der von $\widetilde{A}$ unabhängigen Matrix

$$R = \begin{pmatrix} FA^{-1}H_{yy} - I & 0 & -F \\ -H_R^{-1}B^*A^{-1}H_{yy} & 0 & H_R^{-1}B^* \\ A^{-1}H_{yy}(I - FA^{-1}H_{yy}) & 0 & A^{-1}H_{yy}F - I \end{pmatrix}$$

und der Diagonalmatrix

$$D_{I-\widetilde{I}} = \begin{pmatrix} I-\widetilde{I} & & \\ & I-\widetilde{I} & \\ & & I-\widetilde{I} \end{pmatrix},$$

wobei $\widetilde{I} := A^{-1}\tilde{A}$. Der relative Fehler δ im Newton-Lagrange-Schritt lässt sich damit abschätzen wie folgt:

$$\begin{aligned} \delta = \frac{\|\Delta z - \delta z\|}{\|\delta z\|} &= \frac{\|(K^{-1} - \tilde{K}^{-1})r\|}{\|\delta z\|} \\ &= \frac{\|(K^{-1}\tilde{K} - I)\delta z\|}{\|\delta z\|} \\ &\leq \|K^{-1}\tilde{K} - I\| \\ &= \|RD_{I-\widetilde{I}}\| \\ &\leq \|R\|\,\|D_{I-\widetilde{I}}\| \\ &= \|R\|\,\|I - \tilde{I}\|. \end{aligned} \tag{3.3}$$

R besteht nur aus beschränkten Operatoren und ist somit selbst beschränkt. Daher verschwindet δ für $\|I - \tilde{I}\| \to 0$.

3.3 Lokale Konvergenzresultate

In diesem Abschnitt bezeichnet (V) die folgenden Voraussetzungen: Die Abbildung $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ besitze eine Lösung z^* der Gleichung $F(z) = 0$. In einer Umgebung $U \subset D$ von z^* sei F stetig Fréchet-differenzierbar und $F'(z^*)$ invertierbar.

Löst man das Newton-System inexakt, (3.4), verbleibt ein Fehler $\delta z_k - \Delta z_k$ in der Iterierten, welcher sich durch ein nicht verschwindendes Residuum r_k ausdrückt,

$$F'(z_k)\delta z_k = -F(z_k) + r_k, \qquad z_{k+1} = z_k + \delta z_k, \tag{3.4}$$

mit

$$r_k = F'(z_k)\delta z_k + F(z_k) = F'(z_k)(\delta z_k - \Delta z_k). \tag{3.5}$$

Ein Maß für die Genauigkeit der Lösungen stellt die Folge $\{\eta_k\}$ oberer Schranken der relativen Residuen dar,

$$\frac{\|r_k\|}{\|F(z_k)\|} \leq \eta_k\,. \tag{3.6}$$

Abhängig von den absoluten Größen und der Konvergenz dieser sogenannten „Forcing Sequence“ können Aussagen über die lokale Konvergenz des inexakten Newton-Verfahrens getroffen werden. Offensichtlich erhält man für $\eta_k = 0$ das exakte Newton-Verfahren. Um inexakte Schritte $\delta z_k = 0$ auszuschließen, soll immer $\eta_k < 1$ gelten. Eine leichte Verstärkung dieser Forderung garantiert dann bereits lokale Konvergenz.

Satz 3.3 (Theorem 2.3 in [10]). *Es gelten die Voraussetzungen (V) sowie $\eta_k \leq \eta_{max} < 1$. Dann existiert ein $\varepsilon > 0$, so dass für $\|z_0 - z^*\| \leq \varepsilon$ die durch* (3.4) *definierte Folge $\{z_k\}$ gegen z^* konvergiert.*

Die Konvergenrate (Def. A.8) des äußeren Verfahrens und der relativen Residuen hängt wie folgt zusammen:

Satz 3.4 (Theorem 3.3 in [10]). *Es gelten die Voraussetzungen (V). Die inexakte Newton-Iteration $\{z_k\}$ konvergiere gegen z^*. Superlineare Konvergenz kann genau dann garantiert werden, wenn*

$$\lim_{k\to\infty} \eta_k = 0.$$

Ist weiterhin F' Lipschitz-stetig auf U, so ist die Konvergenz quadratisch genau dann, wenn

$$\eta_k = \mathcal{O}(\|F(z_k)\|), \qquad \text{für } k \to \infty.$$

Die vorgestellte Bedingung (3.6) ist leider nicht affin-invariant und basiert nur auf Größen im Bildraum. Die resultierenden Konvergenzkriterien können bei schlecht konditionierten Funktionalen, wie sie insbesondere bei diskretisierten partiellen Differentialgleichungen auftreten, den Iterationsprozess stark verlangsamen ([12]). In [35] nimmt sich Ypma dieser Problematik an. Dabei wird $\{\eta_k\}$ durch eine affin-invariante Forcing Sequence $\{\nu_k\}$ ersetzt:

$$\frac{\|F'(z_k)^{-1} r_k\|}{\|F'(z_k)^{-1} F(z_k)\|} = \frac{\|\delta z - \Delta z\|}{\|\Delta z\|} \leq \nu_k \leq \nu_{\max}$$

Erneut garantiert die Forderung $\nu_{\max} < 1$ lokal Konvergenz und es können entsprechende Aussagen über die Konvergenzgeschwindigkeit getroffen werden.

Satz 3.5 (Theorem 3.6 in [35]). *Es seien die Voraussetzungen (V) erfüllt. Die inexakte Newton-Folge $\{z_k\}$ konvergiere gegen z^*. Weiterhin sei die Ableitung F' auf ganz U Lipschitz-stetig. Dann konvergiert $\{z_k\}$ superlinear genau dann, wenn*

$$\lim_{k\to\infty} \nu_k = 0$$

Die Konvergenz ist quadratisch genau dann, wenn

$$\nu_k = \mathcal{O}(\|F'(z_k)^{-1}F(z_k)\|), \qquad \textit{für } k \to \infty.$$

Ein für diese Ausarbeitung passenderer Ansatz wurde von Peter Deuflhard in [12] dargestellt. Die Inexaktheit stamme dann nicht mehr vom frühzeitigen Abbruch eines iterativen Lösungsverfahrens, sondern von der Anwendung einer numerischen Approximation der Inversen der Hesse-Matrix

$$M_k^{-1} \approx F'(z_k)^{-1}\,.$$

Der inexakte Newton-Schritt δz_k löst dann das System

$$M_k \delta z_k = F(z_k)\,.$$

In der Literatur wird dieser Ansatz als „Newton-like Method“ bezeichnet. Es definiere

$$\vartheta_k := \frac{\|M_k^{-1} r_k\|}{\|M_k^{-1}F(z_k)\|} = \frac{\|M_k^{-1}(M_k - F'(z_k))\delta z_k\|}{\|\delta z_k\|} \tag{3.7}$$

mit dem Residuum r_k wie in (3.5). Die folgende Konvergenzaussage entstammt [12].

Satz 3.6. *Es gelten die Voraussetzungen (V). Die inexakte Newton-Iteration $\{z_k\}$ konvergiere gegen z^*. Die Approximation M_k der Hesse-Matrix hänge stetig von z ab und sei nichtsingulär. Zudem gelte die Lipschitz-Bedingung*

$$\|M_k^{-1}(F'(z_k + t\delta z_k) - F'(z_k))\,\delta z_k\| \le \omega_k\, t\, \|\delta z_k\|^2\,, \qquad t \in [0,1]\,, \tag{3.8}$$

mit $\omega_k \le \omega_{max}$. Dann wird die Konvergenzrate der inexakten Newton-Iteration asymptotisch durch den Term ϑ_k dominiert:

$$\lim_{k\to\infty} \frac{\|z_{k+1} - z_k\|}{\|z_k - z_{k-1}\|} \le \lim_{k\to\infty} \vartheta_k\,. \tag{3.9}$$

Beweis. Die Taylor-Formel (A.7) liefert die Abschätzung

$$\begin{aligned}
\|z_{k+1} - z_k\| &= \|\delta z_k\| = \|M_k^{-1}F(z_k)\| \\
&= \|M_k^{-1}[F(z_k) - (F(z_{k-1}) + M_{k-1}\delta z_{k-1})]\| \\
&= \left\| M_k^{-1} \int_0^1 (F'(z_{k-1} + t\delta z_{k-1}) - M_{k-1})\delta z_{k-1} dt \right\| \\
&\le \left\| M_k^{-1} \int_0^1 (F'(z_{k-1} + t\delta z_{k-1}) - F'(z_{k-1}))\delta z_{k-1} dt \right\| \\
&\qquad + \|M_k^{-1}(F'(z_{k-1}) - M_{k-1})\delta z_{k-1}\|\,.
\end{aligned}$$

Unter Verwendung der Identität (3.7) und der Voraussetzung (3.8) erhält man

$$\begin{aligned}\frac{\|z_{k+1}-z_k\|}{\|z_k-z_{k-1}\|} = \frac{\|\delta z_k\|}{\|\delta z_{k-1}\|} &\leq \|M_k^{-1}M_{k-1}\| \left(\int_0^1 \omega_{k-1}t\|\delta z_{k-1}\|dt + \vartheta_{k-1}\right) \\ &= \|M_k^{-1}M_{k-1}\| \left(\tfrac{\omega_{k-1}}{2}\|\delta z_{k-1}\| + \vartheta_{k-1}\right).\end{aligned}$$

Aus der Stetigkeit von M_k und der Konvergenz von $\{z_k\}$ folgt schließlich die Aussage (3.9). □

Als Konsequenz ergibt sich lineare Konvergenz, falls $\vartheta_k \leq c < 1$ für hinreichend große k. Für $\vartheta_k \to 0$ ist die Konvergenzrate superlinear. Das folgende Lemma liefert einen Zusammenhang zwischen dem Kontraktionsfaktor c im linearen Konvergenzmodus und der Approximationsgüte von M_k.

Lemma 3.7. *Unter den Voraussetzungen von Satz 3.6 und der Forderung* $\|F'(z_k)^{-1}M_k - I\| \leq \varepsilon < 1$ *gilt*

$$\lim_{k\to\infty} \frac{\|z_{k+1}-z_k\|}{\|z_k-z_{k-1}\|} \leq \frac{\varepsilon}{1-\varepsilon}.$$

Beweis. Aus $\|F'(z_k)^{-1}M_k - I\| \leq \varepsilon$ und dem Störungslemma von Banach A.1 ergibt sich

$$\|M_k^{-1}F'(z_k)\| \leq \frac{1}{1-\varepsilon}.$$

Daher gilt

$$\begin{aligned}\vartheta_k &= \frac{\|M_k^{-1}(M_k - F'(z_k))\delta z_k\|}{\|\delta z_k\|} \\ &= \frac{\|M_k^{-1}F'(z_k)(F'(z_k)^{-1}M_k - I)\delta z_k\|}{\|\delta z_k\|} \\ &\leq \|M_k^{-1}F'(z_k)\|\|F'(z_k)^{-1}M_k - I\| \\ &\leq \frac{\varepsilon}{1-\varepsilon}.\end{aligned}$$

Die Behauptung folgt schließlich aus Satz 3.6. □

Kapitel 4

Berechnung des Tangentialschritts

In diesem Kapitel wird das Rüstzeug für die Entwicklung eines inexakten Composite-Step-Verfahrens zusammengetragen. Als Leitfaden dient dabei eine semilineare Differentialgleichung, welche auch später im Kapitel 5 *Numerische Beispiele* die Nebenbedingung definiert. Es werden die wesentlichen Aspekte der Finiten-Elemente-Methoden wiederholt, um die unendlichdimensionalen Probleme zu diskretisieren. Danach werden einige Abwandlungen des CG-Verfahrens präsentiert sowie ein adaptives Abbruchkriterium. Abschließend wird die konkrete Implementierung formuliert.

4.1 Struktur der Nebenbedingung

Im Folgenden sei immer $V := H^1(\Omega)$ auf einem beschränkten Lipschitz-Gebiet $\Omega \in \mathbb{R}^d$ (Def. A.4). Die Nebenbedingung definiere sich durch das Randwertproblem mit verteilter Steuerung und Robin-Randbedingungen

$$\begin{aligned} -\Delta y + f(y) &= u \qquad \text{in } \Omega \\ n^T \nabla y + \alpha y &= \beta \qquad \text{auf } \partial\Omega. \end{aligned} \tag{4.1}$$

Dabei seien $y \in V$, $u \in L^2(\Omega)$, $0 \leq \alpha \in L^\infty(\partial\Omega)$, $\beta \in L^2(\partial\Omega)$ und $f : V \to L^2(\Omega)$ Fréchet-differenzierbar. Robin-Randbedingungen können als Verallgemeinerung von Dirichlet- und Neumann-Randbedingungen angesehen werden, da durch geeignete Wahl der Parameterfunktionen α und β die gewünschten Randbedingungen erzwungen werden können. Differentialgleichungen der Art (4.1) bezeichnet man als semilinear, da die Funktion $f(y)$ möglicherweise nichtlinear, der Term höchster Dif-

ferentiationsordnung jedoch weiterhin linear ist. Mit dem entsprechend definierten Operator $\mathcal{A} : V \to V^*$ lautet die Variationsformulierung

$$\begin{aligned} \langle \mathcal{A}(y), v \rangle_{V^*,V} &= \int_\Omega \nabla y^T \nabla v \, dt + \int_\Omega f(y) v \, dt + \int_{\partial\Omega} \alpha y v \, ds \\ &= \int_\Omega uv \, dt + \int_{\partial\Omega} \beta v \, ds \qquad \forall v \in V \, . \end{aligned} \tag{4.2}$$

Im Kontext der in dieser Arbeit untersuchten Newton-Verfahren wird vorausgesetzt, dass der lineare abgeleitete Operator

$$\begin{aligned} A &:= \mathcal{A}_y(y_0) : V \to V^* \\ \langle Ay, v \rangle_{V^*,V} &= \int_\Omega \nabla y^T \nabla v \, dt + \int_\Omega f'(y_0) y \, v \, dt + \int_{\partial\Omega} \alpha y v \, ds \end{aligned}$$

invertierbar ist. Der nun folgende Satz gewährleistet die Bijektivität, falls der zweite Integrand die Form $f'(y_0)y = c_0 y$ mit $0 \leq c_0 \in L^\infty$ besitzt.

Satz 4.1 (Theorem 1.19 in [20]). *Sei $\Omega \in \mathbb{R}^d$ ein beschränktes Lipschitz-Gebiet. Die auf $H^1(\Omega) \times H^1(\Omega)$ definierte Bilinearform*

$$a(y, v) = \int_\Omega \nabla y^T \nabla v \, dt + \int_\Omega c_0 y v \, dt + \int_{\partial\Omega} \alpha y v \, ds \, ,$$

definiert einen linearen Operator A durch $\langle Ay, v \rangle_{V^*,V} = a(y, v)$. Unter den Voraussetzungen $0 \leq c_0 \in L^\infty(\Omega)$, $0 \leq \alpha \in L^\infty(\partial\Omega)$ und $\|c_0\|_{L^2(\Omega)} + \|\alpha\|_{L^2(\partial\Omega)} > 0$ ist A beschränkt invertierbar.

Beweis. Die Aussage ist eine direkte Folgerung des Lemmas von Lax und Milgram A.2. Stetigkeit folgt dabei aus dem Spursatz (siehe etwa [7]) und V-Elliptizität kann mit der verallgemeinerten Friedrichs'schen bzw. der verallgemeinerten Ungleichung von Poincaré gezeigt werden. Ein leicht verständlicher Beweis findet sich in dem Lehrbuch von Tröltzsch [29]. □

4.2 Diskretisierungskonzept - Finite Elemente

In diesem Abschnitt wird die Art der Diskretisierung diskutiert. Wegen des Differentialoperators der Nebenbedingung bietet sich die Verwendung von Finite-Elemente-Methoden an. Diese sind deutlich flexibler als etwa Finite-Differenzen-Verfahren, da sie auch bei komplizierten Gebieten Ω eingesetzt werden können und zudem einfach adaptive Verfeinerungen

erlauben. Ausgangspunkt bildet dabei die schwache Formulierung des Randwertproblems

$$a(y, v) = F(v) \qquad \forall v \in V. \tag{4.3}$$

Bei den Finiten Differenzen würde der Differentialoperator selbst diskretisiert werden, wohingegen FE-Methoden auf dem Galerkin-Ansatz beruhen, bei dem man den Lösungsraum diskretisiert. Dabei wird V durch einen endlichdimensionalen Teilraum $V_h \subset V$ ersetzt und die Variationsformulierung auf diesen eingeschränkt:

$$a(y_h, v_v) = F(v_h) \qquad \forall v_h \in V_h. \tag{4.4}$$

Um den Raum V_h zu erzeugen, wird das Gebiet Ω in polytope Teilgebiete $\mathcal{T} = \{T_1, T_2, ..., T_m\}$ zerlegt, sehr geeignet sind Simplizes, es werden aber auch d-dimensionale Quader und in 2D Prismen eingesetzt. Nun definiert man leicht zu handhabende Basisfunktionen φ_i von V_h mit einem möglichst kleinen Träger. Bei linearen Finiten Elementen mit Lagrange-Basis wählt man diese so, dass sie an einem einzigen Knoten den Wert 1 annehmen, in den angrenzenden Elementen T_i lineare Polynome darstellen und auf allen anderen Knoten und Elementen verschwinden. Man erhält

$$V_h = \operatorname{span}\{\varphi_i\} = \{v \in C(\Omega) : v|_{T_i} \in \mathbb{P}_1 \ , \ i = 1, \ldots, m\} \subset H^1(\Omega).$$

Die Raumdimension entspricht dann der Anzahl n der Gitterknoten. Wegen der Linearität von a und F reicht es aus, (4.4) nur für die Basisfunktionen φ_i zu fordern. Stellt man nun die gesuchte diskrete Lösung als Linearkombination der Basisfunktionen dar, $y_h = \sum y_i \varphi_i$, erhält man das lineare Gleichungssystem

$$\sum_{i=1}^{n} y_i a(\varphi_i, \varphi_j) = F(\varphi_j), \qquad j = 1, ..., n,$$

kurz $A\bar{y} = \bar{F}$, mit $A_{ij} = a(\varphi_i, \varphi_j)$, $\bar{y} = (y_i)_{i=1,...,n}$ und $\bar{F} = (F(\varphi_i))_{i=1,...,n}$. Offensichtlich folgt auch hier die eindeutige Lösbarkeit aus dem Lemma von Lax und Milgram. Ein wichtiges Resultat ist die Galerkin-Optimalität, welche besagt, dass die Approximation y_h die bestmögliche im Lösungsraum V_h bezüglich der durch die Bilinearform $a(.,.)$ induzierte Energienorm ist. Aus diesem Grund lässt sich das Finite-Elemente-Verfahren als Projektionsverfahren auffassen. Durch die oben beschriebene Wahl der Basisfunktionen hat die Systemmatrix A die vorteilhafte Eigenschaft, extrem dünn besetzt zu sein, da die Integrale $a(\varphi_i, \varphi_j)$ nur für unmittelbar benachbarte Knoten i und j nicht verschwinden.

Diskrete Normen

Damit das diskretisierte Problem das gleiche Konvergenzverhalten wie das unendlichdimensionale Grundproblem an den Tag legt, wird eine gute Synchronisation der beiden Ansätze angestrebt. Das heißt, dass das Gitter ausreichend fein gewählt werden sollte und dass die Normen der Funktionenräume oder zumindest passende Approximationen benutzt werden. Ein entsprechendes Konvergenzresultat findet sich etwa in [34]. Die Berechnung der $L^2(\Omega)$- und der $H^1(\Omega)$-Normen ist dabei besonders einfach, da bei dem hier angedachten Einsatz Finiter-Elemente-Methoden ohnehin die Systemmatrizen M und A mit

$$M_{ij} = \int_\Omega \varphi_i \varphi_j \, dt\,, \quad A_{ij} = \int_\Omega \nabla \varphi_i^T \nabla \varphi_j \, dt$$

assembliert werden. Für eine diskrete Funktion $v_h \in V_h$ mit dem Koeffizientenvektor $v \in \mathbb{R}^n$ gilt dann bei exakter Integration

$$\|v_h\|^2_{L^2(\Omega)} = v^T M v \quad \text{sowie} \quad \|v\|^2_{H^1(\Omega)} = v^T (A + M) v.$$

Die diskrete duale Norm $\|.\|_{V_h^*}$ eines linearen Funktionals $f : V_h \to \mathbb{R}$ definiert sich über

$$\|f\|_{V_h^*} = \sup_{v_h \in V_h} \frac{|f(v_h)|}{\|v\|_{V_h}}.$$

Liegt f als Vektor vor, d.h. $f(v_h) = b_f^T v$, und die Funktionenraumnorm lässt sich durch eine Matrix S ausdrücken, $\|v_h\|_{V_h} = \|v\|_S$, so lässt sich die diskrete duale Norm explizit berechnen, denn es gilt

$$\begin{aligned}
\|f\|_{V_h^*} = \sup_{v_h \in V_h} \frac{|f(v_h)|}{\|v\|_{V_h}} &= \sup_{v \in \mathbb{R}^n} \frac{|b_f^T v|}{\sqrt{v^T S v}} \\
&= \sup_{w \in \mathbb{R}^n} \frac{|b_f^T (S^{-1} w)|}{\sqrt{(S^{-1}w)^T S (S^{-1} w)}} \\
&= \sup_{w \in \mathbb{R}^n} \frac{|b_f^T S^{-1} w|}{\|w\|_{S^{-1}}} \\
&= \|b_f\|_{S^{-1}},
\end{aligned}$$

unter Verwendung der Ungleichung von Cauchy-Schwarz. Um nicht immer die Berechnung der S-Norm und der noch aufwendigeren S^{-1}-Norm durchführen zu müssen, wurden diese in der algorithmischen Realisierung durch die Matrixnormen $\|.\|_{h^2} := \|.\|_{h^2 I}$ bzw. $\|.\|_{h^{-2}} := \|.\|_{h^{-2} I}$ ersetzt, mit der durch die Gitterweite skalierten Identität $h^2 I$ bzw. $h^{-2} I$. Insbesondere für die diskreten Normen $\|.\|_M$, $\|.\|_{M^{-1}}$ mit der Massenmatrix M sind das gute Alternativen, denn $\|M\| \approx h^2$ und $\|M^{-1}\| \approx h^{-2}$ stimmen wegen des geringen Spektalradius von M sehr genau überein (zumindest bei den hier verwendeten uniformen Gittern).

4.3 Das Vorkonditionierte CG-Verfahren

Unter den Lösern für große lineare Gleichungssysteme $Hp = b$ ist das Verfahren der konjugierten Gradienten (kurz CG-Verfahren) ein wichtiger Vertreter. Für quadratische positiv definite Matrizen $H \in \mathbb{R}^{n\times n}$ ist die Lösung gleichbedeutend mit der Minimierung der quadratischen Form $q(p) = \frac{1}{2}p^T Hp - b^T p$. Beim CG-Verfahren wird dieses Funktional in jedem Schritt über einem affinen Unterraum U_k bezüglich der Energienorm $\|\cdot\|_H = (\cdot, H\cdot)^{1/2}$ minimiert. Aus diesem Grund kann man das Verfahren in die Gruppe der Projektionsverfahren einordnen. Bei exakter Arithmetik liefert die Methode nach höchstens n Schritten die exakte Lösung. Das Verfahren lässt sich noch deutlich beschleunigen, indem man einen Vorkonditionierer zwischenschaltet. Dabei löst man nicht mehr das Originalproblem $Hp = b$, sondern das transformierte Problem $H\mathcal{H}^{-1}\tilde{p} = b$ mit $\tilde{p} = \mathcal{H}p$ mit dem symmetrischen positiv definiten Vorkonditionierer $\mathcal{H}$. Bezüglich des euklidischen Skalarprodukts ist die Matrix $H\mathcal{H}^{-1}$ zwar nicht mehr selbstadjungiert, wohl aber bezüglich des durch $\mathcal{H}^{-1}$ induzierten Produktes $(\cdot, \mathcal{H}^{-1}\cdot)$ ([31]). Mit diesem können wir also den normalen CG-Algorithmus auf das transformierte System anwenden. Eine sparsame Variante, auf die ich mich später noch beziehen werde, findet sich in [31]:

Algorithmus 2 PCG-Algorithmus

$p_k = PCG(H, b, \mathcal{H})$
$p_0 = 0, r_0 = b, s_1 = \bar{r}_0 = \mathcal{H}^{-1}r_0$
for $k = 1, \ldots, k_{\max}$ **do**
$\alpha_k = \frac{r_{k-1}^T \bar{r}_{k-1}}{s_k^T H s_k}$
$p_k = p_{k-1} + \alpha_k s_k$
If Abbruchkriterium positiv **then** Stop.
$r_k = r_{k-1} - \alpha_k H s_k$
$\bar{r}_k = \mathcal{H}^{-1} r_k$
$\beta_k = \frac{r_k^T \bar{r}_k}{r_{k-1}^T \bar{r}_{k-1}}$
$s_{k+1} = \bar{r}_k + \beta_k s_k$
end for

Dabei bleibt der Aufwand für eine Iteration gleich dem des gewöhnlichen CG auf dem Originalproblem bis auf eine Anwendung des Vorkonditionierers. Wir brauchen dabei $\mathcal{H}$ und $\mathcal{H}^{-1}$ gar nicht vorliegen zu haben, es reicht eine Vorschrift für die Anwendung von $\mathcal{H}^{-1}$, wie es zum Beispiel beim Mehrgitteralgorithmus der Fall ist. Der Zusammenhang zwischen der Wahl des Vorkonditionierers und der Konvergenzgeschwindigkeit lässt

sich mit Hilfe der Kondition $\kappa = \kappa(H\mathcal{H}^{-1})$ quantifizieren. Der relative Approximationsfehler des vorkonditionierten CG-Verfahrens nach der k-ten Iteration lässt sich abschätzen gemäß

$$\frac{\|p_k - p^*\|_H}{\|p_0 - p^*\|_H} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k.$$

Für den Beweis siehe etwa [14]. Das nicht vorkonditionierte CG-Verfahren ist gleichbedeutend mit der Wahl $\mathcal{H} = I$, man hat also $\kappa = \kappa(H)$. Hier zeigt sich, was einen guten Vorkonditionierer ausmacht. Er verringert den Quotienten aus größtem und kleinstem Eigenwert und konzentriert somit das Spektrum.

4.3.1 Diskreter Energiefehler

Für gut konditionierte Probleme kann lineare Konvergenz des Energiefehlers angenommen werden. Aus dieser Erkenntnis lässt sich wie folgt ein preiswerter Fehlerschätzer herleiten. Für den Fehler in der Energienorm $\epsilon_k := \|p_k - p^*\|_H^2$ gelte

$$\epsilon_{k+1} \leq \Theta\epsilon_k, \qquad \text{für ein } \Theta < 1.$$

Durch die Orthogonalisierung in jedem CG-Schritt gilt für $\gamma_k = \|p_{k+1} - p_k\|_H^2$ die Identität

$$\epsilon_k = \epsilon_{k+1} + \gamma_k\,.$$

Aus der oberen Ungleichung erhält man die Abschätzung

$$\epsilon_k \leq \frac{\gamma_k}{1-\Theta}. \tag{4.5}$$

Die Werte γ_k lassen sich mit wenig Aufwand zusätzlich berechnen und offensichtlich gilt $\epsilon_k = \sum_{j=k}^{n-1}\gamma_j$. Mit dieser Information lässt sich die Schranke $\Theta \approx \tilde{\Theta}$ schätzen. Für $\epsilon_{k+1} \approx \tilde{\Theta}\epsilon_k$ ist

$$\frac{\epsilon_{k+1}-\epsilon_k}{\epsilon_{k-s+1}-\epsilon_{k-s}} \approx \frac{\tilde{\Theta}^s\epsilon_{k-s+1} - \tilde{\Theta}^s\epsilon_{k-s}}{\epsilon_{k-s+1}-\epsilon_{k-s}} = \tilde{\Theta}^s,$$

mit $s \geq 1$

$$\frac{\epsilon_{k+1}-\epsilon_k}{\epsilon_{k-s+1}-\epsilon_{k-s}} = \frac{\sum\limits_{j=k+1}\gamma_j - \sum\limits_{j=k}\gamma_j}{\sum\limits_{j=k-s+1}\gamma_j - \sum\limits_{j=k-s}\gamma_j} = \frac{\gamma_k}{\gamma_{k-s}}$$

und folglich $\tilde{\Theta} \approx (\frac{\gamma_k}{\gamma_{k-s}})^{\frac{1}{s}}$. Für $s > 1$ wird bereits eine Art Mittelung vollzogen, welche Schwankungen in der Kontraktion ausgleicht. In den numerischen Beispielen wird $s = 2$ gewählt.

4.3.2 Truncated CG für Nichtkonvexe Modelle

Die hier zu lösenden Gleichungssysteme stammen aus Minimierungsproblemen, für die eine geeignete Abstiegsrichtung zu ermitteln ist. Wie im Einleitungskapitel beschrieben, stellt der Newtonschritt eine optimale Suchrichtung dar, falls das Funktional lokal konvex ist. Im Voraus ist für gewöhnlich nicht bekannt, ob diese Eigenschaft erfüllt wird, allerdings wird Nichtkonvexität im Laufe der Iteration durch das CG-Verfahren dedektiert. In diesem Fall ist die Hesse-Matrix $H = J''(x)$ indefinit und man erhält irgendwann ein negatives Energieprodukt $s_k^T H s_k < 0$. Auf Dembo und Steihaug ([11]) geht die Idee zurück, das normale Verfahren frühzeitig abzubrechen, sobald entweder

(1) ein bestimmtes Genauigkeitskriterium erfüllt oder

(2) eine Richtung negativer Krümmung s_k^{neg} erreicht wird.

Bezüglich der Bezeichnungen in Algorithmus 2 stellt die Abfrage

$$\frac{\|r_k\|}{\|r_0\|} \leq \eta$$

ein sinnvolles Abbruchkriterium für (1) dar. Beachte im Rückblick auf Abschnitt 3.3, dass dieses Kriterium der Definition der Forcing-Sequence (3.6) entspricht und sich durch die Wahl von η die asymptotische Konvergenzgeschwindigkeit des äußeren Verfahrens steuern lässt. Dembo und Steihaug weisen darauf hin, dass vor Erreichen einer Richtung negativer Krümmung s_k^{neg} alle Iterierten $p_0, ..., p_k$ sowie s_k^{neg} selbst Abstiegsrichtungen darstellen. Sie schlagen vor, in beiden Fällen (1) und (2) die aktuelle Iterierte p_k zurückzugeben. Das garantiert, dass der Algorithmus im konvexen Fall nicht in die iterative Berechnung der optimalen Newton-Richtung eingreift und im nichtkonvexen Fall zumindest eine Abstiegsrichtung liefert. In dieser Arbeit soll die Methode leicht verändert eingesetzt werden. Bei Iterationsabbruch (1) wird ebenfalls die Iterierte p_k weiterverwendet, im Fall (2) werden dagegen beide Richtungen p_k und s_k^{neg} zurückgegeben. Anschließend wird ein kubisches Modell des zu optimierenden Funktionals auf dem ein- bzw. zweidimensionalen Unterraum minimiert, der durch die Rückgabewerte aufgespannt wird (siehe Abschnitt 2.3).

4.4 Mehrgittervorkonditionierer

Bei der Wahl eines geeigneten Vorkonditionierers $\mathcal{H}$ muss der Nutzer entscheiden, wieviel Informationen er neben der Koeffizientenmatrix H

zur Verfügung stellen kann und will. Bei der Verwendung von Matrixzerlegungsmethoden werden keine weiteren Informationen benötigt, daher sind diese sehr flexibel einsetzbar. Leider wird bei Anwendungen mit einer großen Anzahl an Unbekannten und schlecht konditionierten Matrizen, wie sie im Bereich partieller Differentialgleichungen auftreten der Iterationsprozess nicht ausreichend beschleunigt, da diese Methoden nicht spektraläquvialent sind, das heißt, für die Kondition gilt $\kappa(\mathcal{H}^{-1}H) \gg 1$. Eine Klasse von Vorkonditionierern, welche die Forderung nach Spektraläquvivalenz erfüllen, bilden die Mehrgitterverfahren, häufig auch Multilevelmethoden genannt. Die Entwicklung dieser Techniken begann in den sechziger Jahren des letzten Jahrhunderts und sie wurden besonders durch Wolfgang Hackbusch auf ein stabiles Fundament gestellt (siehe etwa [18],[19]). Heutzutage werden sie in praktisch jedem Gebiet eingesetzt, bei dem partielle Differentialgleichungen mit numerischen Methoden gelöst werden ([36]).

Für die Entwicklung von Mehrgitterverfahren werden die folgenden Erkenntnisse benutzt. Zum einen stellt sich heraus, dass bestimmte iterative Löser, wie zum Beispiel das gedämpfte Jacobiverfahren, hochfrequente Fehleranteile glätten, die niedrigfrequenten jedoch kaum verändern. Zum anderen sind niederfrequente Anteile auf einem gröberen Gitter höherfrequent. Diese Beobachtung lässt sich ausnutzen, indem man sukzessive auf immer gröberen Räumen die jeweils hochfrequenten Fehleranteile herausglättet und schließlich auf dem gröbsten Gitter direkt löst, so dass schließlich das komplette Frequenzspektrum abgedeckt und somit der gesamte Fehler erfasst wird. Die folgende Formulierung des klassischen Mehrgitteralgorithmus sowie der entsprechende Konvergenzbeweis orientieren sich stark an [31]. Vorausgesetzt sei die Existenz einer Folge geschachtelter FE-Räume $S_0 \subset S_1 \subset \cdots \subset S_j \subset H^1(\Omega)$. Die eigentlich trivialen Einbettungen $S_{k-1} \hookrightarrow S_k$ und $S_k^* \hookrightarrow S_{k-1}^*$ sind aufgrund der verschiedenen Basen durch Prolongationsoperatoren I_{k-1}^k bzw. Restriktionsoperatoren I_k^{k-1} zu realisieren. Da man S_k und S_k^* auf kanonische Weise mit $\mathbb{R}^{n_k}$ identifiziert, kann man diese Operatoren durch Basiswechselmatrizen darstellen. Bei den Prolongationen werden den hinzukommenden Knoten interpolierte Werte der alten Knotenwerte zugeordnet, wohingegen die Restriktionen die Koeffizientenvektoren der Residuen aus dem Dualraum S_k^* in den nächst gröberen Raum S_{k-1}^* übertragen. Bei konformen finiten Elementen gilt $I_k^{k-1} = (I_{k-1}^k)^T$. Damit lässt sich rekursiv der klassische Mehrgitteralgorithmus auf Seite 36 darstellen.
Für $\mu = 1$ spricht man vom V-Zyklus, für $\mu = 2$ vom W-Zyklus. In der Praxis hat sich der V-Zyklus als hocheffizient erwiesen, der Beweis für dessen gitterunabhängige Konvergenz (unabhängig vom Elementdurchmesser h) ist allerdings auch deutlich komplizierter als jener für den W-Zyklus,

Algorithmus 3 Klassischer Mehrgitteralgorithmus

$\hat{y}_k = MGM_k(\nu_1, \nu_2, \mu, r_k)$

1. Vorglättung durch ν_1-maliges Anwenden des Glätters G
   $\bar{y}_k = 0$
   $u_k - \bar{y}_k = G^{\nu_1}(y_k - \bar{y}_k)$
2. Grobgitterkorrektur rekursiv durch MGM_{k-1}
   $r_{k-1} = I_k^{k-1}(r_k - A_k\bar{y}_k)$
   if $k = 1$ **then**
      löse direkt $\hat{v}_0 = A_0^{-1}r_0$
   else
      $\hat{v}_{k-1} = 0$
      for $i = 1, \ldots, \mu$ **do**
         $dv = MGM_{k-1}(\nu_1, \nu_2, \mu, r_{k-1} - A_{k-1}\hat{v}_{k-1})$
         $\hat{v}_{k-1} = \hat{v}_{k-1} + dv$
      end for
   end if
3. Nachglättung durch ν_2-maliges Anwenden von G
   $\hat{y}_k = \bar{y}_k + I_{k-1}^k\hat{v}_{k-1}$
   $y_k - \hat{y}_k = G^{\nu_2}(y_k - \hat{y}_k)$

welcher im Folgenden dargestellt wird. Für den Beweis braucht man nicht nur ein Maß für die absolute Größe des Fehlers, etwa die L^2-Norm, sondern auch für dessen Glattheit. Geeignet dafür ist etwa die H^1-Halbnorm $|v|^2_{H^1(\Omega)} = \int \nabla v^T \nabla v$, welche in die Energienorm $\|.\|_A$ eingeht. Diese misst höherfrequente Anteile, welche in der L^2-Norm nicht erfasst werden können. Im Folgenden bezeichnet y_h die Bestapproximation der Unbekannten $y \in H^1(\Omega)$ in V_h in der Energienorm. Die Argumentationslinie beruht nun auf zwei Beobachtungen, welche von Wolfgang Hackbusch erstmals identifiziert und klar getrennt wurden [36]. Da ist zum einen die *Glättungseigenschaft*

$$\|G^m(y_h - v_h)\|_A \leq \frac{c_1}{hm^\gamma}\|y_h - v_h\|_{L^2(\Omega))} \tag{4.6}$$

nach m-maliger Anwendung des Glätters G. Für das CG-Verfahren lässt sich $\gamma = 1$, für das gedämpfte Jacobi- und symmetrische Gauß-Seidel-Verfahren $\gamma = 1/2$ nachweisen. Die Ungleichung verdeutlicht die transiente Glättung hochfrequenter Fehleranteile (offensichtlich bringt der Term $m^{-\gamma}$ nur für kleine m eine starke Reduktion, $\frac{1}{m^\gamma}/\frac{1}{(m+1)^\gamma}$ geht dann sehr schnell gegen 1). Zum anderen benutzt man die sogenannte *Approximationseigenschaft*

$$\|y - y_h\|_{L^2(\Omega)} \leq c_2 h\|y\|_A, \tag{4.7}$$

welche auch als Jackson-Ungleichung bekannt ist. Der folgende Beweis benutzt die Annahme, dass die Glättung eine Reduktion der Energienorm mit sich bringt,

$$\|Gy_h\|_A \leq \|y_h\|_A \,. \tag{4.8}$$

Satz 4.2 (Satz 5.19 in [31]). *Es seien* (4.6), (4.7) *und* (4.8) *erfüllt. Dann erfüllt Algorithmus 3 mit* $\mu = 2$ *die Kontraktionseigenschaft*

$$\|y_k - \hat{y}_k\|_A \leq \rho_k \|y_k\|_A \tag{4.9a}$$

mit dem rekursiv definierten Kontraktionsfaktor

$$\rho_k = \frac{c}{\nu_2^\gamma} + \rho_{k-1}^2 \,, \qquad \rho_0 = 0 \,. \tag{4.9b}$$

Die Folge der ρ_k *ist unter der Bedingung* $\nu_2^\gamma \geq 4c$ *beschränkt:*

$$\rho_k \leq \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{c}{\nu_2^\gamma}} \leq \frac{1}{2} \,. \tag{4.10}$$

Beweis. Der Beweis geht durch vollständige Induktion über das Gitterlevel k. Für $\rho_0 = 0$ ist wegen der direkten Lösung auf dem Grobgitter nichts zu tun. Für $k \geq 1$ und die exakte Lösung y_{h_k} in S_k gilt mit (4.6) und (4.8)

$$\begin{aligned}
\|y_{h_k} - \hat{y}_k\|_A &= \|G^{\nu_2}(y_{h_k} - \bar{y}_k - \hat{v}_{k-1})\|_A \\
&= \|G^{\nu_2}(y_{h_k} - \bar{y}_k - v_{k-1} + v_{k-1} - \hat{v}_{k-1})\|_A \\
&\leq \|G^{\nu_2}(y_{h_k} - \bar{y}_k - v_{k-1})\|_A + \|G^{\nu_2}(v_{k-1} - \hat{v}_{k-1})\|_A \\
&\leq \frac{c_1}{h{\nu_2}^\gamma} \|(y_{h_k} - \bar{y}_k) - v_{k-1}\|_{L^2(\Omega)} + \|v_{k-1} - \hat{v}_{k-1}\|_A \,.
\end{aligned}$$

Dabei bezeichne v_{k-1} die Bestapproximation von $y_{h_k} - \bar{y}_k$ in S_{k-1} in der Energienorm. Die Gleichung (4.7) liefert für den linken Summanden

$$\|(y_{h_k} - \bar{y}_k) - v_{k-1}\|_{L^2(\Omega)} \leq c_2 h \|y_{h_k} - \bar{y}_k\|_A \,.$$

Für den zweiten Term folgt aus der Induktionsvoraussetzung sowie der Optimalität von v_{k-1}

$$\|v_{k-1} - \hat{v}_{k-1}\|_A \leq \rho_{k-1}^2 \|v_{k-1}\|_A \leq \rho_{k-1}^2 \|y_{h_k} - \bar{y}_k\|_A \,.$$

Zusammen erhält man

$$\|y_k - \hat{y}_k\|_A \leq \left(\frac{c}{\nu_2^\gamma} + \rho_{k-1}^2 \right) \|y_{h_k} - \bar{y}_k\|_A$$

mit der neuen Konstante $c = c_1 c_2$. Bei der Vorglättung wird $\bar{y}_k = 0$ initialisiert, daher gilt $y_{h_k} - \bar{y}_k = G^{\nu_1} u_k$. Mit (4.8) ergibt sich die behauptete Kontraktion (4.9). Mit einer einfachen Induktion zeigt man die Beschränkung (4.10) des Kontraktionsfaktors, womit alles gezeigt ist. □

An der obigen Darstellung des Kontraktionsfaktors lässt sich erkennen, wie durch eine größere Anzahl an Nachglättungen die Qualität des Vorkonditionierers verbessert werden kann. Einen vergleichbaren Einfluss hat auch die Anzahl der Vorglättungen, auch wenn das im vorangegangenen Resultat nicht abzulesen ist. Schaltet man den beschriebenen Vorkonditionierer „in Reihe“, kann das Verfahren auch als eigenständiger iterativer Löser betrachtet werden.

Algorithmus 4 Iterierte MGM zur Lösung von $Ay = b$

$y_0 = 0,\ r_0 = b$
for $i = 0, \ldots, k_{\max}$ **do**
  $d_i = MGM(\nu_1, \nu_2, \mu, r_i)$
  $y_{i+1} = y_i + d_i$
  $r_{i+1} = r_i - Ad_i$
end for

Unter Verwendung von Satz 4.2 schließt man

$$\begin{aligned} \|y - y_i\|_A = \|(y - y_{i-1}) - d_{i-1}\|_A &\leq \rho \|y - y_{i-1}\|_A \\ &\leq \ \ldots \ \leq \rho^i \|y - y_0\|_A = \rho^i \|y\|_A \,, \end{aligned} \tag{4.11}$$

womit die Konvergenz bewiesen ist.

4.5 Das Projizierte CG-Verfahren

Die Ermittlung des Tangentialschritts erfordert die Lösung eines Minimierungsproblems

$$\min_{p \in \mathbb{R}^n} q(p) = \frac{1}{2} p^T H p - b^T p, \text{ so dass } Cp = 0,$$

mit $C \in \mathbb{R}^{m \times n}$. Das resultierende KKT-System

$$\begin{pmatrix} H & C^* \\ C & 0 \end{pmatrix} \begin{pmatrix} p \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}, \tag{4.12}$$

kurz $Kz = \bar{b}$, ist eindeutig lösbar, falls C surjektiv und H positiv definit auf dem Kern von C ist (A.9). K selbst ist aber nicht positiv definit und lässt sich daher nicht mit dem gewöhnlichen CG-Verfahren lösen. Abhilfe könnte in der Anwendung eines (stabilisierten) bikonjugierten CG-Verfahrens bestehen. Dabei wird an Stelle von $Kz = \bar{b}$ das System $K^T K z = K^T \bar{b}$ gelöst, welches offensichtlich positiv definit ist. Gegen

diesen Ansatz spricht aber die drastische Verschlechterung der Kondition von K^TK gegenüber K und insbesondere H und die entsprechende Verlangsamung der Konvergenz. Es gibt zwei intuitive Ansätze, welche die Teilmatrizen des Systems gesondert betrachten. Die Schurkomplementmethode fußt auf einer direkten Elimination der Variablen. Unter der Voraussetzung der Invertierbarkeit von H kann x in der ersten Gleichung isoliert und in die zweite eingesetzt werden. Der Lagrange-Multiplikator λ löst dann

$$CH^{-1}C^*\lambda = CH^{-1}b. \tag{4.13}$$

Setzt man diesen wieder in die erste Gleichung ein, kann p aus

$$Hp = b - C^*\lambda \tag{4.14}$$

berechnet werden. In der linearen Algebra wird der Operator $CH^{-1}C^*$ als Schurkomplement von H in K bezeichnet. Die Methode eignet sich für Probleme, bei denen die Anzahl der Nebenbedingungen gering und H gut konditioniert und (leicht) invertierbar ist. In dieser Arbeit soll Invertierbarkeit von H nicht vorausgesetzt werden. Die bessere Alternative ist dann die sogenannte Nullraum-Methode. Für eine surjektive Matrix C bezeichne Z eine Matrix aus Basisvektoren von $\ker C$. Jedes Element $p \in \ker C$ ist dann eindeutig darstellbar durch $p = Zp^Z$ mit $p^Z \in \mathbb{R}^{n-m}$. Nun kann das gleichungsbeschränktes Problem unbeschränkt formuliert werden:

$$\begin{aligned}\min_{p^Z \in \mathbb{R}^{n-m}} \tilde{q}(p^Z) &= q(Zp^Z) \\ &= \frac{1}{2}(Zp^Z)^T H(Zp^Z) - b^T Zp^Z\rangle \\ &= \frac{1}{2}(p^Z)^T H_Z p^Z - (b_Z)^T p^Z,\end{aligned} \tag{4.15}$$

mit der reduzierten Hesse-Matrix $H_Z = Z^THZ$ und der reduzierten rechten Seite $b_Z = Z^Tb$. Offensichtlich ist mit H auch H_Z symmetrisch. Wenn H auf dem Kern von C positiv definit ist, gilt das auch für H_Z. Für die Wohldefiniertheit von (4.15) genügt also bereits diese schwächere Forderung an H. Das Minimierungsproblem kann nun mit dem gewöhnlichen vorkonditionierten CG gelöst werden, mit einem Vorkonditionierer $\mathcal{H}_Z = Z^T\tilde{H}Z$, der H_Z approximiert. Bei den Berechnungen am Ende dieser Arbeit wurde auf den Einsatz eines Vorkonditionierers verzichtet. Für eine mögliche Realisierung wird auf den Artikel [17] von Haber und Ascher verwiesen.

Der in (4.5) beschriebene Fehlerschätzer kann hier benutzt werden, da die Fehler in der reduzierten und in der nichtreduzierten Energienorm

zusammenfallen:

$$\begin{aligned}\|p_Z^k - p_Z^*\|_{H_Z}^2 &= \|p_Z^k - p_Z^*\|_{Z^T H Z} \\ &= \|Z p_Z^k - Z p_Z^*\|_H \\ &= \|p^k - p^*\|_H^2.\end{aligned}$$

Leider ist die Berechnung einer Basis Z meist aufwändig, gerade bei hochdimensionalen Problemen, wie sie bei diskretisierten partiellen DGL auftreten. Erfreulicherweise hat die Nebenbedingung dann häufig die Form $C = [A \;\; B]$ mit einer invertierbaren Matrix A. Hier bildet

$$Z = \begin{bmatrix} -A^{-1}B \\ I \end{bmatrix}$$

eine Basis von $\ker C$. Wenn dieser niedrigdimensional ist ($m \approx n$), kann eine explizite Berechnung in Betracht gezogen werden. Dieser Fall kann auftreten, wenn zum Beispiel Punkt- oder Randsteuerungen vorliegen. Das Problem (4.15) ist dann mit besonders wenig Aufwand zu lösen. Im Allgemeinen braucht man jedoch Z nicht explizit zu berechnen, es reicht aus, Z auf einen Vektor anwenden zu können. In jedem CG-Schritt wird eine Matrixmultiplikation mit der reduzierten Hesse-Matrix benötigt und dementsprechend je eine Anwendung von Z und Z^T.

Alternativ kann das Problem expandiert im Originalraum betrachtet werden. Hierfür kann der CG-Algorithmus 2 verwendet werden, unter der Voraussetzung, dass p_0 bereits in $\ker C$ liegt. Die Projektion in den Kern wird dann durch den Vorkonditionierer, angewendet auf die Abstiegsrichtung r_k, vollzogen:

$$\bar{r}_k = Z(Z^T M Z)^{-1} Z^T r_k, \qquad M \approx H.$$

Dieses Problem ist äquivalent zum Minimierungsproblem

$$\min_{\bar{r} \in \mathbb{R}^n} \tfrac{1}{2} \bar{r}^T M \bar{r} - \bar{r}^T r_k, \text{ so dass } C\bar{r} = 0,$$

dessen Lösung das System

$$\begin{pmatrix} M & C^T \\ C & 0 \end{pmatrix} \begin{pmatrix} \bar{r} \\ \mu \end{pmatrix} = \begin{pmatrix} r_k \\ 0 \end{pmatrix}$$

erfüllt. Ist $(CM^{-1}C^T)^{-1}$ mit verhältnismäßig wenig Aufwand anwendbar (etwa wenn M Diagonalgestalt hat), ist dieses System deutlich leichter zu lösen als das Originalproblem (4.12), zum Beispiel mit dem Schurkomplementansatz. Die beiden beschriebenen Prozesse sind äquvivalent und werden als Projiziertes Vokonditioniertes CG-Verfahren (PPCG) bezeichnet. In dieser Arbeit wird wegen der besseren Übersichtlichkeit der reduzierte, unbeschränkte Ansatz (4.15) gewählt.

4.6 Projektion in den echten Kern

Mit Hilfe des oben formulierten CG-Algorithmus kann preiswert ein inexakter Tangentialschritt auf dem gestörten Unterraum berechnet werden, der durch den Vorkonditionierer charakterisiert wird.

$$\tilde{t} \in \ker \tilde{C} = \ker[\tilde{A} \quad B].$$

In Kapitel 2.2 wurde erläutert, wie der exakte Tangentialschritt lokal optimiert, aber in erster Ordnung nichts an der Zulässigkeit ändert. Um das beschriebene Verhalten gewährleisten zu können, sollte die inexakte Suchrichtung in den korrekten Unterraum $\ker c'(x)$ projiziert werden. Mit den bereits eingeführten Bezeichnungen stellt sich das folgende Problem

$$\min_t \tfrac{1}{2}\langle t - \tilde{t}, M(t - \tilde{t})\rangle, \text{ so dass } Ct = 0,$$

mit einer geeigneten Matrix M. Das KKT-System hierfür

$$\begin{pmatrix} M & C^T \\ C & 0 \end{pmatrix} \begin{pmatrix} t \\ \lambda_t \end{pmatrix} = \begin{pmatrix} M\tilde{t} \\ 0 \end{pmatrix}$$

gleicht dem des Originalproblems (4.12). Die beschriebene Vorgehensweise macht nur dann Sinn, wenn sich dieses System leichter lösen lässt. Die intuitive Wahl der Blockdiagonalmatrix $M = \operatorname{diag}(M_h^Y, M_h^U)$ aus den Massenmatrizen der entsprechenden Funktionenräume Y und U erfüllt das im Allgemeinen nicht, zum Beispiel ist M_h^Y für $Y = H^1(\Omega)$ sehr schlecht konditioniert. Ein Kompromiss zwischen Konvergenzgeschwindigkeit und Problemangepasstheit bildet die Matrix $M = \operatorname{diag}(M_h^{L^2}, M_h^{L^2})$ aus den Massenmatrizen von $L^2(\Omega)$. Weitere Möglichkeiten sind $M = I_{\mathbb{R}^{2n}}$ oder die singuläre Wahl $M = \operatorname{diag}(0, I_{\mathbb{R}^n})$. Bei genauer Betrachtung der letzten Methode erkennt man, dass die inexakte Steuerung $\tilde{u}$ beibehalten und der projizierten Zustand y schlicht über die exakte Gleichung $y = -A^{-1}B\tilde{u}$ berechnet wird. Da nur eine einzige Anwendung von A^{-1} benötigt wird, erfordert diese Art der Projektion sicher den geringsten Aufwand. Diese Ersparnis wird aber durch den Verlust der bisherigen Informationen über y erkauft.

4.7 Äquilibrierung von Iterations- und Projektions-Fehler

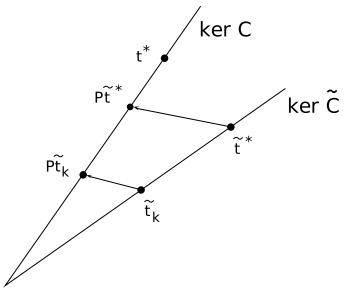

Abbildung 4.1: Projektion des inexakten Tangentialschrittes

Bei den bisherigen Überlegungen stammt die Inexaktheit im Tangentialschritt ausschließlich von der Störung zwischen den Räumen $\ker \tilde{C}$ und $\ker C$ ab. Da der Schritt $\tilde{t}$ aus einem iterativen Prozess $\{\tilde{t}_k\}$ gewonnen wird, welcher aus Aufwandsgründen möglichst früh abgebrochen werden soll, entsteht ein weiterer Beitrag $\|\tilde{t}^* - \tilde{t}_k\|$ zur Inexaktheit. Bei der Projektion $P\tilde{t}_k$ von $\ker \tilde{C}$ nach $\ker C$ entsteht unweigerlich ein Fehler, darum sollte versucht werden, den Iterations- und den Projektionsfehler zu äquilibrieren. Offenbar ist ein sinnvoller Abbruchzeitpunkt für die Iteration $\{\tilde{t}_k\}$ erreicht, sobald $\|\tilde{t}_k - \tilde{t}^*\|$ in die Größenordnung von $\|P\tilde{t}^* - t^*\|$ kommt, da jede Verbesserung darüber hinaus durch die Projektion größtenteils geschluckt wird. Diesem Gedanken folgend wurde versucht, eine obere Schranke für den Fehler $\|P\tilde{t}^* - t^*\|$ zu finden, welche nur von den Größen $\tilde{t}^*$, $P\tilde{t}^*$ sowie von der Metrik H des zu minimierenden Modells $m(t) = \frac{1}{2}\langle t, Ht\rangle - \langle b, t\rangle$ abhängt. Wie sich herausstellt ist das Problem jedoch unbeschränkt, was nun beschrieben werden soll.

Um unnötige Komplexität zu vermeiden, wird angenommen, dass das Modell streng konvex ist, so dass H auch wirklich eine Norm $\|H^{1/2} \cdot \|$ induziert. Außerdem können alle vorkommenden Längen und Winkel in der H-Norm und dem entsprechenden Skalarprodukt gemessen werden. Die Projektion in den echten Kern von C wird ebenfalls bezüglich dieses Skalarproduktes ausgeführt. Unter diesen Bedingungen kann das Problem mit kreisrunden Niveaulinien und echten rechten Winkeln skizziert werden, ohne dabei geometrische Informationen zu verlieren. In Abbildung 4.2 soll der Sachverhalt veranschaulicht werden. Die Skizze soll nicht nur als restringiertes Minimierungsproblem in $\mathbb{R}^2$ interpretiert werden, sondern liefert auch die korrekte Darstellung der Winkel und (gestrichelten) Niveaulinien eines n-dimensionalen Problems, eingeschränkt auf den zweidimensionalen Unterraum, der durch t und t^* aufgespannt wird (vorausgesetzt die Schnittmengen dieses Unterraums und $\ker C$ bzw. $\ker \tilde{C}$ sind tatsächlich eindimensional).

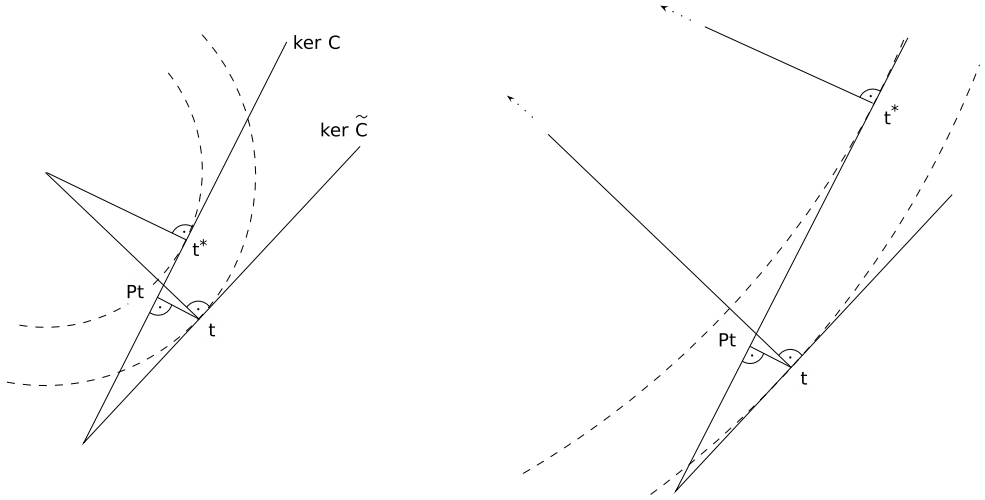

Abbildung 4.2: Der Fehler $\|t^* - Pt\|_H$ zwischen dem exakten und der Projektion des gestörten Optimums ist unbeschränkt

Es ist zu erkennen, dass die Inexaktheit $\|t^* - Pt\|_H$ von der Position des nichtrestringierten Minimierers $H^{-1}b$ abhängt. Je weiter dieser Punkt entfernt ist, umso größer wird auch der Fehler. In der Nähe einer Lösung des äußeren Problems sorgt der optimale Lagrange-Multiplikator eben dafür, dass der restringierte und der unrestringierte Minimierer nah beieinander liegen (vgl. Abb. 1.1, S. 8). Die hier durchgeführten Überlegungen zum Tangentialschritt sind jedoch eher in globalisierten Verfahren von Bedeutung, in denen es aufgrund der Nichlinearität der Nebenbedingung kaum möglich ist, einen derartigen Zusammenhang zu beschreiben.

Dessen ungeachtet gilt aufgrund der H-Projektion die Orthogonalitäts-Beziehung

$$\|\tilde{t}_k - P\tilde{t}_k\|_H^2 + \|P\tilde{t}_k - t^*\|_H^2 = \|\tilde{t}_k - t^*\|_H^2$$

und folglich

$$\|\tilde{t}_k - P\tilde{t}_k\|_H \leq \|\tilde{t}_k - t^*\|_H.$$

Dieser Idee folgend, soll die CG-Iteration $\{\tilde{t}_k\}$ abgebrochen werden, sobald

$$\|\tilde{t}_k - \tilde{t}^*\|_H \approx \|\tilde{t}_k - P\tilde{t}_k\|_H$$

erreicht wird. Natürlich ist der Term auf der rechten Seite im Vorhinein nicht bekannt. Aus diesem Grund sollen Informationen aus der vorhergehenden Composite-Step-Iteration verwendet werden. Es werden die analog definierten Bezeichnungen P^-, H^- und $\tilde{t}^-$ benutzt. Bei Verwendung desselben Vorkonditionierers soll folgende Schätzung angenommen wer-

den:

$$\frac{\|\tilde{t}_k - P\tilde{t}_k\|_H}{\|\tilde{t}_k\|_H} \approx \frac{\|\tilde{t}^- - P^-\tilde{t}^-\|_{H^-}}{\|\tilde{t}^-\|_{H^-}} =: \epsilon^-. \tag{4.16}$$

Insgesamt soll die Iteration abgebrochen werden, sobald

$$\frac{\|\tilde{t}_k - \tilde{t}^*\|_H}{\|\tilde{t}_k\|_H} \leq \epsilon^-.$$

Die Größe auf der linken Seite lässt sich bequem mit dem diskreten Energiefehlerschätzer aus Abschnitt 4.3.1 berechnen. In der Praxis zeigt sich eine sehr gute Übereinstimmung der beiden Seiten in (4.16). Durch die zusätzliche Normierung im Nenner hängen diese Quotienten kaum noch von P und H ab und können als Maß für den „Winkel“ zwischen den beiden Räumen interpretiert werden. Dieser ändert sich aber kaum, wenn in jeder Newton-Iteration der gleiche Vorkonditionierer verwendet wird.

4.8 Realisierung Composite-Step-Algorithmus

An dieser Stelle soll nun geklärt werden, wie die einzelnen Teilprobleme in Algorithmus 1 zu lösen sind.

Berechnung des Normalschritts

Der Normalschritt δn soll zum Einen die Zulässigkeit verbessern, d.h.

$$0 \approx \|c(x_k + \delta n)\| < \|c(x_k)\|\,, \tag{4.17}$$

zum Anderen sollte dafür jedoch nicht zu viel Aufwand betrieben werden, da der anschließende Tangentialschritt ohnehin wieder Einbußen bringt. Einen Kompromiss bildet das Minimierungsproblem erster Ordnung

$$\min_{\delta n} \frac{1}{2}\|\delta n\|_M^2, \quad \text{so dass } c'(x_k)\delta n + c(x_k) = 0, \tag{4.18}$$

mit einer positiv definiten Matrix M. Das resultierende Gleichungssystem

$$\begin{pmatrix} M & c'(x_k)^T \\ c'(x_k) & \end{pmatrix} \begin{pmatrix} \delta n \\ \lambda_{\delta n} \end{pmatrix} + \begin{pmatrix} 0 \\ c(x_k) \end{pmatrix} = 0$$

kann für einfache M über den Schurkomplementansatz aus Abschnitt 4.5 berechnet oder umformuliert durch ein Projiziertes CG-Verfahren gelöst werden. Dies soll aber nicht Thema dieser Ausarbeitung sein und

es wird für nähere Informationen auf das Buch [8] von Conn, Gould und Toint verwiesen. Für stark nichtlineare Nebenbedingungen kann der berechnete Normalschritt zu groß sein und sogar die Zulässigkeit verschlechtern. Es ist daher ratsam, eine Liniensuche anzuschließen, um (4.17) zu gewährleisten. In dem aktuellen Kontext benutzt man häufig eine Armijo-Strategie, wie sie etwa in [22] diskutiert wird. Für unsere Bedürfnisse soll eine einfache Backtracking-Liniensuche genügen.

Algorithmus 5 Backtracking-Liniensuche mit $\theta \in (0,1)$

$\alpha = 1$
while $\|c(x + \alpha\delta n)\| \geq \|c(x)\|$ **do**
  $\alpha = \theta\alpha$
end while

Sofern nicht schon $\|c(x_k)\| = 0$ erfüllt ist, terminiert der Algorithmus nach endlich vielen Schritten und gibt einen Dämpfungsfaktor α zurück, für den $\|c(x_k + \alpha\delta n)\| < \|c(x_k)\|$ gilt, da δn eine Abstiegsrichtung für $\|c(x_k)\|$ darstellt. Mit $c(x_k) = -c'(x_k)\delta n$ gilt nämlich

$$\begin{aligned}
\|c(x_k + \alpha\delta n)\| &= \|c(x_k) + c'(x_k)(\alpha\delta n) + \mathcal{O}(\alpha^2)\| \\
&= \|c(x_k) + \alpha c'(x_k)(\delta n) + \mathcal{O}(\alpha^2)\| \\
&= \|c(x_k) - \alpha c(x_k) + \mathcal{O}(\alpha^2)\| \\
&\leq (1-\alpha)\|c(x_k)\| + \mathcal{O}(\alpha^2) \\
&< \|c(x_k)\|,
\end{aligned}$$

für α hinreichend klein.

Berechnung des Tangentialschritts

Nach der Berechnung des Normalschritts δn_k soll durch den Tangentialschritt δt Abstieg in der Größe $L(x_k + \delta n_k + \delta t, \lambda_k)$ über dem Kern von $c'(x_k)$ erzielt werden. Es bezeichnen

$$\begin{aligned}
g &:= L_x(x_k, \lambda_k)^T, \\
H &:= L_{xx}(x_k, \lambda_k) \text{ und} \\
C &:= [A \;\; B] := [c_y(x_k) \;\; c_u(x_k)].
\end{aligned}$$

Zunächst wird Konvexität angenommen und versucht, das quadratische Modell

$$m_k(p) = L(x_k, \lambda_k) + g^T(\delta n_k + p) + \frac{1}{2}(\delta n_k + p)^T H(\delta n_k + p)$$

zu minimieren, wobei $Cp = 0$. Um den Aufwand so gering wie möglich zu halten, soll das System nicht auf dem echten Kern von C gelöst werden, sondern auf einem durch einen Mehrgittervorkonditionierer $\widetilde{A}^{-1} \approx c_y(x_k)^{-1}$ charakterisierten Kern von $\widetilde{C} := [\widetilde{A} \;\; B]$. Das resultierende KKT-System lautet dann

$$\begin{pmatrix} H & \widetilde{C}^T \\ \widetilde{C} & 0 \end{pmatrix} \begin{pmatrix} p \\ \lambda_p \end{pmatrix} + \begin{pmatrix} g + H\delta n_k \\ 0 \end{pmatrix} = 0\,.$$

Für die Berechnung wird ein PPCG-Verfahren mit einer Abbruchstrategie, wie in Abschnitt 4.3.2 *Truncated CG* versehen. Tritt in dieser Iteration ein negatives Energieprodukt auf, wird nicht nur die aktuelle Iterierte p, sondern auch die Richtung s^{neg} negativer Krümmung zurückgegeben. Damit der Tangentialschritt im echten Kern von C liegt, wie in Paragraph 2.2 über Composite-Step-Verfahren verlangt, soll p und gegebenenfalls s^{neg} in den echten Unterraum $\ker C$ projiziert werden (siehe S.41). Die resultierenden Vektoren seien mit $\bar{p}$ und $\bar{s}$ bezeichnet. Abschließend wird ein kubisches Modell des Lagrange-Funktionals, entwickelt an der Stelle $x_k^n := x_k + \delta n_k$, auf dem durch $\bar{p}$ bzw. $\bar{p}$ und $\bar{s}$ aufgespannten Unterraum $\bar{U}$ minimiert.

$$\delta t_k := \arg\min_{\delta t \in \bar{U}} \psi_\omega(\delta t)$$

mit

$$\psi_\omega(\delta t) = L_x(x_k^n, \lambda_k)\delta t + \tfrac{1}{2} L_{xx}(x_k^n, \lambda_k)(\delta t, \delta t) + \frac{\omega}{6}\|\delta t\|_M^3.$$

Der Wert ω kann mit einem der Schätzer aus Abschnitt 2.3 approximiert werden, mit $\underline{\delta x} = \bar{p}$ bzw. $\bar{p} + \bar{s}$.

Berechnung des Lagrange-Multiplikators

Die Berechnung des Lagrange-Multiplikators wird mit Hilfe der Methode der kleinsten Quadrate durchgeführt:

$$\lambda_k = \arg\min_{\lambda \in \mathbb{R}^n} \|J_x(x_k)^T + c_x(x_k)^T\lambda\|_M^2, \tag{4.19}$$

mit einer geeigneten Norm $\|.\|_M$ (diese sollte der Norm im Dualraum entsprechen). Eine Begründung für dieses Vorgehen findet sich in [28] und [8]. Besitzt die Ableitung $c_x(x_k)$ vollen Rang, dann ist die Lösung von (4.19) eindeutig bestimmt. In diesem Fall ist nämlich $C_k M C_k^T$ positiv definit, mit $C_k = c_x(x_k)$ und das Funktional

$$\begin{aligned} \|J_x(x_k)^T + C_k^T\lambda\|_M^2 &= (J_x(x_k)^T + C_k^T\lambda)^T M (J_x(x_k)^T + C_k^T\lambda) \\ &= J_x(x_k) M J_x(x_k)^T + \lambda^T C_k M C_k^T \lambda + 2 J_x(x_k) M C_k^T \lambda \end{aligned}$$

besitzt ein eindeutiges Minimum λ_k, welches die Gleichung

$$C_k M C_k^T \lambda_k = -C_k M J_x(x_k)^T \tag{4.20}$$

löst. Bei einem Optimum x^* des restringierten Minimierungsproblems stimmt der so berechnete Multiplikator mit jenem, der die KKT-Bedingungen erfüllt überein. Die Matrix M kann durch eine passend skalierte Identität cI ersetzt werden. Dabei kürzt sich in (4.20) der Skalierungsfaktor c heraus. Aus diesem Grund wird M bei der Implementierung nicht beachtet (d.h. als Identität auf $\mathbb{R}^m$ angesehen). Die notwendige Genauigkeit des Lagrange-Multiplikators soll nicht Thema dieser Arbeit sein. Demnach wird das Problem einfach mit dem gewöhnlichen CG-Verfahren ohne Vorkonditionierer ausiteriert.

Abstieg in der Merit-Funktion

In der vorliegenden Implementierung wird die ℓ_1-Merit-Funktion $\Phi(x, \mu) = J(x) + \mu\|c(x)\|_1$ benutzt. Um Abstieg zu garantieren, wurde eine Backtracking-Liniensuche wie in Algorithmus 5 umgesetzt, d.h.

$$\Phi(x_k + \alpha_k(\delta n_k + \delta t_k), \mu_k) < \Phi(x_k, \mu_k),$$

für $\alpha_k \in (0, 1)$ möglichst groß. In [22] werden Strategien zur adaptiven Bestimmung des Penalty-Parameters μ_k vorgeschlagen. In Beispiel 2 des folgenden Kapitels hat bereits die Wahl des konstanten Werts $\mu = 1$ zu einer akzeptablen Performanz geführt.

Kapitel 5

Numerische Beispiele und Resultate

Die bisherigen Ergebnisse sollen in diesem Kapitel an zwei Testproblemen veranschaulicht werden. Als Grundgebiet dient das Einheitsquadrat $\Omega = (0,1) \times (0,1)$ mit Rand $\partial\Omega$. Gesucht wird die Lösung des Minimierungsproblems

$$\min_{\substack{y \in H^1(\Omega) \\ u \in L^2(\Omega)}} J(y,u), \quad \text{so dass } c(y,u) = 0. \tag{5.1}$$

Die Nebenbedingung definiere sich jeweils durch das semilineare Randwertproblem

$$\begin{aligned} -\Delta y + y^3 + y &= u \qquad \text{in } \Omega, \\ n^T \nabla y + y &= \beta \qquad \text{auf } \partial\Omega, \end{aligned}$$

in der schwachen Formulierung

$$\begin{aligned} c(y,u)v &= \int_\Omega \nabla y^T \nabla v \, dt + \int_\Omega (y^3 + y) v \, dt + \int_{\partial\Omega} yv \, , ds - \int_\Omega uv \, dt - \int_{\partial\Omega} \beta v \, ds \\ &= 0, \qquad \forall v \in H^1(\Omega) \, . \end{aligned}$$

Für die Finite-Elemente-Diskretisierung wird Ω uniform in kongruente rechtwinklige Dreiecke zerlegt. Die Räume der Steuerungen $U = L^2(\Omega)$ und der Zustände $Y = H^1(\Omega)$ werden durch den selben Raum linearer Finiter Elemente dikretisiert, $V_h = Y_h = U_h$, mit der Lagrange-Basis $\mathcal{B}_h = \{\varphi_i | i = 1, \ldots, n\}$. Dessen Elemente werden bezeichnet durch

$$y_h = \sum_{i=1}^n y_i \varphi_i \, , \quad u_h = \sum_{i=1}^n u_i \varphi_i \, , \quad y_h^d = \sum_{i=1}^n y_i^d \varphi_i \, .$$

Der Einfachheit halber soll der nichtlineare Teil $f(y_h) = y_h^3 + y_h$ durch eine stückweise lineare Funktion $f_h(y_h)$ interpoliert werden mit

$$f_h(y_h) = \sum_{i=1}^{n} f_i \varphi_i \,, \quad f_i = f(y_i) = y_i^3 + y_i.$$

Analog wird der Randterm $\beta \in L^2(\Omega)$ ersetzt durch

$$\beta_h = \sum_{i=1}^{n} \beta_i \; \varphi_i|_{\partial\Omega} \,.$$

Dieses Vorgehen spart vielfache Neuberechnung der Integrale, geht jedoch auf Kosten des Diskretisierungsfehlers. Zwischen den Elementen im Funktionenraum $v_h \in V_h$ und deren Repräsentanten $v_h = (v_i)_{i=1,\dots,n} \in \mathbb{R}^n$ wird nicht unterschieden. Das soll der Übersichtlichkeit dienen und dürfte nicht zu Verwirrung führen. Die Massenmatrix M, die Steifigkeitsmatrix A und die Randmassenmatrix B sind definiert durch

$$M_{ij} = \int_\Omega \varphi_i \varphi_j \, dt \,, \quad A_{ij} = \int_\Omega \nabla \varphi_i^T \nabla \varphi_j \, dt \,, \quad B_{ij} = \int_{\partial\Omega} \varphi_i \varphi_j \, ds$$

und offenbar allesamt symmetrisch. Mit $v_h \in V_h$ kann nun die Nebenbedingung diskret formuliert werden,

$$c_h(y_h, u_h)v_h = y_h^T A v_h + y_h^T B v_h + f_h(y_h)^T M v_h - u_h^T M v_h - \beta_h^T B v_h,$$

bzw. über dem Dualraum

$$c_h(y_h, u_h)^T = (A + B)y_h + M f_h(y_h) - M u_h - B\beta_h.$$

Die Berechnungen wurden in Matlab, Version R2012a durchgeführt. Die Gittergenerierung sowie die Assemblierung der System- und Prolongationsmatrizen wurden mit Hilfe einer Finite-Elemente-Bibliothek durchgeführt, die im Rahmen der Vorlesung „Inside Finite Elements“ (Martin Weiser, FU Berlin, 2011) entstanden ist. Der verwendete Quellcode findet sich auf der Homepage des Vortragenden
http://www.zib.de/weiser/FEM-2011/.

Beispiel 1. Zur Demonstration des lokalen Konvergenzverhaltens wird zunächst ein quadratisches und konvexes Zielfunktional betrachtet:

$$J(y, u) = \tfrac{1}{2}\|y - y^d\|_{L^2}^2 + \tfrac{\sigma}{2}\|u\|_{L^2}^2,$$

Über den diskretisierten Funktionenräumen kann das Minimierungsproblem (5.1) nun rein algebraisch formuliert werden:

$$\min_{y_h, u_h \in \mathbb{R}^n} J(y_h, u_h) = \tfrac{1}{2}(y_h - y_h^d)^T M (y_h - y_h^d) + \tfrac{\sigma}{2} u_h^T M u_h,$$
$$\text{so dass } c_h(y_h, u_h) = 0.$$

Mit dem Lagrange-Multiplikator $\lambda \in \mathbb{R}^n$ erhält man das Lagrange-Funktional

$$L_h(y_h, u_h, \lambda) = J(y_h, u_h) + c_h(y_h, u_h)\,\lambda.$$

Die Ableitungen nach den Knotenwerten y_i, u_i sowie den Werten λ_i lauten

$$L_h'(y_h, u_h, \lambda)^T = \begin{pmatrix} M(y_h - y_h^d) + (A + B + MD_{f'})^T\lambda \\ \sigma M u_h - M\lambda \\ (A+B)y_h + Mf_h(y_h) - Mu_h - B\beta_h \end{pmatrix},$$

$$L_h''(y_h, u_h, \lambda) = \begin{pmatrix} M + D_{f''} & 0 & (A + B + MD_{f'})^T \\ 0 & \sigma M & -M \\ A + B + MD_{f'} & -M & 0 \end{pmatrix}$$

mit den Diagonalmatrizen $D_{f'}$ und $D_{f''}$ mit dem i-ten Diagonaleintrag $f'(y_i) = 3y_i^2 + 1$ bzw. $f''(y_i)\sum_j M_{ij}\lambda_j = 6y_i \sum_j M_{ij}\lambda_j$.

In diesem ersten Beispiel wird das gewöhnliche Newton-Verfahren eingesetzt. Dabei sollen die Effekte einer exakten und einer inexakten Implementierung verglichen werden, realisiert durch die Verwendung inexakter Hessematrizen. In jedem Schritt muss ein System der Art

$$K\delta z = r$$

gelöst werden, mit den Abkürzungen (im exakten Fall)

$$\begin{aligned} \delta z &= (\delta y, \delta u, \delta\lambda)^T, \\ r &= -L_h'(y_h, u_h, \lambda)^T = (r_1, r_2, r_3)^T \\ K &= L_h''(y_h, u_h, \lambda) = \begin{pmatrix} H_1 & & C_y^T \\ & H_2 & C_u^T \\ C_y & C_u & \end{pmatrix}. \end{aligned}$$

Der folgende heuristische Ansatz ist einem Composite-Step-Verfahren nachempfunden. Dabei wird der gesuchte Newton-Schritt in zwei Teile zerlegt, $\delta z = \delta n + \delta t$, mit $\delta n = (C_y^{-1} r_3, 0, 0)^T$. Nachdem δn bestimmt wurde, bleibt das folgende Problem zu lösen:

$$K\delta t = r - K\delta n = \begin{pmatrix} r_1 - H_1 C_y^{-1} r_3 \\ r_2 \\ 0 \end{pmatrix}.$$

Offenbar liegt der Schritt δt im Kern der linearisierten Nebenbedingungen und kann mit dem Projizierten CG-Verfahren berechnet werden, falls die Submatrix $H = \mathrm{diag}(H_1, H_2)$ positiv definit auf diesem Unterraum ist.

Für den inexakten Newton-Schritt werden δn und δt analog berechnet, bis auf die Anwendung von C_y^{-1}, welche durch den Mehrgittervorkonditionierer $\widetilde{C}_y^{-1}$ ersetzt wird. Dazu wurde der klassische Mehrgitteralgorithmus (S.36) implementiert mit einer gedämpften Jacobi-Iteration als

Glätter und dem Dämpfungsfaktor 1/2. Die notwendigen restringierten Level-Matrizen C_y^j, $j = 1, \ldots, j_{\max}$ sind rekursiv definiert über

$$C_y^{j-1} = I_j^{j-1} C_y^j I_{j-1}^j$$

mit $C_y^{j_{\max}} = C_y = A+B+MD_{f'}$ und den Prolongationen I_{j-1}^j von Gitterlevel $j-1$ nach j bzw. den Restriktionen $I_j^{j-1} = (I_{j-1}^j)^T$ von j nach $j-1$ (vgl. Abschnitt 4.4). Im Folgenden bezeichne $mg(i,j)$ die i-fach iterierte Anwendung des Vorkonditionierers mit je j Vor- und Nachglättungen.

Hier muss bemerkt werden, dass die so formulierte Ableitung $C_y = A+B+MD_{f'}$ nicht symmetrisch ist und die Anwendung des symmetrischen Mehrgitteralgorithmus nicht das antizipierte Verhalten haben muss. Eine Alternative stellt die symmetrisierte Matrix $\bar{C}_y = A + B + D_{f'}^{1/2} M D_{f'}^{1/2}$ dar. Diese Wahl führt jedoch zu keiner nennenswerten Änderung der hier präsentierten Ergebnisse. Das liegt wohl darin begründet, dass $C_y = A + B + MD_{f'}$ annähernd symmtrisch ist. In den durchgeführten Experimenten mit 1089 Knoten lag der Wert $\|C_y - C_y^T\|/\|C_y\|$ durchschnittlich in der Größenordnung 10^{-4}, bei feineren Gittern deutlich darunter und es ist zu erwarten dass dieser asymptotisch mit der Gitterweite verschwindet, da der zugrunde liegende Operator $c'(y,u)$ im unendlichdimensionalem Raum symmetrisch ist. Das gleiche gilt für das hiernach anschließende Beispiel 2.

In Beispiel 1 wurde keine Projektion in den echten Unterraum $\ker C$ durchgeführt. Der Lagrange-Multiplikators wird aus der ersten Zeile des inexakten KKT-Systems berechnet,

$$\lambda_k = \widetilde{C}_y^{-1}(r_1 - H_1 y_k).$$

Das beschriebene Problem hat sich als derart gutartig herausgestellt, dass das exakte und inexakte ungedämpfte Newton-Verfahren für fast alle getesteten Startwerte sowie moderat gewählte Regularisierungsparameter $\sigma \geq 10^{-6}$ konvergiert. Insbesondere war das quadratische Modell immer konvex auf dem Kern der linearisierten Nebenbedingungen. Für die folgenden Grafiken wurden die Startwerte $y_0 = u_0 = \lambda_0 = 10(1, \ldots, 1)^T$ und $\sigma = 1$ gewählt. Eine vollständige Auflistung der in Beispiel 1 verwendeten Parameter findet sich im Anhang B.1.

Die Abbildungen 5.1 - 5.3 zeigen die Reduktion der Residuen $\|L_h'(z_k)\|_{h^{-2}}$, der Längen der Newton-Schritte sowie die Kontraktion der Newton-Schritte als Maß für die lineare Konvergenzgeschwindigkeit. Anfänglich sind kaum Unterschiede festzustellen. Nach etwa sieben Iterationen geht das exakte Verfahren in die quadratische Konvergenzphase über. Nach wenigen weiteren Schritten ist dann bereits die vorgegebene Schwelle $\|L_h'(z_k)\|_{h^{-2}} \leq$

10^{-12} unterschritten. In Abb. 5.3 kann man den Eintritt der inexakten Iteration in die lineare Konvergenzphase sehr genau beobachten. Zunächst beschleunigt sich dabei die Kontraktion der Schrittweiten, schwächt sich danach aber rasch wieder ab, um ein weitgehend stabiles Niveau zu erreichen (vergleiche Satz 3.6). Weiterhin ist sehr gut zu erkennen, wie sich eine größere Genauigkeit des Vorkonditionierers auf die lineare Konvegenzgeschwindigkeit auswirkt.

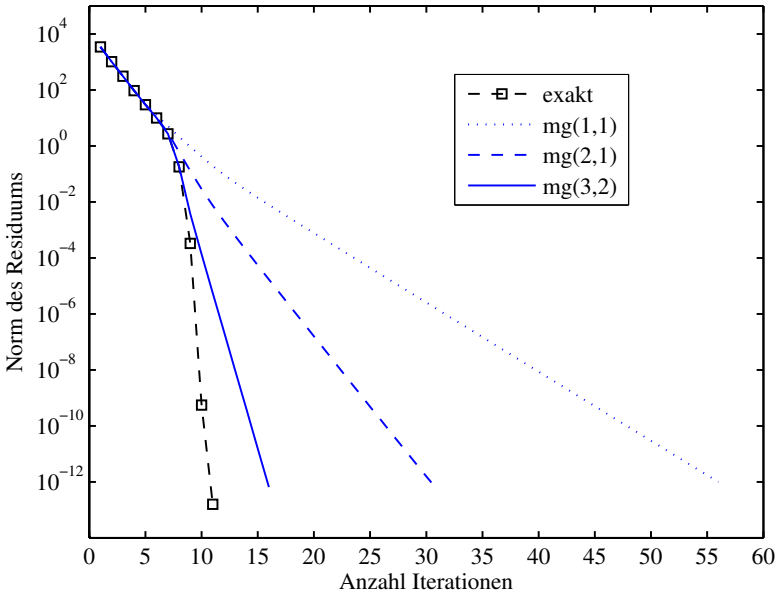

Abbildung 5.1: Vergleich der Residuen $\|L'(z_k)\|_{h^{-2}}$ des exakten und inexakten Newton-Lagrange-Verfahrens.

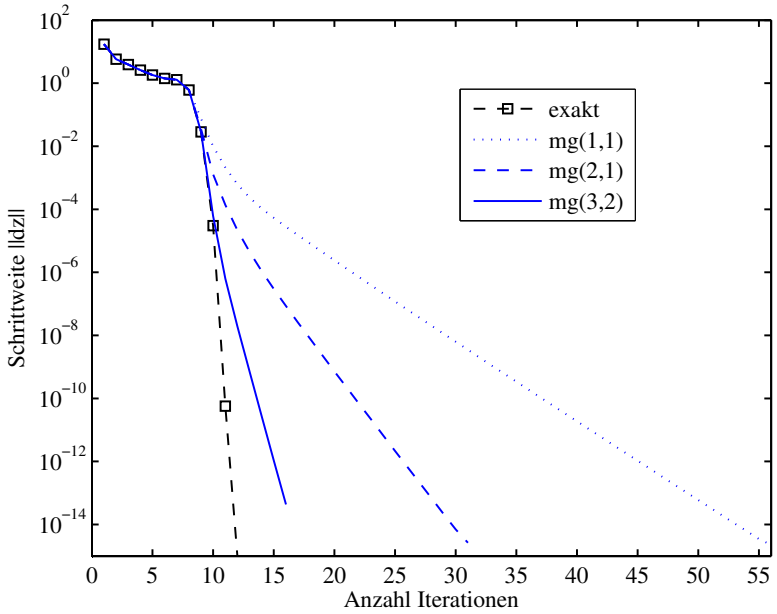

Abbildung 5.2: Vergleich $\|\Delta z_k\|_{h^2}$ bzw. $\|\delta z_k\|_{h^2}$ des exakten und inexakten Newton-Lagrange-Verfahrens.

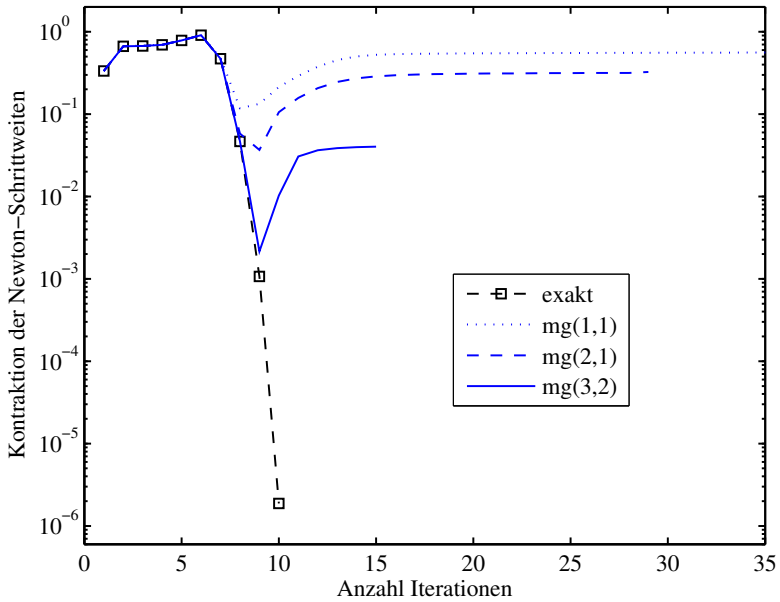

Abbildung 5.3: Vergleich der Kontraktion der Newton-Schrittweiten $\|\Delta z_{k+1}\|_{h^2}/\|\Delta z_k\|_{h^2}$ bzw. $\|\delta z_{k+1}\|_{h^2}/\|\delta z_k\|_{h^2}$ des exakten und inexakten Newton-Lagrange-Verfahrens.

Beispiel 2. Unter Beibehaltung der bisherigen Nebenbedingungen wird nun ein nichtkonvexes Zielfunktional herangezogen.

$$J(y,u) = \int_\Omega g(y-y^d)\,dt + \tfrac{\sigma}{2}\|u\|_{L^2}^2,$$

für $g(y) = \ln(\sqrt{y^2+\varepsilon}+1)$ mit einem fest gewählten $\varepsilon > 0$. Wie in der Nebenbedingung soll auch g hier durch eine lineare Interpolation g_h der Knotenwerte ersetzt werden:

$$g_h(y_h - y_h^d) = \sum_{i=1}^n g_i\varphi_i\ , \quad g_i = g(y_i - y_i^d).$$

Wegen der Identität

$$\int_\Omega g_h(y_h-y_h^d)\,dt = \sum_{i=1}^n g_i \int_\Omega \varphi_i\,dt = g_h^T v_{\mathcal{B}},$$

mit dem Vektor

$$v_{\mathcal{B}} = \sum_{i=1}^n \int_\Omega \varphi_i\,dt = M\,(1,\dots,1)^T,$$

gewinnt man das diskrete Optimalsteuerungsproblem

$$\min_{y_h,u_h\in\mathbb{R}^n} J_h(y_h,u_h) = g_h^T v_{\mathcal{B}} + \tfrac{\sigma}{2}u_h^T M u_h, \tag{5.2a}$$

$$\text{so dass}\ \ c_h(y_h,u_h) = 0. \tag{5.2b}$$

Für das entsprechende Lagrange-Funktional $L_h = J_h + c_h\lambda$ ergeben sich die Ableitungen

$$L_h'(y_h,u_h,\lambda)^T = \begin{pmatrix} D_{g'}v_{\mathcal{B}} + (A+B+MD_{f'})^T\lambda \\ \sigma M u_h - M\lambda \\ (A+B)y_h + Mf_h(y_h) - Mu_h - B\beta_h \end{pmatrix},$$

$$L_h''(y_h,u_h,\lambda) = \begin{pmatrix} D_{g''}+D_{f''} & 0 & (A+B+MD_{f'})^T \\ 0 & \sigma M & -M \\ A+B+MD_{f'} & -M & 0 \end{pmatrix} \tag{5.3}$$

mit den Diagonalmatrizen $D_{g'}$ und $D_{g''}$ mit dem jeweils i-ten Diagonaleintrag $g'(y_i-y_i^d)v_{\mathcal{B},i}$ bzw. $g''(y_i-y_i^d)v_{\mathcal{B},i}$ sowie den oben definierten Matrizen $D_{f'}$ und $D_{f''}$.

Bei der Berechnung des inexakten Tangentialschritts wird analog zu Beispiel 1 die benötigte Inverse des linken unteren Teils der KKT-Matrix C_y durch den Mehrgittervorkonditionierer $\widetilde{C}_y^{-1}$ ersetzt.

Für die Lösung des Optimierungsproblems (5.2) wurde das Composite-Step-Verfahren aus Abschnitt 4.8 eingesetzt. Mit den Parametern σ und ε kann die Größe nichtkonvexer Regionen beeinflusst werden. Es hat sich herausgestellt, dass für alle getesteten ε, zulässigen Startwerte mit $c_h(y_0, u_0) = 0$ und $\sigma > 10^{-4}$ das quadratische Modell an allen Iterierten konvex auf dem Nullraum der Nebenbedingung ist. Nichtkonvexität konnte erst erreicht werden für unzulässige Startwerte und $\sigma < 10^{-4}$. Aus diesem Grund wurden $y_0 = u_0 = 10(1, \ldots, 1)^T$ gewählt mit $c_h(y_0, u_0) \gg 0$ sowie $\varepsilon = 10^{-2}$ und $\sigma = 5 \cdot 10^{-5}$. Die Iteration wird abgebrochen, sobald entweder $\|L_h'(z_k)\|_{h^{-2}} \leq TOL$ oder die maximale Anzahl an Newton-Iterationen $K_{\max}$ erreicht wird. Im ersten Fall wird die Lösung als erfolgreich angesehen. Hier wurde $TOL = 10^{-12}$ und $K_{\max} = 70$ gewählt. Die Projektion des Tangentialschritts in den exakten Kern wurde außer in den Abbildungen 5.8 - 5.10 durch die Projektionsmatrix $I_{\mathbb{R}^{2n}}$ durchgeführt. Das ein- oder zweidimensionale kubische Minimierungsproblem für die Schrittweitenbestimmung des Tangentialschritts wurde in Matlab an die Methode `fminsearch` weitergereicht. Dabei wurde die Lipschitz-Konstante $[\omega]$ übergeben, mit

$$[\omega] = \max([\omega]_2, [\omega]_3),$$

für die ω-Schätzer aus Abschnitt 2.3.

Wie im ersten Beispiel kann man gut die Abhängigkeit der Konvergenzrate von der Genauigkeit des Tangentialschritts erkennen. Bei den Abbildungen 5.4 - 5.6 wurde dabei die Anzahl der Anwendungen und Glättungen des Mehrgitter-Vorkonditionierers variiert und der Berechnung der exakten Newton-Richtungen gegenübergestellt. In Abbildung 5.7 wurden dagegen der Vorkonditionierer beibehalten aber die Anzahl der Iterationen im Projizierten CG für die Tangentialschrittberechnung verglichen.

Obwohl das verwendetet Composite-Step-Verfahren nach Konstruktion monoton fallend in der Merit-Funktion ist (Abb. 5.5) gilt das in der präkonvexen Phase keineswegs für die Residuen und die Newton-Schrittweiten (Abb. 5.4 und 5.6).

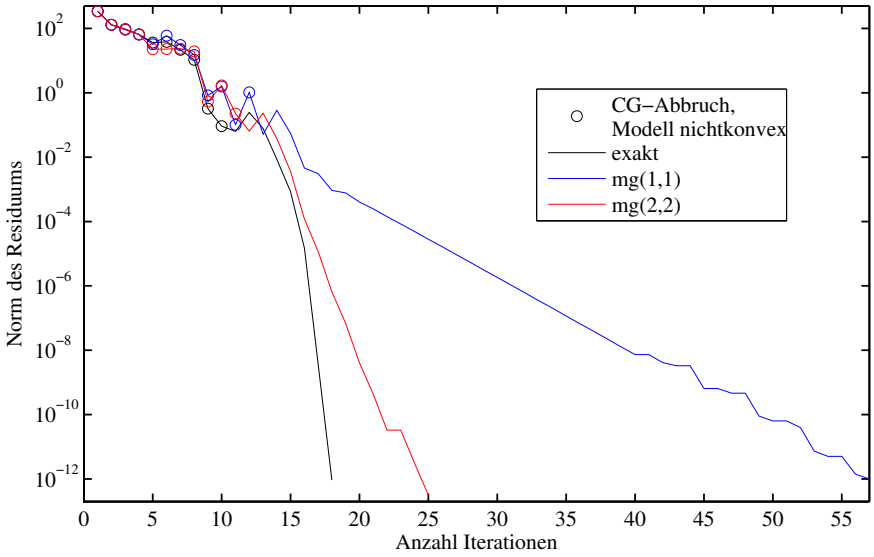

Abbildung 5.4: Konvergenz der Residuen $\|L'(z_k)\|_{h^{-2}}$ des exakten und inexakten Composite-Step-Verfahrens. Die eingekringelten Iterationen weisen darauf hin, dass das lokale Model nichtkonvex auf der linearisierten Nebenbedingung war, wonach das PPCG-Verfahren abgebrochen wurde.

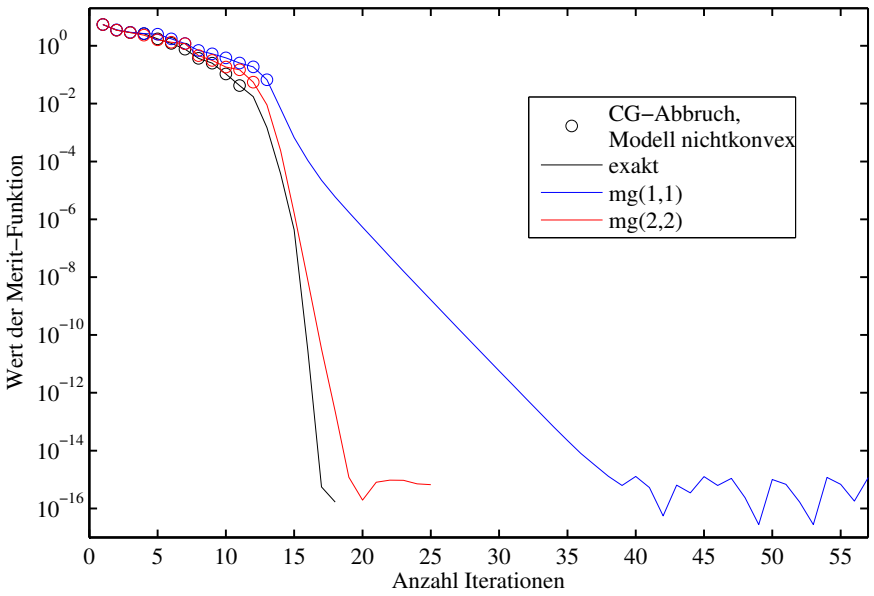

Abbildung 5.5: Wert der Merit-Funktion $\Phi(x_k, \mu) - \Phi(x^*, \mu)$ mit $\Phi(x, \mu) = J(x) + \mu\|c(x)\|_1$ und $\mu = 1$.

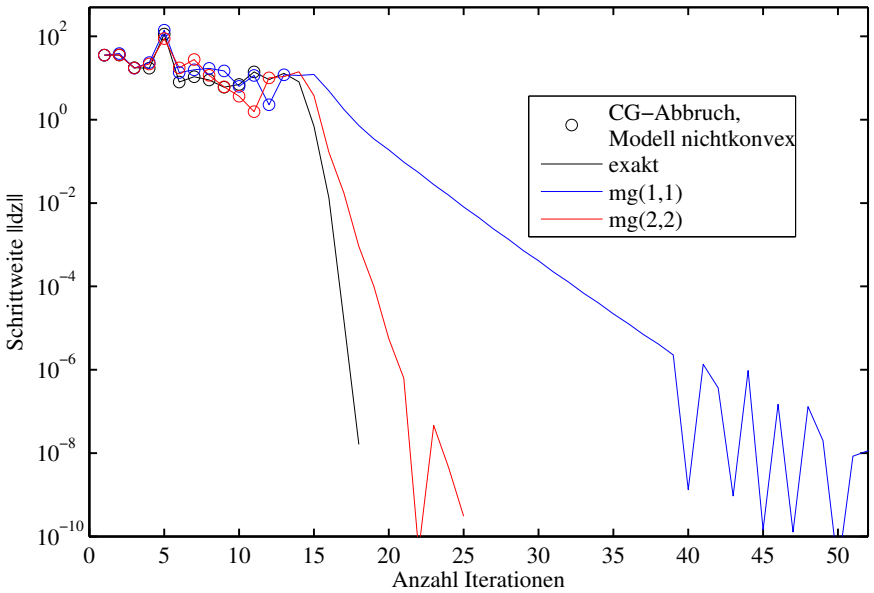

Abbildung 5.6: Entwicklung der Newton-Schrittweiten $\|\Delta z_k\|_{h^2}$ bzw. $\|\delta z_k\|_{h^2}$.

Das folgende Bild zeigt die Auswirkungen, wenn man das innere PPCG-Verfahren frühzeitig und unkontrolliert abbricht. Die Konvergenzgeschwindigkeit der Residuen bleibt zwar näherungsweise linear, allerdings mit deutlich schlechteren Kontraktionsraten, je früher der Abbruch stattfindet.

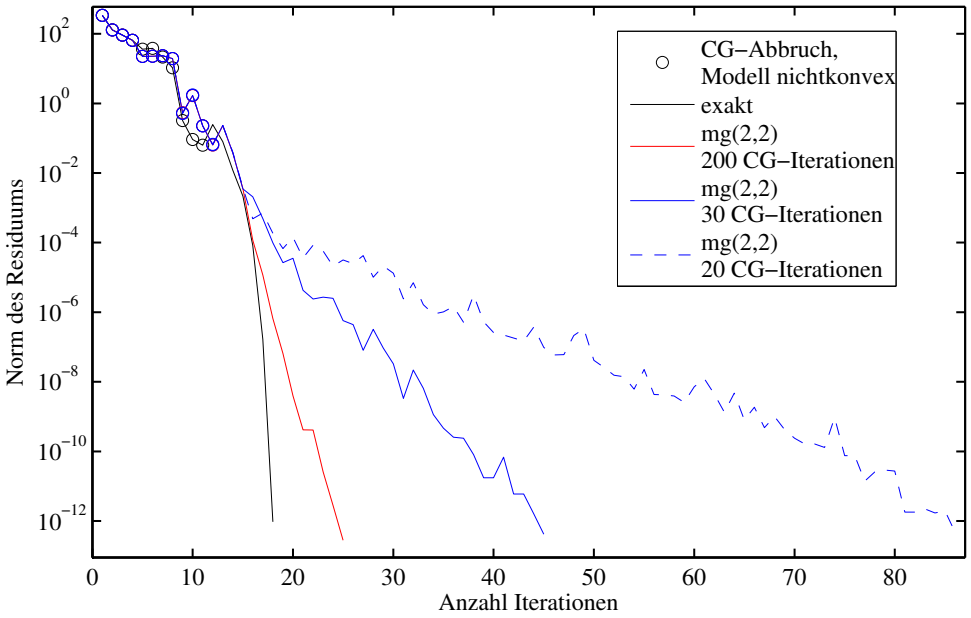

Abbildung 5.7: Konvergenz der Residuen $\|L'(z_k)\|_{h^{-2}}$ des exakten und inexakten Composite-Step-Verfahrens, wobei die CG-Iteration zur Tangentialschrittberechnung auf dem inexakten Kern nach 20, 30 bzw. 200 Iterationen abgebrochen wird.

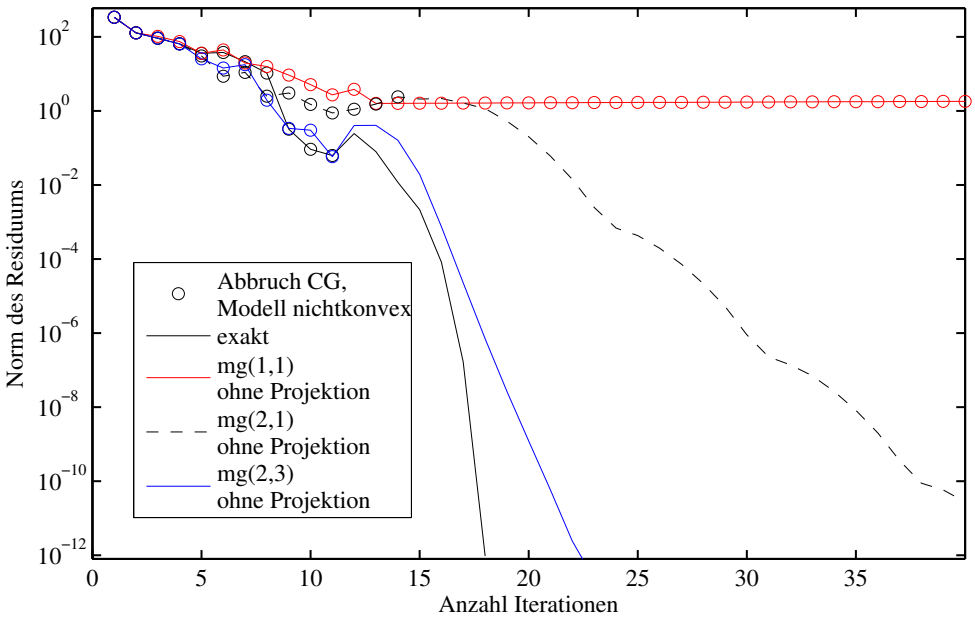

Abbildung 5.8: Je größer die Ungenauigkeit bei der Berechnung des inexakten Tangentialschritts ist, umso mehr Bedeutung kommt der Projektion in den echten Kern der linearisierten Nebenbedingungen zu.

Die Projektion des inexakten Tangentialschritts in den exakten Nullraum der Nebenbedingungen hatte im vorherigen Beispiel keine sichtbaren Auswirkungen auf den Iterationsprozess. Beim zweiten Funktional zeigt sich allerdings, dass das Unterlassen der Projektion die Konvergenz zum Erliegen bringen kann. In Abbildung 5.8 ist zu erkennen, dass das Residuum stagniert, falls der Tangentialschritt nicht schon so genau berechnet wurde, dass er dem Kern der Nebenbedingung ausreichend nahe kommt. Die Ursache dieses Effekts findet sich in der Verwendung der Merit-Funktion. Liegt der Tangentialschritt nicht im Nullraum, können sich die jeweiligen Verbesserungen des Normal- und Tangentialschritts gegenseitig auslöschen. Das bedeutet, dass nicht einmal lokal Abstieg in der Merit-Funktion garantiert werden kann und der kombinierte Schritt übermäßig gedämpft wird. Es wurde ein maximaler Dämpfungsfaktor 10^{-6} durch die Backtracking-Liniensuche implementiert. Aus diesem Grund sackt die Schrittweite der roten Iteration in Abbildung 5.9 nur ungefähr in dieser Größenordnung ab und verharrt dann auf dem gleichen Niveau.

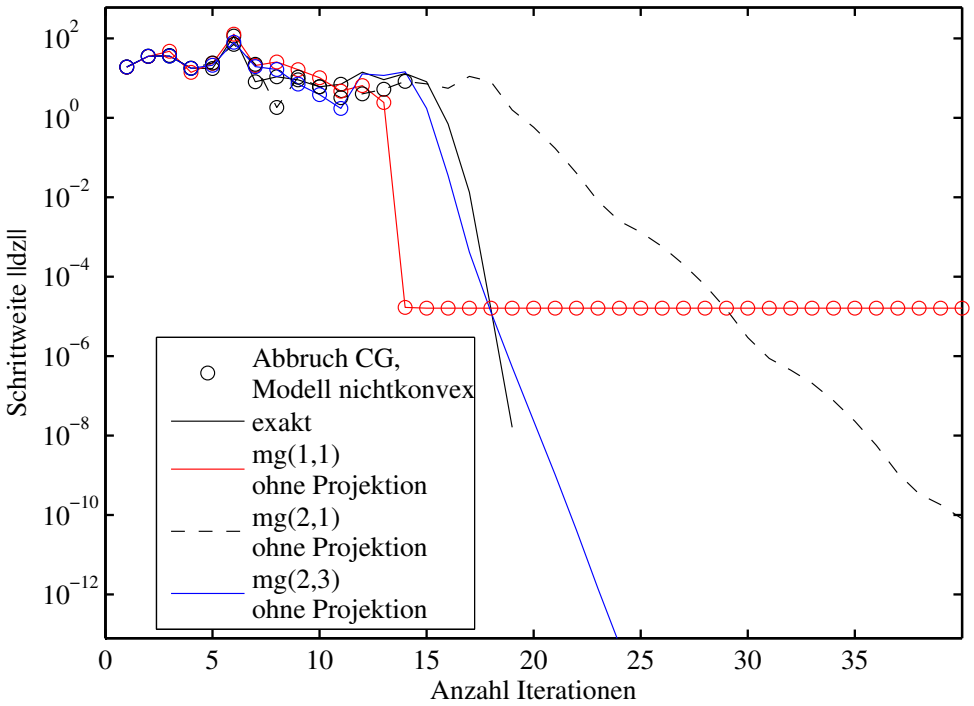

Abbildung 5.9: Wird die Projektion des inexakten Tangentialschritts unterlassen, sorgt die Merit-Funktion dafür, dass der Schritt übermäßig gedämpft und der Iterationsprozess dramatisch abgebremst wird.

Im nächste Diagramm ist zu erkennen, dass es von Vorteil ist, problemangepasste Projektionsmatrizen einzusetzen. Die unveränderlichen Projektionsmatrizen $P = \text{diag}(0, I_{\mathbb{R}^n})$ (grün), $P = I_{\mathbb{R}^{2n}}$ (blau) sowie $P = \text{diag}(M_h^{L^2}, M_h^{L^2})$ zeigen kaum Unterschiede im asymptotischen Verhalten. Die problemangepassten Matrizen $P = H(k) = L_{xx}(z_k)$ und $P = \text{diag}(H(k)) = \text{diag}(L_{xx}(z_k))$ legen dagegen eine etwa doppelt so hohe Konvergenzgeschwindigkeit an den Tag. Die gute Performanz der Diagonalmatrix $\text{diag}(L_{xx}(z_k))$ überrascht hier wenig, da $L_{xx}(z_k)$ nur aus einer dominanten Diagonalmatrix und einer sehr niedrig skalierten L^2-Massenmatrix besteht (siehe Gleichung 5.3).

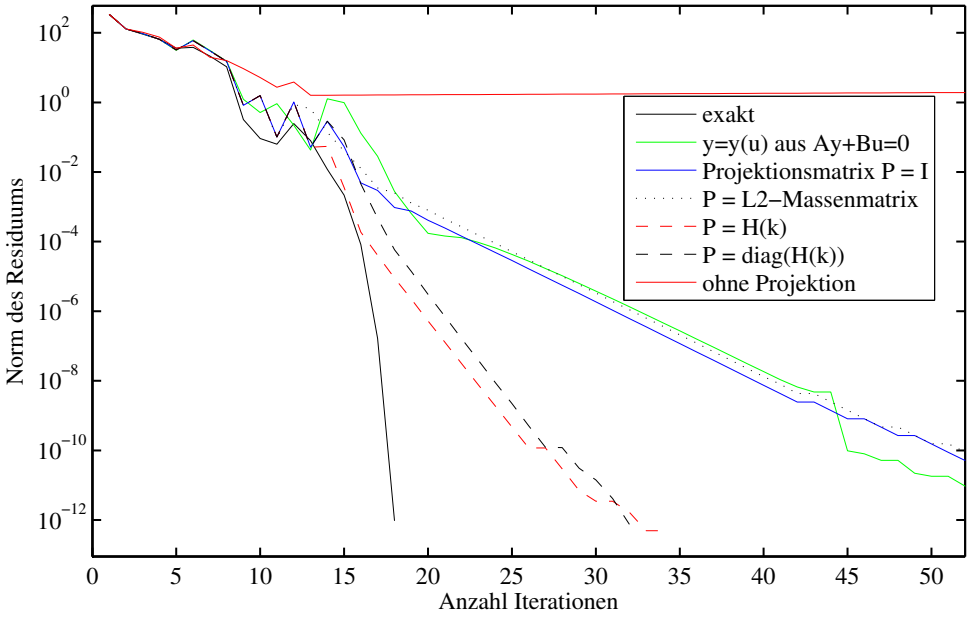

Abbildung 5.10: Hier charakterisiert der Vorkonditionierer mg(1,1) den inexakten Tangentialschritt. Die grüne Linie zeigt den Iterationsprozess, wenn man den Zustand direkt aus der inexakten Steuerung berechnet, gemäß $y = -C_y^{-1} C_u \tilde{u}$, was äquivalent zur Verwendung der Projektionsmatrix $P = \text{diag}(0, I_{\mathbb{R}^n})$ ist.

Abschließend wird das in Abschnitt 4.7 beschriebene adaptive Abbruchkriterium getestet. Dabei wird das PPCG-Verfahren abgebrochen, sobald der relative Energiefehler den relativen Projektionsfehler des vorgehenden Tangentialschritts unterschreitet:

$$\frac{\|\tilde{t}_k - \tilde{t}^*\|_H}{\|\tilde{t}_k\|_H} \leq \epsilon^- := \frac{\|\tilde{t}^- - P^-\tilde{t}^-\|_{H^-}}{\|\tilde{t}^-\|_{H^-}} \tag{5.4}$$

Die H-Normen sind durch Hesse-Matrizen der Lagrange-Funktion an der aktuellen und der vorhergehenden Iterierten definiert. Als Projektionsmatrix wurde dagegen immer die Identität in $\mathbb{R}^{2n}$ benutzt. Diese Wahl widerspricht zwar der theoretischen Motivation für das Abbruchkriterium (5.4), bringt aber dennoch eine beachtliche Ersparnis an teueren PPCG-Iterationen mit sich ohne die Konvergenzgeschwindigkeit merklich zu beeinflussen.

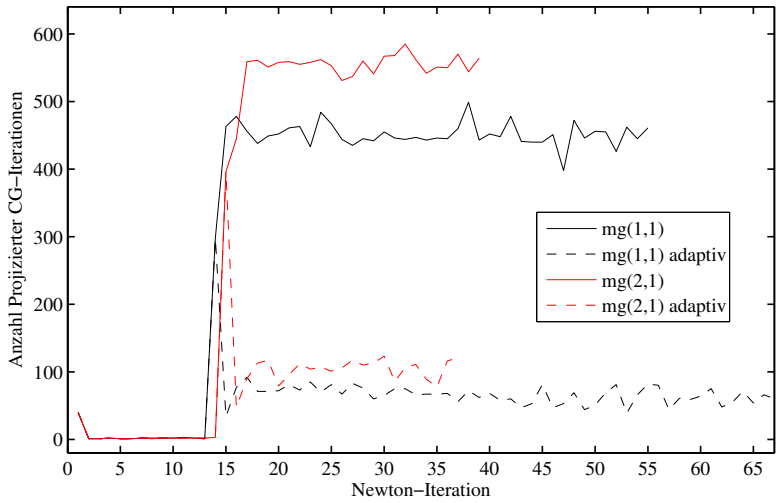

Abbildung 5.11: Die Anzahl innerer PPCG-Iterationen bei der Berechnung des Tangentialschritts. Die maximale Anzahl wurde hier auf 1000 gesetzt

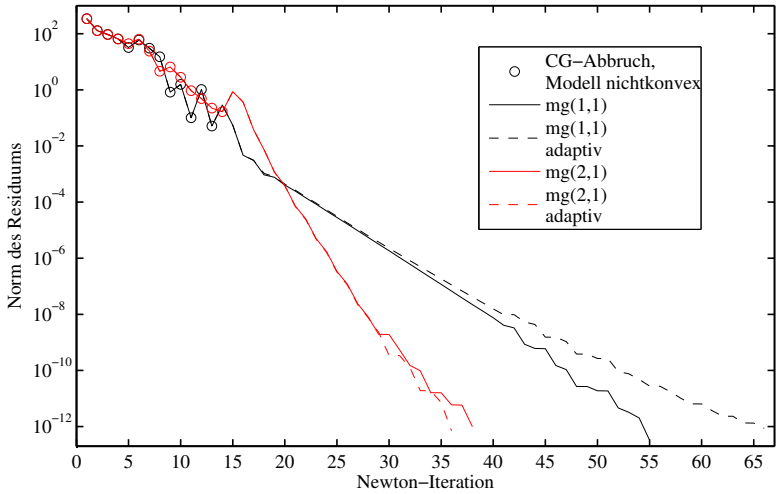

Abbildung 5.12: Die adaptive Abriegelung des inneren PPCG-Verfahrens hat trotz der beachtlichen Aufwandsreduktion kaum Einfluss auf die asymptotische Konvergenzgeschwindigkeit.

Kapitel 6

Schlussbemerkung

Es hat sich gezeigt, dass die theoretischen Aussagen über die lokale Konvergenzgeschwindigkeit des exakten und des inexakten Newton-Verfahrens in der Praxis sehr gut abgebildet werden. In der Nähe der Lösung sollte daher versucht werden sehr genau zu arbeiten und somit eine hohe lineare Konvergenzrate zu erzielen. In der präkonvexen Phase bringt es kaum Vorteile, viel Arbeit in die Bestimmung eines Schritts zu investieren. Es ist dann nur von Bedeutung, dass der Tangentialschritt im Nullraum der linearisierten Nebenbedingungen liegt. Die Art und Weise der Projektion scheint dabei ebenfalls erst in der asymptotischen Konvergenzphase relevant zu werden. Dann zeigt sich, dass problemangepasste Prjektionsmatrizen die Konvergenzgeschwindigkeit positiv beeinflussen. Eine mögliche Verbesserung des Composite-Step-Verfahrens könnte darin bestehen, die Qualität des Vorkonditionierers und der Projektionsmatrix adaptiv zu steuern.

Anhang A

Sätze und Definitionen

Lemma A.1 (Störungslemma von Banach, Lemma 2.3.2 in [24]). *Seien $A, C : V \to V$ lineare Operatoren auf dem Banachraum V. A sei invertierbar mit der Schranke $\|A^{-1}\| \leq \alpha$. Falls $\|A - C\| \leq \beta$ und $\alpha\beta < 1$, dann ist C ebenfalls invetierbar und es gilt*

$$\|C^{-1}\| \leq \frac{\alpha}{1 - \alpha\beta}\,.$$

Lemma A.2 (Lemma von Lax und Milgram, [20]). *Sei V ein reeller Hilbertraum auf dem eine stetige und V-elliptische Bilinearform $a(.,.)$ definiert ist, d.h. es existieren Konstanten α_0, $\beta_0 > 0$, so dass*

$$\begin{aligned} |a(y,v)| &\leq \alpha_0 \|y\|_V \|v\|_V && \forall y, v \in V && \textit{(Stetigkeit),} \\ a(y,y) &\geq \beta_0 \|y\|_V^2 && \forall y \in V && \textit{(V-Elliptizität).} \end{aligned}$$

Dann besitzt die Gleichung

$$a(y,v) = F(v), \quad \forall v \in V$$

für jedes stetige, lineare Funktional $F \in V^*$ eine eindeutige Lösung $y \in V$. Diese Lösung erfüllt

$$\|y\|_V \leq \frac{1}{\beta_0} \|F\|_{V^*}, \quad \forall v \in V\,.$$

Satz A.3 (Verallgemeinerte Friedrichs'sche Ungleichung, Lemma 2.5 in [29]). *Ist $\Omega \subset \mathbb{R}^n$ ein beschränktes Lipschitzgebiet mit Rand $\partial\Omega$, so existiert eine von $y \in H^1(\Omega)$ unabhängige Konstante c, so dass*

$$\|y\|_{H^1(\Omega)}^2 \leq c\Big(\int_\Omega \nabla y^T \nabla y \, dt + \int_{\partial\Omega} y^2 \, ds \Big)$$

für alle $y \in H^1(\Omega)$ erfüllt ist.

Definition A.4 (Lipschitz-Gebiet, [16]). *Ein beschränktes Gebiet Ω heißt Lipschitz-Gebiet, wenn es endlich viele offene Kugeln K_i gibt mit:*
(a) $\cup K_i \supset \partial\Omega$, $K_i \cap \partial\Omega \neq 0$.
(b) Es gibt eine Lipschitz-stetige Funktion $y = f^{(i)}(x)$, welche die Kugel K_i eineindeutig auf ein Gebiet im $\mathbb{R}^n$ abbildet. Dabei geht $\partial\Omega \cap \overline{K}_i$ in einen Teil der Ebene $y_n = 0$ und $\Omega \cap K_i$ in ein einfach zusammenhängendes Gebiet im Halbraum $\{y : y_n > 0\}$ über. Die Funktionaldeterminante

$$\frac{\partial(f_1^{(i)}(x), \ldots, f_n^{(i)}(x))}{\partial(x_1, \ldots, x_n)}$$

sei von Null verschieden für $x \in \overline{K}_i$.

Definition A.5 (Konvexität). *Die Menge X heißt konvex, falls für beliebige $s \in (0,1)$ und $x, y \in X$ gilt $x + s(y - x) \in X$. Eine Funktion $f : X \to \mathbb{R}$ heißt konvex, falls*

$$f(sx + (1-s)y) \leq sf(x) + (1-s)f(y), \qquad \forall s \in (0,1) \text{ und } \forall x, y \in X.$$

f ist strikt konvex, wenn die obige Gleichung für $<$ statt $\leq$ gilt. f heißt gleichmäßig konvex, falls ein $\lambda > 0$ existiert, so dass

$$f(y) \geq f(x) + f'(x)(y - x) + \tfrac{1}{2}\lambda\|y - x\|^2 \qquad \forall x, y \in X.$$

Definition A.6 (Fréchet-Differential). *Sei $F : U \subset X \to Y$ eine Abbildung zwischen Banachräumen X und Y mit einer nicht leeren, offenen Menge U. F heißt Fréchet-differenzierbar an der Stelle u, falls ein beschränkter, linearer Operator $A : U \to Y$ existiert, so dass*

$$\frac{\|F(u+h) - F(u) - Ah\|_Y}{\|h\|_X} \to 0, \qquad \text{für } \|h\|_X \to 0$$

für $h \in U$ gilt.

Satz A.7 (Taylor-Formel für Banachräume mit Restglied in Integralform). *Das Funktional $J : U \to Y$ sei $n+1$-mal stetig Fréchet-differenzierbar auf der offenen, konvexen Teilmenge U des Banachraums X. Dann gilt für alle x, $x + h \in U$*

$$J(x+h) - J(x) = \sum_{k=1}^{n} \tfrac{1}{k!}\, J^{(k)}(x)h^k + R_n(x)$$

mit dem Restglied

$$R_n(x) = \int_0^1 \tfrac{(1-t)^n}{n!}\, J^{(n+1)}(x + th)h^{n+1}\, dt$$

und der Bezeichnung

$$J^{(k)}(x)h^k = J^{(k)}(x)(h, \ldots, h)\,.$$

Definition A.8 (Konvergenzrate)**.** *Die Folge $\{x_k\}$ konvergiere gegen x^*. Die Konvergenz heißt superlinear, falls*

$$\|x_{k+1} - x^*\| = o(\|x_k - x^*\|), \quad \text{für } k \to \infty.$$

Die Konvergenz ist von Ordnung p (quadratisch für $p = 2$), falls

$$\|x_{k+1} - x^*\| = \mathcal{O}(\|x_k - x^*\|^p), \quad \text{für } k \to \infty.$$

Lemma A.9. *Falls C surjektiv ist und H positiv definit auf dem Kern von C, dann ist die KKT-Matrix*

$$K = \begin{pmatrix} H & C^T \\ C & 0 \end{pmatrix}$$

invertierbar.

Beweis. Sei $z = (x, y)^T$ eine Lösung der linearen Gleichungssystems $Kz = 0$. Dann ist $x \in \ker C$. Aus $Hx + C^T y = 0$ folgt

$$0 = x^T(Hx + C^T y) = x^T Hx + (Cx)^T y = x^T Hx,$$

was nur für $x = 0$ gelten kann. Es folgt $C^T y = 0$, also auch $\langle C^T y, v\rangle = \langle y, Cv\rangle = 0$ für alle v. Die Surjektivität von C zeigt schließlich $y = 0$ und damit $z = 0$. □

Anhang B

Parameterübersicht Beispiele

Zielfunktional und Nebendbedingung		
σ	Regularisierungsparameter Zielfunktional	1
y^d	Angestrebter Zustand Zielfunktional	$\sin(2\pi x_1)\cos(\pi x_2)$
β	Rechte Seite Robin-Randbedingung	$n^T \nabla y^d + y^d$
y_0, u_0, λ_0	Startwerte	$10 \cdot (1, \ldots, 1)^T$
Parameter für den Löser		
TOL	Toleranz Newton-Verfahren $\|L'(z_k)\|_{h^{-1}I}$	10^{-12}
$K_{\max}$	Maximale Anzahl Newton-Iterationen	70
tol	Toleranz Energiefehler PPCG	10^{-12}
$k_{\max}$	Maximale Anzahl PPCG-Iterationen	40
Verwendetes Gitter		
Ω	Grundgebiet	$(0,1) \times (0,1)$
n	Anzahl Gitterpunkte	1089
$j_{\max}$	Anzahl Gitterebenen	6
h	Maximaler Elementdurchmesser	$0,0442$

Tabelle B.1: Parameter für Beispiel 1.

Zielfunktional und Nebendbedingung		
σ	Regularisierungsparameter Zielfunktional	$5 \cdot 10^{-5}$
ε	Parameter Nichtkonvexität Zielfunktional	10^{-2}
y^d	Angestrebter Zustand Zielfunktional	$\sin(2\pi x_1)\cos(\pi x_2)$
β	Rechte Seite Robin-Randbedingung	$n^T \nabla y^d + y^d$
y_0, u_0, λ_0	Startwerte	$10 \cdot (1, \ldots, 1)^T$
Parameter für den Löser		
TOL	Toleranz Newton-Verfahren $\|L'(z_k)\|_{h^{-1}I}$	10^{-12}
$K_{\max}$	Maximale Anzahl Newton-Iterationen	70
tol	Toleranz Energiefehler PPCG	10^{-12}
$k_{\max}$	Maximale Anzahl PPCG-Iterationen	200
	Projektionsmatrix in echten Kern	$I_{\mathbb{R}^{2n}}$
μ	Penalty-Parameter für ℓ_1-Merit-Funktion	1
θ	Backtracking Parameter Normalschritt/Merit-Funktion	$0,5$
	Backtracking Maximale Anzahl Schrittreduktionen Normalschritt/Merit-Funktion	20
Verwendetes Gitter		
Ω	Grundgebiet	$(0,1) \times (0,1)$
n	Anzahl Gitterpunkte	1089
$j_{\max}$	Anzahl Gitterebenen	6
h	Maximaler Elementdurchmesser	$0,0442$

Tabelle B.2: Parameter für Beispiel 2.

Literaturverzeichnis

[1] W. Alt: *Nichtlineare Optimierung. Eine Einführung in Theorie, Verfahren und Anwendungen.* Vieweg, 2002.

[2] L. Angermann, P. Knabner: *Numerik partieller Differentialgleichungen. Eine anwendungsorientierte Einführung.* Springer, 2000.

[3] I. Argyros: *Computational Theory Of Iterative Methods.* Elsevier, 2007.

[4] A. Battermann: *Mathematical Optimization Methods for the Remediation of Ground Water Contaminations.* PhD thesis, Universtität Trier, 2000.

[5] G. Biros, O. Ghattas: *Parallel Lagrange-Newton-Krylov-Schur Methods for PDE-Constrained Optimization.* SIAM Journal of Scientific Computing 27 (2), 2005.

[6] S. C. Brenner, L. R. Ridgway: *The Mathematical Theory of Finite Element Methods* Springer, 1996.

[7] D. Braess: *Finite Elemente. Theorie, Schnelle Löser und Anwendungen in der Elastizitätstheorie.* Springer, 1997.

[8] A. R. Conn, N. I. L. Gould, P. L. Toint: *Trust-Region Methods.* MPS-SIAM Series on Optimization, 2000.

[9] F. E. Curtis: *Inexact Newton Methods and PDE-Constrained Optimization.* Präsentation in COPTA Lecture Series, Universität Wisconsin, Madison, April 2009.

[10] R. S. Dembo, S. C. Eisenstat, T. Steihaug: *Inexact Newton Methods.* SIAM Journal Numerical Analysis 19 (2), 1982.

[11] R. S. Dembo, T. Steihaug: *Truncated-Newton Algorithms for Large-Scale Unconstrained Optimization.* Mathematical Programmming 26, 1983.

[12] P. Deuflhard: *Global Inexact Newton Methods for Very Large Scale Nonlinear Problems.* IMPACT of Computing in Science and Engineering 3, 1991.

[13] P. Deuflhard: *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms.* Springer Series in Computational Mathematics, 2004.

[14] P. Deuflhard, A. Hohmann: *Numerische Mathematik. Eine algorithmisch orientierte Einführung.* De Gruyter, 1991.

[15] A. Griewank: *The Modification of Newton's Method for Unconstrained Optimization by Bounding Cubic Terms.* DAMTP, Universität Cambridge, Unveröffentlicht, 1981.

[16] C. Großmann, H. Roos: *Numerische Behandlung partieller Differentialgleichungen.* Teubner, 2005.

[17] E. Haber, U. M. Ascher: *Preconditioned All-At-Once Methods for Large, Sparse Parameter Estimation Problems.* Inverse Problems 17, 2001.

[18] W. Hackbusch: *Ein Iteratives Verfahren zur schnellen Auflösung Elliptischer Randwertprobleme.* Mathematisches Institut der Universität Köln, Report 76-12, 1976.

[19] W. Hackbusch, U. Trottenberg: *Multigrid Methods, Proceedings, Köln 1981.* Lecture Notes in Mathematics 960, Springer, 1982.

[20] M. Hinze, R. Pinnau, M. Ulbrich, S. Ulbrich: *Optimization with PDE Constraints.* Springer, 2009.

[21] K. Ito, K. Kunisch, V. Schulz and I. Ghermann: *Approximate Nullspace Iterations for KKT Systems in Model Based Optimization.* SIAM Journal of Matrix Analysis and Applications 31, 2010.

[22] J. Nocedal, S. J. Wright: *Numerical Optimization.* Springer, 1999.

[23] H. J. Oberle: *Optimierung.* Lecture notes, Universität Hamburg, 2011.

[24] J. M. Ortega, W. C. Rheinboldt: *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, London, 1970.

[25] H. H. Pennes: *Analysis of tissue and arterial blood temperatures in the resting human forearm.* Journal Applied Physics 1, 1948.

[26] Yousef Saad: *Iterative Methods for Sparse Linear Systems.* PWS, 2000.

[27] A. Schiela: *Interior Point Methods in Functionspace for State Constraints - Inexact Newton and Adaptivity.* ZIB-Report 09-01, Januar 2009.

[28] A. Schiela: *Nonlinear Optimization.* FU Berlin, Vorlesungsskript, 2010.

[29] F. Tröltzsch: *Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen.* Vieweg, 2005.

[30] M. Weiser: *On goal-oriented adaptivity for elliptic otpimal control problems.* ZIB-Report 09-08, Dezember 2009.

[31] M. Weiser, P. Deuflhard: *Numerische Mathematik 3. Adaptive Lösung partieller Differentialgleichungen.* De Gruyter, 2011.

[32] M. Weiser, P. Deuflhard, B. Erdmann: *Affine conjugate adaptive Newton methods for nonlinear elastomechanics.* Optimization Methods and Software 22 (3), 2007.

[33] M. Weiser, T. Gänzler, A. Schiela: *A Control Reduced Primal Interior Point Method for PDE Constrained Optimization.* ZIB-Report 04-38, September 2004.

[34] M. Weiser, A. Schiela, P. Deuflhard: *Asymptotic Mesh Independence of Newton's Method Revisited.* SIAM Journal Numerical Analysis 42 (5), 2005.

[35] T. J. Ypma: *Local Convergence of Inexact Newton Methods.* SIAM Journal Numerical Analysis 21 (3), 1984.

[36] H. Yserentant: *Old and New Convergence Proofs for Multigrid Methods.* Acta Numerica 1993.