Thema

Ein inexaktes Newton-Verfahren im Banachraum und seine Anwendung bei der Lösung semilinearer partieller Differentialgleichungen

D I P L O M A R B E I T

Technische Universität Chemnitz
Fakultät für Mathematik

eingereicht von **Frank Schmidt**
geb. am **10. Dezember 1984** in **Oschatz**

Betreuer: **Prof. Dr. C. Helmberg**
Dr. R. Griesse

Chemnitz, den **5. Dezember 2007**

Aufgabenstellung

In der Arbeit sollen zunächst aus der Literatur bekannte Resultate für das Newton-Verfahren mit inexakter Bestimmung des Newton-Schrittes zusammengestellt werden. Im Anschluss daran ist vorgesehen, das Verfahren zur Lösung einer semilinearen (elliptischen) partiellen Differentialgleichung einzusetzen. Hierbei soll untersucht werden, wie sich die erlaubten Inexaktheiten ausnutzen lassen, um das Problem effizient zu lösen. Angedacht ist insbesondere die Wahl der Diskretisierung, die die Genauigkeit des Newton-Schrittes beeinflusst, durch das Verfahren zu steuern. Es ist dabei die Verwendung eines a-posteriori-Fehlerschätzers und einer adaptiven Verfeinerung anzustreben.

Zusammenfassung

In dieser Arbeit wird ein inexaktes, affin kovariantes Newton-Verfahren im Banachraum vorgestellt. Neben unterschiedlichen Aussagen zur lokalen Konvergenz wird auch auf einen Ansatz zur Globalisierung eingegangen. Anhand einer semilinearen partiellen Differentialgleichung wird die numerische Nutzbarkeit belegt.

Inhaltsverzeichnis

Abbildungsverzeichnis **6**

Tabellenverzeichnis **7**

1 Einleitung **8**

2 Grundlagen **11**

2.1 Grundlagen im Banachraum 11

2.1.1 Ableitungen in Banachräumen 11

2.1.2 Integration im Banachraum 13

2.1.3 Der Satz von Banach 15

2.2 Sobolevräume 15

2.3 Einbettungen 21

2.4 Der Nemyzki-Operator 23

3 Der lokale Algorithmus **27**

3.1 Konvergenzaussagen 30

3.2 Algorithmische Umsetzung 38

3.2.1 Quadratischer Konvergenzmodus 38

3.2.2 Linearer Konvergenzmodus 41

4 Die globale Phase **43**

4.1 Theoretische Aussagen 43

4.2 Algorithmische Umsetzung 54

4.3 Ein naiver Globalisierungsansatz 57

5 Beispiel **58**

5.1 Problemstellung 58

5.2 Vorbetrachtungen für das Newton-Verfahren 61
5.3 Fehlerschätzer 66
5.3.1 Konstruktion des Fehlerschätzers 67
5.3.2 Diskussion der Konstanten 72
5.4 Verfeinerungsstrategie 74
5.5 Numerische Resultate 76
5.5.1 Konvergenz auf einem festen Gitter 77
5.5.2 Akzeptieren jedes Schrittes 78
5.5.3 Quadratischer Konvergenzmodus 79
5.5.4 Lineare Konvergenz 83

6 Zusammenfassung und Ausblick **85**

Literaturverzeichnis **87**

Danksagung **89**

Eidesstattliche Erklärung **90**

Abbildungsverzeichnis

2.1 Beispiele für Lipschitz-Gebiete 16
2.2 Beispiele für keine Lipschitz-Gebiete 16

3.1 Motivation für die Lipschitz-Bedingung 34
3.2 Die Menge $\left\{\delta x \in X : \frac{||\Delta x - \delta x||}{||\delta x||} \leq \delta\right\}$ mit $\delta = \frac{1}{2}$ 35
3.3 Algorithmus für den quadratischen Konvergenzmodus 40

4.1 Globaler Algorithmus 56

5.1 Gebiete ω_T und ω_E 69
5.2 Klassifizierung von Punkten 73
5.3 reguläre Verfeinerung 74
5.4 Verfeinerung nach [RS75] 76
5.5 Lösung $y^*(x)$ in Ω 77

Tabellenverzeichnis

5.1 Normales Newton-Verfahren auf einem festen Gitter mit 24833 Knoten . . 78
5.2 Erreichte Genauigkeit für verschiedene Gittergrößen 78
5.3 Iterationsverlauf, falls jeder Schritt akzeptiert wird 80
5.4 Quadratischer Konvergenzmodus mit $y_0 = 0$ 81
5.5 Quadratischer Konvergenzmodus, bei dem y_0 die Newton-Lösung auf einem Gitter mit 1601 Knoten ist 82
5.6 Linearer Konvergenzmodus mit $y_0 = 0$ 83
5.7 Linearer Konvergenzmodus, bei dem y_0 die Newton-Lösung auf einem Gitter mit 1601 Knoten ist 83

1 Einleitung

Bei der Modellierung vieler physikalischer Vorgänge in der Natur treten partielle Differentialgleichungen auf. Zum Beispiel lassen sich verschiedene Prozesse der Wärmeleitung oder der Strömungsmechanik mit ihnen beschreiben. Außerdem können sie als Nebenbedingung bei Aufgaben der optimalen Steuerung vorkommen.

Da häufig keine analytische Lösung solcher Probleme angegeben werden kann, wird auf numerische Näherungsverfahren zurückgegriffen. Dazu wird das Problem zunächst diskretisiert und anschließend gelöst. Eines der ältesten Näherungsverfahren zur Lösung solcher nichtlinearen Gleichungen ist das Newton-Verfahren, welches für stetig differenzierbare Funktionen $\mathrm{F} : \mathbb{R}^N \to \mathbb{R}^N$ leicht durchgeführt werden kann. Im Verfahren wird ein x^* mit $\mathrm{F}(x^*) = 0$ gesucht. Dazu linearisiert man die Funktion F an einem Startpunkt x_0, d.h. sie wird in eine Taylorreihe erster Ordnung entwickelt:

$$\mathrm{F}(x) \approx \mathrm{F}(x_0) + \mathrm{F}'(x_0)\,(x - x_0) \ .$$

Dabei bezeichnet $\mathrm{F}'(x_0)$ die Jacobi-Matrix von F an der Stelle x_0. Die Nullstelle der linearen Funktion auf der rechten Seite ist im Falle der Invertierbarkeit von $\mathrm{F}'(x_0)$ durch

$$x_{nst} = x_0 - \mathrm{F}'(x_0)^{-1}\,\mathrm{F}(x_0)$$

gegeben und wird als neuer Iterationspunkt x_1 verwendet. Mit $\Delta x_0 := -\,\mathrm{F}'(x_0)^{-1}\,\mathrm{F}(x_0)$ wird die Newton-Richtung bezeichnet. Aus der mehrfachen Wiederholung dieses Vorgehens mit dem aktuellen Iterationspunkt x_n an Stelle von x_0 resultiert eine Folge x_n mit der Bildungsvorschrift:

$$\begin{aligned} \Delta x_n &:= -\,\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n) \\ x_{n+1} &:= x_n + \Delta x_n \ . \end{aligned}$$

Bemerkenswert ist, dass die Skalierung der Funktion F zu A F mit einer invertierbaren Matrix $A \in \mathbb{R}^{N\times N}$ keinen Einfluss auf die Iterationsfolge x_n hat, denn es gilt:

$$x_{n+1} = x_n - \left(A\,F'(x_n)\right)^{-1} A\,F(x_n) = x_n - F'(x_n)^{-1}\,F(x_n)\ .$$

Diese Eigenschaft wird als die affine Invarianz des Newton-Verfahrens bezeichnet.

In einer Umgebung des Entwicklungspunktes x_n stimmt das lineare Modell annähernd mit der Funktion F überein. Unter geeigneten Voraussetzungen kann die quadratische Konvergenz der Folge x_n gegen eine Lösung x^* mit $F(x^*) = 0$ gezeigt werden. Zu diesen zählt neben einem kleinen Abstand von x_0 zu einer Lösung x^* auch eine Lipschitz-Bedingung. Die zum Newton-Verfahren bekannten Konvergenzaussagen hängen meist von der auftretenden Lipschitz-Konstante ab, wobei die ihr zugrunde liegende Lipschitz-Bedingung in ihrer Gestalt sehr variieren kann. Es existieren Lipschitz-Bedingungen, die keine affine Invarianz aufweisen. Das heißt, durch eine Skalierung von F mit einer invertierbaren Matrix $A \in \mathbb{R}^{N\times N}$ kann die Lipschitz-Konstante beliebig groß und infolgedessen die Aussagen einiger Sätze beliebig schwach werden. Um also die natürliche Eigenschaft der affinen Invarianz des Newton-Verfahrens auch in den zugehörigen Aussagen nutzen zu können, sind affin invariante Voraussetzungen notwendig.

In der Literatur existiert eine nahezu unüberschaubare Anzahl an Aussagen und Modifikationen des Newton-Verfahrens. Dazu gehören neben dem bekannten inexakten Newton-Verfahren ([DES82]) auch das Quasi-Newton-Verfahren (z.B. [Kel99]) sowie Aussagen zum Verfahren in allgemeinen Banachräumen (z.B. [KA64]). In Banachräumen versteht man unter der Funktion F eine Abbildung von einem Banachraum X in einen Banachraum Y. Außerdem geht die Jacobi-Matrix mit Hilfe der Fréchet-Ableitung in einen linearen Operator über.

Neben der Diskretisierung von Problemen besteht die Möglichkeit, sie in einem unendlichdimensionalen Funktionenraum zu formulieren. Dieser Ansatz ist für partielle Differentialgleichungen und für Aufgaben der optimalen Steuerung realisierbar. Eine Lösung x^* mit $F(x^*) = 0$ kann dann als schwache Lösung der Differentialgleichung oder bei Aufgaben der optimalen Steuerung als Tripel von Funktionen interpretiert werden. In diesen Fällen handelt es sich bei X um einen Funktionenraum und bei dem Bildraum Y um den Dualraum eines Funktionenraumes. Bei der numerischen Lösung ist es, selbst nach der Diskretisierung des Raumes X, im Allgemeinen nicht möglich, für ein Element x des

diskretisierten Raumes die als Supremum definierte Norm von $F(x)$ zu berechnen. Dies steht im direkten Zusammenhang mit der affinen Invarianz. Denn falls es erlaubt ist, die Funktion F mit einem linearen, invertierbaren Operator A zu skalieren, verliert die Norm in Y ihre Bedeutung. Sie kann durch die Skalierung fast beliebig verändert werden und ist somit für theoretische Aussagen ungeeignet.

Ein exaktes Newton-Verfahren im unendlich-dimensionalen Banachraum ist nicht von praktischer Bedeutung, denn seine Elemente lassen sich von einem Rechner nicht vollständig erfassen. Eine einfache Anpassung des inexakten Newton-Verfahrens von [DES82] für allgemeine Banachräume ist aufgrund der fehlenden affinen Invarianz für die genannten Problemstellungen ungeeignet. Abhilfe schafft hier zunächst das in [Ypm84] beschriebene inexakte, affin invariante Verfahren für den endlich-dimensionalen Fall. Die algorithmische Nutzung eines sehr ähnlichen Verfahrens ist in [Deu04] zu finden. Darüber hinaus werden darin drei weitere Invarianzklassen dargestellt, woraufhin die affine Invarianz in die affine Kovarianz umbenannt wurde.

Die vorliegende Arbeit erweitert einige algorithmisch nutzbare Aussagen aus [Deu04] und verallgemeinert diese auf Banachräume. Damit ist es möglich, die angesprochenen Probleme im Funktionenraum zu formulieren und beim numerischen Lösen dieser, durch eine geeignete Kontrolle des aus der Diskretisierung resultierenden Fehlers, Konvergenz im Funktionenraum zu erhalten.

2 Grundlagen

2.1 Grundlagen im Banachraum

In diesem Abschnitt werden zunächst einige bekannte Begriffe und Sätze der Analysis in endlich-dimensionalen Räumen auf Banachräume verallgemeinert. Im Folgenden werden mit X und Y stets reelle Banachräume und mit $||\cdot||_X$ bzw. $||\cdot||_Y$ ihre Normen bezeichnet. Die Darstellungen richten sich nach [KA64].

2.1.1 Ableitungen in Banachräumen

Definition 2.1 Existiert für die Funktion[1] $\mathrm{F} : X \to Y$ und $\delta x \in X$ der Grenzwert

$$\delta\,\mathrm{F}(x_0, \delta x) := \lim_{t \downarrow 0} \frac{\mathrm{F}(x_0 + t\delta x) - \mathrm{F}(x_0)}{t} \ ,$$

so ist $\delta\,\mathrm{F}(x_0, \delta x)$ die Richtungsableitung von F im Punkt $x_0 \in X$ in Richtung δx.
F heißt an der Stelle $x_0 \in X$ Fréchet-differenzierbar, wenn es einen beschränkten[2] linearen Operator $\mathrm{A} = \mathrm{A}(x_0) \in \mathcal{L}(X, Y)$ mit

$$\lim_{\triangle x \to 0} \frac{||\mathrm{F}(x_0 + \triangle x) - \mathrm{F}(x_0) - \mathrm{A}\,\triangle x||_Y}{||\triangle x||_X} = 0$$

gibt[3]. In diesem Fall ist A die Ableitung von F an der Stelle x_0 und wird mit $\mathrm{F}'(x_0)$ bezeichnet.
F heißt in X Fréchet-differenzierbar, wenn F für alle $x \in X$ Fréchet-differenzierbar ist.

[1] Die Begriffe Funktion und Operator werden gleichwertig verwendet.
[2] Es sei daran erinnert, dass die Begriffe Beschränktheit und Stetigkeit für lineare Operatoren äquivalent sind.
[3] In der Literatur existieren Definitionen, die auf die Beschränktheit von A verzichten.

Ist F auf X Fréchet-differenzierbar, so ist die Ableitung F' eine Abbildung von X in den Raum der linearen stetigen Operatoren (d.h. $\mathrm{F}' : X \to \mathcal{L}(X, Y)$).
Eine Fréchet-differenzierbare Funktion bzw. ihre Ableitung hat ähnliche Eigenschaften, wie sie aus dem endlich-dimensionalen Fall bekannt sind. Einige wichtige Eigenschaften sind in der folgenden Bemerkung festgehalten.

Bemerkung 2.2

(i) Eine im Punkt x_0 Fréchet-differenzierbare Funktion F *ist in x_0 stetig, denn für eine Folge x_n mit $||x_n - x_0||_X \to 0$ gilt:*

$$||\mathrm{F}(x_n) - \mathrm{F}(x_0)||_Y \leq \big|\big|\mathrm{F}'(x_0)(x_n - x_0)\big|\big|_Y + ||\mathrm{r}(x_0, x_n - x_0)||_Y$$

mit

$$\mathrm{r}(x_0, x_n - x_0) = \mathrm{F}(x_n) - \mathrm{F}(x_0) - \mathrm{F}'(x_0)(x_n - x_0)$$

und

$$||\mathrm{r}(x_0, x_n - x_0)||_Y \to 0 \quad \textit{für} \quad ||x_n - x_0||_X \to 0 \ .$$

Zusammen mit

$$\big|\big|\mathrm{F}'(x_0)(x_n - x_0)\big|\big|_Y \leq \big|\big|\mathrm{F}'(x_0)\big|\big|_{\mathcal{L}(X,Y)} ||(x_n - x_0)||_X \to 0 \quad \textit{für} \quad ||x_n - x_0||_X \to 0$$

folgt damit auch $||\mathrm{F}(x_n) - \mathrm{F}(x_0)||_Y \to 0$ und somit die Stetigkeit von F *in x_0.*

(ii) Es sei $\mathrm{F} = \alpha_1 \mathrm{F}_1 + \alpha_2 \mathrm{F}_2$ ($\alpha_1, \alpha_2 \in \mathbb{R}$). Falls F_1 und F_2 in x_0 Fréchet-differenzierbar sind, so ist auch F *in x_0 Fréchet-differenzierbar und es gilt $\mathrm{F}'(x_0) = \alpha_1 \mathrm{F}_1'(x_0) + \alpha_2 \mathrm{F}_2'(x_0)$.*

(iii) Falls F *linear und beschränkt ist (d.h. $\mathrm{F} \in \mathcal{L}(X, Y)$), so ist* F *in X Fréchet-differenzierbar und es gilt $\mathrm{F}'(x_0) = \mathrm{F}$.*

(iv) Für Fréchet-differenzierbare Funktionen gilt die Kettenregel.

Definition 2.3 Es sei $\mathrm{F} : X \to Y$ in einer Umgebung von x_0 Fréchet-differenzierbar. Ist zusätzlich die Abbildung

$$X \ni x \mapsto \mathrm{F}'(x) \in \mathcal{L}(X, Y)$$

an der Stelle x_0 stetig, dann heißt F in x_0 stetig Fréchet-differenzierbar. F heißt stetig Fréchet-differenzierbar in X, falls F für alle $x \in X$ stetig Fréchet-differenzierbar ist.

Lemma 2.4 *Ist eine Funktion* $\mathrm{F} : (X, ||\cdot||_X) \to (Y, ||\cdot||_Y)$ *in* $x_0 \in X$ *Fréchet-differenzierbar, so kann die Norm* $||\cdot||_X$ *durch eine stärkere Norm* $||\cdot||_{\tilde{X}}$ *(d.h.* $||x||_X \leq c\,||x||_{\tilde{X}}$*) und die Norm* $||\cdot||_Y$ *durch eine schwächere Norm* $||\cdot||_{\tilde{Y}}$ *(d.h.* $||y||_{\tilde{Y}} \leq \bar{c}\,||y||_Y$*) ersetzt werden, ohne dass sich etwas an der Differenzierbarkeit von* F *ändert. Das heißt,* $\mathrm{F} : \left(X, ||\cdot||_{\tilde{X}}\right) \to \left(Y, ||\cdot||_{\tilde{Y}}\right)$ *ist in* x_0 *Fréchet-differenzierbar und hat dieselbe Ableitung, denn es gilt:*

$$\begin{aligned} & \frac{||\mathrm{F}(x_0 + \triangle x) - \mathrm{F}(x_0) - \mathrm{A}\,\triangle x||_{\tilde{Y}}}{||\triangle x||_{\tilde{X}}} \\ \leq \quad c\ \bar{c}\ & \frac{||\mathrm{F}(x_0 + \triangle x) - \mathrm{F}(x_0) - \mathrm{A}\,\triangle x||_Y}{||\triangle x||_X} \to 0 \quad \textit{für} \quad \triangle x \to 0\ . \end{aligned}$$

Eine analoge Aussage gilt auch für die stetige Fréchet-Differenzierbarkeit.

Die zuvor genannte Definition 2.1 eignet sich auch für Ableitungen höherer Ordnung. Der Raum $\mathcal{L}(X, Y)$ ist mit der induzierten Operatornorm wieder ein Banachraum (weil Y ein Banachraum ist). Für die zweite Ableitung von F in x_0 gilt $\mathrm{F}''(x_0) \in \mathcal{L}(X, \mathcal{L}(X, Y))$ bzw. $\mathrm{F}'' : X \to \mathcal{L}(X, \mathcal{L}(X, Y))$. Zweite und höhere Ableitungen werden im Rahmen dieser Arbeit nicht benötigt.

2.1.2 Integration im Banachraum

Definition 2.5 Es sei $\mathrm{F} : [a, b] \to Y$ mit $[a, b] \subset \mathbb{R}$ und Y ein Banachraum. Das Integral der Funktion F wird als Grenzwert der Integralsumme definiert, d.h.

$$\int_a^b \mathrm{F}(t)\ \mathrm{d}t := \lim_{\substack{a=t_0<t_1<\cdots<t_n=b \\ \max\limits_k (t_{k+1}-t_k)\to 0}} \quad \sum_{k=0}^{n-1} \mathrm{F}(t_k)(t_{k+1} - t_k)\ ,$$

falls die rechte Seite existiert.

Analog zur Fréchet-Ableitung besitzt auch das hier definierte Integral viele der aus der endlich-dimensionalen Analysis bekannten Eigenschaften.

Bemerkung 2.6

(i) Ist $\mathrm{F} : [a, b] \to Y$ *stetig, so existiert das hier definierte Integral.*

(ii) Für ein integrierbares F*, einen weiteren Banachraum* Z *und* $\mathrm{A} \in \mathcal{L}(Y, Z)$ *gilt:*

$$\int_a^b \mathrm{A}\left(\mathrm{F}(t)\right)\ \mathrm{d}t = \mathrm{A}\left(\int_a^b \mathrm{F}(t)\ \mathrm{d}t\right)\ .$$

(iii) Für F *integrierbar gilt:*

$$\left|\left|\int_a^b \mathrm{F}(t)\ \mathrm{d}t\right|\right|_Y \leq \int_a^b ||\mathrm{F}(t)||_Y\ \mathrm{d}t\ .$$

Neben Definition 2.5 gibt es außerdem die Möglichkeit, das Integral auf stückweise konstanten Funktionen zu definieren und es anschließend auf den Raum der integrierbaren Funktionen fortzusetzen.

In der folgenden Definition wird der eingeführte Integralbegriff für eine weitere Klasse von Funktionen erklärt.

Definition 2.7 Es sei der Operator $\mathrm{A} : X \to \mathcal{L}(X, Y)$ und $\triangle x \in X$ gegeben. Das Integral $\int_{x_0}^{x_0+\triangle x} \mathrm{A}(x)\ \mathrm{d}x$ wird definiert als

$$\int_{x_0}^{x_0+\triangle x} \mathrm{A}(x)\ \mathrm{d}x := \int_0^1 \mathrm{A}(x_0 + t\triangle x)\triangle x\ \mathrm{d}t\ .$$

Ist A stetig, so existiert das eben definierte Integral.

Der nächste Satz ist ein wichtiges Resultat, um die Konvergenz des Newton-Verfahrens zu beweisen. Er stellt eine Verallgemeinerung des endlich-dimensionalen Hauptsatzes der Differential- und Integralrechnung dar.

Satz 2.8 *Es sei* $\mathrm{F} : X \to Y$ *eine Fréchet-differenzierbare Funktion mit einer in* $[x_1, x_2] := \{x \in X : x = x_1 + t\,(x_2 - x_1)$ *für* $t \in [0, 1]\}$ *stetigen Ableitung* F'*. Dann gilt:*

$$\int_{x_1}^{x_2} \mathrm{F}'(x)\ \mathrm{d}x = \mathrm{F}(x_2) - \mathrm{F}(x_1)\ .$$

2.1.3 Der Satz von Banach

Satz 2.9 (**Satz von Banach**) *Es sei $U \in \mathcal{L}(X,X)$ mit $||U||_X \leq q < 1$. Dann hat $I-U$ eine beschränkte Inverse und es gilt*

$$\left|\left|(I-U)^{-1}\right|\right|_X \leq \frac{1}{1-q} \ .$$

Definition 2.10 $\bar{\mathcal{L}}(X,Y)$ bezeichnet den Raum der linear beschränkten Operatoren, deren Inverse existiert und beschränkt ist. Das heißt:

$$\bar{\mathcal{L}}(X,Y) := \left\{ \mathrm{A} \in \mathcal{L}(X,Y) : \mathrm{A}^{-1} \text{ existiert und } \mathrm{A}^{-1} \in \mathcal{L}(Y,X) \right\} \ .$$

Satz 2.11 *Für $U_0 \in \bar{\mathcal{L}}(X,Y)$ hat $V = U_0 + U$ mit $||U||_{\mathcal{L}(X,Y)} < \frac{1}{\left|\left|U_0^{-1}\right|\right|_{\mathcal{L}(Y,X)}}$ eine beschränkte Inverse V^{-1}, so dass V in $\bar{\mathcal{L}}(X,Y)$ liegt und es gilt*

$$\left|\left|V^{-1}\right|\right|_{\mathcal{L}(Y,X)} \leq \frac{\left|\left|U_0^{-1}\right|\right|_{\mathcal{L}(Y,X)}}{1-\left|\left|U_0^{-1}U\right|\right|_{\mathcal{L}(X,X)}} \leq \frac{\left|\left|U_0^{-1}\right|\right|_{\mathcal{L}(Y,X)}}{1-\left|\left|U_0^{-1}\right|\right|_{\mathcal{L}(Y,X)} ||U||_{\mathcal{L}(X,Y)}} \ .$$

Folgerung 2.12 *Es sei $\mathrm{F} : X \to Y$ eine Fréchet-differenzierbare Funktion mit $\mathrm{F}'(x_0) \in \bar{\mathcal{L}}(X,Y)$ und F' stetig in x_0. Dann existiert eine Umgebung $B_\delta(x_0) := \{x \in X : ||x||_X \leq \delta\}$ um x_0, so dass $\mathrm{F}'(x)$ für alle $x \in B_\delta(x_0)$ beschränkt invertierbar ist.*

2.2 Sobolevräume

Um später mit den schwachen Lösungen von Differentialgleichungen arbeiten zu können, ist es notwendig, Sobolevräume zu definieren. Dafür werden der Gaußsche Integralsatz und die Formel der partiellen Integration benötigt. Diese gelten nur auf „schönen“ Gebieten Ω. Auch für die Friedrichsche Ungleichung und die Spurabbildung sind Gebiete mit gewisser Regularität erforderlich wie beispielsweise die Lipschitz-Gebiete. Die Ausführungen des verbleibenden Teils dieses Kapitels orientieren sich an [Trö05] und [Gri07].

Definition 2.13 Ein offenes, beschränktes Gebiet $\Omega \subset \mathbb{R}^N$ mit Rand Γ gehört zur Klasse der $C^{k,1}$-Gebiete ($k \in \mathbb{N}_0$), falls sich Γ in endlich viele Abschnitte Γ_i zerteilen

lässt, so dass jedes Γ_i der Graph einer auf einem $N-1$ dimensionalen Würfel k-mal stetig differenzierbaren Funktion mit Lipschitz-stetigen Ableitungen ist. Des Weiteren darf Ω lokal stets nur auf einer Seite des Randes liegen. Gebiete der Klasse $C^{0,1}$ heißen Lipschitz-Gebiete.

Diese Definition ist mathematisch nicht exakt formuliert[1]. In dieser Arbeit wird jedoch keine genauere benötigt.

Beispiele für Lipschitz-Gebiete sind in Abbildung 2.1 dargestellt. Bei den abgebildeten Gebieten in 2.2 handelt es sich jedoch um keine Lipschitz-Gebiete.

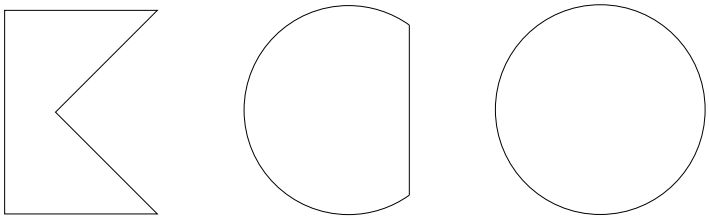


Abbildung 2.1: Beispiele für Lipschitz-Gebiete

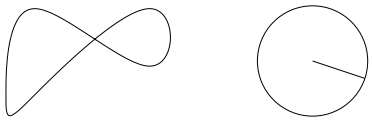


Abbildung 2.2: Beispiele für keine Lipschitz-Gebiete

Definition 2.14 Für eine Menge E und p mit $1 \leq p < \infty$ werden die Lebesgue-Räume $L^p(E)$ durch

$$L^p(E) = \{\mathrm{f} : E \to \overline{\mathbb{R}} : \mathrm{f}\,\text{messbar und} \int\limits_E |\mathrm{f}(x)|^p \ \mathrm{d}x < \infty\}$$

definiert und mit der Norm $||\mathrm{f}||_{L^p(E)} = \left(\int\limits_E |\mathrm{f}(x)|^p \ \mathrm{d}x\right)^{\frac{1}{p}}$ versehen. Zusätzlich wird der

[1] Eine exaktere Definition kann z.B. in [Trö05] gefunden werden.

Raum $L^\infty(E)$ durch

$$L^\infty(E) = \{\mathrm{f} : E \to \overline{\mathbb{R}} : \mathrm{f\ messbar\ und\ } \operatorname*{ess\,sup}_{x\in E} |\mathrm{f}(x)| < \infty\}$$

erklärt und mit der Norm

$$||\mathrm{f}||_{L^\infty(E)} = \operatorname*{ess\,sup}_{x\in E} |\mathrm{f}(x)| := \inf_{\mathrm{N\ Nullmenge}} \sup_{x\in E\setminus N} |\mathrm{f}(x)|$$

versehen. Für $k \in \mathbb{N} \cup \{\infty\}$ wird mit $\mathcal{C}^k(E)$ die Menge der k-mal stetig differenzierbaren Funktionen auf E bezeichnet und mit $\mathcal{C}_0^k(E)$ die Menge der k-mal stetig differenzierbaren Funktionen mit kompaktem Träger.

Für ein Lipschitz-Gebiet $\Omega \subset \mathbb{R}^N$ (mit Abschluss $\overline{\Omega}$) und die Funktionen $y, v \in \mathcal{C}^1(\overline{\Omega})$ gilt die Formel der partiellen Integration, d.h.

$$\int_\Omega v(x) D_i y(x)\ \mathrm{d}x = \int_{\partial\Omega} v(x) y(x) n_i(x)\ \mathrm{d}s - \int_\Omega y(x) D_i v(x)\ \mathrm{d}x\ .$$

Hierbei bezeichnet $n_i(x)$ die i-te Komponente der nach außen gerichteten Einheitsnormalen $n(x)$ im Punkt $x \in \partial\Omega$, $\mathrm{d}s$ das Lebesguesche Oberflächenmaß auf $\partial\Omega$ und $D_i = \frac{\partial}{\partial x_i} = D^{(0,\ldots,0,1,0,\ldots,0)}$. Für einen Multiindex $\alpha = (a_1, \ldots, a_N)$ bezeichnet D^α den Differentialoperator $\frac{\partial^{|\alpha|}}{(\partial x_1)^{a_1}\ldots(\partial x_N)^{a_N}}$ und ist für $|\alpha|$-mal stetig differenzierbare Funktionen nach dem Satz von Fubini wohldefiniert.
Gilt zusätzlich $v = 0$ auf $\partial\Omega$, so folgt:

$$\int_\Omega y(x) D_i v(x)\ \mathrm{d}x = -\int_\Omega v(x) D_i y(x)\ \mathrm{d}x\ .$$

Nach mehrmaligem Anwenden dieser Beziehung ergibt sich für $y \in \mathcal{C}^k(\overline{\Omega})$ sowie $v \in \mathcal{C}_0^k(\Omega)$ und $|\alpha| \le k$

$$\int_\Omega y(x) D^\alpha v(x)\ \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega v(x) D^\alpha y(x)\ \mathrm{d}x\ .$$

Daraus motiviert ist eine Verallgemeinerung des klassischen Ableitungsbegriffs auf die „große“ Menge der $L^1_{loc}(\Omega)$ Funktionen. Die Menge der lokal integrierbaren Funktionen

$L^1_{loc}(\Omega)$ ist definiert als

$$L^1_{loc}(\Omega) := \left\{ y : \Omega \to \overline{\mathbb{R}} : y \in L^1(K) \text{ für alle } K \subset \Omega \text{ kompakt} \right\} .$$

Definition 2.15 Falls für eine Funktion $y \in L^1_{loc}(\Omega)$ und einen Multiindex α eine Funktion $w \in L^1_{loc}(\Omega)$ existiert, so dass

$$\int_\Omega y(x) D^\alpha v(x)\, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega v(x) w(x)\, \mathrm{d}x \qquad \text{für alle } v \in \mathcal{C}_0^\infty(\Omega)$$

gilt, heißt w die schwache Ableitung der Ordnung α von y und wird auch mit $w = D^\alpha y$ bezeichnet.

Die schwache Ableitung bildet also die Formel der partiellen Integration nach. Mit Hilfe der schwachen Ableitung ist es nun möglich, die Sobolevräume zu definieren.

Definition 2.16 Für p mit $1 \leq p < \infty$ und $k \in \mathbb{N}_0$ definiert man den Sobolevraum $W^{k,p}(\Omega)$ als

$$W^{k,p}(\Omega) := \{ y \in L^p(\Omega) : D^\alpha y \text{ existiert und } D^\alpha y \in L^p(\Omega) \text{ für alle } \alpha \text{ mit } |\alpha| \leq k \}$$

und versieht ihn mit der Norm

$$||y||_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha| \leq k} ||D^\alpha y||^p_{L^p(\Omega)} \right)^{\frac{1}{p}} .$$

Die Räume $W^{k,p}(\Omega)$ mit der Norm $||\cdot||_{W^{k,p}(\Omega)}$ sind Banachräume. Für $p = 2$ handelt es sich sogar um Hilberträume, die mit $H^k(\Omega) := W^{k,2}(\Omega)$ bezeichnet werden.

Bei partiellen Differentialgleichungen spielen häufig die Randwerte einer Funktion eine besondere Rolle. Für partielle Differentialgleichungen mit homogenen Dirichlet-Randbedingungen, wie beispielsweise die in Kapitel 5 behandelte, ist es von Vorteil, die Räume $W^{k,p}_0(\Omega)$ zu definieren und zu verwenden.

Definition 2.17 Mit $W^{k,p}_0(\Omega)$ bezeichnet man den Abschluss von $\mathcal{C}_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$ (bzgl. $||\cdot||_{W^{k,p}(\Omega)}$). Dieser Raum wird ebenfalls mit der Norm $||\cdot||_{W^{k,p}(\Omega)}$ versehen und ist

(per Definition) ein abgeschlossener Teilraum von $W^{k,p}(\Omega)$. Auch in diesem Fall werden für $p = 2$ die Räume $H_0^k(\Omega)$ als $H_0^k(\Omega) := W_0^{k,2}(\Omega)$ definiert.

Für Funktionen aus $W_0^{k,p}(\Omega)$ gilt, dass die Randwerte aller schwachen Ableitungen bis zur Ordnung $k-1$ verschwinden. Für inhomogene Dirichlet-Randbedingungen ist die später definierte Spurabbildung wichtig.

Von besonderem Interesse ist der Raum $H_0^1(\Omega)$. Dieser ist als abgeschlossener Teilraum von $H^1(\Omega)$ selbst ein Hilbertraum mit dem Skalarprodukt:

$$(y, v)_{H^1(\Omega)} = \int_\Omega yv \, \mathrm{d}x + \int_\Omega \nabla y \cdot \nabla v \, \mathrm{d}x \ .$$

Dieses induziert die $H^1(\Omega)$-Norm $||\cdot||_{H^1(\Omega)}$

$$||y||_{H^1(\Omega)} = \sqrt{(y,y)_{H^1(\Omega)}} = \left(\int_\Omega \left(y^2 + |\nabla y|^2 \right) \, \mathrm{d}x \right)^{\frac{1}{2}} \ .$$

Auf dem Raum $H_0^1(\Omega)$ lässt sich ein weiteres Skalarprodukt durch

$$(y, v)_{H_0^1(\Omega)} = \int_\Omega \nabla y \cdot \nabla v \, \mathrm{d}x$$

definieren. Dies erzeugt die $H_0^1(\Omega)$-Norm

$$||y||_{H_0^1(\Omega)} = \sqrt{(y,y)_{H_0^1(\Omega)}} = \left(\int_\Omega |\nabla y|^2 \, \mathrm{d}x \right)^{\frac{1}{2}} \ .$$

Auf dem ganzen Raum $H^1(\Omega)$ bildet sie nur eine Halbnorm, da eine konstante Funktion $c \in H^1(\Omega)$ die „Norm“ 0 hätte, aber im $H^1(\Omega)$-Sinne nicht der Nullfunktion entspricht. Sie wird daher häufig als $H^1(\Omega)$-Seminorm bezeichnet. Auf dem Raum $H_0^1(\Omega)$ ist die $H_0^1(\Omega)$-Norm äquivalent zur $H^1(\Omega)$-Norm, denn es gilt:

$$||y||^2_{H^1(\Omega)} = ||y||^2_{L^2(\Omega)} + ||\nabla y||^2_{L^2(\Omega)} \leq \left(1 + c_\Omega^2\right) ||y||^2_{H_0^1(\Omega)} \ . \tag{2.1}$$

Dabei bezeichnet c_Ω die Konstante aus der Friedrichschen-Ungleichung (Lemma 2.18). Die zweite Beziehung $||y||_{H_0^1(\Omega)} \leq ||y||_{H^1(\Omega)}$ gilt offensichtlich.

Lemma 2.18 (**Friedrichsche Ungleichung**) *Es sei Ω ein Lipschitz-Gebiet, dann existiert eine nur von Ω abhängige Konstante c_Ω, so dass für alle $y \in H_0^1(\Omega)$ die Ungleichung*

$$||y||_{L^2(\Omega)} \leq c_\Omega \, ||y||_{H_0^1(\Omega)}$$

gilt.

Für einfache Gebiete Ω kann man sogar den Wert der Konstante c_Ω angeben. Sie ist als Lösung eines Eigenwert-Problems gegeben und beträgt $c_\Omega = \frac{1}{j_{0,1}}$ für den Fall, dass Ω der Einheitskreis ist ([KS84]). Mit $j_{0,1}$ ist die erste Nullstelle der Bessel-Funktion nullter Ordnung gemeint und es gilt $j_{0,1} \approx 2{,}4048$. Für Gebiete, auf denen die Lösung des Eigenwertproblems nicht bekannt ist, lässt sich $\mathrm{diam}(\Omega)$ als obere Schranke für c_Ω angeben ([Ran06]). Dabei bezeichnet $\mathrm{diam}(\Omega)$ den Durchmesser des Gebietes Ω, also den größtmöglichen Abstand zweier Punkte aus $\overline{\Omega}$.

Es ist möglich, die Randwerte einer Funktion $y \in L^p(\Omega)$ zu ändern, ohne dass sich y im Sinne des Raumes $L^p(\Omega)$ ändert, da der Rand $\partial\Omega$ bzgl. des $\mathbb{R}^N$-Maßes eine Nullmenge ist. Für später ist es notwendig, die Randwerte einer Funktion aus $W^{1,p}(\Omega)$ sinnvoll zu definieren. Dies leistet die Spurabbildung τ aus dem nachfolgenden Satz.

Satz 2.19 (**Spurabbildung**) *Es sei Ω ein Lipschitz-Gebiet mit Rand $\Gamma = \partial\Omega$ und $p \geq 1$. Dann existiert eine stetige, lineare Abbildung $\tau : W^{1,p}(\Omega) \to L^p(\Gamma)$, so dass für alle $y \in \mathcal{C}(\overline{\Omega})$ gilt:*

$$(\tau y)(x) = y(x) \qquad \textit{fast überall auf } \Gamma.$$

So ist also die Randbedingung

$$y = g \qquad \text{auf } \Gamma$$

als

$$y\big|_\Gamma := \tau y = g \qquad \text{fast überall}$$

zu verstehen.

Bemerkung 2.20 *Sowohl die Räume $W^{k,p}(\Omega)$ als auch $W_0^{k,p}(\Omega)$ können in ähnlicher Weise für $p = \infty$ definiert werden.*

Definition 2.21 Mit $H^{-k}(\Omega)$ wird der Dualraum[1] von $H_0^k(\Omega)$ bezeichnet, kurz: $H^{-k}(\Omega) := \left(H_0^k(\Omega)\right)^*$

Das folgende Lemma ist notwendig, um im Kapitel 5 zur schwachen Formulierung einer partiellen Differentialgleichung mit Laplace-Operator überzugehen.

Lemma 2.22 *Es sei Ω ein Lipschitz-Gebiet mit Rand $\Gamma = \partial\Omega$. Für $u, v \in H^1(\Omega)$ gilt*

$$\int_\Omega u(x) D_i v(x)\, \mathrm{d}x = \int_\Gamma u(x) v(x) n_i(x)\, \mathrm{d}s - \int_\Omega v(x) D_i u(x)\, \mathrm{d}x$$

für alle $i = 1, \ldots, N$.
Für $u \in H^2(\Omega)$ und $v \in H^1(\Omega)$ gilt:

$$\int_\Omega \triangle u(x) v(x)\, \mathrm{d}x = \int_\Gamma \nabla u(x) \cdot n(x) v(x)\, \mathrm{d}s - \int_\Omega \nabla u(x) \cdot \nabla v(x)\, \mathrm{d}x \ .$$

2.3 Einbettungen

In diesem Abschnitt werden einige Einbettungen angegeben, um verschiedene Beziehungen unter den eingeführten Räumen herzustellen.

Definition 2.23 Für zwei normierte lineare Räume $(X, ||\cdot||_X)$ und $(Y, ||\cdot||_Y)$ heißt X stetig eingebettet in Y (Schreibweise: $X \hookrightarrow Y$), falls X ein Unterraum von Y ist und

$$||x||_Y \leq c\, ||x||_X$$

für alle $x \in X$ und eine Konstante $c > 0$ gilt. Man sagt, X hat die „stärkere“ Norm.

Aus dieser Definition folgt sofort die Transitivität der stetigen Einbettung, d.h. aus $X \hookrightarrow Y$ und $Y \hookrightarrow Z$ folgt $X \hookrightarrow Z$.

Lemma 2.24 *Aus $X \hookrightarrow Y$ folgt $Y^* \hookrightarrow X^*$.*

[1] Der Dualraum eines reellen, linear normierten Raumes U ist als $\{\varphi : U \to \mathbb{R} : \varphi \text{ linear und stetig}\}$ definiert und wird U^* genannt.

Beweis. Es gilt

$$Y^* = \{\varphi : Y \to \mathbb{R} \text{ linear, stetig}\} \subset \{\varphi : X \to \mathbb{R} \text{ linear, stetig}\} = X^*$$

und

$$||\varphi||_{X^*} = \sup_{\substack{x \in X \\ x \neq 0}} \frac{|\varphi(x)|}{||x||_X} \leq \sup_{\substack{x \in Y \\ x \neq 0}} c\, \frac{|\varphi(x)|}{||x||_Y} = ||\varphi||_{Y^*}$$

für alle $\varphi \in Y^*$. □

Lemma 2.25 (**verallgemeinerte Hölder-Ungleichung**) *Es sei E eine messbare Menge und $p_i \geq 1$ $(i = 1, \ldots, m)$ gegeben. Dann folgt für Funktionen $v_i \in L^{p_i}(E)$ $(i = 1, \ldots, m)$*

$$\prod_{i=1}^{m} v_i \in L^r(E)$$

und es gilt die Abschätzung

$$\left|\left| \prod_{i=1}^{m} v_i \right|\right|_{L^r(E)} \leq \prod_{i=1}^{m} ||v_i||_{L^{p_i}(E)}$$

mit $\frac{1}{r} := \sum_{i=1}^{m} \frac{1}{p_i}$. Für $m = 2$ und $p_1 = p_2 = 2$ erhält man die bekannte Cauchy-Schwarz-Ungleichung in $L^2(E)$.

Einfache stetige Einbettungen sind zum Beispiel:

(i) $W^{k,p}(\Omega) \hookrightarrow L^p(\Omega)$ für $k \in \mathbb{N}_0$ und $1 \leq p < \infty$.

(ii) $L^p(\Omega) \hookrightarrow L^q(\Omega)$ für ein beschränktes Gebiet Ω und $1 \leq q \leq p \leq \infty$.

(iii) $H_0^1(\Omega) \hookrightarrow H^1(\Omega)$ für ein Lipschitz-Gebiet Ω.

Beispiel (i) gilt offensichtlich, Beispiel (ii) mit Hilfe der Hölder-Ungleichung 2.25 und Beispiel (iii) wegen Gleichung (2.1).

Zu schwierigeren Einbettungen macht der folgende Satz eine Aussage. Die Einbettungen hängen dabei von der Dimension des Raumes und den Regularitäten k, p des entsprechenden Sobolevraumes $W^{k,p}(\Omega)$ ab. Ein Beweis dafür ist zum Beispiel in [AF03] zu finden.

Satz 2.26 (**Sobolevscher Einbettungssatz**) *Es sei $\Omega \subset \mathbb{R}^N$ ein Lipschitz-Gebiet, $1 < p < \infty$ sowie $k \in \mathbb{N}_0$, dann gelten die folgenden stetigen Einbettungen.*

(i) *Für $kp < N$: $W^{k,p}(\Omega) \hookrightarrow L^q(\Omega)$ für alle q mit $1 \leq q \leq \frac{Np}{N-kp}$*

(ii) *Für $kp = N$: $W^{k,p}(\Omega) \hookrightarrow L^q(\Omega)$ für alle q mit $1 \leq q < \infty$*

(iii) *Für $kp > N$: $W^{k,p}(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$*

Abschließend werden in diesem Abschnitt einige, sich aus dem letzten Satz ergebende, wichtige Einbettungen als Beispiel angegeben.

Beispiel 2.27

(i) *Für $\Omega \subset \mathbb{R}$ gilt: $H^1(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$.*

(ii) *Für $\Omega \subset \mathbb{R}^2$ gilt: $H^1(\Omega) \hookrightarrow L^q(\Omega)$ für alle q mit $1 \leq q < \infty$.*

(iii) *Für $\Omega \subset \mathbb{R}^3$ gilt: $H^1(\Omega) \hookrightarrow L^6(\Omega)$.*

2.4 Der Nemyzki-Operator

Der Nemyzki-Operator, auch Superpositionsoperator genannt, ist bei dem in Kapitel 5 betrachteten Beispiel von Bedeutung. Insbesondere spiegeln sich Eigenschaften wie seine Stetigkeit bzw. Differenzierbarkeit in der Funktion F, auf welche das Newton-Verfahren angewendet wird, wieder.

Definition 2.28 Es sei $E \subset \mathbb{R}^N$ eine beschränkte, messbare Menge und $\varphi : E \times \mathbb{R} \to \mathbb{R}$ eine reellwertige Funktion. Die Abbildung ϕ mit

$$\phi(y) = \varphi(\cdot, y(\cdot))$$

heißt Nemyzki-Operator. Dabei wird einer Funktion $y : E \to \mathbb{R}$, die durch $z(x) = \varphi(x, y(x))$ definierte Funktion $z : E \to \mathbb{R}$ zugeordnet.

Im Folgenden wird mit E stets eine beschränkte und messbare Menge bezeichnet. Später tritt das Lipschitz-Gebiet Ω an die Stelle von E.

Definition 2.29 Eine Funktion $\varphi : E \times \mathbb{R} \to \mathbb{R}$ erfüllt die Carathéodory-Bedingung, falls $x \mapsto \varphi(x, y)$ für alle $y \in \mathbb{R}$ messbar und $y \mapsto \varphi(x, y)$ stetig ist für fast alle $x \in E$.

Beispiel 2.30 *Die Funktion $\varphi(x, y) = y^3$ erfüllt die Carathéodory-Bedingung. Der dazugehörige Nemyzki-Operator ist durch*

$$\phi(y)(x) = (y(x))^3$$

gegeben.

Der nachstehende Satz aus [Trö05] charakterisiert, unter welchen Bedingungen der Nemyzki-Operator von $L^p(E)$ nach $L^q(E)$ abbildet bzw. stetig abbildet.

Satz 2.31 *Die Funktion φ genüge der Carathéodory-Bedingung. Dann bildet der Nemyzki-Operator $\phi(y) = \varphi(\cdot, y(\cdot))$ für $1 \leq q \leq p < \infty$ von $L^p(E)$ nach $L^q(E)$ ab, genau dann wenn die Wachstumsbedingung*

$$|\varphi(x, y)| \leq \alpha(x) + \beta(x) \, |y|^{\frac{p}{q}} \tag{2.2}$$

mit Funktionen $\alpha \in L^q(E)$ und $\beta \in L^\infty(E)$ erfüllt ist. Des Weiteren ist der Operator ϕ für $q < \infty$ automatisch stetig, wenn er überhaupt $L^p(E)$ in $L^q(E)$ abbildet.

Die Richtungsableitung von $\phi(y)$ in Richtung δy ist durch

$$\begin{aligned}
\phi'(y)\delta y =& \lim_{t \downarrow 0} \frac{\phi(y + t\delta y) - \phi(y)}{t} \\
=& \lim_{t \downarrow 0} \frac{\varphi(\cdot, y(\cdot) + t\delta y(\cdot)) - \varphi(\cdot, y(\cdot))}{t} \\
=& \varphi_y(\cdot, y(\cdot))\delta y(\cdot)
\end{aligned}$$

gegeben. Unter welchen Bedingungen es sich dabei tatsächlich um die Fréchet-Ableitung handelt, zeigt der nächste Satz ([Trö05]).

Satz 2.32 *Zusätzlich zum vorigen Satz sei $\varphi(x, y)$ für fast alle $x \in E$ partiell nach y differenzierbar und der Nemyzki-Operator $\tilde{\phi}(y) = \varphi_y(\cdot, y(\cdot))$ bilde $L^p(E)$ in $L^r(E)$ ab mit $1 \leq q < p < \infty$ und*

$$r = \frac{pq}{p - q} \ .$$

Dann ist $\phi(y) = \varphi(\cdot, y(\cdot))$ von $L^p(E)$ nach $L^q(E)$ Fréchet-differenzierbar und es gilt

$$\left(\phi'(y)\delta y\right)(x) = \varphi_y(x, y(x))\delta y(x) \ .$$

Um die Voraussetzungen des soeben genannten Satzes zu erfüllen, muss also insbesondere die Bedingung (2.2) mit $q := r$ und $\varphi := \varphi_y$ gelten, d.h.

$$|\varphi_y(x, y)| \leq \tilde{\alpha}(x) + \tilde{\beta}(x)\,|y|^{\frac{p}{r}}$$

für Funktionen $\tilde{\alpha} \in L^r(E)$ und $\tilde{\beta} \in L^\infty(E)$. Die Größe r ist dabei so gewählt, dass die Funktion $\left(\phi'(y)\delta y\right)(\cdot) = \varphi_y(\cdot, y(\cdot))\delta y(\cdot)$ gerade noch in $L^q(E)$ liegt, denn es gilt:

$$\frac{1}{r} + \frac{1}{p} = \frac{1}{q} \ .$$

Der folgende Satz sichert die stetige Fréchet-Differenzierbarkeit und entstammt eigenen Überlegungen.

Satz 2.33 *Unter den Voraussetzungen des Satzes 2.32 ist der Nemyzki-Operator $\phi(y) = \varphi(\cdot, y(\cdot))$ von $L^p(E)$ nach $L^q(E)$ sogar stetig Fréchet-differenzierbar.*

Beweis. Es ist zu zeigen, dass aus $||y_n - \bar{y}||_{L^p(E)} \to 0$ stets $||\phi'(y_n) - \phi'(\bar{y})||_{\mathcal{L}(L^p(E), L^q(E))} \to 0$ folgt. Aus

$$\begin{aligned} \left|\left|\phi'(y_n) - \phi'(\bar{y})\right|\right|_{\mathcal{L}(L^p(E), L^q(E))} &= \sup_{\substack{\delta y \in L^p(E) \\ ||\delta y||_{L^p(E)} = 1}} \left|\left|\left(\phi'(y_n) - \phi'(\bar{y})\right)\delta y\right|\right|_{L^q(E)} \\ \text{(Hölder-Ungleichung)} \qquad &\leq ||\varphi_y(\cdot, y_n(\cdot)) - \varphi_y(\cdot, \bar{y}(\cdot))||_{L^r(E)} \end{aligned}$$

folgt zusammen mit der Stetigkeit von $\tilde{\phi}(y) = \varphi_y(\cdot, y(\cdot))$ von $L^p(E)$ nach $L^r(E)$ die Behauptung. □

Beispiel 2.34 *Betrachtet man wieder $\varphi(x, y) = y^3$ und wendet die letzten drei Sätze 2.31, 2.32 und 2.33 mit $p = 6$, $q = 2$ und $r = 3$ an, so ist also $\phi(y)(\cdot) = (y(\cdot))^3$ von $L^6(E)$ nach $L^2(E)$ stetig Fréchet-differenzierbar und es gilt:*

$$\left(\phi'(y)\delta y\right)(x) = 3\,(y(x))^2\,\delta y(x) \ .$$

Diese Eigenschaft lässt sich auch einfach überprüfen. Die Fréchet-Differenzierbarkeit gilt wegen

$$\begin{aligned}
&\lim_{\delta y \to 0} \frac{||\phi(y+\delta y)(\cdot) - \phi(y)(\cdot) - (\phi'(y)\delta y)(\cdot)||_{L^2(E)}}{||\delta y||_{L^6(E)}} \\
&= \lim_{\delta y \to 0} \frac{\left|\left|((y+\delta y)(\cdot))^3 - (y(\cdot))^3 - 3(y(\cdot))^2 \delta y(\cdot)\right|\right|_{L^2(E)}}{||\delta y||_{L^6(E)}} \\
&= \lim_{\delta y \to 0} \frac{\left|\left|3y(\cdot)(\delta y(\cdot))^2 + (\delta y(\cdot))^3\right|\right|_{L^2(E)}}{||\delta y||_{L^6(E)}} \\
\textit{(Hölder-Ungleichung)} \qquad &\leq \lim_{\delta y \to 0} \frac{3||y||_{L^6(E)} \left(||\delta y||_{L^6(E)}\right)^2 + \left(||\delta y||_{L^6(E)}\right)^3}{||\delta y||_{L^6(E)}} \\
&= 0 \ .
\end{aligned}$$

Die stetige Fréchet-Differenzierbarkeit von $L^6(E)$ nach $L^2(E)$ gilt wegen

$$\begin{aligned}
\left|\left|\phi'(y_n) - \phi'(\bar{y})\right|\right|_{\mathcal{L}(L^6(E),L^2(E))} &= \sup_{\substack{\delta y \in L^6(E) \\ ||\delta y||_{L^6(E)}=1}} \left|\left|3\left(y_n(\cdot)^2 - (\bar{y}(\cdot)^2\right)\delta y(\cdot)\right|\right|_{L^2(E)} \\
\textit{(Hölder-Ungleichung)} \qquad &\leq \sup_{\substack{\delta y \in L^6(E) \\ ||\delta y||_{L^6(E)}=1}} 3||y_n + \bar{y}||_{L^6(E)} ||y_n - \bar{y}||_{L^6(E)} ||\delta y||_{L^6(E)} \\
&= 3||y_n + \bar{y}||_{L^6(E)} ||y_n - \bar{y}||_{L^6(E)} \to 0 \ \textit{für} \ y_n \to \bar{y} \ \textit{bzgl.} \ ||\cdot||_{L^6(E)} \ .
\end{aligned}$$

3 Der lokale Algorithmus

In diesem Kapitel wird der lokale Teil eines inexakten, affin kovarianten Verfahrens hergeleitet. Lokal bedeutet hierbei, dass der Startpunkt x_0 hinreichend nah an einer Lösung x^* sein muss. Mit „nah“ ist nicht der direkte Abstand zur Lösung x^* gemeint, sondern dass die Norm des Newton-Schrittes $\Delta x_0 = -\,\mathrm{F}'(x_0)^{-1}\,\mathrm{F}(x_0)$ klein ist. Um welche Norm es sich jeweils handelt, ergibt sich in diesem und dem nächsten Kapitel aus dem Zusammenhang und wird nur selten besonders gekennzeichnet.

Es wird eine Variante eines inexakten Newton-Verfahrens für eine Funktion $\mathrm{F} : X \to Y$ betrachtet. Dabei bezeichnet Δx_n die exakte Newton-Richtung im n-ten Iterationsschritt, d.h.

$$\Delta x_n = -\,\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n)\ ,$$

wobei $\mathrm{F}'(x_n)^{-1}$ der lineare inverse Operator von der Fréchet-Ableitung F' in x_n ist. Der exakte Newton-Schritt hat allerdings „nur“ theoretische Bedeutung. Im Allgemeinen ist X ein unendlich-dimensionaler Banachraum und Δx_n steht in praktischen Rechnungen nicht zur Verfügung. Eine berechenbare Näherung für Δx_n wird δx_n genannt. Der dabei auftretende Fehler kann auf zweierlei Art ausgedrückt werden, einerseits durch einen Fehler r_n in Y mit

$$r_n := \mathrm{F}'(x_n)\delta x_n + \mathrm{F}(x_n) = \mathrm{F}'(x_n)\,(\delta x_n - \Delta x_n)\ ,$$

andererseits durch den Fehler $\Delta x_n - \delta x_n$ im Raum X. Es ist zu klären, welche Bedingungen an den Fehler gestellt werden müssen, um die Konvergenz zu sichern. Wie sich herausstellen wird, ist die relative Genauigkeit δ_n bzgl. δx_n ein wichtiges Maß für die Konvergenztheorie. Sie ist definiert als

$$\delta_n := \frac{||\Delta x_n - \delta x_n||}{||\delta x_n||}\ .$$

Da nur δx_n zur Verfügung steht, kann auch nur dieser inexakte Schritt zum Aufdatieren genutzt werden. Damit ergibt sich x_{n+1} als

$$x_{n+1} := x_n + \delta x_n = x_0 + \sum_{i=0}^{n} \delta x_n \ .$$

Definition 3.1 Eine Folge x_n mit $x_n \to x^*$ konvergiert mit linearer Konvergenzgeschwindigkeit gegen x^*, falls für ein $\theta < 1$ (Konvergenzfaktor) die Beziehung

$$||x_{n+1} - x^*|| \leq \theta \, ||x_n - x^*||$$

für alle n gilt. Sie konvergiert mit Ordnung p ($p > 1$), falls für ein $c \geq 0$ und alle n die Beziehung

$$||x_{n+1} - x^*|| \leq c \, ||x_n - x^*||^p$$

gilt. Für $p = 2$ spricht man von quadratischer Konvergenz.

Das folgende Lemma entstammt eigenen Überlegungen und folgert aus geeigneter linearer Konvergenz des Schrittes δx_n die lineare Konvergenz von $||x_n - x^*||$.

Lemma 3.2 *Die Folge $||\delta x_n||$ konvergiere linear gegen 0, d.h. $||\delta x_{n+1}|| \leq \theta \, ||\delta x_n||$ mit $\theta < 1$. Dann besitzt die Reihe $x_0 + \sum_{i=0}^{\infty} \delta x_i$ einen Grenzwert x^*. Für $\theta < \frac{1}{3}$ gilt sogar die lineare Konvergenz von $||x_n - x^*||$, d.h.*

$$||x_{n+1} - x^*|| \leq \tilde{\theta} \, ||x_n - x^*||$$

mit $\tilde{\theta} < 1$ und $x_n := x_0 + \sum_{i=0}^{n-1} \delta x_i$.

Beweis. Die Existenz eines Grenzwertes folgt direkt aus dem Quotientenkriterium. Der

zweite Teil kann wie folgt bewiesen werden. Es gilt:

$$\begin{aligned} ||x_n - x^*|| = \left|\left| \sum_{i=n}^{\infty} \delta x_i \right|\right| &\leq \sum_{i=n}^{\infty} ||\delta x_i|| \\ &\leq \sum_{i=n}^{\infty} \theta^{i-n+1} \, ||\delta x_{n-1}|| \\ &\leq \frac{\theta}{1-\theta} \, ||\delta x_{n-1}|| \ . \end{aligned} \tag{3.1}$$

Zusammen mit

$$||x_n - x_{n-1}|| \leq ||x_n - x^*|| + ||x_{n-1} - x^*|| \ ,$$

der Dreiecksungleichung und der Beziehung (3.1) ergibt sich:

$$\begin{aligned} ||x_{n-1} - x^*|| &\geq ||x_n - x_{n-1}|| - ||x_n - x^*|| \\ &\geq ||\delta x_{n-1}|| - \frac{\theta}{1-\theta} \, ||\delta x_{n-1}|| \\ &= \left(\frac{1-2\theta}{1-\theta} \right) ||\delta x_{n-1}|| \ . \end{aligned} \tag{3.2}$$

Aus (3.1) und (3.2) folgt schließlich

$$\begin{aligned} ||x_n - x^*|| \leq \frac{\theta}{1-\theta} \, ||\delta x_{n-1}|| &\leq \left(\frac{1-\theta}{1-2\theta} \right) \left(\frac{\theta}{1-\theta} \right) ||x_{n-1} - x^*|| \\ &= \underbrace{\left(\frac{\theta}{1-2\theta} \right)}_{=:\tilde{\theta}<1, \text{ für } \theta<\frac{1}{3}} \quad ||x_{n-1} - x^*|| \end{aligned}$$

und damit die Behauptung. □

Bemerkung 3.3

(i) Konvergiert die Folge $||\delta x_n||$ mit der Ordnung $p > 1$ (d.h. $||\delta x_{n+1}|| \leq c\,||\delta x_n||^p$ mit $c \geq 0$), so konvergiert sie ab einem $n_0 \in \mathbb{N}$ insbesondere linear, wobei der Konvergenzfaktor für große n beliebig klein wird. Eine solche Folge erfüllt also die Voraussetzungen des vorigen Lemmas für $n > n_0$.

(ii) Für eine Folge δx_n *mit* $||\delta x_{n+1}|| \leq \theta\, ||\delta x_n||$ *und* $0 \leq \theta < \frac{1}{2}$ *vermutet der Autor, dass*

$$||x_{n+1} - x^*|| \leq \tilde{\theta}\, ||x_n - x^*|| \qquad \forall n \geq n_0$$

mit einem $n_0 \in \mathbb{N}$ *und* $\tilde{\theta} < 1$ *gilt.*

3.1 Konvergenzaussagen

Verschiedene Sätze aus [Deu04][1] bilden die Grundlage für die folgenden Konvergenzaussagen. Bewiesen werden die lokal quadratische und die lokal lineare Konvergenz des beschriebenen inexakten Newton-Verfahrens. Ein Vorteil des folgenden Satzes besteht in seiner guten numerischen Nutzbarkeit.

Satz 3.4 *Es sei* $\mathrm{F} : X \to Y$ *eine in* X *stetig Fréchet-differenzierbare Funktion und* $x_0 \in X$*. Außerdem existiere* $\mathrm{F}'(x)^{-1}$ *für alle* $x \in X$ *und es gelte die affin invariante Lipschitz-Bedingung*

$$\left|\left|\mathrm{F}'(z)^{-1}\left(\mathrm{F}'(y) - \mathrm{F}'(x)\right) v\right|\right| \leq \omega\, ||y - x||\, ||v|| \tag{3.3}$$

für kollineare $x, y, z \in X$ *und für alle* $v \in X$*. Des Weiteren gelte für die relative Genauigkeit*

$$\delta_n \leq \bar{\delta} < 1 \quad \textit{und} \quad \delta_n \leq \frac{\rho}{2} \frac{h_n^\delta}{1 + h_n^\delta} \tag{3.4}$$

und für den Startpunkt x_0 *die Bedingung*

$$||\Delta x_0|| < \frac{2\left(1 - \bar{\delta}\right)^2}{(1 + \rho)\, \omega} \tag{3.5}$$

für eine Konstante $\rho > 0$*. Dabei bezeichnen* $h_n := \omega\, ||\Delta x_n||$ *bzw.* $h_n^\delta := \omega\, ||\delta x_n||$ *die Kantorovitsch Größen und* $\delta_n := \frac{||\Delta x_n - \delta x_n||}{||\delta x_n||}$ *die relative Genauigkeit bzgl.* δx_n*.*
Dann konvergiert die Folge $x_n = x_0 + \sum_{i=0}^{n-1} \delta x_i$ *gegen einen Punkt* x^* *mit* $\mathrm{F}(x^*) = 0$*. Weiter*

[1] Satz 2.10 und 2.11, Seite 67-72

gilt $||\Delta x_n|| \to 0$ *mit*

$$\frac{||\Delta x_{n+1}||}{||\Delta x_n||} \leq \frac{1+\rho}{2\left(1-\delta_n\right)^2}\,\omega\,||\Delta x_n|| \leq \frac{1+\rho}{2\left(1-\bar{\delta}\right)^2}\,\omega\,||\Delta x_n|| \tag{3.6}$$

für die exakten Richtungen Δx_n*. Für die inexakten Richtungen* δx_n *gilt* $||\delta x_n|| \to 0$ *mit*

$$\frac{||\delta x_{n+1}||}{||\delta x_n||} \leq \frac{1+\rho}{2\left(1-\delta_{n+1}\right)}\,\omega\,||\delta x_n|| \leq \frac{1+\rho}{2\left(1-\bar{\delta}\right)}\,\omega\,||\delta x_n|| \ . \tag{3.7}$$

Beweis. Aus der Dreiecksungleichung folgt

$$\begin{aligned} & ||\Delta x_n|| - ||\delta x_n - \Delta x_n|| \leq ||\delta x_n|| \leq ||\delta x_n - \Delta x_n|| + ||\Delta x_n|| \\ \iff \quad & ||\Delta x_n|| - \delta_n\,||\delta x_n|| \leq ||\delta x_n|| \leq \delta_n\,||\delta x_n|| + ||\Delta x_n|| \\ \iff \quad & \frac{||\Delta x_n||}{1+\delta_n} \leq ||\delta x_n|| \leq \frac{||\Delta x_n||}{1-\delta_n}\ . \end{aligned} \tag{3.8}$$

Es darf also $||\delta x_n||$ nur begrenzt von $||\Delta x_n||$ abweichen (siehe auch Proposition 3.7). Diese Beziehung ist von großer Bedeutung und wird in dieser Arbeit häufig verwendet. Weiter gilt:

$$\begin{aligned} \Delta x_{n+1} &= -\,\mathrm{F}'(x_{n+1})^{-1}\,\mathrm{F}(x_{n+1}) \\ &= -\,\mathrm{F}'(x_{n+1})^{-1}\left(\mathrm{F}(x_{n+1}) - \mathrm{F}(x_n) - \mathrm{F}'(x_n)\delta x_n + r_n\right) \\ &= -\int_0^1 \mathrm{F}'(x_{n+1})^{-1}\left(\mathrm{F}'(x_n + t\,\delta x_n) - \mathrm{F}'(x_n)\right)\delta x_n\,\mathrm{d}t - \mathrm{F}'(x_{n+1})^{-1} r_n \end{aligned}$$

und damit

$$||\Delta x_{n+1}|| \leq \underbrace{\left|\left|\int_0^1 \mathrm{F}'(x_{n+1})^{-1}\left(\mathrm{F}'(x_n + t\,\delta x_n) - \mathrm{F}'(x_n)\right)\delta x_n\,\mathrm{d}t\right|\right|}_{(I)} + \underbrace{\left|\left|\mathrm{F}'(x_{n+1})^{-1} r_n\right|\right|}_{(II)}\ . \tag{3.9}$$

Hierbei lässt sich (I) mit (3.3) durch

$$(I) \leq \int_0^1 t\omega\,||\delta x_n||^2\ \mathrm{d}t = \frac{1}{2} h_n^\delta\,||\delta x_n||$$

und (II) durch

$$\begin{aligned}(II) &= \left|\left|\mathrm{F}'(x_{n+1})^{-1}\,\mathrm{F}'(x_n)\,(\delta x_n - \Delta x_n)\right|\right| \\ &= \left|\left|\mathrm{F}'(x_{n+1})^{-1}\left(\mathrm{F}'(x_n) - \mathrm{F}'(x_{n+1})\right)(\delta x_n - \Delta x_n) + (\delta x_n - \Delta x_n)\right|\right| \\ &\leq (\omega\,||\delta x_n|| + 1)\,||\delta x_n - \Delta x_n|| \\ &= \left(1 + h_n^\delta\right)||\delta x_n - \Delta x_n||\end{aligned}$$

abschätzen. Damit ergibt sich aus (3.9) der zentrale Zusammenhang

$$\frac{||\Delta x_{n+1}||}{||\delta x_n||} \leq \frac{1}{2}h_n^\delta + \left(1 + h_n^\delta\right)\delta_n \tag{3.10}$$

und mit der Bedingung (3.4) sogar

$$\frac{||\Delta x_{n+1}||}{||\delta x_n||} \leq \frac{1}{2}\,(1+\rho)\,h_n^\delta\ . \tag{3.11}$$

Daraus folgt zusammen mit (3.8) die behauptete Ungleichung (3.6), denn es gilt:

$$\begin{aligned}\frac{||\Delta x_{n+1}||}{||\Delta x_n||} \leq \frac{1}{1-\delta_n}\frac{||\Delta x_{n+1}||}{||\delta x_n||} &\leq \frac{1+\rho}{2\,(1-\delta_n)}h_n^\delta \\ &\leq \frac{1+\rho}{2\,(1-\delta_n)^2}\omega\,||\Delta x_n|| \\ &\leq \frac{1+\rho}{2\left(1-\bar{\delta}\right)^2}\omega\,||\Delta x_n||\ .\end{aligned}$$

Mit der Startbedingung (3.5) ergibt dies die quadratische Konvergenz von $||\Delta x_n||$ gegen 0. Analog folgt (3.7), denn es gilt:

$$\begin{aligned}\frac{||\delta x_{n+1}||}{||\delta x_n||} \leq \frac{1}{1-\delta_{n+1}}\frac{||\Delta x_{n+1}||}{||\delta x_n||} &\leq \frac{1+\rho}{2\,(1-\delta_{n+1})}h_n^\delta \\ &= \frac{1+\rho}{2\,(1-\delta_{n+1})}\omega\,||\delta x_n|| \\ &\leq \frac{1+\rho}{2\left(1-\bar{\delta}\right)}\omega\,||\delta x_n||\ .\end{aligned}$$

Aus (3.8) und der Startbedingung (3.5) folgt $||\delta x_0|| < \frac{2\left(1-\bar{\delta}\right)}{(1+\rho)\omega}$ und damit auch die quadratische Konvergenz von $||\delta x_n||$ gegen 0.

Dass die Folge x_n einen Grenzwert x^* besitzt, resultiert aus dem Quotientenkriterium bzw. Lemma 3.2. Für den Funktionswert des Grenzwertes x^* gilt:

$$\begin{aligned}
||\mathrm{F}(x^*)|| &= \lim_{n\to\infty} ||\mathrm{F}(x_n)|| && \text{, weil F und } ||\cdot|| \text{ stetig sind} \\
&= \lim_{n\to\infty} \left|\left|\mathrm{F}'(x_n)\,\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n)\right|\right| \\
&\leq \lim_{n\to\infty} \left|\left|\mathrm{F}'(x_n)\right|\right| ||\Delta x_n|| \\
&\leq \lim_{n\to\infty} C \left|\left|\mathrm{F}'(x^*)\right|\right| ||\Delta x_n|| && \text{, weil F}' \text{ stetig ist} \\
&= 0 \ .
\end{aligned}$$

□

Bemerkung 3.5

(i) Es reicht aus, die stetige Fréchet-Differenzierbarkeit und die Regularität von F *in einer Umgebung* D *der Iterierten zu fordern. Für die Lipschitz-Bedingung genügt es, wenn sie, gemäß der entsprechenden Stellen im Beweis, auf den Verbindungsstrecken der Iterierten gilt.*

(ii) Für eine Fréchet-differenzierbare Funktion mit einer Lipschitz-stetigen Ableitung mit Lipschitz-Konstante L *gilt in einer Umgebung von* x^* *die Beziehung:*

$$\left|\left|\mathrm{F}'(z)^{-1}\left(\mathrm{F}'(y)-\mathrm{F}'(x)\right)v\right|\right| \leq \left|\left|\mathrm{F}'(z)^{-1}\right|\right| L\,||y-x||\,||v|| \ .$$

Folgerung 2.12 besagt, dass $\mathrm{F}'(z)^{-1}$ *in einer Umgebung von* x^* *existiert und beschränkt ist, sofern* $\mathrm{F}'(x^*)$ *in* $\bar{\mathcal{L}}(X,Y)$ *liegt. Das heißt,* F *erfüllt die von Satz 3.4 geforderte Lipschitz-Bedingung* (3.3)*. Die Lipschitz-Stetigkeit der Ableitung bedeutet, anschaulich dargestellt, dass die Funktion eine nicht zu große Krümmung haben darf. Abbildung 3.1 zeigt dies in einem einfachen eindimensionalen Fall. Die Funktion* F *darf dabei die gekennzeichnete Fläche nicht verlassen und hat somit zwangsläufig eine Nullstelle, gegen die das Verfahren konvergiert.*

(iii) Mit $||\Delta x_n|| \to 0$ *bzw.* $||\delta x_n|| \to 0$ *konvergiert auch* h_n *bzw.* h_n^δ *gegen* 0*. Zusammen mit Bedingung* (3.4) *ergibt dies, dass die relative Genauigkeit* δ_n *im Laufe der Iteration gegen* 0 *streben muss.*

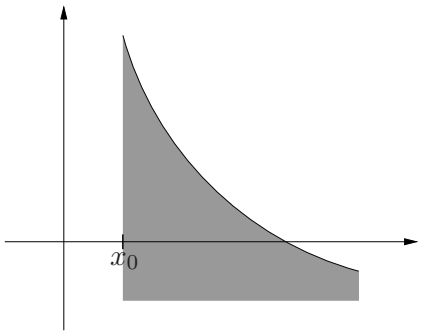


Abbildung 3.1: Motivation für die Lipschitz-Bedingung

In [Ypm84] spielt die relative Genauigkeit bzgl. $||\Delta x_n||$ eine wichtige Rolle. Sie ist durch $\epsilon_n = \frac{||\Delta x_n - \delta x_n||}{||\Delta x_n||}$ definiert. Zwischen ihr und der hier verwendeten relativen Genauigkeit δ_n besteht der folgende direkte Zusammenhang.

Korollar 3.6 *Aus Ungleichung* (3.8) *ergibt sich der folgende Zusammenhang zwischen der relativen Genauigkeit bzgl.* $||\delta x||$ *(*$\delta = \frac{||\Delta x - \delta x||}{||\delta x||}$*) und der relativen Genauigkeit bzgl.* $||\Delta x||$ *(*$\epsilon = \frac{||\Delta x - \delta x||}{||\Delta x||}$*):*

$$\frac{\epsilon}{1+\epsilon} \leq \delta \leq \frac{\epsilon}{1-\epsilon} \ .$$

Falls also δ klein ist, so ist auch ϵ klein und umgekehrt.

Die folgende Proposition dient zur Veranschaulichung der zulässigen Menge für δx_n.

Proposition 3.7 *Sofern die Norm von einem Skalarprodukt* $(\cdot,\cdot)$ *induziert wird (d.h.* X *ist ein Hilbertraum), beschreibt die Menge* $\left\{\delta x \in X : \frac{||\Delta x - \delta x||}{||\delta x||} \leq \delta\right\}$ *(mit* $\delta < 1$*) eine Kugel um den Mittelpunkt* $\frac{1}{1-\delta^2}\Delta x$ *mit dem Radius* $\frac{\delta}{1-\delta^2}\,||\Delta x||$.

Beweis. Es gilt:

$$\begin{aligned}
& \frac{||\Delta x - \delta x||}{||\delta x||} \leq \delta \\
\iff \quad & ||\delta x||^2 - (\delta x, \Delta x) - (\Delta x, \delta x) + ||\Delta x||^2 \leq \delta^2 \, ||\delta x||^2 \\
\iff \quad & ||\delta x||^2 - \frac{1}{1-\delta^2}\,(\delta x, \Delta x) - \frac{1}{1-\delta^2}\,(\Delta x, \delta x) + \frac{1}{1-\delta^2}\,||\Delta x||^2 \leq 0
\end{aligned}$$

$$\iff \qquad \left|\left|\delta x - \frac{1}{1-\delta^2}\Delta x\right|\right|^2 + \left(\frac{1}{1-\delta^2} - \frac{1}{(1-\delta^2)^2}\right)||\Delta x||^2 \leq 0$$

$$\iff \qquad \left|\left|\delta x - \frac{1}{1-\delta^2}\Delta x\right|\right| \leq \frac{\delta}{1-\delta^2}\,||\Delta x|| \ .$$

□

Insbesondere wird damit auch die im Satz 3.4 verwendete Ungleichung (3.8) deutlich, wie in Abbildung 3.2 dargestellt.

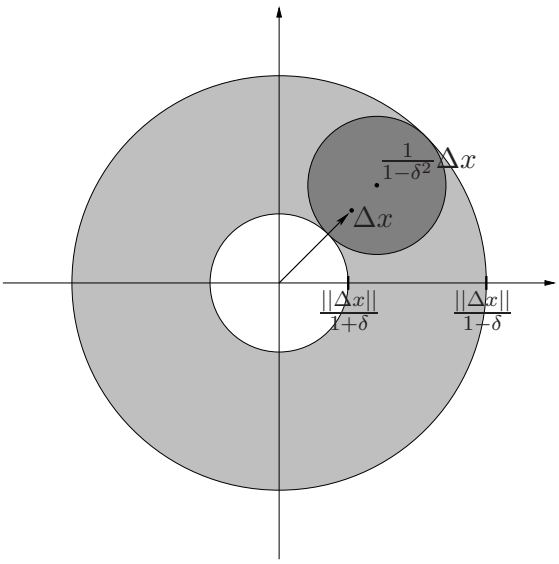


Abbildung 3.2: Die Menge $\left\{\delta x \in X : \frac{||\Delta x - \delta x||}{||\delta x||} \leq \delta\right\}$ mit $\delta = \frac{1}{2}$

Die Forderung an die relative Genauigkeit δ_n in Satz 3.4 kann sich als zu stark erweisen. Der folgende Satz schwächt die Forderung ab und sichert die lineare Konvergenz von $||\Delta x_n||$ und $||\delta x_n||$ gegen 0.

Satz 3.8 *Es sei* $\mathrm{F} : X \to Y$ *eine in* X *stetig Fréchet-differenzierbare Funktion und* $x_0 \in X$*. Außerdem existiere* $\mathrm{F}'(x)^{-1}$ *für alle* $x \in X$ *und es gelte die affin invariante Lipschitz-Bedingung*

$$\left|\left|\mathrm{F}'(z)^{-1}\left(\mathrm{F}'(y) - \mathrm{F}'(x)\right)v\right|\right| \leq \omega\,||y-x||\,||v||$$

für kollineare $x, y, z \in X$ und für alle $v \in X$. Des Weiteren sei die Iteration so gesteuert, dass

$$\vartheta(h_n^\delta, \delta_n) := \frac{\frac{1}{2}h_n^\delta + \left(1 + h_n^\delta\right)\delta_n}{1 - \delta_n} \leq \theta \tag{3.12}$$

und

$$\bar{\vartheta}(h_n^\delta, \delta_n, \delta_{n+1}) := \frac{\frac{1}{2}h_n^\delta + \left(1 + h_n^\delta\right)\delta_n}{1 - \delta_{n+1}} \leq \theta \tag{3.13}$$

für ein $\theta < 1$ gilt.
Dann folgt die lineare Konvergenz von $||\Delta x_n||$ und $||\delta x_n||$ gegen 0 mit Konvergenzfaktor θ und die Folge $x_n = x_0 + \sum_{i=0}^{n-1} \delta x_i$ konvergiert gegen ein x^* mit $\mathrm{F}(x^*) = 0$.

Beweis. Analog zum Beweis von Satz 3.4 besitzen hier Gleichung (3.8) und (3.10) Gültigkeit. Aus ihnen folgt:

$$\frac{||\Delta x_{n+1}||}{||\Delta x_n||} \leq \frac{\frac{1}{2}h_n^\delta + \left(1 + h_n^\delta\right)\delta_n}{1 - \delta_n} = \vartheta(h_n^\delta, \delta_n) \leq \theta$$

sowie

$$\frac{||\delta x_{n+1}||}{||\delta x_n||} \leq \frac{\frac{1}{2}h_n^\delta + \left(1 + h_n^\delta\right)\delta_n}{1 - \delta_{n+1}} = \bar{\vartheta}(h_n^\delta, \delta_n, \delta_{n+1}) \leq \theta$$

und damit die behauptete lineare Konvergenz. Der verbleibende Teil der Behauptung (also $\mathrm{F}(x^*) = 0$) kann wie im Beweis von Satz 3.4 gezeigt werden. □

In den Bedingungen (3.12) und (3.13) ist eine „versteckte“ Startbedingung enthalten. So kann es passieren, dass die Bedingungen trotz $\delta_n = 0$ nicht erfüllt werden können. Aus numerischer Sicht lässt sich dieser Satz nur schwer nutzen, da sich die Terme $\vartheta(h_n^\delta, \delta_n)$ und $\bar{\vartheta}(h_n^\delta, \delta_n, \delta_{n+1})$, wegen der meist unbekannten Lipschitz-Konstante, weder berechnen noch zweckmäßig schätzen lassen. Für den Fall, dass das Verfahren konvergiert, gilt $h_n^\delta \to 0$. Dies macht sich der folgende Satz zu Nutze.

Satz 3.9 *Es sei $\mathrm{F} : X \to Y$ eine in X stetig Fréchet-differenzierbare Funktion und $x_0 \in X$. Außerdem existiere $\mathrm{F}'(x)^{-1}$ für alle $x \in X$ und es gelte die affin invariante Lipschitz-Bedingung*

$$\left|\left|\mathrm{F}'(z)^{-1}\left(\mathrm{F}'(y) - \mathrm{F}'(x)\right)v\right|\right| \leq \omega\,||y - x||\,||v|| \tag{3.14}$$

für kollineare $x, y, z \in X$ und für alle $v \in X$. Des Weiteren gelte für $\theta < 1$ und $\bar{\delta} < \frac{1}{2}$ die Beziehung

$$\theta \geq \frac{\bar{\delta}}{1-\bar{\delta}} \qquad \left(\Longleftrightarrow \bar{\delta} \leq \frac{\theta}{1+\theta} \right) . \tag{3.15}$$

Zusätzlich wird $\delta_n \leq \bar{\delta}$ für alle n als Bedingung an die relative Genauigkeit gestellt und es gelte die Startbedingung

$$||\Delta x_0|| \leq \frac{\left(1-\bar{\delta}\right)\left(\theta\left(1-\bar{\delta}\right)-\bar{\delta}\right)}{\omega\left(\frac{1}{2}+\bar{\delta}\right)} . \tag{3.16}$$

Dann folgt die lineare Konvergenz von $||\Delta x_n||$ und $||\delta x_n||$ gegen 0 mit Konvergenzfaktor θ und die Folge $x_n = x_0 + \sum_{i=0}^{n-1} \delta x_i$ konvergiert gegen ein x^* mit $\mathrm{F}(x^*) = 0$.

Beweis. Gezeigt wird, dass $\vartheta(h_n^\delta, \delta_n) \leq \theta$ und $\bar{\vartheta}(h_n^\delta, \delta_n, \delta_{n+1}) \leq \theta$ für alle n gilt. Dann folgt aus dem vorigen Satz die lineare Konvergenz von $||\Delta x_n||$ und $||\delta x_n||$ mit Konvergenzfaktor θ. Zusammen mit Gleichung (3.8) ergibt sich:

$$\begin{aligned}\vartheta(h_n^\delta, \delta_n) = \frac{\frac{1}{2}h_n^\delta + \left(1+h_n^\delta\right)\delta_n}{1-\delta_n} &\leq \frac{\frac{1}{2}\frac{h_n}{1-\delta_n} + \left(1+\frac{h_n}{1-\delta_n}\right)\delta_n}{1-\delta_n} \\ &\leq \frac{\frac{1}{2}\frac{h_n}{1-\bar{\delta}} + \left(1+\frac{h_n}{1-\bar{\delta}}\right)\bar{\delta}}{1-\bar{\delta}} .\end{aligned}$$

Analog folgt:

$$\bar{\vartheta}(h_n^\delta, \delta_n, \delta_{n+1}) \leq \frac{\frac{1}{2}\frac{h_n}{1-\bar{\delta}} + \left(1+\frac{h_n}{1-\bar{\delta}}\right)\bar{\delta}}{1-\bar{\delta}} .$$

Aus einfachem Umstellen resultiert:

$$\begin{aligned} & \frac{\frac{1}{2}\frac{h_n}{1-\bar{\delta}} + \left(1+\frac{h_n}{1-\bar{\delta}}\right)\bar{\delta}}{1-\bar{\delta}} \leq \theta \\ \Longleftrightarrow \quad & \left(\frac{1}{2}\frac{1}{1-\bar{\delta}} + \frac{\bar{\delta}}{1-\bar{\delta}}\right) h_n \leq \theta\left(1-\bar{\delta}\right) - \bar{\delta} \\ \Longleftrightarrow \quad & \left(\frac{1}{2}+\bar{\delta}\right) h_n \leq \left(1-\bar{\delta}\right)\left(\theta\left(1-\bar{\delta}\right)-\bar{\delta}\right) \\ \Longleftrightarrow \quad & ||\Delta x_n|| \leq \frac{\left(1-\bar{\delta}\right)\left(\theta\left(1-\bar{\delta}\right)-\bar{\delta}\right)}{\omega\left(\frac{1}{2}+\bar{\delta}\right)} . \end{aligned}$$

Für $n = 0$ gilt dies nach Voraussetzung (3.16). Daraus ergibt sich $||\Delta x_1|| \leq \theta\,||\Delta x_0||$ und $||\delta x_1|| \leq \theta\,||\delta x_0||$. Durch Induktion folgt schließlich die lineare Konvergenz von $||\Delta x_n||$ und $||\delta x_n||$ gegen 0 mit Konvergenzfaktor θ. Auch hier kann der verbleibende Teil der Behauptung (also $\mathrm{F}(x^*) = 0$) wie im Beweis von Satz 3.4 gezeigt werden. □

Die Beziehung (3.15) stellt einen direkten Zusammenhang zwischen der geforderten relativen Genauigkeit und dem damit verbundenen Konvergenzfaktor dar. Wird zu gegebenen $\bar{\delta}$ der „bestmögliche“ Konvergenzfaktor $\theta = \frac{\bar{\delta}}{1-\bar{\delta}}$ gewählt, dann verlangt die Startbedingung (3.16), dass $||\Delta x_0|| = 0$ (d.h. $x_0 = x^*$) gilt. Ist der Abstand zwischen θ und $\frac{\bar{\delta}}{1-\bar{\delta}}$ größer, kommt dies der Startbedingung zu Gute.

3.2 Algorithmische Umsetzung

Ziel ist, die Aussagen von den Sätzen 3.4, 3.8 und 3.9 algorithmisch zu nutzen. Dafür werden hier ein $\bar{\delta} < 1$ und $\theta < 1$ vorausgesetzt und es wird davon ausgegangen, dass die Newton-Gleichung

$$\mathrm{F}'(x_n)\Delta x_n = -\,\mathrm{F}(x_n)$$

beliebig genau gelöst werden kann. Außerdem muss δ_n selbst nicht zur Verfügung stehen. Es genügt, wenn eine obere Schätzung[1] $[\delta_n]$ für die relative Genauigkeit

$$\delta_n = \frac{||\Delta x_n - \delta x_n||}{||\delta x_n||} \leq \bar{\delta} < 1$$

vorhanden ist (d.h. $\delta_n \leq [\delta_n]$).

3.2.1 Quadratischer Konvergenzmodus

Für einen Algorithmus ist es notwendig, die Größe h_n^δ zu schätzen. Der erste Teil von Gleichung (3.7) lautet:

$$\frac{||\delta x_{n+1}||}{||\delta x_n||} \leq \frac{1+\rho}{2\,(1-\delta_{n+1})} h_n^\delta\ .$$

[1] Größen, die mit eckigen Klammern versehen sind, stehen für Schätzungen dieser. In der Regel gilt eine der Beziehungen $[a] \leq a$ oder $[a] \geq a$.

Daraus ist die „a-postposteriori“ Schätzung

$$\left[h_n^\delta\right]_1 := 2\frac{||\delta x_{n+1}||}{||\delta x_n||}\frac{1-[\delta_{n+1}]}{1+\rho} = 2\tilde{\theta}_n\frac{1-[\delta_{n+1}]}{1+\rho} \leq h_n^\delta$$

für h_n^δ mit dem Kontraktionsfaktor $\tilde{\theta}_n := \frac{||\delta x_{n+1}||}{||\delta x_n||}$ und $\delta_{n+1} \leq [\delta_{n+1}] < 1$ motiviert. „A-postposteriori“ meint, dass die Größe $\left[h_n^\delta\right]_1$ erst nach der Berechnung von δx_{n+1} zur Verfügung steht, also nachdem der Schritt in Richtung δx_n bereits vollzogen ist und die neue Richtung δx_{n+1} berechnet wurde. Dies soll durch den niedergestellten Index 1 angedeutet werden. Mit einem kleinen Trick lässt sich daraus eine a-posteriori Schätzung für h_n^δ angeben. Diese ist aus $h_n^\delta = \tilde{\theta}_{n-1}h_{n-1}^\delta$ durch

$$\left[h_n^\delta\right] := \tilde{\theta}_{n-1}\left[h_{n-1}^\delta\right]_1 = 2\tilde{\theta}_{n-1}^2\frac{1-[\delta_n]}{1+\rho} \leq h_n^\delta$$

motiviert. Da es nicht gelingt, eine echte a-priori Schätzung für h_n^δ anzugeben, kann im Algorithmus erst nach Berechnung der Richtung δx_n festgestellt werden, welche Genauigkeit benötigt wird und ob δx_n dieser genügt. Dafür wird die mittlere Ungleichung in

$$\delta_n \leq [\delta_n] \leq \frac{\rho}{2}\frac{\left[h_n^\delta\right]}{1+\left[h_h^\delta\right]} \leq \frac{\rho}{2}\frac{h_n^\delta}{1+h_n^\delta}$$

überprüft. Die hintere Ungleichung gilt, weil $\frac{\rho}{2}\frac{h}{1+h}$ monoton wachsend in h und $\left[h_n^\delta\right] \leq h_n^\delta$ ist. Falls diese Bedingung an die Genauigkeit nicht erfüllt ist, wird δx_n verworfen und ein genaueres berechnet. Mit Blick auf Diskretisierungen von Differentialgleichungen kann dies die Vergrößerung des Ansatzraumes bzw. eine Verfeinerung des Gitters bedeuten. In diesem Fall kann $[\delta_n]$ mit Hilfe eines PDE-Fehlerschätzers definiert werden.

Der lokale Teil eines Algorithmus könnte wie in Abbildung 3.3 dargestellt, aussehen. Erkennbar an $\langle\cdot,\cdot\rangle$ wird hierbei der Newton-Schritt bereits im „Galerkin-Sinn“ bzgl. des Unterraumes X_n gelöst. Die Wahl der Unterräume X_n steuert dabei die Genauigkeit δ_n.

Schritt 1 wird so lange wiederholt, bis ein δx_n mit ausreichender Genauigkeit bestimmt wurde. Dies gelingt in jedem Fall, wenn die Vergrößerungen des Ansatzraumes so gewählt werden, dass $[\delta x_n]$ gegen 0 „konvergiert“. Die geforderte Genauigkeit $\frac{\rho}{2}\frac{\left[h_n^\delta\right]}{1+\left[h_n^\delta\right]}$ ist automatisch größer Null (sofern $\Delta x_n \neq 0$ gilt). Für $n = 0$ kann nur die „a-postposteriori“ Schätzung $\left[h_0^\delta\right]_1$ verwendet werden. In dieser Hinsicht unterscheidet sich der erste Iterationsschritt von den anderen.

Eingabe: x_0, welches die Voraussetzungen von Satz 3.4 erfüllt.
Ausgabe: Folge x_n, welche die von Satz 3.4 vorausgesagten Eigenschaften hat.

Schritt 0 Wähle ρ und den Ansatzraum X_0
Löse $\langle \mathrm{F}'(x_0)\delta x_0, \varphi\rangle = \langle -\mathrm{F}(x_0), \varphi\rangle \qquad \forall \varphi \in X_0$ mit $\delta x_0 \in X_0$
Schätze $[\delta_0]$
Setze $x_1 := x_0 + \delta x_0$, $X_1 := X_0$ und $n := 1$

Schritt 1 Löse $\langle \mathrm{F}'(x_n)\delta x_n, \varphi\rangle = \langle -\mathrm{F}(x_n), \varphi\rangle \qquad \forall \varphi \in X_n$ mit $\delta x_n \in X_n$
Schätze $[\delta_n]$
Setze $\tilde{\theta}_{n-1} := \frac{||\delta x_n||}{||\delta x_{n-1}||}$ und $\left[h_n^\delta\right] := 2\tilde{\theta}_{n-1}^2 \frac{1-[\delta_n]}{1+\rho}$
Falls $n = 1$,
Setze $\left[h_0^\delta\right]_1 := 2\tilde{\theta}_0 \frac{1-[\delta_1]}{1+\rho}$
Falls $[\delta_0] > \frac{\rho}{2} \frac{\left[h_0^\delta\right]_1}{1+\left[h_0^\delta\right]_1}$, gehe zu Schritt 0 und beginne neu mit größerem X_0
Falls $[\delta_n] > \frac{\rho}{2} \frac{\left[h_n^\delta\right]}{1+\left[h_n^\delta\right]}$, vergrößere X_n und wiederhole Schritt 1

Schritt 2 Setze $x_{n+1} := x_n + \delta x_n$ und $X_{n+1} := X_n$
Falls $||\delta x_n|| < \tau_{rel}\,||\delta x_0|| + \tau_{abs}$, beende die Iteration erfolgreich
Setze $n := n + 1$ und gehe zu Schritt 1

Abbildung 3.3: Algorithmus für den quadratischen Konvergenzmodus

Da in praktischen Rechnungen meist nicht klar ist, ob die Voraussetzungen von Satz 3.4 gelten oder nicht, empfiehlt sich die Verwendung eines Konvergenzmonitors, der eine eventuelle Divergenz feststellt. Es wäre ideal, beobachten zu können, ob $\theta_n = \frac{||\Delta x_{n+1}||}{||\Delta x_n||} \leq \bar{\theta} < 1$ gilt. Die Größen Δx_{n+1} und Δx_n stehen jedoch nicht zur Verfügung, da der Newton-Schritt nur inexakt gelöst wird. Die Ungleichung (3.8) motiviert die folgenden beiden Konvergenzmonitore, die zur Feststellung der Konvergenz oder einer möglichen Divergenz dienen:

- **Notwendiger Konvergenzmonitor**

$$\hat{\theta}_n := \frac{1-[\delta_{n+1}]}{1+[\delta_n]} \frac{||\delta x_{n+1}||}{||\delta x_n||} \leq \theta_n \tag{3.17}$$

- **Hinreichender Konvergenzmonitor**

$$\theta_n \leq \hat{\hat{\theta}}_n := \frac{1+[\delta_{n+1}]}{1-[\delta_n]} \frac{||\delta x_{n+1}||}{||\delta x_n||} \tag{3.18}$$

Falls im Verlauf der Iteration ein $\hat{\theta}_n > 1$ auftritt, dann sollte die Iteration als „nicht konvergent“ beendet werden.

Eventuell zahlt es sich aus, nur die „a-postposteriori“ Schätzung $\left[h_n^\delta\right]_1$ zur Überprüfung der geforderten Genauigkeit zu verwenden. Der Vorteil ist, dass $\left[h_n^\delta\right]_1$ eventuell näher an h_n^δ liegt als $\left[h_n^\delta\right]$ und somit noch Schritte akzeptiert werden, welche sonst als „zu ungenau“ abgewiesen werden würden. Als nachteilig erweist sich, dass immer gleich zwei Schritte verworfen werden müssten, falls der Newton-Schritt nicht genau genug gelöst wurde.

Für die Wahl von ρ wird $\rho = 1$ vorgeschlagen. Eine kleine Wahl von ρ verlangt, dass der Newton-Schritt mit einer höheren Genauigkeit gelöst wird, aber erlaubt, dass $||\Delta x_0||$ größer ist als bei einer großen Wahl von ρ.

Das Abbruchkriterium $||\delta x_n|| < \tau_{rel}\,||\delta x_0|| + \tau_{abs}$ ist durch folgende Abschätzung motiviert. Mit $x^* = \lim\limits_{n\to\infty} x_n$ gilt:

$$\begin{aligned}
||x_n - x^*|| &= \left|\left|\sum_{i=n}^{\infty} \delta x_i\right|\right| \\
&\leq \sum_{i=n}^{\infty} ||\delta x_i|| \\
&\leq \sum_{i=n}^{\infty} \left(\frac{1+\rho}{2\left(1-\bar{\delta}\right)}\omega\right)^{2^{i-n}-1} ||\delta x_n||^{2^{i-n}} \qquad \text{mit (3.7)} \\
&= \frac{2\left(1-\bar{\delta}\right)}{(1+\rho)\,\omega} \sum_{i=0}^{\infty} \left(\frac{1+\rho}{2\left(1-\bar{\delta}\right)}\omega\,||\delta x_n||\right)^{2^i} .
\end{aligned}$$

Das heißt $||x_n - x^*||$ ist klein, falls $\frac{1+\rho}{2\left(1-\bar{\delta}\right)}\omega\,||\delta x_n||$ klein ist. Da für ω keine obere Schätzung existiert, kann auch für $\frac{1+\rho}{2\left(1-\bar{\delta}\right)}\omega$ keine angegeben werden. Es wird daher nur auf $||\delta x_n||$ zurückgegriffen.

3.2.2 Linearer Konvergenzmodus

Zwar gilt

$$\vartheta(h_n^\delta, \delta_n) \leq \vartheta(h_n^\delta, [\delta_n]) \quad \text{und} \quad \bar{\vartheta}(h_n^\delta, \delta_n, \delta_{n+1}) \leq \bar{\vartheta}(h_n^\delta, [\delta_n], [\delta_{n+1}]) \;,$$

jedoch sind ω und damit h_n^δ meist unbekannt und es lassen sich nur schwer obere Schätzungen dafür finden. Daher ist eine direkte algorithmische Umsetzung von Satz 3.8 nicht möglich, Abhilfe schafft Satz 3.9. Er besagt, dass es genügt, neben einer Startbedingung, die relative Genauigkeit δ_n unter einem Schwellwert $\bar{\delta} < \frac{1}{2}$ zu halten, um lineare Konvergenz mit dem Konvergenzfaktor $\theta \geq \frac{\bar{\delta}}{1-\delta}$ zu sichern. Entfernt man aus dem obigen Algorithmus die fürs Schätzen benötigten Stellen und ersetzt das Überprüfen der Genauigkeit durch

$$[\delta_n] \leq \bar{\delta} \ ,$$

so erhält man einen Algorithmus für den linearen Konvergenzmodus.

4 Die globale Phase

In diesem Kapitel wird eine Methode vorgestellt, die auf die Eingabe von „guten“ Startpunkten x_0 mit kleinem $||\Delta x_0||$ verzichtet. Das dabei motivierte Verfahren ist keineswegs global konvergent. Jedoch lässt es sich leichter anwenden, da häufig keine guten Startpunkte zur Verfügung stehen. Die Notation in diesem Kapitel richtet sich nach [Deu04]. Im Anschluss an einige Aussagen aus diesem Buch folgen Überlegungen zu einer eigenen Strategie.

Die affine Invarianz bzw. affine Kovarianz steht wieder im Vordergrund. Ein Grund für die Betrachtung der affinen Kovarianz besteht darin, dass sich die Aussagen der Sätze nicht ändern sollen, falls F mit einem linearen Operator $\mathrm{A} \in \bar{\mathcal{L}}(Y, Z)$ zu $\mathrm{A}\,\mathrm{F}$ skaliert wird (Z bezeichnet neben X und Y einen weiteren Banachraum). Das steht in direktem Zusammenhang mit der hier gestellten Bedingung, die Norm des Raumes Y ($\mathrm{F} : X \to Y$) nicht zu verwenden. Das Newton-Verfahren selbst ist invariant gegenüber einer solchen Skalierung, d.h. die Folge der Iterierten x_n ist unabhängig von A.

4.1 Theoretische Aussagen

Für die globale Phase stellt sich die Frage, unter welcher Bedingung ein neuer Punkt x_{n+1} als Fortschritt gegenüber x_n anzusehen ist. Ein Kriterium der Form $||\Delta x_{n+1}|| < ||\Delta x_n||$ ist nur in der lokalen Phase sinnvoll. Das häufig verwendete Abstiegskriterium für das Residuum $||\mathrm{F}(x_{n+1})|| < ||\mathrm{F}(x_n)||$ kann nicht benutzt werden, denn es benötigt die Norm in Y und ist nicht affin kovariant. Vielmehr setzt es voraus, dass F angemessen skaliert ist. Um die affine Kovarianz zu gewährleisten, ist es notwendig, die Bedingung

$$||\mathrm{A}\,\mathrm{F}(x_{n+1})|| < ||\mathrm{A}\,\mathrm{F}(x_n)|| \tag{4.1}$$

für alle $\mathrm{A} \in \bar{\mathcal{L}}(Y, Z)$ zu fordern. Zur Vereinfachung der Notation wird zunächst die allgemeine Levelfunktion eingeführt.

Definition 4.1 Für $\mathrm{F} : X \to Y$ und $\mathrm{A} \in \bar{\mathcal{L}}(Y, Z)$ bezeichnet

$$\mathrm{T}(\cdot|\,\mathrm{A}) : X \to Z, x \mapsto \frac{1}{2}\,||\mathrm{A}\,\mathrm{F}(x)||^2$$

die allgemeine Levelfunktion. Die dazugehörige Levelmenge $\mathrm{G}(z|\,\mathrm{A})$ ist durch

$$\mathrm{G}(z|\,\mathrm{A}) := \{x \in X : \mathrm{T}(x|\,\mathrm{A}) \leq \mathrm{T}(z|\,\mathrm{A})\}$$

definiert.

Damit lässt sich Gleichung (4.1) äquivalent

$$\mathrm{T}(x_{n+1}|\,\mathrm{A}) < \mathrm{T}(x_n|\,\mathrm{A}) \tag{4.2}$$

ausdrücken. Es ist nicht klar, ob neben x^* mit $\mathrm{F}(x^*) = 0$ weitere Punkte existieren, die diese Bedingung für alle $\mathrm{A} \in \bar{\mathcal{L}}(Y, Z)$ erfüllen. Für den endlich-dimensionalen Fall mit $\mathrm{F} : \mathbb{R}^N \to \mathbb{R}^N$ und invertierbarem $\mathrm{A} \in \mathbb{R}^{N\times N}$ wurde in [Deu04] unter geeigneten Voraussetzungen die Existenz des sogenannten Newton-Pfades $\bar{x}(\lambda)$ mit $\lambda \in [0, 1]$ gezeigt. Für diesen Pfad gilt:

$$\begin{aligned}\bar{x}(0) &= x_0\ ,\\ \bar{x}(1) &= x^*\ ,\\ \mathrm{F}(\bar{x}(\lambda)) &= (1-\lambda)\,\mathrm{F}(x_0)\end{aligned}$$

und damit

$$\mathrm{T}(\bar{x}(\lambda)|\,\mathrm{A}) = (1-\lambda)^2\,\mathrm{T}(x_0|\,\mathrm{A}) \tag{4.3}$$

für alle invertierbaren $\mathrm{A} \in \mathbb{R}^{N\times N}$. Er ist als Lösung der gewöhnlichen Differentialgleichung

$$\frac{\mathrm{d}\bar{x}}{\mathrm{d}\lambda} = -\,\mathrm{F}'(\bar{x})^{-1}\,\mathrm{F}(x_0) \qquad \text{mit} \qquad \bar{x}(0) = x_0$$

gegeben und lässt sich numerisch nur bedingt bestimmen. Die Gleichung (4.3) sichert den

Abstieg für jede Levelfunktion entlang des Pfades. Jeder Punkt des Newton-Pfades erfüllt die Gleichung (4.2) für alle $\mathrm{A} \in \mathbb{R}^{N \times N}$ und kann (sofern verfügbar) als neuer Iterationspunkt akzeptiert werden. Der Beweis verwendet neben dem Satz über implizite Funktionen auch ein Kompaktheitsargument sowie die Eigenvektorzerlegung der Matrix $\mathrm{A}^T\,\mathrm{A}$. Deshalb lässt er sich nicht problemlos auf den allgemeineren unendlich-dimensionalen Fall übertragen.

Die Forderung, dass Gleichung (4.2) für alle $\mathrm{A} \in \bar{\mathcal{L}}(Y, Z)$ gilt, ist hier also zu streng. Abgeschwächt wird sie, indem die folgenden Betrachtungen für ein beliebiges, aber festes $\mathrm{A} \in \bar{\mathcal{L}}(Y, Z)$ gemacht werden.

Das nächste Lemma besagt, dass für $\mathrm{A} \in \bar{\mathcal{L}}(Y, Z)$ die Newton-Richtung Δx stets eine Abstiegsrichtung bzgl. $\mathrm{T}(x|\,\mathrm{A})$ darstellt.

Lemma 4.2 *Es sei Z ein reeller Hilbertraum mit Skalarprodukt $(\cdot,\cdot)$. Des Weiteren sei die Funktion $\mathrm{F} : X \to Y$ in x Fréchet-differenzierbar mit $\mathrm{F}(x) \neq 0$ und $\mathrm{F}'(x) \in \bar{\mathcal{L}}(X, Y)$. Dann gilt für $\mathrm{A} \in \bar{\mathcal{L}}(Y, Z)$*

$$\mathrm{T}'(x|\,\mathrm{A})\Delta x < 0$$

mit $\Delta x = -\,\mathrm{F}'(x)^{-1}\,\mathrm{F}(x)$.

Beweis. Mit der Kettenregel für die Fréchet-Differenzierbarkeit ergibt sich:

$$\mathrm{T}'(x|\,\mathrm{A})\Delta x = \big(\mathrm{A}\,\mathrm{F}(x), \mathrm{A}\,\mathrm{F}'(x)\Delta x\big) = -\,(\mathrm{A}\,\mathrm{F}(x), \mathrm{A}\,\mathrm{F}(x)) = -\,||\mathrm{A}\,\mathrm{F}(x)||^2 < 0\ .$$

□

Dieses Lemma motiviert die Betrachtung des gedämpften Newton-Verfahrens, denn der volle Newton-Schritt Δx ist unter Umständen zu lang. Es ist durch

$$x_{n+1} = x_n + \lambda_n \Delta x_n$$

mit $\Delta x_n = -\,\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n)$ und $\lambda_n \in (0, 1]$ sowie $x_0 \in X$ erklärt.

Das nächste Lemma und der darauf folgende Satz zeigen die globale Konvergenz eines „geeignet gedämpften“ Newton-Verfahrens. Die globale Konvergenz ist bezüglich $\mathrm{T}(x|\,\mathrm{A})$

für ein festes $\mathrm{A} \in \bar{\mathcal{L}}(Y,Z)$ zu verstehen. Das Vorgehen richtet sich dabei nach den Ausführungen in [Deu04][1].

Lemma 4.3 *Es sei* $\mathrm{F} : X \to Y$ *eine in* X *stetig Fréchet-differenzierbare Funktion mit* $\mathrm{F}'(x) \in \bar{\mathcal{L}}(X,Y)$ *für alle* $x \in X$. *Des Weiteren gelte die affin kovariante Lipschitz-Bedingung*

$$\left|\left|\mathrm{F}'(x)^{-1}\left(\mathrm{F}'(y)-\mathrm{F}'(x)\right)(y-x)\right|\right| \leq \omega\,||y-x||^2$$

für alle $x, y \in X$.
Dann gilt für einen Punkt $x \in X$ *und* $\mathrm{A} \in \bar{\mathcal{L}}(Y,Z)$ *die Beziehung*

$$\mathrm{T}(x+\lambda\Delta x|\,\mathrm{A}) \leq t(\lambda|\,\mathrm{A})^2\,\mathrm{T}(x|\,\mathrm{A})$$

mit dem Polynom

$$t(\lambda|\,\mathrm{A}) = 1-\lambda+\frac{1}{2}\tilde{h}\lambda^2\ .$$

Hierbei bezeichnet $\tilde{h} = h\,||\mathrm{A}\,\mathrm{F}'(x)||\left|\left|\left(\mathrm{A}\,\mathrm{F}'(x)\right)^{-1}\right|\right|$ *und* $h = \omega\,||\Delta x||$ *sowie* $\Delta x = -\,\mathrm{F}'(x)^{-1}\,\mathrm{F}(x)$.

Beweis. Es gilt:

$$\begin{aligned}
||\mathrm{A}\,\mathrm{F}(x+\lambda\Delta x)|| &= \left|\left|\mathrm{A}\left(\mathrm{F}(x+\lambda\Delta x)-\mathrm{F}(x)-\mathrm{F}'(x)\Delta x\right)\right|\right| \\
&= \left|\left|\mathrm{A}\left(\int_0^\lambda \left(\mathrm{F}'(x+t\Delta x)-\mathrm{F}'(x)\right)\Delta x\,\mathrm{d}t-(1-\lambda)\,\mathrm{F}'(x)\Delta x\right)\right|\right| \\
&\leq (1-\lambda)\left|\left|\mathrm{A}\,\mathrm{F}'(x)\Delta x\right|\right| + \left|\left|\mathrm{A}\,\mathrm{F}'(x)\int_0^\lambda \mathrm{F}'(x)^{-1}\left(\mathrm{F}'(x+t\Delta x)-\mathrm{F}'(x)\right)\Delta x\,\mathrm{d}t\right|\right| \\
&\leq (1-\lambda)\,||\mathrm{A}\,\mathrm{F}(x)|| + \left|\left|\mathrm{A}\,\mathrm{F}'(x)\right|\right|\frac{1}{2}\omega\lambda^2\,||\Delta x||^2 \\
&\leq (1-\lambda)\,||\mathrm{A}\,\mathrm{F}(x)|| + \frac{1}{2}h\lambda^2\left|\left|\mathrm{A}\,\mathrm{F}'(x)\right|\right|\left|\left|\left(\mathrm{A}\,\mathrm{F}'(x)\right)^{-1}\mathrm{A}\,\mathrm{F}(x)\right|\right| \\
&\leq (1-\lambda)\,||\mathrm{A}\,\mathrm{F}(x)|| + \frac{1}{2}\tilde{h}\lambda^2\,||\mathrm{A}\,\mathrm{F}(x)|| \\
&\leq \left(1-\lambda+\frac{1}{2}\tilde{h}\lambda^2\right)||\mathrm{A}\,\mathrm{F}(x)||\ .
\end{aligned}$$

□

[1] Satz 3.12 und 3.13, Seite 135-137

Das Lemma schätzt den Fortschritt bzgl. $\mathrm{T}(x|\,\mathrm{A})$ entlang der Newton-Richtung Δx ab. Das dabei auftretende Polynom $t(\lambda|\,\mathrm{A})$ nimmt an der Stelle $\tilde{\lambda}(\mathrm{A}) := \frac{1}{\tilde{h}}$ sein Minimum an. Falls in jedem Iterationsschritt ein Mindestfortschritt garantiert werden kann, konvergiert $\mathrm{T}(x_n|\,\mathrm{A})$ gegen 0. Das ist im folgenden Satz formuliert.

Satz 4.4 *Es sei* $\mathrm{A} \in \bar{\mathcal{L}}(Y,Z)$ *und* $\mathrm{F} : X \to Y$ *eine in* X *stetig Fréchet-differenzierbare Funktion mit* $\mathrm{F}'(x) \in \bar{\mathcal{L}}(X,Y)$ *für alle* $x \in X$*. Außerdem gelte die affin kovariante Lipschitz-Bedingung*

$$\left|\left|\mathrm{F}'(x)^{-1}\left(\mathrm{F}'(y)-\mathrm{F}'(x)\right)(y-x)\right|\right| \leq \omega\,||y-x||^2$$

für alle $x, y \in X$*. Zusätzlich sei ein* $x_0 \in X$ *gegeben und*

$$H := \omega \sup_{x\in \mathrm{G}(x_0|\,\mathrm{A})} \left|\left|\mathrm{F}'(x)^{-1}\,\mathrm{F}(x)\right|\right| \left|\left|\mathrm{A}\,\mathrm{F}'(x)\right|\right| \left|\left|\left(\mathrm{A}\,\mathrm{F}'(x)\right)^{-1}\right|\right| < \infty\ . \tag{4.4}$$

Dann existiert ein $\varepsilon > 0$*, so dass für das gedämpfte Newton-Verfahren mit* $\lambda_n \in [\varepsilon, 2\tilde{\lambda}_n(\mathrm{A})-\varepsilon]$

$$\lim_{n\to\infty} \mathrm{T}(x_n|\,\mathrm{A}) = 0$$

gilt. Dabei bezeichnet $\tilde{\lambda}_n(\mathrm{A}) = \frac{1}{\tilde{h}_n}$ *und* $\tilde{h}_n = h_n\,||\mathrm{A}\,\mathrm{F}'(x_n)||\left|\left|\left(\mathrm{A}\,\mathrm{F}'(x_n)\right)^{-1}\right|\right|$ *sowie* $h_n = \omega\,||\Delta x_n||$*.*

Beweis. Es wird gezeigt, dass

$$\mathrm{T}(x_{n+1}|\,\mathrm{A}) \leq \theta\,\mathrm{T}(x_n|\,\mathrm{A}) \tag{4.5}$$

für alle n mit $\theta < 1$ gilt. Daraus folgt $\mathrm{G}(x_{n+1}|\,\mathrm{A}) \subset \mathrm{G}(x_n|\,\mathrm{A})$ und per Induktion die Behauptung.
Nach Lemma 4.3 gilt

$$\mathrm{T}(x_n + \lambda\Delta x_n|\,\mathrm{A}) \leq t_n(\lambda|\,\mathrm{A})^2\,\mathrm{T}(x_n|\,\mathrm{A}) \tag{4.6}$$

mit $t_n(\lambda|\,\mathrm{A}) = 1 - \lambda + \frac{1}{2}\tilde{h}_n\lambda^2$. Das Polynom $t_n(\lambda|\,\mathrm{A})$ lässt sich durch

$$t_n(\lambda|\,\mathrm{A}) \leq \begin{cases} 1 - \frac{1}{2}\lambda & \text{, für } 0 \leq \lambda \leq \frac{1}{\tilde{h}_n} \\ 1 + \frac{1}{2}\lambda - \frac{1}{\tilde{h}_n} & \text{, für } \frac{1}{\tilde{h}_n} \leq \lambda \leq \frac{2}{\tilde{h}_n} \end{cases}$$

abschätzen. Wegen Voraussetzung (4.4) gilt $\frac{1}{\tilde{h}_n} \geq \frac{1}{H}$ für alle n. Daher kann ein ε mit $0 < \varepsilon < \frac{1}{H}$ gewählt werden. Mit diesem gilt:

$$t_n(\lambda_n|\,\mathrm{A}) \leq 1 - \varepsilon\ .$$

Daraus folgt (4.5) mit $\theta = (1-\varepsilon)^2 < 1$ und somit die Behauptung. □

Die Bedingung (4.4) wirkt ungewöhnlich. Sie entsteht bei der Übertragung des Satzes auf Banachräume. Im endlich-dimensionalen Fall genügt es, wenn $\mathrm{G}(x_0|\,\mathrm{A}) \subset \mathrm{D}_0$ mit kompakten D_0 gilt, um daraus die Bedingung (4.4) folgern zu können. Zusätzlich besagt der Satz von Bolzano-Weierstraß, dass die Folge x_n dann mindestens einen Häufungspunkt besitzt.

Insgesamt zeigt Satz 4.4, dass das gedämpfte Newton-Verfahren für ein beliebiges, aber festes $\mathrm{A} \in \bar{\mathcal{L}}(Y,Z)$ unter der Beschränktheitsvoraussetzung (4.4) bzgl. $\mathrm{T}(x_n|\,\mathrm{A})$ konvergiert. Für die affine Kovarianz müsste eine solche Aussage vielmehr für alle $\mathrm{A} \in \bar{\mathcal{L}}(Y,Z)$ gelten. Wie jedoch die folgenden Überlegungen nahelegen, ist dies nicht ohne Weiteres möglich. Der Beweis des Satzes benutzt die Voraussetzung (4.4), um zu kurze Schrittweiten auszuschließen. Die aus der Abschätzung (4.6) resultierende optimale Schrittweite für Δx_n ist das Minimum des Polynoms $t_n(\lambda|\,\mathrm{A})$ und durch

$$\tilde{\lambda}_n(\mathrm{A}) = \frac{1}{\tilde{h}_n} = \frac{1}{\omega\,||\Delta x_n||\,\kappa(\mathrm{A}\,\mathrm{F}'(x_n))}$$

mit der Kondition $\kappa(\mathrm{A}\,\mathrm{F}'(x_n)) = ||\mathrm{A}\,\mathrm{F}'(x_n)||\,\left|\left|(\mathrm{A}\,\mathrm{F}'(x_n))^{-1}\right|\right|$ gegeben. Die Kondition des Operators $\mathrm{A}\,\mathrm{F}'(x_n)$ ist für die optimale Schrittweite ausschlaggebend. Im Zusammenhang mit der affinen Kovarianz ergibt es keinen Sinn, eine gute Kondition zu erwarten. Erlaubt man jedoch A vom Iterationsindex n abzuhängen und wählt $\mathrm{A}_n = \mathrm{F}'(x_n)^{-1}$ $(Z = X)$, gilt $\kappa(\mathrm{A}_n\,\mathrm{F}'(x_n)) = 1$. Für diese Wahl tritt die größte optimale Schrittweite auf. Wie bereits in Lemma 4.2 erwähnt, ist die Newton-Richtung für alle $\mathrm{A} \in \bar{\mathcal{L}}(Y,X)$ (mit X Hilbertraum)

eine Abstiegsrichtung bzgl. $\mathrm{T}(x|\,\mathrm{A})$. Für $\mathrm{A} = \mathrm{F}'(x_n)^{-1}$ handelt es sich an der Stelle $x = x_n$ sogar um die Richtung des steilsten Abstiegs, denn für eine Richtung $\delta x \in X$ gilt:

$$\mathrm{T}'(x_n|\,\mathrm{F}'(x_n)^{-1})\delta x = \left(\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n), \mathrm{F}'(x_n)^{-1}\,\mathrm{F}'(x_n)\delta x\right) = -\left(\Delta x_n, \delta x\right) \ .$$

Damit ist $-\Delta x_n$ der Riesz-Repräsentant von $\mathrm{T}'(x_n|\,\mathrm{F}'(x_n)^{-1})$ bzgl. $(\cdot,\cdot)$.
Der Fortschritt der Funktion $\mathrm{T}(x_n|\,\mathrm{A}_n)$ in Richtung Δx_n lässt sich in der Form

$$\begin{aligned} & \mathrm{T}(x_n + \lambda\Delta x_n|\,\mathrm{F}'(x_n)^{-1}) < \mathrm{T}(x_n|\,\mathrm{F}'(x_n)^{-1}) \\ \iff \quad & \left|\left|\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n + \lambda\Delta x_n)\right|\right| < \left|\left|\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n)\right|\right| \\ \iff \quad & ||\mathrm{F}(x_n + \lambda\Delta x_n)||_{\mathrm{F}'(x_n)^{-1}} < ||\mathrm{F}(x_n)||_{\mathrm{F}'(x_n)^{-1}} \end{aligned}$$

darstellen und wird auch das natürliche Abstiegskriterium genannt. Dabei bezeichnet $||\cdot||_{\mathrm{F}'(x_n)^{-1}} := \left|\left|\mathrm{F}'(x_n)^{-1}\cdot\right|\right|$ die lokale Norm in x_n. Eine Interpretation ist, dass im neuen Punkt das „alte“ lineare Modell der Funktion F einen kürzeren Schritt verspricht als im aktuellen Punkt x_n. Im eindimensionalen Fall stimmt es mit dem weit verbreiteten Abstiegskriterium für das Residuum (d.h. $||\mathrm{F}(x_n + \lambda\Delta x_n)|| < ||\mathrm{F}(x_n)||$) überein. Wie in [AO87] gezeigt wird, kann für das natürliche Abstiegskriterium im Allgemeinen keine globale Konvergenz gelten. Möglicherweise wiederholen sich die Iterationen zyklisch, d.h. das Newton-Verfahren „kreiselt“. Dennoch wird es häufig als Kriterium für den Fortschritt herangezogen. So beruhen auch die folgenden Überlegungen auf der Annahme, dass eine Reduktion bzgl. der lokalen Norm $||\cdot||_{\mathrm{F}'(x_n)^{-1}}$ erstrebenswert ist.

Ziel bei dem Entwurf eines Verfahrens ist, ausgehend vom aktuellen Iterationspunkt x_n, einen neuen Punkt x_{n+1} zu finden, so dass der Kontraktionsterm $\hat{\theta}_n := \frac{\left|\left|\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_{n+1})\right|\right|}{||\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n)||}$ möglichst klein wird. Genauso wie in Kapitel 3 muss die Newton-Gleichung $\Delta x_n = -\,\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n)$ nicht exakt lösbar sein und damit auch nicht $\overline{\Delta x}_{n+1} := -\,\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_{n+1})$. Schwierigkeiten bestehen in der Steuerung der Genauigkeit, mit der diese Gleichungen gelöst werden, und im Finden einer Schrittweite λ_n, die dem natürlichen Abstiegskriterium genügt.
Analog zum Kapitel 3 bezeichnet δx_n die inexakte Lösung der Newton-Richtung. Sie ist zusammen mit dem Fehler r_n durch

$$\mathrm{F}'(x_n)\delta x_n = -\,\mathrm{F}(x_n) + r_n$$

gegeben. Für die relative Genauigkeit wird wieder δ_n mit

$$\delta_n := \frac{||\Delta x_n - \delta x_n||}{||\delta x_n||} = \frac{\left|\left|\mathrm{F}'(x_n)^{-1} r_n\right|\right|}{||\delta x_n||}$$

verwendet, wobei Δx_n die exakte Newton-Richtung $-\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n)$ ist. Anschließend wird entlang der Strecke $x_n + \lambda\delta x_n$ ein $\lambda_n \in (0,1]$ gesucht, so dass

$$\hat{\theta}_n := \frac{\left|\left|\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n + \lambda_n \delta x_n)\right|\right|}{||\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n)||} < 1$$

gilt und $x_n + \lambda_n \delta x_n$ als neue Iterierte x_{n+1} akzeptiert wird. Wie bereits erwähnt, kann auch $\overline{\Delta x}_{n+1} = -\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_{n+1})$ nicht exakt gelöst werden. Daher bezeichnet $\overline{\delta x}_{n+1}$ die Näherungslösung, welche zusammen mit dem Fehler $\bar{r}_{n+1}$ durch

$$\mathrm{F}'(x_n)\overline{\delta x}_{n+1} = -\mathrm{F}(x_{n+1}) + \bar{r}_{n+1}$$

gegeben ist. Analog wird die relative Genauigkeit $\bar{\delta}_{n+1}$ mit

$$\bar{\delta}_{n+1} := \frac{\left|\left|\overline{\Delta x}_{n+1} - \overline{\delta x}_{n+1}\right|\right|}{\left|\left|\overline{\delta x}_{n+1}\right|\right|} = \frac{\left|\left|\mathrm{F}'(x_n)^{-1}\bar{r}_{n+1}\right|\right|}{\left|\left|\overline{\delta x}_{n+1}\right|\right|}$$

definiert.

Das folgende Lemma bildet die Grundlage für die Entwicklung einer Strategie zur Steuerung der Genauigkeiten δ_n, $\bar{\delta}_{n+1}$ und der Wahl der Schrittweite λ_n.

Lemma 4.5 *Es sei* F *eine stetig Fréchet-differenzierbare Funktion mit* $\mathrm{F}'(x) \in \bar{\mathcal{L}}(X,Y)$ *für alle* $x \in X$*. Außerdem gelte die affin kovariante Lipschitz-Bedingung*

$$\left|\left|\mathrm{F}'(x)^{-1}\left(\mathrm{F}'(y) - \mathrm{F}'(x)\right)(y-x)\right|\right| \leq \omega\,||y-x||^2 \tag{4.7}$$

für alle $x, y \in X$*.*
Dann gilt für das gedämpfte Newton-Verfahren:

$$\frac{\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} \leq 1 + \delta_n - \lambda + \frac{1}{2}\lambda^2 h_n^\delta$$

mit $h_n^\delta = \omega\,||\delta x_n||$ *und* $\overline{\Delta x}_{n+1}(\lambda) := -\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n + \lambda\delta x_n)$*.*

Beweis. Es gilt:

$$\begin{aligned}
\overline{\Delta x}_{n+1}(\lambda) &= -\,\mathrm{F}'(x_n)^{-1}\,\mathrm{F}(x_n+\lambda\delta x_n)\\
&= \Delta x_n - \mathrm{F}'(x_n)^{-1}\left(\mathrm{F}(x_n+\lambda\delta x_n)-\mathrm{F}(x_n)\right)\\
&= \Delta x_n - \mathrm{F}'(x_n)^{-1}\int_0^\lambda \mathrm{F}'(x_n+s\delta x_n)\delta x_n\,\mathrm{d}s\\
&= \Delta x_n - \lambda\delta x_n - \mathrm{F}'(x_n)^{-1}\int_0^\lambda \left(\mathrm{F}'(x_n+s\delta x_n)-\mathrm{F}'(x_n)\right)\delta x_n\,\mathrm{d}s
\end{aligned}$$

und damit

$$\begin{aligned}
\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right| &\le \lambda\,||\Delta x_n-\delta x_n|| + (1-\lambda)\,||\Delta x_n||\\
&\qquad + \left|\left|\int_0^\lambda \mathrm{F}'(x_n)^{-1}\left(\mathrm{F}'(x_n+s\delta x_n)-\mathrm{F}'(x_n)\right)\delta x_n\,\mathrm{d}s\right|\right|\\
&\le \lambda\delta_n\,||\delta x_n|| + (1-\lambda)\,(1+\delta_n)\,||\delta x_n|| + \frac{1}{2}\omega\lambda^2\,||\delta x_n||^2 \qquad \text{(mit (3.8))}\\
&= (1+\delta_n)\,||\delta x_n|| - \lambda\,||\delta x_n|| + \frac{1}{2}\omega\lambda^2\,||\delta x_n||^2\ .
\end{aligned}$$

□

Bemerkung 4.6 *Es genügt, wenn an Stelle der globalen Lipschitz-Bedingung* (4.7) *die lokale, affin kovariante Lipschitz-Bedingung*

$$\left|\left|\mathrm{F}'(x_n)^{-1}\left(\mathrm{F}'(x_n+s\delta x_n)-\mathrm{F}'(x_n)\right)\delta x_n\right|\right| \le \omega_n s\,||\delta x_n||^2 \tag{4.8}$$

für $s\in[0,1]$, $\delta x_n \in \mathcal{B}_{\frac{4}{3}||\Delta x_n||}(\Delta x_n)$ und alle Iterierten x_n gilt. Hierbei bezeichnet $\mathcal{B}_\delta(x)$ die abgeschlossene Kugel um x mit Radius δ, d.h. $\mathcal{B}_\delta(x)=\{y\in X : ||y-x||\le\delta\}$. Falls X ein Hilbertraum ist, genügt es, die Bedingung (4.8) *nur für alle $\delta x_n \in \mathcal{B}_{\frac{4}{15}||\Delta x_n||}(\frac{16}{15}\Delta x)$ zu fordern. Dabei handelt es sich um die in Proposition 3.7 beschriebenen Kugeln mit $\delta=\frac{1}{4}$. An späterer Stelle wird erklärt, warum $\delta_n\le\frac{1}{4}$ eine vernünftige Forderung ist.*

Insgesamt ergibt sich aus Bedingung (4.8) *die Abschätzung*

$$\frac{\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} \leq 1 + \delta_n - \lambda + \frac{1}{2}\lambda^2 h_n^\delta \tag{4.9}$$

mit $h_n^\delta = \omega_n \, ||\delta x_n||$*, von der im Folgenden ausgegangen wird.*
Zudem reicht es aus, wenn der inverse Operator von $\mathrm{F}'(x)$ *nur an den Iterationspunkten* x_n *existiert.*

Für den exakten Fall (d.h. $\delta_n = 0$) ergibt sich

$$\frac{\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right|}{||\Delta x_n||} \leq 1 - \lambda + \frac{1}{2}\lambda^2 h_n$$

mit $h_n = \omega_n \, ||\Delta x_n||$. Selbst für das exakte Lösen muss nicht $\frac{\left|\left|\overline{\Delta x}_{n+1}(1)\right|\right|}{||\Delta x_n||} < 1$ gelten. Jedoch existiert ein $\bar{\lambda}_n$, so dass

$$\frac{\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right|}{||\Delta x_n||} \leq 1 - \lambda + \frac{1}{2}\lambda^2 h_n < 1$$

für $\lambda \in (0, \bar{\lambda}_n)$ gilt. Diese Eigenschaft soll für den hier behandelten inexakten Fall erhalten bleiben. Dafür ist es vorteilhaft, die Genauigkeit δ_n in Abhängigkeit von der Schrittweite λ zu steuern. An dieser Stelle wird die in [Deu04] verfolgte Strategie verlassen und an δ_n die Forderung

$$\delta_n \leq \rho\lambda \tag{4.10}$$

mit $0 \leq \rho < 1$ gestellt. Der Einfachheit halber soll auch für $\bar{\delta}_{n+1}$

$$\bar{\delta}_{n+1} \leq \rho\lambda \tag{4.11}$$

gelten. Genauso wie in Kapitel 3 müssen die Größen δ_n und $\bar{\delta}_{n+1}$ nicht zur Verfügung stehen. Es genügt, wenn die Schätzungen $[\delta_n]$ und $\left[\bar{\delta}_{n+1}\right]$ vorhanden sind, d.h.

$$\delta_n \leq [\delta_n] \qquad \text{und} \qquad \bar{\delta}_{n+1} \leq \left[\bar{\delta}_{n+1}\right] \ .$$

Die Genauigkeitsbedingungen (4.10) und (4.11) gehen in die überprüfbaren Kriterien

$$[\delta_n] \leq \rho\lambda \qquad \text{und} \qquad \left[\bar{\delta}_{n+1}\right] \leq \rho\lambda \tag{4.12}$$

über. Wegen $\lambda \leq 1$ gilt damit $\delta_n \leq [\delta_n] \leq \rho$ und $\bar{\delta}_{n+1} \leq \left[\bar{\delta}_{n+1}\right] \leq \rho$. Zusammen mit der häufig verwendeten Ungleichung (3.8) und der Beziehung (4.9) ergibt dies:

$$\begin{aligned}
\hat{\theta}_n(\lambda) := \frac{\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right|}{||\Delta x_n||} &\leq \frac{1+\bar{\delta}_{n+1}}{1-\delta_n} \frac{\left|\left|\overline{\delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} \\
&\leq \frac{1+\left[\bar{\delta}_{n+1}\right]}{1-[\delta_n]} \frac{\left|\left|\overline{\delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} =: \tilde{\theta}_n(\lambda) \\
&\leq \frac{1}{1-[\delta_n]} \frac{1+\left[\bar{\delta}_{n+1}\right]}{1-\left[\bar{\delta}_{n+1}\right]} \frac{\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} \\
&\leq \frac{1}{1-[\delta_n]} \frac{1+\left[\bar{\delta}_{n+1}\right]}{1-\left[\bar{\delta}_{n+1}\right]} \left(1+\delta_n-\lambda+\frac{1}{2}\lambda^2 h_n^\delta\right) \\
&\leq \frac{1}{1-[\delta_n]} \frac{1+\left[\bar{\delta}_{n+1}\right]}{1-\left[\bar{\delta}_{n+1}\right]} \left(1+\delta_n-\lambda+\frac{1}{2\,(1-\rho)}\lambda^2 h_n\right) \\
&\leq \frac{1+\rho\lambda}{(1-\rho\lambda)^2} \left(1+(\rho-1)\,\lambda+\frac{1}{2\,(1-\rho)}\lambda^2 h_n\right) =: \varphi_n(\lambda)\ .
\end{aligned} \tag{4.13}$$

Eine Schrittweite λ wird akzeptiert, falls die überprüfbare Bedingung

$$\tilde{\theta}_n(\lambda) = \frac{1+\left[\bar{\delta}_{n+1}\right]}{1-[\delta_n]} \frac{\left|\left|\overline{\delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} < 1$$

erfüllt ist. Damit ist nach obiger Abschätzung das gewünschte Kriterium

$$\hat{\theta}_n(\lambda) = \frac{\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right|}{||\Delta x_n||} < 1$$

gesichert. Auch die Frage, ob diese Bedingung für $\lambda \to 0$ stets eingehalten werden kann, wird beantwortet. Denn eine kleine Rechnung zeigt, dass $\varphi_n(0) = 1$ und $\varphi_n'(0) = 4\rho - 1$ gilt, daraus ergibt sich $\varphi_n'(0) < 0$ für $\rho < \frac{1}{4}$ und folglich $\tilde{\theta}_n(\lambda) < 1$ für kleine λ. An dieser Stelle wird auch die in Bemerkung 4.6 verwendete Mindestgenauigkeit von $\frac{1}{4}$ klar.

Die vorgeschlagene Genauigkeitssteuerung (4.12) verlangt für kleine Schrittweiten λ eine genauere Lösung der Newton-Gleichung als für große Schrittweiten.

In anderen Globalisierungsansätzen sind Schrittweiten mit $\lambda > 1$ zulässig, zum Beispiel, wenn nach Punkten gesucht wird, die die Armijo- bzw. Wolfe-Bedingung erfüllen. Schritte mit $\lambda > 1$ werden hier nicht betrachtet. Denn zum einen ist das Testen eines λ auf Zulässigkeit (d.h. $\tilde{\theta}_n(\lambda) < 1$) ungefähr so aufwändig wie das Bestimmen einer

neuen Schrittrichtung und zum anderen wird der Fortschritt nur bzgl. der lokalen Norm $||\cdot||_{\mathrm{F}'(x_n)^{-1}}$ gemessen.

4.2 Algorithmische Umsetzung

Bevor ein möglicher Algorithmus angegeben wird, soll untersucht werden, wo es sich lohnen kann, nach zulässigen Punkten zu suchen. Analog zu den Abschätzungen (4.13) ergibt sich die folgende Ungleichungskette:

$$\begin{aligned}\frac{\left|\left|\overline{\delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} &\leq \frac{1}{1-\bar{\delta}_{n+1}} \frac{\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} \\ &\leq \frac{1}{1-\bar{\delta}_{n+1}} \left(1+\delta_n-\lambda+\frac{1}{2}\lambda^2 h_n^\delta\right) \\ &\leq \frac{1}{1-\left[\bar{\delta}_{n+1}\right]} \left(1+[\delta_n]-\lambda+\frac{1}{2}\lambda^2 h_n^\delta\right) \ .\end{aligned}$$

Daraus lässt sich eine Schätzung $\left[h_n^\delta\right]$ für h_n^δ durch

$$\left[h_n^\delta\right] := \frac{2}{\lambda^2}\left(\left(1-\left[\bar{\delta}_{n+1}\right]\right)\frac{\left|\left|\overline{\delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} - 1 - [\delta_n] + \lambda\right) \leq h_n^\delta$$

gewinnen. Ist $\left[h_n^\delta\right] \leq 0$, kann es nicht für die folgenden Überlegungen herangezogen werden.
Weiter gilt:

$$\begin{aligned}\frac{\left|\left|\overline{\Delta x}_{n+1}(\lambda)\right|\right|}{||\Delta x_n||} &\leq \frac{1+\bar{\delta}_{n+1}}{1-\delta_n} \frac{\left|\left|\overline{\delta x}_{n+1}(\lambda)\right|\right|}{||\delta x_n||} \\ &\leq \frac{1+\rho\lambda}{(1-\rho\lambda)^2}\left(1+(\rho-1)\,\lambda+\frac{1}{2}\lambda^2 h_n^\delta\right) =: \bar{\varphi}_n(\lambda)\end{aligned}$$

Die Funktion $\bar{\varphi}_n(\lambda)$ hat die Form:

$$\bar{\varphi}_n(\lambda) = \frac{p_1(\lambda)}{p_2(\lambda)}$$

mit $p_1 \in \mathcal{P}_3$ und $p_2 \in \mathcal{P}_2$. Dabei ist $\mathcal{P}_n$ die Menge aller Polynome p mit $\operatorname{grad} p \leq n$. Für die Ableitung von $\bar{\varphi}_n(\lambda)$ gilt:

$$\bar{\varphi}_n'(\lambda) = \frac{p_3(\lambda)}{p_4(\lambda)}$$

mit $p_3, p_4 \in \mathcal{P}_3$. In diesem Fall lassen sich die Extrempunkte von $\bar{\varphi}_n(\lambda)$ analytisch leicht bestimmen. Wegen ihrer Größe wird hier auf die Angabe einer Formel verzichtet. Insbesondere kann

$$m(h_n^\delta) := \operatorname*{argmin}_{\lambda \in [0,1]} \bar{\varphi}_n(\lambda)$$

berechnet werden. Ersetzt man h_n^δ durch die Schätzung $\left[h_n^\delta\right]$, kann die Beziehung

$$m(h_n^\delta) \leq m(\left[h_n^\delta\right]) \tag{4.14}$$

gezeigt werden. In einem Algorithmus ist das zur Verkürzung der Suche nach einem zulässigen λ nützlich. Gleichung (4.14) sichert, dass „eher zu große λ“ auf Zulässigkeit getestet werden.

Eine mögliche algorithmische Umsetzung ist in Abbildung 4.1 angegeben.

Zunächst wird immer die Schrittweite $\lambda_n = 1$ auf Zulässigkeit überprüft. Wenn dies der Fall ist, wird $x_{n+1} = x_n + \delta x_n$ als neuer Punkt akzeptiert. Anderenfalls erfolgt der Versuch mit kleineren λ_n. Falls bei einer oder zwei aufeinander folgenden Iterationen die volle Schrittweite $\lambda_n = 1$ akzeptiert wird, empfiehlt sich der Übergang in die lokale Phase aus Kapitel 3. Dabei kann die letzte Iterierte vom Algorithmus in Abbildung 4.1 als Startpunkt x_0 des lokalen Algorithmus (Abbildung 3.3) verwendet werden. Dieser Ansatz lässt sich wie folgt begründen. In der Nähe einer Lösung x^* ist $||\delta x_n||$ und damit auch h_n^δ klein. Gilt zusätzlich die Lipschitz-Bedingung in einer Umgebung der Lösung, dann folgt $\tilde{\theta}_n(1) < 1$ (d.h. $\lambda_n = 1$ ist zulässig) aus der Gestalt der Funktion $\bar{\varphi}_n(\lambda)$.

Falls ein Konvergenzmonitor (siehe (3.17) und (3.18)) verwendet wird und dieser fehlschlägt, sollte in den globalen Teil zurückgesprungen werden.

Auf einen expliziten Übergang in die lokale Phase kann allerdings auch verzichtet werden, denn es gilt $\delta_n \leq \rho < \frac{1}{4}$. Damit erfüllt die Genauigkeit die Bedingung für den lokalen linearen Konvergenzmodus aus Abschnitt 3.2.2. Der lokale Algorithmus unterscheidet sich lediglich durch die Berechnung von $\overline{\delta x}_{n+1}$ vom globalen Algorithmus.

Eingabe: Startpunkt x_0 und Funktion F, welche die Lipschitz-Bedingung (4.8) erfüllt.
Ausgabe: Folge x_n, welche bzgl. des natürlichen Abstiegskriteriums einen Fortschritt verspricht.

Schritt 0 Wähle $\rho < \frac{1}{4}$ und den „Linesearch-Parameter“ β mit $0 < \beta < 1$
Setze $n := 0$ und $\lambda_n := 1$

Schritt 1 Bestimme δx_n als Lösung von $\mathrm{F}'(x_n)\Delta x_n = -\mathrm{F}(x_n)$
mit Genauigkeit $\delta_n \leq [\delta_n] \leq \rho\lambda_n$
Bestimme $\overline{\delta x}_{n+1}$ als Lösung von $\mathrm{F}'(x_n)\overline{\Delta x}_{n+1} = -\mathrm{F}(x_n + \lambda_n \delta x_n)$
mit Genauigkeit $\bar{\delta}_{n+1} \leq \left[\bar{\delta}_{n+1}\right] \leq \rho\lambda_n$
Falls $\tilde{\theta}_n := \frac{1+\left[\bar{\delta}_{n+1}\right]}{1-[\delta_n]} \, \frac{\left|\left|\overline{\delta x}_{n+1}\right|\right|}{||\delta x_n||} < 1$, gehe zu Schritt 2
Sonst,
Setze $\left[h_n^\delta\right] := \frac{2}{\lambda_n^2}\left(\left(1 - \left[\bar{\delta}_{n+1}\right]\right)\frac{\left|\left|\overline{\delta x}_{n+1}\right|\right|}{||\delta x_n||} - 1 - [\delta_n] + \lambda_n\right)$
Falls $\left[h_n^\delta\right] > 0$, setze $\lambda_n := \min\left\{\beta\lambda_n, m(\left[h_n^\delta\right])\right\}$ und wiederhole Schritt 1
Sonst, setze $\lambda_n := \beta\lambda_n$ und wiederhole Schritt 1

Schritt 2 Setze $x_{n+1} := x_n + \lambda_n \delta x_n$
Setze $n := n + 1$
Setze $\lambda_n := 1$ und gehe zu Schritt 1

Abbildung 4.1: Globaler Algorithmus

Der im Algorithmus verwendete „Linesearch-Parameter“ β sichert eine Mindestreduktion von λ_n und damit des Suchintervalls. Beispielsweise wird mit $\beta = \frac{1}{2}$ jedes nicht akzeptierte λ_n mindestens halbiert.

Falls nach einer Reduktion von λ_n das alte δx_n die neue Genauigkeitsforderung erfüllt, kann es wieder verwendet werden. Eine analoge Möglichkeit besteht für $\overline{\delta x}_{n+1}$, falls es der neuen Genaugigkeitsanforderung genügt und sich δx_n nicht geändert hat.

Im Zusammenhang mit partiellen Differentialgleichungen bedeutet genaueres Lösen häufig die Verfeinerung des eingesetzten Gitters (mehr dazu folgt im nächsten Kapitel). Es kann sich lohnen, zwischen verschiedenen Iterationen das Gitter zu vergröbern, um keine unnötigen Informationen mitzuführen und das Gitter während der globalen Phase möglichst klein zu halten. Problematisch dabei ist, die Vergröberung so zu steuern, dass nicht zu viel notwendige Information verloren geht.

4.3 Ein naiver Globalisierungsansatz

Ein völlig anderer Ansatz zur Globalisierung besteht darin, das Problem zunächst zu diskretisieren und es anschließend mit einem normalen endlich-dimensionalen Verfahren zu lösen. Die Lösung dient als Startpunkt für die lokale Phase aus Kapitel 3. Dabei gibt es keine Garantie dafür, dass die Diskretisierung gut genug war, um einen zulässigen Startpunkt zu liefern.

Diese Variante findet teilweise im nächsten Kapitel Anwendung. Das „normale“ Newton-Verfahren dient dabei dazu, das diskretisierte Problem zu lösen. Es handelt sich hierbei also nicht um eines der zuvor gemeinten global konvergenten Verfahren, ist jedoch für die betrachtete Problemstellung in Kapitel 5 ausreichend.

5 Beispiel

In diesem Kapitel wird das vorgestellte lokale Newton-Verfahren angewandt. Dazu wird eine kleine Klasse von nichtlinearen elliptischen partiellen Differentialgleichungen betrachtet. Für diese werden kurz Fragen zur Existenz und Eindeutigkeit einer Lösung beantwortet, anschließend ein Beispiel dieser Klasse ausgewählt sowie verschiedene numerische Ergebnisse vorgestellt. Sie resultieren aus der Umsetzung des vorgeschlagenen lokalen Algorithmus (Abbildung 3.3) mit Matlab. Die Grundlage dafür bildet die PDE-Toolbox, von der Quelltext modifiziert und aus lizenzrechtlichen Bedenken auf eine Abgabe der Dateien verzichtet wurde.

5.1 Problemstellung

Gesucht ist die Lösung y der folgenden semilinearen[1] partiellen Differentialgleichung

$$\begin{aligned} -\triangle y + \mathrm{d}(\cdot, y) &= \mathrm{f} \qquad \text{in } \Omega \\ y &= 0 \qquad \text{auf } \Gamma = \partial\Omega \end{aligned} \tag{5.1}$$

mit gegebenem $\mathrm{f} \in L^2(\Omega)$, einer Funktion $\mathrm{d} = \mathrm{d}(x, y) : \Omega \times \mathbb{R} \to \mathbb{R}$ und einem Lipschitz-Gebiet Ω.

Der Übergang zur schwachen Formulierung der Differentialgleichung gelingt durch Multiplikation der Gleichung mit einer Funktion v aus $H_0^1(\Omega)$ und anschließender Integration. Unter Ausnutzung von Lemma 2.22 liefert dies schließlich:

$$(\nabla y, \nabla v)_{L^2(\Omega)} + (\mathrm{d}(\cdot, y), v)_{L^2(\Omega)} = (\mathrm{f}, v)_{L^2(\Omega)} \ .$$

[1] Semilineare Differentialgleichungen sind solche, bei denen alle Ableitungen von höchster Ordnung linear auftreten.

Definition 5.1 Eine Funktion $y \in H_0^1(\Omega)$ heißt schwache Lösung von Aufgabe (5.1), falls

$$(\nabla y, \nabla v)_{L^2(\Omega)} + (\mathrm{d}(\cdot, y), v)_{L^2(\Omega)} = (\mathrm{f}, v)_{L^2(\Omega)} \tag{5.2}$$

für alle $v \in H_0^1(\Omega)$ gilt und alle Integrale existieren.

Die homogene Dirichlet-Randbedingung tritt bereits bei der Wahl des Lösungsraumes $H_0^1(\Omega)$ auf und muss somit nicht extra gefordert werden. Ohne Restriktionen an die Funktion d kann nicht die Existenz einer schwachen Lösung erwartet werden. Daher sollen die folgenden Voraussetzungen gelten.

Voraussetzung 5.2 *Es sei $\Omega \subset \mathbb{R}^N$ ($N \geq 2$) ein Lipschitz-Gebiet. Des Weiteren sei die Funktion $\mathrm{d} = \mathrm{d}(x, y) : \Omega \times \mathbb{R} \to \mathbb{R}$ für jedes feste y messbar bzgl. $x \in \Omega$ (d.h. $x \mapsto \mathrm{d}(x, y)$ ist messbar für alle $y \in \mathbb{R}$) und $\mathrm{d}(\cdot, 0) \in L^2(\Omega)$. Außerdem sei $y \mapsto \mathrm{d}(x, y)$ stetig und monoton wachsend für fast jedes $x \in \Omega$ und erfülle die lokale Lipschitz-Bedingung. Das heißt, für jedes $M > 0$ existiert eine Konstante $L(M)$, so dass*

$$|\mathrm{d}(x, y_1) - \mathrm{d}(x, y_2)| \leq L(M)\,|y_1 - y_2| \qquad \forall y_1, y_2 \in [-M, M]$$

für fast alle $x \in \Omega$ gilt.

Im Folgenden wird stets angenommen, dass die Funktion $\mathrm{d}(x, y)$ diesen Voraussetzungen genügt.

Beispiel 5.3 *Jede monoton wachsende Funktion $\mathrm{d} = \mathrm{d}(y) \in \mathcal{C}^1(\mathbb{R})$ erfüllt offensichtlich die Messbarkeitsvoraussetzung, die Monotonie-Bedingung sowie $\mathrm{d}(\cdot, 0) \in L^2(\Omega)$ und die lokale Lipschitz-Bedingung, denn für ein gegebenes M gilt:*

$$|\mathrm{d}(y_1) - \mathrm{d}(y_2)| = \left| \int_{y_2}^{y_1} \mathrm{d}'(y)\,\mathrm{d}y \right| \leq \underbrace{\max_{y \in [-M,M]} \left|\mathrm{d}'(y)\right|}_{=:L(M)} \cdot |y_1 - y_2| \ .$$

Für das später betrachtete konkrete Beispiel wird $\mathrm{d}(y) = y^3 \in \mathcal{C}^1(\mathbb{R})$ verwendet.

Lemma 5.4 *Es sei $y \in \mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$ eine starke Lösung von Aufgabe (5.1), dann ist y insbesondere eine schwache Lösung.*

Beweis. Multipliziert man Gleichung (5.1) mit einem $v \in H_0^1(\Omega)$ und integriert anschließend unter Ausnutzung von Lemma 2.22, liefert dies die Gleichung (5.2). Dabei ist noch zu zeigen, dass alle auftretenden Integrale existieren. Die Existenz von $(y,v)_{L^2(\Omega)}$ und $(\mathrm{f},v)_{L^2(\Omega)}$ ist offensichtlich. Lediglich für $(\mathrm{d}(\cdot,y),v)_{L^2(\Omega)}$ bedarf es einer Überprüfung. Mit $y \in \mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$ liegt y insbesondere auch in $L^\infty(\Omega)$. Es gibt also ein M mit $|y(x)| \leq M$ (fast überall). Zusammen mit der lokalen Lipschitz-Bedingung ergibt sich:

$$\begin{aligned}
\left|(\mathrm{d}(\cdot,y),v)_{L^2(\Omega)}\right| &\leq \int_\Omega |\mathrm{d}(x,y(x))|\,|v(x)|\;\mathrm{d}x \\
&\leq \int_\Omega |\mathrm{d}(x,y(x)) - \mathrm{d}(x,0)|\,|v(x)|\;\mathrm{d}x + \int_\Omega |\mathrm{d}(x,0)|\,|v(x)|\;\mathrm{d}x \\
&\leq \int_\Omega L(M)\,|y(x)|\,|v(x)|\;\mathrm{d}x + \int_\Omega |\mathrm{d}(x,0)|\,|v(x)|\;\mathrm{d}x\;.
\end{aligned}$$

Die Existenz beider auftretender Integrale resultiert aus $y \in \mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$, $v \in H_0^1(\Omega)$ und $\mathrm{d}(\cdot,0) \in L^2(\Omega)$. Zusammen mit der Messbarkeit folgt die Behauptung. □

Zur Existenz und Eindeutigkeit einer schwachen Lösung sei nun der folgende Satz gegeben, der sich durch Modifikation der Aussagen aus [Trö05, Kapitel 4.1] gewinnen lässt.

Satz 5.5 *Es gelten die Voraussetzungen 5.2. Außerdem sei* $\mathrm{f} \in L^r(\Omega)$ *und* $\mathrm{d}(\cdot,0) \in L^r(\Omega)$ *mit* $r > \frac{N}{2}$. *Dann existiert für die Aufgabe* (5.1) *eine eindeutige schwache Lösung* $y \in H_0^1(\Omega) \cap L^\infty(\Omega)$ *und es gilt die folgende Abschätzung für eine von* f *und* d *unabhängige Konstante c:*

$$||y||_{H^1(\Omega)} + ||y||_{L^\infty(\Omega)} \leq c\left(||\mathrm{f}||_{L^r(\Omega)} + ||\mathrm{d}(\cdot,0)||_{L^r(\Omega)}\right)\;.$$

Für die Existenz und Eindeutigkeit einer Lösung in $H_0^1(\Omega)$ verwendet der Beweis den Hauptsatz über monotone Operatoren (Browder und Minty). Der Beweis, welcher zeigt, dass unter den gegebenen Voraussetzungen sogar eine Lösung in $L^\infty(\Omega)$ existiert, ist jedoch sehr aufwändig und wird hier nicht vorgeführt.

5.2 Vorbetrachtungen für das Newton-Verfahren

Mit dem vorgestellten Newton-Verfahren wird nach einer schwachen Lösung der Aufgabe (5.1) gesucht. Die Funktion F ist durch

$$\mathrm{F} : H_0^1(\Omega) \to H^{-1}(\Omega), y \mapsto (\nabla y, \nabla \cdot)_{L^2(\Omega)} + (\mathrm{d}(\cdot_x, y), \cdot)_{L^2(\Omega)} - (\mathrm{f}, \cdot)_{L^2(\Omega)}$$

gegeben. In der Notation wird zwischen $\cdot_x$ und $\cdot$ unterschieden. Mit $(\mathrm{d}(\cdot_x, y), v)_{L^2(\Omega)}$ ist $\int_\Omega \mathrm{d}(x, y(x))v(x)\,\mathrm{d}x$ gemeint. Die Definition von F erfolgt auf dem ganzen Raum $H_0^1(\Omega)$, denn zusammen mit $1 \leq N \leq 3$ und der später gestellten Wachstumsbedingung (5.4) existieren die auftretenden Integrale. Dies wird an dortiger Stelle genauer diskutiert. Gesucht ist ein $y \in H_0^1(\Omega)$ mit $\mathrm{F}(y) = 0$. Für die Existenz der Fréchet-Ableitung von F kann die Nichtlinearität $(\mathrm{d}(\cdot_x, y), \cdot)_{L^2(\Omega)}$ Probleme bereiten. Für die Abbildung

$$H_0^1(\Omega) \ni y \mapsto (\mathrm{d}(\cdot_x, y), \cdot)_{L^2(\Omega)} \in H^{-1}(\Omega)$$

kommt als Ableitung in Richtung $\delta y \in H_0^1(\Omega)$ nur

$$\begin{aligned}
&\lim_{t\to 0} \frac{(\mathrm{d}(\cdot_x, y + t\delta y), \cdot)_{L^2(\Omega)} - (\mathrm{d}(\cdot_x, y), \cdot)_{L^2(\Omega)}}{t} \\
&= \left(\lim_{t\to 0} \frac{\mathrm{d}(\cdot_x, y + t\delta y) - \mathrm{d}(\cdot_x, y)}{t}, \cdot \right)_{L^2(\Omega)} \\
&= (\mathrm{d}_y(\cdot_x, y)\delta y, \cdot)_{L^2(\Omega)}
\end{aligned}$$

in Frage. Dabei bezeichnet $\mathrm{d}_y(x, y)$ die partielle Ableitung von $\mathrm{d}(x, y)$ nach y. Der lineare Anteil $(\nabla y, \nabla \cdot)_{L^2(\Omega)}$ und der konstante Anteil $(\mathrm{f}, \cdot)_{L^2(\Omega)}$ in F bereiten keine Schwierigkeiten beim Ableiten. Es ist nachzurechnen, unter welchen Bedingungen es sich bei

$$\mathrm{F}'(y)\delta y = (\nabla \delta y, \nabla \cdot)_{L^2(\Omega)} + (\mathrm{d}_y(\cdot_x, y)\delta y, \cdot)_{L^2(\Omega)} \in H^{-1}(\Omega)$$

tatsächlich um die Fréchet-Ableitung handelt. Es gilt:

$$\begin{aligned}
\mathrm{F}(y + \delta y) - \mathrm{F}(y) - \mathrm{F}'(y)\delta y = {} & (\nabla (y + \delta y), \nabla \cdot)_{L^2(\Omega)} + (\mathrm{d}(\cdot_x, y + \delta y), \cdot)_{L^2(\Omega)} - (\mathrm{f}, \cdot)_{L^2(\Omega)} \\
& - \Big((\nabla y, \nabla \cdot)_{L^2(\Omega)} + (\mathrm{d}(\cdot_x, y), \cdot)_{L^2(\Omega)} - (\mathrm{f}, \cdot)_{L^2(\Omega)} \Big) \\
& - \Big((\nabla \delta y, \nabla \cdot)_{L^2(\Omega)} + (\mathrm{d}_y(\cdot_x, y)\delta y, \cdot)_{L^2(\Omega)} \Big)
\end{aligned}$$

$$= (\mathrm{d}(\cdot_x, y+\delta y) - \mathrm{d}(\cdot_x, y) - \mathrm{d}_y(\cdot_x, y)\delta y, \cdot)_{L^2(\Omega)} \ .$$

Zu zeigen ist:

$$\lim_{\delta y \to 0} \frac{\left|\left|(\mathrm{d}(\cdot_x, y+\delta y) - \mathrm{d}(\cdot_x, y) - \mathrm{d}_y(\cdot_x, y)\delta y, \cdot)_{L^2(\Omega)}\right|\right|_{H^{-1}(\Omega)}}{||\delta y||_{H^1(\Omega)}} = 0 \ .$$

Mit $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ und Lemma 2.24 gilt:

$$\begin{aligned}
&\lim_{\delta y \to 0} \frac{\left|\left|(\mathrm{d}(\cdot_x, y+\delta y) - \mathrm{d}(\cdot_x, y) - \mathrm{d}_y(\cdot_x, y)\delta y, \cdot)_{L^2(\Omega)}\right|\right|_{H^{-1}(\Omega)}}{||\delta y||_{H^1(\Omega)}} \\
&\leq \lim_{\delta y \to 0} \frac{\left|\left|(\mathrm{d}(\cdot_x, y+\delta y) - \mathrm{d}(\cdot_x, y) - \mathrm{d}_y(\cdot_x, y)\delta y, \cdot)_{L^2(\Omega)}\right|\right|_{(L^2(\Omega))^*}}{||\delta y||_{H^1(\Omega)}} \\
&= \lim_{\delta y \to 0} \frac{||\mathrm{d}(\cdot_x, y+\delta y) - \mathrm{d}(\cdot_x, y) - \mathrm{d}_y(\cdot_x, y)\delta y||_{L^2(\Omega)}}{||\delta y||_{H^1(\Omega)}} \ . \qquad (5.3)
\end{aligned}$$

Die Gleichung (5.3) sagt aus, dass $\mathrm{F} : H_0^1(\Omega) \to H^{-1}(\Omega)$ Fréchet-differenzierbar ist, falls der Nemyzki-Operator

$$y \mapsto \mathrm{d}(\cdot_x, y)$$

von $H_0^1(\Omega)$ nach $L^2(\Omega)$ Fréchet-differenzierbar ist. Nach Beispiel 2.27 bettet $H_0^1(\Omega)$ für $\Omega \subset \mathbb{R}^N$ mit $1 \leq N \leq 3$ stetig in $L^6(\Omega)$ ein. Zusammen mit Lemma 2.4 und Beispiel 2.34 folgt damit für F mit $\mathrm{d}(x, y) = y^3$ die Fréchet-Differenzierbarkeit. Dies gilt auch für andere Funktionen $\mathrm{d}(x, y)$, für welche der zugehörige Nemyzki-Operator von $L^6(\Omega)$ nach $L^2(\Omega)$ Fréchet-differenzierbar ist. Nach Satz 2.32 sind das Funktionen $\mathrm{d}(x, y)$, die neben den „üblichen“ Voraussetzungen die Wachstumsbedingung (2.2) für $\mathrm{d}(x, y)$ und $\mathrm{d}_y(x, y)$ erfüllen, d.h.

$$|\mathrm{d}(x, y)| \leq \alpha(x) + \beta(x)\,|y|^3 \qquad (5.4)$$

und

$$|\mathrm{d}_y(x, y)| \leq \tilde{\alpha}(x) + \tilde{\beta}(x)\,|y|^2 \qquad (5.5)$$

mit $\alpha \in L^2(\Omega)$, $\tilde{\alpha} \in L^3(\Omega)$ sowie $\beta, \tilde{\beta} \in L^\infty(\Omega)$. Beschränkt man sich auf den Fall $N = 2$, ergeben sich mit ähnlicher Diskussion die schwächeren Wachstumsbedingungen

$$|\mathrm{d}(x, y)| \leq \alpha(x) + \beta(x)\,|y|^p$$

und

$$|\mathrm{d}_y(x,y)| \leq \tilde{\alpha}(x) + \tilde{\beta}(x)\,|y|^p$$

mit $\alpha \in L^2(\Omega)$, $\tilde{\alpha} \in L^{2+\varepsilon}(\Omega)$, $\beta, \tilde{\beta} \in L^\infty(\Omega)$, $p < \infty$ sowie $\varepsilon > 0$.
In analoger Weise zeigt sich, dass F mit $\mathrm{d}(x,y) = y^3$ sogar stetig Fréchet-differenzierbar ist.

Die Newton-Richtung[1] δy_n im Punkt y_n ist als Lösung von

$$\mathrm{F}'(y_n)\delta y_n = -\,\mathrm{F}(y_n) \tag{5.6}$$

$$\begin{aligned}\iff \quad (\nabla \delta y_n, \nabla\cdot)_{L^2(\Omega)} + (\mathrm{d}_y(\cdot_x, y_n)\delta y_n, \cdot)_{L^2(\Omega)} = -\Big((\nabla y_n, \nabla\cdot)_{L^2(\Omega)} \\ + (\mathrm{d}(\cdot_x, y_n), \cdot)_{L^2(\Omega)} - (\mathrm{f}, \cdot)_{L^2(\Omega)}\Big)\end{aligned} \tag{5.7}$$

gegeben. Um die Existenz und Eindeutigkeit eines solchen δy_n bzw. die Invertierbarkeit von $\mathrm{F}'(y_n)$ zu zeigen, erweist sich die direkte Bestimmung des neuen Punktes $y_{n+1} = y_n + \delta y_n$ als günstig. Die Richtung δy_n ist dann einfach als $\delta y_n = y_{n+1} - y_n$ gegeben. Dazu werden die Gleichungen (5.6) und (5.7) entsprechend umgeformt:

$$\mathrm{F}'(y_n)y_{n+1} = -\,\mathrm{F}(y_n) + \mathrm{F}'(y_n)y_n \tag{5.8}$$

$$\iff \quad (\nabla y_{n+1}, \nabla\cdot)_{L^2(\Omega)} + (\mathrm{d}_y(\cdot_x, y_n)y_{n+1}, \cdot)_{L^2(\Omega)} = (\mathrm{d}_y(\cdot_x, y_n)y_n - \mathrm{d}(\cdot_x, y_n) + \mathrm{f}, \cdot)_{L^2(\Omega)}\ . \tag{5.9}$$

Dies ist die schwache Formulierung von:

$$\begin{aligned} -\triangle y_{n+1} + \mathrm{d}_y(\cdot_x, y_n)y_{n+1} &= \mathrm{d}_y(\cdot_x, y_n)y_n - \mathrm{d}(\cdot_x, y_n) + \mathrm{f} && \text{in } \Omega \\ y_{n+1} &= 0 && \text{auf } \Gamma = \partial\Omega\ . \end{aligned}$$

In jedem Schritt muss also eine lineare elliptische partielle Differentialgleichung gelöst werden. Sie lässt sich in der Form

$$\begin{aligned} -\triangle y_{n+1} + \mathrm{c}_n\, y_{n+1} &= \mathrm{g}_n && \text{in } \Omega \\ y_{n+1} &= 0 && \text{auf } \Gamma \end{aligned} \tag{5.10}$$

[1] In diesem Abschnitt wird δy anstelle von Δy (vgl. Kapitel 3) verwendet, um eine Verwechslung mit dem Laplace-Operator zu vermeiden. Die inexakte Richtung wird mit $\bar{\delta} y$ bezeichnet.

mit $\mathrm{c}_n(x) = \mathrm{d}_y(x, y_n(x))$ und $\mathrm{g}_n(x) = \mathrm{d}_y(x, y_n(x))y_n(x) - \mathrm{d}(x, y_n(x)) + \mathrm{f}(x)$ schreiben. Die Monotonie von $\mathrm{d}(x, y)$ bzgl. y (siehe Voraussetzung 5.2) sichert, dass $\mathrm{c} \geq 0$ fast überall gilt. Dadurch ist die dazugehörige Bilinearform

$$\mathrm{a}_n : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, (w, v) \mapsto (\nabla w, \nabla v)_{L^2(\Omega)} + (\mathrm{c}_n\, w, v)_{L^2(\Omega)}$$

koerziv, denn mit der Äquivalenz von $||\cdot||_{H_0^1(\Omega)}$ und $||\cdot||_{H^1(\Omega)}$ auf $H_0^1(\Omega)$ (siehe Ungleichung 2.1) gilt:

$$\mathrm{a}_n[v, v] \geq (\nabla v, \nabla v)_{L^2(\Omega)} = ||v||^2_{H_0^1(\Omega)} \geq \frac{1}{1 + c_\Omega^2} ||v||^2_{H^1(\Omega)} \ .$$

Des Weiteren ist a_n beschränkt, denn mit der Hölder-Ungleichung ergibt sich für $w, v \in H_0^1(\Omega)$:

$$\begin{aligned} |\mathrm{a}_n[w, v]| &\leq \int_\Omega |\nabla w \cdot \nabla v| \ \mathrm{d}x + \int_\Omega |\mathrm{c}_n\, wv| \ \mathrm{d}x \\ &\leq ||w||_{H^1(\Omega)} ||v||_{H^1(\Omega)} + ||\mathrm{c}_n||_{L^3(\Omega)} ||w||_{L^3(\Omega)} ||v||_{L^3(\Omega)} \\ &\leq \left(1 + c\, ||\mathrm{c}_n||_{L^3(\Omega)}\right) ||w||_{H^1(\Omega)} ||v||_{H^1(\Omega)} \ . \end{aligned}$$

Für $y_n \in H_0^1(\Omega)$ und $1 \leq N \leq 3$ liegt y_n auch in $L^6(\Omega)$ (siehe Beispiel 2.27). Die Wachstumsbedingung (5.5) sichert somit, dass $\mathrm{c}_n = \mathrm{d}_y(\cdot, y_n(\cdot))$ in $L^3(\Omega)$ liegt.

Das zur rechten Seite von (5.10) gehörende lineare Funktional

$$\mathrm{G}_n : H_0^1(\Omega) \to \mathbb{R}, w \mapsto (g_n, w)_{L^2(\Omega)}$$

ist stetig, denn für $w \in H_0^1(\Omega)$ gilt:

$$\begin{aligned} |\mathrm{G}_n(w)| &\leq \int_\Omega |\mathrm{d}_y(x, y_n(x))y_n(x)w(x)| \ \mathrm{d}x + \int_\Omega |\mathrm{d}(x, y_n(x))w(x)| \ \mathrm{d}x + \int_\Omega |\mathrm{f}(x)w(x)| \ \mathrm{d}x \\ &\leq \left(||\mathrm{d}_y(\cdot, y_n(\cdot))y_n(\cdot)||_{L^2(\Omega)} + ||\mathrm{d}(\cdot, y_n(\cdot))||_{L^2(\Omega)} + ||\mathrm{f}||_{L^2(\Omega)}\right) ||w||_{L^2(\Omega)} \\ &\leq \left(||\mathrm{d}_y(\cdot, y_n(\cdot))y_n(\cdot)||_{L^2(\Omega)} + ||\mathrm{d}(\cdot, y_n(\cdot))||_{L^2(\Omega)} + ||\mathrm{f}||_{L^2(\Omega)}\right) ||w||_{H^1(\Omega)} \end{aligned}$$

und damit

$$||\mathrm{G}_n||_{H^{-1}(\Omega)} \leq ||\mathrm{d}_y(\cdot, y_n(\cdot))y_n(\cdot)||_{L^2(\Omega)} + ||\mathrm{d}(\cdot, y_n(\cdot))||_{L^2(\Omega)} + ||\mathrm{f}||_{L^2(\Omega)} \ .$$

Wie bereits erwähnt, liegt $\mathrm{d}_y(\cdot, y_n(\cdot))$ in $L^3(\Omega)$. Damit folgt aus $y_n \in H_0^1(\Omega) \hookrightarrow L^6(\Omega)$ und der Hölder-Ungleichung (Lemma 2.25), dass $\mathrm{d}_y(\cdot, y_n(\cdot))y_n(\cdot)$ in $L^2(\Omega)$ liegt. Analog ergibt sich mit (5.4), dass $\mathrm{d}(\cdot, y_n(\cdot))$ zum Raum $L^2(\Omega)$ gehört. Außerdem gilt $\mathrm{f} \in L^2(\Omega)$ nach Voraussetzung.

Das Lemma von Lax-Milgram sichert nun die Existenz eines eindeutigen y_{n+1}, welches die Gleichung (5.9) bzw. die schwache Form der linearen partiellen Differentialgleichung (5.10) erfüllt. Mit der Existenz eines neuen Punktes y_{n+1} ist offensichtlich auch die Existenz eines $\delta y_n = y_{n+1} - y_n$ garantiert.

Lemma 5.6 (**Lax-Milgram, [Gri07]**) *Es sei V ein Hilbertraum über $\mathbb{R}$ und $\mathrm{a} : V \times V \to \mathbb{R}$ eine Bilinearform mit den folgenden beiden Eigenschaften:*

(i) a *ist beschränkt, d.h. es gibt eine Konstante M mit*

$$|\mathrm{a}[v, w]| \leq M\, ||v||_V\, ||w||_V \qquad \forall v, w \in V \ ,$$

(ii) a *ist koerziv, d.h. es gibt eine Konstante m mit*

$$|\mathrm{a}[v, v]| \geq m\, ||v||_V^2 \qquad \forall v \in V \ .$$

Dann existiert zu jedem $\mathrm{G} \in V^*$ genau ein $y \in V$ mit

$$\mathrm{a}[y, v] = \mathrm{G}(v) \qquad \forall v \in V$$

und es gilt die Abschätzung

$$||y||_V \leq c\, ||\mathrm{G}||_{V^*} \tag{5.11}$$

für eine Konstante c. Des Weiteren ist die Abbildung

$$V^* \ni \mathrm{G} \mapsto y \in V$$

linear und wegen Ungleichung (5.11) *auch stetig.*

Dieses Lemma verallgemeinert die Aussage des Riezschen Darstellungssatzes auf beschränkte, koerzive Bilinearformen. Eine analoge Aussage gilt für den Körper der komplexen Zahlen.

Aus dem Lemma von Lax-Milgram folgt sofort, dass die Gleichung (5.9) bzw. (5.8) eine eindeutige Lösung hat. Zudem ist der Operator $\mathrm{F}'(y_n) \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ invertierbar sowie seine Inverse linear und beschränkt.

5.3 Fehlerschätzer

Um das Newton-Verfahren aus Kapitel 3 tatsächlich durchzuführen, ist eine Schätzung des Fehlers $\left|\left|\delta y_n - \tilde{\delta} y_n\right|\right|_{H^1(\Omega)}$ notwendig, der durch die Diskretisierung von Gleichung (5.8) auftritt. Die relative Genauigkeit δ_n ist definiert als

$$\delta_n = \frac{\left|\left|\delta y_n - \tilde{\delta} y_n\right|\right|_{H^1(\Omega)}}{\left|\left|\tilde{\delta} y_n\right|\right|_{H^1(\Omega)}} ,$$

wobei δy_n die exakte und $\tilde{\delta} y_n$ die inexakte Lösung von (5.6) bezeichnet. Hier lässt sich die relative Genauigkeit δ_n auch in der Form

$$\delta_n = \frac{||y_{n+1} - \tilde{y}_{n+1}||_{H^1(\Omega)}}{||\tilde{y}_{n+1} - \tilde{y}_n||_{H^1(\Omega)}}$$

ausdrücken, wobei mit y_{n+1} wieder die exakte und mit $\tilde{y}_{n+1}$ die inexakte Lösung von (5.8) gemeint ist. Der Nenner des Bruches lässt sich leicht ausrechnen, da nur bekannte Größen vorkommen. Der Zähler ist jedoch nicht berechenbar, da die exakte Lösung y_{n+1} unbekannt ist. Beim Schätzen des Fehlers ist es nicht üblich, die $H^1(\Omega)$-Norm zu verwenden. Daher soll jetzt zur kleineren $H_0^1(\Omega)$-Norm übergegangen werden. Es wird daran erinnert, dass auf dem hier betrachteten Hilbertraum $H_0^1(\Omega)$ die $H^1(\Omega)$-Norm und die $H_0^1(\Omega)$-Norm äquivalent sind (siehe (2.1)). Aus diesem Grund gelten alle bisher gemachten Überlegungen auch für die $H_0^1(\Omega)$-Norm. Dementsprechend würde der Fehlerschätzer auch für die $H^1(\Omega)$-Norm funktionieren.

5.3.1 Konstruktion des Fehlerschätzers

Der im Folgenden konstruierte a-posteriori Residuum-basierte Schätzer bestimmt eine obere Schätzung für $||y - y_h||_{H_0^1(\Omega)}$. Zur Vereinfachung wird der Index n weggelassen. Die Darstellung richtet sich nach [Ver96]. Hierbei ist y die exakte schwache Lösung von

$$\begin{aligned} -\triangle y + \mathrm{c}\, y &= \mathrm{g} && \text{in } \Omega \\ y &= 0 && \text{auf } \Gamma \end{aligned} \tag{5.12}$$

und y_h die diskrete inexakte Lösung, welche mit Hilfe der Finiten-Elemente-Methode[1] (kurz: FEM) für die Triangulierung $\mathcal{T}_h$ des Gebietes Ω berechnet wurde. Die Funktionen c und g seien aus dem Raum $L^2(\Omega)$ gegeben. Außerdem gelte fast überall $c \geq 0$. Die zur schwachen Formulierung gehörende Bilinearform ist durch

$$\mathrm{a} : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, (w, v) \mapsto (\nabla w, \nabla v)_{L^2(\Omega)} + (\mathrm{c}\, w, v)_{L^2(\Omega)}$$

gegeben. Offensichtlich gilt

$$\mathrm{a}[v, v] \geq ||v||^2_{H_0^1(\Omega)} \ ,$$

d.h. a ist bzgl. $||\cdot||_{H_0^1(\Omega)}$ koerziv mit Konstante 1.
Daraus folgt:

$$\begin{aligned} ||y - y_h||_{H_0^1(\Omega)} &\leq \frac{1}{||y - y_h||_{H_0^1(\Omega)}} \mathrm{a}[y - y_h, y - y_h] \\ &\leq \sup_{\substack{v \in H_0^1(\Omega) \\ ||v||_{H_0^1(\Omega)} = 1}} \mathrm{a}[y - y_h, v] \\ &= \sup_{\substack{v \in H_0^1(\Omega) \\ ||v||_{H_0^1(\Omega)} = 1}} (\nabla (y - y_h), \nabla v)_{L^2(\Omega)} + (\mathrm{c}\, (y - y_h), v)_{L^2(\Omega)} \\ &= \sup_{\substack{v \in H_0^1(\Omega) \\ ||v||_{H_0^1(\Omega)} = 1}} (\mathrm{g}, v)_{L^2(\Omega)} - (\mathrm{c}\, y_h, v)_{L^2(\Omega)} - (\nabla y_h, \nabla v)_{L^2(\Omega)} \ . \end{aligned} \tag{5.13}$$

[1] In dieser Arbeit wird sich auf die konforme FEM mit linearen Ansatz- und Testfunktionen beschränkt.

Die Beziehung (5.13) besagt, dass die Norm des Fehlers $||y - y_h||_{H_0^1(\Omega)}$ kleiner ist als die Norm des Residuums $(\mathrm{g}, \cdot)_{L^2(\Omega)} - (\mathrm{c}\, y_h, \cdot)_{L^2(\Omega)} - (\nabla y_h, \nabla \cdot)_{L^2(\Omega)}$.

Bevor das Residuum weiter abgeschätzt wird, ist es zunächst notwendig, weitere Notationen einzuführen. Ab sofort erfolgt die Beschränkung auf den Fall $N = 2$. Ähnliche Überlegungen gelten auch für $N = 3$, wobei dann Dreiecke durch Tetraeder, Kanten durch Flächen usw. zu ersetzen sind. Der endlich-dimensionale Ansatzraum entspricht dem Raum der Testfunktionen und wird mit X_h bezeichnet. Eine Basis dieses Raumes sollen die linearen, stetigen Funktionen v_i mit $v_i(x_k) = \delta_{ik}$ für die Knoten x_k bilden. Für ein Dreieck T und eine Kante E der Triangulierung $\mathcal{T}_h$ meint $\mathcal{E}(T)$ die Kanten von T, $\mathcal{N}(T)$ die Knoten von T und $\mathcal{N}(E)$ die Knoten von E. Man definiert die Kantenmenge durch

$$\mathcal{E}_h := \bigcup_{T \in \mathcal{T}_h} \mathcal{E}(T)$$

und unterteilt diese in die Menge der inneren Kanten $\mathcal{E}_{h,\Omega}$ sowie die der Randkanten $\mathcal{E}_{h,\Gamma}$ mit

$$\mathcal{E}_{h,\Gamma} := \{E \in \mathcal{E}_h : E \subset \Gamma\}$$

und

$$\mathcal{E}_{h,\Omega} := \mathcal{E}_h \setminus \mathcal{E}_{h,\Gamma}\ .$$

Hierbei ist die Inklusion $E \subset \Gamma$ „nicht streng“ zu interpretieren, denn es soll $E \subset \Gamma$ auch für Kanten E gelten, die nach der Triangulierung des Gebietes Ω zu Kanten des Randes geworden sind.

Des Weiteren definiert man zu einem Dreieck $T \in \mathcal{T}_h$ und einer Kante $E \in \mathcal{E}_h$ die Gebiete ω_T und ω_E mit

$$\omega_T := \bigcup_{\substack{T' \in \mathcal{T}_h \text{ mit} \\ \mathcal{N}(T) \cap \mathcal{N}(T') \neq \emptyset}} T' \qquad \text{und} \qquad \omega_E := \bigcup_{\substack{T' \in \mathcal{T}_h \text{ mit} \\ \mathcal{N}(E) \cap \mathcal{N}(T') \neq \emptyset}} T'\ .$$

Sie sind in Abbildung 5.1 dargestellt.

Zu jeder inneren Kante $E \in \mathcal{E}_{h,\Omega}$ sei ein normierter Vektor n_E definiert, welcher

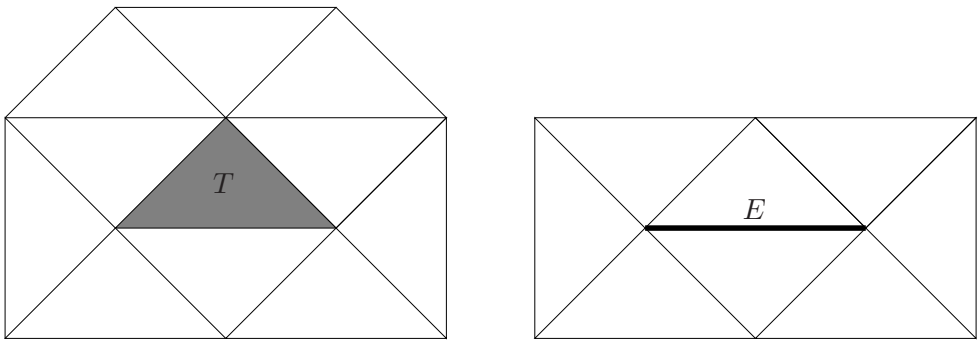


Abbildung 5.1: Gebiete ω_T und ω_E

orthogonal auf E steht (die Richtung ist dabei später nicht von Bedeutung). Für eine innere Kante $E \in \mathcal{E}_{h,\Omega}$, welche von den beiden Dreiecken T_1 und T_2 begrenzt wird, sowie für eine Funktion $v \in L^2(T_1 \cup T_2)$ mit $v\big|_{T_1} \in \mathcal{C}(T_1)$ und $v\big|_{T_2} \in \mathcal{C}(T_2)$, bezeichnet $[v]_E$ den Sprung von v entlang E in Richtung n_E, d.h.

$$[v]_E(x) := \lim_{t \downarrow 0} v(x + tn_E) - \lim_{t \downarrow 0} v(x - tn_E) \ .$$

Schließlich wird noch der Interpolationsoperator $I_h : H_0^1(\Omega) \to X_h$ von Clément benötigt ([Cle75]). Dieser kann eindeutig definiert werden und hat „schöne“ Eigenschaften, wie das folgende Lemma verdeutlicht. Dabei wird mit h_T die längste Kante eines Dreiecks $T \in \mathcal{T}_h$ und mit h_E die Länge der Kante $E \in \mathcal{E}_h$ bezeichnet.

Lemma 5.7 *Für ein beliebiges $v \in H_0^1(\Omega)$ gelten die beiden Abschätzungen:*

$$||v - I_h v||_{L^2(T)} \leq c_1 h_T \, ||v||_{H^1(\omega_T)} \qquad \forall T \in \mathcal{T}_h \tag{5.14}$$

und

$$||v - I_h v||_{L^2(E)} \leq c_2 h_E^{\frac{1}{2}} \, ||v||_{H^1(\omega_E)} \qquad \forall E \in \mathcal{E}_h \ . \tag{5.15}$$

Die Konstanten c_1 und c_2 hängen dabei nur vom kleinsten Winkel der Triangulierung $\mathcal{T}_h$ ab.

Die linke Seite der Beziehung (5.15) ist bzgl. des Spur-Operators auf E (vgl. Satz 2.19) zu verstehen.

Nun kann das Residuum umgeformt und die Ungleichung (5.13) weiter abgeschätzt werden. Zusammen mit Lemma 2.22 ergibt sich für $v \in H_0^1(\Omega)$:

$$\begin{aligned}
&(\mathrm{g} - \mathrm{c}\, y_h, v)_{L^2(\Omega)} - (\nabla y_h, \nabla v)_{L^2(\Omega)} \\
&= \int_\Omega (\mathrm{g} - \mathrm{c}\, y_h)\, v \,\mathrm{d}x - \int_\Omega \nabla y_h \cdot \nabla v \,\mathrm{d}x \\
&= \int_\Omega (\mathrm{g} - \mathrm{c}\, y_h)\, v \,\mathrm{d}x - \sum_{T \in \mathcal{T}_h} \int_T \nabla y_h \cdot \nabla v \,\mathrm{d}x \\
&= \int_\Omega (\mathrm{g} - \mathrm{c}\, y_h)\, v \,\mathrm{d}x + \sum_{T \in \mathcal{T}_h} \left(\int_T \underbrace{\triangle y_h}_{=0} v \,\mathrm{d}x - \int_{\partial T} n_T \cdot \nabla y_h v \,\mathrm{d}s \right) \\
&= \sum_{T \in \mathcal{T}_h} \int_T (\mathrm{g} - \mathrm{c}\, y_h)\, v \,\mathrm{d}x - \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [n_E \cdot \nabla y_h]_E \, v \,\mathrm{d}s \,. \qquad (5.16)
\end{aligned}$$

Hierbei ist mit n_T die äußere Normale von dem Dreieck T gemeint. Die diskrete Lösung y_h erfüllt die sogenannte „Galerkin-Orthogonalität“

$$(\mathrm{g} - \mathrm{c}\, y_h, v)_{L^2(\Omega)} - (\nabla y_h, \nabla v)_{L^2(\Omega)} = 0 \qquad \forall v \in X_h \,.$$

Damit folgt für $v \in H_0^1(\Omega)$ zusammen mit der Gleichung (5.16), dem Lemma 5.7, der Cauchy-Schwarzschen-Ungleichung und der Beziehung (2.1):

$$\begin{aligned}
&(\mathrm{g} - \mathrm{c}\, y_h, v)_{L^2(\Omega)} - (\nabla y_h, \nabla v)_{L^2(\Omega)} \\
&= (\mathrm{g} - \mathrm{c}\, y_h, v - I_h v)_{L^2(\Omega)} - (\nabla y_h, \nabla (v - I_h v))_{L^2(\Omega)} \\
&= \sum_{T \in \mathcal{T}_h} \int_T (\mathrm{g} - \mathrm{c}\, y_h)\, (v - I_h v) \,\mathrm{d}x - \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [n_E \cdot \nabla y_h]_E \, (v - I_h v) \,\mathrm{d}s \\
&\leq \sum_{T \in \mathcal{T}_h} c_1 h_T \, ||\mathrm{g} - \mathrm{c}\, y_h||_{L^2(T)} \, ||v||_{H^1(\omega_T)} + \sum_{E \in \mathcal{E}_{h,\Omega}} c_2 h_E^{\frac{1}{2}} \, ||[n_E \cdot \nabla y_h]_E||_{L^2(E)} \, ||v||_{H^1(\omega_E)} \\
&\leq \max\{c_1, c_2\} \left(\sum_{T \in \mathcal{T}_h} h_T^2 \, ||\mathrm{g} - \mathrm{c}\, y_h||_{L^2(T)}^2 + \sum_{E \in \mathcal{E}_{h,\Omega}} h_E \, ||[n_E \cdot \nabla y_h]_E||_{L^2(E)}^2 \right)^{\frac{1}{2}} \\
&\qquad \cdot \left(\sum_{T \in \mathcal{T}_h} ||v||_{H^1(\omega_T)}^2 + \sum_{E \in \mathcal{E}_{h,\Omega}} ||v||_{H^1(\omega_E)}^2 \right)^{\frac{1}{2}} \qquad (5.17)
\end{aligned}$$

$$\leq c_{\mathcal{T}_h} \max\{c_1, c_2\} \left(\sum_{T \in \mathcal{T}_h} h_T^2 \, ||\mathrm{g} - \mathrm{c}\, y_h||^2_{L^2(T)} + \sum_{E \in \mathcal{E}_{h,\Omega}} h_E \, ||[n_E \cdot \nabla y_h]_E||^2_{L^2(E)} \right)^{\frac{1}{2}} ||v||_{H^1(\Omega)}$$

$$\leq c_{\mathcal{T}_h} \max\{c_1, c_2\} \sqrt{1 + c_\Omega^2}$$

$$\cdot \left(\sum_{T \in \mathcal{T}_h} h_T^2 \, ||\mathrm{g} - \mathrm{c}\, y_h||^2_{L^2(T)} + \sum_{E \in \mathcal{E}_{h,\Omega}} h_E \, ||[n_E \cdot \nabla y_h]_E||^2_{L^2(E)} \right)^{\frac{1}{2}} ||v||_{H_0^1(\Omega)} \ .$$

Die Konstante $c_{\mathcal{T}_h}$ ist ein Maß dafür, wie oft maximal über ein Dreieck integriert wird. Sie hängt nur vom kleinsten Winkel der Triangulierung ab und kann für ein gegebenes Gitter leicht bestimmt werden. Näheres wird im nächsten Abschnitt erläutert.

Mit der Ungleichung (5.13) ergibt sich die wichtige Beziehung

$$||y - y_h||_{H_0^1(\Omega)} \leq \underbrace{c_{\mathcal{T}_h} \max\{c_1, c_2\} \sqrt{1 + c_\Omega^2}}_{=:\bar{c}}$$

$$\cdot \underbrace{\left(\sum_{T \in \mathcal{T}_h} h_T^2 \, ||\mathrm{g} - \mathrm{c}\, y_h||^2_{L^2(T)} + \sum_{E \in \mathcal{E}_{h,\Omega}} h_E \, ||[n_E \cdot \nabla y_h]_E||^2_{L^2(E)} \right)^{\frac{1}{2}}}_{=:\eta_h} \ . \tag{5.18}$$

Hierbei steht auf der linken Seite der Abstand der diskreten Lösung y_h zur „unbekannten“ exakten Lösung y. Im Gegensatz dazu befinden sich auf der rechten Seite nur berechenbare Größen (bis auf die Konstante $\bar{c}$).

Für ein Dreieck $T \in \mathcal{T}_h$ definiert man den lokalen Fehlerindikator $\eta_{h,T}$ durch

$$\eta_{h,T} := \left(h_T^2 \, ||\mathrm{g} - \mathrm{c}\, y_h||^2_{L^2(T)} + \frac{1}{2} \sum_{E \in \mathcal{E}(T) \cap \mathcal{E}_{h,\Omega}} h_E \, ||[n_E \cdot \nabla y_h]_E||^2_{L^2(E)} \right)^{\frac{1}{2}} \ .$$

Damit wird der Fehler entlang einer Kante auf die beiden angrenzenden Dreiecke aufgeteilt. Es gilt die Beziehung:

$$\eta_h = \left(\sum_{T \in \mathcal{T}_h} \eta_{h,T}^2 \right)^{\frac{1}{2}} \ .$$

Insgesamt ergibt sich:

$$||y - y_h||_{H_0^1(\Omega)} \leq \bar{c}\,\eta_h \leq \bar{c} \sum_{T \in \mathcal{T}_h} \eta_{h,T} \ .$$

Die lokalen Fehlerindikatoren $\eta_{h,T}$ werden eingesetzt, um die Dreiecke T mit großem Fehler (d.h. mit großem $\eta_{h,T}$) zu erkennen und für eine mögliche Verfeinerung zu markieren.

Bemerkung 5.8 *Für $h := \max\limits_{T \in \mathcal{T}_h} h_T \to 0$ kann gezeigt werden, dass auch $\bar{c}\,\eta_h$ gegen 0 strebt, sofern der kleinste Winkel der Triangulierungen $\mathcal{T}_h$ von unten beschränkt bleibt.*

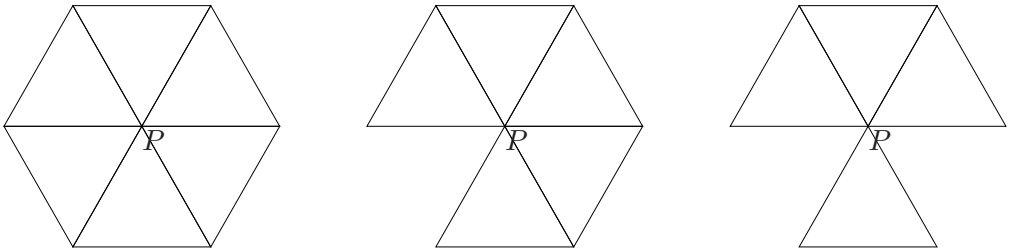
5.3.2 Diskussion der Konstanten

Im Schätzer (5.18) für $||y - y_h||_{H_0^1(\Omega)}$ tauchen die vier Konstanten c_Ω, $c_{\mathcal{T}_h}$, c_1 und c_2 auf. Bis auf c_Ω hängen sie von der jeweiligen Triangulierung ab, insbesondere vom kleinsten Winkel der Triangulierung. Falls der kleinste Winkel über alle Triangulierungen von unten beschränkt bleibt, so sind die Konstanten $c_{\mathcal{T}_h}$, c_1 und c_2 von oben beschränkt. Wie dies realisiert werden kann, ist in Abschnitt 5.4 beschrieben.

Die Konstante c_Ω bereitet hier keine Probleme, denn für Ω wird der Einheitskreis, d.h. $\Omega = \left\{x = (x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 < 1\right\}$ verwendet. Aus den Bemerkungen zur Friedrichschen Ungleichung (Lemma 2.18) ergibt sich $c_\Omega \approx \frac{1}{2{,}4048}$.

Wie oben bereits angedeutet, ist die Konstante $c_{\mathcal{T}_h}$ zu einer gegebenen Triangulierung leicht bestimmbar. Dazu wird nochmals die Summe in (5.17) betrachtet und daran erinnert, dass mit den Gebieten ω_T bzw. ω_E alle an $T \in \mathcal{T}_h$ bzw. $E \in \mathcal{E}_h$ angrenzenden Dreiecke gemeint sind. Es gilt:

$$\begin{aligned}
&\left(\sum_{T \in \mathcal{T}_h} ||v||^2_{H^1(\omega_T)} + \sum_{E \in \mathcal{E}_{h,\Omega}} ||v||^2_{H^1(\omega_E)} \right)^{\frac{1}{2}} \\
&= \left(\sum_{T \in \mathcal{T}_h} c_{a,T}\, ||v||^2_{H^1(T)} + \sum_{T \in \mathcal{T}_h} c_{b,T}\, ||v||^2_{H^1(T)} \right)^{\frac{1}{2}} \\
&\leq \underbrace{\left(2 \max_{T \in \mathcal{T}_h} \{c_{a,T}, c_{b,T}\} \right)^{\frac{1}{2}}}_{=: c_{\mathcal{T}_h}} ||v||_{H^1(\Omega)} \ .
\end{aligned}$$



(a) innerer Punkt (b) einfacher Randpunkt (c) zweifacher Randpunkt

Abbildung 5.2: Klassifizierung von Punkten

Hierbei zählt $c_{a,T}$ bzw. $c_{b,T}$, wie oft über das Dreieck T integriert wird. Beispielsweise gilt $c_{a,T} = 11$ in Abbildung 5.1. Es setzt sich aus der Anzahl der angrenzenden Dreiecke und T selbst zusammen. Für ein Dreieck T mit Eckpunkten P_1, P_2 und P_3 lässt sich $c_{a,T}$ in der Form

$$c_{a,T} = 1 + (N_1 - 1) + (N_2 - 1) + (N_3 - 1) - j = N_1 + N_2 + N_3 - 2 - j$$

ausdrücken. Dabei gibt N_i an, wie viele Dreiecke den Punkt P_i als Eckpunkt besitzen und j bezeichnet die Anzahl der Randkanten von T. Offensichtlich ist $N_1 + N_2 + N_3$ eine obere Schranke für $c_{a,T}$.
Analog ergibt sich für $c_{b,T}$

$$c_{b,T} = 3 + (M_1 - 2) + (M_2 - 2) + (M_3 - 2) = M_1 + M_2 + M_3 - 3 \ ,$$

wobei M_i die Anzahl der Kanten mit dem Punkt P_i bezeichnet. Aus den verfügbaren Datenstrukturen der Matlab PDE-Toolbox ist M_i nur schwer zu berechnen. Jedoch besteht der folgende einfache Zusammenhang zwischen M_i und N_i. Es gilt $M_i = N_i$, falls P_i ein innerer Punkt ist, $M_i = N_i + 1$, falls P_i ein einfacher Randpunkt ist und allgemein $M_i = N_i + n$, falls P_i ein n-facher Randpunkt ist. Worum es sich bei dieser Klassifizierung der Punkte handelt, soll durch Abbildung 5.2 verdeutlicht werden. Unter der gemachten Voraussetzung, dass Ω ein Lipschitz-Gebiet ist, kann es nur innere Punkte und einfache Randpunkte geben. Daraus ergibt sich:

$$c_{b,T} \leq N_1 + N_2 + N_3 \ .$$

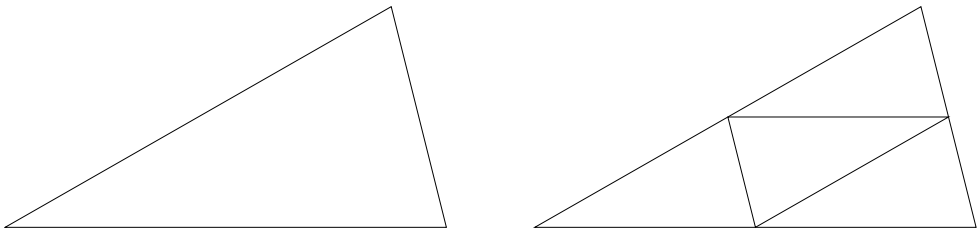


Abbildung 5.3: reguläre Verfeinerung

Größere Schwierigkeiten bereiten die zwei Konstanten c_1 und c_2 bzw. das benötigte Maximum der beiden. Es wurden verschiedene Tests mit linearen partiellen Differentialgleichungen vom Typ (5.12) durchgeführt. Die Lösung y und eine Funktion c wurden vorgegeben und daraus die Funktion g berechnet. Anschließend erfolgte das Lösen der Aufgabe und der Vergleich des Abstandes der berechneten zur exakten, bekannten Lösung, d.h. $||y - y_h||_{H_0^1(\Omega)}$. Hierbei spielt der Effektivitäts-Index $\eta_{\text{eff}} = \frac{\bar{c}\,\eta_h}{||y-y_h||_{H_0^1(\Omega)}}$ eine wichtige Rolle. Es wurde festgestellt, dass sich η_{eff} für $h = \max\limits_{T \in \mathcal{T}_h} h_T \to 0$ einem festen Wert nähert. Daraus motiviert ist die Wahl für c_1 und c_2. Es wird

$$c_1 = c_2 = 0{,}04 = \max\{c_1, c_2\}$$

verwendet. Für diese Wahl war η_{eff} stets größer 1, so dass von der Gültigkeit der Ungleichung (5.18) ausgegangen wird.

5.4 Verfeinerungsstrategie

Wie im Algorithmus 3.3 beschrieben, muss der Raum X_h verfeinert werden, falls die Lösung in diesem Raum nicht den Genauigkeitsanforderungen genügt. Das kann auf verschiedene Weise geschehen.

Die einfachste Möglichkeit besteht in der globalen, regulären Verfeinerung des Gitters. Darunter versteht man das Aufteilen aller Dreiecke in vier kongruente Dreiecke. Wie in Abbildung 5.3 dargestellt, werden dazu die Seitenmittelpunkte als neue Punkte eingefügt und miteinander verbunden. Da die vier neuen Dreiecke ähnlich zum alten Dreieck sind, haben sie die gleichen Winkel, so dass sich insbesondere der kleinste Winkel nicht

verändert. Bei einer solchen Verfeinerung vervierfacht sich ungefähr die Anzahl der Punkte bzw. Freiheitsgrade, der Gitterparameter $h = \max_{T \in \mathcal{T}_h} h_T$ halbiert sich.

Eine andere Möglichkeit ist die Verfeinerung ausgewählter Dreiecke. Die Auswahl solcher Dreiecke erfolgt anhand der lokalen Fehlerindikatoren $\eta_{h,T}$ aus dem vorherigen Abschnitt 5.3. So können Dreiecke mit einem hohen Anteil am Gesamtfehler ausgewählt und verfeinert werden. Dreiecke T, die die Bedingung

$$\eta_{h,T} \geq \beta \max_{T \in \mathcal{T}_h} \eta_{h,T} \tag{5.19}$$

mit dem Parameter β erfüllen, werden für die Verfeinerung markiert. Dies ist eine gängige Auswahlstrategie, die auch hier mit $\beta = 0{,}8$ angewandt wird. Die markierten Dreiecke werden, wie bei der regulären Verfeinerung, in vier kongruente Dreiecke aufgeteilt. Dabei werden die drei Kanten in ihrer Mitte geteilt und auf jeder Kante entsteht ein neuer Punkt. Eventuell ist das Gitter dann nicht mehr zulässig. Um dies zu beheben, wird die Strategie aus [RS75] verwendet. In dieser erfolgt nach der regulären Verfeinerung der markierten Dreiecke die Teilung der längsten Kante derjenigen Dreiecke, die eine geteilte Kante besitzen. Dieser Schritt wird solange wiederholt, bis keine weiteren Kanten mehr geteilt werden. Danach beginnt die Bildung neuer Dreiecke aus den geteilten Kanten bzw. den neuen Punkten. Ein Dreieck mit drei geteilten Kanten wird, wie bei der regulären Verfeinerung, in vier neue Dreiecke aufgespalten. Hat ein Dreieck nur zwei geteilte Kanten, wird der Mittelpunkt der längsten Kante mit der gegenüberliegenden Ecke und dem verbleibenden neuen Punkt verbunden. Bei Dreiecken mit nur einer geteilten Kante wird der neue Punkt mit der gegenüberliegenden Ecke verbunden. In Abbildung 5.4 ist ein Beispiel dargestellt, bei dem die beiden Dreiecke T_1 und T_2 für die Verfeinerung markiert sind.

Ein großer Vorteil besteht darin, dass sich der kleinste Winkel im Verlauf mehrerer solcher Verfeinerungen höchstens halbiert. Das Verfahren ist bereits in der hier verwendeten Matlab PDE-Toolbox implementiert.

Nachteilig am Auswahlkriterium (5.19) erweist sich, dass unter Umständen nur ein Dreieck für die Verfeinerung markiert wird. Dadurch kann der Fortschritt sehr gehemmt sein. Eine andere Möglichkeit ist die Verfeinerung eines festen Anteils an Dreiecken. So könnten beispielsweise die $0{,}2 \cdot |\mathcal{T}_h|$ schlechtesten Dreiecke ausgewählt und verfeinert werden.

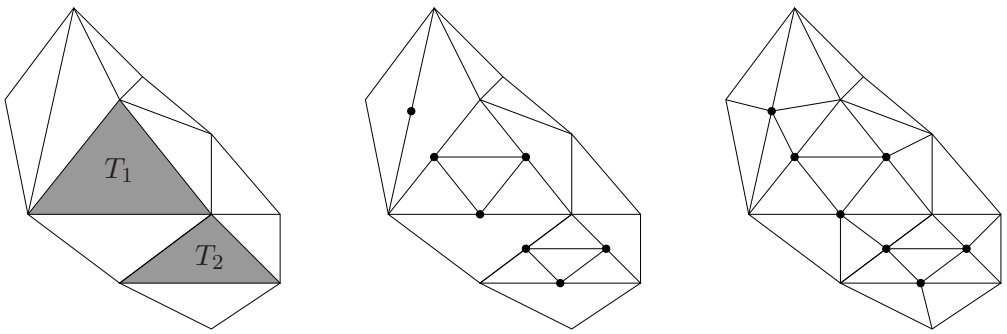


Abbildung 5.4: Verfeinerung nach [RS75]

Auch andere Verfeinerungsstrategien wurden getestet. Zum Beispiel wurde versucht, die ausgewählten Dreiecke gleich mehrfach zu verfeinern oder den Parameter β anhand der geschätzten noch benötigten Genauigkeit zu steuern. Diese Strategie erwies sich jedoch als nicht erfolgreich, denn in den meisten Fällen wurde β so klein gewählt, dass eine reguläre Verfeinerung stattfand. Das werden auch die Iterationsverläufe im nächsten Abschnitt belegen.

5.5 Numerische Resultate

In diesem Abschnitt sollen verschiedene numerische Ergebnisse des Verfahrens für ein konkretes Beispiel vom Typ (5.1) wiedergegeben werden. Für die nichtlineare Funktion d wird $\mathrm{d}(x, y) = y^3$ gewählt. Sie erfüllt alle gestellten Voraussetzungen. Wie bereits in Abschnitt 5.3.2 erwähnt, wird für Ω der Einheitskreis verwendet. Die Lösung y^* wird vorgegeben und dazu die entsprechende Funktion f aus

$$\mathrm{f} := -\triangle y^* + {y^*}^3$$

berechnet. Damit ist es möglich, den Abstand der Iterierten zur Lösung y^* zu ermitteln. Es wird

$$y^*(x = (x_1, x_2)) = \left(1 - x_1^2 - x_2^2\right) \sin(20 x_1 x_2) \exp(4 x_1)$$

gewählt. Die Lösung y^* ist in Abbildung 5.5 dargestellt und erfüllt offensichtlich die Randbedingung $y^* = 0$ auf $\partial\Omega$. Sie besteht aus einem flachen und aus einem „turbulenteren“ Teil. Für die $H_0^1(\Omega)$-Norm von y^* gilt: $||y^*||_{H_0^1(\Omega)} \approx 52{,}26$. Damit erhält man

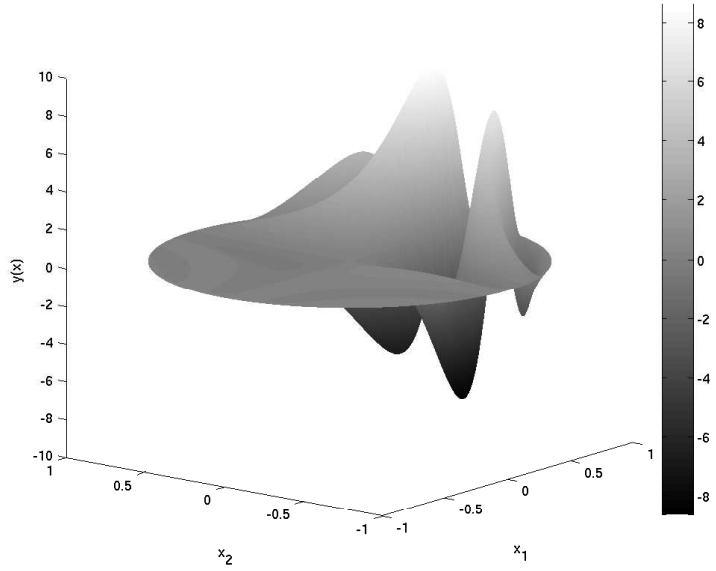


Abbildung 5.5: Lösung $y^*(x)$ in Ω

ein Gefühl für den relativen Fehler $\frac{||y^*-y_n||_{H_0^1(\Omega)}}{||y^*||_{H_0^1(\Omega)}}$ von y_n. Um das vorgeschlagene Verfahren vergleichen zu können, werden zunächst Ergebnisse von zwei einfachen Ansätzen gezeigt.

5.5.1 Konvergenz auf einem festen Gitter

Im Folgenden wird das „normale“ Newton-Verfahren für eine feste Diskretisierung (d.h. ein festes Gitter) angewandt. Dies geschieht auf unterschiedlich feinen Gittern, welche aus einem Anfangsgitter nach entsprechend häufiger globaler, regulärer Verfeinerung hervorgingen. Als Startlösung wird $y_0 = 0$ verwendet.

In Tabelle 5.1 ist der Verlauf des Newton-Verfahrens auf einem Gitter mit 24833 Knoten dargestellt. Gut zu erkennen ist die erwartete quadratische Konvergenz des Verfahrens, d.h. $\frac{||\delta y_n||_{H_0^1(\Omega)}}{||\delta y_{n-1}||^2_{H_0^1(\Omega)}}$ bleibt von oben beschränkt. Nach 6 Schritten hat das Iterationsverfahren eine Genauigkeit von `5.24E+00` erreicht und ist fertig konvergiert.

Iteration n	$\|\|y^* - y_{n+1}\|\|_{H_0^1(\Omega)}$	$\|\|\delta y_n\|\|_{H_0^1(\Omega)}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|^2_{H_0^1(\Omega)}}$
0	9.75E+00	5.89E+01	
1	5.31E+00	7.39E+00	2.13E-03
2	5.24E+00	7.65E-01	1.40E-02
3	5.24E+00	1.08E-02	1.85E-02
4	5.24E+00	2.26E-06	1.92E-02
5	5.24E+00	2.46E-13	4.82E-02

Tabelle 5.1: Normales Newton-Verfahren auf einem festen Gitter mit 24833 Knoten

Knoten p	benötigte Iterationen m	$\|\|y^* - y_m\|\|_{H_0^1(\Omega)}$
33	8	4.57E+01
113	6	4.99E+01
417	6	3.67E+01
1601	6	2.03E+01
6273	6	1.04E+01
24833	6	5.24E+00
98817	6	2.62E+00
394241	6	1.31E+00

Tabelle 5.2: Erreichte Genauigkeit für verschiedene Gittergrößen

Diese Vorgehensweise wurde für verschiedenen Gittergrößen wiederholt. Tabelle 5.2 stellt die erreichte Genauigkeit auf dem entsprechenden Gitter nach dem Fertig-Iterieren des Verfahrens dar. Bemerkenswert ist, dass die Genauigkeit auf dem Gitter mit 113 Knoten größer als auf dem gröberen Gitter mit nur 33 Knoten ist. Dies spricht für den Einfluss der Nichtlinearität. Ab einer Gittergröße von 417 ist zu erkennen, dass die Genauigkeit (bzw. der Fehler) $||y^* - y_m||_{H_0^1(\Omega)}$ in der Ordnung $\mathcal{O}(h)$ liegt[1]. Dies entspricht der erwarteten Konvergenzordnung für lineare partielle Differentialgleichungen. Die Nichtlinearität ist hier kaum noch von Bedeutung.

5.5.2 Akzeptieren jedes Schrittes

Eine weitere Möglichkeit besteht in der Akzeptanz jedes Schrittes. Die Verfeinerung des Gitters erfolgt wie in Abschnitt 5.4 beschrieben. Zusätzlich wird der neue Punkt stets

[1] Eine reguläre Verfeinerung vervierfacht die Anzahl der Knoten und halbiert h sowie den Fehler.

akzeptiert, unabhängig davon, ob der jeweilige Schritt der Genauigkeitsforderung des Verfahrens genügt. Im Vergleich zu den Ergebnissen des vorigen Abschnittes sieht man in Tabelle 5.3 ungefähr die doppelte Genauigkeit bei ähnlicher Gittergröße. Zum Beispiel beträgt diese hier bei einem Gitter mit ca. 400 000 Knoten $||y^* - y_{48}||_{H_0^1(\Omega)} =$ `7.55E-01` im Gegensatz zu $||y^* - y_6||_{H_0^1(\Omega)} =$ `1.31E+00` aus dem vorhergehenden Abschnitt. Um auf einem festen Gitter eine Genauigkeit von `7.55E-01` zu erreichen, werden fast 1 600 000 Knoten benötigt. Allerdings ist dieser Vorteil nicht so groß, wie es zunächst erscheint, da deutlich mehr Iterationen benötigt werden. Diese sind zur adaptiven Anpassung des Gitters notwendig, um den Vorteil in der Genauigkeit zu erreichen. Welches der beiden bisherigen Verfahren effektiv schneller ist, wird in dieser Arbeit nicht diskutiert. Denn die Geschwindigkeit hängt vor allem davon ab, wie schnell ein lineares Gleichungssystem mit p Unbekannten gelöst werden kann. Da es möglich ist, nur ein Dreieck zur Gitterverfeinerung zu markieren, erlaubt das hier verwendete Auswahlkriterium keine Aussage über einen Mindestfortschritt der Gittergröße, was die Gegenüberstellung der Geschwindigkeiten zusätzlich erschwert.

Der wohl größte Nachteil im Vergleich zum Verfahren aus Kapitel 3 ist die fehlende Konvergenzkontrolle. Anhand von $\tilde{\theta}_{n-1} = \frac{||\delta y_n||_{H_0^1(\Omega)}}{||\delta y_{n-1}||_{H_0^1(\Omega)}}$ lässt sich eine „schwache“ lineare Konvergenz feststellen. Mit „schwach“ ist gemeint, dass der Kontraktionfaktor $\tilde{\theta}_n$ meist knapp unter 1, jedoch manchmal über 1 liegt. An der letzten Spalte ist deutlich erkennbar, dass es sich nicht um quadratische Konvergenz handelt. Zu erwarten ist, dass das in diesem Abschnitt beschriebene Verfahren die beste Genauigkeit liefert.

5.5.3 Quadratischer Konvergenzmodus

Es wird das in Kapitel 3 vorgestellte Verfahren mit dem quadratischen Konvergenzmodus angewandt (siehe Algorithmus 3.3). Dabei wird auch hier das Gitter, wie in Abschnitt 5.4 beschrieben, verfeinert. Anders als im Algorithmus wird auf die „a-postposteriori“ Überprüfung der Genauigkeit von δy_0 verzichtet. Der erste Schritt wird stets akzeptiert, um einen Stillstand zu vermeiden. Nachteilig erweist sich, dass δy_0 nicht mehr zur Theorie passen muss. Das setzt sich durch $\tilde{\theta}_0$ bis zur nächsten Iteration fort. Zusätzlich erfolgt die Unterscheidung, ob auf einem kleinen Gitter mit Startpunkt $y_0 = 0$ begonnen wird oder ob y_0 als Lösung des normalen Newton-Verfahrens (auf einem festen Gitter, vgl. Abschnitt 5.5.1) hervorgegangen ist.

Iteration n	Knoten p	$\|\|y^* - y_{n+1}\|\|_{H_0^1(\Omega)}$	$\|\|\delta y_n\|\|_{H_0^1(\Omega)}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|_{H_0^1(\Omega)}}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|^2_{H_0^1(\Omega)}}$
0	33	5.17E+01	3.81E+01		
1	46	5.31E+01	2.92E+01	7.67E-01	2.01E-02
2	50	5.29E+01	1.82E+01	6.22E-01	2.13E-02
3	58	5.07E+01	1.35E+01	7.44E-01	4.09E-02
4	72	4.76E+01	1.43E+01	1.06E+00	7.84E-02
5	93	4.35E+01	1.81E+01	1.26E+00	8.77E-02
6	123	3.77E+01	1.58E+01	8.73E-01	4.83E-02
7	132	3.69E+01	5.37E+00	3.41E-01	2.16E-02
8	157	3.44E+01	9.94E+00	1.85E+00	3.44E-01
9	216	3.03E+01	1.15E+01	1.16E+00	1.17E-01
10	286	2.75E+01	9.04E+00	7.83E-01	6.78E-02
11	315	2.61E+01	5.50E+00	6.08E-01	6.73E-02
12	427	2.26E+01	8.55E+00	1.56E+00	2.83E-01
⋮	⋮	⋮	⋮	⋮	⋮
18	1641	1.17E+01	3.72E+00	8.35E-01	1.88E-01
⋮	⋮	⋮	⋮	⋮	⋮
24	6268	6.00E+00	1.77E+00	8.16E-01	3.77E-01
⋮	⋮	⋮	⋮	⋮	⋮
31	25019	3.00E+00	8.84E-01	9.47E-01	1.01E+00
⋮	⋮	⋮	⋮	⋮	⋮
38	92059	1.56E+00	4.11E-01	8.80E-01	1.88E+00
⋮	⋮	⋮	⋮	⋮	⋮
47	387654	7.55E-01	1.81E-01	8.41E-01	3.91E+00

Tabelle 5.3: Iterationsverlauf, falls jeder Schritt akzeptiert wird

Iteration n	Knoten p	$\|\|y^* - y_{n+1}\|\|_{H_0^1(\Omega)}$	$\|\|\delta y_n\|\|_{H_0^1(\Omega)}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|_{H_0^1(\Omega)}}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|^2_{H_0^1(\Omega)}}$
0 (0)	33	5.17E+01	3.81E+01		
1 (20)	1655	3.48E+01	4.37E+01	1.15E+00	3.01E-02
	1655	2.17E+01	1.84E+01		
2 (24)	100115	1.86E+01	1.98E+01	4.52E-01	1.03E-02
3 (0)	100115	7.46E+00	1.40E+01	7.09E-01	3.59E-02
	100115	2.24E+00	7.66E+00		
4 (18)	437709	1.77E+00	7.71E+00	5.51E-01	3.93E-02
	437709	7.10E-01	1.59E+00		

Tabelle 5.4: Quadratischer Konvergenzmodus mit $y_0 = 0$

Der Verlauf beider Varianten ist in den Tabellen 5.4 und 5.5 dargestellt. Die geklammerte Zahl hinter dem Iterationszähler gibt an, wie oft versucht wurde, einen zulässigen Schritt zu finden, der jedoch wegen zu geringer Genauigkeit verworfen werden musste. Steht in einer Zeile kein Iterationszähler, handelt es sich um berechnete Schritte, die wegen zu geringer Genauigkeit nicht akzeptiert worden sind. Sie werden ausgegeben, um die einzelnen Strategien besser vergleichen zu können. In der dritten Spalte befindet sich dann der Abstand zur Lösung, falls der Schritt gegangen worden wäre. Von besonderem Interesse ist der erste versuchte Schritt nach einem akzeptierten Schritt δy_n. Dieser wird mit $\bar{\delta y}_{n+1}$ bezeichnet. Die Größen δy_n und $\bar{\delta y}_{n+1}$ werden dabei auf demselben Gitter berechnet. Eine wichtige Rolle spielt das Verhältnis von $||\delta y_n||_{H_0^1(\Omega)}$ zur Schrittlänge des Kandidaten $\left|\left|\bar{\delta y}_{n+1}\right|\right|_{H_0^1(\Omega)}$. Es zeigt an, inwieweit das Newton-Verfahren auf diesem Gitter konvergiert ist. Ist $\frac{||\bar{\delta y}_{n+1}||_{H_0^1(\Omega)}}{||\delta y_n||_{H_0^1(\Omega)}}$ klein, spricht das für eine fast abgeschlossene Konvergenz.

Insgesamt werden nur wenige Iterationen gemacht. Für einen einzelnen Iterationsschritt sind allerdings viele Fehlversuche notwendig. Auffällig ist, dass in Tabelle 5.4 für ein Gitter mit 100115 Knoten direkt zwei Schritte hintereinander akzeptiert werden. Das legt die Vermutung nahe, dass die Nichtlinearität aufgrund der wenigen Iterationsschritte noch nicht überwunden ist. Dafür spricht ebenfalls das oben genannte Verhältnis $\frac{\left|\left|\bar{\delta y}_{n+1}\right|\right|_{H_0^1(\Omega)}}{||\delta y_n||_{H_0^1(\Omega)}}$ nach jedem akzeptierten Schritt, welches hier bis einschließlich der letzten Iteration relativ groß ist. Im Gegensatz zu dem Verfahren aus dem vorigen Abschnitt 5.5.2 kann hier noch ein Fortschritt ohne Verfeinerung des Gitters erzielt werden. Erst mit dem

Iteration n	Knoten p	$\|\|y^* - y_{n+1}\|\|_{H_0^1(\Omega)}$	$\|\|\delta y_n\|\|_{H_0^1(\Omega)}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|_{H_0^1(\Omega)}}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|^2_{H_0^1(\Omega)}}$
0 (0)	1601	2.03E+01	2.31E-14		
1 (13)	7965	5.46E+00	1.18E+01	5.11E+14	2.21E+28
	7965	5.46E+00	2.38E-02		
	26469	2.93E+00	2.82E+00		
	104375	1.46E+00	3.22E+00		
	404380	7.40E-01	3.32E+00		

Tabelle 5.5: Quadratischer Konvergenzmodus, bei dem y_0 die Newton-Lösung auf einem Gitter mit 1601 Knoten ist

letzten Schritt würde eine größere Genauigkeit als mit dem normalen Newton-Verfahren (Abschnitt 5.5.1) erreicht werden. Sie ist mit 7.10E-01 in etwa so groß wie bei der in Abschnitt 5.5.2 verfolgten Strategie. Die geforderte Genauigkeit hindert das Verfahren stark am Voranschreiten. Ob es sich tatsächlich um quadratische Konvergenz handelt, ist unklar, da die geringe Anzahl an Iterationen darüber keine Aussage erlaubt. Erschwerend kommt hinzu, dass δy_0 wegen des oben beschrieben Problems nicht zur Theorie passen muss und damit auch $\frac{||\delta y_1||_{H_0^1(\Omega)}}{||\delta y_0||_{H_0^1(\Omega)}}$ und $\frac{||\delta y_1||_{H_0^1(\Omega)}}{||\delta y_0||^2_{H_0^1(\Omega)}}$ „fehlerbehaftet“ sein können.

In Tabelle 5.5 sind die Auswirkungen des „fehlerbehafteten“ δy_0 besonders deutlich zu erkennen. Da mit der Newton-Lösung auf einem Gitter mit 1601 Knoten gestartet wird, ist die erste Schrittlänge mit $||\delta y_0||_{H_0^1(\Omega)} =$ 2.31E-14 sehr klein. Das hat entsprechende Auswirkung auf die nachfolgende Iteration. Insgesamt wird nur noch ein weiterer Schritt akzeptiert. Wie die letzten drei Zeilen nahe legen, resultiert ein möglicher Fortschritt nur aus der Verfeinerung des Gitters. Das heißt, dass die normale Newton-Iteration auf dem festen Gitter die Nichtlinearität überwunden hat. Die im Verhältnis zu Zeile 2 kleine Schrittlänge in Zeile 3 untermauert das. Das eigentliche Verfahren startet mit einem „fast lineraren Problem“. Um für eine lineare partielle Differentialgleichung (für die der Fehler der diskreten Lösung mit der Ordnung $\mathcal{O}(h)$ gegen 0 strebt) eine quadratische Konvergenz zu erreichen, muss h quadratisch gegen 0 konvergieren[1]. Dies stellt eine enorme Forderung dar. Der lineare Konvergenzmodus im folgenden Abschnitt versucht, diese abzuschwächen.

[1]Das ist nicht ganz exakt. Der Diskretisierungsfehler einer linearen partiellen Differentialgleichung wird mit der Schrittlänge des Verfahrens in Verbindung gesetzt.

Iteration n	Knoten p	$\|\|y^* - y_{n+1}\|\|_{H_0^1(\Omega)}$	$\|\|\delta y_n\|\|_{H_0^1(\Omega)}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|_{H_0^1(\Omega)}}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|^2_{H_0^1(\Omega)}}$
0 (15)	873	1.81E+01	5.70E+01		
	873	1.60E+01	6.81E+00		
1 (16)	19658	3.44E+00	1.20E+01	2.10E-01	3.69E-03
	19658	3.39E+00	5.52E-01		
	23512	3.10E+00	1.01E+00		
	99440	1.50E+00	1.95E+00		
	408819	7.34E-01	2.11E+00		
2 (21)	653645	5.84E-01	2.13E+00	1.78E-01	1.49E-02
	653645	5.84E-01	5.39E-03		

Tabelle 5.6: Linearer Konvergenzmodus mit $y_0 = 0$

Iteration n	Knoten p	$\|\|y^* - y_{n+1}\|\|_{H_0^1(\Omega)}$	$\|\|\delta y_n\|\|_{H_0^1(\Omega)}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|_{H_0^1(\Omega)}}$	$\frac{\|\|\delta y_n\|\|_{H_0^1(\Omega)}}{\|\|\delta y_{n-1}\|\|^2_{H_0^1(\Omega)}}$
0 (17)	17295	3.63E+00	1.21E+01		
	17295	3.63E+00	2.62E-02		
	26458	2.93E+00	1.33E+00		
	104295	1.46E+00	2.04E+00		
	404281	7.40E-01	2.19E+00		
1 (21)	520382	6.52E-01	2.20E+00	1.81E-01	1.50E-02
	520382	6.52E-01	4.20E-05		

Tabelle 5.7: Linearer Konvergenzmodus, bei dem y_0 die Newton-Lösung auf einem Gitter mit 1601 Knoten ist

5.5.4 Lineare Konvergenz

In diesem Teil erfolgt die Vorstellung der numerischen Ergebnisse, der in Abschnitt 3.2.2 beschriebenen Methode. Dazu wird $\bar{\delta} = \frac{1}{3}$ gewählt. Analog zum vorigen Abschnitt werden beide Varianten für den Startpunkt y_0 betrachtet und, wie in Abschnitt 5.4 beschrieben, verfeinert. Die Verläufe sind in den Tabellen 5.6 und 5.7 wiedergegeben. Die im quadratischen Konvergenzmodus angesprochene Problematik mit δy_0 tritt hier nicht auf. Die Forderung $\delta_n \leq \bar{\delta}$ wird auch für $n = 0$ erfüllt.

Genauso wie beim quadratischen Konvergenzmodus ist die Iterationsanzahl gering. Gegensätzlich ist, dass im Fall $y_0 = 0$ bereits mit der zweiten Iteration (gemeint ist Zeile 3 in Tabelle 5.6) der „übliche“ Fehler aus Abschnitt 5.5.2 erreicht wird. Ähnlich wie im vorigen

Abschnitt spiegelt sich dies auch beim Start mit der Newton-Lösung wieder (Tabelle 5.7). In diesen Fällen hängt der Fortschritt wieder fast ausschließlich von der Verfeinerung des Gitters ab. Gestärkt wird diese Aussage von den kleinen Verhältnissen $\frac{||\bar{\delta y}_{n+1}||_{H_0^1(\Omega)}}{||\delta y_n||_{H_0^1(\Omega)}}$ in beiden Tabellen. Wieder ist aufgrund der geringen Iterationsanzahl schwer einzuschätzen, um welche Konvergenzgeschwindigkeit es sich tatsächlich handelt.

6 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein inexaktes, affin invariantes Newton-Verfahren im Banachraum vorgestellt. Dabei wurde sowohl auf eine lokale als auch auf eine globale Phase eingegangen. Für Erstere konnte die lineare und die quadratische Konvergenz bewiesen werden. Im quadratischen Konvergenzmodus sind die Kantorovitsch-Größen h_n^δ ausschlaggebend für die geforderte Genauigkeit. Nachteilig daran ist, dass diese meist nur geschätzt werden können und dabei keine Mindestgüte garantiert werden kann. Aus einer schlechten Schätzung resultiert eine übermäßig starke Genauigkeitsanforderung. Der lineare Konvergenzmodus hat diesen Nachteil nicht. Für ihn genügt es, wenn die relativen Genauigkeiten δ_n unterhalb einer Grenze $\bar{\delta}$ bleiben.

Die im Anschluss daran vorgestellte globale Phase hat ebenso diese vorteilhafte Eigenschaft. Sie nutzt das natürliche Abstiegskriterium als Maß für den Fortschritt und erfüllt damit die wichtige affine Invarianz, ist jedoch nicht global konvergent. Da die Anforderungen an die relative Genauigkeit denen im linearen Konvergenzmodus der lokalen Phase ähneln, ist ein einfacher Übergang in die lokale Phase möglich.

In Kapitel 5 wurde der lokale Teil des Verfahrens am Beispiel einer semilinearen partiellen Differentialgleichung angewandt. Insgesamt erscheint das gewählte Beispiel zum Untersuchen der lokalen Phase etwas unpassend. Denn nach wenigen Iterationen hatte es die Eigenschaften einer linearen partiellen Differentialgleichung, was wiederum nicht verwunderlich ist, da es sich um die lokale Phase handelt. Problematisch dabei ist, dass nur mit viel rechenintensivem Aufwand der von der linearen bzw. quadratischen Konvergenz geforderte Fortschritt realisiert werden kann.
Der vom vorgeschlagenen Verfahren erreichte Vorteil in der erzielten Genauigkeit gegenüber dem Newton-Verfahren, welches mit einer festen Diskretisierung arbeitet, gelang vor allem durch die adaptive Steuerung der Diskretisierung bzw. des Gitters.

Für die Zukunft empfiehlt sich daher die Implementierung der vorgeschlagenen globalen Phase, um damit auch andere Problemstellungen behandeln zu können. Solche sind zum Beispiel neben Aufgaben der optimalen Steuerung auch endlich-dimensionale Probleme wie die bekannte „Rosenbrock-Funktion“. Eine detailliertere Analyse des Verfahrens ist bei endlich-dimensionalen Problemen möglich, weil die relative Genauigkeit exakt bestimmt und sogar eingestellt werden kann. Bei unendlich-dimensionalen Aufgaben muss besondere Sorgfalt auf die angesprochene Vergröberung der Diskretisierung gelegt werden.

Um das Verfahren auf Aufgaben der optimalen Steuerung anwenden zu können, ist ein entsprechender Fehlerschätzer notwendig. In vielen Fehlerschätzern tritt eine häufig unbekannte Konstante auf. Denkbar ist, dass unter zusätzlichen Annahmen auf diese Konstante verzichtet werden kann, indem nur auf die Quotienten des Fehlerschätzers aus verschiedenen Iterationen zurückgegriffen wird. Es stellt sich die Frage, ob in diesem Fall trotzdem eine Konvergenz gesichert werden kann.

Zu guter Letzt ist es noch möglich, die durch die Diskretisierung auftretenden großen linearen Gleichungssysteme nur inexakt zu lösen. Hierfür kommen verschiedene CG-artige Algorithmen in Frage. Es wäre zu untersuchen, inwieweit man den zusätzlichen Fehler in der relativen Genauigkeit des Verfahrens verbergen kann und welche Algorithmen geeignet sind.

Literaturverzeichnis

[AF03] Adams, R. und J. Fournier: *Sobolev Spaces*, 2003.

[AO87] Ascher, U. und M.R. Osborne: *A note on solving nonlinear equations and the natural criterion function.* Journal of Optimization Theory and Applications, 55(1):147–152, 1987.

[Cle75] Clement, P.: *Approximation by finite element functions using local regularization.* Rev. Francaise Automat. Informat. Recherche Operationnelle Ser. Rouge Anal. Numer, 9:77–84, 1975.

[DES82] Dembo, R.S., S.C. Eisenstat und T. Steihaug: *Inexact Newton Methods.* SIAM Journal on Numerical Analysis, 19(2):400–408, 1982.

[Deu91] Deuflhard, P.: *Global inexact Newton methods for very large scale nonlinear problems.* Impact of Computing in Science and Engineering, 3:366–393, 1991.

[Deu04] Deuflhard, P.: *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms.* Springer Verlag, 2004.

[Gri07] Griesse, R.: *Skript zur Vorlesung Optimale Steuerung partieller Differentialgleichungen*, 2007.

[KA64] Kantorovich, L.V. und G.P. Akilov: *Funktionalanalysis in normierten Räumen.* Akademie-Verlag, 1964.

[Kel99] Kelley, C.T.: *Iterative methods for optimization.* SIAM Philadelphia, 1999.

[KS84] Kuttler, J.R. und V.G. Sigillito: *Eigenvalues of the Laplacian in Two Dimensions.* SIAM Review, 26(2):163–193, 1984.

[Ran06] Rannacher, R.: *Skript zur Vorlesung Numerische Mathematik 2 (Numerik Partieller Differentialgleichungen)*, 2006.

[RS75] Rosenberg, I.G. und F. Stenger: *A Lower Bound on the Angles of Triangles Constructed by Bisecting the Longest Side*. Mathematics of Computation, 29(130):390–395, 1975.

[Trö05] Tröltzsch, F.: *Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen*. Vieweg Verlag, 2005.

[Ver96] Verfürth, R.: *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley-Teubner, 1996.

[Ypm84] Ypma, T.J.: *Local Convergence of Inexact Newton Methods*. SIAM Journal on Numerical Analysis, 21(3):583–590, 1984.

Danksagung

Ich möchte mich beim RICAM[1] und seinen Mitarbeitern für die sehr gute Zusammenarbeit und die vielen gemeinsamen Aktivitäten bedanken. Mein besonderer Dank geht an meinen Betreuer Dr. R. Griesse, der mir während der gesamten Zeit mit Rat und Vorschlägen zur Seite stand. Ebenso bedanke ich mich bei Prof. Dr. C. Helmberg, der die Betreuung der Arbeit an der Fakultät für Mathematik in Chemnitz übernahm.

[1] Johann Radon Institute for Computational and Applied Mathematics (www.ricam.oeaw.ac.at)

Eidesstattliche Erklärung

Ich erkläre an Eides Statt, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Chemnitz, den 5. Dezember 2007

Frank Schmidt