



Fachbereich Mathematik, Naturwissenschaften und Datenverarbeitung  
Studiengang Wirtschaftsmathematik

## Abschlussarbeit

zur Erlangung des akademischen Grades

Bachelor of Science

# Unrestringierte Optimierungsprobleme: Gradienten- und Trust-Region-Verfahren im Vergleich

Vorgelegt von Büsra Karaoglan am 14. Februar 2017  
(Matrikelnummer: 5056046)

Referent Prof. Dr. Rigger  
Korreferent Prof. Dr. Müller

## Zusammenfassung

Das Ziel dieser Arbeit ist es, ein unrestringiertes nichtlineares Optimierungsproblem zu lösen. Hierfür werden zwei geeignete numerische Verfahren, das Gradientenverfahren und Trust-Region-Verfahren, zur Lösung unrestringierter nichtlinearer Optimierungsaufgaben vorgestellt. Zudem werden diese Verfahren an verschiedenen Beispielen ausführlich ausgearbeitet und konkretisiert. Das Gradientenverfahren, zur nähерungsweisen Lösung des unrestringierten Optimierungsproblems, setzt sich aus einer Richtungs- und einer Schrittweitenstrategie zusammen. Von einer aktuellen Näherung ausgehend, wird in eine Abstiegsrichtung  $s^k$  gegangen. In diese wird ein Schritt gemacht, dessen Länge  $\sigma_k$  durch die gewählte Schrittweitenstrategie in dieser Arbeit mit der Armijo-Regel festgelegt ist.

Beim Trust-Region-Verfahren ist der Ansatz genau umgekehrt. Die Idee besteht im Wesentlichen darin, die Zielfunktion lokal auf einer  $\Delta$ -Kugel um eine aktuelle Näherung durch ein einfacheres Modell zu ersetzen, etwa einer linearen oder quadratischen Approximation der Zielfunktion. Dann wird ein Minimum des Modells beziehungsweise der vereinfachten Zielfunktion auf dieser  $\Delta$ -Kugel bestimmt. Wird eine Verminderung des Zielfunktionswertes entweder nicht erreicht oder ist diese eher enttäuschend gering, so wird dem Modell auf einer zu großen Kugel um die aktuelle Näherung vertraut. Diese wird daher verkleinert und auf dieser verkleinerten Kugel wird dann erneut ein Minimum der Modelfunktion bestimmt. Andernfalls wird dieses Minimum als neue aktuelle Näherung akzeptiert und der Radius  $\Delta$  wird vergrößert, wenn ein verschärfter Test auf hinreichende Verminderung des Zielfunktionswertes erfolgreich bestanden wird. Insgesamt wird also eine Schrittweite  $\Delta$  vorgegeben und dann dazu eine geeignete Suchrichtung  $s^k$  bestimmt [WJ92].

Die Arbeit beschäftigt sich mit der Frage, welche der beiden vorgestellten Verfahren effizienter arbeitet. Zum Abschluss werden die numerischen Resultate beider Verfahren für die bekannte Rosenbrock-Funktion ausgewertet und verglichen.

# Inhaltsverzeichnis

---

<b>1 Einleitung</b>	<b>2</b>
1.1 Motivation . . . . .	2
1.2 Zielsetzung . . . . .	3
1.3 Aufbau der Arbeit . . . . .	3
<b>2 Mathematische Grundlagen</b>	<b>4</b>
2.1 Mathematische Notationen . . . . .	4
2.2 Optimalitätsbedingungen . . . . .	6
2.3 Konvexität . . . . .	9
<b>3 Das Gradientenverfahren</b>	<b>14</b>
3.1 Allgemeines Abstiegsverfahren . . . . .	14
3.2 Richtung des steilsten Abstiegs . . . . .	15
3.3 Die Armijo-Schrittweitenregel . . . . .	16
3.4 Globale Konvergenz des Gradientenverfahrens . . . . .	17
3.5 Konvergenzgeschwindigkeit des Gradientenverfahrens . . . . .	19
<b>4 Trust-Region-Verfahren</b>	<b>26</b>
4.1 Einleitung . . . . .	26
4.2 Globale Konvergenz . . . . .	36
4.3 Charakterisierung der Lösungen des Teilproblems . . . . .	40
4.4 Schnelle lokale Konvergenz . . . . .	45
<b>5 Numerische Resultate</b>	<b>50</b>
<b>A Anhang</b>	<b>60</b>

# 1 Einleitung

---

## 1.1 Motivation

Optimierungsaufgaben treten in zahlreichen Anwendungsproblemen in den Natur- und Ingenieurwissenschaften, der Wirtschaft oder der Industrie auf. Beispielsweise versuchen Transportunternehmen die Fahrt- oder Flugkosten zu minimieren und dabei sicherzustellen, dass alle Aufträge ausgeführt werden. Ebenso führt die numerische Simulation vieler physikalischer Vorgänge in den Naturwissenschaften auf Optimierungsprobleme, da das zugrundeliegende mathematische Modell oftmals auf dem Prinzip der Energieminimierung beruht [HH11].

Diese Arbeit beschäftigt sich mit unrestringierten Optimierungsaufgaben. Das bedeutet, es werden Aufgaben betrachtet, bei denen keine Nebenbedingungen beziehungsweise Restriktionen vorliegen. Hier ist eine Zielfunktion  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  gegeben. Das Problem besteht darin, einen Punkt zu finden, in dem diese Zielfunktion minimal ist. Das heißt es wird ein Problem der Form

$$\text{Minimiere } f(x), \quad x \in \mathbb{R}^n \tag{1.1}$$

betrachtet. Es wird zwischen

- einer *globalen Lösung* von (1.1), das heißt einem Punkt  $\bar{x} \in \mathbb{R}^n$  mit  $f(\bar{x}) \leq f(x)$  für alle  $x \in \mathbb{R}^n$ ,
- und einer *lokalen Lösung* von (1.1), das heißt einem Punkt  $\bar{x} \in \mathbb{R}^n$ , zu dem es eine Umgebung  $U$  gibt, so dass  $f(\bar{x}) \leq f(x)$  wenigstens für alle  $x \in U$  ist,

unterschieden [WJ92].

Nichtlineare Optimierungsprobleme können in der Regel nicht analytisch, sondern nur numerisch gelöst werden. Dementsprechend wird hier mit Hilfe von Iterationsverfahren Lösungen von (1.1) approximiert. Für das Problem (1.1) wird dabei ausgehend von einem Startpunkt  $x^0$  eine Folge  $\{x^k\}$ ,  $k = 1, 2, \dots$  berechnet, mit dem Ziel, dass die Folge  $x^k$  einen Grenzwert  $\bar{x}$  hat, der die lokale Lösung ist.

Ist das Minimum einer Funktion gesucht, so ist es bei gegebenem  $x^k$  naheliegend, bei der Berechnung von  $x^{k+1}$  das Ziel

$$f(x^{k+1}) < f(x^k)$$

anzustreben. Verfahren, die eine solche Strategie realisieren, werden *Abstiegsverfahren* genannt. Auch wenn das Verfahren nicht gegen ein (lokales) Minimum konvergiert, so wird doch in jeder Iteration der Funktionswert verkleinert und damit ein besserer Punkt berechnet, was in der Praxis oft schon zufriedenstellend ist.

Ein Abstiegsverfahren benutzt zur Berechnung von  $x^{k+1}$  eine Abstiegsrichtung. Das ist eine Richtung  $s^k$  mit der Eigenschaft

$$f(x^k + \sigma s^k) < f(x^k) \quad \forall \sigma \in ]0, \bar{\sigma}]$$

mit einem  $\bar{\sigma} > 0$ . Es gibt zwei prinzipielle Vorgehensweisen zur Konstruktion von Abstiegsverfahren:

- Verfahren mit Schrittweitensteuerung: Hier wird zunächst eine Abstiegsrichtung  $s^k$  aufgrund lokaler Informationen über die Zielfunktion im aktuellen Iterationspunkt  $x^k$  bestimmt. Dann wird eine Schrittweite  $\sigma_k \in ]0, \bar{\sigma}]$  berechnet, mit der ein möglichst großer Abstieg erzielt wird, und setzt  $x^{k+1} = x^k + \sigma_k s^k$ .

- Trust-Region-Verfahren: Hier wird basierend auf einem lokalen Modell der Zielfunktion (beispielsweise einer quadratischen Approximation der Zielfunktion) eine Trust-Region (Vertrauensbereich) berechnet, auf der das lokale Modell die Zielfunktion hinreichend gut approximiert. Das lokale Modell erlaubt dann die Berechnung einer Abstiegsrichtung  $s^k$ , und es wird  $x^{k+1} = x^k + s^k$  gesetzt [AW02].

Der zentrale Gegenstand der Arbeit ist eine ausführliche Einführung in die Theorie der nichtlinearen Optimierung sowie wichtiger Lösungsverfahren zu geben. Hier kann natürlich nur eine Auswahl aus diesem sehr umfangreichen Gebiet präsentiert werden. Diese Ausführungen werden durch mehrere Beispiele ergänzt. Die Besonderheit der Arbeit liegt darin, am Ende die ausgewählten Verfahren an einer Testfunktion miteinander zu vergleichen, und somit die typischen Verhaltensweisen der Algorithmen zu analysieren. Die folgende Arbeit bezieht sich hauptsächlich auf das Buch *Nichtlineare Optimierung* von Michael und Stefan Ulbrich aus dem Jahre 2012 [UU12].

## 1.2 Zielsetzung

Im Rahmen dieser Bachelorarbeit sollen die Methoden der Optimierung auf die *Rosenbrock-Funktion* angewendet werden. Die Funktion dient zur guten praktischen Veranschaulichung der Problematik von numerischen Verfahren und deren Konvergenzverhalten. Hierbei sind keine Restriktionen zu berücksichtigen und es werden dementsprechend auch nur unrestringierte Optimierungsverfahren herangezogen. Ziel dieser Arbeit ist es, die Vorstellung und Implementierung der geeigneten Verfahren zur nichtlinearen Optimierung auszuarbeiten. Das Bestreben der Thesis ist darauf ausgerichtet, die Beschreibung und Analyse der ausgewählten Verfahren detailliert und verständlich zu instruieren. Für die technische Umsetzung und Realisierung wurde hier mit *Mathematica*, einer der bekanntesten kommerziellen Softwarepakete für Mathematiker, gearbeitet.

## 1.3 Aufbau der Arbeit

Einleitend werden in dieser Arbeit die Grundlagen der mathematischen Optimierung vorgestellt. In den zwei nachfolgenden Kapiteln werden numerische Verfahren erörtert, welche den Hauptteil dieser Arbeit umfassen. In Kapitel 3 wird das Gradientenverfahren und die Armijo-Schrittweitenregel detailliert vorgestellt. Im Anschluss daran wird in Kapitel 4 das Trust-Region-Verfahren untersucht. Zum besseren Verständnis werden zu den jeweiligen Verfahren innerhalb der Kapitel Beispiele angeführt. Die in Kapitel 3 und 4 vorgestellten Verfahren werden in Kapitel 5 anhand der Rosenbrock-Funktion ausgewertet. Schließlich werden die Aspekte der Konvergenzgeschwindigkeit der implementierten Verfahren analysiert und diskutiert. Die Auswertungen erlauben Rückschlüsse über die Stärken und Schwächen der jeweiligen Verfahren beziehungsweise Vorgehensweisen zu ziehen. Im Anhang A befindet sich die kommentierte Implementierung der beiden Verfahren und deren Ergebnisse für das Testbeispiel der Rosenbrock-Funktion.

## 2 Mathematische Grundlagen

---

In diesem einleitenden Kapitel werden die grundlegenden Begriffe, Bezeichnungen und Notationen der Mathematik eingeführt. Der Schwerpunkt liegt dabei auf der Optimalitätsbedingung, die in den folgenden Kapiteln Anwendung finden wird.

### 2.1 Mathematische Notationen

Vektoren  $x \in \mathbb{R}^n$  sind grundsätzlich als Spaltenvektor zu verstehen und  $x^T$  ist der durch Transposition entstehende Zeilenvektor.

Eine **Norm** [GK99] ist eine Abbildung  $\|\cdot\|$  von einem Vektorraum  $V$  über dem Körper  $\mathbb{K}$  der reellen oder der komplexen Zahlen in die Menge der nichtnegativen reellen Zahlen  $\mathbb{R}_0^+$ ,

$$\|\cdot\| : V \rightarrow \mathbb{R}_0^+, x \mapsto \|x\|,$$

die für alle Vektoren  $x, y \in V$  und alle Skalare  $\alpha \in \mathbb{K}$  die folgenden drei Axiome erfüllt:

- 1.) Definitheit  $\|x\| = 0 \Rightarrow x = 0$
- 2.) Homogenität  $\|\alpha x\| = |\alpha| \cdot \|x\|$
- 3.) Dreiecksungleichung  $\|x + y\| \leq \|x\| + \|y\|$

Die wichtigsten Beispiele von Normen im  $\mathbb{R}^n$  sind:

- $l_1$ - oder Summennorm  $\|x\|_1 = \sum_{i=1}^n |x_i|$
- $l_2$ - oder euklidische Norm  $\|x\|_2 = \sqrt{x^T x} = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$
- $l_\infty$ - oder Maximumnorm  $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$

Von besonderer Bedeutung ist häufig die durch die euklidische Vektornorm induzierte Matrixnorm

$$\|A\|_2 := \max_{\|x\|_2=1} \|Ax\|_2$$

diese wird üblicherweise als *Spektralnorm* bezeichnet und lässt sich durch

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$$

charakterisieren, wobei  $\lambda_{\max}(A^T A)$  der größte Eigenwert der symmetrischen und positiv semidefiniten Matrix  $A^T A$  ist. Für eine symmetrische Matrix  $A \in \mathbb{R}^{n \times n}$  ergibt sich hieraus insbesondere

$$\|A\|_2 = |\lambda_{\max}(A)|.$$

#### Die Kondition einer Matrix [GK99]

Ist  $A \in \mathbb{R}^{n \times n}$  eine reguläre Matrix und  $\|\cdot\|_2$  die Spektralnorm im  $\mathbb{R}^{n \times n}$ , so wird

$$Kond(A) = \kappa(A) := \|A\|_2 \|A^{-1}\|_2$$

als die *Spektral-Kondition* der Matrix bezeichnet.

Ist  $A$  überdies symmetrisch, so ergibt sich offenbar die Darstellung

$$\kappa(A) = \lambda_{\max}(A) \cdot \lambda_{\min}(A^{-1}) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

wobei  $\lambda_{\max}(A)$  und  $\lambda_{\min}(A)$  wieder den größten beziehungsweise kleinsten Eigenwert von  $A$  bezeichnen.

Sei  $\varepsilon > 0$  und  $\bar{x} \in \mathbb{R}^n$ . Die Menge

$$B_\varepsilon(\bar{x}) = \{x \in \mathbb{R}^n; \|x - \bar{x}\| < \varepsilon\}$$

heißt **offene  $\varepsilon$ -Kugel** um  $\bar{x}$ .

Ist  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  stetig differenzierbar, dann bezeichnet

$$\nabla f(x) = \left( \frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)^T \in \mathbb{R}^n$$

den **Gradienten** von  $f$  im Punkt  $x$ . Dabei ist der Gradient ein Spaltenvektor.

Ist  $f$  zweimal stetig differenzierbar, dann bezeichnet

$$\nabla^2 f(x) = \left( \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{i,j=1,\dots,n} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{pmatrix} \in \mathbb{R}^{n \times n}$$

die **Hesse-Matrix** von  $f$  im Punkt  $x$ . Diese ist symmetrisch, da die Stetigkeit von  $\nabla^2 f$  vorausgesetzt wurde.

Eine  $n \times n$ -Matrix  $A$  mit Elementen aus dem Körper  $\mathbb{K}$  heißt invertierbar oder umkehrbar, falls es eine **inverse Matrix**  $A^{-1}$  gibt mit  $AA^{-1} = E_n = A^{-1}A$ , wobei  $E_n$  die Einheitsmatrix der Größe  $n \times n$  darstellt [MM13].

Die Inverse einer Matrix kann für den allgemeinen Fall mit Determinantenformel wie folgt berechnet werden:

$$A^{-1} = \frac{1}{\det A} (A_{adj})^T$$

dabei ist  $A_{adj} = ((-1)^{i+j} \det A_{ij})$  und  $(-1)^{i+j} \det A_{ij}$  das algebraische Komplement von  $a_{ij}$ : Die Matrix  $A_{ij}$  entsteht aus  $A = (a_{ij})$  durch Streichen der  $i$ -ten Zeile und der  $j$ -ten Spalte. Für  $2 \times 2$ -Matrizen ergibt sich damit die explizite Formel:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad A^{-1} = \frac{1}{\det A} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Für die Bewertung von iterativen Verfahren ist die Geschwindigkeit, mit der eine Iterationsfolge  $\{x^k\}_{k \in \mathbb{N}} \subset \mathbb{R}^n$  gegen eine (lokale oder globale) Lösung  $\bar{x} \in \mathbb{R}^n$  des Optimierungsproblems konvergiert, ein wichtiges Kriterium [RH13]. Daher wird nun im Folgenden der Konvergenzbegriff definiert:

#### Definition 2.1.1 $q$ -Konvergenzgeschwindigkeit

Es seien  $\{x^k\}_{k \in \mathbb{N}} \subset \mathbb{R}^n$  und  $\lim_{k \rightarrow \infty} x^k = \bar{x}$ . Die Folge  $\{x^k\}_{k \in \mathbb{N}}$  konvergiert gegen  $\bar{x}$

- $q$ -linear mit dem Konvergenzfaktor  $C$ , wenn ein  $C \in (0, 1)$  und ein  $k_0 \in \mathbb{N}$  existieren, sodass

$$\|x^{k+1} - \bar{x}\| \leq C \|x^k - \bar{x}\|$$

- *q-superlinear*, wenn eine positive Nullfolge  $\{c_k\}_{k \in \mathbb{N}}$  und ein  $k_0 \in \mathbb{N}$  existieren, sodass

$$\|x^{k+1} - \bar{x}\| \leq c_k \|x^k - \bar{x}\|$$

- *q-quadratisch*, wenn ein  $C > 0$  und ein  $k_0 \in \mathbb{N}$  existieren, sodass

$$\|x^{k+1} - \bar{x}\| \leq C \|x^k - \bar{x}\|^2$$

für alle  $k \in \mathbb{N}$  mit  $k \geq k_0$  gilt.

Die aufgeführten Definitionen zur q-Konvergenzgeschwindigkeit basieren auf dem Quotientenkriterium zur absoluten Konvergenz von Reihen. Analog gibt es noch Definitionen zur r-Konvergenzgeschwindigkeit, die sich auf das Wurzelkriterium beziehen, die jedoch hier nicht weiter ausgeführt werden.

## 2.2 Optimalitätsbedingungen

Der folgende Satz gibt notwendige Bedingungen erster Ordnung für das Vorliegen eines lokalen Minimums von  $f$  an:

**Satz 2.2.1 Notwendige Optimalitätsbedingung erster Ordnung.**

Sei  $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$  differenzierbar auf der offenen Menge  $U \subset \mathbb{R}^n$  und  $\bar{x} \in U$  ein lokales Minimum von  $f$ . Dann gilt  $\nabla f(\bar{x}) = 0$ .

*Beweis* Dies ist aus der Analysis bekannt. Der Nachweis kann durch Betrachten des Differenzenquotienten  $[f(\bar{x} + td) - f(\bar{x})]/t$  erfolgen. Für beliebiges  $d \in \mathbb{R}^n$  und hinreichend kleine  $t > 0$  ist dieser nichtnegativ. Grenzwertbildung  $t \rightarrow 0^+$  ergibt  $\nabla f(\bar{x})^T d \geq 0$  und die Wahl  $d = -\nabla f(\bar{x})$  liefert  $\nabla f(\bar{x}) = 0$ .

□

Dieser Optimalitätsbedingung wird nun ein Name zugewiesen.

**Definition 2.2.1** Sei  $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$  differenzierbar in einer Umgebung von  $\bar{x} \in U$ . Der Punkt  $\bar{x}$  heißt stationärer Punkt von  $f$ , falls  $\nabla f(\bar{x}) = 0$  gilt.

Die Stationaritätsbedingung ist notwendig, aber nicht hinreichend für ein lokales Minimum. Denn wegen  $\nabla(-f) = -\nabla f$  ist jeder stationäre Punkt von  $f$  auch ein stationärer Punkt von  $-f$ . Daher kann der Stationaritätsbegriff zwischen Maxima und Minima nicht unterscheiden. Ein stationärer Punkt kann auch weder Minimum noch Maximum sein:

**Definition 2.2.2** Ein stationärer Punkt  $\bar{x}$  von  $f$ , der weder lokales Minimum noch lokales Maximum ist, heißt Sattelpunkt.

Um zwischen Minima, Maxima und Sattelpunkten unterscheiden zu können, muss das Krümmungsverhalten der Funktion betrachtet werden:

**Satz 2.2.2 Notwendige Optimalitätsbedingungen zweiter Ordnung.**

Sei  $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$  zweimal stetig differenzierbar auf der offenen Menge  $U \subset \mathbb{R}^n$  und  $\bar{x} \in U$  ein lokales Minimum von  $f$ . Dann gilt:

(i)  $\nabla f(\bar{x}) = 0$  (das heißt,  $\bar{x}$  ist stationärer Punkt von  $f$ )

(ii) Die Hesse-Matrix  $\nabla^2 f(\bar{x})$  ist positiv semidefinit:

$$d^T \nabla^2 f(\bar{x}) d \geq 0 \quad \forall d \in \mathbb{R}^n.$$

*Beweis* Bedingung (i) wurde bereits in Satz 2.2.1 nachgewiesen. Zum Nachweis von (ii) sei nun  $d \in \mathbb{R}^n \setminus \{0\}$  beliebig. Wird  $\tau = \tau(d) > 0$  hinreichend klein gewählt, so liefert Taylor-Entwicklung für alle  $t \in (0, \tau]$ :

$$0 \leq f(\bar{x} + td) - f(\bar{x}) = t\nabla f(\bar{x})^T d + \frac{t^2}{2} d^T \nabla^2 f(\bar{x}) d + \rho(t) \quad \text{mit} \quad \rho(t) = o(t^2),$$

wobei in der ersten Ungleichung verwendet wurde, dass  $\bar{x}$  lokales Minimum von  $f$  ist. Dies liefert wegen (i):

$$d^T \nabla^2 f(\bar{x}) d \geq -2 \frac{\rho(t)}{t^2}.$$

Die rechte Seite strebt für  $t \rightarrow 0$  gegen 0. Daraus folgt die Behauptung.  $\square$

Die Bedingungen (i) und (ii) aus Satz 2.2.2 sind notwendig, aber nicht hinreichend, wie der Sattelpunkt  $\bar{x} = 0$  von  $f(x) = x^3$  zeigt. Die notwendigen Bedingungen aus Satz 2.2.2 werden durch eine Verschärfung von (ii) hinreichend gemacht:

**Satz 2.2.3 Hinreichende Optimalitätsbedingungen zweiter Ordnung.**

Sei  $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$  zweimal stetig differenzierbar auf der offenen Menge  $U \subset \mathbb{R}^n$  und  $\bar{x} \in U$  ein Punkt, in dem gilt:

- (i)  $\nabla f(\bar{x}) = 0$  (das heißt,  $\bar{x}$  ist stationärer Punkt von  $f$ ),
- (ii) Die Hesse-Matrix  $\nabla^2 f(\bar{x})$  ist positiv definit:

$$d^T \nabla^2 f(\bar{x}) d > 0 \quad \forall d \in \mathbb{R}^n \setminus \{0\}.$$

Dann ist  $\bar{x}$  ein striktes lokales Minimum von  $f$ .

*Beweis* Seien (i) und (ii) erfüllt. Dann gibt es wegen (ii)  $\mu > 0$  mit

$$d^T \nabla^2 f(\bar{x}) d \geq \mu d^T d = \mu \|d\|^2 \quad \forall d \in \mathbb{R}^n,$$

dabei ist  $\mu$  zum Beispiel der kleinste Eigenwert von  $\nabla^2 f(\bar{x})$ . Durch Taylor-Entwicklung und das Ausnutzen der Stationarität können ein  $\varepsilon > 0$  gefunden werden, dass für alle  $d \in B_\varepsilon(0)$  gilt:

$$\begin{aligned} f(\bar{x} + d) &= f(\bar{x}) + \nabla f(\bar{x}) d + \frac{1}{2} d^T \nabla^2 f(\bar{x}) d + o(\|d\|^2) \\ f(\bar{x} + d) &= f(\bar{x}) + \frac{1}{2} d^T \nabla^2 f(\bar{x}) d + o(\|d\|^2) \\ \Rightarrow f(\bar{x} + d) - f(\bar{x}) &= \frac{1}{2} d^T \nabla^2 f(\bar{x}) d + o(\|d\|^2) \geq \frac{\mu}{2} \|d\|^2 + o(\|d\|^2) \geq \frac{\mu}{4} \|d\|^2. \end{aligned}$$

Somit ist  $\bar{x}$  striktes lokales Minimum wie behauptet.  $\square$

Es werden nun die fünf möglichen Definitheits-Fälle, die eine Matrix haben kann, vorgestellt.

**Definition 2.2.3** Sei  $A \in \mathbb{R}^{n \times n}$  symmetrische Matrix, das heißt  $A$  hat reelle Eigenwerte.

- (i)  $A$  heißt positiv definit, falls  $d^T A d > 0 \quad \forall d \in \mathbb{R}^n \setminus \{0\}$ .
- (ii)  $A$  heißt positiv semidefinit, falls  $d^T A d \geq 0 \quad \forall d \in \mathbb{R}^n$ .
- (iii)  $A$  heißt negativ definit, falls  $d^T A d < 0 \quad \forall d \in \mathbb{R}^n \setminus \{0\}$ .
- (iv)  $A$  heißt negativ semidefinit, falls  $d^T A d \leq 0 \quad \forall d \in \mathbb{R}^n \setminus \{0\}$ .
- (v)  $A$  heißt indefinit, falls es  $d \in \mathbb{R}^n$  und  $w \in \mathbb{R}^n$  gibt mit  $d^T A d > 0$  und  $w^T A w < 0$ .

Mit Hilfe der Definitheit der Hesse-Matrix kann bei den stationären Punkten einer zweimal stetig differenzierbaren Funktion bestimmt werden, welche davon Maxima, Minima oder Sattelpunkte sind.

**Satz 2.2.4** Sei  $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$  zweimal stetig differenzierbar auf der offenen Menge  $U \subset \mathbb{R}^n$  und  $\nabla f(\bar{x}) = 0$  für ein  $\bar{x} \in U$ , dann gilt

- (i) Die Hesse-Matrix  $\nabla^2 f(\bar{x})$  ist positiv definit  $\Rightarrow$   $f$  hat in  $\bar{x}$  ein lokales Minimum.
- (ii) Die Hesse-Matrix  $\nabla^2 f(\bar{x})$  ist negativ definit  $\Rightarrow$   $f$  hat in  $\bar{x}$  ein lokales Maximum.
- (iii) Die Hesse-Matrix  $\nabla^2 f(\bar{x})$  ist indefinit  $\Rightarrow$   $f$  hat in  $\bar{x}$  einen Sattelpunkt.

Es werden in Abbildung 2.1 die drei möglichen Fälle (i)-(iii) aus Satz 2.2.4 im  $\mathbb{R}^2$  veranschaulicht.

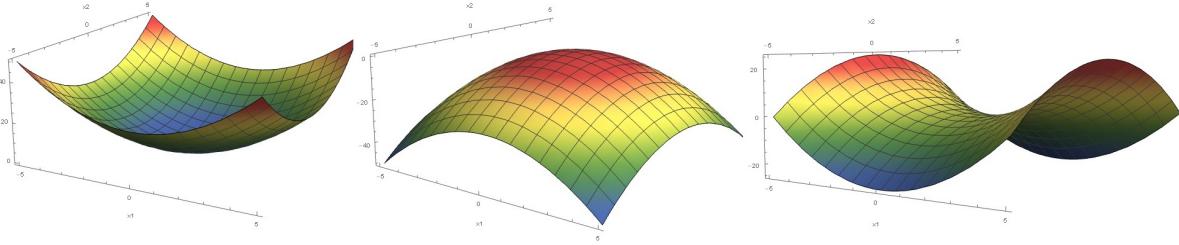


Abbildung 2.1: Die Darstellung der stationären Punkte je nach Definitheit der Hesse-Matrix

**Beispiel 2.2.1** Die Funktion  $f(x_1, x_2) = -\frac{1}{x_1} + \frac{1}{x_2} - 4x_1 + x_2$  hat den Gradienten  $\nabla f(x_1, x_2) = (\frac{1}{x_1^2} - 4; -\frac{1}{x_2^2} + 1)^T$ . Die notwendige Optimalitätsbedingung lautet:

$$\nabla f(x_1; x_2) \stackrel{!}{=} 0 \Rightarrow \text{(I)} \frac{1}{x_1^2} - 4 = 0 \text{ und } \text{(II)} -\frac{1}{x_2^2} + 1 = 0$$

Aus den beiden Gleichungen folgt

$$\frac{1}{x_1^2} - 4 = 0 \Leftrightarrow \frac{1}{x_1^2} = 4 \Leftrightarrow x_1^2 = \frac{1}{4} \Leftrightarrow x_1 = \pm \sqrt{\frac{1}{4}} = \pm \frac{1}{2}$$

und

$$-\frac{1}{x_2^2} + 1 = 0 \Leftrightarrow -\frac{1}{x_2^2} = -1 \Leftrightarrow \frac{1}{x_2^2} = 1 \Leftrightarrow x_2^2 = 1 \Leftrightarrow x_2 = \pm 1.$$

Es existieren also vier Kandidaten  $(\vec{x}_1; \vec{x}_2)_1^T = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$ ,  $(\vec{x}_1; \vec{x}_2)_2^T = \begin{pmatrix} \frac{1}{2} \\ -1 \end{pmatrix}$ ,  $(\vec{x}_1; \vec{x}_2)_3^T = \begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix}$

und  $(\vec{x}_1; \vec{x}_2)_4^T = \begin{pmatrix} -\frac{1}{2} \\ -1 \end{pmatrix}$  für Extremstellen.

Um die hinreichende Bedingung zu prüfen, wird zusätzlich die Hesse-Matrix benötigt. Mit

$$\frac{\partial^2 f}{\partial x_1 \partial x_1}(x) = -\frac{2}{x_1^3}, \quad \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) = 0 \text{ und } \frac{\partial^2 f}{\partial x_2 \partial x_2}(x) = \frac{2}{x_2^3}$$

gilt  $\nabla^2 f(x_1; x_2) = \begin{pmatrix} -\frac{2}{x_1^3} & 0 \\ 0 & \frac{2}{x_2^3} \end{pmatrix}$ . Nun wird die Hesse-Matrix an den vier Stellen geprüft, ob diese positiv, negativ oder indefinit ist.

- Mit  $(\vec{x}_1; \vec{x}_2)_1^T = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$  ist die Hesse-Matrix  $\nabla^2 f(\frac{1}{2}; 1) = \begin{pmatrix} -\frac{2}{\frac{1}{2}^3} & 0 \\ 0 & \frac{2}{1^3} \end{pmatrix} = \begin{pmatrix} -16 & 0 \\ 0 & 2 \end{pmatrix}$  indefinit und die Funktion  $f$  hat in  $(\vec{x}_1; \vec{x}_2)_1^T = (\frac{1}{2}; 1)^T$  einen Sattelpunkt.

- Mit  $(\vec{x}_1; \vec{x}_2)_2^T = \begin{pmatrix} \frac{1}{2} \\ -1 \end{pmatrix}$  ist die Hesse-Matrix  $\nabla^2 f(\frac{1}{2}; -1) = \begin{pmatrix} -\frac{2}{\frac{1}{2}^3} & 0 \\ 0 & \frac{2}{(-1)^3} \end{pmatrix} = \begin{pmatrix} -16 & 0 \\ 0 & -2 \end{pmatrix}$  negativ definit und die Funktion  $f$  hat in  $(\vec{x}_1; \vec{x}_2)_2^T = (\frac{1}{2}; -1)^T$  ein Maximum.
- Mit  $(\vec{x}_1; \vec{x}_2)_3^T = \begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix}$  ist die Hesse-Matrix  $\nabla^2 f(-\frac{1}{2}; 1) = \begin{pmatrix} -\frac{2}{(-\frac{1}{2})^3} & 0 \\ 0 & \frac{2}{1^3} \end{pmatrix} = \begin{pmatrix} 16 & 0 \\ 0 & 2 \end{pmatrix}$  positiv definit und die Funktion  $f$  hat in  $(\vec{x}_1; \vec{x}_2)_3^T = (-\frac{1}{2}; 1)^T$  ein Minimum.
- Mit  $(\vec{x}_1; \vec{x}_2)_4^T = \begin{pmatrix} -\frac{1}{2} \\ -1 \end{pmatrix}$  ist die Hesse-Matrix  $\nabla^2 f(-\frac{1}{2}; -1) = \begin{pmatrix} -\frac{2}{(-\frac{1}{2})^3} & 0 \\ 0 & \frac{2}{(-1)^3} \end{pmatrix} = \begin{pmatrix} 16 & 0 \\ 0 & -2 \end{pmatrix}$  indefinit und die Funktion  $f$  hat in  $(\vec{x}_1; \vec{x}_2)_4^T = (-\frac{1}{2}; -1)^T$  einen Sattelpunkt.

## 2.3 Konvexität

Es wird nun eine wichtige Klasse von Funktionen betrachtet, deren lokale Minima stets auch globale Minima sind: die konvexen Funktionen.

**Definition 2.3.1** Die Menge  $X \subset \mathbb{R}^n$  heißt konvex, falls für alle  $x, y \in X$  und alle  $\lambda \in [0, 1]$  gilt:

$$(1 - \lambda)x + \lambda y \in X.$$

Das bedeutet also, liegen  $x$  und  $y$  in  $X$ , so muss auch ihre Verbindungsstrecke in  $X$  liegen.

In der folgenden Abbildung 2.2 wird eine konvexe und eine nicht konvexe Teilmenge des  $\mathbb{R}^2$  dargestellt.

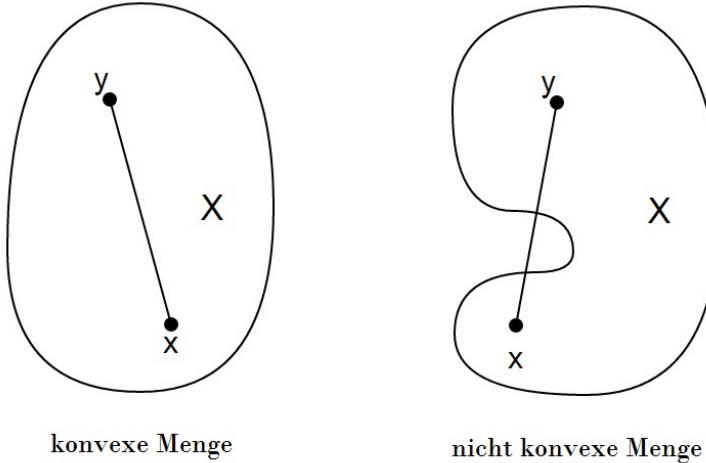


Abbildung 2.2: Veranschaulichung der konvexen und nicht konvexen Menge

**Definition 2.3.2** Sei  $f : X \rightarrow \mathbb{R}$  eine auf einer konvexen Menge  $X \subset \mathbb{R}^n$  definierte Funktion. Dann heißt  $f$

- konvex, falls für alle  $x, y \in X$  und alle  $\lambda \in [0, 1]$  gilt:

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y).$$

- streng (oder strikt) konvex, falls für alle  $x, y \in X$  mit  $x \neq y$  und alle  $\lambda \in (0, 1)$  gilt:

$$f((1 - \lambda)x + \lambda y) < (1 - \lambda)f(x) + \lambda f(y).$$

- gleichmäßig konvex, falls es  $\mu > 0$  gibt, so dass für alle  $x, y \in X$  und alle  $\lambda \in [0, 1]$  gilt:

$$f((1 - \lambda)x + \lambda y) + \mu \lambda(1 - \lambda) \|y - x\|^2 \leq (1 - \lambda)f(x) + \lambda f(y).$$

Die Unterschiede zwischen konvexen und nicht konvexen Funktionen ist in Abbildung 2.3 illustriert.

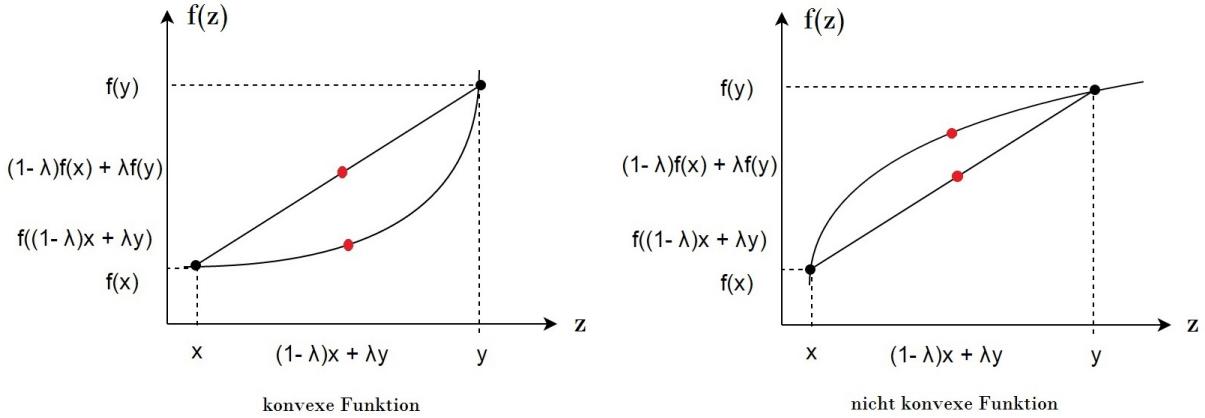


Abbildung 2.3: Veranschaulichung der konvexen und nicht konvexen Funktion

Ist  $f$  differenzierbar, so kann die Konvexität auch auf eine andere Weise charakterisiert werden:

**Satz 2.3.1** Sei  $f : X \rightarrow \mathbb{R}$  stetig differenzierbar auf einer offenen Umgebung der konvexen Menge  $X \subset \mathbb{R}^n$ . Dann gilt:

1. Die Funktion  $f$  ist konvex genau dann, wenn für alle  $x, y \in X$  gilt:

$$\nabla f(x)^T(y - x) \leq f(y) - f(x). \quad (2.1)$$

2. Die Funktion  $f$  ist strikt konvex genau dann, wenn für alle  $x, y \in X$  mit  $x \neq y$  gilt:

$$\nabla f(x)^T(y - x) < f(y) - f(x).$$

3. Die Funktion  $f$  ist gleichmäßig konvex genau dann, wenn es  $\mu > 0$  gibt, so dass für alle  $x, y \in X$  gilt:

$$\nabla f(x)^T(y - x) + \mu \|y - x\|^2 \leq f(y) - f(x).$$

Beweis zu 1:

" $\Rightarrow$ ": Sei  $f$  konvex. Dann gilt für beliebige  $x, y \in X$  und alle  $0 < \lambda \leq 1$

$$\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq \frac{(1 - \lambda)f(x) + \lambda f(y) - f(x)}{\lambda} = f(y) - f(x).$$

Übergang zum Limes  $\lambda \rightarrow 0^+$  liefert nun (2.1), denn

$$\nabla f(x)^T(y - x) = \lim_{\lambda \rightarrow 0^+} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda}.$$

" $\Leftarrow$ ": Es gelte (2.1). Für beliebige  $x, y \in X$  und  $0 \leq \lambda \leq 1$ . Es wird nun die verkürzte Schreibweise  $x_\lambda = (1 - \lambda)x + \lambda y$  verwendet. Es muss gezeigt werden, dass gilt

$$(1 - \lambda)f(x) + \lambda f(y) - f(x_\lambda) \geq 0.$$

Um dies nachzuweisen, wird für die folgende Berechnung (2.1) verwendet

$$\begin{aligned} (1 - \lambda)f(x) + \lambda f(y) - f(x_\lambda) &= (1 - \lambda)(f(x) - f(x_\lambda)) + \lambda(f(y) - f(x_\lambda)) \\ &\geq (1 - \lambda)\nabla f(x_\lambda)^T(x - x_\lambda) + \lambda\nabla f(x_\lambda)^T(y - x_\lambda) \\ &= \nabla f(x_\lambda)^T((1 - \lambda)x + \lambda y - x_\lambda) = 0. \end{aligned} \quad (2.2)$$

zu 2:

" $\Rightarrow$ ": Sei  $f$  streng konvex. Für  $x, y \in X, x \neq y$  und  $z = \frac{x+y}{2}$  gilt dann

$$f(z) < \frac{1}{2}(f(x) + f(y)), \quad \text{also} \quad f(z) - f(x) < \frac{1}{2}(f(y) - f(x)).$$

Weiter gilt wegen 1:

$$f(z) - f(x) \geq \nabla f(x)^T(z - x) = \frac{1}{2}\nabla f(x)^T(y - x).$$

Insgesamt liefert dies:

$$\nabla f(x)^T(y - x) \leq 2(f(z) - f(x)) < f(y) - f(x).$$

" $\Leftarrow$ ": Folgt direkt durch Verwenden von  $>$  in (2.2).

zu 3:

" $\Rightarrow$ ": Wie in 1. wird der Differenzenquotient betrachtet:

$$\begin{aligned} \nabla f(x)^T(y - x) &= \lim_{\lambda \rightarrow 0^+} \frac{f(x_\lambda) - f(x)}{\lambda} \\ &\leq \lim_{\lambda \rightarrow 0^+} \frac{(1 - \lambda)f(x) + \lambda f(y) - \mu\lambda(1 - \lambda)\|y - x\|^2 - f(x)}{\lambda} \\ &= f(y) - f(x) - \mu\|y - x\|^2. \end{aligned}$$

" $\Leftarrow$ ": Es werden die Eigenschaften der euklidischen Norm benutzt

$$\begin{aligned} \|x - x_\lambda\| &= \|x - ((1 - \lambda)x + \lambda y)\| = \|x - (x - \lambda x + \lambda y)\| \\ &= \|x - x + \lambda x - \lambda y\| = \|\lambda x - \lambda y\| = |\lambda| \|y - x\| = \lambda \|y - x\| \end{aligned}$$

und

$$\begin{aligned} \|y - x_\lambda\| &= \|y - ((1 - \lambda)x + \lambda y)\| = \|y - (x - \lambda x + \lambda y)\| \\ &= \|y - x + \lambda x - \lambda y\| = \|(1 - \lambda)y - (1 - \lambda)x\| = |1 - \lambda| \|y - x\| = (1 - \lambda) \|y - x\|. \end{aligned}$$

Ähnlich wie in (2.2) gilt nun

$$\begin{aligned} (1 - \lambda)f(x) + \lambda f(y) - f(x_\lambda) &= (1 - \lambda)(f(x) - f(x_\lambda)) + \lambda(f(y) - f(x_\lambda)) \\ &\geq (1 - \lambda)\nabla f(x_\lambda)^T(x - x_\lambda) + \mu\|x - x_\lambda\|^2 + \lambda\nabla f(x_\lambda)^T(y - x_\lambda) + \mu\|y - x_\lambda\|^2 \\ &= \nabla f(x_\lambda)^T((1 - \lambda)x + \lambda y - x_\lambda) + \mu(1 - \lambda)\|x - x_\lambda\|^2 + \lambda\|y - x_\lambda\|^2 \\ &= \mu((1 - \lambda)\lambda^2 + \lambda(1 - \lambda)^2)\|y - x\|^2 \\ &= \mu\lambda(1 - \lambda)\|y - x\|^2. \end{aligned}$$

□

Ist  $f$  zweimal stetig differenzierbar, so lässt sich die Konvexität von  $f$  in Verbindung bringen mit der Definitheit der Hesse-Matrix von  $f$ :

**Satz 2.3.2** Sei  $f : X \rightarrow \mathbb{R}$  zweimal stetig differenzierbar auf der offenen konvexen Menge  $X \subset \mathbb{R}$ . Dann gilt:

1. Die Funktion  $f$  ist konvex genau dann, wenn die Hesse-Matrix  $\nabla^2 f(x)$  für alle  $x \in X$  positiv semidefinit ist, das heißt, genau dann, wenn gilt

$$\forall x \in X, \forall d \in \mathbb{R}^n : \quad d^T \nabla^2 f(x) d \geq 0.$$

2. Die Funktion  $f$  ist strikt konvex, falls die Hesse-Matrix  $\nabla^2 f(x)$  für alle  $x \in X$  positiv definit ist, das heißt, falls gilt

$$\forall x \in X, \forall d \in \mathbb{R}^n \setminus \{0\} : \quad d^T \nabla^2 f(x) d > 0.$$

3. Die Funktion  $f$  ist genau dann gleichmäßig konvex, wenn die Hesse-Matrix  $\nabla^2 f(x)$  für alle  $x \in X$  gleichmäßig positiv definit ist, das heißt, genau dann, wenn es  $\mu > 0$  gibt, so dass gilt:

$$\forall x \in X, \forall d \in \mathbb{R}^n : \quad d^T \nabla^2 f(x) d \geq \mu \|d\|^2.$$

*Beweis* zu 1:

" $\Rightarrow$ ": Sei  $f$  konvex. Seien  $x \in X$  und  $d \in \mathbb{R}^n$  beliebig. Da  $X$  offen ist, gibt es  $\tau = \tau(x, d) > 0$  mit  $x + td \in X$  für alle  $t \in [0, \tau]$ . Für  $0 < t \leq \tau$  ergibt mit Satz 2.3.1, 1 und Taylor-Entwicklung:

$$0 \leq f(x + td) - f(x) - t\nabla f(x)^T d = \frac{t^2}{2} d^T \nabla^2 f(x) d + o(t^2).$$

Die Multiplikation mit  $\frac{2}{t^2}$  und Grenzübergang  $t \rightarrow 0^+$  liefert die Behauptung.

" $\Leftarrow$ ": Für beliebige  $x, y \in X$  liefert Taylor-Entwicklung ein  $\sigma \in [0, 1]$  mit

$$\begin{aligned} f(y) - f(x) &= \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x + \sigma(y - x)) (y - x) \\ &\geq \nabla f(x)^T (y - x). \end{aligned} \tag{2.3}$$

Nach Satz 2.3.1, 1 ist also  $f$  konvex.

zu 2:

Für  $x, y \in X$  mit  $x \neq y$  folgt die Ungleichung in (2.3) mit  $>$  statt  $\geq$ .

zu 3:

" $\Rightarrow$ ": Wie in 1. gibt für alle  $x \in X$  und  $d \in \mathbb{R}^n \setminus \{0\}$  ein  $\tau = \tau(x, d) > 0$ , so dass für alle  $0 < t \leq \tau$  gilt:

$$0 \leq f(x + td) - f(x) - t\nabla f(x)^T d - \mu \|td\|^2 = \frac{t^2}{2} d^T \nabla^2 f(x) d - t^2 \mu \|d\|^2 + o(t^2).$$

Multiplikation mit  $\frac{2}{t^2}$  und Grenzübergang  $t \rightarrow 0^+$  liefert

$$d^T \nabla^2 f(x) d \geq 2\mu \|d\|^2.$$

" $\Leftarrow$ ": Mit den Notationen von 1. resultiert

$$\begin{aligned} f(y) - f(x) &= \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x + \sigma(y - x)) (y - x) \\ &\geq \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2. \end{aligned}$$

Nach Satz 2.3.1, 3 ist also  $f$  gleichmäßig konvex.

□

Die große Bedeutung der Konvexität für die Optimierung besteht unter anderem in den folgenden Resultaten, die hier gleich für den restriktierten Fall bewiesen werden:

**Satz 2.3.3** *Sei  $f : X \rightarrow \mathbb{R}$  konvex auf der konvexen Menge  $X \subset \mathbb{R}^n$ . Dann gilt:*

1. *Jedes lokale Minimum von  $f$  auf  $X$  ist auch globales Minimum von  $f$  auf  $X$ .*
2. *Ist  $f$  sogar streng konvex, so besitzt  $f$  höchstens ein lokales Minimum auf  $X$  und dieses ist dann (falls es existiert) das strikte globale Minimum von  $f$  auf  $X$ .*
3. *Ist  $X$  offen,  $f$  stetig differenzierbar und  $\bar{x} \in X$  ein stationärer Punkt von  $f$ , so ist  $\bar{x}$  globales Minimum von  $f$  auf  $X$ .*

*Beweis* zu 1:

Sei  $\bar{x}$  ein lokales Minimum von  $f$  auf  $X$ . Angenommen, es gibt  $x \in X$  mit  $f(x) < f(\bar{x})$ . Dann gilt für alle  $t \in (0, 1]$ :

$$f(\bar{x} + t(x - \bar{x})) \leq (1 - t)f(\bar{x}) + tf(x) < (1 - t)f(\bar{x}) + tf(\bar{x}) = f(\bar{x}).$$

Dies ist ein Widerspruch zur lokalen Optimalität von  $\bar{x}$  auf  $X$ .

zu 2:

Angenommen,  $f$  besitzt zwei verschiedene lokale Minima  $\bar{x}$  und  $\bar{y}$  auf  $X$ . Nach 1. sind dann sowohl  $\bar{x}$ , als auch  $\bar{y}$  globale Minima von  $f$  auf  $X$ , das heißt es gilt

$$f(x) \geq f(\bar{x}) = f(\bar{y}) \forall x \in X.$$

Wegen der strikten Konvexität von  $f$  gilt dann speziell für  $x = \frac{\bar{x} + \bar{y}}{2} \in X$ :

$$f(x) < \frac{f(\bar{x}) + f(\bar{y})}{2} = f(\bar{x}) \leq f(x).$$

Dies ist ein Widerspruch.

Die Funktion  $f$  kann daher höchstens ein lokales Minimum auf  $X$  haben, und dieses ist dann nach 1. das eindeutige globale Minimum auf  $X$ .

zu 3:

Für alle  $x \in X$  gilt nach Satz 2.3.1, 1:

$$f(x) - f(\bar{x}) \geq \nabla f(\bar{x})^T (x - \bar{x}) = 0.$$

Somit ist  $\bar{x}$  globales Minimum von  $f$  auf  $X$ .

□

## 3 Das Gradientenverfahren

---

In diesem Kapitel wird das *Gradientenverfahren*, häufig auch als *Verfahren des steilsten Abstiegs* bezeichnet, vorgestellt. Dieses Verfahren wurde bereits 1847 von Cauchy [CA47] untersucht. Es bestimmt im Punkt  $x^k$  diejenige Richtung  $s^k$ , in der die Funktion  $f$  am stärksten abnimmt. In Abschnitt 3.1 wird zunächst der Modellalgorithmus für ein allgemeines Abstiegsverfahren vorgestellt. Im Anschluss daran wird in Abschnitt 3.2 die Abstiegsrichtung und in Abschnitt 3.3 die Schrittweite für das Gradientenverfahren untersucht. Schließlich wird in Abschnitt 3.4 die globale Konvergenz und in Abschnitt 3.5 die Konvergenzgeschwindigkeit des Gradientenverfahrens behandelt. Zum Ende des Kapitels wird das Gradientenverfahren anhand eines Beispiels ausführlich vorgeführt.

### 3.1 Allgemeines Abstiegsverfahren

Es wird das unrestringierte Minimierungsproblem

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

mit einer stetig differenzierbaren Zielfunktion  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  betrachtet. Die verschiedenen Abstiegsverfahren haben alle die folgende Struktur und unterscheiden sich nur durch die Wahl der Suchrichtung und der Schrittweite:

#### Algorithmus 3.1.1 Modellalgorithmus für ein Abstiegsverfahren.

0. Wähle einen Startpunkt  $x^0 \in \mathbb{R}^n$ .  
Für  $k = 0, 1, 2, \dots$ :
1. Prüfe auf Abbruch (meist: STOP, falls  $x^k$  stationär ist).
2. Berechne eine Abstiegsrichtung  $s^k \in \mathbb{R}^n$ , das heißt eine Richtung mit  $\nabla f(x^k)^T s^k < 0$ .
3. Bestimme eine Schrittweite  $\sigma_k > 0$ , so dass  $f(x^k + \sigma_k s^k) < f(x^k)$  gilt und die Abnahme der Zielfunktion, das heißt der Ausdruck  $f(x^k) - f(x^k + \sigma_k s^k)$ , hinreichend groß ist.
4. Setze  $x^{k+1} = x^k + \sigma_k s^k$ .

Die zentrale Idee ist hier die Verwendung von *Abstiegsrichtungen*. Der Vektor  $s \in \mathbb{R}^n \setminus \{0\}$  heißt Abstiegsrichtung der stetig differenzierbaren Funktion  $f$  im Punkt  $x$ , falls im Punkt  $x$  die Steigung von  $f$  in Richtung  $s$  negativ ist. Die Steigung von  $f$  in Richtung  $s$  ist gegeben durch den Differenzenquotienten

$$\lim_{t \rightarrow 0^+} \frac{f(x + ts) - f(x)}{\|ts\|} = \frac{\nabla f(x)^T s}{\|s\|}.$$

Damit ergibt sich die folgende Definition:

#### Definition 3.1.1 Abstiegsrichtung.

Der Vektor  $s \in \mathbb{R}^n \setminus \{0\}$  heißt Abstiegsrichtung der stetig differenzierbaren Funktion  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  im Punkt  $x$ , falls  $\nabla f(x)^T s < 0$  gilt.

## 3.2 Richtung des steilsten Abstiegs

Die naheliegendste Abstiegsrichtung ist wohl die Richtung des steilsten Abstiegs.

**Definition 3.2.1** Sei  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  stetig differenzierbar und  $x \in \mathbb{R}^n$  beliebig mit  $\nabla f(x) \neq 0$ . Weiter bezeichne  $d \in \mathbb{R}^n$  die Lösung des Problems

$$\min_{\|d\|=1} \nabla f(x)^T d. \quad (3.2)$$

Jeder Vektor der Form  $s = \lambda d, \lambda > 0$  heißt dann Richtung des steilsten Abstiegs von  $f$  in  $x$ .

Die Nebenbedingung ist erforderlich, damit das Problem eine Lösung besitzt, denn die Zielfunktion ist nicht nach unten beschränkt [AW02].

**Satz 3.2.1** Sei  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  stetig differenzierbar und  $x \in \mathbb{R}^n$  beliebig mit  $\nabla f(x) \neq 0$ . Dann besitzt das Problem (3.2) die eindeutige Lösung

$$d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

Insbesondere ist  $s$  eine Richtung des steilsten Abstiegs von  $f$  in  $x$  genau dann, wenn es  $\lambda > 0$  gibt mit  $s = -\lambda \nabla f(x)$ .

*Beweis* Die Cauchy-Schwarzsche Ungleichung besagt für beliebige  $v, w \in \mathbb{R}^n$ :

$$|v^T w| \leq \|v\| \|w\|$$

mit Gleichheit genau dann, wenn  $v$  und  $w$  linear abhängig sind. Für  $d \in \mathbb{R}^n$  mit  $\|d\| = 1$  gilt daher

$$\nabla f(x)^T d \geq -|\nabla f(x)^T d| \geq -\|\nabla f(x)\| \underbrace{\|d\|}_{=1} = -\|\nabla f(x)\|$$

mit Gleichheit genau dann, wenn  $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$  gilt. Da  $d$  eingesetzt ergibt genau

$$\nabla f(x)^T \cdot d = \nabla f(x)^T \cdot -\frac{\nabla f(x)}{\|\nabla f(x)\|} = -\frac{\|\nabla f(x)\|^2}{\|\nabla f(x)\|} = -\|\nabla f(x)\|.$$

Dies beweist die erste Aussage.

Die zweite Aussage folgt nun aus der Definition der Richtungen des steilsten Abstiegs.

□

An dieser Stelle sei vermerkt, dass die Aussage von Satz 3.2.1 nur für die Norm  $\|x\| = \sqrt{x^T x}$  richtig ist. Wird eine andere Norm in (3.2) gewählt, so ergibt sich eine andere Richtung des steilsten Abstiegs.

Bei der Normierung  $\|d\| = 1$  stellt sich zudem die Frage, ob die Lösung des Problems (3.2), also die Richtung  $-\nabla f(x)$ , von der gewählten Länge abhängt. Dazu wird das Problem

$$\min_{\|d\|=r} \nabla f(x)^T d \quad (3.3)$$

betrachtet. Analog zu Satz 3.2.1 kann gezeigt werden, dass für ein beliebiges  $r > 0$  der Vektor  $d_r = r \cdot d$  genau dann die Lösung des Minimierungsproblems (3.3) ist, wenn  $d$  die Lösung des Minimierungsproblems (3.2) ist [AW02].

**Beispiel 3.2.1** Für die Funktion  $f(x, y) = x^2 + y^2$  mit  $x^0 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$  soll nun  $d$  berechnet werden:

Es gilt  $\nabla f(x, y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}$  und somit

$$d = -\frac{\nabla f(x, y)}{\|\nabla f(x, y)\|} = -\frac{(2 \cdot 2; 2 \cdot 2)^T}{\|(2 \cdot 2; 2 \cdot 2)^T\|} = -\frac{(4; 4)^T}{\|(4; 4)^T\|} = -\frac{(4; 4)^T}{\sqrt{4^2 + 4^2}} = -\frac{(4; 4)^T}{\sqrt{32}} = -\frac{1}{\sqrt{32}} \begin{pmatrix} 4 \\ 4 \end{pmatrix}.$$

Beim Gradientenverfahren wird nun als Abstiegsrichtung der negative Gradient  $s^k = -\nabla f(x^k)$  verwendet, welcher auch als die Richtung des steilsten Abstiegs bezeichnet wird.

### 3.3 Die Armijo-Schrittweitenregel

Die Wahl der Suchrichtung wird nun durch eine geeignete Schrittweite  $\sigma_k$  ergänzt. Die folgende Strategie zur Bestimmung der Schrittweite, die sogenannte *Armijo-Regel*, ist einfach zu implementieren und kann für beliebige Abstiegsrichtungen  $s^k$  angewendet werden.

Bei einer linearen Approximation wird erwartet, dass für  $\sigma_k > 0$  und  $s^k \in \mathbb{R}^n$  gilt [HB15]

$$f(x^{k+1}) = f(x^k + \sigma_k s^k) \approx f(x^k) + \sigma_k \nabla f(x^k)^T s^k$$

Die Idee der Armijo-Schrittweitenregel ist es, die tatsächliche Abnahme der Zielfunktion  $f(x^k + \sigma_k s^k) - f(x^k)$  mit der aus der linearen Approximation erwarteten Abnahme  $\sigma_k \nabla f(x^k)^T s^k$  zu vergleichen. Erst wenn die tatsächliche Abnahme einen vorgegebenen Bruchteil (zum Beispiel  $\gamma = 10^{-2}$ ) der erwarteten Abnahme erreicht, wird der Schritt durchgeführt. Ist dies aber nicht der Fall, so wird die Schrittweite als zu groß eingeschätzt und muss verkürzt, zum Beispiel halbiert, werden [HB15].

#### Armijo-Schrittweitenregel:

Seien  $\beta \in (0, 1)$  (z.B.  $\beta = \frac{1}{2}$ ) und  $\gamma \in (0, 1)$  (z.B.  $\gamma = 10^{-2}$ ) fest gewählte Parameter. Bestimme die größte Zahl  $\sigma_k \in 1, \beta, \beta^2, \dots$ , für die gilt:

$$\underbrace{f(x^k + \sigma_k s^k) - f(x^k)}_{:= \phi_k(\sigma)} \leq \sigma_k \gamma \nabla f(x^k)^T s^k. \quad (3.4)$$

Anhand der folgenden Abbildung 3.1 wird die Armijo-Schrittweitenregel verdeutlicht. Hierbei erfüllen alle  $\sigma_k = \sigma > 0$  die Ungleichung (3.4), für die der Graph der Funktion  $\phi_k(\sigma) = f(x^k + \sigma s^k) - f(x^k)$  ( $--\cdots-$ ) auf oder unterhalb der Geraden  $\gamma \phi'_k(0)\sigma = \sigma \gamma \nabla f(x^k)^T s^k$  ( $\cdot\cdot\cdot\cdot\cdot$ ) liegt.

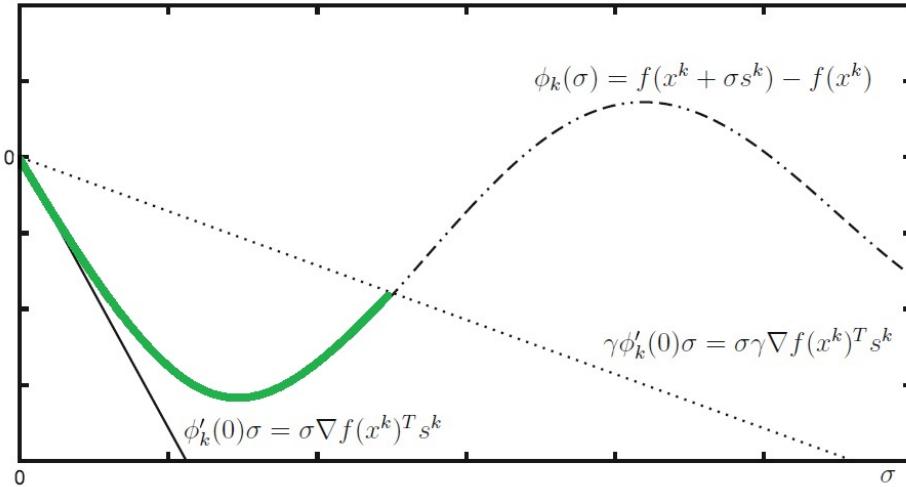


Abbildung 3.1: Die Armijo-Schrittweitenregel

Es muss sichergestellt werden, dass diese Schrittweitenregel immer durchführbar ist. Daher folgt auch:

**Lemma 3.3.1** Sei  $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$  stetig differenzierbar auf der offenen Menge  $U$ . Weiter sei  $\gamma \in (0, 1)$  gegeben. Ist nun  $x \in U$  ein Punkt und  $s \in \mathbb{R}^n$  eine Abstiegsrichtung von  $f$  in  $x$ , so gibt es ein  $\bar{\sigma} > 0$  mit

$$f(x + \sigma s) - f(x) \leq \sigma \gamma \nabla f(x)^T s \quad \forall \sigma \in [0, \bar{\sigma}]. \quad (3.5)$$

*Beweis* Für  $\sigma = 0$  ist die Ungleichung in (3.5) offensichtlich erfüllt.

Sei nun  $\sigma > 0$  hinreichend klein. Dann gilt  $x + \sigma s \in U$  und

$$\frac{f(x + \sigma s) - f(x)}{\sigma} - \gamma \nabla f(x)^T s \xrightarrow{\sigma \rightarrow 0^+} \nabla f(x)^T s - \gamma \nabla f(x)^T s = (1 - \gamma) \nabla f(x)^T s < 0.$$

Daher kann  $\bar{\sigma} > 0$  so klein gewählt werden, dass dann Folgendes gilt:

$$\frac{f(x + \sigma s) - f(x)}{\sigma} - \gamma \nabla f(x)^T s \leq 0 \quad \forall \sigma \in (0, \bar{\sigma}].$$

Für dieses  $\bar{\sigma}$  ist dann (3.5) erfüllt.

□

## 3.4 Globale Konvergenz des Gradientenverfahrens

Mit den beschriebenen Festlegungen der Suchrichtungswahl und der Schrittweitenregel ergibt sich aus Algorithmus 3.1.1 das folgende Verfahren:

**Algorithmus 3.4.1 Gradientenverfahren, Verfahren des steilsten Abstiegs.**

0. Wähle  $\beta \in (0, 1)$ ,  $\gamma \in (0, 1)$  und einen Startpunkt  $x^0 \in \mathbb{R}^n$ .  
Für  $k = 0, 1, 2, \dots$ :
1. Falls  $\nabla f(x^k) = 0$ , STOP.
2. Setze  $s^k = -\nabla f(x^k)$ .
3. Bestimme die Schrittweite  $\sigma_k > 0$  mithilfe der Armijo-Regel (3.4).
4. Setze  $x^{k+1} = x^k + \sigma_k s^k$ .

In der Praxis wird die Abbruchbedingung in Schritt 1 durch eine Bedingung der Form  $\|\nabla f(x^k)\| \leq \varepsilon$ , mit einer zu Beginn festgelegten Toleranz  $\varepsilon > 0$  (zum Beispiel  $\varepsilon = 10^{-8}$ ), ersetzt.

Der folgende Satz beweist unter Verwendung der Armijo-Schrittweitenregel, dass die Iterierten gegen einen stationären Punkt konvergieren, wenn diese eine konvergente Teilfolge besitzen.

**Satz 3.4.1** Sei  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  stetig differenzierbar. Dann terminiert Algorithmus 3.4.1 entweder endlich mit einem stationären Punkt  $x^k$ , oder er erzeugt eine unendliche Folge  $(x^k)$  mit folgenden Eigenschaften:

1. Für alle  $k$  gilt  $f(x^{k+1}) < f(x^k)$ .
2. Jeder Häufungspunkt von  $(x^k)$  ist ein stationärer Punkt von  $f$ .

*Beweis* Es wird nur der Fall betrachtet, in dem der Algorithmus nicht endlich abbricht. Aufgrund der Armijo-Schrittweitenregel gilt

$$\begin{aligned} f(x^k + \sigma_k s^k) - f(x^k) &\leq \sigma_k \gamma \nabla f(x^k)^T \underbrace{s^k}_{= -\nabla f(x^k)} \\ &\leq -\sigma_k \gamma \|\nabla f(x^k)\|^2 \end{aligned}$$

Nach Lemma 3.3.1 erzeugt das Verfahren dann unendliche Folgen  $(x^k)$  und  $(\sigma_k) \subset (0, 1]$  mit  $\nabla f(x^k) \neq 0$  und

$$f(x^{k+1}) - f(x^k) = f(x^k + \sigma_k s^k) - f(x^k) \leq -\sigma_k \gamma \|\nabla f(x^k)\|^2 < 0.$$

Daraus folgt 1.

zu 2: Sei  $\bar{x}$  ein Häufungspunkt von  $(x^k)$  und  $(x^k)_K$  eine Teilfolge mit  $(x^k)_K \rightarrow \bar{x}$ . Die Folge  $(f(x^k))$  ist monoton fallend und besitzt daher einen Grenzwert  $\varphi \in \mathbb{R} \cup \{-\infty\}$ . Daraus folgt insbesondere  $(f(x^k))_K \rightarrow \varphi$ . Wegen der Stetigkeit von  $f$  und  $(x^k)_K \rightarrow \bar{x}$  gilt aber auch  $(f(x^k))_K \rightarrow f(\bar{x})$ . Daher folgt  $\varphi = f(\bar{x})$  und

$$f(x^k) \rightarrow f(\bar{x}).$$

Wegen der Armijo-Regel

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\leq -\sigma_k \gamma \|\nabla f(x^k)\|^2 \\ f(x^k) - f(x^{k+1}) &\geq \sigma_k \gamma \|\nabla f(x^k)\|^2 \end{aligned}$$

und unter Verwendung der Teleskopsumme

$$\sum_{k=0}^n f(x^k) - f(x^{k+1}) = f(x^0) - f(x^n), \text{ für } n \rightarrow \infty \text{ gilt } f(x^n) = f(\bar{x})$$

gilt

$$f(x^0) - f(\bar{x}) = \sum_{k=0}^{\infty} (f(x^k) - f(x^{k+1})) \geq \gamma \sum_{k=0}^{\infty} \sigma_k \|\nabla f(x^k)\|^2.$$

Aus Analysis ist die notwendige Bedingung für Konvergenz von Reihen bekannt. Diese besagt, wenn die Reihe  $\sum_{k=m}^{\infty} a_k$  konvergent ist, dann ist die Folge  $(a_n)_{n \geq m}$  ihrer Glieder eine Nullfolge. Daher folgt auch

$$\sigma_k \|\nabla f(x^k)\|^2 \rightarrow 0. \quad (3.6)$$

Den Rest des Beweises wird per Widerspruch geführt. Angenommen, es gilt  $\nabla f(\bar{x}) \neq 0$ . Wegen der Stetigkeit von  $\nabla f$  und  $(x^k)_K \rightarrow \bar{x}$  gibt es dann  $l \in K$  mit

$$\|\nabla f(x^k)\| \geq \frac{\|\nabla f(\bar{x})\|}{2} > 0 \quad \forall k \in K, k \geq l.$$

Wegen (3.6) folgt dann

$$(\sigma_k)_K \rightarrow 0.$$

Insbesondere gibt es  $l' \in K$ ,  $l' \geq l$  mit  $\sigma_k \leq \beta$  für alle  $k \in K$ ,  $k \geq l'$ . Gemäß der Armijo-Schrittweitenregel (3.4) und der Schrittweite gilt dann

$$f(x^k + \beta^{-1} \sigma_k s^k) - f(x^k) > -\gamma \beta^{-1} \sigma_k \|\nabla f(x^k)\|^2 \quad \forall k \in K, k \geq l'. \quad (3.7)$$

Sei nun  $(t_k)_K = (\beta^{-1} \sigma_k)_K$ . Dann ist  $(t_k)_K$  eine Nullfolge. Nach dem Mittelwertsatz gibt es  $\tau_k \in [0, t_k]$  mit

$$\begin{aligned} \lim_{K \ni k \rightarrow \infty} \frac{f(x^k + t_k s^k) - f(x^k)}{t_k} &= \lim_{K \ni k \rightarrow \infty} \frac{t_k \nabla f(x^k + \tau_k s^k)^T s^k}{t_k} = -\|\nabla f(\bar{x})\|^2, \\ \lim_{K \ni k \rightarrow \infty} \|\nabla f(x^k)\|^2 &= \|\nabla f(\bar{x})\|^2. \end{aligned}$$

Daher ergibt sich aus (3.7) der Widerspruch

$$\begin{aligned} 0 &> -\|\nabla f(\bar{x})\|^2 \geq -\gamma \|\nabla f(\bar{x})\|^2 \\ 0 &> (\gamma - 1) \|\nabla f(\bar{x})\|^2 \geq 0 \\ 0 &< (1 - \gamma) \|\nabla f(\bar{x})\|^2 \leq 0. \end{aligned}$$

Somit war die Annahme  $\nabla f(\bar{x}) \neq 0$  falsch und der Beweis ist beendet.

□

Die Armijo-Regel kann auf vielfache Weise geeignet modifiziert werden, ohne dass der Konvergenzsatz seine Gültigkeit verliert. Wesentlich ist, dass im Fall  $(\sigma_k)_K \rightarrow 0$  eine Nullfolge  $(t_k)_K$  von Schrittweiten konstruiert werden kann, für die die Armijo-Bedingung verletzt ist. Deshalb wurde im Beweis die Schrittweite  $\sigma_k$  um den Faktor  $\beta$  verkleinert und als neue Schrittweite  $(t_k)_K$  definiert.

### 3.5 Konvergenzgeschwindigkeit des Gradientenverfahrens

Die Konvergenzgeschwindigkeit des Gradientenverfahrens ist im Allgemeinen sehr unbefriedigend. Im Folgenden wird dies stichhaltig begründet für den Fall, dass  $f$  streng konvex und quadratisch ist. Dabei ist diese Beispielfunktion wohl bedacht, da der Fall einer streng konvexen quadratischen Zielfunktion der schönste denkbare Fall eines unrestringierten Optimierungsproblems ist. Sei also  $f$  streng konvex und quadratisch:

$$f(x) = c_0 + c^T x + \frac{1}{2} x^T C x$$

mit  $c_0 \in \mathbb{R}$ ,  $c \in \mathbb{R}^n$  und  $C \in \mathbb{R}^{n \times n}$  symmetrisch und positiv definit. Es wird entlang der Suchrichtung das Minimum gesucht. Hierbei wird nun die *Minimierungsregel* als Schrittweiten-Regel verwendet:

$$\text{Bestimme } \sigma_k > 0 \text{ mit } f(x^k + \sigma_k s^k) = \min_{\sigma > 0} f(x^k + \sigma s^k).$$

Für die Funktion  $f(x(\sigma)) = f(x + \sigma s) = c_0 + c^T(x + \sigma s) + \frac{1}{2}(x + \sigma s)^T C(x + \sigma s)$  gilt mit Hilfe der mehrdimensionalen Kettenregel

$$\begin{aligned} \frac{d}{d\sigma} f(x(\sigma)) &= D_x f(x(\sigma)) \cdot \frac{d}{d\sigma} x(\sigma) \\ &= c^T \left( D_x(x + \sigma s) \cdot \frac{d}{d\sigma} x(\sigma) \right) + \frac{1}{2} \left( D_x((x + \sigma s)^T) \cdot \frac{d}{d\sigma} x(\sigma) \right) C(x + \sigma s) \\ &\quad + \frac{1}{2} (x + \sigma s)^T C \left( D_x(x + \sigma s) \cdot \frac{d}{d\sigma} x(\sigma) \right) \\ &= c^T \cdot 1 \cdot s + \frac{1}{2} ((1 \cdot s^T) \cdot C(x + \sigma s)) + \frac{1}{2} ((x + \sigma s)^T C \cdot (1 \cdot s)) \\ &= c^T s + \frac{1}{2} \underbrace{(s^T \cdot C(x + \sigma s))}_{(s^T C(x + \sigma s))^T = (x + \sigma s)^T C^T s} + \frac{1}{2} ((x + \sigma s)^T C \cdot s) \\ &= c^T s + \frac{1}{2} ((x + \sigma s)^T C^T s + (x + \sigma s)^T C s) \\ &= \left( c^T + \frac{1}{2} ((x + \sigma s)^T C^T + (x + \sigma s)^T C) \right) \cdot s \\ &\stackrel{\text{C symmetrisch}}{=} \left( c^T + \frac{1}{2} 2(x + \sigma s)^T C \right) \cdot s \\ &= (c^T + (x + \sigma s)^T C) \cdot s \\ &= (c + C(x + \sigma s))^T \cdot s \end{aligned}$$

wobei in der dritten und in der letzten Gleichung die Rechenregeln für das Transponieren einer Summe  $(A+B)^T = A^T + B^T$  beziehungsweise eines Produkts zweier Matrizen  $(AB)^T = B^T A^T$  verwendet wurde. Zudem gilt  $C^T = C$ , da  $C$  symmetrisch ist.

Für die Hesse-Matrix von  $f(x(\sigma)) = f(x + \sigma s)$  gilt

$$\begin{aligned} \frac{d^2}{d\sigma d\sigma} f(x(\sigma)) &= \frac{d}{d\sigma} (c + C(x + \sigma s))^T \cdot s \\ &= \frac{d}{d\sigma} (c^T + (x + \sigma s)^T C) \cdot s \\ &= s^T C s. \end{aligned}$$

Zusammenfassend folgt für die Iterationspunkte

$$\begin{aligned} \frac{d}{d\sigma_k} f(x^k(\sigma_k)) &= (c + C(x^k + \sigma_k s^k))^T \cdot s^k, \\ \frac{d^2}{d\sigma_k d\sigma_k} f(x^k(\sigma_k)) &= s^{k^T} C s^k > 0. \end{aligned}$$

Insbesondere ist  $f(x^k + \sigma_k s^k)$  streng konvex. Daher ist  $\sigma_k$  charakterisiert durch

$$\frac{d}{d\sigma_k} f(x^k(\sigma_k)) = (c + C(x^k + \sigma_k s^k))^T s^k \stackrel{!}{=} 0.$$

Daraus folgt zunächst

$$\begin{aligned} (c + Cx^k + C\sigma_k s^k)^T s^k &= 0 \\ C^T s^k + Cx^{kT} s^k + C\sigma_k s^{kT} s^k &= 0 \\ C\sigma_k s^{kT} s^k &= -c^T s^k - (Cx^k)^T s^k \\ \sigma_k &= \frac{-c^T s^k - (Cx^k)^T s^k}{s^{kT} C s^k} \end{aligned}$$

und somit

$$\sigma_k = -\frac{(c + Cx^k)^T s^k}{s^{kT} C s^k} = -\frac{\nabla f(x^k)^T s^k}{s^{kT} C s^k} = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^T C \nabla f(x^k)} = \frac{\|s^k\|^2}{s^{kT} C s^k}. \quad (3.8)$$

Es bietet sich also im Falle einer quadratischen und streng konvexen Zielfunktion an, die Armijo-Regel durch die optimale Schrittweite (3.8) zu ersetzen.

Es wird nun das Verhalten des Gradientenverfahrens veranschaulicht. Hierzu wird eine Iterierte  $x^k$  betrachtet, die nicht stationär ist. Die Suchrichtung ist dabei  $s^k = -\nabla f(x^k)$ . Ziel ist es, zu zeigen, dass der Gradient (und damit  $s^k$ ) senkrecht auf der Niveaumenge (im  $\mathbb{R}^2$ : Höhenlinie) durch  $x^k$  steht. Bezeichne  $L_k = \{x; f(x) = f(x^k)\}$  diese Niveaumenge. Da  $f$  stetig differenzierbar ist und  $\nabla f(x^k) \neq 0$  gilt, ist  $L_k$  eine stetig differenzierbare Hyperfläche (im  $\mathbb{R}^2$ : Kurve). Es wird nun eine beliebige  $C^1$ -Kurve  $\gamma: (-1, 1) \rightarrow L_k$  mit  $\gamma(0) = x^k$  betrachtet, die in der Niveaumenge verläuft. Somit wird die Niveaumenge  $L_k$  parametrisiert, beziehungsweise nur die Kurven betrachtet, die in  $L_k$  liegen. Im Folgenden wird gezeigt, dass der Gradient senkrecht zu allen Kurven ist, welche in  $L_k$  liegen, und somit auch senkrecht auf der Tangentialebene ist.

Differenzieren der Identität  $f(\gamma(t)) = f(x^k)$  ergibt mit der mehrdimensionalen Kettenregel

$$\frac{d}{dt} f((\gamma(t)) = D_{\gamma} f(\gamma(t)) \cdot \frac{d}{dt} \gamma(t)) = \nabla f(\gamma(t))^T \dot{\gamma}(t) = 0,$$

und somit auch

$$\nabla f(\gamma(0))^T \dot{\gamma}(0) = \nabla f(x^k)^T \dot{\gamma}(0) = 0.$$

Die Suchrichtung  $s^k = -\nabla f(x^k)$  steht also senkrecht auf  $L_k$ . Folglich wird im Punkt  $x_k$  senkrecht zur Höhenlinie  $L_k$  gestartet und dieser Richtung so lange bergab gefolgt, bis das globale Minimum  $\sigma_k > 0$  der Funktion  $f(x(\sigma)) := f(x^k + \sigma s^k)$  erreicht ist. Dort gilt

$$f'(x^k + \sigma_k s^k) = \nabla f(x^k + \sigma_k s^k)^T s^k = 0.$$

Im neuen Punkt  $x^{k+1} = x^k + \sigma_k s^k$  steht also  $s^k$  senkrecht auf  $\nabla f(x^{k+1})$ , der Normalen zur Niveaumenge  $L_{k+1}$ . Im Punkt  $x^{k+1}$  gilt daher  $s^k \parallel L_{k+1}$ . Die neue Richtung  $s^{k+1}$  ist dann wieder senkrecht zu  $L_{k+1}$  (und damit zu  $s^k$ ) und so weiter. Daher ist der durch das Verfahren erzeugte Polygonzug eine Zickzacklinie mit rechten Winkeln. Wenn das Gradientenverfahren in einem schmalen Tal verläuft, so führt die obige Überlegung dazu, dass die Iteration nur wenige Fortschritte machen wird (in Kapitel 5 wird dies ausführlich erläutert).

Des Weiteren ist die Funktion  $f$  streng konvex und die Hesse-Matrix  $\nabla^2 f(x) = C$  positiv definit. Somit folgt aus Satz 2.3.3 2. die Eigenschaft, dass die Funktion  $f$  höchstens ein lokales Minimum besitzt, welches dann gleichzeitig das strikte globale Minimum von  $f$  ist. Das global eindeutige Minimum  $\bar{x}$  von  $f$  errechnet sich durch folgende Umformung, die Dank der positiven Definitheit und somit auch der Invertierbarkeit von  $\nabla^2 f(x) = C$  erfolgt:

$$\begin{aligned} \nabla f(\bar{x}) &= c + C\bar{x} \stackrel{!}{=} 0 \\ C\bar{x} &= -c \\ \bar{x} &= -C^{-1}c \end{aligned}$$

**Beispiel 3.5.1** Gegeben sei  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x) = c_0 + c^T x + \frac{1}{2} x^T C x$  mit  $c_0 \in \mathbb{R}$ ,  $c \in \mathbb{R}^n$  und  $C \in \mathbb{R}^{n \times n}$  symmetrisch und positiv definit.

Dann gilt für  $c_0 = 1$ ,  $c^T = (1; 2)$  und  $C = \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}$ :

$$\begin{aligned} f(x) &= 1 + (1; 2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2} (x_1; x_2) \cdot \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= 1 + x_1 + 2x_2 + \frac{1}{2} (x_1; x_2) \cdot \begin{pmatrix} 3x_1 \\ 4x_2 \end{pmatrix} \\ &= 1 + x_1 + 2x_2 + \frac{3}{2} x_1^2 + 2x_2^2 \end{aligned}$$

Sei  $x^0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  und somit

$$f(x^0) = f(1; 1) = 1 + 1 + 2 \cdot 1 + \frac{3}{2}(1)^2 + 2(1)^2 = \frac{15}{2} = 7,5.$$

Es folgt

$$\nabla f(x^k) = c + Cx^k \Rightarrow \nabla f(x^0) = c + Cx^0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \end{pmatrix}$$

und

$$\sigma_k = \frac{\|s^k\|^2}{s^{kT} C s^k} \Rightarrow \sigma_0 = \frac{(4; 6) \cdot \begin{pmatrix} 4 \\ 6 \end{pmatrix}}{(4; 6) \cdot \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 6 \end{pmatrix}} = \frac{13}{48} = 0,2708.$$

Es gilt also

$$x^{k+1} = x^k + \sigma_k s^k \Rightarrow x^1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{13}{48} \cdot \begin{pmatrix} -4 \\ -6 \end{pmatrix} = \begin{pmatrix} -\frac{1}{12} \\ -\frac{5}{8} \end{pmatrix} = \begin{pmatrix} -0,0833 \\ -0,625 \end{pmatrix}.$$

Eingesetzt in

$$f(x^1) = f\left(-\frac{1}{12}; -\frac{5}{8}\right) = 1 + \left(-\frac{1}{12}\right) + 2\left(-\frac{5}{8}\right) + \frac{3}{2}\left(-\frac{1}{12}\right)^2 + 2\left(-\frac{5}{8}\right)^2 = \frac{11}{24} = 0,4583,$$

$$\nabla f(x^1) = c + Cx^1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix} \cdot \begin{pmatrix} -\frac{1}{12} \\ -\frac{5}{8} \end{pmatrix} = \begin{pmatrix} \frac{3}{4} \\ -\frac{1}{2} \end{pmatrix}$$

und in

$$\sigma_1 = \frac{\left(\frac{3}{4}; -\frac{1}{2}\right) \cdot \begin{pmatrix} \frac{3}{4} \\ -\frac{1}{2} \end{pmatrix}}{\left(\frac{3}{4}; -\frac{1}{2}\right) \cdot \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix} \cdot \begin{pmatrix} \frac{3}{4} \\ -\frac{1}{2} \end{pmatrix}} = \frac{\frac{13}{16}}{\frac{43}{16}} = \frac{13}{43} = 0,3023$$

So folgt

$$x^2 = \begin{pmatrix} -\frac{1}{12} \\ -\frac{5}{8} \end{pmatrix} + \frac{13}{43} \cdot \begin{pmatrix} -\frac{3}{4} \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} -\frac{40}{129} \\ -\frac{163}{344} \end{pmatrix} = \begin{pmatrix} -0,3101 \\ -0,4738 \end{pmatrix}$$

und

$$f(x^2) = f\left(-\frac{40}{129}; -\frac{163}{344}\right) = 1 + \left(-\frac{40}{129}\right) + 2\left(-\frac{163}{344}\right) + \frac{3}{2}\left(-\frac{40}{129}\right)^2 + 2\left(-\frac{163}{344}\right)^2 = \frac{1385}{4128} = 0,3355.$$

Es wird nun die Lösung des Gradientenverfahren nach der 2. Iteration mit dem eigentlichen globalen Minimum der Funktion verglichen:

$$\bar{x} = -C^{-1}c \Rightarrow \bar{x} = -\begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = -\frac{1}{12} \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = -\frac{1}{12} \begin{pmatrix} 4 \\ 6 \end{pmatrix} = \begin{pmatrix} -\frac{1}{3} \\ -\frac{1}{2} \end{pmatrix}$$

und

$$f(\bar{x}) = f\left(-\frac{1}{3}; -\frac{1}{2}\right) = 1 + \left(-\frac{1}{3}\right) + 2\left(-\frac{1}{2}\right) + \frac{3}{2}\left(-\frac{1}{3}\right)^2 + 2\left(-\frac{1}{2}\right)^2 = \frac{1}{3} = 0,3333.$$

Es folgen nun wichtige Ungleichungen für die Beurteilung der Konvergenzgeschwindigkeit des Gradientenverfahrens. Dabei werden die beiden Lemmata hier ohne Beweis geführt, die entsprechenden Beweise können aber in [GK99] nachgelesen werden.

**Lemma 3.5.1 Kantorovich-Ungleichung.**

Sei  $C \in \mathbb{R}^{n \times n}$  eine symmetrische, positiv definite Matrix. Dann gilt:

$$\frac{\|d\|^4}{(d^T C d)(d^T C^{-1} d)} \geq \frac{4\lambda_{\min}(C)\lambda_{\max}(C)}{(\lambda_{\min}(C) + \lambda_{\max}(C))^2} \quad \forall d \in \mathbb{R}^n \setminus \{0\}.$$

**Lemma 3.5.2** Sei  $A \in \mathbb{R}^{n \times n}$  eine symmetrische und positiv definite Matrix. Sind  $\lambda_{\min}$  und  $\lambda_{\max}$  der kleinste beziehungsweise größte Eigenwert von  $A$ , so gilt

$$\lambda_{\min} u^T u \leq u^T A u \leq \lambda_{\max} u^T u$$

für alle  $u \in \mathbb{R}^n$ .

Mittels dieser Ungleichungen wird nun der folgende Satz bezüglich der Konvergenzgeschwindigkeit für das Gradientenverfahren mit optimaler Schrittweite und streng konvexer quadratischer Zielfunktion bewiesen.

**Satz 3.5.1** Sei  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  streng konvex und quadratisch. Weiter seien die Folgen  $(x^k)$  und  $(\sigma_k)$  durch das Gradientenverfahren mit Minimierungsregel erzeugt. Dann gilt:

$$f(x^{k+1}) - f(\bar{x}) \leq \left( \frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)} \right)^2 (f(x^k) - f(\bar{x})), \quad (3.9)$$

$$\|x^k - \bar{x}\| \leq \sqrt{\frac{\lambda_{\max}(C)}{\lambda_{\min}(C)}} \left( \frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)} \right)^k \|x^0 - \bar{x}\|, \quad (3.10)$$

wobei  $\bar{x} = -C^{-1}c$  das globale Minimum von  $f$  bezeichnet und  $\lambda_{\max}(C)$  beziehungsweise  $\lambda_{\min}(C)$  der maximale beziehungsweise minimale Eigenwert von  $C$  sind.

*Beweis* Da  $f$  quadratisch ist, liefert Taylor-Entwicklung um  $\bar{x}$

$$\begin{aligned} f(x) &= f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2} (x - \bar{x})^T C (x - \bar{x}) \\ f(x) - f(\bar{x}) &= \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2} (x - \bar{x})^T C (x - \bar{x}) \\ f(x) - f(\bar{x}) &= \frac{1}{2} (x - \bar{x})^T C (x - \bar{x}) \end{aligned}$$

wobei  $\nabla f(\bar{x}) = 0$  verwendet wurde. Werden beide Seiten abgeleitet, so folgt wieder mit  $\nabla f(\bar{x}) = 0$  und analog zu der Vorgehensweise mit der mehrdimensionalen Kettenregel zur Einleitung der Kapitel 3.5 folgt:

$$\nabla f(x) = C(x - \bar{x})$$

Weiter liefert die Taylor-Entwicklung um  $x^k$

$$\begin{aligned} f(x^{k+1}) &= f(x^k) + \sigma_k \underbrace{\nabla f(x^k)^T s^k}_{= -s^k} + \frac{\sigma_k^2}{2} s^k C s^k \\ &= f(x^k) - \sigma_k \|s^k\|^2 + \frac{\sigma_k^2}{2} s^k C s^k \end{aligned}$$

zudem wird (3.8) verwendet

$$\begin{aligned}
f(x^{k+1}) - f(\bar{x}) &= f(x^k) - f(\bar{x}) - \sigma_k \cdot \|s^k\|^2 + \sigma_k^2 \cdot \frac{1}{2} s^k T C s^k \\
&= f(x^k) - f(\bar{x}) - \frac{\|s^k\|^2}{s^k T C s^k} \cdot \|s^k\|^2 + \left( \frac{\|s^k\|^2}{s^k T C s^k} \right)^2 \cdot \frac{1}{2} s^k T C s^k \\
&= f(x^k) - f(\bar{x}) - \frac{\|s^k\|^4}{s^k T C s^k} + \frac{\|s^k\|^4}{(s^k T C s^k)^2} \cdot \frac{1}{2} s^k T C s^k \\
&= f(x^k) - f(\bar{x}) - \frac{\|s^k\|^4}{s^k T C s^k} + \frac{1}{2} \frac{\|s^k\|^4}{s^k T C s^k} \\
&= f(x^k) - f(\bar{x}) - \frac{1}{2} \frac{\|s^k\|^4}{s^k T C s^k}.
\end{aligned}$$

Nun gilt

$$\begin{aligned}
f(x^k) - f(\bar{x}) &= \frac{1}{2} (x^k - \bar{x})^T C (x^k - \bar{x}) \\
&= \frac{1}{2} (x^k - \bar{x})^T C C^{-1} C (x^k - \bar{x}) \\
&= \frac{1}{2} (C(x^k - \bar{x}))^T C^{-1} (C(x^k - \bar{x})) \\
&= \frac{1}{2} s^k T C^{-1} s^k,
\end{aligned}$$

wobei hier ein weiteres Mal die Rechenregeln für das Transponieren eines Produktes zweier Matrizen und die Symmetrieeigenschaft von  $C$  verwendet wird. Zudem werden  $I = C^{-1}C$  und  $s^k = -\nabla f(x^k) = -C(x^k - \bar{x})$  in der obigen Gleichung angewendet. Daher gilt auch

$$1 = \frac{f(x^k) - f(\bar{x})}{\frac{1}{2} s^k T C^{-1} s^k}$$

und somit

$$\begin{aligned}
f(x^{k+1}) - f(\bar{x}) &= f(x^k) - f(\bar{x}) - \frac{1}{2} \frac{\|s^k\|^4}{s^k T C s^k} \\
&= f(x^k) - f(\bar{x}) - \frac{1}{2} \frac{\|s^k\|^4}{s^k T C s^k} \cdot \frac{f(x^k) - f(\bar{x})}{\frac{1}{2} s^k T C^{-1} s^k} \\
&= \left( 1 - \frac{1}{2} \frac{2 \frac{\|s^k\|^4}{(s^k T C s^k)(s^k T C^{-1} s^k)}}{(s^k T C s^k)(s^k T C^{-1} s^k)} \right) (f(x^k) - f(\bar{x})) \\
&= \left( 1 - \frac{\|s^k\|^4}{(s^k T C s^k)(s^k T C^{-1} s^k)} \right) (f(x^k) - f(\bar{x})).
\end{aligned}$$

Die Kantorovich-Ungleichung (Lemma 3.5.1) wird angewendet

$$\begin{aligned}
1 - \frac{\|s^k\|^4}{(s^k T C s^k)(s^k T C^{-1} s^k)} &\leq 1 - \frac{4\lambda_{min}(C)\lambda_{max}(C)}{(\lambda_{min}(C) + \lambda_{max}(C))^2} \\
&\leq \frac{(\lambda_{min}(C) + \lambda_{max}(C))^2}{(\lambda_{min}(C) + \lambda_{max}(C))^2} - \frac{4\lambda_{min}(C)\lambda_{max}(C)}{(\lambda_{min}(C) + \lambda_{max}(C))^2} \\
&\leq \frac{\lambda_{min}^2(C) + 2\lambda_{min}(C)\lambda_{max}(C) + \lambda_{max}^2(C) - 4\lambda_{min}(C)\lambda_{max}(C)}{(\lambda_{min}(C) + \lambda_{max}(C))^2} \\
&\leq \frac{\lambda_{min}^2(C) - 2\lambda_{min}(C)\lambda_{max}(C) + \lambda_{max}^2(C)}{(\lambda_{min}(C) + \lambda_{max}(C))^2} \\
&\leq \frac{(\lambda_{min}(C) - \lambda_{max}(C))^2}{(\lambda_{min}(C) + \lambda_{max}(C))^2} \\
&\leq \left( \frac{\lambda_{min}(C) - \lambda_{max}(C)}{\lambda_{min}(C) + \lambda_{max}(C)} \right)^2
\end{aligned}$$

und liefert (3.9)

$$\begin{aligned} f(x^{k+1}) - f(\bar{x}) &= \left(1 - \frac{\|s^k\|^4}{(s^{kT}Cs^k)(s^{kT}C^{-1}s^k)}\right) (f(x^k) - f(\bar{x})) \\ &\leq \left(\frac{\lambda_{\min}(C) - \lambda_{\max}(C)}{\lambda_{\min}(C) + \lambda_{\max}(C)}\right)^2 (f(x^k) - f(\bar{x})). \end{aligned}$$

Es wird nun für  $f(x) - f(\bar{x}) = \frac{1}{2}(x - \bar{x})^T C(x - \bar{x})$  das Lemma 3.5.2 verwendet

$$\begin{aligned} \frac{1}{2}\lambda_{\min}(C)(x - \bar{x})^T(x - \bar{x}) &\leq f(x) - f(\bar{x}) = \frac{1}{2}(x - \bar{x})^T C(x - \bar{x}) \leq \frac{1}{2}\lambda_{\max}(C)(x - \bar{x})^T(x - \bar{x}), \\ \frac{1}{2}\lambda_{\min}(C)\|x - \bar{x}\|^2 &\leq f(x) - f(\bar{x}) = \frac{1}{2}(x - \bar{x})^T C(x - \bar{x}) \leq \frac{1}{2}\lambda_{\max}(C)\|x - \bar{x}\|^2, \end{aligned}$$

somit gilt

$$f(x) - f(\bar{x}) = \frac{1}{2}(x - \bar{x})^T C(x - \bar{x}) \begin{cases} \leq \frac{\lambda_{\max}(C)}{2}\|x - \bar{x}\|^2, & (\star) \\ \geq \frac{\lambda_{\min}(C)}{2}\|x - \bar{x}\|^2. \end{cases}$$

Die Aussage (3.10) folgt aus der Umformung

$$\begin{aligned} \frac{1}{2}\lambda_{\min}(C)\|x^{k+1} - \bar{x}\|^2 &\leq f(x^{k+1}) - f(\bar{x}) \\ \|x^{k+1} - \bar{x}\|^2 &\leq \frac{2}{\lambda_{\min}(C)} (f(x^{k+1}) - f(\bar{x})) \\ \|x^{k+1} - \bar{x}\|^2 &\stackrel{(3.9)}{\leq} \frac{2}{\lambda_{\min}(C)} \left(\frac{\lambda_{\min}(C) - \lambda_{\max}(C)}{\lambda_{\min}(C) + \lambda_{\max}(C)}\right)^2 (f(x^k) - f(\bar{x})) \\ \|x^{k+1} - \bar{x}\|^2 &\stackrel{(\star)}{\leq} \frac{2}{\lambda_{\min}(C)} \left(\frac{\lambda_{\min}(C) - \lambda_{\max}(C)}{\lambda_{\min}(C) + \lambda_{\max}(C)}\right)^2 \frac{\lambda_{\max}(C)}{2} \|x^k - \bar{x}\|^2 \\ \|x^{k+1} - \bar{x}\| &\leq \sqrt{\frac{\lambda_{\max}(C)}{\lambda_{\min}(C)}} \left(\frac{\lambda_{\min}(C) - \lambda_{\max}(C)}{\lambda_{\min}(C) + \lambda_{\max}(C)}\right) \|x^k - \bar{x}\| \end{aligned}$$

und zusätzlich mit Hilfe der Induktion

$$\|x^k - \bar{x}\| \leq \sqrt{\frac{\lambda_{\max}(C)}{\lambda_{\min}(C)}} \left(\frac{\lambda_{\min}(C) - \lambda_{\max}(C)}{\lambda_{\min}(C) + \lambda_{\max}(C)}\right)^k \|x^0 - \bar{x}\|.$$

□

Mit der *spektralen Konditionszahl*  $\kappa := \kappa(\nabla^2 f) = \kappa(C) = \frac{\lambda_{\max}}{\lambda_{\min}}$  (für die positiv definite Matrix  $C$ ) lassen sich die Abschätzungen aus 3.5.1 auch wie folgt formulieren:

$$\begin{aligned} f(x^{k+1}) - f(\bar{x}) &\leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 (f(x^k) - f(\bar{x})) \\ \|x^k - \bar{x}\| &\leq \sqrt{\kappa} \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x^0 - \bar{x}\| \end{aligned}$$

Somit gilt unter den Voraussetzungen des Satzes 3.5.1 q-lineare Konvergenz für die Folge der Funktionswerte  $(f(x^k))_{k \in \mathbb{N}}$ . Für die Folge der Iterierten  $(x^k)_{k \in \mathbb{N}}$  folgt mit [RH13], Satz 3.3] aber nur die r-lineare Konvergenz. Die Konvergenz des Verfahrens des steilsten Abstiegs mit perfekter Schrittweite bei streng konvexer quadratischer Zielfunktion  $f(x) = c_0 + c^T x + \frac{1}{2}x^T C x$  ist also umso langsamer, je größer die spektrale Konditionszahl der Matrix  $C$  ist [RH13]. Denn je größer die Konditionszahl der Matrix  $C$  ist, desto näher ist der Bruch auf der rechten Seite an 1, und desto langsamer konvergiert die Folge auf der linken Seite gegen Null. Tatsächlich kann also die Konvergenz des Gradientenverfahrens beliebig langsam sein - und das bereits im Idealfall einer streng konvexen, quadratischen Zielfunktion [CC16].

### Beispiel 3.5.2

Als Beispiel wird das Verfahren des steilsten Abstiegs mit optimaler Schrittweite auf einer streng konvexen und quadratischen Funktion  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x) = x_1^2 + 10x_2^2$  und der Startpunkt  $x^0 = (10, 1)^T$  angewendet. Es ergibt sich

$$\nabla f(x) = (2x_1; 20x_2)^T \quad \text{und} \quad \nabla^2 f(x) = C = \begin{pmatrix} 2 & 0 \\ 0 & 20 \end{pmatrix}.$$

Mit den Eigenwerten  $\lambda_{\max} = 20$  und  $\lambda_{\min} = 2$  von  $C$  erfolgt die Konditionszahl  $\kappa(C) = \frac{20}{2} = 10$  und somit

$$f(x^{k+1}) - f(\bar{x}) \leq \left( \frac{10-1}{10+1} \right)^2 \cdot (f(x^k) - f(\bar{x})) = 0,6694 \cdot (f(x^k) - f(\bar{x})).$$

In Abbildung 3.2 ist das Gradientenverfahren mit optimaler Schrittweite, angewandt auf die oben erläuterte Funktion, dargestellt.

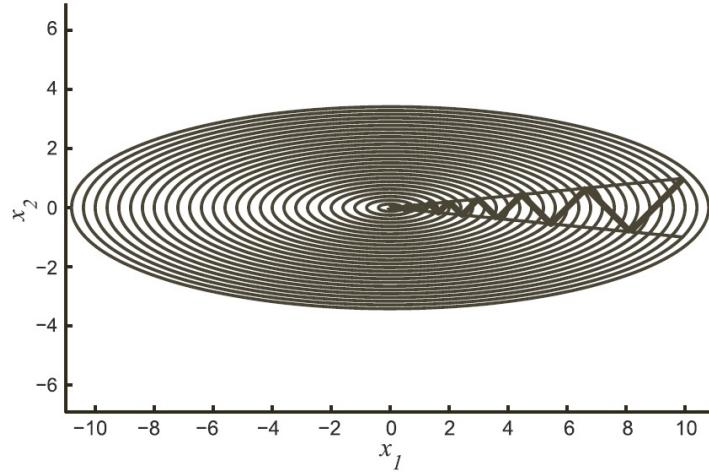


Abbildung 3.2: Gradientenverfahren mit optimaler Schrittweite

Wenn ein Problem gut konditioniert ist, zum Beispiel  $f(x) = x_1^2 + x_2^2$ , zeigt der negative Gradient in die Richtung des Minimums  $\bar{x}$ . Denn es gilt

$$\nabla f(x) = (2x_1; 2x_2)^T, \quad \nabla^2 f(x) = C = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad \kappa = \frac{2}{2} = 1 \quad \text{und} \quad \left( \frac{1-1}{1+1} \right)^2 = 0$$

Das heißt im Falle von  $\lambda_{\min} = \lambda_{\max}$  konvergiert das Verfahren sogar in einem einzigen Schritt zur optimalen Lösung. Wenn andererseits das Problem schlecht konditioniert ist und der negative Gradient nicht in die Nähe des Minimums  $\bar{x}$  weist, wie oben im Beispiel, konvergiert das Gradientenverfahren sehr langsam (Zick-Zack-Effekt). Die Abbildung 3.3 illustriert die beiden erläuterten Fälle [GK13].

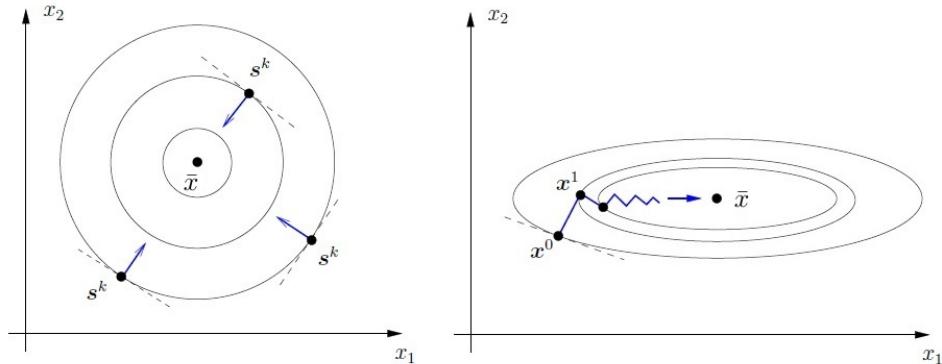


Abbildung 3.3: Beispiel eines gut und eines schlecht konditionierten Problems für das Gradientenverfahren

## 4 Trust-Region-Verfahren

---

In diesem Kapitel wird das *Trust-Region-Verfahren* erläutert. Es wird eine quadratische Modellfunktion  $q_k$  der Zielfunktion  $f$  optimiert. Dabei werden nur Schrittweiten innerhalb des Trust-Region-Radius  $\Delta$  zugelassen, damit das Modell der Zielfunktion zuverlässig annähern kann. In Abschnitt 4.1 wird zunächst die wesentliche Idee des Trust-Region-Algorithmus vorgestellt und in Abschnitt 4.2 wird dann dessen Konvergenzeigenschaften untersucht. Im Anschluss daran wird in Abschnitt 4.3 das im Allgemeinen nicht konvexe Trust-Region-Teilproblem im Detail behandelt und eine Charakterisierung für ein globales Minimum vorgeführt. Schließlich wird in Abschnitt 4.4 bewiesen, unter welchen Voraussetzungen eine schnelle Konvergenz zu erwarten ist.

### 4.1 Einleitung

Trust-Region-Verfahren wurden in den 70er Jahren erstmals vorgestellt und zählen seitdem zu den leistungsfähigsten Globalisierungsverfahren der nichtlinearen Optimierung. Das Buch von Andrew Conn, Nicholas Gould und Philippe Toint [CG00] bietet einen umfangreichen Überblick.

Es wird wieder das unrestringierte Optimierungsproblem

$$\min_{x \in \mathbb{R}^n} f(x) \quad (4.1)$$

mit der Zielfunktion  $f : \mathbb{R} \rightarrow \mathbb{R}^n$  betrachtet.

Bei dem Gradientenverfahren wurde zuerst eine Abstiegsrichtung  $s^k$  und dann mit Hilfe eines Schrittweitenverfahrens (zum Beispiel über die Armijo-Regel) eine Schrittweite  $\sigma_k$  bestimmt, die einen hinreichend großen Abstieg in Richtung  $s^k$  versichert. Bei den Trust-Region-Verfahren wird jedoch erst eine Schrittweite  $\Delta$  vorgegeben und dann wird eine Richtung  $s$  gesucht. Zur Motivation des Verfahrens wird zusätzlich von einer zweimal stetig differenzierbaren Zielfunktion ausgegangen. Bevor aber die Idee des Verfahrens erläutert wird, folgt nun eine kurze Wiederholung des Gedankengangs beim Abstiegsverfahren.

Ein generelles Prinzip zur Bestimmung einer Abstiegsrichtung besteht darin, die Funktion  $f$  lokal in der Nähe der Iterierten  $x^k$  durch eine Approximation  $\hat{f}(y)$  zu ersetzen und diese dann zu minimieren. Bei dem Gradientenverfahren erfolgte dies durch eine lineare Approximation [GM16].

#### Lineare Approximation:

Linearisierung von  $f$  in  $x$  ergibt die Funktion

$$\hat{f}(y) = f(x) + \nabla f(x)^T(y - x).$$

Allerdings besitzt diese lineare Funktion im Allgemeinen kein Minimum auf  $\mathbb{R}^n$ . Daher werden zusätzlich die Werte von  $y$  eingeschränkt, indem

$$\|y - x\| \leq 1$$

gefordert wird. In Kapitel 3 wurde bereits ausgearbeitet, dass die Lösung  $\hat{y}$  des Minimierungsproblems

$$\hat{f}(y) \rightarrow \min \quad \text{unter} \quad \|y - x\| \leq 1$$

gerade auf die (normierte) Suchrichtung  $d = \hat{y} - x = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$  führt. Es entsteht also das Gradientenverfahren.

Es ist nun naheliegend,  $f$  quadratisch zu approximieren, um ein weiteres Verfahren zu erhalten.

### Quadratische Approximation:

$f$  wird lokal durch

$$\hat{f}(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)$$

approximiert. Ist die Hesse-Matrix  $\nabla^2 f(x)$  positiv definit, so besitzt  $\hat{f}$  ein eindeutig bestimmtes Minimum  $\hat{y}$ , welches durch das lineare Gleichungssystem

$$\nabla \hat{f}(\hat{y}) = \nabla f(x) + \nabla^2 f(x)(\hat{y} - x) = 0$$

bestimmt ist. Minimierung der quadratischen Approximation  $\hat{f}(y)$  führt also auf die Suchrichtung

$$d = \hat{y} - x = -\nabla^2 f(x)^{-1} \nabla f(x).$$

Wegen

$$\nabla f(x)^T d = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0$$

für  $\nabla f(x) \neq 0$  ist  $d$  eine Abstiegsrichtung von  $f$  in  $x$ .

Mit  $x^k \in \mathbb{R}^n$  wird nun die aktuelle Iterierte bezeichnet und die Berechnung eines Schrittes  $s^k$  zur Bestimmung der neuen Iterierten  $x^{k+1} = x^k + s^k$  basiert auf folgender Idee:

Durch Taylor-Entwicklung von  $f(x^k + s)$  um  $s = 0$  entsteht ein *quadratisches Modell*

$$q_k(s) = f_k + g^k s + \frac{1}{2} s^T H_k s$$

mit  $f_k = f(x^k)$ ,  $g^k = \nabla f(x^k)$  und  $H_k = \nabla^2 f(x^k)$ . Das quadratische Modell kann aber auch in dieser Form betrachtet werden

$$q_k(x) = f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T H_k(x - x^k).$$

Ist die Hesse-Matrix nicht positiv definit, so besitzt  $q_k$  unter Umständen kein Minimum. Dieses Problem ist auch schon einmal beim Gradientenverfahren für die lineare Funktion  $f(x) + \nabla f(x)^T d$  aufgetreten. Dort wurde zusätzlich die Nebenbedingung  $\|d\| \leq 1$  eingeführt, um ein wohldefiniertes Minimierungsproblem zu erhalten. Es ist daher einleuchtend, auch für  $s$  eine Nebenbedingung zu formulieren. Aus diesem Grunde wird auch der Schritt  $s^k$  durch (restringierte) Minimierung von  $q_k$  gewonnen.

Das Modell  $q_k(s)$  stimmt in einer Umgebung von  $s = 0$  gut mit  $f(x^k + s)$  überein, denn nach dem Satz von Taylor gilt:  $f(x^k + s) = q_k(s) + o(\|s\|^2)$ . Das heißt es soll

$$q_k(0) = f(x^k) = f_k$$

gelten. Ist aber  $\|s\|$  groß, so muss dies natürlich nicht mehr gelten. Daher ist es sinnvoll, dem Modell  $q_k$  nur auf einem *Vertrauensbereich* (einer *Trust-Region*)  $\{s; \|s\| \leq \Delta_k\}$  zu trauen. Hierbei bezeichnet  $\Delta_k > 0$  den *Trust-Region-Radius*. Die Schrittberechnung erfolgt nun durch Lösen des **Trust-Region Teilproblems**:

$$\min_{s \in \mathbb{R}^n} \{q_k(s); \|s\| \leq \Delta_k\}. \quad (4.2)$$

Die geometrische Idee des Trust-Region-Verfahrens entspricht der Abbildung 4.1 [CA07].

Somit wird das Modell  $q_k$  beim Trust-Region-Verfahren nicht im  $\mathbb{R}^n$  sondern in einer Kugel  $B_{\Delta_k}(x^k)$  zu gegebenem  $\Delta_k > 0$  und der Nebenbedingung  $\|s^k\| = \|x^{k+1} - x^k\| \leq \Delta_k$  minimiert. Mit Hilfe von Satz von Weierstraß, einer der Hauptsätze der Analysis, besitzt das Minimierungsproblem  $q_k$  stets eine Lösung. Denn die Funktion  $q_k$  ist stetig und mit  $\|s\| \leq \Delta_k$  sogar kompakt, somit wird die Funktion  $q_k$  im Definitionsbereich ihr Maximum sowie Minimum annehmen. Speziell auch für eine nicht positiv definite Hesse-Matrix.

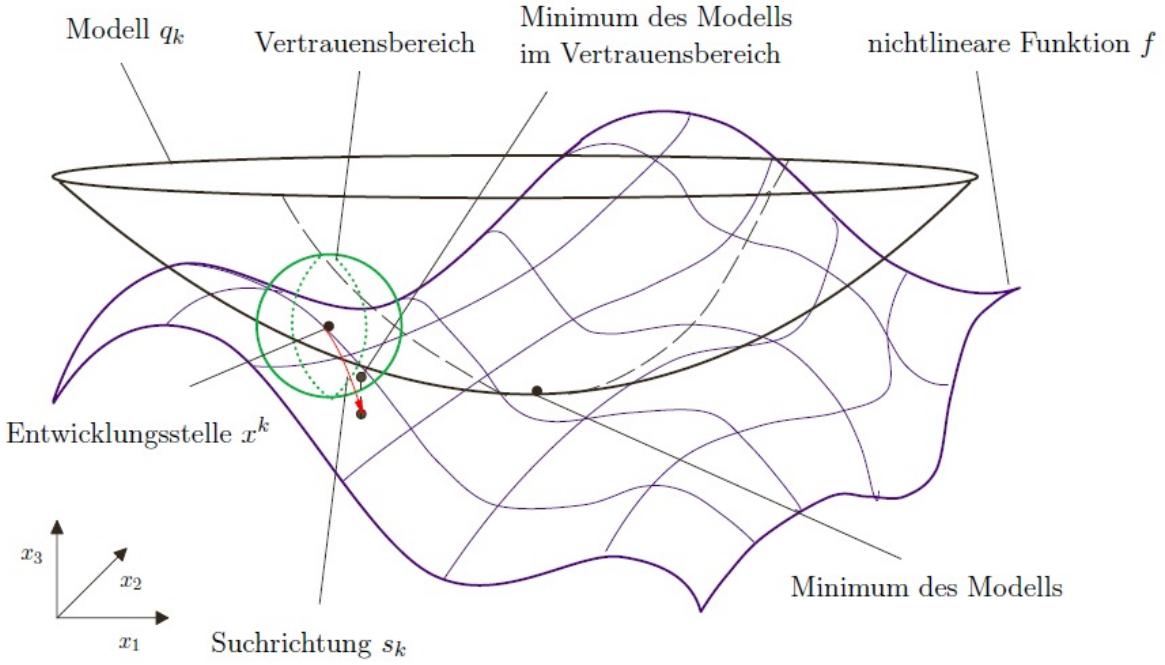


Abbildung 4.1: Die geometrische Darstellung des Trust-Region-Verfahrens.

#### Anpassung des Vertrauensbereichs:

Es stellt sich nun die Frage, wie  $\Delta_k$  gewählt werden sollte. Hierfür erfolgt die Bewertung der Qualität des berechneten Schritts durch Vergleich der Abnahme der Modelfunktion  $q_k$  (predicted reduction)

$$pred_k(s^k) = q_k(0) - q_k(s^k) = f_k - q_k(s^k)$$

und der tatsächlichen Abnahme (actual reduction) der Zielfunktion

$$ared_k(s^k) = f_k - f(x^k + s^k).$$

Der Quotient

$$\rho_k(s^k) = \frac{ared_k(s^k)}{pred_k(s^k)} \quad (4.3)$$

dient somit als Qualitätsmaß für die Übereinstimmung von  $pred_k$  und  $ared_k$ .

Unabhängig vom aktuellen Iterationsschritt werden Parameter  $0 < \eta_1 < \eta_2 < 1$  gewählt (in [CG00] werden beispielsweise die Werte  $\eta_1 = 0,1$  und  $\eta_2 = 0,9$  vorgeschlagen). Dann wird wie folgt verfahren:

- Ist  $\rho_k$  sehr klein (oder sogar negativ), das heißt  $\rho_k \leq \eta_1$ , so war  $q_k$  kein gutes Modell für  $f$  im Vertrauensbereich. Wenn die  $f$ -Abnahme  $ared_k(s^k)$  unbefriedigend im Vergleich zur Modellabnahme  $pred_k(s^k)$  ist, wird das darauf zurückgeführt, dass der Vertrauensbereich zu groß gewählt wurde. Daher wird auch der Schritt  $s^k$  verworfen und der Trust-Region-Radius wird reduziert:  $x^{k+1} = x^k$ ,  $\Delta_{k+1} < \Delta_k$ .
- Ist der Quotient klein aber nicht zu sehr klein, das heißt ist  $\rho_k \in (\eta_1, \eta_2]$ , so stimmt das Modell  $q_k$  im Vertrauensbereich hinreichend gut mit der Funktion  $f$  überein und die neue Näherung ist somit zufriedenstellend. Der Schritt  $x^{k+1} = x^k + s^k$  wird akzeptiert und der Radius wird beibehalten.
- Ist  $\rho_k$  ungefähr 1 oder gilt der verschärzte Test  $\rho_k > \eta_2$ , so ist die Übereinstimmung des quadratischen Modells  $q_k$  auf dem Vertrauensbereich mit der nichtlinearen Funktion  $f$  sehr gut. In diesem Fall wird nicht nur der Schritt  $x^{k+1} = x^k + s^k$  akzeptiert, sondern auch der Radius wird vergrößert:  $\Delta_{k+1} > \Delta_k$ .

Im Allgemeinen wird also der Schritt  $s^k$  akzeptiert, falls  $\rho_k \geq \eta_1$  gilt. Andernfalls wird der berechnete Schritt verworfen. So wird also mit Hilfe der Parameter  $0 < \eta_1 < \eta_2$  entschieden, ob der Radius  $\Delta_k$  vergrößert, beibehalten oder verkleinert werden soll. Der Radius wird mit einem der Parameter  $0 < \gamma_0 < \gamma_1 < 1 < \gamma_2$  multipliziert und somit der neue Radius verkleinert beziehungsweise vergrößert. Überdies wird durch den Parameter  $\Delta_{min} \geq 0$  verhindert, dass der Radius zu klein wird.

Die am häufigsten verwendete Norm ist die euklidische Norm ( $l_2$ -Norm) und dabei wird das Modell  $q_k$  um einen kugel-, oder ellipsenförmigen Vertrauensbereich optimiert. Werden andere Normen, wie zum Beispiel die Summennorm ( $l_1$ -Norm) oder Maximumsnorm ( $l_\infty$ -Norm), gewählt, so entstehen anders geformte Vertrauensbereiche und dies führt natürlich zu einem anderen Trust-Region-Verfahren. In dieser Arbeit wird jedoch nur die euklidische Norm betrachtet. In Abbildung 4.2 sind von links nach rechts die  $l_1$ -,  $l_2$ -, und  $l_\infty$ -Normen dargestellt [CG00].

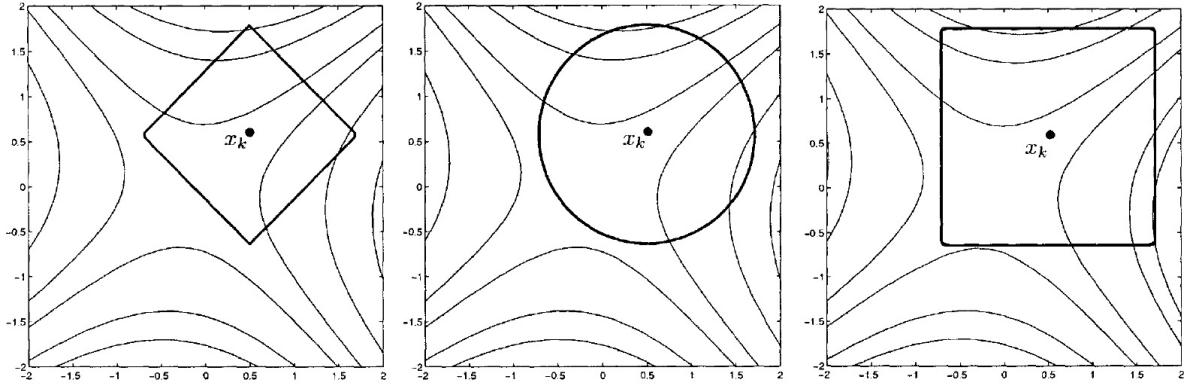


Abbildung 4.2: Unterschiedliche Gestalt des Trust-Region-Radius

Damit sind die grundlegenden Ideen des *Trust-Region Newton Verfahrens* beschrieben. Der Name Newton weist darauf hin, dass hier mit der exakten Hesse-Matrix gearbeitet wird. Im Falle  $\|s^k\| < \Delta_k$  ist dann  $s^k$  ein globales Minimum von  $q_k$ , und  $s^k$  ist somit der Newton-Schritt:

Auch bei Newton-Verfahren wird bei der Schrittberechnung in der  $k$ -ten Iteration das quadratische Modell

$$q_k(s) = f(x^k) + \nabla f(x^k)^T s + \frac{1}{2} s^T \nabla^2 f(x^k) s$$

minimiert. Der Vektor  $s^k$  ist das globale Minimum von  $q_k$  und berechnet sich gemäß

$$\nabla q_k(s^k) = \nabla f(x^k)^T + \nabla^2 f(x^k) s^k \stackrel{!}{=} 0.$$

Wird es die Gleichung nach  $s^k$  aufgelöst, so folgt

$$s^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k).$$

Es wird hier abermals darauf hingewiesen, dass  $H_k$  zwar eine symmetrische, aber nicht notwendig positiv definite Matrix ist. Klar ist, dass dieses Hilfsproblem eine globale Lösung  $s^*$  besitzt, da wie oben schon erwähnt, eine stetige Funktion auf einer nichtleeren, kompakten Menge ihr Minimum annimmt. Da  $H_k$  aber nicht als positiv definit vorausgesetzt wurde, so dass die Zielfunktion im Hilfsproblem nicht notwendig gleichmäßig konvex ist, kann (4.2) auch von einer globalen Lösung verschiedene lokale Lösungen besitzen. In Kapitel 4.3 Charakterisierung der Lösungen des Teilproblems werden notwendige und hinreichende Bedingungen dafür angegeben, dass ein  $s^* \in \mathbb{R}^n$  eine globale Lösung von (4.2) ist [WJ92]. Steht aber die Hesse-Matrix von  $f$  nicht zur Verfügung zum Beispiel weil  $f$  nicht zweimal differenzierbar oder die Berechnung von  $\nabla^2 f$  zu aufwendig ist, so kann für die Hesse-Matrix  $H_k$  des Modells  $q_k$  eine geeignete symmetrische Approximation der Hesse-Matrix, zum Beispiel eine Quasi-Newton-Approximation verwendet werden.

Die Funktion  $f$  muss nach unten beschränkt und die Hesse-Matrix  $H_k$  muss beschränkt sein, damit der Nachweis globaler Konvergenz erfolgen kann:

**Voraussetzung 4.1.1** Die Zielfunktion  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  ist stetig differenzierbar und nach unten beschränkt.

**Voraussetzung 4.1.2** Es gibt eine Konstante  $C_H > 0$ , so dass für alle  $k$  gilt:

$$\|H_k\| \leq C_H.$$

Die Abbildung 4.3 zeigt einen möglichen Schritt eines Trust-Region-Verfahrens im Vergleich zu einem Schritt eines Suchrichtungsverfahrens (zum Beispiel Gradientenverfahren) mit gleicher Schrittweite. Hier in diesem Beispiel führt der Trust-Region-Schritt näher an das Minimum von  $f$ , als der Suchrichtungsschritt [SO04].

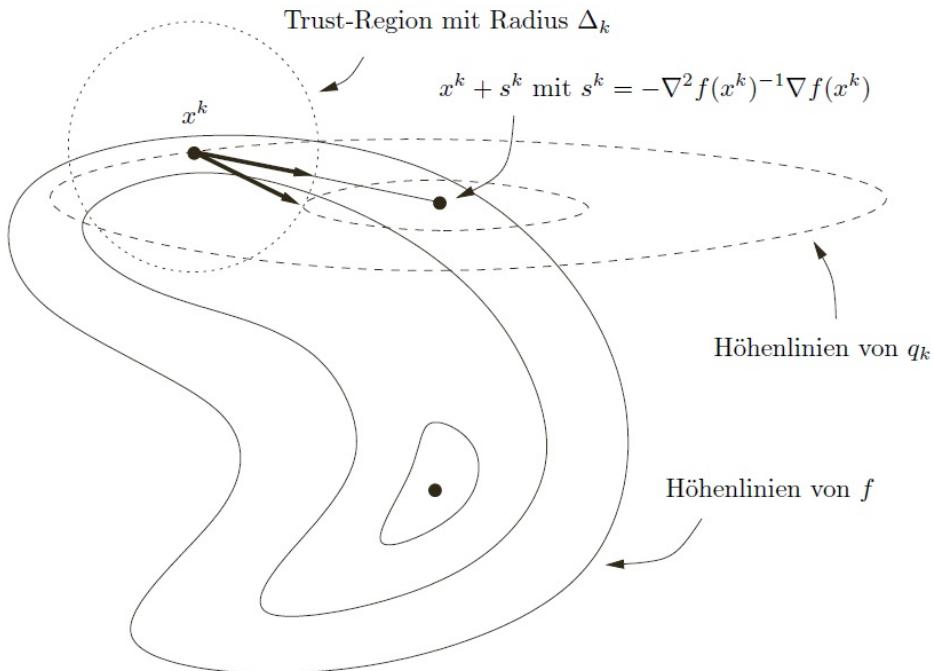


Abbildung 4.3: Schritte von Suchrichtungs- und Trust-Region-Verfahren

Es wird wieder der Fall einer symmetrischen Matrix  $H_k \neq 0$  betrachtet. Das Trust-Region-Teilproblem (4.2) zu lösen, kann schwer sein, insbesondere für eine indefinite Matrix  $H_k$ . Zum Nachweis der globalen Konvergenz des  $q_k$  reicht es aber auch nur entlang des negativen Gradienten zu minimieren. Somit entsteht für  $H_k = 0$  das Trust-Region-Gradientenverfahren. Hierbei wird der zulässige Bereich stark verkleinert, da nur noch die bestimmte Suchrichtung  $s_l$  zugelassen wird und es entsteht eine spezielle Schrittweitensteuerung  $\tau$ .

Ausgangspunkt zur näherungsweisen Lösung von (4.2) ist der *Cauchy-Punkt* (definiert als  $s_c = \tau \cdot s_l$ ), dessen Konstruktion in drei Schritten erfolgt [HH11].

- 1.) Bestimme  $s_l^k$  als Lösung des Optimierungsproblems

$$\min_{\|s\| \leq \Delta_k} f(x^k) + \nabla f(x^k)^T s.$$

Aus Satz 3.2.1 folgt für dessen Lösung sofort

$$s_l^k = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|} \cdot \Delta_k = -\frac{g^k}{\|g^k\|} \cdot \Delta_k$$

2.) Berechne eine Cauchy-Schrittweite  $\tau_k$  als Lösung von

$$\min_{s \in \mathbb{R}^n} q_k(s) \text{ u. d. N. } \|s\| \leq \Delta_k, \quad s = \tau \cdot s_l \quad \text{mit } \tau \geq 0$$

Setze dann  $s_c^k := \tau_k \cdot s_l^k$ .

3.) Dieses Problem ist offenbar äquivalent zu

$$\min_{\tau \geq 0} q_k(\tau \cdot s_l) \text{ u. d. N. } \tau \leq 1$$

und damit zu

$$\min_{\tau \in [0,1]} \phi(\tau)$$

mit

$$\phi(\tau) := q_k(\tau \cdot s_l^k) = f(x^k) + \tau g^{kT} s_l^k + \frac{1}{2} \tau^2 s_l^{kT} H_k s_l^k.$$

Um  $\tau_k$  explizit anzugeben, wird eine Fallunterscheidung durchgeführt:

(i) Für  $g^{kT} H_k g^k \leq 0$  gilt

$$\begin{aligned} \frac{d}{d\tau_k} \phi &= g^{kT} s_l^k + \tau_k s_l^{kT} H_k s_l^k = g^{kT} \cdot \left( -\frac{g^k \cdot \Delta_k}{\|g^k\|} \right) + \tau_k \cdot \left( -\frac{g^k \cdot \Delta_k}{\|g^k\|} \right)^T \cdot H_k \cdot \left( -\frac{g^k \cdot \Delta_k}{\|g^k\|} \right) \\ &= \underbrace{-\Delta_k \|g^k\|}_{< 0} + \tau_k \cdot \underbrace{\frac{\Delta_k^2}{\|g^k\|^2} \cdot g^{kT} H_k g^k}_{\leq 0} \end{aligned}$$

Dann ist die Funktion  $\phi(\tau) := q_k(\tau \cdot s_l^k)$  streng monoton fallend auf  $\tau \in [0, 1]$  und besitzt somit ihr Minimum in  $\tau_k = 1$ .

(ii) Im Falle  $g^{kT} H_k g^k > 0$  gilt entweder  $\tau_k = 1$  oder  $\tau_k < 1$  ergibt sich als freies Minimum der Funktion  $\phi$ . Im zweiten Fall gilt

$$\frac{d}{d\tau_k} \phi = -\Delta_k \|g^k\| + \tau_k \cdot \frac{\Delta_k^2}{\|g^k\|^2} \cdot g^{kT} H_k g^k \stackrel{!}{=} 0$$

und damit

$$\tau_k = \frac{\Delta_k \cdot \|g^k\| \cdot \|g^k\|^2}{\Delta_k^2 \cdot g^{kT} H_k g^k} = \frac{\|g^k\|^3}{\Delta_k g^{kT} H_k g^k}$$

Also gilt insgesamt

$$\tau_k = \begin{cases} 1, & \text{falls } g^{kT} H_k g^k \leq 0 \\ \min \left\{ 1, \frac{\|g^k\|^3}{\Delta_k g^{kT} H_k g^k} \right\}, & \text{sonst.} \end{cases}$$

Die Abbildung 4.4 zeigt die Cauchy-Schrittweite innerhalb des Trust-Region-Radius [NW06].

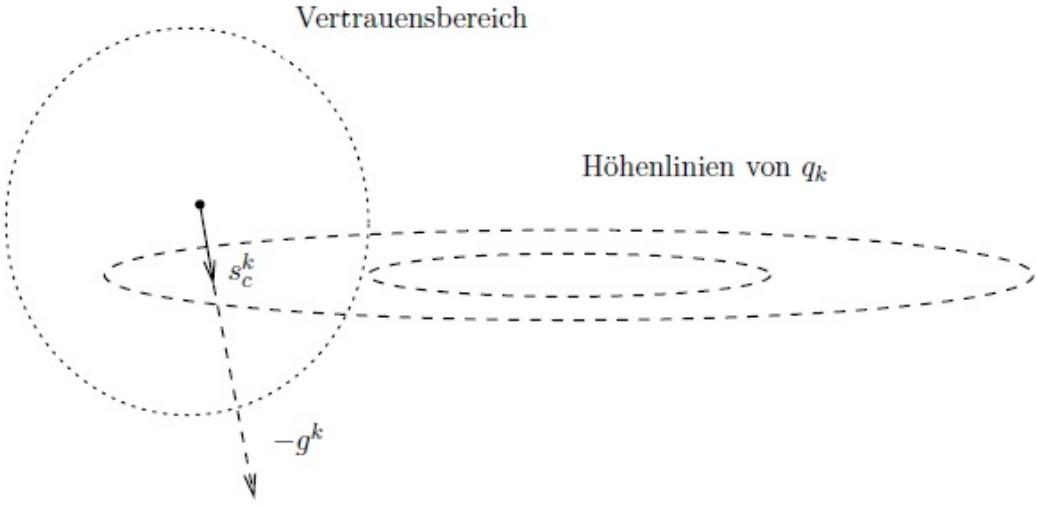


Abbildung 4.4: Die Cauchy-Schrittweite

Wie oben schon erläutert ist die Berechnung exakter Lösungen des Trust-Region-Teilproblems zwar relativ effizient möglich, für große Optimierungsprobleme aber häufig trotzdem zu aufwendig. Das Trust-Region-Verfahren ist aber bereits dann global konvergent, wenn alle berechneten Schritte  $s^k$  folgender Bedingung genügen:

**Cauchy-Abstiegsbedingung (Fraction of Cauchy Decrease):**

Es gibt von  $k$  unabhängige Konstanten  $\alpha \in (0, 1]$  und  $\beta \geq 1$  mit

$$\|s^k\| \leq \beta \Delta_k, \quad \text{pred}_k(s^k) \geq \alpha \cdot \text{pred}_k(s_c^k), \quad (4.4)$$

wobei der *Cauchy-Schritt*  $s_c^k$  die eindeutige Lösung des folgenden eindimensionalen Minimierungsproblems ist:

$$\min q_k(s) \quad \text{u.d.N} \quad s = \tau \cdot s_l^k, \quad \tau \geq 0, \quad \|s\| \leq \Delta_k. \quad (4.5)$$

In seinen Einzelheiten wird das Trust-Region-Verfahren nun im folgenden Algorithmus beschrieben.

#### Algorithmus 4.1.1 Trust-Region-Verfahren.

Wähle Parameter  $\alpha \in (0, 1]$ ,  $\beta \geq 1$ ,  $0 < \eta_1 < \eta_2 < 1$ ,  $0 < \gamma_0 < \gamma_1 < 1 < \gamma_2$  und  $\Delta_{\min} \geq 0$ .

Wähle einen Startpunkt  $x^0 \in \mathbb{R}^n$  und einen Trust-Region-Radius  $\Delta_0 > 0$  mit  $\Delta_0 \geq \Delta_{\min}$ .

Für  $k = 0, 1, 2, \dots$ :

1. Falls  $g^k = 0$ , dann STOP mit Resultat  $x^k$ .
2. Wähle eine symmetrische Matrix  $H_k \in \mathbb{R}^{n \times n}$ .
3. Berechne einen Schritt  $s^k$ , der die Cauchy-Abstiegsbedingung (4.4) erfüllt.
4. Berechne  $\rho_k(s^k)$  gemäß (4.3).
5. Falls  $\rho_k(s^k) > \eta_1$ , dann akzeptiere den Schritt  $s^k$ , das heißt setze  $x^{k+1} = x^k + s^k$ . Andernfalls verwirfe den Schritt, das heißt setze  $x^{k+1} = x^k$ .
6. Berechne  $\Delta_{k+1}$  gemäß Algorithmus 4.1.2.

#### Algorithmus 4.1.2 Update des Trust-Region-Radius $\Delta_k$ .

Seien  $\eta_1, \eta_2$  und  $\gamma_0, \gamma_1, \gamma_2$  wie in Algorithmus 4.1.1 gewählt.

1. Falls  $\rho_k(s^k) \leq \eta_1$ , so wähle  $\Delta_{k+1} \in [\gamma_0 \Delta_k, \gamma_1 \Delta_k]$ .
2. Falls  $\rho_k(s^k) \in (\eta_1, \eta_2]$ , so wähle  $\Delta_{k+1} \in [\max\{\Delta_{min}, \gamma_1 \Delta_k\}, \max\{\Delta_{min}, \Delta_k\}]$ .
3. Falls  $\rho_k(s^k) > \eta_2$ , so wähle  $\Delta_{k+1} \in [\max\{\Delta_{min}, \Delta_k\}, \max\{\Delta_{min}, \gamma_2 \Delta_k\}]$ .

Die Schritte  $s^k$  werden in zwei Klassen eingeteilt, die erfolgreichen und die verworfenen Schritte:

**Definition 4.1.1** Der Schritt  $s^k$  heißt erfolgreich, falls  $\rho_k(s^k) > \eta_1$  gilt und somit  $x^{k+1} = x^k + s^k$  gesetzt wird. Mit  $\mathcal{S} \subset \mathbb{N}_0$  wird die Indexmenge aller erfolgreichen Schritte bezeichnet.

**Beispiel 4.1.1** Sei

$$f(x) = x^4 - 10x^3 + 2x^2 + 10x + 2$$

die zu minimierende Funktion. Es handelt sich also um ein eindimensionales Optimierungsproblem. Deshalb wird hier auch anstatt der Norm  $\|\cdot\|$  der Absolutbetrag  $|\cdot|$  verwendet. Zudem werden bei einer eindimensionalen Optimierung einfacheheitshalber die  $n$ -te Ableitungsstriche, also  $f^n(x)$ , anstelle des nabla-Operators  $\nabla^n f(x)$ , verwendet. Mit dem quadratischen Modell

$$q_k(s) = f(x^k) + f'(x^k)s + \frac{1}{2}f''(x^k)s^2$$

und mit

$$f'(x) = 4x^3 - 30x^2 + 4x + 10 \text{ und } f''(x) = 12x^2 - 60x + 4$$

wird das Minimum der Funktion  $f$  mit Hilfe des Trust-Region-Verfahrens approximiert. Der Cauchy-Schritt ist als  $s_c^k = \tau_k \cdot s_l^k$  mit  $s_l^k = -\frac{f'(x^k)}{|f'(x^k)|} \cdot \Delta_k$  und

$$\tau_k = \begin{cases} 1, & \text{falls } f'(x^k) \cdot f''(x^k) \cdot f'(x^k) \leq 0 \\ \min \left\{ 1, \frac{|f'(x^k)|^3}{\Delta_k f'(x^k) \cdot f''(x^k) \cdot f'(x^k)} \right\}, & \text{sonst} \end{cases}$$

definiert. In diesem Beispiel sind die folgenden Parameter fest gewählt:  $\eta_1 = 0,1$ ,  $\eta_2 = 0,9$ ,  $\gamma_0 = 0,25$ ,  $\gamma_1 = 0,75$  und  $\gamma_2 = 2$  und die Startwerte seien  $x^0 = 1,5$  und  $\Delta_0 = 2$ .

1. Es gilt:  $x^0 = 1,5$  und  $\Delta_0 = 2$ .

Bestimme  $\tau_0$  aus  $\min q_0(s_c^0)$  mit  $s_c^0 = \tau_0 \cdot s_l^0$  unter der Nebenbedingung  $|s_c^0| \leq \Delta_0$ . Prüfe für  $\tau_0$  ob  $f'(x^k) \cdot f''(x^k) \cdot f'(x^k) \leq 0$  oder  $> 0$  folgt. Es gilt

$$f'(x^0) = f'(1,5) = 4 \cdot (1,5)^3 - 30 \cdot (1,5)^2 + 4 \cdot (1,5) + 10 = -38 \neq 0$$

und

$$f''(x^0) = f''(1,5) = 12 \cdot (1,5)^2 - 60 \cdot (1,5) + 4 = -59.$$

Somit folgt

$$f'(x^0) \cdot f''(x^0) \cdot f'(x^0) = (-38) \cdot (-59) \cdot (-38) = -85196 \leq 0,$$

also gilt  $\tau_0 = 1$ . Daher ergibt sich  $s_c^0 = \tau_0 \cdot s_l^0 = s_l^0 = -\frac{-38}{|-38|} \cdot 2 = 2$ . Eingesetzt in  $x^{k+1}$ :

$$x^1 = x^0 + s_c^0 = 1,5 + 2 = 3,5$$

Prüfe nun das Qualitätsmaß  $\rho_0(s_c^0) = \frac{\text{ared}_0(s_c^0)}{\text{pred}_0(s_c^0)}$ . Mit

$$\text{ared}_0(s_c^0) = f(x^0) - f(x^0 + s_c^0) = f(1,5) - f(3,5) = -\frac{115}{16} - \left( -\frac{3475}{16} \right) = 210$$

und

$$\begin{aligned} \text{pred}_0(s_c^0) &= f(x^0) - q_0(s_c^0) = f(x^0) - \left( f(x^0) + f'(x^0)s_c^0 + \frac{1}{2} \cdot f''(x^0)s_c^{0,2} \right) \\ &= -f'(x^0) \cdot s_c^0 - \frac{1}{2}f''(x^0) \cdot s_c^{0,2} = -(-38) \cdot 2 - \frac{1}{2} \cdot (-59) \cdot 2^2 = 76 + 118 = 194 \end{aligned}$$

folgt

$$\rho_0(s_c^0) = \frac{\text{ared}_0(s_c^0)}{\text{pred}_0(s_c^0)} = \frac{210}{194} = 1,08247 > 0,9 = \eta_2.$$

Das bedeutet, dass die tatsächliche Reduktion (Zähler) mit der vorhergesagten Reduktion (Nenner) gut übereinstimmen und dass der Radius  $\Delta_0$  sogar vergrößert werden kann

$$\Delta_1 = \gamma_2 \cdot \Delta_0 = 2 \cdot 2 = 4.$$

Und da  $\rho_0(s_c^0) = 1,08247 > 0,1 = \eta_1$  ist, wird auch der Schritt  $s_c^0$  akzeptiert und es gilt  $x^1 = 3,5$ .

2. Es gilt  $x^1 = 3,5$  und  $\Delta_1 = 4$ .

Bestimme  $\tau_1$  aus  $\min q_1(s_c^1)$  mit  $s_c^1 = \tau_1 \cdot s_l^1$  unter der Nebenbedingung  $|s_c^1| \leq \Delta_1$ . Prüfe für  $\tau_1$  ob  $f'(x^k) \cdot f''(x^k) \cdot f'(x^k) \leq 0$  oder  $> 0$  folgt. Es gilt

$$f'(x^1) = f'(3,5) = -172 \neq 0 \text{ und } f''(x^1) = f''(3,5) = -59$$

und somit auch

$$f'(x^1) \cdot f''(x^1) \cdot f'(x^1) = (-172) \cdot (-59) \cdot (-172) = -1745456 \leq 0$$

also gilt  $\tau_1 = 1$ . Daher ergibt sich  $s_c^1 = \tau_1 \cdot s_l^1 = s_l^1 = -\frac{-172}{-172} \cdot 4 = 4$ . Eingesetzt in  $x^{k+1}$ :

$$x^2 = x^1 + s_c^1 = 3,5 + 4 = 7,5$$

Mit

$$\text{ared}_1(s_c^1) = f(x^1) - f(x^1 + s_c^1) = f(3,5) - f(7,5) = -\frac{3475}{16} - \left(-\frac{13843}{16}\right) = 648$$

und

$$\text{pred}_1(s_c^1) = -f'(x^1) \cdot s_c^1 - \frac{1}{2} \cdot f''(x^1) \cdot s_c^{12} = -(-172) \cdot 4 - \frac{1}{2} \cdot (-59) \cdot 4^2 = 688 + 472 = 1160$$

folgt

$$\rho_1(s_c^1) = \frac{\text{ared}_1(s_c^1)}{\text{pred}_1(s_c^1)} = \frac{648}{1160} = 0,558621 \in (\eta_1, \eta_2].$$

Die neue Näherung ist zufriedenstellend, denn es gilt  $\rho_1(s_c^1) = 0,558621 > 0,1 = \eta_1$  und wird daher auch akzeptiert, also es gilt  $x^2 = 7,5$ . Jedoch wird hier der Radius  $\Delta_1$  mit dem Faktor  $\gamma_1$  multipliziert und der Vertrauensbereich wird verkleinert

$$\Delta_2 = \gamma_1 \cdot \Delta_1 = 0,75 \cdot 4 = 3.$$

3. Es gilt  $x^2 = 7,5$  und  $\Delta_2 = 3$ .

Bestimme  $\tau_2$  aus  $\min q_2(s_c^2)$  mit  $s_c^2 = \tau_2 \cdot s_l^2$  unter der Nebenbedingung  $|s_c^2| \leq \Delta_2$ . Prüfe für  $\tau_2$  ob  $f'(x^k) \cdot f''(x^k) \cdot f'(x^k) \leq 0$  oder  $> 0$  folgt. Es gilt

$$f'(x^2) = f'(7,5) = 40 \neq 0 \text{ und } f''(x^2) = f''(7,5) = 229$$

und somit auch

$$f'(x^2) \cdot f''(x^2) \cdot f'(x^2) = 40 \cdot 229 \cdot 40 = 366400 > 0.$$

Dies bedeutet, es muss noch zusätzlich

$$\frac{|f'(x^2)|^3}{\Delta_2 \cdot f'(x^2) \cdot f''(x^2) \cdot f'(x^2)} < 1 \text{ oder } \geq 1$$

geprüft werden. Es gilt

$$\frac{|f'(x^2)|^3}{\Delta_2 \cdot f'(x^2) \cdot f''(x^2) \cdot f'(x^2)} = \frac{40^3}{3 \cdot 366400} = 0,0582242 < 1.$$

Somit wird  $\tau_2 = \frac{|f'(x^2)|^3}{\Delta_2 \cdot f'(x^2) \cdot f''(x^2) \cdot f'(x^2)}$  in  $s_c^2 = \tau_2 \cdot s_l^2$  eingesetzt. Daher ergibt sich

$$\begin{aligned} s_c^2 &= \tau_2 \cdot s_l^2 = \frac{|f'(x^2)|^3}{\Delta_2 \cdot f'(x^2) \cdot f''(x^2) \cdot f'(x^2)} \cdot -\frac{f'(x^2)}{|f'(x^2)|} \cdot \Delta_2 \\ &= -\frac{|f'(x^2)|^3}{f'(x^2) \cdot f''(x^2) \cdot f'(x^2)} = -\frac{40^3}{366400} = -0,174672. \end{aligned}$$

Eingesetzt in  $x^{k+1}$ :

$$x^3 = x^2 + s_c^2 = 7,5 - 0,174672 = 7,32533$$

Prüfe nun das Qualitätsmaß  $\rho_2(s_c^2) = \frac{\text{ared}_2(s_c^2)}{\text{pred}_2(s_c^2)}$ . Mit

$$\text{ared}_2(s_c^2) = f(x^2) - f(x^2 + s_c^2) = f(7,5) - f(7,32533) = 3,5991$$

und

$$\text{pred}_2(s_c^2) = -f'(x^2) \cdot s_c^2 - \frac{1}{2} \cdot f''(x^2) \cdot s_c^{2^2} = -40 \cdot (-0,174672) - \frac{1}{2} \cdot 229 \cdot (-0,174672)^2 = 3,49345$$

folgt

$$\rho_2(s_c^2) = \frac{\text{ared}_2(s_c^2)}{\text{pred}_2(s_c^2)} = \frac{3,5991}{3,49345} = 1.03024 > 0,9 = \eta_2.$$

Hier ist, wie in der 1. Iteration, sogar der verschärzte Test erfolgreich, der aussagt, dass die vorhergesagte und die tatsächliche Verminderung gut übereinstimmt. Daher wird auch das Modell  $q_k$  auf einer größeren Kugel vertraut:

$$\Delta_3 = \gamma \cdot \Delta_2 = 2 \cdot 3 = 6$$

Und da  $\rho_2(s_c^2) = 1.03024 > 0,1 = \eta_1$  ist, wird auch der Schritt  $s_c^2$  akzeptiert und es gilt  $x^3 = 7,32533$ .

Die Funktion  $f$  hat bei  $\bar{x} = 7,3166$  das globale Minimum. Es wurde also hier ohne jegliche Voraussetzungen schon in der 3. Iteration die globale Lösung sehr gut angenähert.

### Bemerkung [HB15]

Die Armijo-Schrittweitenregel für das Gradientenverfahren aus Abschnitt 3.3 kann auch als spezieller Fall eines Trust-Region-Verfahrens interpretiert werden: In der aktuellen Iterierten  $x^k$  wird  $f$  durch seine lineare Näherung

$$q_k(s) = f(x^k) + \nabla f(x^k)^T (x - x^k) = f(x^k) + \nabla f(x^k)^T s$$

ersetzt und diese Näherung wird jedoch nur in einer Kugel  $\|x - x^k\| \leq \Delta_k$  vertraut. Dann wird das restringierte Minimierungsproblem

$$f(x^k) + \nabla f(x^k)^T (x - x^k) \rightarrow \min! \quad \text{u. d. N.} \quad x \in B_{\Delta_k}(x^k)$$

wie oben schon erläutert gelöst und es folgt die Lösung  $x^{k+1} = x^k + s_l^k$  mit  $s_l^k = -\frac{\Delta_k \cdot \nabla f(x^k)}{\|\nabla f(x^k)\|}$ .

Um zu entscheiden, ob diese Lösung akzeptiert werden soll, wird die vorhergesagte Abnahme der Zielfunktion (predicted reduction) mit der tatsächlichen Abnahme (actual reduction) verglichen

$$\text{pred}_k(s^k) = f(x^k) - q_k(s^k) = -\nabla f(x^k)^T s^k \quad \text{und} \quad \text{ared}_k(s^k) = f(x^k) - f(x^{k+1}).$$

Erreicht die tatsächliche Abnahme wenigstens einen vorher festgelegten Bruchteil des Vorhergesagten,

$$\frac{\text{ared}_k(s^k)}{\text{pred}_k(s^k)} \geq \gamma$$

so wird der Schritt  $s_l^k$  akzeptiert und die Iteration mit  $x^{k+1}$  weitergeführt. Wird dieser Bruchteil nicht erreicht, so wird der Schritt verworfen und mit einer verkleinerten Trust-Region wiederholt. Wird der Quotient umgeformt

$$\frac{\text{ared}_k(s^k)}{\text{pred}_k(s^k)} = \frac{f(x^k) - f(x^{k+1})}{-\nabla f(x^k)^T s^k} = \frac{f(x^{k+1}) - f(x^k)}{\nabla f(x^k)^T s^k} \geq \gamma$$

und mit  $\nabla f(x^k)^T s^k < 0$  multipliziert, so folgt die Armijo-Regel

$$f(x^{k+1}) - f(x^k) \leq \gamma \nabla f(x^k)^T s^k.$$

## 4.2 Globale Konvergenz

In diesem Abschnitt wird gezeigt, dass Algorithmus 4.4.1 unter den Voraussetzungen 4.1.1 und 4.1.2 global konvergiert. Es wird zunächst überlegt, wann der Algorithmus 4.4.1 wohldefiniert ist. Dies ist offenbar genau dann der Fall, wenn die in der Definition des Quotienten  $\rho_k$  auftretenden Nenner immer von Null verschieden sind. Da  $s^k$  aber das Trust-Region-Teilproblem (4.2) löst und der Nullvektor für dieses Problem zulässig ist, ist zumindest  $f(x^k) - q_k(s^k) = q_k(0) - q_k(s^k) \geq 0$  für alle  $k \in \mathbb{N}$ . Das folgende Lemma besagt insbesondere, dass diese Differenz nur dann gleich Null werden kann, wenn  $x^k$  bereits ein stationärer Punkt von  $f$  wäre, so dass der Algorithmus 4.4.1 in Schritt 1 hätte abbrechen müssen. Also ist der Algorithmus 4.4.1 für jede zweimal stetig differenzierbare Funktion  $f$  tatsächlich wohldefiniert [GK99].

Die Verwendung des Cauchy-Punktes gestattet die folgende Abschätzung des Modellabstiegs  $pred_k(s^k)$ :

**Lemma 4.2.1** *Es gelten die Voraussetzungen 4.1.1 und 4.1.2. Ist dann  $g^k \neq 0$  und genügt  $s^k$  der Cauchy-Abstiegsbedingung (4.4), so gilt:*

$$pred_k(s^k) \geq \frac{\alpha}{2} \|g^k\| \min \left\{ \Delta_k, \frac{\|g^k\|}{C_H} \right\}$$

*Beweis*

a)  $g^{k^T} H_k g^k \leq 0$ . Dann ist  $\tau_k = 1$ ,  $s_c^k = 1 \cdot s_l^k = -\frac{\Delta_k g^k}{\|g^k\|}$  und es folgt

$$\begin{aligned} pred_k(s_c^k) &= q_k(0) - q_k(s_c^k) = f(x^k) - \left( f(x^k) + g^{k^T} s_c^k + \frac{1}{2} s_c^{k^T} H_k s_c^k \right) \\ &= -g^{k^T} \cdot \left( -\frac{\Delta_k g^k}{\|g^k\|} \right) - \frac{1}{2} \left( -\frac{\Delta_k g^k}{\|g^k\|} \right)^T \cdot H_k \cdot \left( -\frac{\Delta_k g^k}{\|g^k\|} \right) \\ &= \frac{\Delta_k \|g^k\|^2}{\|g^k\|} - \underbrace{\frac{1}{2} \frac{\Delta_k^2}{\|g^k\|^2} \cdot g^{k^T} H_k g^k}_{\leq 0} \geq \|g^k\| \Delta_k. \end{aligned}$$

b) Im Fall  $0 < g^{k^T} H_k g^k \leq \frac{\|g^k\|^3}{\Delta_k}$  ergibt sich wieder  $s_c^k = s_l^k = -\frac{\Delta_k g^k}{\|g^k\|}$  mit  $\tau_k = 1$  wegen

$$\frac{\|g^k\|^3}{\Delta_k g^{k^T} H_k g^k} \geq 1 \Rightarrow g^{k^T} H_k g^k \leq \frac{\|g^k\|^3}{\Delta_k}.$$

Dann gilt

$$\begin{aligned} pred_k(s_c^k) &= q_k(0) - q_k(s_c^k) = f(x^k) - \left( f(x^k) + g^{k^T} s_c^k + \frac{1}{2} s_c^{k^T} H_k s_c^k \right) \\ &= -g^{k^T} \cdot \left( -\frac{\Delta_k g^k}{\|g^k\|} \right) - \frac{1}{2} \left( -\frac{\Delta_k g^k}{\|g^k\|} \right)^T \cdot H_k \cdot \left( -\frac{\Delta_k g^k}{\|g^k\|} \right) \\ &= \frac{\Delta_k \|g^k\|^2}{\|g^k\|} - \underbrace{\frac{1}{2} \frac{\Delta_k^2}{\|g^k\|^2} \cdot g^{k^T} H_k g^k}_{\leq \frac{\|g^k\|^3}{\Delta_k}} \\ &\geq \|g^k\| \Delta_k - \frac{1}{2} \frac{\Delta_k^2}{\|g^k\|^2} \cdot \frac{\|g^k\|^3}{\Delta_k} = \|g^k\| \Delta_k - \frac{1}{2} \|g^k\| \Delta_k \\ &= \frac{1}{2} \|g^k\| \Delta_k. \end{aligned}$$

c) Im Fall  $g^{k^T} H_k g^k > \frac{\|g^k\|^3}{\Delta_k}$  mit  $\tau_k = \frac{\|g^k\|^3}{\Delta_k g^{k^T} H_k g^k}$  wegen

$$\frac{\|g^k\|^3}{\Delta_k g^{k^T} H_k g^k} < 1 \Rightarrow g^{k^T} H_k g^k > \frac{\|g^k\|^3}{\Delta_k}$$

gilt für

$$s_c^k = \tau_k \cdot s_l^k = \frac{\|g^k\|^3}{\Delta_k g^{kT} H_k g^k} \cdot \left( -\frac{\Delta_k g^k}{\|g^k\|} \right) = -\frac{\|g^k\|^2}{g^{kT} H_k g^k} g^k$$

und somit

$$\begin{aligned} pred_k(s_c^k) &= q_k(0) - q_k(s_c^k) = f(x^k) - \left( f(x^k) + g^{kT} s_c^k + \frac{1}{2} s_c^{kT} H_k s_c^k \right) \\ &= -g^{kT} \cdot \left( -\frac{\|g^k\|^2}{g^{kT} H_k g^k} g^k \right) \\ &\quad - \frac{1}{2} \left( -\frac{\|g^k\|^2}{g^{kT} H_k g^k} g^k \right)^T \cdot H_k \cdot \left( -\frac{\|g^k\|^2}{g^{kT} H_k g^k} g^k \right) \\ &= \frac{\|g^k\|^4}{g^{kT} H_k g^k} - \frac{1}{2} \frac{\|g^k\|^4}{(g^{kT} H_k g^k)^2} g^{kT} H_k g^k \\ &= \frac{\|g^k\|^4 - \frac{1}{2} \|g^k\|^4}{g^{kT} H_k g^k} = \frac{1}{2} \frac{\|g^k\|^4}{g^{kT} H_k g^k} \\ &\geq \frac{1}{2} \frac{\|g^k\|^4}{\|H_k\| \|g^k\|^2} = \frac{1}{2} \frac{\|g^k\|^2}{\|H_k\|} \geq \frac{1}{2} \frac{\|g^k\|^2}{C_H}. \end{aligned}$$

Hierbei wurde zusätzlich die Voraussetzung 4.1.2 mit  $\|H_k\| \leq C_H$  und der daraus folgende Ungleichung  $\frac{1}{\|H_k\|} \geq \frac{1}{C_H}$  verwendet.

Die Kombination der Ungleichungen liefert

$$pred_k(s_c^k) \geq \frac{1}{2} \|g^k\| \min \left\{ \Delta_k, \frac{\|g^k\|}{C_H} \right\}$$

und unter Beachtung der Cauchy-Abstiegsbedingung (4.4) folgt hieraus

$$pred_k(s^k) \geq \alpha \cdot pred_k(s_c^k) \geq \frac{\alpha}{2} \|g^k\| \min \left\{ \Delta_k, \frac{\|g^k\|}{C_H} \right\}.$$

□

Als Nächstes wird vorgeführt, dass die Suche nach einem erfolgreichen Schritt stets erfolgreich ist. Hierzu wird das folgende Lemma, das auch später noch hilfreich sein wird, benötigt:

**Lemma 4.2.2** *Sei  $f$  stetig differenzierbar und  $x \in \mathbb{R}^n$  ein Punkt mit  $\nabla f(x) \neq 0$ . Dann gibt es zu jedem  $\eta \in (0, 1)$  Konstanten  $\delta = \delta(x, \eta) > 0$  und  $\Delta = \Delta(x, \eta) > 0$ , so dass*

$$\rho_k(s^k) > \eta$$

gilt für alle  $x^k \in \mathbb{R}^n$  mit  $\|x^k - x\| \leq \delta$  und alle  $\Delta_k \in (0, \Delta]$ ,  $s^k \in \mathbb{R}^n$  und  $H_k = H_k^T \in \mathbb{R}^{n \times n}$ , die den Voraussetzungen 4.1.2 und der Bedingung (4.4) genügen.

*Beweis* Nach dem Mittelwertsatz gibt es  $\tau \in [0, 1]$  mit

$$\begin{aligned} pred_k(s^k) - ared_k(s^k) &= f_k - q_k(s^k) - (f_k - f(x^k + \tau s^k)) = f(x^k + \tau s^k) - q_k(s^k) \\ &= (f(x^k) + \nabla f(x^k + \tau s^k) s^k) - \left( f(x^k) + g^{kT} s^k + \frac{1}{2} s^{kT} H_k s^k \right) \\ &= \nabla f(x^k + \tau s^k)^T s^k - g^{kT} s^k - \frac{1}{2} s^{kT} H_k s^k \\ &= s^k \left( \nabla f(x^k + \tau s^k)^T - g^{kT} \right) - \frac{s^{kT} H_k s^k}{2} \\ &\leq \beta \|\nabla f(x^k + \tau s^k) - g^k\| \Delta_k + \frac{\beta^2 C_H}{2} \Delta_k^2, \end{aligned}$$

wobei die Voraussetzungen  $\|s^k\| \leq \beta\Delta_k$  und  $\|H_k\| \leq C_H$ , hier explizit  $-\frac{1}{2}\|H_k\| \leq \frac{1}{2}\|H_k\| \leq \frac{1}{2}C_H$ , verwendet wurden. Weiter ergeben (4.4) und Lemma 4.2.1:

$$pred_k(s^k) \geq \frac{\alpha}{2}\|g^k\| \min\left\{\Delta_k, \frac{\|g^k\|}{C_H}\right\}.$$

Aufgrund der Stetigkeit von  $\nabla f$  kann  $\delta > 0$  so klein gewählt werden, dass  $\|g^k\| \stackrel{(*)}{\geq} \varepsilon := \frac{\|\nabla f(x)\|}{2}$  gilt für alle  $x^k$  mit  $\|x^k - x\| \leq \delta$ . Für  $0 < \Delta \leq \frac{\varepsilon}{C_H}$  und alle  $\Delta_k \leq \Delta$  ergibt sich somit

$$pred_k(s^k) \geq \frac{\alpha}{2}\|g^k\|\Delta_k \stackrel{(*)}{\geq} \frac{\alpha}{2}\varepsilon\Delta_k,$$

denn für  $0 < \Delta \leq \Delta_k \leq \frac{\varepsilon}{C_H} \leq \Delta_k \leq \frac{\|g^k\|}{C_H}$  gilt  $\min\left\{\Delta_k, \frac{\|g^k\|}{C_H}\right\} = \Delta_k$ .

Wegen

$$\|x^k - x\| \leq \delta, \quad \|x^k + \tau s^k - x\| \leq \|x^k - x\| + \|s^k\| \leq \delta + \beta\Delta_k \leq \delta + \beta\Delta$$

konvergieren für  $\delta + \Delta \rightarrow 0$  sowohl  $g^k = \nabla f(x^k)$  als auch  $\nabla f(x^k + \tau s^k)$  gegen  $\nabla f(x)$ .

Wird also  $\delta$  und  $\Delta$  hinreichend klein gewählt, so gilt:

$$\beta\|\nabla f(x^k + \tau s^k) - g^k\| + \frac{\beta^2 C_H}{2}\Delta_k < (1 - \eta)\frac{\alpha}{2}\varepsilon.$$

Es ergibt sich dann

$$\begin{aligned} \rho_k(s^k) &= \frac{ared_k(s^k)}{pred_k(s^k)} = \frac{pred_k - pred_k + ared_k}{pred_k} = \frac{pred_k}{pred_k} - \frac{pred_k - ared_k}{pred_k} \\ &= 1 - \frac{pred_k - ared_k}{pred_k} > 1 - \frac{(1 - \eta)\frac{\alpha}{2}\varepsilon\Delta_k}{\frac{\alpha}{2}\varepsilon\Delta_k} = \frac{\frac{\alpha}{2}\varepsilon\Delta_k - (1 - \eta)\frac{\alpha}{2}\varepsilon\Delta_k}{\frac{\alpha}{2}\varepsilon\Delta_k} = \frac{\eta\frac{\alpha}{2}\varepsilon\Delta_k}{\frac{\alpha}{2}\varepsilon\Delta_k} = \eta. \end{aligned}$$

□

Daraus ergibt sich unmittelbar:

**Korollar 4.2.1** Algorithmus 4.4.1 terminiere nicht endlich und es gelten die Voraussetzungen 4.1.1 und 4.1.2. Dann erzeugt der Algorithmus unendlich viele erfolgreiche Schritte.

*Beweis* Angenommen, die Aussage ist falsch und es gibt nur endlich viele erfolgreiche Iterationsschritte. Dann existiert ein Index  $l \geq 0$  mit  $x^k = x^l$  und  $\rho_k(s^k) \leq \eta_1$  für alle  $k \geq l$ . Weiter folgt

$$\Delta_k \leq \gamma_1\Delta_{k-1} \leq \dots \gamma_1^{k-l}\Delta_l \rightarrow 0 \quad k \rightarrow \infty.$$

Da  $g_l = \nabla f(x^l) \neq 0$  gilt, kann das Lemma 4.2.2 mit  $x = x^l$  sowie  $\eta = \eta_1$  angewendet werden und es gibt somit ein  $\Delta > 0$ , so dass  $\rho_k(s^k) > \eta_1$  gilt für alle  $k \geq l$  mit  $\Delta_k \leq \Delta$  (beachte  $x^k = x^l$  für alle  $k \geq l$ ). Wegen  $\Delta_k \rightarrow 0$  existiert ein kleinstes solches  $k$ . Der Schritt  $s^k$  wäre dann erfolgreich, im Widerspruch zur Annahme.

□

Als Nächstes wird folgendes untersucht:

**Lemma 4.2.3** Es gelte der Algorithmus 4.4.1 unter den Voraussetzungen 4.1.1 und 4.1.2. Ist dann  $\mathcal{K} \subset \mathcal{S}$  eine unendliche Menge mit  $\|g^k\| = \|\nabla f(x^k)\| \geq \varepsilon > 0$  für alle  $k \in \mathcal{K}$ , dann gilt

$$\sum_{k \in \mathcal{K}} \Delta_k < \infty.$$

*Beweis* Für alle  $k \in \mathcal{K} \subset \mathcal{S}$  ist der Schritt  $s^k$  erfolgreich und mit

$$\rho_k(s^k) > \eta_1 \Leftrightarrow \frac{ared_k(s^k)}{pred_k(s^k)} > \eta_1 \Leftrightarrow ared_k(s^k) > \eta_1 pred_k(s^k) \Leftrightarrow f(x^k) - f(x^k + s^k) > \eta_1 pred_k(s^k)$$

gilt:

$$\begin{aligned} f(x^k) - f(x^{k+1}) &= \text{ared}_k(s^k) > \eta_1 \text{pred}_k(s^k) \geq \eta_1 \frac{\alpha}{2} \|g^k\| \min \left\{ \Delta_k, \frac{\|g^k\|}{C_H} \right\} \\ &\geq \eta_1 \frac{\alpha}{2} \varepsilon \min \left\{ \Delta_k, \frac{\varepsilon}{C_H} \right\}. \end{aligned}$$

Wegen  $f(x^k) \geq f(x^{k+1})$  für alle  $k$  ergibt sich

$$\begin{aligned} f(x^0) - f(x^k) &= \sum_{l \in \mathcal{S}, l < k} (f(x^l) - f(x^{l+1})) \geq \sum_{l \in \mathcal{K}, l < k} (f(x^l) - f(x^{l+1})) \\ &\geq \eta_1 \frac{\alpha}{2} \varepsilon \sum_{l \in \mathcal{K}, l < k} \min \left\{ \Delta_l, \frac{\varepsilon}{C_H} \right\} =: S_k. \end{aligned}$$

Im Falle  $\sum_{l \in \mathcal{K}} \Delta_l = \infty$  wäre die Folge  $(S_k)$  unbeschränkt und es würde  $f(x^k) \rightarrow -\infty$  gelten, im Widerspruch zur Voraussetzung 4.1.1.

□

Nun ist Dank des Korollars 4.2.1 und des Lemmas 4.2.3 bekannt, wie sich der Algorithmus 4.4.1 beim Erzeugen von unendlich vielen erfolgreichen Schritten verhält. Somit kann die globale Konvergenz des Verfahrens bewiesen werden:

**Satz 4.2.1** Unter den Voraussetzungen 4.1.1 und 4.1.2 terminiert Algorithmus 4.4.1 entweder mit einem stationären Punkt  $x^k$ , oder er erzeugt eine unendliche Folge  $(x^k)$  mit

$$\liminf_{k \rightarrow \infty} \|g^k\| = 0. \quad (4.6)$$

Ist zudem  $\nabla f$  gleichmäßig stetig auf einer Menge  $\Omega \subset \mathbb{R}^n$  mit  $(x^k) \subset \Omega$ , dann gilt sogar:

$$\lim_{k \rightarrow \infty} \|g^k\| = 0. \quad (4.7)$$

*Beweis* Aus Korollar 4.2.1 ist bekannt, dass der Algorithmus wohldefiniert ist und unendlich viele erfolgreiche Schritte erzeugt, falls er nicht endlich terminiert.

Nachweis von (4.6): Angenommen, (4.6) gilt nicht. Dann gibt es ein  $\varepsilon > 0$  mit  $\|g^k\| = \|\nabla f(x^k)\| \geq \varepsilon$  für alle  $k \geq 0$ . Aus Lemma 4.2.3 folgt dann

$$\sum_{k \in \mathcal{S}} \Delta_k < \infty.$$

Insbesondere liefert dies  $\Delta_k \rightarrow 0$  für  $\mathcal{S} \ni k \rightarrow \infty$ . Darüber hinaus gilt für alle  $k > l$

$$\|x^k - x^l\| \leq \sum_{i \in \mathcal{S}, l \leq i < k} \|s^i\| \leq \beta \sum_{i \in \mathcal{S}, l \leq i < k} \Delta_i \leq \beta \sum_{i \in \mathcal{S}, i \geq l} \Delta_i \rightarrow 0 \quad (\text{für } l \rightarrow \infty).$$

Somit ist  $(x^k)$  eine Cauchy-Folge und konvergiert daher gegen einen Grenzwert  $\bar{x}$ . Aus Stetigkeitsgründen gilt  $\|\nabla f(\bar{x})\| \geq \varepsilon$ .

Anwenden von Lemma 4.2.2 mit  $x = \bar{x}$  und  $\eta = \eta_2$  liefert dann (wegen  $x^k \rightarrow \bar{x}$ ) Konstanten  $L > 0$  und  $\Delta > 0$ , so dass  $\rho_k(s^k) > \eta_2$  für alle  $k \geq L$  mit  $\Delta_k \leq \Delta$ .

Es wird nun induktiv gezeigt, dass daraus folgt:

$$\Delta_k \geq \min\{\Delta_L, \gamma_0 \Delta\} \quad \forall k \geq L. \quad (4.8)$$

Im Falle  $k = L$  ist dies klar. Sei nun  $k \geq L$  und es gelte (4.8).

Gilt  $\Delta_k > \Delta$ , so folgt  $\Delta_{k+1} \geq \gamma_0 \Delta_k > \gamma_0 \Delta$ .

Ist andererseits  $\Delta_k \leq \Delta$ , so gilt  $\rho_k(s^k) > \eta_2$  und deshalb  $\Delta_{k+1} \geq \Delta_k \geq \min\{\Delta_L, \gamma_0 \Delta\}$ .

Damit ist (4.8) nachgewiesen, im Widerspruch zu  $\lim_{S \ni k \rightarrow \infty} \Delta_k = 0$ .

Nachweis von (4.7): Sei nun  $\nabla f$  gleichmäßig stetig auf einer Menge  $\Omega \supset (x^k)$ . Wie bereits gezeigt, gibt es unendlich viele erfolgreiche Schritte, und es gilt (4.6).

Angenommen, (4.7) gilt nicht. Dann gibt es  $\varepsilon > 0$  mit  $\|g^k\| \geq 2\varepsilon$  für unendlich viele  $k \in \mathcal{S}$ . Wegen (4.6) gilt aber auch  $\|g^k\| < \varepsilon$  für unendlich viele  $k \in \mathcal{S}$ . Es kann daher aufsteigende Folgen  $(k_i), (l_i) \subset \mathcal{S}$  gewählt werden mit

$$k_1 < l_1 < k_2 < l_2 < \dots, \quad \|g_{k_i}\| \geq 2\varepsilon, \quad \|g^k\| \geq \varepsilon, \quad k \in \mathcal{K}_i, \quad \|g_{l_i}\| < \varepsilon,$$

wobei  $\mathcal{K}_i = \{k_i, \dots, l_i - 1\} \cap \mathcal{S}$  ist. Die Menge  $\mathcal{K} = \bigcup_{i=1}^{\infty} \mathcal{K}_i$  umfasst unendlich viele erfolgreiche Indizes, und es gilt  $\|g^k\| \geq \varepsilon$  für alle  $k \in \mathcal{K}$ . Daher ist Lemma 4.2.3 anwendbar und liefert:

$$\sum_{k \in \mathcal{K}} \Delta_k < \infty.$$

Da  $\mathcal{K}$  die disjunkte Vereinigung der Mengen  $\mathcal{K}_i$  ist, folgt  $\sum_{k \in \mathcal{K}_i} \Delta_k \rightarrow 0$  für  $i \rightarrow \infty$ . Dies ergibt

$$\|x^{l_i} - x^{k_i}\| \leq \sum_{k \in \mathcal{K}_i} \|s^k\| \leq \beta \sum_{k \in \mathcal{K}_i} \Delta_k \rightarrow 0 \quad (\text{für } i \rightarrow \infty).$$

Andererseits gilt aber  $\|g_{l_i} - g_{k_i}\| \geq |\|g_{l_i}\| - \|g_{k_i}\|| > \varepsilon$  für alle  $i$  im Widerspruch zur gleichmäßigen Stetigkeit von  $\nabla f$ .

□

Es wird noch ein weiteres Konvergenzresultat angegeben, das die Wahl  $\Delta_{min} > 0$  erfordert:

**Satz 4.2.2** *Es gelten die Voraussetzungen 4.1.1 und 4.1.2, und es sei  $\Delta_{min} > 0$ . Dann terminiert Algorithmus 4.4.1 entweder mit einem stationären Punkt  $x^k$ , oder er erzeugt eine unendliche Folge  $(x^k)$ , deren Häufungspunkte stationäre Punkte sind.*

*Beweis* Im Falle, dass der Algorithmus nicht terminiert, werden unendlich viele erfolgreiche Schritte durchgeführt. Sei nun  $\bar{x}$  ein Häufungspunkt und  $\mathcal{K} \subset \mathcal{S}$  so, dass  $(x^k)_{\mathcal{K}}$  gegen  $\bar{x}$  konvergiert.

Angenommen, es gilt  $\nabla f(\bar{x}) \neq 0$ . Dann gibt es  $l > 0$  und  $\varepsilon > 0$  mit  $\|g^k\| = \|\nabla f(x^k)\| \geq \varepsilon$  für alle  $k \in \mathcal{K}$  mit  $k \geq l$ . Insbesondere gilt dann nach Lemma 4.2.3

$$\sum_{k \in \mathcal{K}} \Delta_k < \infty.$$

Weiter gibt es nach Lemma 4.2.2 Zahlen  $\delta > 0$  und  $\Delta > 0$  mit  $k \in \mathcal{S}$ , falls  $x^k \in \bar{B}_{\delta}(\bar{x})$  und  $\Delta_k \leq \Delta$ .

Sei nun  $k \in \mathcal{K}$  hinreichend groß. Dann gilt  $x^k \in \bar{B}_{\delta}(\bar{x})$ . Im Falle  $k-1 \in \mathcal{S}$  folgt dann  $\Delta_k \geq \Delta_{min} > 0$ . Ist andererseits  $k-1 \notin \mathcal{S}$ , dann gilt  $x^{k+1} = x^k$  und daher muss  $\Delta_{k-1} > \Delta$  sein (sonst wäre  $s^{k-1}$  erfolgreich). Dies ergibt

$$\Delta_k \geq \gamma_0 \Delta_{k-1} > \gamma_0 \Delta.$$

Für große  $k \in \mathcal{K}$  ist daher  $\Delta_k \geq \min\{\gamma_0 \Delta, \Delta_{min}\}$  im Widerspruch zu  $(\Delta_k)_{\mathcal{K}} \rightarrow 0$ .

□

### 4.3 Charakterisierung der Lösungen des Teilproblems

Es wurde schon mit Hilfe des Cauchy-Punktes eine näherungsweise Lösung für die Optimierung der quadratischen Modellfunktion  $q_k$  erläutert. Nun soll eine nahezu exakte Lösung des Trust-Region Teilproblems (4.2) bestimmt werden. Wie in Kapitel 2.2 Optimalitätsbedingungen schon erläutert, muss ein lokales Minimum  $\bar{x}$  einer zweimal stetig differenzierbaren Funktion den notwendigen Optimalitätsbedingungen erster Ordnung,  $\nabla f(\bar{x}) = 0$  und zweiter Ordnung,  $\nabla^2 f(\bar{x})$  positiv semi definit genügen. Ist die Hesse-Matrix  $\nabla^2 f(\bar{x})$  positiv definit, so ist  $\bar{x}$  sogar ein globales Minimum. Zudem ist es bekannt, dass jede symmetrisch positiv definite Matrix invertierbar (auch regulär genannt) ist. Dementsprechend

hat die Modellfunktion  $q_k$  nur dann eine positive Krümmung, und somit ein Minimum mit der Newton-Suchrichtung  $s_n^k = -H_k^{-1}g^k$ , wenn die Hesse-Matrix  $H_k$  positiv definit ist. Es müssten aber zusätzlich noch die Fälle betrachtet werden, wenn die Hesse-Matrix negativ definit oder indefinit ist.

Zunächst sei bemerkt, dass das Problem (4.2) stets eine Lösung besitzt, da die Zielfunktion  $q_k$  stetig ist und der zulässige Bereich offenbar eine kompakte Menge darstellt. Allerdings handelt es sich im Allgemeinen um ein nichtkonvexes Problem, denn hier wird keine Definitheitsbedingung an die Matrix  $H_k$  gestellt. Denn falls  $H_k$  nicht positiv definit ist, ist  $q_k$  auch nicht mehr gleichmäßig konvex. Ohne eine Konvexitätsvoraussetzung erscheint es zunächst allerdings sehr schwer, eine Lösung, also ein globales Minimum, des Trust-Region Teilproblems (4.2) zu finden [GK99].

Einen Hinweis darauf, wie dies vielleicht doch geschehen kann, liefert der folgende Satz. Es werden notwendige und hinreichende Bedingungen dafür angegeben, dass  $s^k$  eine Lösung des Problems (4.2) ist, und dass es ein globales Minimum besitzt.

#### Satz 4.3.1

1. Das Problem (4.2) besitzt mindestens eine (globale) Lösung.

2. Der Vektor  $s^k \in \mathbb{R}^n$  ist Lösung von (4.2) genau dann, wenn es  $\lambda \in \mathbb{R}$  gibt, so dass gilt:

$$\|s^k\| \leq \Delta_k \quad (4.9)$$

$$\lambda \geq 0, \quad \lambda(\|s^k\| - \Delta_k) = 0 \quad (4.10)$$

$$(H_k + \lambda I)s^k = -g^k \quad (4.11)$$

$$H_k + \lambda I \text{ ist prositiv semidefinit} \quad (4.12)$$

3. Gilt (4.9) - (4.11) und ist die Matrix  $H_k + \lambda I$  positiv definit, so ist  $s^k$  die eindeutige Lösung von (4.2).

*Beweis* zu 1.: Die Trust-Region ist kompakt und die Funktion  $q^k$  ist stetig. Somit besitzt (4.2) eine Lösung.

zu 2. " $\Rightarrow$ ": Sei  $s^k$  globale Lösung von (4.2). Setze  $y^k = \nabla q_k(s^k) = H_k s^k + g^k$ . Offensichtlich ist (4.9) erfüllt.

zu (4.10) und (4.11): Im Falle  $\|s^k\| < \Delta_k$  ist  $s^k$  lokales Minimum von  $q_k$  und daher gelten (4.10) und (4.11) mit  $\lambda = 0$ . Nebenbei bemerkt:  $q_k$  ist dann also konvex und  $s^k$  ist globales Minimum von  $q_k$  auf  $\mathbb{R}^n$ . Sei nun  $\|s^k\| = \Delta_k$ . Angenommen, es gibt kein  $\lambda$ , für das (4.10) und (4.11) gelten. Dann gilt  $y^k \neq 0$  und  $\alpha_k := \angle(y^k, s^k) \neq \pi$ , also

$$\cos(\alpha_k) = \frac{y^{k^T} s^k}{\|y^k\| \|s^k\|} > -1.$$

Setze  $v^k = -\frac{y^k}{\|y^k\|} - \frac{s^k}{\|s^k\|}$  (Winkelhalbierende zwischen  $-y^k$  und  $-s^k$ ) und berechne

$$y^{k^T} v^k = -\frac{y^{k^T} y^k}{\|y^k\|} - \frac{y^{k^T} s^k}{\|s^k\|} = -\|y^k\|(1 + \cos(\alpha_k)) < 0.$$

Somit ist  $v^k$  eine Abstiegsrichtung für  $q_k$  im Punkt  $s^k$ . Weiter gilt

$$\left[ \frac{d}{dt} \frac{1}{2} \|s^k + tv^k\|^2 \right]_{t=0} = v^{k^T} s^k = -\left( \frac{y^{k^T} s^k}{\|y^k\|} + \|s^k\| \right) = -\|s^k\|(\cos(\alpha_k) + 1) < 0.$$

Damit ergibt sich  $\|s^k + tv^k\| < \|s^k\| \leq \Delta_k$  für kleine  $t > 0$ . Da  $v^k$  eine Abstiegsrichtung ist, folgt somit ein Widerspruch zur Optimalität von  $s^k$ .

zu (4.12): Es genügt zu zeigen, dass

$$d^T (H_k + \lambda I)d \geq 0 \quad \forall d \in \mathbb{R}^n \text{ mit } d^T s^k < 0$$

gilt, da es auf das Vorzeichen von  $d$  nicht ankommt und der Fall  $d^T s^k = 0$  aus Stetigkeitsgründen folgt (denn ist  $d^T s^k = 0$  und  $d^T (H_k + \lambda I)d < 0$  so gilt für  $d(t) = d - ts^k$  und kleine  $t > 0$ :  $d(t)^T (H_k + \lambda I)d(t) < 0$

sowie  $d(t)^T s^k < 0$ .

Betrachte ein beliebiges  $d$  mit  $d^T s^k < 0$  und setze  $t = -\frac{2d^T s^k}{\|d\|^2} > 0$ . Dann gilt

$$\begin{aligned}\|s^k + td\|^2 &= \|s^k\|^2 + 2ts^{k^T} d + t^2\|d\|^2 = \|s^k\|^2 + 2\left(-\frac{2d^T s^k}{\|d\|^2}\right)s^{k^T} d + \left(-\frac{2d^T s^k}{\|d\|^2}\right)^2\|d\|^2 \\ &= \|s^k\|^2 - \frac{4\|s^k\|^2\|d\|^2}{\|d\|^2} + \frac{4\|s^k\|^2\|d\|^2}{\|d\|^4}\|d\|^2 = \|s^k\|^2 \leq \Delta_k^2\end{aligned}$$

und mit der Definition von  $t$  und mit (4.11) und deren Umformungen

$$H_k s^k + \lambda s^k = -g^k \Leftrightarrow (\spadesuit) -\lambda s^k = H_k s^k + g^k = y^k \text{ und } t = -\frac{2d^T s^k}{\|d\|^2} \Leftrightarrow (\clubsuit) s^k = -\frac{t\|d\|^2}{2d^T}$$

folgt

$$\begin{aligned}0 &\leq q_k(s^k + td) - q_k(s^k) = f_k + y^{k^T}(s^k + td) + \frac{1}{2}(s^k + td)^T H_k(s^k + td) - f_k - y^{k^T} s^k - \frac{1}{2}s^{k^T} H_k s^k \\ &= y^{k^T}(s^k + td - s^k) + \frac{1}{2}(s^k + td - s^k)^T H_k(s^k + td - s^k) = ty^{k^T} d + \frac{t^2}{2}d^T H_k d \\ (\clubsuit) \quad &-t\lambda s^{k^T} d + \frac{t^2}{2}d^T H_k d \stackrel{(\clubsuit)}{=} \frac{t^2}{2}\lambda\|d\|^2 + \frac{t^2}{2}d^T H_k d = \frac{t^2}{2}d^T(H_k + \lambda I)d.\end{aligned}$$

Damit ist auch (4.12) nachgewiesen.

zu 2. " $\Leftarrow$ ":

Für  $s^k \in \mathbb{R}^n$  und  $\lambda \in \mathbb{R}$  seien (4.9) - (4.12) erfüllt.

Sei  $h \in \mathbb{R}^n$  beliebig mit  $\|h\| \leq \Delta_k$  und  $d = h - s^k$ . Sei wie oben  $y^k = H_k s^k + g^k$  und mit gleichen Umformungen folgt

$$\begin{aligned}q_k(h) - q_k(s^k) &= y^{k^T} d + \frac{1}{2}d^T H_k d = -\lambda s^{k^T} d + \frac{1}{2}d^T H_k d \geq -\lambda s^{k^T} d - \frac{1}{2}\lambda\|d\|^2 \\ &= -\frac{\lambda}{2}(2s^{k^T} d + \|d\|^2) = -\frac{\lambda}{2}(\|d + s^k\|^2 - \|s^k\|^2) \\ &= -\frac{\lambda}{2}(\|h\|^2 - \|s^k\|^2) \stackrel{(*)}{\geq} 0.\end{aligned}\tag{4.13}$$

Die erste Abschätzung folgt durch (14.58):  $\frac{1}{2}d^T H_k d = \underbrace{\frac{1}{2}d^T(H_k + \lambda I)d}_{\geq 0} - \frac{1}{2}d^T \lambda d \geq -\frac{1}{2}\lambda\|d\|^2$ . Die Abschätzung (\*) ist für  $\lambda = 0$  klar und im Falle  $\lambda > 0$  folgt  $\|h\| \leq \Delta_k = \|s^k\|$  und (\*) gilt wiederum. Damit ist  $s$  Lösung von (4.2).

zu 3.: Es gelte (4.9) - (4.11) und  $H_k + \lambda I$  sei positiv definit, daraus folgt dann in (4.13):

$$q_k(h) - q_k(s^k) = \dots = -\lambda s^{k^T} d + \frac{1}{2}d^T H_k d > -\lambda s^{k^T} d - \frac{1}{2}\lambda\|d\|^2 = \dots \geq 0.$$

□

### Bemerkung:

Die Aussage, dass (4.9) - (4.12) notwendig für die Optimalität von  $s^k$  sind, könnte auch durch die Karush-Kuhn-Tucker-Optimalitätstheorie nachgewiesen werden. Die KKT-Bedingungen sind ein notwendiges Optimalitätskriterium erster Ordnung für restringierte Optimierungsprobleme und es wird im Rahmen dieser Arbeit nicht detaillierter darauf eingegangen.

Es folgen nun Abbildungen, die die Modellfunktion  $q_k$  je nach der Definitheit von  $H_k$  darstellen. Zudem wird im Verlauf auch das Regularisierungsverfahren behandelt [CA07].

In Abbildung 4.5 hat das Modell  $q_k$  eine positive Krümmung und somit auch eine positiv definite  $H_k$ . Es ist zu erkennen, dass das Minimum der quadratischen Modellfunktion  $q_k$  im Trust-Region Radius liegt. Liegt das Minimum im oder am Rand des Vertrauensbereichs, so wird das Trust-Region-Verfahren mit dem Newton-Schritt gelöst.

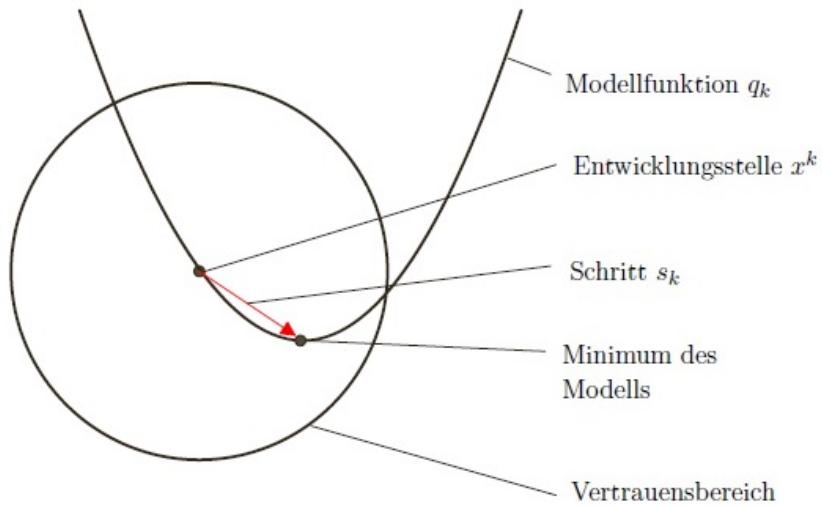


Abbildung 4.5: Positiv definite  $H_k$  und Minimum im Vertrauensbereich

In Abbildung 4.6 liegt zwar eine positive Krümmung, und somit auch eine positiv definite  $H_k$ , vor, doch das Minimum des Modells liegt nicht im Trust-Region Radius. Daher ist hier auch nur noch approximative Lösungen möglich.

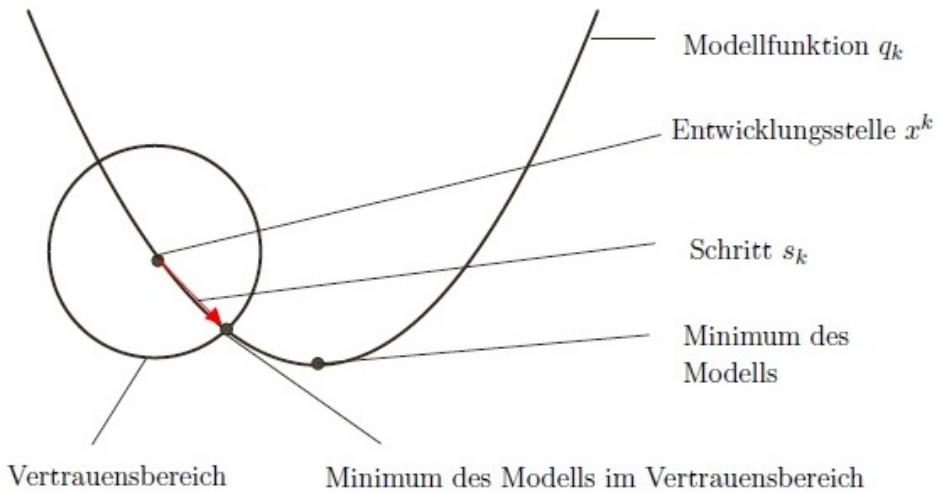


Abbildung 4.6: Positiv definite  $H_k$  und Minimum außerhalb des Vertrauensbereiches

Ist  $H_k$  negativ definit oder indefinit, so wie auf dem Bild 4.7, so hat die Modellfunktion  $q_k$  infolgedessen eine negative Krümmung.

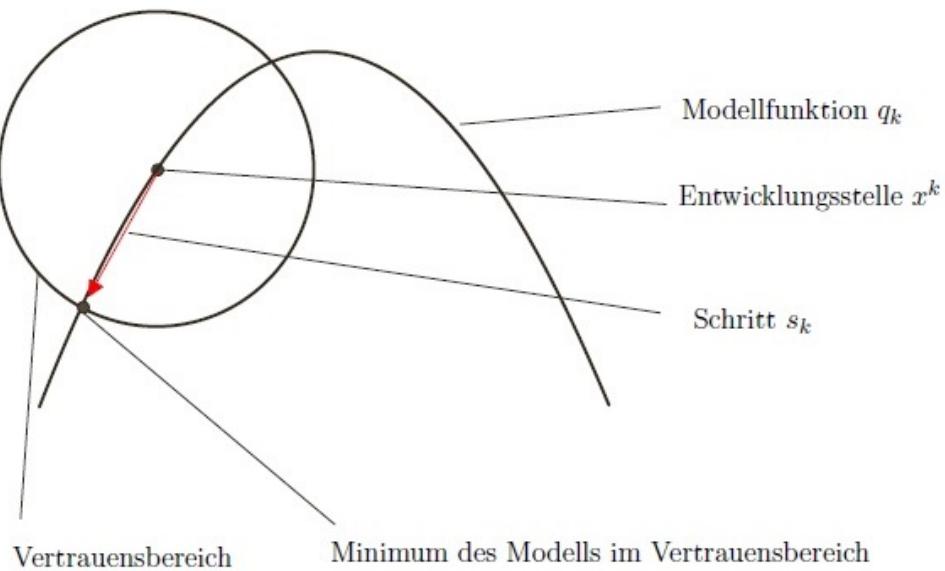


Abbildung 4.7: Negativ definite oder indefinite  $H_k$

Es folgt nun die geometrische Darstellung des Satzes 4.3.1. Dabei wird mit Hilfe der Regularisierungsparameter  $\lambda$  die negativ definite oder indefinite Hesse-Matrix  $H_k$  reguliert, sprich positiv definit gemacht. Somit existiert mit dieser Regularisierungsmethode eine Hesse-Matrix  $H_k + \lambda I$ , welches das Trust-Region Teilproblem (4.2) nahezu exakt mit einem globalen Minimum löst.

Die Abbildung 4.8 zeigt die Wirkung von  $\lambda$  auf das quadratische Modellfunktion  $q_k$  mit einer zwar positiven Krümmung, aber ein Minimum außerhalb des Vertrauensbereichs.

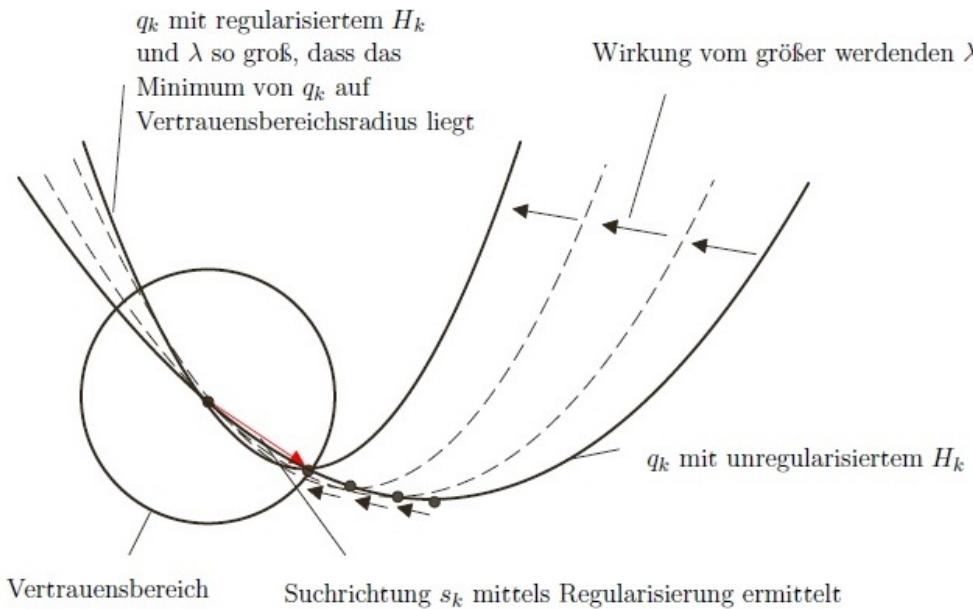


Abbildung 4.8: Die Wirkung der Parameter auf positiv definite  $H_k$

In Abbildung 4.9 wird wieder die Wirkung der Regularisierungsparameter auf das Modell  $q_k$  dargestellt. Hier hat das Modell aber eine negative Krümmung und somit auch negativ definite oder indefinite  $H_k$ .

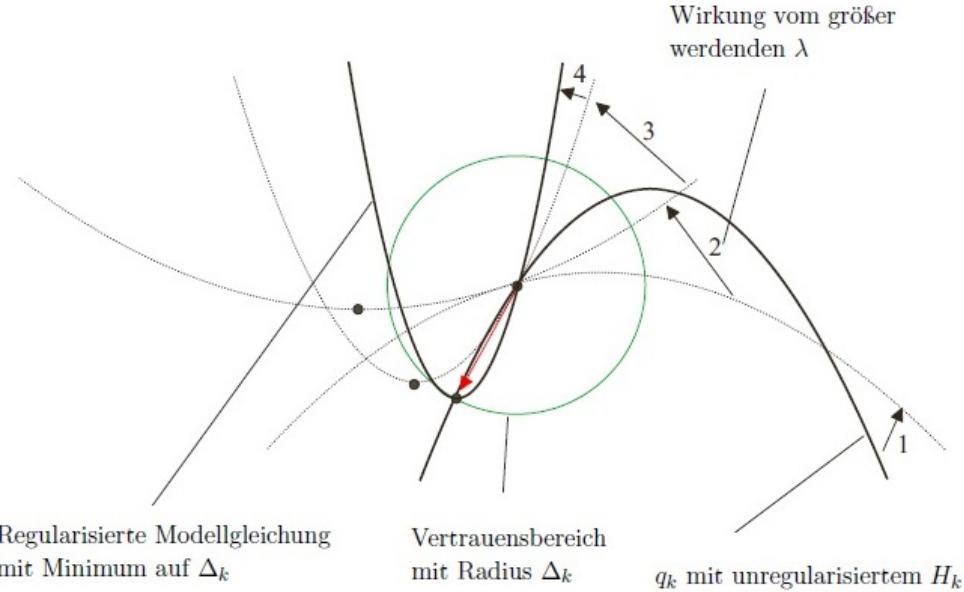


Abbildung 4.9: Die Wirkung der Parameter auf negativ definite oder indefinite  $H_k$

## 4.4 Schnelle lokale Konvergenz

Es wird nun untersucht, unter welchen Voraussetzungen schnelle Konvergenz zu erwarten ist. Wie üblich wird hierzu das Verhalten des Verfahrens nahe eines stationären Punktes  $\bar{x}$  betrachtet, der die hinreichende Optimalitätsbedingungen 2. Ordnung erfüllt, d.h.  $\nabla f(\bar{x}) = 0$ ,  $\nabla^2 f(\bar{x})$  ist positiv definit. Es wird sich hier nur auf den Fall beschränkt, dass die exakte Hesse-Matrix für das Modell herangezogen wird, und stets den Newton-Schritt verwendet, falls dieser wohldefiniert ist und der Cauchy-Abstiegsbedingung genügt:

**Algorithmus 4.4.1 Trust-Region-Newton-Verfahren.** Algorithmus 4.4.1, wobei die Schritte 2 und 3 wie folgt implementiert sind:

2. Berechne  $H_k = \nabla^2 f(x^k)$
3. Falls der Newton-Schritt  $s_n^k = -H_k^{-1}g^k$  existiert und die Cauchy-Abstiegsbedingung (4.4) erfüllt, so wähle  $s^k = s_n^k$ . Sonst bestimme einen Schritt  $s^k$ , der (4.4) genügt.

**Beispiel 4.4.1** Sei

$$f(x_1, x_2) = 2x_1^2 + x_2^2 - 3$$

mit

$$\nabla f(x_1, x_2) = \begin{pmatrix} 4x_1 \\ 2x_2 \end{pmatrix} \quad \text{und} \quad \nabla^2 f(x_1, x_2) = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$$

die zu minimierende mehrdimensionale Funktion. Die Startwerte seien  $x^0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  und  $\Delta_0 = 2$ . Die Parameter sind  $\eta_1 = 0,1$ ,  $\eta_2 = 0,9$ ,  $\gamma_0 = 0,25$ ,  $\gamma_1 = 0,75$ ,  $\gamma_2 = 2$  und zusätzlich die Parameter  $\alpha = 0,5$  und  $\beta = 1$  für die Cauchy-Abstiegsbedingung (4.4). Es wird nun mit dem Cauchy-Schritt  $s_c^k$  und dem Newton-Schritt  $s_n^k$  die Cauchy-Abstiegsbedingung geprüft, und dann entschieden, welche der beiden Schritte besser geeignet ist. Unabhängig der Schrittauswahl gilt für die folgende Berechnung

$$pred_k(s^k) = q_k(0) - q_k(s^k) = f_k - q_k(s^k) = f_k - \left( f_k + g^{kT} s + \frac{1}{2} s^T H_k s \right) = -g^{kT} s - \frac{1}{2} s^T H_k s,$$

$$g^0 = \nabla f(x_1^0, x_2^0) = \begin{pmatrix} 4 \\ 2 \end{pmatrix} \quad \text{und} \quad H_0 = \nabla^2 f(x_1^0, x_2^0) = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}.$$

• **Der Cauchy-Schritt  $s_c^k$ :**

Prüfe zuerst, ob  $g^k^T H_k g^k > 0$  oder  $\leq 0$  ist:

$$(4; 2) \cdot \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 2 \end{pmatrix} = (4; 2) \cdot \begin{pmatrix} 16 \\ 4 \end{pmatrix} = 64 + 8 = 72 > 0$$

Dann wird zusätzlich geprüft, ob

$$\frac{\|g^k\|^3}{\Delta_k g^k^T H_k g^k} < 1 \text{ oder } \geq 1$$

ist. Es gilt

$$\frac{\|g^0\|^3}{\Delta_0 g^0^T H_0 g^0} = \frac{\left\| \begin{pmatrix} 4 \\ 2 \end{pmatrix} \right\|^3}{2 \cdot 72} = \frac{(\sqrt{4^2 + 2^2})^3}{144} = \frac{5 \cdot \sqrt{5}}{18} = 0,6211 < 1$$

also wird  $\tau_0 = \frac{\|g^0\|^3}{\Delta_0 g^0^T H_0 g^0}$  gewählt. Der Cauchy-Schritt ist nun

$$s_c^0 = \tau_0 \cdot s_l^0 = \frac{\|g^0\|^3}{\Delta_0 g^0^T H_0 g^0} \cdot -\frac{g^0}{\|g^0\|} \Delta_0 = -\frac{\|g^0\|^2}{g^0^T H_0 g^0} \cdot g^0 = -\frac{(\sqrt{4^2 + 2^2})^2}{72} \cdot \begin{pmatrix} 4 \\ 2 \end{pmatrix} = \begin{pmatrix} -\frac{10}{9} \\ -\frac{5}{9} \end{pmatrix}.$$

Dabei gilt für den Cauchy-Schritt

$$\|s_c^0\| = \sqrt{\left(-\frac{10}{9}\right)^2 + \left(-\frac{5}{9}\right)^2} = \frac{5 \cdot \sqrt{5}}{9} = 1,2423 \leq 2 = 1 \cdot 2 = \beta \cdot \Delta_0$$

und

$$\begin{aligned} pred_0(s_c^0) &= -(4; 2) \cdot \begin{pmatrix} -\frac{10}{9} \\ -\frac{5}{9} \end{pmatrix} - \frac{1}{2} \left( -\frac{10}{9}; -\frac{5}{9} \right) \cdot \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} -\frac{10}{9} \\ -\frac{5}{9} \end{pmatrix} \\ &= \frac{40}{9} + \frac{10}{9} - \frac{1}{2} \left( -\frac{10}{9}; -\frac{5}{9} \right) \cdot \begin{pmatrix} -\frac{40}{9} \\ -\frac{10}{9} \end{pmatrix} = \frac{50}{9} - \frac{25}{9} = \frac{25}{9} = 2, \bar{7}. \end{aligned}$$

• **Der Newton-Schritt  $s_n^k$ :**

Für den Newton-Schritt wird zusätzlich die Inverse der  $H_k$  benötigt. Die Inverse von  $H_k$  ist zu berechnen, da hier wie oben schon berechnet  $g^k^T H_k g^k > 0$  gilt, kann  $H_k$  invertiert werden:

$$H_0^{-1} = \frac{1}{4 \cdot 2 - 0 \cdot 0} \cdot \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} = \frac{1}{8} \cdot \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}.$$

Der Newton-Schritt ist

$$s_n^0 = -H_0^{-1} g^0 = -\frac{1}{8} \cdot \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 2 \end{pmatrix} = -\frac{1}{8} \cdot \begin{pmatrix} 8 \\ 8 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix},$$

es gilt zusätzlich

$$\|s_n^0\| = \sqrt{(-1)^2 + (-1)^2} = \sqrt{2} = 1,41 \leq 2 = \Delta_0$$

und

$$\begin{aligned} pred_0(s_n^0) &= -(4; 2) \cdot \begin{pmatrix} -1 \\ -1 \end{pmatrix} - \frac{1}{2}(-1; -1) \cdot \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ -1 \end{pmatrix} \\ &= -(-6) - \frac{1}{2}(-1; -1) \cdot \begin{pmatrix} -4 \\ -2 \end{pmatrix} = 6 - \frac{1}{2} \cdot 6 = 3. \end{aligned}$$

Nun wird die Cauchy-Abstiegsbedingung (4.4)  $\text{pred}_0(s_n^0) \stackrel{?}{\geq} \alpha \cdot \text{pred}_0(s_c^0)$  geprüft:

$$\text{pred}_0(s_n^0) = 3 \geq 1,38 = 0,5 \cdot \frac{25}{9} = \alpha \cdot \text{pred}_0(s_c^0)$$

Es wird hier erkannt, dass beim Newton-Schritt eine bessere Abnahme der Modellfunktion  $q_k$  erreicht wird. Aus diesem Grund wird auch der Newton-Schritt gewählt und es gilt  $x^1 = x^0 + s_n^0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ . Der Trust-Region Algorithmus wird dann weiter verfahren:

$$\text{ared}_0(s_n^0) = f(x^0) - f(x^0 + s_n^0) = (2 \cdot (1)^2 + (1)^2 - 3) - (2 \cdot (0)^2 + (0)^2 - 3) = 0 - (-3) = 3$$

Zuletzt wird die

$$\rho_0(s_n^0) = \frac{\text{ared}_0(s_n^0)}{\text{pred}_0(s_n^0)} = \frac{3}{3} = 1 > 0,9 = \eta_2$$

berechnet. Somit ist die  $f$ -Abnahme befriedigend im Vergleich zur Modellabnahme, daher gilt auch

$$\Delta_0 = \gamma_2 \cdot \Delta_0 = 2 \cdot 2 = 4.$$

Und da  $\rho_0(s_n^0) = 1 > 0,1 = \eta_1$  gilt, wird der Schritt  $s_n^0$  akzeptiert und es gilt  $x^1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ . Bei der nächsten Iteration gilt aber schon

$$g^1 = \nabla f(x^1) = \begin{pmatrix} 4 \cdot 0 \\ 2 \cdot 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Somit wäre das Abbruchkriterium erfüllt und die Minimalstelle lautet:  $x^1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ . Es hat also nur einen Schritt mit der Newton-Richtung gereicht, um das Minimum der Funktion zu finden.

Es folgen nun fünf Hilfsresultate - 4.4.1, 4.4.2, 4.4.2, 4.4.1, 4.4.3 - die später für die Gültigkeit des Satzes 4.4.2 benötigt werden. Für weitergehenden diesbezüglichen Ausführungen wird auf Abschnitt 10 in [UU12] verwiesen.

**Lemma 4.4.1** Sei  $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  stetig differenzierbar. Weiter sei  $\nabla f(\bar{x}) = 0$  und  $\nabla^2 f(\bar{x})$  sei invertierbar. Dann gibt es  $\varepsilon > 0$  und  $\gamma > 0$  mit

$$\|\nabla f(x)\| \geq \gamma \|x - \bar{x}\| \quad \forall x \in B_\varepsilon(\bar{x}).$$

Insbesondere ist  $\bar{x}$  eine isolierte Nullstelle von  $\nabla f$ .

**Algorithmus 4.4.2 Lokales Newton-Verfahren für Optimierungsprobleme.**

0. Wähle einen Startpunkt  $x^0 \in \mathbb{R}^n$ .

Für  $k = 0, 1, 2, \dots$ :

1. STOP, falls  $\nabla f(x^k) = 0$ .

2. Berechne den Newton-Schritt  $s^k \in \mathbb{R}^n$  durch Lösen der Newton-Gleichung

$$\nabla^2 f(x^k) s^k = -\nabla f(x^k).$$

3. Setze  $x^{k+1} = x^k + s^k$ .

**Lemma 4.4.2** Sei  $A \in \mathbb{R}^{n \times n}$  symmetrisch und positiv definit. Dann gilt für alle  $\mu \in (0, \lambda_{\min}(A))$  und alle symmetrischen Matrizen  $B \in \mathbb{R}^{n \times n}$  mit  $\|B\| \leq \lambda_{\min}(A) - \mu$ :

$$\lambda_{\min}(A + B) \geq \mu.$$

**Satz 4.4.1 Lokale Konvergenz des Newton-Verfahrens für Optimierungsprobleme.**

Sei  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  zweimal stetig differenzierbar und  $\bar{x} \in \mathbb{R}^n$  ein lokales Minimum von  $f$ , in dem die hinreichenden Bedingungen 2. Ordnung gelten. Dann gibt es  $\delta > 0$  und  $\mu > 0$ , so dass gilt:

1.  $\bar{x}$  ist der einzige stationäre Punkt auf  $B_\delta(\bar{x})$ .
2.  $\lambda_{\min}(\nabla^2 f(x)) \geq \mu$  für alle  $x \in B_\delta(\bar{x})$ .
3. Für alle  $x^0 \in B_\delta(\bar{x})$  terminiert Algorithmus 4.4.2 entweder mit  $x^k = \bar{x}$ , oder er erzeugt eine Folge  $(x^k) \subset B_\delta(\bar{x})$ , die  $q$ -superlinear gegen  $\bar{x}$  konvergiert.
4. Ist  $\nabla^2 f$  Lipschitz-stetig auf  $B_\delta(\bar{x})$  mit Konstante  $L$ , d.h. gilt

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \quad \forall x, y \in B_\delta(\bar{x}),$$

so ist die Konvergenzrate (falls das Verfahren nicht endlich terminiert) sogar  $q$ -quadratisch:

$$\|x^{k+1} - \bar{x}\| \leq \frac{L}{2\mu} \|x^k - \bar{x}\|^2 \quad \forall k \geq 0.$$

**Lemma 4.4.3** Sei  $\bar{x}$  ein isolierter Häufungspunkt der Folge  $(x^k) \subset \mathbb{R}^n$ . Für jede gegen  $\bar{x}$  konvergente Teilfolge  $(x^k)_K$  gelte  $(x^{k+1} - x^k)_K \rightarrow 0$ . Dann konvergiert die gesamte Folge  $(x^k)$  gegen  $\bar{x}$ .

Die Hilfsresultate werden nun für den Beweis des folgenden Satzes verwendet.

**Satz 4.4.2** Die Funktion  $f$  sei zweimal stetig differenzierbar und die Niveaumenge

$$N_0 = \{x; f(x) \leq f(x^0)\}$$

sei kompakt. Algorithmus 4.4.1 erzeuge eine Folge, die einen Häufungspunkt  $\bar{x}$  besitzt, in dem die Hesse-Matrix  $\nabla^2 f(\bar{x})$  positiv definit ist. Dann gilt:

- a)  $x^k \rightarrow \bar{x}$ ,  $g^k \rightarrow 0$  ( $k \rightarrow \infty$ ),  $\nabla f(\bar{x}) = 0$ .
- b) Es gibt  $l \geq 0$ , so dass für alle  $k \geq l$  jeder Schritt  $s^k$  erfolgreich ist und  $\Delta_k \geq \Delta_l > 0$  gilt.
- c) Es gibt  $l' \geq l$  mit  $s^k = s_n^k$  für alle  $k \geq l'$  und all diese Schritte sind erfolgreich. Algorithmus 4.4.1 geht also für  $k \geq l'$  über in das Newton-Verfahren. Insbesondere konvergiert  $(x^k)$   $q$ -superlinear gegen  $\bar{x}$  und sogar  $q$ -quadratisch, falls  $\nabla^2 f$  Lipschitz-stetig in einer Umgebung von  $\bar{x}$  ist.

*Beweis* Die stetige Funktion  $f$  ist auf der kompakten Niveaumenge  $N_0$  nach unten beschränkt. Somit gilt die Voraussetzung 4.1.1.

Nach der Konstruktion des Verfahrens gilt  $f_{k+1} \leq f_k$  für alle  $k$  und somit  $(x^k) \subset N_0$ . Da die stetige Funktion  $\nabla^2 f$  auf dem Kompaktum  $N_0$  beschränkt ist, gilt wegen  $H_k = \nabla^2 f(x^k)$  auch die Voraussetzung 4.1.2.

Schließlich ist die stetige Funktion  $\nabla f$  gleichmäßig stetig auf dem Kompaktum  $N_0$ .

Somit ist Satz 4.2.1 anwendbar und liefert

$$\lim_{k \rightarrow \infty} g^k = 0.$$

Da  $\bar{x}$  ein Häufungspunkt von  $(x^k)$  ist, folgt  $\nabla f(\bar{x}) = 0$  wegen der Stetigkeit von  $\nabla f$ . Da  $\nabla^2 f(\bar{x})$  positiv definit ist, gibt es wegen Lemma 4.4.2  $\varepsilon > 0$  und  $\mu > 0$ , so dass gilt:

$$d^T \nabla^2 f(x) d \geq \mu \|d\|^2 \quad \forall x \in \bar{B}_\varepsilon(\bar{x}), \quad \forall d \in \mathbb{R}^n.$$

Somit gilt wegen Lemma 4.4.1

$$\nabla f(x) \neq 0 \text{ für alle } x \in \bar{B}_\varepsilon(\bar{x}) \setminus \{\bar{x}\}. \tag{4.14}$$

Wegen  $g^k \rightarrow 0$  zeigt dies, dass  $\bar{x}$  der einzige Häufungspunkt von  $(x^k)$  in  $\bar{B}_\varepsilon(\bar{x})$  ist. Weiter gilt mit

$$q_k(s^k) \leq q_k(0) \Leftrightarrow 0 \leq q_k(0) - q_k(s^k) = pred_k(s^k) \Leftrightarrow 0 > -pred_k(s^k)$$

für alle  $x^k \in \bar{B}_\varepsilon(\bar{x})$

$$\begin{aligned} 0 &> -pred_k(s^k) = f_k + g^{kT} + \frac{1}{2}s^{kT}H_k s^k - f_k = g^{kT}s^k + \frac{1}{2}s^{kT}H_k s^k \\ &\geq g^{kT}s^k + \frac{1}{2}\mu\|s^k\|^2 \geq (-\|g^k\| + \frac{\mu}{2}\|s^k\|)\|s^k\|, \end{aligned}$$

also  $0 > (-\|g^k\| + \frac{\mu}{2}\|s^k\|)\|s^k\| \Leftrightarrow 0 > (-\|g^k\| + \frac{\mu}{2}\|s^k\|)$  und somit

$$0 < \frac{\mu}{2}\|s^k\| < \|g^k\|. \quad (4.15)$$

Ist  $(x^k)_\mathcal{K}$  eine Teilfolge mit  $(x^k)_\mathcal{K} \rightarrow \bar{x}$ , so folgt  $(g^k)_\mathcal{K} \rightarrow 0$  und daher wegen (4.15) auch  $(s^k)_\mathcal{K} \rightarrow 0$ . Da stets  $x^{k+1} - x^k \in \{0, s^k\}$  gilt, wird hieraus  $(x^{k+1} - x^k)_\mathcal{K} \rightarrow 0$  gefolgert. Damit liefert Lemma 4.4.3:

$$\lim_{k \rightarrow \infty} x^k = \bar{x}.$$

b) Es wird nun gezeigt, dass  $\rho_k(s^k) \rightarrow 1$  ( $k \rightarrow \infty$ ) gilt. Nach a) bleibt  $(x^k)$  schließlich in  $\bar{B}_\varepsilon(\bar{x})$ . Sei also  $x^k \in \bar{B}_\varepsilon(\bar{x})$  und wegen  $\|s^k\| \leq \beta\Delta_k \Leftrightarrow \Delta_k \geq \frac{\|s^k\|}{\beta}$  und  $\|g^k\| > \frac{\mu}{2}\|s^k\|$  gilt nun

$$pred_k(s^k) \geq \frac{\alpha}{2}\|g^k\| \min \left\{ \Delta_k, \frac{\|g^k\|}{C_H} \right\} \geq \frac{\alpha\mu}{4}\|s^k\| \min \left\{ \frac{\|s^k\|}{\beta}, \frac{\mu}{2C_H}\|s^k\| \right\} = c\|s^k\|^2$$

mit  $c = \frac{\alpha\mu}{4} \min \left\{ \frac{1}{\beta}, \frac{\mu}{2C_H} \right\}$ . Aus dem Mittelwertsatz folgt die Existenz eines Zwischenpunktes  $\xi_k$  auf der Verbindungsstrecke von  $x^k$  und  $x^k + s^k$ . Daher gilt auch für die Taylor-Entwicklung von  $f$  um  $x^k$

$$f(x^k + s^k) = f(x^k) + \nabla f(x^k)^T s^k + \frac{1}{2}s^{kT}\nabla^2 f(\xi_k)s^k$$

für ein  $\xi_k \in [x^k, x^k + s^k]$ . Damit ergibt sich

$$\begin{aligned} |\rho_k(s^k) - 1| &= \left| \frac{ared_k - pred_k}{pred_k} \right| = \frac{|ared_k(s^k) - pred_k(s^k)|}{pred_k(s^k)} \\ &= \frac{|(f_k - f(x^k + s^k)) - (f_k - q_k(s^k))|}{pred_k(s^k)} = \frac{|q_k(s^k) - f(x^k + s^k)|}{pred_k(s^k)} \\ &= \frac{\left| \left( f(x^k) + \nabla f(x^k)^T s^k + \frac{1}{2}s^{kT}\nabla^2 f(x^k)s^k \right) - \left( f(x^k) + \nabla f(x^k)^T s^k + \frac{1}{2}s^{kT}\nabla^2 f(\xi_k)s^k \right) \right|}{pred_k(s^k)} \\ &\leq \frac{|s^{kT}(\nabla^2 f(x^k) - \nabla^2 f(\xi_k))s^k|}{2c\|s^k\|^2} \leq \frac{1}{2c}\|\nabla^2 f(x^k) - \nabla^2 f(\xi_k)\| \rightarrow 0 \quad (k \rightarrow \infty), \end{aligned}$$

da  $x^k \rightarrow \bar{x}$  und  $s^k \rightarrow 0$  nach a).

Daher gibt es  $l > 0$  mit  $\rho_k(s^k) > \eta_2$  für alle  $k \geq l$ . Dies zeigt, dass für alle  $k \geq l$  der Schritt  $s^k$  erfolgreich ist und dass  $\Delta_k \geq \Delta_l > 0$  gilt.

c) Für hinreichend große  $k$  gilt  $x^k \in \bar{B}_\varepsilon(\bar{x})$  und somit ist  $H_k$  invertierbar mit  $\|H_k^{-1}\| \leq \frac{1}{\mu}$ . Dabei wurde hier wieder Lemma 4.4.2 mit  $\|H_k\| \geq \mu \Leftrightarrow \|H_k^{-1}\| \leq \frac{1}{\mu}$  verwendet. Für den Newton-Schritt  $s_n^k$  ergibt sich also:

$$\|s_n^k\| \leq \frac{1}{\mu}\|g^k\| \rightarrow 0 \quad (k \rightarrow \infty).$$

Daher existiert  $l' \geq l$  mit

$$x^k \in B_\varepsilon(\bar{x}), \quad \|s_n^k\| \leq \Delta_l \leq \Delta_k$$

für alle  $k \geq l'$ . Somit ist dann  $s_n^k$  die Lösung von (4.2) und erfüllt also insbesondere die Cauchy-Abstiegsbedingung. Daher wird für  $k \geq l'$  stets  $s^k = s_n^k$  gewählt und nach b) ist der Schritt erfolgreich. Es wird also für  $k \geq l'$  das Newton-Verfahren durchgeführt. Die lokalen Konvergenzaussagen folgen nun aus Satz 4.4.1.

□

## 5 Numerische Resultate

---

In diesem Abschnitt werden die im Kapitel 3 und 4 vorgestellten Gradienten- und Trust-Region-Verfahren und deren Konvergenzverhalten in der Praxis überprüft. Hierfür werden die Verfahren in *Mathematica* implementiert. Zur Verifikation des Optimierungsalgorithmus wird die viel zitierte Rosenbrock-Funktion als Testfunktion herangezogen. Infolgedessen werden mit Hilfe der erzielten numerischen Resultate die Besonderheiten der beiden Verfahren herausgearbeitet.

Die nichtkonvexe zweidimensionale *Rosenbrock-Funktion*

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}, \quad f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

ist ein Polynom vierten Grades. Der erste Term  $100(x_2 - x_1^2)^2$  definiert hier das Tal beziehungsweise die Parabel und durch  $(1 - x_1)^2$  wird dies leicht verschoben. Daher besitzt ihr Graph auch von oben betrachtet ein im Wesentlichen parabelförmiges Tal mit seitlich sehr steil ansteigenden Funktionswerten und im Vergleich hierzu eine nur sehr geringer Steigung des Talbodens, siehe den Flächen- und Höhenlinienplot in Abbildung 5.1.

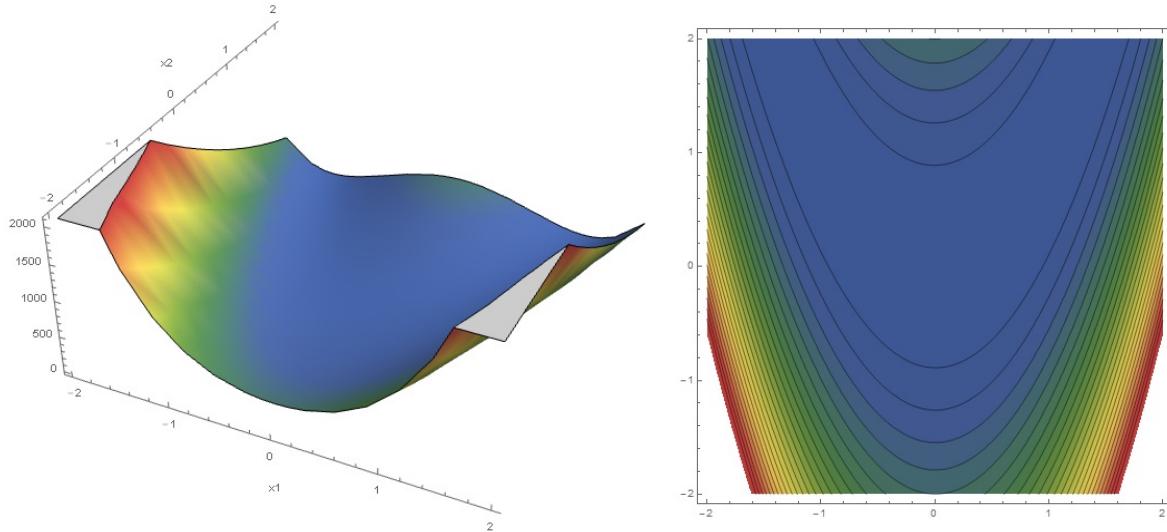


Abbildung 5.1: Veranschaulichung der Rosenbrock-Funktion

Es gilt

$$\nabla f(x) = \begin{pmatrix} 400x_1^3 - 400x_1x_2 + 2x_1 - 2 \\ -200x_1^2 + 200x_2 \end{pmatrix} \quad \text{und} \quad \nabla^2 f(x) = \begin{pmatrix} 1200x_1^2 - 400x_2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}.$$

Es wird das Optimalitätskriterium  $\nabla f(x) \stackrel{!}{=} 0$  angewendet und die Gleichungen werden nach  $x_1$  und  $x_2$  aufgelöst. Daraus ergibt sich der stationäre Punkt  $\bar{x} = (1, 1)^T$ . Zusätzlich wird der Optimalpunkt  $\bar{x}$  in die Hesse-Matrix  $\nabla^2 f(x)$  eingesetzt und es wird geprüft, ob die Eigenwerte der Hesse-Matrix positiv definit sind:

$$\nabla^2 f(\bar{x}) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$$

mit Einheitsmatrix  $E_n$  folgt

$$\begin{aligned} \det(\nabla^2 f(\bar{x}) - \lambda \cdot E_2) &= \det \begin{pmatrix} 802 - \lambda & -400 \\ -400 & 200 - \lambda \end{pmatrix} = (802 - \lambda) \cdot (200 - \lambda) - (-400 \cdot -400) \\ &= \lambda^2 - 1002\lambda + 400 \quad \Rightarrow \quad \lambda_1 \approx 1001,6 \text{ und } \lambda_2 \approx 0,4. \end{aligned}$$

Die positive Definitheit der Hesse-Matrix  $\nabla^2 f(x)$  im Punkt  $\bar{x} = (1, 1)^T$  ist ein Indiz dafür, dass  $\bar{x}$  das einzige globale Minimum der Funktion ist und dass  $f$  keine weiteren stationären Punkte besitzt. Wird nun das globale Minimum in die Funktion eingesetzt, so ist der Minimalwert der Funktion  $f(\bar{x}) = 0$ . Zudem wird hier darauf hingewiesen, dass  $\nabla^2 f(x)$  offenbar auch negative Eigenwerte besitzen kann und somit ist die Rosenbrock-Funktion nicht auf dem gesamten  $\mathbb{R}^n$  konvex.

Es wird zunächst mit dem Gradientenverfahren begonnen. Das in Kapitel 3 vorgestellte Verfahren mit der Armijo-Schrittweitenregel wird auf die oben angegebene Rosenbrock-Funktion angewendet. Es werden dabei folgende Daten verwendet:

$x^0 = (-1, 9; 2)^T$  (Startpunkt),  $\varepsilon = 10^{-3}$  (Abbruchbedingung),  $\beta = 0,5$  und  $\gamma = 10^{-4}$  (Armijo-Regel).

In der Tabelle 5.1 wird der Verlauf des Gradientenverfahrens vorgeführt. Es werden 4031 Iterationen benötigt, um das erforderliche Abbruchkriterium zu erzielen.

$k$	$x^k$	$f(x^k)$
0	(-1, 9; 2)	267,62
1	(-1, 29971; 2, 15723)	27,1899
2	(-1, 53281; 2, 06582)	14,4632
3	(-1, 35800; 2, 12123)	13,2361
4	(-1, 50037; 2, 06712)	9,63696
5	(-1, 38765; 2, 10305)	8,85048
6	(-1, 47919; 2, 06839)	7,57699
:	:	:
42	(-1, 41995; 2, 02053)	5,85799
43	(0, 997336; 1, 16747)	2,98557
44	(1, 06466; 1, 13372)	0,00418556
45	(1, 06362; 1, 13231)	0,00415279
46	(1, 06392; 1, 13211)	0,00408911
:	:	:
4030	(1, 00078; 1, 00157)	$6,13034 \cdot 10^{-7}$
4031	(1, 00078; 1, 00157)	$6,12058 \cdot 10^{-7}$

Tabelle 5.1: Verlauf des Gradientenverfahrens mit der Armijo-Schrittweitenregel bei Anwendung auf die Rosenbrock-Funktion mit Startpunkt  $x^0 = (-1, 9; 2)^T$

Für den Großteil der erzeugten Iterierten gilt, dass sie zwar in einer relativ kleinen Umgebung des Talbodens liegen, aber diesen nicht exakt treffen. In diesem Fall weicht dann die Richtung  $s^k = -\nabla f(x^k)$  deutlich von der Tangente des Talbodens ab. Die Forderung der  $f$ -Abnahme erzwingt nun einen sehr kurzen Schritt  $\sigma_k s^k$ , da die Funktion vom Talboden weg sehr schnell wächst. Auf diese Weise wird ein Zickzack-Pfad in einer kleinen Umgebung des Talbodens erzeugt, der aus sehr kurzen Schritten zusammengesetzt ist. Für  $f(x^k) \rightarrow f(\bar{x}) = 0$  zieht sich die Niveaumenge  $\{x; f(x) \leq f(x^k)\}$  zunehmend enger um den Talboden zusammen. Die Iterierten haben für ihren Zickzack-Pfad also zunehmend weniger Platz. Dies führt dazu, dass die Konvergenz des Gradientenverfahrens sehr langsam ist.

Diese Überlegung wird auch durch den Iterationsablauf des Verfahrens bestätigt, denn obwohl das Verfahren nach 43 Iterationsschritten bereits in der Nähe des Minimums  $\bar{x} = (1, 1)^T$  ist, braucht es noch über 1800 Iterationen um dem vorgegebenem Abbruchkriterium zu genügen.

Die Abbildung 5.2 illustriert den Iterationsverlauf des Gradientenverfahrens. Es wurden nicht nur die Iterationspunkte gezeichnet, sondern auch zwei aufeinanderfolgende Iterationspunkte mit einer Linie verbunden.

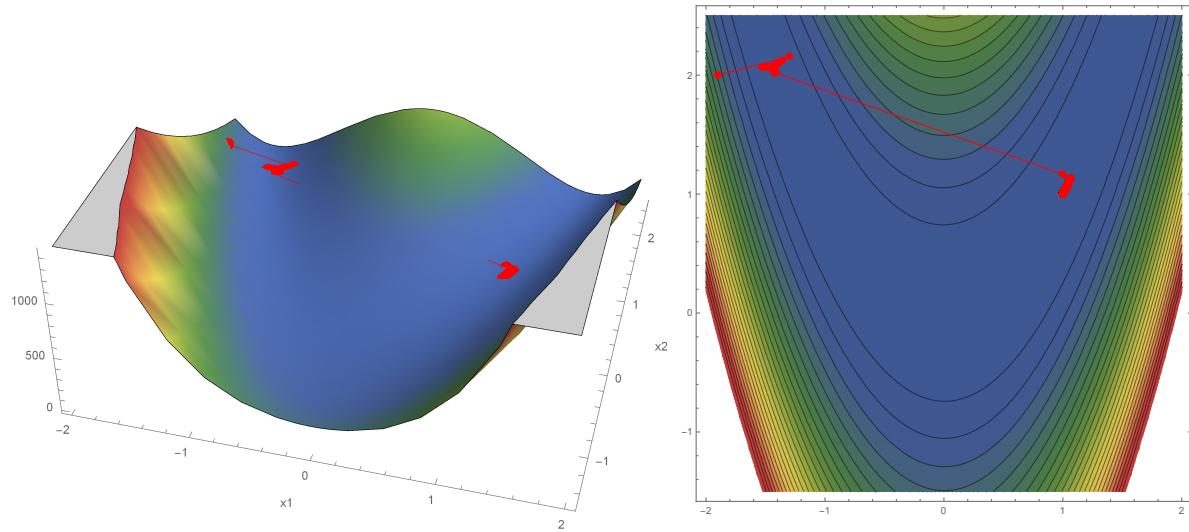


Abbildung 5.2: Gradientenverfahren: Iterationsverlauf bei der Rosenbrock-Funktion mit Startpunkt  $x^0 = (-1, 9; 2)^T$

Die Abbildung 5.3 zeigt den Polygonzug der Iterierten  $x^k$ . Das linke Bild illustriert zunächst den großen Sprung von der Iterierten  $x^0$  auf das Iterierte  $x^1$  und dann nach 41 weitere kleinere Schritte erfolgt wieder eine große Schrittweite von der Iterierten  $x^{42}$  auf das Iterierte  $x^{43}$ . Das rechte Bild zeigt hier das zuvor beschriebene Zickzackverhalten der Iterierten  $x^k$  mit sehr kurzen Schrittweiten in den letzten Iterationen des Gradientenverfahrens.

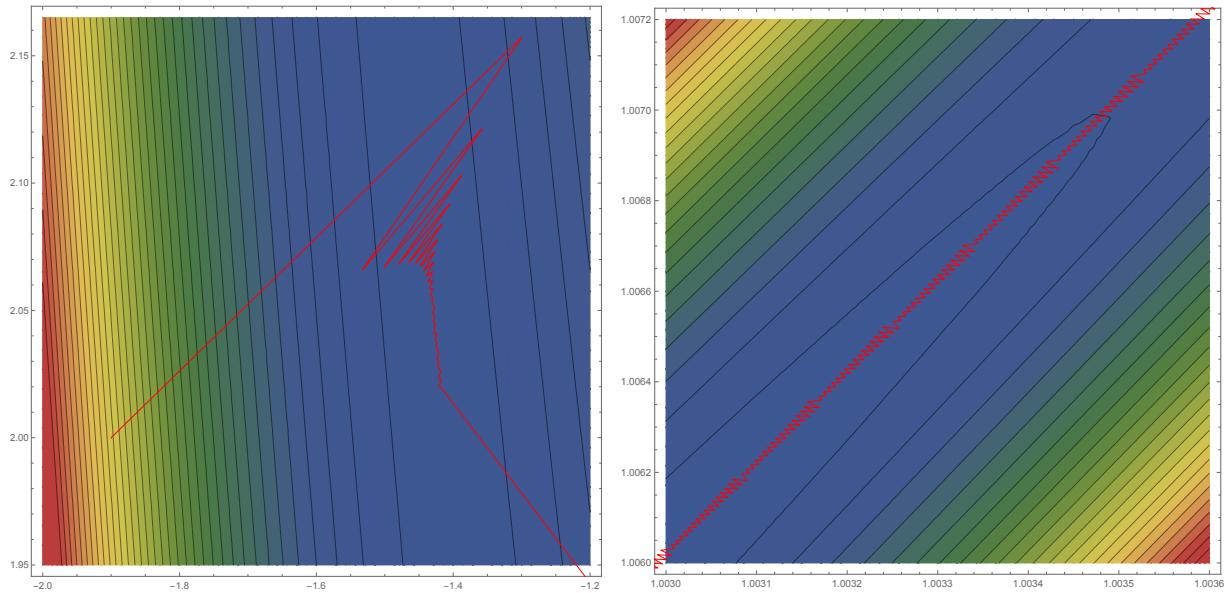


Abbildung 5.3: Iterationsverlauf der Rosenbrock-Funktion mit Startpunkt  $x^0 = (-1, 9; 2)^T$

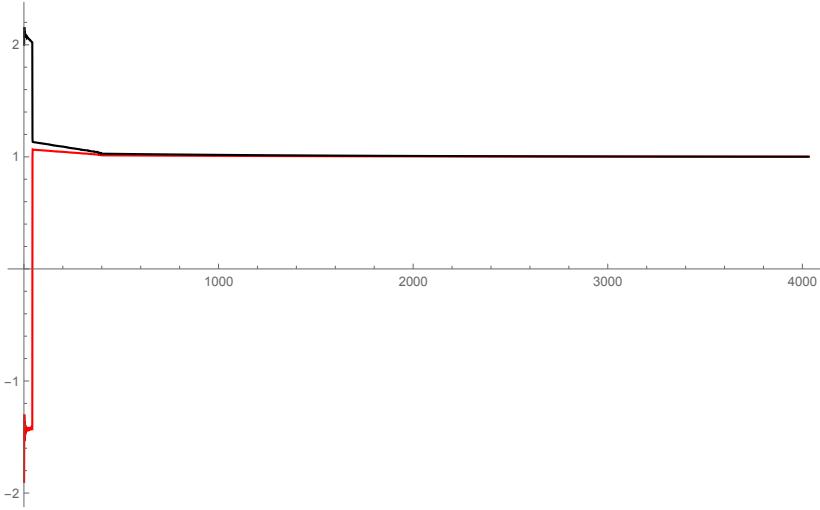


Abbildung 5.4: Gradientenverfahren: Die Folge der Iterierten  $(x^k)$  mit  $x_1$  (rot) und  $x_2$  (schwarz) mit Startpunkt  $x^0 = (-1, 9; 2)^T$

Für die Konditionszahl der Rosenbrock-Funktion gilt

$$\kappa(\nabla^2 f(\bar{x})) = \frac{\lambda_{\max}(\nabla^2 f)}{\lambda_{\min}(\nabla^2 f)} = 2508,01.$$

Somit ist die Kondition der Hesse-Matrix sehr schlecht und die Konvergenz des Gradientenverfahrens wird dementsprechend langsam sein.

Es werden verschiedene Werte von  $\varepsilon$  für das Abbruchkriterium  $\|\nabla f(x^k)\| \leq \varepsilon$  getestet. Die Tabelle 5.2 gibt an, wie viele Iterationsschritte das Gradientenverfahren benötigt werden um der Abbruchbedingung zu genügen.

$\varepsilon$	Iterationen $k$	$x^k$	$\ x^k - \bar{x}\ $
$10^{-2}$	1150	(1, 0078446346; 1, 0157662290)	0,01761
$10^{-3}$	4031	(1, 0007809416; 1, 0015671709)	0,00175097
$10^{-4}$	6884	(1, 0000774985; 1, 0001554737)	0,000173718
$10^{-5}$	9705	(1, 0000078593; 1, 0000157657)	0,0000176161
$10^{-6}$	12545	(1, 0000007851; 1, 0000015751)	$1,75996 \cdot 10^{-6}$
$10^{-7}$	15385	(1, 0000000784; 1, 0000001573)	$1,7583 \cdot 10^{-7}$
$10^{-8}$	18225	(1, 0000000078; 1, 0000000157)	$1,75664 \cdot 10^{-8}$

Tabelle 5.2: Minimierung der Rosenbrock-Funktion mittels des Gradientenverfahrens

Werden ab der ersten Zeile die Ergebnisse mit den Ergebnissen in der Zeile davor verglichen, so folgt bei genauerer Betrachtung

$$\begin{aligned} \frac{0,00175097}{0,01761} &= 0,09943, & \frac{0,000173718}{0,00175097} &= 0,09921, & \frac{0,0000176161}{0,000173718} &= 0,10140, \\ \frac{1,75996 \cdot 10^{-6}}{0,0000176161} &= 0,09990, & \frac{1,7583 \cdot 10^{-7}}{1,75996 \cdot 10^{-6}} &= 0,09990, & \frac{1,75664 \cdot 10^{-8}}{1,7583 \cdot 10^{-7}} &= 0,09990, \end{aligned}$$

dass der Abstand zwischen dem Vektor  $x^k$ , mit dem das Gradientenverfahren abbricht, und dem Minimum  $\bar{x}$  immer circa um einen Faktor von 0,1 verkleinert wird. Zusätzlich werden die Iterationsanzahlen wieder ab der ersten Zeile betrachtet

$$4031 - 1150 = 2881, \quad 6884 - 4031 = 2853, \quad 9705 - 6884 = 2821, \\ 12545 - 9705 = 2840, \quad 15385 - 12545 = 2840, \quad 18225 - 15385 = 2840$$

und es folgt, dass bei Veränderung des Abbruchkriteriums eine sehr starke Erhöhung, circa um 2840, der benötigten Anzahl der Iteration bewirkt.

Diese Gesetzmäßigkeit beruht auf das im Kapitel 3 vorgestellte Konvergenzrate bei quadratischen Funktionen, denn die Aussage des Satzes 3.5.1 lässt sich auch auf den Fall nichtquadratischer Funktionen übertragen [SP93]. Es gilt daher auch

$$\left(\frac{\kappa - 1}{\kappa + 1}\right)^k = \left(\frac{2508,01 - 1}{2508,01 + 1}\right)^{2840} = 0,103857.$$

Es wird nun die Leistungsfähigkeit des Trust-Region-Verfahren demonstriert. Wie beim Gradientenverfahren wird hier das Konvergenzverhalten des Trust-Region-Verfahrens mit der in der Optimierungsliteratur sehr beliebten Rosenbrock-Funktion getestet.

Für das Trust-Region-Verfahren werden folgende Parameter verwendet:

$x^0 = (-1, 9; 2)^T$  und  $\Delta_0 = 2$  (Startwerte),  $\varepsilon = 10^{-3}$  (Abbruchbedingung),

$\beta = 2$  und  $\alpha = 0,5$  (Cauchy-Abstiegsbedingung),

$\eta_1 = 0,1$ ,  $\eta_2 = 0,9$ ,  $\gamma_1 = 0,5$ ,  $\gamma_2 = 2$  und  $\Delta_{min} = 0,01$  (Update des Trust-Region-Radius).

Die folgende Tabelle 5.3 gibt nicht nur die Iterierten  $x^k$  sondern auch  $\rho_k(s^k) = \frac{ared_k(s^k)}{pred_k(s^k)}$  an. Somit kann zusätzlich aus der dritten Spalte abgelesen werden, ob ein Schritt erfolgreich war, also  $\rho_k(s^k) > \eta_1$  und somit der Schritt  $s^k$  akzeptiert und  $x^{k+1} = x^k + s^k$  gesetzt wurde, oder nicht. Wie schon im Kapitel 5 erläutert, muss das Trust-Region-Verfahren die Cauchy-Abstiegsbedingung prüfen und entscheiden, ob ein Cauchy-Schritt  $s_c^k$  oder ein Newton-Schritt  $s_n^k$  durchgeführt wird. Da dies eine wichtige Information ist, wird es hier zusätzlich in der Tabelle mit aufgelistet. Hierbei bedeutet  $N$  ein Newton- und  $C$  ein Cauchy-Schritt.

$k$	$x^k$	$\rho_k$	$s^k$
0	(-1, 9; 2)	1.0001	$N$
1	(-1, 89102; 3, 57588)	1,00108	$N$
2	(-1, 88898; 3, 57589)	0,999999	$C$
3	(-1, 88891; 3, 56772)	-712,817	$N$
4	(-1, 88891; 3, 56772)	-712,817	$N$
5	(-1, 88891; 3, 56772)	1,0011	$N$
6	(-1, 88683; 3, 56773)	0,999996	$C$
7	(-1, 88669; 3, 55914)	-634,505	$N$
8	(-1, 88669; 3, 55914)	-634,505	$N$
9	(-1, 88669; 3, 55914)	1,00113	$N$
10	(-1, 88456; 3, 55918)	0,99999	$C$
11	(-1, 88436; 3, 55011)	-560,602	$N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1188	(0, 983646; 0, 966281)	1,17659	$N$
1189	(0, 996671; 0, 993183)	1,10757	$N$
1190	(0, 999891; 0, 999771)	1,02508	$N$
1191	(1; 1)	1,00109	$N$

Tabelle 5.3: Verlauf des Trust-Region-Verfahrens bei Anwendung auf die Rosenbrock-Funktion mit Startpunkt  $x^0 = (-1, 9; 2)^T$

Wie in der Tabelle 5.3 dargestellt, werden nur 1191 Iterationen benötigt und somit ist das Trust-Region-Verfahren für die Rosenbrock-Funktion mit dem Startwert  $x^0 = (-1, 9; 2)^T$  sehr viel schneller als das Gradientenverfahren.

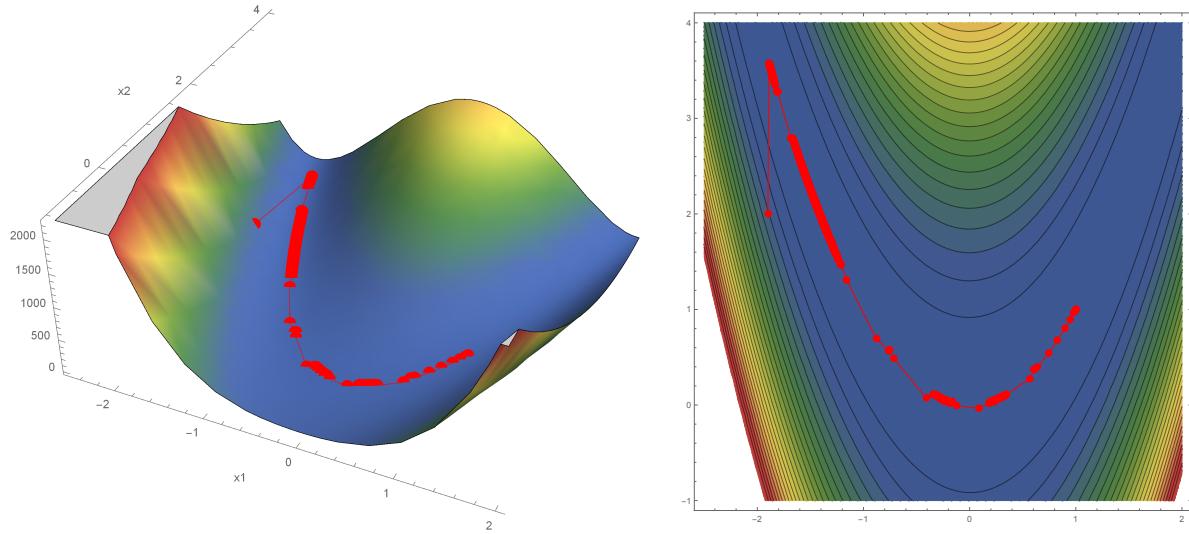


Abbildung 5.5: Trust-Region-Verfahren: Iterationsverlauf bei der Rosenbrock-Funktion mit Startpunkt  $x^0 = (-1, 9; 2)^T$

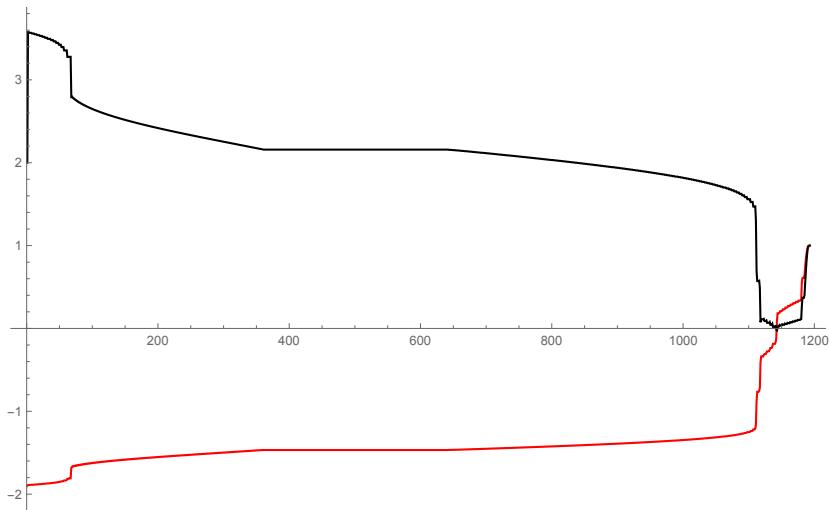


Abbildung 5.6: Trust-Region-Verfahren: Die Folge der Iterierten  $(x^k)$  mit  $x_1$  (rot) und  $x_2$  (schwarz) mit Startpunkt  $x^0 = (-1, 9; 2)^T$

Noch deutlicher wird das unterschiedliche Konvergenzverhalten beider Verfahren anhand des Startwertes  $x^0 = (-1, 2; 1)^T$ . Das Trust-Region-Verfahren für die Rosenbrock-Funktion konvergiert weitaus schneller, als das Gradientenverfahren. Hierbei braucht das Gradientenverfahren 5231 Iterationen während das Trust-Region-Verfahren nur 121 benötigt.

$k$	$x^k$	$f(x^k)$
0	(-1, 2; 1)	24, 2
1	(-0, 989453; 1, 08594)	5.10111
2	(-1, 06433; 1, 04417)	5.04701
3	(-1, 02345; 1, 06148)	4.11404
4	(-1, 02677; 1, 056)	4.10809
5	(-1, 02026; 1, 05532)	4.10215
6	(-1, 02384; 1, 04969)	4.09613
7	(-1, 01709; 1, 04913)	4.09012
8	(-1, 02085; 1, 04347)	4.08401
9	(-1, 01397; 1, 04291)	4.07792
10	(-1, 01781; 1, 03714)	4.0717
11	(-1, 01088; 1, 03667)	4.065511
$\vdots$	$\vdots$	$\vdots$
5228	(0, 999209; 0, 998414)	$6, 27816 \cdot 10^{-7}$
5229	(0, 999208; 0, 998416)	$6, 26806 \cdot 10^{-7}$
5230	(0, 99921; 0, 998417)	$6, 25801 \cdot 10^{-7}$
5231	(0, 99921; 0, 998418)	$6, 24799 \cdot 10^{-7}$

Tabelle 5.4: Verlauf des Gradientenverfahrens mit der Armijo-Schrittweitenregel bei Anwendung auf die Rosenbrock-Funktion mit Startpunkt  $x^0 = (-1, 2; 1)^T$

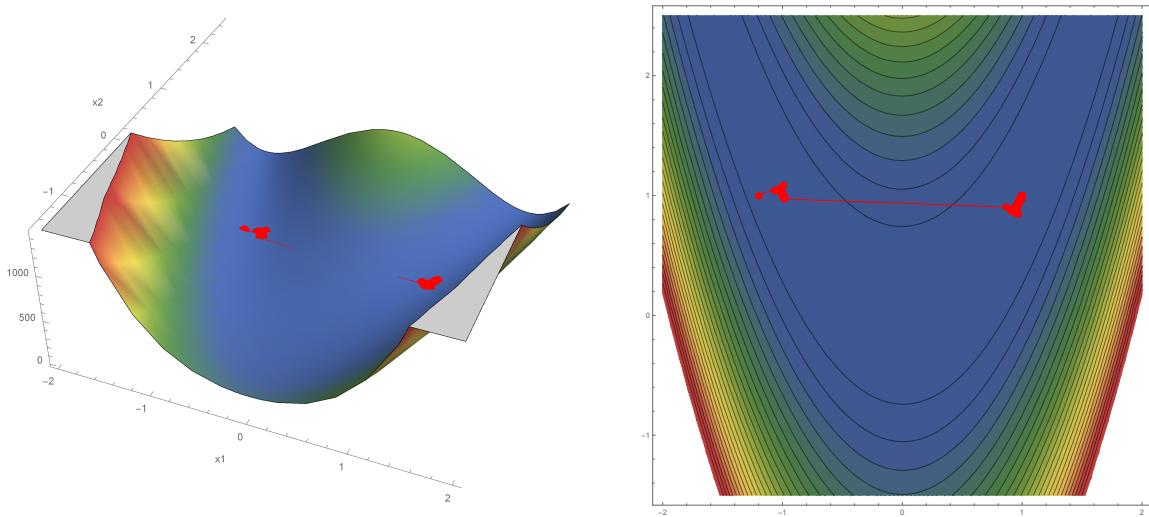


Abbildung 5.7: Gradientenverfahren: Iterationsverlauf bei der Rosenbrock-Funktion mit Startpunkt  $x^0 = (-1, 2; 1)^T$

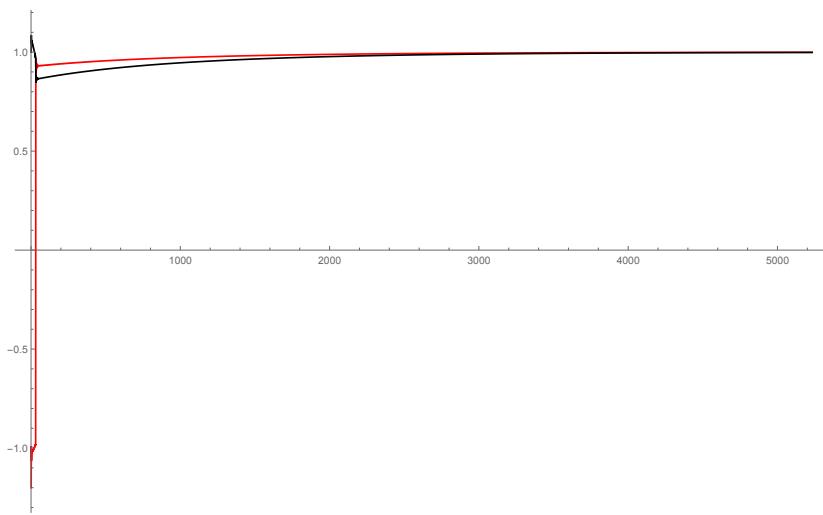


Abbildung 5.8: Gradientenverfahren: Die Folge der Iterierten  $(x^k)$  mit  $x_1$  (rot) und  $x_2$  (schwarz) mit Startpunkt  $x^0 = (-1, 2; 1)^T$

$k$	$x^k$	$\rho_k$	$s^k$
0	(-1, 2; 1)	1, 00277	N
1	(-1, 17528; 1, 38067)	-333, 709	N
2	(-1, 17528; 1, 38067)	1, 00344	N
3	(-1, 17118; 1, 38078)	0, 999962	C
4	(-1, 17074; 1, 3697)	-279, 185	N
5	(-1, 17074; 1, 3697)	-279, 185	N
6	(-1, 17074; 1, 3697)	1, 00353	N
7	(-1, 16653; 1, 36986)	0, 999922	C
8	(-1, 16589; 1, 35798)	-230, 577	N
$\vdots$	$\vdots$	$\vdots$	$\vdots$
111	(0, 654209; 0, 422874)	0, 983496	N
112	(0, 825134; 0, 65163)	0, 101357	N
113	(0, 850688; 0, 723016)	1, 042	N
114	(0, 850688; 0, 723016)	-0, 424058	N
115	(0, 850688; 0, 723016)	-0, 424058	N
116	(0, 850688; 0, 723016)	-0, 424058	N
117	(0, 890641; 0, 791286)	0, 984182	N
118	(0, 969268; 0, 933298)	0, 8434854	N
119	(0, 983009; 0, 966118)	1, 05417	N
120	(0, 999382; 0, 998496)	1, 01043	N
121	(0, 999969; 0, 999937)	1, 00244	N

Tabelle 5.5: Verlauf des Trust-Region-Verfahrens bei Anwendung auf die Rosenbrock-Funktion mit Startpunkt  $x^0 = (-1, 2; 1)^T$

Das Trust-Region-Verfahren löst dieses Problem sehr effizient. Allerdings muss natürlich festgehalten werden, dass hier in jeder Iteration die Hesse-Matrix berechnet und diese dann invertiert werden muss.

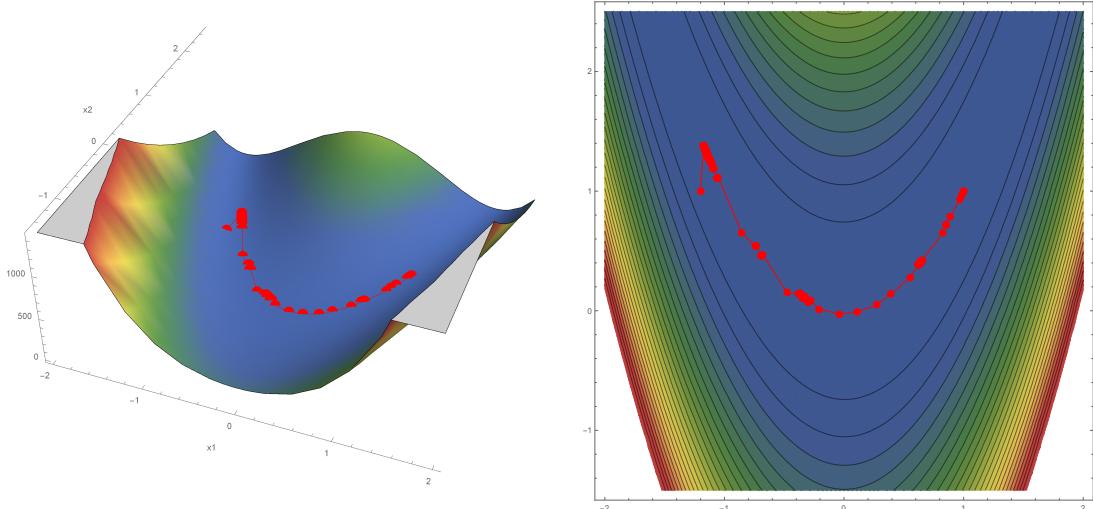


Abbildung 5.9: Trust-Region-Verfahren: Iterationsverlauf mit Startpunkt  $x^0 = (-1, 2; 1)^T$

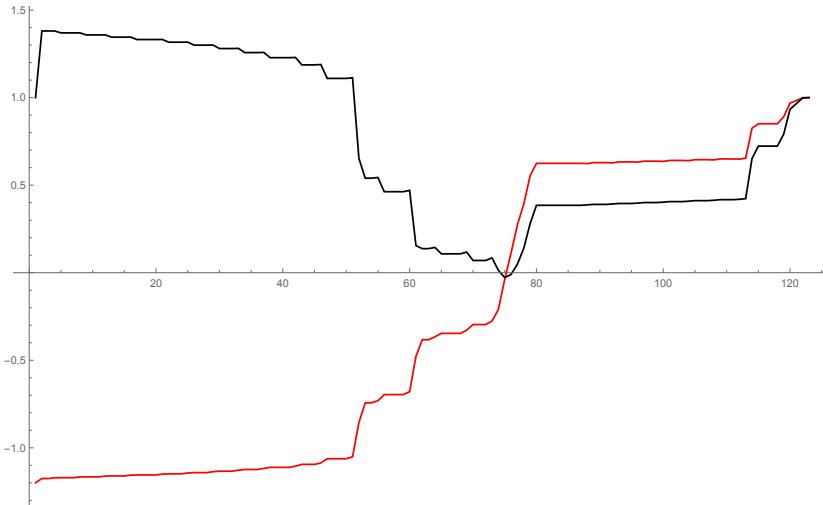


Abbildung 5.10: Trust-Region-Verfahren: Die Folge der Iterierten ( $x^k$ ) mit  $x_1$  (rot) und  $x_2$  (schwarz) mit Startpunkt  $x^0 = (-1, 2; 1)^T$

Zusammenfassend kann bei Anwendung der Verfahren bezüglich der Rosenbrock-Funktion gesagt werden, dass das Trust-Region-Verfahren offensichtlich schneller konvergiert als das Gradientenverfahren. Bei gleichen Startwerten und gleicher Abbruchbedingung hat das Gradientenverfahren mit der Armijo-Schrittweitenregel 4031 Iterationen, siehe Tabelle 5.1, benötigt während das Trust-Region-Newton-Verfahren schon nach 1191 Iterationen, siehe Tabelle 5.3, terminiert. Für den Startwert  $x^0 = (-1, 2; 1)^T$  hat das Gradientenverfahren sogar 5231 Iterationen berechnet, siehe Tabelle 5.4, während das Trust-Region-Newton-Verfahren nur 121 Iterationen, siehe Tabelle 5.5, benötigt hat. Es wird darüberhinaus bei dem Trust-Region-Verfahren mit Startpunkt  $x^0 = (-1, 2; 1)^T$  in den letzten Iterationsschritten eine schnelle lokale Konvergenz sichtbar.

Das Resultat ist sinnvoll, da das Trust-Region-Verfahren nicht nur die Informationen über den Gradienten  $\nabla f(x)$  sondern auch zusätzlich die Hesse-Matrix  $\nabla^2 f(x)$  auswertet. Zudem wurde bei dem Trust-Region-Verfahren durch die Cauchy-Abstiegsbedingung eine Art Absicherung eingeführt, bei dem das Verfahren mindestens mit der Gradientenrichtung übereinstimmen soll. Überdies sei bemerkt, dass sich

das oszillatorische Verhalten des Gradientenverfahrens nachteilig im Konvergenzverhalten gegenüber dem Trust-Region-Verfahren auswirkt.

Die grundlegende Problemstellung dieser Arbeit war das Gradienten- und das Trust-Region-Verfahren zueinander ins Verhältnis zu setzen. Während das Gradientenverfahren als Ansatz lediglich eine lineare Approximation an einem nichtlinearen Problem verwendet, verwertet das Trust-Region-Verfahren die Informationen aus der Hesse-Matrix  $\nabla^2 f(x)$ , indem es eine quadratische Approximation bestimmt. Somit können die numerischen Resultate eindeutig bestätigen, dass die anfangs erläuterten Annahmen bezüglich der Konvergenzverhalten der Verfahren beim Anwenden auf die Rosenbrock-Funktion zutreffen.

# A Anhang

---

## Minimierung der Rosebrock-Funktion mittels des Gradientenverfahrens

```
(* Die Parameter *)
γ = 10^(-4);
β = 1/2;
ε = 10^(-3);

(* Die Gradientenrichtung und Armijo-Schrittweitenregel *)
AbstiegsRichtung[f_, x_] :=
  -{D[f[{x01, x02}], x01], D[f[{x01, x02}], x02]} /.
    {x01 → x[[1]], x02 → x[[2]]}
  [Leite ab] [Leite ab]

ArmijoSchrittweite[f_, x_, s_] :=
Module[{σ = 1}, While[f[x + σ s] - f[x] > σ γ (-AbstiegsRichtung[f, x]).s, σ = σ β];
  [Modul] [Solange]
  σ]

(* An das Verfahren wird eine Testfunktion f, ein Startwert x0,
Höchstanzahl an Iterationen n und ein Abbruchbedingung ε übergeben. Es
berechnet zuerst den Gradientenrichtung s. Diese wird zusätzlich
noch in Armijo-Schrittweitenregel benötigt. Zuerst wird getestet,
ob die Armijo-Bedingung mit der Schrittweite σ =
(β=1/2)^0 = 1 erfüllt ist. Wenn dies nicht zutrifft,
wird die Schrittweite so oft mit β=
1/2 multipliziert bis die Armijo-Bedingung erfüllt ist. Zuletzt
wird die neue Iterierte x^(k+1) = x^k + σ*s gesetzt *)
GradientenVerfahren[f_, x0_, n_, ε_] :=
Module[{x = x0, z = {f[x0]}, X = {x0}, σ = 1, s, i = 0},
  [Modul]
  While[ (Norm[{D[f[{x01, x02}], x01], D[f[{x01, x02}], x02}].  

    [Solange] [Norm] [Leite ab] [Leite ab]  

    {x01 → x[[1]], x02 → x[[2]]}] > ε) && i ≤ n,  

    {s = AbstiegsRichtung[f, x];  

     σ = ArmijoSchrittweite[f, x, s];  

     x = x + σ s; z = f[x]; i = i + 1;  

     z = Append[z, z],  

     [Hänge an]  

     x = Append[x, x]  

     [Hänge an]  

    }];
  {x, z}];

f[{x1_, x2_}] := 100 (x2 - x1^2)^2 + (1 - x1)^2;
{x, z} = GradientenVerfahren[f, x0 = {-1.9, 2.}, 20000, ε] // N
  [ln]

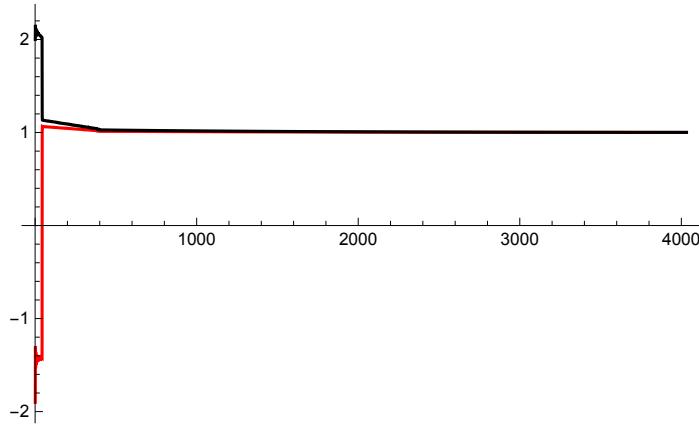
Table[{X[[i]], z[[i]]}, {i, 1, Length[X]}] // TableForm
  [Tabelle] [Länge] [Tabellendarstellung]
```

```

Length[Table[{X[[i]], Z[[i]]}, {i, 1, Length[X]}]]
|Länge|Tabelle|Länge

(* Die Iterationsverlauf von x1 und x2 *)
ListLinePlot[Transpose[X], PlotStyle -> {Red, Black}, PlotRange -> All]
|Liniengrafik einer ...|transponiere|Grafikstil|rot|schwarz|Definitions- und alle

```



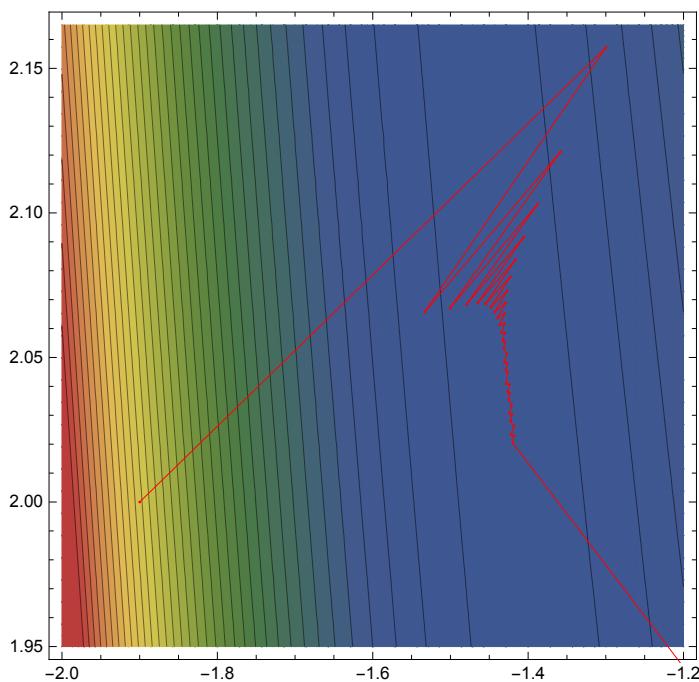
```

(* Die Iterierten x^0 bis x^42 *)
ContourPlot[f[{x1, x2}], {x1, -2, -1.2}, {x2, 1.95, 2.165},
|Konturgraphik

Epilog -> {Red, PointSize[0.0015], Map[Point, X], Line[X]},
|Epilog|rot|Punktgöße|w...|Punkt|Linie

ColorFunction -> "DarkRainbow", Contours -> 35]
|Farbfunktion|Konturen

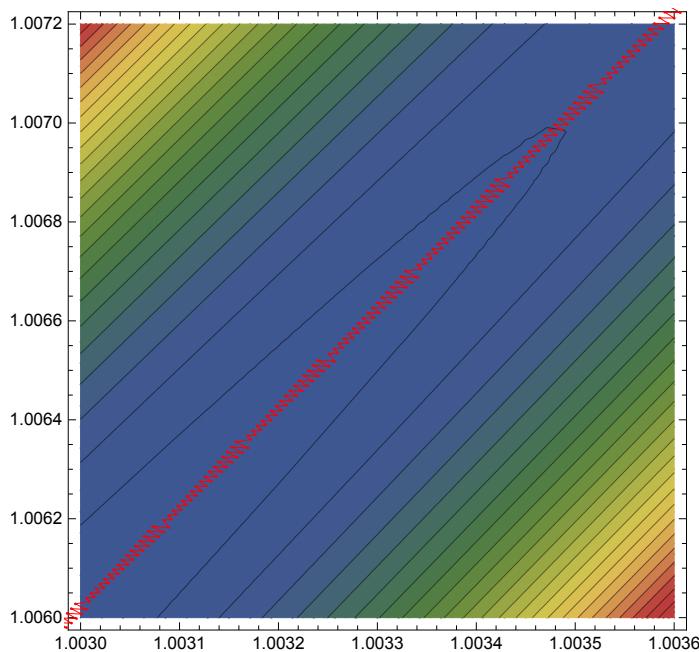
```



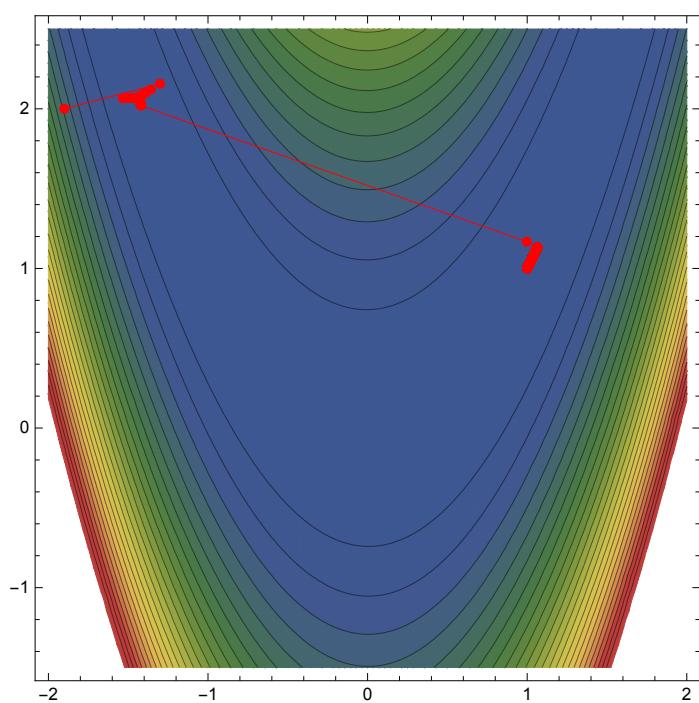
```
(* Iterationsverlauf kurz vor Minimum (1,1) *)
ContourPlot[f[{x1, x2}], {x1, 1.003, 1.0036}, {x2, 1.006, 1.0072},
|Konturgraphik
```

```
Epilog -> {Red, PointSize[0.0015], Map[Point, X], Line[X]},
|Epilog |rot |Punktgröße |w... |Punkt |Linie
```

```
ColorFunction -> "DarkRainbow", Contours -> 25]
|Farbfunktion |Konturen
```



```
ContourPlot[f[{x1, x2}], {x1, -2, 2}, {x2, -1.5, 2.5},  
|Konturgraphik  
Epilog -> {Red, PointSize[0.015], Map[Point, x], Line[x]},  
|Epilog |rot |Punktgröße |w... |Punkt |Linie  
ColorFunction -> "DarkRainbow", Contours -> 25]  
|Farbfunktion |Konturen
```



```

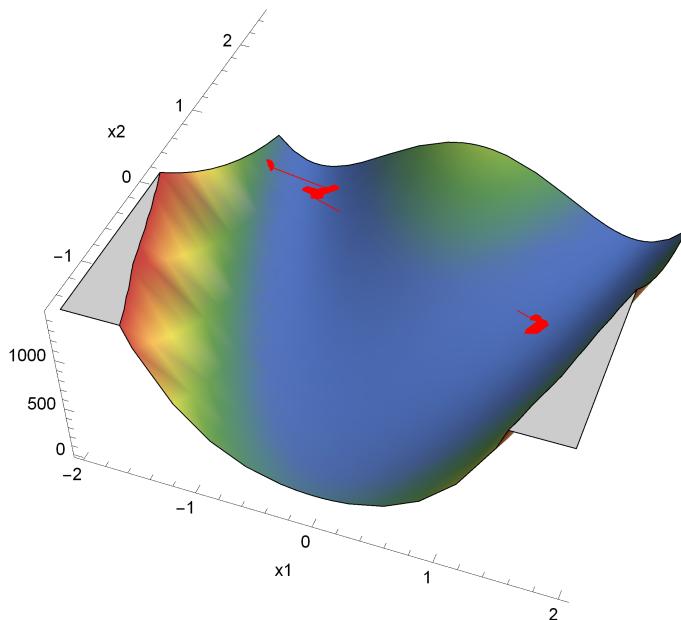
G1 = Plot3D[f[{x1, x2}], {x1, -2, 2}, {x2, -1.5, 2.5}, ColorFunction →
  |grafische Darstellung 3D|Farbfunktion
  "DarkRainbow", Boxed → False, AxesLabel → Automatic, Mesh → None];
  |einger...|falsch|Achsenbeschr...|automatisch|Netz|keine

G2 = Graphics3D[{PointSize[0.02], Red, Point[Table[
  |3D-Grafik|Punktgröße|rot|Punkt|Tabelle
  {X[[i, 1]], X[[i, 2]], f[{X[[i, 1]], X[[i, 2]]}]}, {i, 1, Length[X]}]]}]];
  |Länge

G3 = Graphics3D[{Red, Line[Table[{X[[i, 1]], X[[i, 2]],
  |3D-Grafik|rot|Linie|Tabelle
  f[{X[[i, 1]], X[[i, 2]]}]}, {i, 1, Length[X]}]]}];
  |Länge

Show[
  |zeige an
  G1,
  G2,
  G3]

```



## Konvergenzgeschwindigkeit bei versch. $\epsilon$ -Werte

$\epsilon = 10^{-2}$ , 1150 Iterationen

```

Norm[{1.007844634602773, 1.0157662290974787} - {1., 1.}]
|Norm
0.01761

```

$\epsilon = 10^{-3}$ , 4031 Iterationen

```

Norm[{1.0007809416256455, 1.0015671709533085} - {1., 1.}]
|Norm

```

0.00175097

$\epsilon = 10^{-4}$ , 6884 Iterationen

```
Norm[{1.0000774985140735, 1.0001554737613876} - {1., 1.}]
|Norm
```

0.000173718

$\epsilon = 10^{-5}$ , 9705 Iterationen

```
Norm[{1.0000078593615962, 1.0000157657353972} - {1., 1.}]
|Norm
```

0.0000176161

$\epsilon = 10^{-6}$ , 12545 Iterationen

```
Norm[{1.0000007851861243, 1.0000015751028364} - {1., 1.}]
|Norm
```

$1.75996 \times 10^{-6}$

$\epsilon = 10^{-7}$ , 15385 Iterationen

```
Norm[{1.0000000784444656, 1.0000001573614579} - {1., 1.}]
|Norm
```

$1.7583 \times 10^{-7}$

$\epsilon = 10^{-8}$ , 18225 Iterationen

```
Norm[{1.000000007837046, 1.0000000157212594} - {1., 1.}]
|Norm
```

$1.75664 \times 10^{-8}$

---

# Minimierung der Rosebrock - Funktion mittels des Trust-Region-Verfahrens

```
(Debug) In[1]:= (* Die Parameter *)
γ1 = 0.5;
γ2 = 2;
η1 = 0.1;
η2 = 0.9;
β = 2;
α = 0.5;
ε = 10^(-3);

(* Cauchy-Schritt(sc = tau * s1) und Newton-Schritt sn *)
Tau[G_, H_, Δ_] := If[G.H.G ≤ 0, 1, Min[(Norm[G]^3) / (Δ * G.H.G), 1]];
    | wenn          | klein... Norm
CauchySchritt[G_, H_, Δ_] := -Tau[G, H, Δ] * ((Δ * G) / (Norm[G]));
    | Norm
NewtonSchritt[G_, H_] := -Inverse[H].G;
    | inverse Matrix

(* Die Cauchy-Abstiegsbedingung entscheidet, ob Newton- oder Cauchy-
Schritt gewählt werden muss. Zusätzlich werden diese Schritte auch
jeweils mit C für Cauchy- und mit N für Newton-Schritt benannt. *)
    | Konstante      | numerischer Wert
CABedingung[G_, H_, N_, C_, Δ_] := Module[{Tmp, Tmp2},
    | Modul
        Tmp = If[(-G.N - 0.5 * N.H.N ≥ α * (-G.C - 0.5 * C.H.C)),
            | wenn      | numerisc... | n... | numerischer... | Konstante | Ko... | Konstante
                Tmp2 = "N"; N,
                    | n... | numerischer Wert
                Tmp2 = "C"; C;
                    | K... | Konstante
                Tmp = If[Norm[Tmp] ≤ β Δ, Tmp, C];
                    | ... | Norm           | Konstante
                {Tmp, Tmp2}
            ];
        (* Rho dient für die Bewertung der Qualität des berechneten Schrittes *)
        Rho[x_, s_, G_, H_, f_] := (f[x] - f[x + s]) / (-G.s - 0.5 * s.H.s);
        (* Je nach Rho ρ wird entweder der Radius Δ verkleinert,
        beibehalten oder vergrößert *)
        UpdateTrustRegion[Δ_, rho_] := Module[{Δout},
            | Modul
                If[rho ≤ η1, Δout = γ1 Δ];
                | wenn
                If[η1 < rho ≤ η2, Δout = 0.75 Δ];
                | wenn
```

```

LWWCHH
If[ $\eta_2 < \rho$ ,  $\Delta_{out} = \gamma_2 \Delta$ ] ;
|wenn
 $\Delta_{out}$  ;

(* Ist Rho rho mindestens größer als  $\eta_1$ , so wird der Schritt akzeptiert *)
AkzeptiereSchritt[x_, s_, rho_] := If[rho >  $\eta_1$ , x + s, x];
|wenn

(* An das Verfahren wird eine Testfunktion f,
ein Startwert x0 und höchstanzahl an Iterationen n
übergeben. Zuerst werden die erste und zweite Ableitungen,
bzw. G und H, der Tesfunktion f an der Stelle x berechnet. Diese
werden dann später an oben aufgeföhren Definitionen übergeben und
an der Stelle x ausgewärtet. Zunächst werden Cauchy- und Newton-
Schritt berechnet und mit Hilfe von Cauchy-Abstiegsbedingung geprüft,
welche der beiden gewählt werden muss. Danach wird
Rho berechnet und mit AkzeptiereSchritt geprüft,
ob die neue  $x^{(k+1)}$  die Anforderung genügt. Zuletzt wird mit UpdateTrustRegion
die am Anfang als  $\Delta=2$  festgesetzte Radius aktualisiert. *)
TrustRegionVerfahren[f_, x0_, n_] := Module[{G, H, x = x0, X = {x0},
|Modul

s,  $\Delta = 2$ , rho, k, x01, x02, Gnow, Hnow, Taunow, Cnow, Nnow, Kind},
G[{x1_, x2_}] := Evaluate[{D[f[{x01, x02}], x01], D[f[{x01, x02}], x02}] /.
|werte aus |leite ab |leite ab
{x01 → x1, x02 → x2}];

H[{x1_, x2_}] := Evaluate[{D[G[{x01, x02}], x01], D[G[{x01, x02}], x02}] /.
|werte aus |leite ab |leite ab
{x01 → x1, x02 → x2}];

Catch[
|fange ab

Do[
|iteriere

If[Norm[G[x]] < ε, Throw[Print[k]]];
|... |Norm |werfe |gebe aus

Gnow = G[x];
Hnow = H[x];
Taunow = Tau[x, Gnow, Hnow, Δ];
Cnow = CauchySchritt[Gnow, Hnow, Δ];
Nnow = NewtonSchritt[Gnow, Hnow];
{s, Kind} = CABedingung[Gnow, Hnow, Nnow, Cnow, Δ];
rho = Rho[x, s, Gnow, Hnow, f];
x = AkzeptiereSchritt[x, s, rho];
Δ = UpdateTrustRegion[Δ, rho];
X = Append[X, x],
|hänge an
{k, 0, n}];

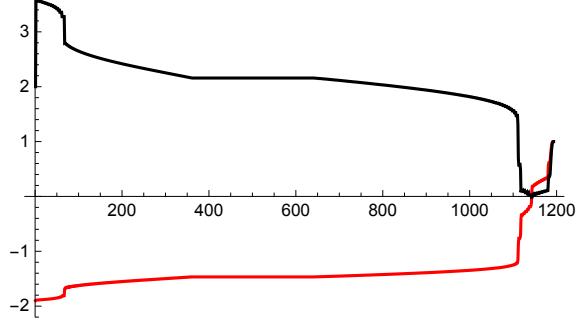
];

```

```
f[{x1_, x2_}] := 100 (x2 - x1^2)^2 + (1 - x1)^2;
X = TrustRegionVerfahren[f, x0 = {-1.9, 2.}, 10000] // N
```

(\* Die Iterationsverlauf der x1 und x2 \*)
ListLinePlot[Transpose[X], PlotStyle -> {Red, Black}, PlotRange -> All]

Liniengrafik einer ... transponiere    Grafikstil    rot    schwarz    Definitions- un... alle



```
ContourPlot[f[{x1, x2}], {x1, -2.5, 2}, {x2, -1., 4},
```

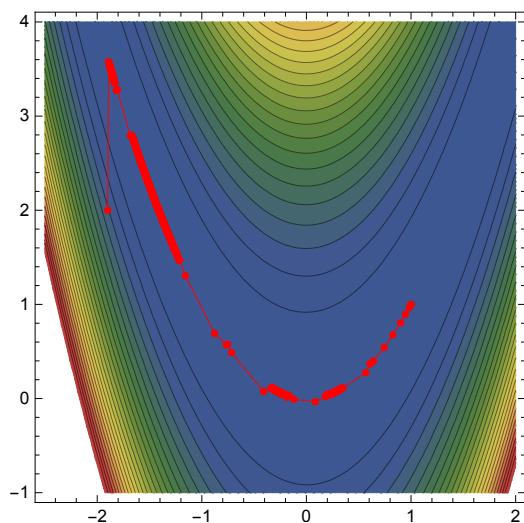
Konturgraphik

Epilog -> {Red, PointSize[0.015], Map[Point, X], Line[X]},

Epilog    rot    Punktgröße    w...    Punkt    Linie

ColorFunction -> "DarkRainbow", Contours -> 25]

Farbfunktion    Konturen



```

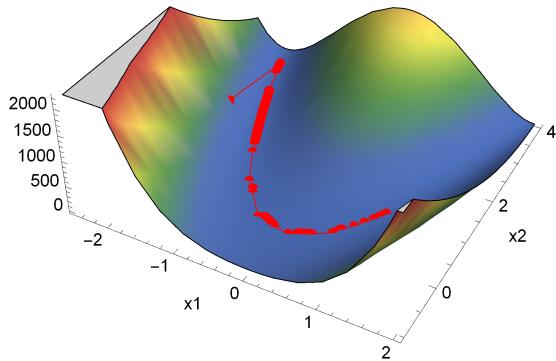
T1 = Plot3D[f[{x1, x2}], {x1, -2.5, 2}, {x2, -1, 4}, ColorFunction -> "DarkRainbow",
  grafische Darstellung 3D Farbfunktion
  Boxed -> False, AxesLabel -> Automatic, Mesh -> None];
einger... falsch Achsenbeschr... automatisch Netz keine

T2 = Graphics3D[{PointSize[0.02], Red, Point[Table[
  3D-Graphik Punktgrö... rot Punkt Tabelle
  {X[[i, 1]], X[[i, 2]], f[{X[[i, 1]], X[[i, 2]]}]}, {i, 1, Length[X]}]]]];
Länge

T3 = Graphics3D[{Red, Line[Table[{X[[i, 1]], X[[i, 2]],
  3D-Graphik rot Linie Tabelle
  f[{X[[i, 1]], X[[i, 2]]}]}, {i, 1, Length[X]}]]];
Länge

Show[
  zeige an
  T1,
  T2,
  T3]

```



# Literaturverzeichnis

---

- [AW02] Alt, Walter: *Nichtlineare Optimierung*, Vieweg & Sohn Verlagsgesellschaft, Braunschweig/Wiesbaden, 2002.
- [CA47] Cauchy, Augustin-Louis: *Méthode Générale pour la Résolution des Systèmes d'Équations Simultanées.*, Comptes Rendus de l'Academie des Sciences, Paris, 1847.
- [CA07] Clausner, André: *Möglichkeiten zur Steuerung von Trust-Region Verfahren im Rahmen der Parameteridentifikation*, Technische Universität Chemnitz, 2007.
- [CC16] Clason, Christian: *Nichtlineare Optimierung*, Universität Duisburg-Essen, 2016.
- [CG00] Conn, Gould & Toint: *Trust-region methods*, SIAM, 2000.
- [GK99] Geiger & Kanzow: *Numerische Verfahren zur Lösung unrestrictierter Optimierungsaufgaben*, Springer-Verlag, Berlin Heidelberg, 1999.
- [GK13] Graichen, Knut: *Methoden der Optimierung und optimalen Steuerung*, Universität Ulm, 2013.
- [GM16] Gerdts, Matthias: *Optimierung*, Universität der Bundeswehr München, 2016.
- [HB15] Harrach, Bastian: *Einführung in die Optimierung*, Universität Stuttgart, 2015.
- [HH11] Harbrecht, Helmut: *Nichtlineare Optimierung*, Universität Basel, 2011.
- [MM13] Merziger, Mühlbach, Wille & Wirth: *Formeln + Hilfen Höhere Mathematik*, Binomi Verlag, 2013.
- [NW06] Nocedal & Wright: *Numerical Optimization*, Springer Science+Business Media, 2006.
- [RH13] Reinhardt, Hoffmann & Gerlach: *Nichtlineare Optimierung*, Springer-Verlag, Berlin Heidelberg, 2013.
- [SP93] Spellucci, Peter: *Numerische Verfahren der nichtlinearen Optimierung*, Birkhäuser, Basel, 1993.
- [SO04] Stein, Oliver: *Optimierung III*, Universität Duisburg-Essen, 2004.
- [UU12] Michael & Stefan Ulbrich: *Nichtlineare Optimierung*, Springer Basel AG, Berlin, 2012.
- [WJ92] Werner, Jochen: *Numerische Mathematik 2*, Vieweg-Studium, Wiesbaden, 1992.

# Abbildungsverzeichnis

---

2.1	Die Darstellung der stationären Punkte je nach Definitheit der Hesse-Matrix . . . . .	8
2.2	Veranschaulichung der konvexen und nicht konvexen Menge . . . . .	9
2.3	Veranschaulichung der konvexen und nicht konvexen Funktion . . . . .	10
3.1	Die Armijo-Schrittweitenregel . . . . .	16
3.2	Gradientenverfahren mit optimaler Schrittweite . . . . .	25
3.3	Beispiel eines gut und eines schlecht konditionierten Problems für das Gradientenverfahren	25
4.1	Die geometrische Darstellung des Trust-Region-Verfahrens. . . . .	28
4.2	Unterschiedliche Gestalt des Trust-Region-Radius . . . . .	29
4.3	Schritte von Suchrichtungs- und Trust-Region-Verfahren . . . . .	30
4.4	Die Cauchy-Schrittweite . . . . .	32
4.5	Positiv definite $H_k$ und Minimum im Vertrauensbereich . . . . .	43
4.6	Positiv definite $H_k$ und Minimum außerhalb des Vertrauensbereiches . . . . .	43
4.7	Negativ definite oder indefinit $H_k$ . . . . .	44
4.8	Die Wirkung der Parameter auf positiv definite $H_k$ . . . . .	44
4.9	Die Wirkung der Parameter auf negativ definite oder indefinit $H_k$ . . . . .	45
5.1	Veranschaulichung der Rosenbrock-Funktion . . . . .	50
5.2	Gradientenverfahren: Iterationsverlauf bei der Rosenbrock-Funktion mit Startpunkt $x^0 = (-1, 9; 2)^T$ . . . . .	52
5.3	Iterationsverlauf der Rosenbrock-Funktion mit Startpunkt $x^0 = (-1, 9; 2)^T$ . . . . .	52
5.4	Gradientenverfahren: Die Folge der Iterierten $(x^k)$ mit $x_1$ (rot) und $x_2$ (schwarz) mit Startpunkt $x^0 = (-1, 9; 2)^T$ . . . . .	53
5.5	Trust-Region-Verfahren: Iterationsverlauf bei der Rosenbrock-Funktion mit Startpunkt $x^0 = (-1, 9; 2)^T$ . . . . .	55
5.6	Trust-Region-Verfahren: Die Folge der Iterierten $(x^k)$ mit $x_1$ (rot) und $x_2$ (schwarz) mit Startpunkt $x^0 = (-1, 9; 2)^T$ . . . . .	55
5.7	Gradientenverfahren: Iterationsverlauf bei der Rosenbrock-Funktion mit Startpunkt $x^0 = (-1, 2; 1)^T$ . . . . .	56
5.8	Gradientenverfahren: Die Folge der Iterierten $(x^k)$ mit $x_1$ (rot) und $x_2$ (schwarz) mit Startpunkt $x^0 = (-1, 2; 1)^T$ . . . . .	57
5.9	Trust-Region-Verfahren: Iterationsverlauf mit Startpunkt $x^0 = (-1, 2; 1)^T$ . . . . .	58
5.10	Trust-Region-Verfahren: Die Folge der Iterierten $(x^k)$ mit $x_1$ (rot) und $x_2$ (schwarz) mit Startpunkt $x^0 = (-1, 2; 1)^T$ . . . . .	58

## Tabellenverzeichnis

---

5.1	Verlauf des Gradientenverfahrens mit der Armijo-Schrittweitenregel bei Anwendung auf die Rosenbrock-Funktion mit Startpunkt $x^0 = (-1, 9; 2)^T$	51
5.2	Minimierung der Rosebrock-Funktion mittels des Gradientenverfahrens	53
5.3	Verlauf des Trust-Region-Verfahrens bei Anwendung auf die Rosenbrock-Funktion mit Startpunkt $x^0 = (-1, 9; 2)^T$	54
5.4	Verlauf des Gradientenverfahrens mit der Armijo-Schrittweitenregel bei Anwendung auf die Rosenbrock-Funktion mit Startpunkt $x^0 = (-1, 2; 1)^T$	56
5.5	Verlauf des Trust-Region-Verfahrens bei Anwendung auf die Rosenbrock-Funktion mit Startpunkt $x^0 = (-1, 2; 1)^T$	57

## **Erklärung**

Ich erkläre, dass ich die eingereichte Ausarbeitung selbstständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den herangezogenen Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Friedberg, den 14. Februar 2017

---

Unterschrift