Optimierung und inverse Probleme

Prof. Dr. Bastian von Harrach

Goethe-Universität Frankfurt am Main
Institut für Mathematik

Wintersemester 2016/17

http://numerical.solutions

Vorläufige Version, wird während der Vorlesungszeit laufend ergänzt und korrigiert.
Letzte Aktualisierung: 6. Oktober 2016

Inhaltsverzeichnis

Kapitel 1

Motivation und Einleitung

1.1 Optimierung und Identifikation

Diese Vorlesung widmet sich der Aufgabe, das Minimum einer Funktion

$$f: X \to \mathbb{R},$$

zu finden. Solche Probleme heißen *Minimierungsaufgaben*, oder (da durch Negation von f offensichtlich auch Maximierungsaufgaben in diese Form gebracht werden können) *Optimierungsaufgaben*. f heißt *Zielfunktional*, X heißt *zulässiger Bereich*, die Elemente $x \in X$ heißen *zulässige Punkte*, und die Forderung $x \in X$ nennt man auch *Nebenbedingung*.

Das Problem

„Finde einen (lokalen oder globalen) Minimierer $x \in X$ von f!"

schreiben wir auch kurz als

$$f(x) \to \min! \quad \text{u.d.N. } x \in X$$

oder

$$\min_{x \in X} f(x).$$

Wir setzen in dieser Vorlesung stets $X \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$ voraus, wobei X Abschluss einer offenen Menge ist. Für unsere Algorithmen werden wir Differenzierbarkeitsforderungen an f stellen, und wir werden uns meist darauf beschränken, lokale Minima zu suchen. In diesem Sinne betrachten wir also *endlich-dimensionale*, *kontinuierliche*, *glatte* und *lokale* Optimierungsaufgaben.

Optimierungsaufgaben werden im Englischen auch als *program* bezeichnet. Entsprechend nennt man z.B. die Lösung linearer Optimierungsaufgaben *linear programming*.

Ein wichtiger Spezialfall tritt bei der Lösung von Gleichungssystemen

$$F(x) = y$$

mit $F: \mathbb{R}^n \to \mathbb{R}^m$ auf. Solche *inversen Probleme* (auch: *Identifikationsprobleme* oder *Ausgleichsprobleme*) besitzen in der Praxis aufgrund von Mess- und Modellierungsungenauigkeiten oft keine eindeutige Lösung. Stattdessen begnügt man sich mit bestmöglichen Lösungen, also Minima von

$$f(x) \to \min!$$

mit $f(x) := \|F(x) - y\|^2$.

1.2 Einfache mathematische Beispiele

1.2.1 Unrestringierte lineare Optimierung

Für $X = \mathbb{R}^n$ spricht man auch von *unrestringierter* Optimierung. Betrachten wir den einfachsten Fall, dass die Zielfunktion linear ist, also

$$f: \mathbb{R}^n \to \mathbb{R}, \quad f(x) := b^T x \quad \forall x \in \mathbb{R}^n$$

mit einem $b \in \mathbb{R}^n$.

Für $x_k = -kb$ gilt $f(x_k) = -k\|b\|^2 \to -\infty$, die Minimierungsaufgabe

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

besitzt also keine (globale) Lösung. f besitzt auch kein lokales Minimum (siehe Übungsaufgabe 1.1).

1.2.2 Lineare Ausgleichsrechnung

Lässt sich ein lineares Gleichungssystem

$$Ax = b,$$

mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ nicht exakt lösen, so begnügt man sich in der Praxis oft mit der sogenannten *Kleinsten-Quadrate-Lösung*, d.h. der Lösung des Minimierungsproblems

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

mit $f(x) = \|Ax - b\|^2$. Die globalen Minima von f sind genau die Lösungen der *Normalengleichungen*

$$A^T Ax = A^T b$$

(siehe Übungsaufgabe 1.2).

1.3 Komplexere Anwendungsbeispiele

Wir skizzieren noch drei mit bedeutenden Preisen verbundene Anwendungsbeispiele für Optimierungsprobleme.

1.3.1 Minimalflächen[1]

Zieht man einen geschlossenen Draht durch Seifenlauge, so bildet sich ein Film dessen Flächeninhalt minimal wird. Mathematisch können wir dies modellieren, indem wir zu einem Gebiet $\Omega \subseteq \mathbb{R}^2$ und Randdaten $r : \partial\Omega \to \mathbb{R}$ eine Funktion

$$u : \overline{\Omega} \to \mathbb{R}$$

suchen, die auf dem Rand mit r übereinstimmt und deren Graph minimale Oberfläche besitzt (vgl. die in der Vorlesung gemalten Skizze).

Zur näherungsweise numerischen Lösung dieses Problems diskretisieren wir die gesuchte Funktion u durch einen Vektor $U \in \mathbb{R}^n$ und stellen einen funktionalen Zusammenhang

$$A : \mathbb{R}^n \to \mathbb{R}$$

zwischen der Diskretisierung U und dem Flächeninhalt $A(U)$ des zugehörigen Graphen auf. (U enthält beispielsweise die Auswertungen von u in den Knotenpunkten einer regulären Triangulierung von Ω und $A(U)$ ist der Flächeninhalt des Graphen der zugehörigen stetigen und stückweise linearen Funktion, vgl. die in der Vorlesung gemalten Skizzen).

[1] Jesse Douglas erhielt 1936 die Fields-Medaille für seinen Existenzbeweis für das Minimalflächenproblem.

Die diskretisierte Version des Minimalflächenproblem führt also auf ein Optimierungsproblem der Form

$$A(U) \to \min! \quad \text{u.d.N. } U \in \mathbb{R}^n$$

wobei $n \in \mathbb{N}$ die Anzahl der Freiheitsgrade in der Diskretisierung beschreibt und damit (bei immer besserer Diskretisierung) beliebig groß werden kann.

Wird die Seifenhaut über ein Hindernis gespannt, so ergeben sich entsprechend restringierte Optimierungsprobleme.

1.3.2 Portfoliotheorie nach Markowitz[2]

Ein Fondsmanager versucht n Finanzprodukte (Anleihen, Aktien, Derivate, ...) so zu kombinieren, dass er möglichst viel Gewinn bei möglichst geringem Risiko macht. Ein einfaches Modell ist, dass der zukünftige Wert (etwa nach Ablauf eines Jahres) der n Finanzprodukte zufällig verteilt ist und dass der Manager (aufgrund seiner Markterwartung oder aus historischen Daten) die Erwartungswerte μ_j der Renditen kennt. Investiert er x_j Euro in das j-te Finanzprodukt, so wird der zukünftige Wert des Portfolios zufallsverteilt sein mit Erwartungswert $E(x) := \sum_{j=1}^n \mu_j x_j$.

Um einen möglichst großen (erwarteten) Gewinn zu machen, betrachtet der Fondsmanager also das Optimierungsproblem

$$E(x) = \mu^T x = \sum_{j=1}^n \mu_j x_j \to \max!$$

unter der Nebenbedingung $\sum_{j=1}^n x_j = 1$, wobei wir o.B.d.A. den den zur Verfügung stehenden Betrag auf 1 normiert haben. Dieses Optimierungsproblem wird offenbar trivialerweise dadurch gelöst, alles in das Finanzprodukt mit der höchsten erwarteten Rendite zu investieren.

Kennt der Fondsmanager (wiederum aufgrund seiner Markterwartung oder aus historischen Daten) auch die Kovarianzmatrix $\Sigma = (\sigma_{jk})_{j,k=1}^n \in \mathbb{R}^{n\times n}$ der Renditen, so ist die Varianz des zukünftigen Wert seines Portfolios gegeben durch

$$V(x) = x^T \Sigma x = \sum_{j,k=1}^n x_j \Sigma_{jk} x_k.$$

[2]Harry M. Markowitz erhielt 1990 für seine Theorie der Portfolio-Auswahl den Wirtschaftsnobelpreis zusammen mit Merton H. Miller und William Sharpe.

Akzeptieren wir die Varianz als Maß für das Risiko der gewählten Investmentstrategie, dann führt die Maximierung des Gewinns bei einer gegebenen Risikobereitschaft auf Optimierungsprobleme der Form

$$\mu^T x \to \max! \quad \text{u.d.N. } x \in \mathbb{R}^n,\ x^T \Sigma x \leq R$$

und die Minimierung des Risikos bei einer festgesetzten Gewinnabsicht auf Optimierungsprobleme der Form

$$x^T \Sigma x \to \min! \quad \text{u.d.N. } x \in \mathbb{R}^n,\ \mu^T x \geq r.$$

Weitere Nebenbedingungen können aus der Art der Finanzprodukte erwachsen (z.B. $x_j \geq 0$).

Es ist naheliegend, nur Portfolio zu verwenden die beide obigen Optimierungsprobleme lösen, da ansonsten der gleiche erwartete Gewinn mit geringerem Risiko oder ein höherer Gewinn bei gleichen Risiko möglich wäre. Solche Portfolios heißen *effizient*.

1.3.3 Computertomographie[3]

In der Computertomographie wird ein Schnittbild des Inneren eines Patienten erstellt, indem der Patient von verschiedenen Seiten mit Röntgenstrahlen durchleuchtet wird. In einer einfachen diskreten Modellierung unterteilen wir das gesuchte Schnittbild in n Pixel und nehmen an, dass die Gesamtabsorption eines Röntgenstrahls durch den Körper sich als Summe der Absorptionskoeffizienten aller passierten Pixel gewichtet mit der Strahllänge durch den Pixel ergibt (vgl. die in der Vorlesung gemalten Skizzen). So erhalten wir ein großes lineares Gleichungssystem

$$Ax = b.$$

Dabei ist x_j der Absorptionskoeffizienten des j-ten Pixels, b_i ist die Gesamtabschwächung des i-ten Röntgenstrahls und die Einträge a_{ij} ($i = 1, ..., m$, $j = 1, ..., n$) der Matrix A ergeben sich aus der Länge des Weges des i-ten Röntgenstrahls durch das j-te Pixel. Wählen wir weniger Pixel als Strahlen ($n < m$), so ist das LGS überbestimmt und wir können (aufgrund der Mess- und Modellierungsfehler) nicht erwarten, dass eine exakte Lösung existiert.

[3]Allan M. Cormack und Godfrey Hounsfield erhielten 1979 für die Entwicklung der Computertomographie den Nobelpreis für Medizin. Die mathematischen Grundlagen der heute genutzten Algorithmen wurden 1917 durch den österreichischen Mathematiker Johann Radon entwickelt.

Wie in Abschnitt 1.2.2 führt das Problem, eine bestmögliche Lösung des LGS zu finden auf das Ausgleichsproblem

$$\|Ax - b\| \to \min! \quad \text{u.d.N. } x \in \mathbb{R}^n.$$

Neuartige Tomographieverfahren beruhen häufig auf nicht-linearen Zusammenhängen zwischen den gesuchten physikalischen Koeffizienten und den gemessenen Größen und damit auf nicht-lineare Ausgleichsprobleme der Form

$$\|F(x) - b\|^2 \to \min! \quad \text{u.d.N. } x \in \mathbb{R}^n.$$

Es muss an dieser Stelle erwähnt werden, dass Messfehler die *optimal zu den gemessenen Daten passenden Lösungen* in der Praxis häufig bis zur völligen Unbrauchbarkeit verfälschen. Auswege aus diesem sogenannten Problem der *Schlechtgestelltheit* sind Gegenstand der im nächsten Semester angebotenen Vorlesung *Regularisierung inverser Probleme*.

Kapitel 2

Unrestringierte nichtlineare Optimierung

Dieses Kapitel folgt dem sehr empfehlenswerten Lehrbuch von Michael und Stefan Ulbrich [Ulbrich]. Wir betrachten das unrestringierte Optimierungsproblem

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

für eine allgemeine (hinreichend glatte) Zielfunktion $f: \mathbb{R}^n \to \mathbb{R}$.

2.1 Grundlagen und Optimalitätsbedingungen

Wir wiederholen einige Grundbegriffe aus der Analysis und der Linearen Algebra. In diesem Abschnitt sei $U \subseteq \mathbb{R}^n$ stets eine offene Menge.

2.1.1 Grundlagen

Für Vektoren $x \in \mathbb{R}^n$ und Matrizen $A \in \mathbb{R}^{m \times n}$ bezeichnen wir mit

$$\|x\| := \sqrt{x^T x} \quad \text{und} \quad \|A\| = \max_{\|x\|=1} \|Ax\|$$

die Euklidische Vektornorm und die dadurch induzierte Matrixnorm.

Für eine auf U stetig differenzierbare Funktion

$$F: U \to \mathbb{R}^m, \quad F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}$$

bezeichnen wir die Jacobi-Matrix mit

$$F'(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

und für eine skalare, auf U ein- bzw. zweimal stetig differenzierbare Funktion $f : U \to \mathbb{R}$ bezeichne

$$\nabla f(x) = f'(x)^T = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} \in \mathbb{R}^n,$$

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial f(x)}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial f(x)}{\partial x_n^2} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

den *Gradient* bzw. die *Hesse-Matrix* von f.[1]

Für jedes $d \in \mathbb{R}^n$ können wir die *Richtungsableitung* schreiben als

$$\frac{\partial f}{\partial d} = \lim_{t \to 0} \frac{f(x + td) - f(x)}{t} = \nabla f(x)^T d. \tag{2.1}$$

Definition 2.1
Zu zwei Punkten $x_0, x_1 \in \mathbb{R}^n$ heißt

$$[x_0, x_1] := \{x_t := (1 - t)x_0 + tx_1 \ : \ t \in [0, 1]\}$$

***Verbindungsstrecke** zwischen x_0 und x_1.*

Satz 2.2
Sei $f : U \to \mathbb{R}$ stetig differenzierbar und $x_0, x_1 \in U$ seien Punkte mit $[x_0, x_1] \subseteq U$.

Die Funktion

$$g : \ t \mapsto f(x_t), \quad x_t := (1 - t)x_0 + tx_1$$

ist auf einer Umgebung von $[0, 1]$ stetig differenzierbar und es gilt

$$g'(t) = \nabla f(x_t)^T (x_1 - x_0).$$

[1] Achtung: Die Notation ∇^2 wird in der Literatur nicht einheitlich verwendet. Im Kontext partieller Differentialgleichung findet sich ∇^2 auch als Notation für den Laplace-Operator.

Ist f zweimal stetig differenzierbar, dann ist auch g zweimal stetig differenzierbar und es gilt

$$g''(t) = (x_1 - x_0)^T \nabla^2 f(x_t)(x_1 - x_0).$$

Beweis: Übungsaufgabe 1.4. □

Satz 2.3 (Lineare & quadratische Taylor-Formel skalarer Fkt.)
Sei $f : U \to \mathbb{R}$ stetig differenzierbar, $x \in U$ und $\epsilon > 0$, so dass $B_\epsilon(x) \subseteq U$.

(a) Für alle $d \in B_\epsilon(0)$ existiert ein $s \in [0, 1]$ mit

$$f(x + d) = f(x) + \nabla f(x + sd)^T d.$$

Außerdem gilt für alle $d \in B_\epsilon(0)$

$$f(x + d) = f(x) + \nabla f(x)^T d + \rho(d)$$

und das Restglied $\rho : \overline{B_\epsilon(0)} \to \mathbb{R}$ *erfüllt $\rho(d) = o(\|d\|)$, d.h.*

$$\lim_{d \to 0} \frac{|\rho(d)|}{\|d\|} = 0.$$

(b) Ist $f : U \to \mathbb{R}$ zweimal stetig differenzierbar, dann existiert für alle $d \in B_\epsilon(0)$ ein $s \in [0, 1]$ mit

$$f(x + d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x + sd) d$$

Außerdem gilt für alle $d \in B_\epsilon(0)$

$$f(x + d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d + \rho(d)$$

und das Restglied $\rho : \overline{B_\epsilon(0)} \to \mathbb{R}$ *erfüllt $\rho(d) = o(\|d\|^2)$, d.h.*

$$\lim_{d \to 0} \frac{|\rho(d)|}{\|d\|^2} = 0.$$

Beweis: Das folgt mit Satz 2.2 sofort aus den eindimensionalen Taylor-Formeln. □

Für vektorwertige Funktionen existieren die Zwischenwerte in Satz 2.3 im Allgemeinen nicht mehr. Mit dem folgenden Satz lässt sich aber dennoch das Verhalten der Funktion zwischen zwei Werten x und $x + d$ durch das Verhalten ihrer Ableitung zwischen diesen Werten abschätzen.

Lemma 2.4
Sei $F : U \to \mathbb{R}^m$ stetig differenzierbar, $x \in U$ und $\epsilon > 0$, so dass $B_\epsilon(x) \subseteq U$. Für alle $d \in B_\epsilon(0)$ gilt

$$F(x+d) = F(x) + \int_0^1 F'(x+td)d \,\mathrm{d}t = F(x) + \left(\int_0^1 F'(x+td) \,\mathrm{d}t \right) d,$$

also insbesondere

$$\|F(x+d) - F(x)\| \leq \|d\| \sup_{t\in[0,1]} \|F'(x+td)\| .$$

Außerdem ist

$$F(x+d) = F(x) + F'(x)d + \rho(d),$$

und das Restglied $\rho : \overline{B_\epsilon(0)} \to \mathbb{R}^m$ *erfüllt* $\rho(d) = o(\|d\|)$, *d.h.*

$$\lim_{d\to 0} \frac{\|\rho(d)\|}{\|d\|} = 0.$$

Beweis: Da U offen ist, existiert $\delta > 0$ so dass $[x - \delta d, x + (1+\delta)d] \subset U$. Wir definieren

$$f : \ (-\delta, 1+\delta) \to \mathbb{R}^m, \quad f(t) := F(x+td)$$

Offenbar ist $f(t) = F(g(t))$, wobei

$$g : (-\delta, 1+\delta) \to \mathbb{R}^n, \quad g(t) := x + td$$

und es ist $g'(t) = d$.

Aus der mehrdimensionalen Kettenregel folgt

$$f'(t) = F'(g(t))g'(t) = F'(x+td)d$$

und mit dem Hauptsatz der Differential- und Integralrechnung erhalten wir

$$F(x+d) - F(x) = f(1) - f(0) = \int_0^1 f'(t) \,\mathrm{d}t = \int_0^1 F'(x+td))d \,\mathrm{d}t$$

und damit die erste Behauptung.

Die zweite Behauptung ist nichts anderes als die Definition der totalen Differenzierbarkeit. □

Nach dem Satz von Schwarz ist die Hesse-Matrix einer zweimal stetig differenzierbaren Funktion symmetrisch.

Definition 2.5
Eine symmetrische Matrix $A = (a_{ij})_{i,j=1,\dots,n} \in \mathbb{R}^{n\times n}$ heißt

(a) ***positiv semidefinit****, falls $x^T Ax \geq 0$ für alle $x \in \mathbb{R}^n$*

(b) ***positiv definit****, falls $x^T Ax > 0$ für alle $0 \neq x \in \mathbb{R}^n$*

A heißt ***negativ (semi-)definit****, falls $-A$ positiv (semi-)definit ist.*

Satz 2.6
Zu einer symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$ existiert eine ONB aus Eigenvektoren

$$v_1, \dots, v_n \in \mathbb{R}^n$$

mit zugehörigen Eigenwerten $\lambda_1, \dots, \lambda_n \in \mathbb{R}$, also

$$v_j^T v_k = \delta_{jk} \quad \text{und} \quad Av_k = \lambda_k v_k.$$

Es ist also

$$A = \sum_{k=1}^{n} \lambda_k v_k v_k^T$$

und mit $\lambda_{min}(A) := \min_{k=1,\dots,n} \lambda_k$, $\lambda_{max}(A) := \max_{k=1,\dots,n} \lambda_k$ gilt

$$\lambda_{min}(A) = \min_{0\neq x\in\mathbb{R}^n} \frac{x^T Ax}{\|x\|^2} \leq \frac{x^T Ax}{\|x\|^2} \leq \max_{0\neq x\in\mathbb{R}^n} \frac{x^T Ax}{\|x\|^2} = \lambda_{max}(A).$$

Insbesondere ist eine symmetrische Matrix A also genau dann positiv semidefinit (bzw. positiv definit / negativ semidefinit / negativ definit), wenn alle Eigenwerte nicht-negativ (bzw. positiv / nicht-positiv / negativ) sind.

Für symmetrische, positiv semidefinite $A \in \mathbb{R}^{n\times n}$ gilt außerdem

$$\|A\| = \max_{k=1,\dots,n} \lambda_k.$$

Beweis: Wir setzen dies als bekannt voraus und verweisen auf die Grundvorlesungen zur linearen Algebra. □

Definition 2.7
Für eine invertierbare Matrix $A \in \mathbb{R}^{n\times n}$ definieren wir die ***Kondition*** *durch*

$$\kappa(A) = ||A||\ ||A^{-1}||.$$

Für symmetrische positiv definite Matrizen gilt offenbar

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

Bemerkung 2.8
Der Name Kondition *kommt daher, dass $\kappa(A)$ wegen*

$$\frac{\|A^{-1}(b+\delta) - A^{-1}b\|}{\|A^{-1}b\|} = \left\|A^{-1}\delta\right\| \left\|A^{-1}b\right\|^{-1}$$
$$\leq \|A\| \left\|A^{-1}\right\| \frac{\|\delta\|}{\|A\| \, \|A^{-1}b\|} \leq \kappa(A) \frac{\|\delta\|}{\|b\|}$$

ein Maß für die relative Fehlerverstärkung bei Inversion von A ist.

2.1.2 Optimalitätsbedingungen

Definition 2.9
Sei $\emptyset \neq X \subseteq \mathbb{R}^n$ und $f: X \to \mathbb{R}$. Ein Punkt $x \in X$ heißt

(a) ***globales Minimum*** *von f in X, falls*

$$f(x) \leq f(y) \quad \text{für alle } y \in X.$$

(b) ***lokales Minimum*** *von f in X, falls ein $\epsilon > 0$ existiert, so dass x globales Minimum von f in $X \cap B_\epsilon(x)$ ist, also*

$$f(x) \leq f(y) \quad \text{für alle } y \in X \cap B_\epsilon(x).$$

(c) ***striktes***[2] globales Minimum *von f, falls*

$$f(x) < f(y) \quad \text{für alle } x \neq y \in X.$$

(d) ***striktes*** lokales Minimum *von f, falls ein $\epsilon > 0$ existiert, so dass x striktes globales Minimum von f in $X \cap B_\epsilon(x)$ ist, also*

$$f(x) < f(y) \quad \text{für alle } x \neq y \in X \cap B_\epsilon(x).$$

Analog definieren wir (strikte) globale oder lokale Maxima.

Definition und Satz 2.10 (Notwendige Optimalitätsbed. 1. Ordnung)
Sei $\emptyset \neq U \subseteq \mathbb{R}^n$ offen und $f: U \to \mathbb{R}$ stetig differenzierbar. Ist $x \in U$ ein lokales Minimum von f, dann gilt $\nabla f(x) = 0$.

Punkte mit $\nabla f(x) = 0$ heißen ***stationäre Punkte****.*

[2] auch: *strenges* oder *isoliertes*

Beweis: Für jedes $d \in \mathbb{R}^n$ gilt gemäß (2.1)

$$\nabla f(x)^T d = \lim_{t \to 0^+} \frac{f(x+td) - f(x)}{t} \geq 0.$$

Für $d := -\nabla f(x)$ folgt, dass $-\|\nabla f(x)\|^2 \geq 0$ und damit $\nabla f(x) = 0$. □

Satz 2.11 (Notwendige Optimalitätsbed. 2. Ordnung)
Sei $\emptyset \neq U \subseteq \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ zweimal stetig differenzierbar. Ist $x \in U$ ein lokales Minimum von f, dann ist x ein stationärer Punkt und $\nabla^2 f(x)$ ist positiv semidefinit, also

(a) $\nabla f(x) = 0$

(b) $d^T \nabla^2 f(x) d \geq 0$ *für alle* $d \in \mathbb{R}^n$

Beweis: Definition und Satz 2.10 liefert (a). Zum Beweis von (b) sei $d \in \mathbb{R}^n$ und $\epsilon > 0$ sei so klein, dass $B_\epsilon(x) \subseteq U$ und x globales Minimum in $B_\epsilon(x)$ ist.

Wir verwenden die quadratische Taylor-Formel in Satz 2.3 mit td anstelle von d

$$f(x+td) = f(x) + t\nabla f(x)^T d + \frac{t^2}{2} d^T \nabla^2 f(x) d + \rho(td) \quad \forall |t| < \frac{\epsilon}{\|d\|}.$$

Mit $f(x+td) \geq f(x)$ und $\nabla f(x) = 0$ folgt, dass

$$d^T \nabla^2 f(x) d \geq -\frac{2}{t^2} \rho(td) \quad \forall |t| < \frac{\epsilon}{\|d\|}$$

und mit $\frac{|\rho(td)|}{|t|^2} \to 0$ erhalten wir, dass $d^T \nabla^2 f(x) d \geq 0$. □

Satz 2.12 (Hinreichende Optimalitätsbed. 2. Ordnung)
Sei $\emptyset \neq U \subseteq \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ zweimal stetig differenzierbar. Ist $x \in U$ ein stationärer Punkt mit positiv definiter Hesse-Matrix, also

(a) $\nabla f(x) = 0$

(b) $d^T \nabla^2 f(x) d > 0$ *für alle* $0 \neq d \in \mathbb{R}^n$

dann ist x ein striktes lokales Minimum von f.

Beweis: Wegen (b) und Satz 2.6 ist $\mu := \lambda_{\min}(\nabla^2 f(x)) > 0$ und

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

Wir verwenden wieder die quadratische Taylor-Formel aus Satz 2.3. Für hinreichend kleine $\epsilon > 0$ ist $B_\epsilon(x) \subseteq U$,

$$f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d + \rho(d) \quad \forall d \in B_\epsilon(0)$$

und

$$|\rho(d)| \leq \frac{\mu}{4} \|d\|^2 \quad \forall d \in B_\epsilon(0).$$

Wegen $\nabla f(x) = 0$ gilt also

$$f(x+td) \geq f(x) + \frac{\mu}{4} \|d\|^2 \quad \forall d \in B_\epsilon(0),$$

so dass x ein striktes lokales Minimum ist. □

Beispiel 2.13

(a) Die notwendige Optimalitätsbedingung 1. Ordnung ist nicht hinreichend. Nicht jeder stationäre Punkt ist ein Minimum. Für $f(x) = -x^2$, $x \in \mathbb{R}$ ist der einzige stationäre Punkt $x = 0$ ein Maximum. Für $f(x) = x^3$ ist der einzige stationäre Punkt weder ein Minimum noch ein Maximum (solche Punkte heißen auch Sattelpunkte*, vgl. die in der Vorlesung gemalte Skizze).*

(b) Die notwendige Optimalitätsbedingung 2. Ordnung ist nicht hinreichend. Für

$$f(x) := x^3 \quad \textit{ist} \quad \nabla f(0) = f'(0) = 0, \quad \nabla^2 f(0) = f''(0) = 0.$$

Die Hesse-Matrix ist im stationären Punkt $x = 0$ also positiv semidefinit, aber $x = 0$ ist weder Minimum noch Maximum.

(c) Die hinreichende Optimalitätsbedingung 2. Ordnung ist nicht notwendig. Für $f(x) = x^4$ ist der stationäre Punkt $x = 0$ ein Minimum, in dem die Hesse-Matrix $\nabla^2 f(0) = f''(0) = 0$ nicht positiv definit ist.

2.1.3 Konvexität

Wir werden in dieser Vorlesung Algorithmen entwickeln, mit denen stationäre Punkte gefunden werden können. Für konvexe Funktionen stimmen diese mit den globalen Minima überein.

Definition 2.14
*Eine Menge $X \subset \mathbb{R}^n$ heißt **konvex**, falls alle Verbindungsstrecken zweier Punkte in X verlaufen, also*

$$(1-\lambda)x_0 + \lambda x_1 \in X \quad \forall x_0, x_1 \in X,\ \lambda \in [0,1].$$

Zu gegebenen $x_0, x_1 \in X$ schreiben wir auch kurz

$$x_\lambda := (1-\lambda)x_0 + \lambda x_1 \quad \forall \lambda \in [0,1].$$

Definition 2.15
Sei $X \subset \mathbb{R}^n$ eine konvexe Menge. Eine Funktion $f: X \to \mathbb{R}$ heißt

(a) ***konvex**, falls*

$$f((1-\lambda)x_0 + \lambda x_1) \leq (1-\lambda)f(x_0) + \lambda f(x_1)$$

für alle $x_0, x_1 \in X$, $\lambda \in [0,1]$.

(b) ***strikt**[3] **konvex**, falls*

$$f((1-\lambda)x_0 + \lambda x_1) < (1-\lambda)f(x_0) + \lambda f(x_1)$$

für alle $x_0, x_1 \in X$, $x_0 \neq x_1$, $\lambda \in (0,1)$.

(c) ***gleichmäßig konvex**, falls ein $\mu > 0$ existiert mit*

$$f((1-\lambda)x_0 + \lambda x_1) + \mu\lambda(1-\lambda)\left\|x_1 - x_0\right\|^2 \leq (1-\lambda)f(x_0) + \lambda f(x_1)$$

für alle $x_0, x_1 \in X$, $\lambda \in [0,1]$.

Satz 2.16
Sei $U \subseteq \mathbb{R}^n$ offen und $f: U \to \mathbb{R}$ stetig differenzierbar. f ist auf einer konvexen Menge $X \subseteq U$

(a) genau dann konvex, wenn

$$\nabla f(x_0)^T(x_1 - x_0) \leq f(x_1) - f(x_0) \quad \forall x_0, x_1 \in X.$$

(b) genau dann strikt konvex, wenn

$$\nabla f(x_0)^T(x_1 - x_0) < f(x_1) - f(x_0) \quad \forall x_0, x_1 \in X,\ x_0 \neq x_1.$$

[3] auch: streng

(c) genau dann gleichmäßig konvex, wenn ein $\mu > 0$ existiert mit

$$\nabla f(x_0)^T(x_1 - x_0) + \mu \|x_1 - x_0\|^2 \leq f(x_1) - f(x_0) \quad \forall x_0, x_1 \in X.$$

Beweis: (a) „$\Longrightarrow$": Ist f konvex, dann ist

$$\begin{aligned}\nabla f(x_0)^T(x_1 - x_0) &= \lim_{t\to 0} \frac{f(x_0 + t(x_1 - x_0)) - f(x_0)}{t} \\ &= \lim_{t\to 0} \frac{f((1-t)x_0 + tx_1) - f(x_0)}{t} \\ &\leq \lim_{t\to 0} \frac{(1-t)f(x_0) + tf(x_1) - f(x_0)}{t} = f(x_1) - f(x_0).\end{aligned}$$

(a) „$\Longleftarrow$": Sei $\lambda \in [0,1]$ und $x_0, x_1 \in X$. Für $x_\lambda := (1-\lambda)x_0 + \lambda x_1$ gilt nach Voraussetzung

$$\begin{aligned}f(x_0) - f(x_\lambda) &\geq \nabla f(x_\lambda)^T(x_0 - x_\lambda) \\ f(x_1) - f(x_\lambda) &\geq \nabla f(x_\lambda)^T(x_1 - x_\lambda)\end{aligned}$$

und damit

$$\begin{aligned}&(1-\lambda)f(x_0) + \lambda f(x_1) \\ &= (1-\lambda)(f(x_0) - f(x_\lambda)) + \lambda(f(x_1) - f(x_\lambda)) + f(x_\lambda) \\ &\geq (1-\lambda)\nabla f(x_\lambda)^T(x_0 - x_\lambda) + \lambda \nabla f(x_\lambda)^T(x_1 - x_\lambda) + f(x_\lambda) \\ &= \nabla f(x_\lambda)^T((1-\lambda)x_0 + \lambda x_1 - x_\lambda) + f(x_\lambda) = f(x_\lambda).\end{aligned}$$

(b) „$\Longrightarrow$": Sei f strikt konvex und $x_0, x_1 \in X$, $x_0 \neq x_1$. Für $x_{\frac{1}{2}} = \frac{x_0 + x_1}{2}$ gilt

$$\begin{aligned}\nabla f(x_0)^T(x_1 - x_0) = 2\nabla f(x_0)^T(x_{\frac{1}{2}} - x_0) &\overset{(a)}{\leq} 2\left(f(x_{\frac{1}{2}}) - f(x_0)\right) \\ &< 2\left(\frac{f(x_0) + f(x_1)}{2} - f(x_0)\right) = f(x_1) - f(x_0).\end{aligned}$$

(b) „$\Longleftarrow$": wie (a) „$\Longleftarrow$" mit „$<$" statt „$\leq$"

(c) „$\Longrightarrow$": Ist f gleichmäßig konvex, dann folgt wie in (a)

$$\begin{aligned}&\nabla f(x_0)^T(x_1 - x_0) = \lim_{t\to 0} \frac{f(x_0 + t(x_1 - x_0)) - f(x_0)}{t} \\ &= \lim_{t\to 0} \frac{f((1-t)x_0 + tx_1) - f(x_0)}{t} \\ &\leq \lim_{t\to 0} \frac{(1-t)f(x_0) + tf(x_1) - f(x_0) - \mu t(1-t)\|x_1 - x_0\|^2}{t} \\ &= f(x_1) - f(x_0) - \mu \|x_1 - x_0\|^2.\end{aligned}$$

(c) „⟸": Sei $\lambda \in [0,1]$, $x_0, x_1 \in X$ und $x_\lambda := (1-\lambda)x_0 + \lambda x_1$. Mit

$$x_0 - x_\lambda = \lambda(x_0 - x_1) \quad \text{und} \quad x_\lambda - x_1 = (1-\lambda)(x_0 - x_1)$$

folgt wie in (a)

$$\begin{aligned}
&(1-\lambda)f(x_0) + \lambda f(x_1) \\
&= (1-\lambda)(f(x_0) - f(x_\lambda)) + \lambda(f(x_1) - f(x_\lambda)) + f(x_\lambda) \\
&\geq (1-\lambda)\left(\nabla f(x_\lambda)^T(x_0 - x_\lambda) + \mu \|x_0 - x_\lambda\|^2\right) \\
&\qquad + \lambda\left(\nabla f(x_\lambda)^T(x_1 - x_\lambda) + \mu \|x_1 - x_\lambda\|^2\right) + f(x_\lambda) \\
&= \left((1-\lambda)\mu\lambda^2 + \lambda\mu(1-\lambda)^2\right)\|x_1 - x_0\|^2 + f(x_\lambda) \\
&= \mu(1-\lambda)\lambda \|x_1 - x_0\|^2 + f(x_\lambda),
\end{aligned}$$

womit alle Behauptungen gezeigt sind. □

Satz 2.17
Sei $X \subseteq \mathbb{R}^n$ offen und konvex. Eine zweimal stetig differenzierbare Funktion $f: X \to \mathbb{R}$ ist

(a) genau dann konvex, wenn $\nabla^2 f(x)$ für alle $x \in X$ positiv semidefinit ist, also

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

(b) strikt konvex, falls $\nabla^2 f(x)$ für alle $x \in X$ positiv definit ist, also

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

(c) genau dann gleichmäßig konvex, wenn $\nabla^2 f(x)$ auf X gleichmäßig positiv definit ist, d.h. ein $\mu > 0$ existiert, so dass

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

Beweis: (a) „⟹": Sei f konvex und $d \in \mathbb{R}^n$. Für alle hinreichend kleinen $t > 0$ gilt unter Verwendung der quadratischen Taylor-Formel in Satz 2.3 und Satz 2.16[4]

$$\frac{t^2}{2} d^T \nabla^2 f(x) d + \rho(td) = f(x+td) - f(x) - t\nabla f(x)^T d \geq 0,$$

also

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

und mit $\frac{\rho(td)}{t^2} \to 0$ folgt $d^T \nabla^2 f(x) d \geq 0$.

[4]Hier wird die Voraussetzung, dass X offen ist, benötigt, damit $x + td \in X$ für hinreichend kleine $t > 0$ gilt.

(a) „$\Longleftarrow$": Seien $x_0, x_1 \in X$. Nach Satz 2.3 existiert ein $s \in [0,1]$ mit

$$\begin{aligned} f(x_1) &= f(x_0) + \nabla f(x_0)^T d + \frac{1}{2}(x_1 - x_0)^T \nabla^2 f(x_0 + s(x_1 - x_0))(x_1 - x_0) \\ &\geq f(x_0) + \nabla f(x_0)^T (x_1 - x_0), \end{aligned}$$

so dass aus Satz 2.16 die Konvexität von f folgt.

(b) „$\Longleftarrow$": analog (a) „$\Longleftarrow$" mit „$>$" statt „$\geq$"

(c) „$\Longrightarrow$": Analog (a) „$\Longrightarrow$" erhalten wir

$$\begin{aligned} &\frac{t^2}{2} d^T \nabla^2 f(x) d + \rho(td) - \mu \left\| td \right\|^2 \\ &\quad = f(x + td) - f(x) - t \nabla f(x)^T d - \mu \left\| td \right\|^2 \geq 0, \end{aligned}$$

also

$$d^T \nabla^2 f(x) d \geq \mu \left\| d \right\|^2 - \frac{2\rho(td)}{t^2}$$

und mit $\frac{\rho(td)}{t^2} \to 0$ folgt $d^T \nabla^2 f(x) d \geq \mu \left\| d \right\|^2$.

(c) „$\Longleftarrow$": Analog (a) „$\Longleftarrow$" existiert zu $x_0, x_1 \in X$ ein $s \in [0,1]$ mit

$$\begin{aligned} f(x_1) &= f(x_0) + \nabla f(x_0)^T d + \frac{1}{2}(x_1 - x_0)^T \nabla^2 f(x_0 + s(x_1 - x_0))(x_1 - x_0) \\ &\geq f(x_0) + \nabla f(x_0)^T (x_1 - x_0) + \mu \left\| x_1 - x_0 \right\|^2, \end{aligned}$$

so dass aus Satz 2.16 die gleichmäßige Konvexität folgt. □

Beispiel 2.18
$f(x) = x^4$ ist strikt konvex, aber $\nabla^2 f(x) = f''(x) = 12x^2$ ist in $x = 0$ nicht positiv definit. Die fehlende Implikation in Satz 2.17(b) gilt also tatsächlich nicht.

Satz 2.19
Sei $X \subseteq \mathbb{R}^n$ eine konvexe Menge und $f: X \to \mathbb{R}$ eine konvexe Funktion.

(a) Jedes lokale Minimum von f auf X ist auch globales Minimum von f auf X.

(b) Ist f strikt konvex, dann besitzt f höchstens ein lokales Minimum auf X. Falls ein lokales Minimum existiert, so ist es nach (a) das strikte globale Minimum.

(c) Ist f stetig differenzierbar auf einer X enthaltenden offenen Menge, dann ist jeder stationäre Punkt in X ein globales Minimum von f in X.

Beweis: (a) Sei $x \in X$ kein globales Minimum von f. Dann existiert ein $y \in X$ mit $f(y) < f(x)$. Aus der Konvexität folgt, dass

$$f(x + t(y - x)) \leq (1 - t)f(x) + tf(y) < f(x) \quad \forall t \in (0, 1],$$

so dass x kein lokales Minimum sein kann.

(b) Seien $x, y \in X$ zwei lokale Minima von f. Nach (a) sind x und y dann auch globale Minima und insbesondere ist $f(x) = f(y)$. Wäre $x \neq y$, so folgt aus der strikten Konvexität, dass

$$f(\frac{x + y}{2}) < \frac{f(x) + f(y)}{2} = f(x) = f(y),$$

und damit ein Widerspruch zur globalen Minimalität. Es ist also $x = y$.

(c) Ist $x \in X$ ein stationärer Punkt, dann gilt nach Satz 2.16

$$f(y) - f(x) \geq \nabla f(x)^T (y - x) = 0.$$

Jeder stationäre Punkt ist also globales Minimum. □

Zur Konvergenzanalyse von Minimierungsverfahren ist außerdem das folgende Lemma hilfreich.

Lemma 2.20
Ist $f : \mathbb{R}^n \to \mathbb{R}$ konvex und stetig und besitzt f ein eindeutiges globales Minimum, so ist für jedes $C > 0$ die zugehörige Niveaumenge

$$f^{-1}(]-\infty, C]) := \{x \in \mathbb{R}^n : \ f(x) \leq C\}$$

von f beschränkt.

Beweis: Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex und stetig und $\hat{x} \in \mathbb{R}^n$ sei das eindeutige globale Minimum von f. Da f stetig ist, können wir zu jedem Radius $r > 0$

$$C_r := \min\{f(x) : \ x \in \partial B_r(\hat{x})\}.$$

definieren und aufgrund der Eindeutigkeit des globalen Minimums gilt $C_r > f(\hat{x})$ für alle $r > 0$.

Wir werden zeigen, dass

(a) $f^{-1}(]-\infty, C_r]) \subseteq \overline{B_r(\hat{x})}$,

(b) $C_r \to \infty$ für $r \to \infty$,

womit dann offenbar die Behauptung bewiesen ist.
Sei $r > 0$ und $x \notin \overline{B_r(\hat{x})}$. Dann gilt für $t := \frac{r}{\|x-\hat{x}\|}$

$$\hat{x} + t(x - \hat{x}) \in \partial B_r(\hat{x}) \quad \text{und} \quad 0 < t < 1.$$

Da $\hat{x}$ das einzige globale Minimum ist, gilt $f(x) > f(\hat{x})$. Mit der Konvexität von f folgt, dass

$$C_r \leq f(\hat{x}+t(x-\hat{x})) \leq (1-t)f(\hat{x})+tf(x) < (1-t)f(x)+tf(x) = f(x). \quad (2.2)$$

Damit ist gezeigt, dass $f(x) > C_r$ für alle $x \notin \overline{B_r(\hat{x})}$ gilt, woraus durch Kontraposition (a) folgt.
Für alle $x \notin \overline{B_r(\hat{x})}$ folgt aus der in (2.2) gezeigten Ungleichung

$$C_r \leq (1-t)f(\hat{x}) + tf(x),$$

außerdem dass

$$f(x) \geq \frac{1}{t}(C_r - f(\hat{x})) + f(\hat{x}) = \frac{\|x-\hat{x}\|}{r}(C_r - f(\hat{x})) + f(\hat{x}).$$

Es gilt also für alle $R > r$

$$C_R = \min\{f(x) : \ x \in \partial B_R(\hat{x})\} \geq \frac{R}{r}(C_r - f(\hat{x})) + f(\hat{x}).$$

so dass wegen $C_r > f(\hat{x})$ auch (b) und damit die Behauptung folgt. □

2.2 Das Gradientenverfahren

In diesem Abschnitt sei

$$f : \ \mathbb{R}^n \to \mathbb{R}$$

stets eine stetig differenzierbare Funktion. Zur Vereinfachung der Schreibweise nummerieren wir die Folgenglieder einer Folge von Vektoren in dieser Vorlesung üblicherweise mit tiefgestelltem Index, also

$$x_0, x_1, x_2, \ldots \in \mathbb{R}^n.$$

Wenn es in konkreten Beispielen zweckmäßig ist auf die Komponenten der Vektoren zuzugreifen, so wechseln wir fallweise zur Notation mit hochgestelltem Index, also

$$x^{(0)} = \begin{pmatrix} x_1^{(0)} \\ \vdots \\ x_n^{(0)} \end{pmatrix}, x^{(1)} = \begin{pmatrix} x_1^{(1)} \\ \vdots \\ x_n^{(1)} \end{pmatrix}, \ldots \in \mathbb{R}^n$$

2.2.1 Richtung des steilsten Abstiegs

In der Umgebung eines Punktes $x_0 \in \mathbb{R}^n$ gilt nach der linearen Taylor-Formel

$$f(x) \approx f(x_0) + \nabla f(x_0)^T (x - x_0).$$

In einer Kugel mit Radius r um x_0 nimmt die lineare Taylor-Approximation ihr Minimum an bei

$$x = x_0 - \frac{r}{\|\nabla f(x_0)\|} \nabla f(x_0)$$

(siehe Übungsaufgabe 3.1). In der linearen Approximation ist $x - x_0 = -s\nabla f(x_0)$, $s > 0$ also die *Richtung des steilsten Abstiegs* von f. Eine naheliegende Idee zur numerischen Lösung des Minimierungsproblems

$$f(x) \to \min!$$

ist es daher, beginnend mit einem Startwert x_0, eine Folge von Iterierten $x_k \in$ durch

$$x_{k+1} = x_k - s_k \nabla f(x_k)$$

mit noch zu bestimmenden Schrittweiten $s_k > 0$ zu wählen. Dies ist das sogenannte *Gradientenverfahren*, siehe Algorithmus 2.

2.2.2 Die Armijo-Schrittweitenregel

In der linearen Approximation erwarten wir dass für $s_k > 0$ und $d_k \in \mathbb{R}^n$ (im Gradientenvefahren: $d_k = -\nabla f(x_k)$) gilt

$$f(x_{k+1}) = f(x_k + s_k d_k) \approx f(x_k) + s_k \nabla f(x_k)^T d_k.$$

In der linearen Approximation würden wir die Schrittweite s_k möglichst groß wählen, damit $f(x_{k+1})$ möglichst klein wird. Auf der anderen Seite ist die

lineare Approximation aber nur in einer kleiner Umgebung von x_k eine gute Approximation an f.

Die Idee der *Armijo-Schrittweitenregel* ist es daher, die tatsächlich erzielte Abnahme der Zielfunktion

$$f(x_k) - f(x_k + s_k d_k)$$

mit der aus der linearen Approximation erwarteten Abnahme

$$-s_k \nabla f(x_k)^T d_k$$

zu vergleichen. Nur wenn die tatsächlich erzielte Abnahme einen vorgegebenen Bruchteil (z.B. $\gamma = 1\%$) der erwarteten erreicht, wird der Schritt durchgeführt. Ansonsten wird die Schrittweite als zu groß eingeschätzt (die lineare Approximation scheint für diese Schrittweite keine gute Näherung mehr zu sein) und die Schrittweite verkürzt (z.B. halbiert, $\beta = 0.5$).

Algorithmus 1 bestimmt zu gegebenen Parametern $\beta \in (0,1)$ und $\gamma \in (0,1)$ eine Schrittweite s_k, die die *Armijo-Bedingung*

$$f(x_{k+1}) \leq f(x_k) + \gamma s_k \nabla f(x_k)^T d_k \tag{2.3}$$

erfüllt.

Algorithm 1 Armijo-Schrittweitenregel

Gegeben: Parameter $\beta \in (0,1)$ und $\gamma \in (0,1)$ (z.B. $\beta = 0.5$, $\gamma = 0.01$)
Gegeben: aktuelle Iterierte x_k, Richtung $d_k \in \mathbb{R}^n$
$s_k := 1/\beta$
repeat
  $s_k := \beta s_k$
  $x_{k+1} := x_k + s_k d_k$
until $f(x_{k+1}) \leq f(x_k) + \gamma s_k \nabla f(x_k)^T d_k$
return x_{k+1}

Algorithmus 1 ist für allgemeine Suchrichtungen formuliert. Die Kombination aus Gradientenverfahren und Armijo-Schrittweitenregel fassen wir in Algorithmus 2 zusammen.

Algorithm 2 Gradientenverfahren mit Armijo-Schrittweitenregel

Gegeben: Startwert $x_0 \in \mathbb{R}^n$
Gegeben: Armijo-Parameter $\beta \in (0,1)$ und $\gamma \in (0,1)$ (z.B. $\beta = 0.5$, $\gamma = 0.01$)
for $k := 0, 1, 2, \ldots$ **do**
  if $\nabla f(x_k) = 0$ **then**
    STOP
  else
    $s_k := 1/\beta$
    repeat
      $s_k := \beta s_k$
      $x_{k+1} := x_k - s_k \nabla f(x_k)$
    until $f(x_{k+1}) \leq f(x_k) - \gamma s_k \|\nabla f(x_k)\|^2$
  end if
end for
return $x_0, x_1, x_2, \ldots$

Wir müssen noch untersuchen, ob Algorithmus 1 terminiert, also die Armijo-Schrittweitenregel tatsächlich eine Schrittweite finden wird. Dies zeigen wir nicht nur für das Gradientenverfahren (mit $d_k = -\nabla f(x_k)$) sondern gleich für allgemeine Abstiegsverfahren, siehe Abschnitt 2.3.

Definition 2.21
*Ein Vektor $0 \neq d \in \mathbb{R}^n$ heißt **Abstiegsrichtung** von f im Punkt x falls $\nabla f(x)^T d < 0$.*

Satz 2.22
Ist d eine Abstiegsrichtung für f im Punkt x, so existiert für jedes $\gamma \in (0,1)$ ein $t > 0$ mit

$$f(x + sd) \leq f(x) + \gamma s \nabla f(x)^T d \quad \forall s \in [0, t].$$

Insbesondere terminiert die Armijo-Schrittweitenregel in Algorithmus 1 für jede Abstiegsrichtung nach endlich vielen Schritten.

Beweis: Für $s \to 0$ ist

$$\frac{f(x + sd) - f(x) - \gamma s \nabla f(x)^T d}{s} \to (1 - \gamma) \nabla f(x)^T d < 0.$$

Die Armijo-Bedingung ist also für hinreichend kleine s erfüllt. □

Beispiel 2.23
Satz 2.22 zeigt, dass in Richtung einer Abstiegsrichtung eine Funktion lokal abnimmt. Die Umkehrung gilt nicht. $f(x) = -x^2$ *nimmt im Punkt* $x = 0$ *in die Richtung* $d = 1$ *ab, aber* $d = 1$ *ist keine Abstiegsrichtung für* f *in* $x = 0$.

Bemerkung 2.24
Die Armijo-Schrittweitenregel für das Gradientenverfahren können wir auch als speziellen Fall eines Trust-Region-Verfahren *interpretieren: In der aktuellen Iterierten* x_k *ersetzen wir* f *durch seine lineare Näherung*

$$f(x) \approx f(x_k) + \nabla f(x_k)^T (x - x_k)$$

und vertrauen *dieser Näherung jedoch nur in einer Kugel* $\|x - x_k\| \leq \Delta_k$ *(der* Trust-Region*). Dann lösen wir das restringierte Minimierungsproblem*

$$f(x_k) + \nabla f(x_k)^T (x - x_k) \to \text{min! } \textit{u.d.N. } x \in B_{\Delta_k}(x_k)$$

und erhalten die Lösung $x_{k+1} = x_k - s_k \nabla f(x_k)$ *(mit* $s_k = \Delta_k / \|\nabla f(x_k)\|$*).*

*Um zu entscheiden, ob wir diese Lösung akzeptieren vergleichen wir die vorhergesagte Abnahme der Zielfunktion (*predicted reduction*) mit der tatsächlichen Abnahme (*actual reduction*)*

$$pred_k = \nabla f(x_k)^T (x_{k+1} - x_k) = s_k \|\nabla f(x_k)\|^2, \qquad ared_k = f(x_k) - f(x_{k+1}).$$

Erreicht die tatsächliche Abnahme wenigstens den vorher festgelegten Bruchteil γ *der vorhergesagten Abnahme,*

$$\frac{ared_k}{pred_k} \geq \gamma,$$

so wird der Schritt 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.*

2.2.3 Konvergenz des Gradientenverfahren

Wir können nicht erwarten, dass das Gradientenverfahren in Algorithmus 2 ohne zusätzliche Annahmen gegen ein Minimum von f konvergiert. Die Iterierten können divergieren, z.B. für die lineare Zielfunktion aus Abschnitt 1.2.1. Außerdem terminiert das Verfahren, wenn ein stationärer Punkt erreicht ist, dies kann auch ein Sattellpunkt oder sogar ein Maximum sein.

Der folgende Satz zeigt aber zumindest, dass wenn die Iterierten eine konvergente Teilfolge besitzen, diese gegen einen stationären Punkt konvergiert. Wir nennen solche Resultate in dieser Vorlesung *Präkonvergenz.*

Satz 2.25 (Präkonvergenz des Gradientenverfahrens)
Algorithmus 2 terminiert entweder nach endlich vielen Schritten an einem stationären Punkt x_k oder er erzeugt eine Folge $(x_k)_{k\in\mathbb{N}}$ von Iterierten mit den Eigenschaften

(a) $f(x_{k+1}) < f(x_k)$

(b) Jeder Häufungspunkt von x_k ist ein stationärer Punkt von f.

Beweis: Terminiert Algorithmus 2 an einem x_k so gilt $\nabla f(x_k) = 0$. Terminiert der Algorithmus nicht, so ist $\nabla f(x_k) \neq 0$ für alle $k \in \mathbb{N}$. Aufgrund der Armijo-Schrittweitenregel erfüllen die Iterierten

$$f(x_{k+1}) \leq f(x_k) - \gamma s_k \nabla f(x_k)^T \nabla f(x_k) < f(x_k),$$

womit (a) gezeigt ist.

Zum Beweis von (b) sei $x \in \mathbb{R}^n$ ein Häufungspunkt von x_k, d.h. es existiert eine unendliche Indexmenge $L \subseteq \mathbb{N}$, so dass

$$\lim_{L\ni l\to\infty} x_l = x$$

Wegen der Stetigkeit von f gilt

$$\lim_{L\ni l\to\infty} f(x_l) = f(x)$$

und da $f(x_k)$ monoton fällt, folgt aus der Konvergenz dieser Teilfolge die Konvergenz der ganzen Folge. Es gilt also

$$\lim_{k\to\infty} f(x_k) = f(x)$$

und damit insbesondere $f(x_k) - f(x_{k+1}) \to 0$.

Alle Schritte erfüllen die Armijo-Bedingung

$$\gamma s_k \left\|\nabla f(x_k)\right\|^2 \leq f(x_k) - f(x_{k+1}) \to 0.$$

Es gilt also

$$\lim_{k\to\infty} s_k \left\|\nabla f(x_k)\right\|^2 = 0$$

und damit insbesondere auch

$$\lim_{L\ni l\to\infty} s_l \left\|\nabla f(x_l)\right\|^2 = 0.$$

Wegen der Stetigkeit von ∇f konvergiert $(\nabla f(x_l))_{l\in L}$ gegen $\nabla f(x)$ und es folgt dass entweder

$$\lim_{L\ni l\to\infty} s_l = 0 \quad \text{oder} \quad \nabla f(x) = 0.$$

Im zweiten Fall ist die Behauptung bewiesen. Es sei also $(s_l)_{l\in L}$ eine Nullfolge. Insbesondere muss dann für fast alle (o.B.d.A. für alle) $l \in L$ eine Verkürzung der Schrittweite bei der Armijo-Regel in Algorithmus 1 stattgefunden haben. Die Schrittweite s_l/β erfüllte also die Armijo-Bedingung noch nicht, d.h.

$$f\left(x_l - \frac{s_l}{\beta}\nabla f(x_l)\right) > f(x_l) - \gamma\frac{s_l}{\beta}\left\|\nabla f(x_l)\right\|^2 \quad \forall l \in L.$$

Nach Satz 2.3 existiert für jedes $l \in L$ ein $\sigma_l \in [0, \frac{s_l}{\beta}]$ mit

$$\begin{aligned}&-\gamma\frac{s_l}{\beta}\left\|\nabla f(x_l)\right\|^2\\ &< f\left(x_l - \frac{s_l}{\beta}\nabla f(x_l)\right) - f(x_l) = -\frac{s_l}{\beta}\nabla f\left(x_l - \sigma_l\nabla f(x_l)\right)^T\nabla f(x_l).\end{aligned}$$

Da $\sigma_l \to 0$ für $L \ni l \to \infty$ erhalten wir $\gamma\left\|\nabla f(x)\right\|^2 \geq \left\|\nabla f(x)\right\|^2$ und wegen $\gamma < 1$ folgt auch in diesem Fall $\nabla f(x) = 0$. □

Unter zusätzlichen Annahmen (insbesondere strikter Konvexität) folgt aus der Präkonvergenz in Satz 2.25 tatsächliche Konvergenz. Um aus der Konvergenz von Teilfolgen die Konvergenz der ganzen Folge zu erhalten, ist der folgende Hilfssatz nützlich.

Lemma 2.26
Sei $\hat{x} \in \mathbb{R}^n$ und $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$.

(a) Besitzt jede Teilfolge der Folge $(x_k)_{k\in\mathbb{N}}$ eine gegen $\hat{x} \in \mathbb{R}^n$ konvergente (Teil-)Teilfolge, so konvergiert die gesamte Folge $(x_k)_{k\in\mathbb{N}}$ gegen $\hat{x}$.

(b) Ist $(x_k)_{k\in\mathbb{N}}$ beschränkt und ist $\hat{x}$ ihr einziger Häufungspunkt, so konvergiert die gesamte Folge $(x_k)_{k\in\mathbb{N}}$ gegen $\hat{x}$.

Beweis: (a) Angenommen $(x_k)_{k\in\mathbb{N}}$ konvergiert nicht gegen $\hat{x}$. Dann gibt es eine Umgebung von $\hat{x}$, außerhalb derer unendliche viele Folgenglieder liegen. Diese bilden dann aber eine Teilfolge, die keine gegen $\hat{x}$ konvergente (Teil-)Teilfolge besitzt.

(b) Ist $(x_k)_{k\in\mathbb{N}}$ beschränkt, so besitzt jede Teilfolge nach dem Satz von Bolzano-Weierstraß eine konvergente (Teil-)Teilfolge und deren Grenzwert ist $\hat{x}$, da dies der einzige Häufungspunkt ist. Die Behauptung folgt damit aus (a). □

Folgerung 2.27

(a) Sei $C > 0$, so dass die Niveaumenge

$$f^{-1}(]-\infty, C]) := \{x \in \mathbb{R}^n : \ f(x) \leq C\}$$

beschränkt ist und in ihr genau ein stationärer Punkt liegt. Dann ist dieser stationäre Punkt das eindeutige globale Minimum von f und das Gradientenverfahren konvergiert für jeden Startwert x_0 mit $f(x_0) \leq C$ gegen das Minimum.

(b) Ist f konvex, dann terminiert Algorithmus 2 entweder nach endlich vielen Schritten an einem globalen Minimum oder es erzeugt eine Folge, deren Häufungspunkte globale Minima von f sind.

(c) Ist f konvex und besitzt ein eindeutiges Minimum, dann konvergiert die Folge aus Algorithmus 2 für jeden Startwert $x_0 \in \mathbb{R}^n$ gegen das globale Minimum von f.

(d) Ist f strikt konvex und besitzt ein Minimum, dann konvergiert die Folge aus Algorithmus 2 für jeden Startwert $x_0 \in \mathbb{R}^n$ gegen das globale Minimum von f.

Beweis: (a) Da f stetig ist, besitzt f auf der kompakten Niveaumenge $f^{-1}(]-\infty, C])$ (mindestens) ein Minimum. Jedes solche Minimum ist stationärer Punkt von f, so dass es nur eines geben kann. Dieses ist nach Konstruktion offenbar auch das eindeutige globale Minimum von f.

Satz 2.25(a) zeigt, dass die Iterierten des Gradientenverfahren mit Startwert x_0, $f(x_0) \leq C$, alle in der Niveaumenge $f^{-1}(]-\infty, C])$ liegen. Da die Niveaumenge beschränkt ist und nach Satz 2.25(b) das globale Minimum der einzige Häufungspunkt der Iterierten ist, folgt die Behauptung aus Lemma 2.26(b).

(b) folgt, da nach Satz 2.19(c) jeder stationäre Punkt ein globales Minimum ist.

(c) folgt aus (a), da es wegen Satz 2.19(c) genau einen stationären Punkt gibt und nach Satz 2.20 alle Niveaumengen beschränkt sind.

(d) folgt, da nach Satz 2.19(b) strikt konvexe Funktionen höchstens ein Minimum besitzen. □

2.2.4 Nachteile des Gradientenverfahren

Die Verwendung der Richtung des steilsten Abstiegs ist sehr naheliegend. Schon für einfache quadratische Zielfunktionen ist diese Richtung jedoch nicht optimal, vgl. die in der Vorlesung gemalten Skizzen. Die daraus resultierende langsame Konvergenzgeschwindigkeit des Verfahrens ist Gegenstand des 4. Übungsblattes.

2.3 Allgemeine Abstiegsverfahren

Es sei weiterhin

$$f: \mathbb{R}^n \to \mathbb{R}$$

stets eine differenzierbare Funktion. Gemäß Übungsblatt 4 ist ein Abstieg in Richtung des steilsten Abstiegs nicht optimal. Wir untersuchen deshalb jetzt, wann allgemeine Abstiegsverfahren (siehe Algorithmus 3) konvergieren.

Algorithm 3 Allgemeines Abstiegsverfahren

Gegeben: Startwert $x_0 \in \mathbb{R}^n$
for $k := 0, 1, \ldots$ **do**
  if $\nabla f(x_k) = 0$ **then**
    STOP
  else
    Wähle Suchrichtung d_k
    Wähle Schrittweite s_k
    $x_{k+1} := x_k + s_k d_k$
  end if
end for
return $x_1, x_2, x_3, \ldots$

2.3.1 Ein allgemeines Konvergenzresultat

Eine natürliche Minimalforderung[5] an die Suchrichtungen und Schrittweiten ist, dass d_k Abstiegsrichtungen sind, und dass sich der Zielfunktionalwert in jedem Schritt verringern soll.

$$\nabla f(x_k)^T d_k < 0 \quad \text{und} \quad f(x_k + s_k d_k) < f(x_k) \quad \forall k \in \mathbb{N}. \tag{2.4}$$

[5] In der Literatur wird der Begriff Abstiegsverfahren oft nur für solche Verfahren verwendet, die diese Minimalforderungen erfüllen.

Durch zwei weitere Forderungen an die Suchrichtungen und Schrittweiten können wir (Prä-)Konvergenzresultate in der Form von Satz 2.25 zeigen. Wir formulieren die Bedingungen zunächst abstrakt:

Definition 2.28
Das allgemeine Abstiegsverfahrens in Algorithmus 3 besitzt

(a) ***zulässige Schrittweiten****, falls*

$$f(x_k + s_k d_k) < f(x_k)$$

und für jede konvergente Teilfolge $(x_l)_{l\in L}$ der Iterierten gilt

$$\lim_{L\ni l\to\infty} \frac{\nabla f(x_l)^T d_l}{\|d_l\|} = 0. \tag{2.5}$$

(b) ***zulässige Suchrichtungen****, falls*

$$\nabla f(x_k)^T d_k < 0$$

und für jede konvergente Teilfolge $(x_l)_{l\in L}$ der Iterierten

$$\lim_{L\ni l\to\infty} \frac{\nabla f(x_l)^T d_l}{\|d_l\|} = 0 \quad \Longrightarrow \quad \lim_{L\ni l\to\infty} \nabla f(x_l) = 0. \tag{2.6}$$

Des Beweis des folgenden Satzes ist mit dieser Definition trivial.

Satz 2.29 (Präkonvergenz allgemeiner Abstiegsverfahren)
Das allgemeine Abstiegsverfahrens in Algorithmus 3 besitze zulässige Schrittweiten und Suchrichtungen. Dann terminiert das Verfahren entweder nach endlich vielen Schritten an einem stationären Punkt x_k oder es erzeugt eine Folge $(x_k)_{k\in\mathbb{N}}$ von Iterierten mit den Eigenschaften

(a) $f(x_{k+1}) < f(x_k)$

(b) Jeder Häufungspunkt von x_k ist ein stationärer Punkt von f.

Beweis: Dies folgt sofort aus Definition 2.28. □

Bemerkung 2.30
Zum Nachweis der Zulässigkeit von Schrittweiten ist die folgende Beobachtung nützlich, die wir bereits im Beginn vom Beweis von Satz 2.25 ausgenutzt haben. Erfüllen die Schrittweiten die Bedingung

$$f(x_k + s_k d_k) < f(x_k),$$

und existiert eine konvergente Teilfolge der Iterierten $(x_k)_{k\in\mathbb{N}}$, so bilden die Zielfunktionalwerte $(f(x_k))_{k\in\mathbb{N}}$ eine monoton fallende Folge mit konvergenter Teilfolge. Es muss also die gesamte Folge $(f(x_k))_{k\in\mathbb{N}}$ konvergieren und insbesondere gilt

$$f(x_k) - f(x_{k+1}) \to 0.$$

2.3.2 Interpretation der Zulässigkeitsbedingungen

Zur Interpretation der Zulässigkeitsbedingungen suchen wir nach intuitiven Eigenschaften, die diese abstrakten Bedingungen sicherstellen. Hierfür beginnen wir mit den Suchrichtungen.

Zulässigkeit der Suchrichtungen. Die Bedingung

$$\|\nabla f(x_k)\| \, \|d_k\| \cos \angle(-\nabla f(x_k), d_k) = -\nabla f(x_k)^T d_k > 0$$

besagt, dass der Winkel zwischen Suchrichtung und der Richtung des steilsten Abstiegs betragsmäßig unter $\pi/2$ liegt. Es ist naheliegend zu fordern, dass die Suchrichtungen sich auch nicht für $k \to \infty$ diesem Winkel annähern sollten, also

$$\exists \alpha > 0 : \ \cos \angle(-\nabla f(x_k), d_k) = \frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \, \|d_k\|} \geq \alpha. \tag{2.7}$$

Diese Forderung heißt *Winkelbedingung.*

Satz 2.31
Seien $(d_k)_{k\in\mathbb{N}}$ die Suchrichtungen eines allgemeinen Abstiegsverfahrens. Erfüllen die Suchrichtungen die Winkelbedingung (2.7), so sind sie zulässig.

Beweis: Aus (2.7) folgt $\nabla f(x_k)^T d_k < 0$ und

$$\|\nabla f(x_k)\| \leq \frac{1}{\alpha} \frac{-\nabla f(x_k)^T d_k}{\|d_k\|},$$

so dass die Zulässigkeitsbedingungen erfüllt sind. □

Bemerkung 2.32
Die Suchrichtungen sind auch bereits zulässig, wenn sie verallgemeinerte Winkelbedingungen *der Form*

$$\exists \alpha > 0, \ p > -1 : \ \frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \, \|d_k\|} \geq \alpha \, \|\nabla f(x_k)\|^p$$

oder

$$\exists \alpha_1, \alpha_2 > 0, \ p > -1 : \ \frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \, \|d_k\|} \geq \min\{\alpha_1, \alpha_2 \, \|\nabla f(x_k)\|^p\}$$

erfüllen (siehe Übungsaufgabe 5.2). Insbesondere zeigt der Fall $p > 0$, dass der Winkel zwischen Suchrichtung und der Richtung des steilsten Abstiegs betragsmäßig beliebig nah an $\pi/2$ herankommen darf, solange gleichzeitig der Gradient gegen Null konvergiert. Dies werden wir im Abschnitt über globalisierte Newton-Verfahren ausnutzen.

Beispiel 2.33

(a) Die Suchrichtungen $d_k = -\nabla f(x_k)$ erfüllen offensichtlich die Winkelbedingung und sind daher zulässig.

(b) Seien $A_k \in \mathbb{R}^{n\times n}$ symmetrische und positiv definite Matrizen mit

$$0 < c \leq \lambda_{min}(A_k) \leq \lambda_{max}(A_k) \leq C$$

Wir betrachten die (beim Newton-Verfahren in Abschnitt 2.4 verwendeten) Suchrichtungen

$$d_k = -A_k^{-1}\nabla f(x_k).$$

Es ist

$$\begin{aligned}\frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\|\,\|d_k\|} &= \frac{\nabla f(x_k)^T A_k^{-1}\nabla f(x_k)}{\|\nabla f(x_k)\|\,\left\|A_k^{-1}\nabla f(x_k)\right\|} \\ &\geq \frac{\lambda_{min}(A_k^{-1})\,\|\nabla f(x_k)\|^2}{\|\nabla f(x_k)\|^2\,\left\|A_k^{-1}\right\|} = \frac{\lambda_{min}(A_k^{-1})}{\lambda_{max}(A_k^{-1})} = \frac{\lambda_{min}(A_k)}{\lambda_{max}(A_k)} \geq \frac{c}{C}.\end{aligned}$$

Diese Suchrichtungen erfüllen also die Winkelbedingung und sind damit zulässig.

Zulässigkeit der Schrittweiten. Zur Interpretation der Forderung an die Schrittweiten, schätzen wir zunächst ab, wie weit sich f durch einen Schritt in die Richtung d verkleinern lässt. Für zweimal stetig differenzierbares f folgt aus Satz 2.3, dass ein $\sigma \in [0, s]$ existiert mit

$$f(x + sd) = f(x) + s\nabla f(x)^T d + \frac{s^2}{2} d^T \nabla^2 f(x + \sigma d) d$$

Mit $C := \max_{\sigma\in[0,s]} \|\nabla^2 f(x + \sigma d)\|$ ist also

$$f(x + sd) \leq f(x) + s\nabla f(x)^T d + \frac{s^2}{2} C\,\|d\|^2$$

und das Polynom auf der rechten Seite besitzt offenbar sein Minimum in

$$s^* = \frac{-\nabla f(x)^T d}{C\,\|d\|^2}.$$

und dort ist

$$f(x + s^* d) \leq f(x) - \frac{1}{2C}\frac{|\nabla f(x)^T d|^2}{\|d\|^2}.$$

Eine Verkleinerung von f um $\frac{1}{2C}\frac{|\nabla f(x)^T d|^2}{\|d\|^2}$ ist also möglich. Wenn in einem Abstiegsverfahren diese mögliche Verkleinerung zumindest bis auf eine Konstante erreicht wird, so heißen die verwendeten Schrittweiten *effizient*.

Definition 2.34
*Seien $(d_k)_{k\in\mathbb{N}}$ und $(s_k)_{k\in\mathbb{N}}$ die Suchrichtungen und Schrittweiten eines allgemeinen Abstiegsverfahrens. Die Schrittweiten s_k heißen **effizient**, falls ein $\theta > 0$ existiert, so dass*

$$f(x_k + s_k d_k) \leq f(x_k) - \theta \frac{|\nabla f(x_k)^T d_k|^2}{\|d_k\|^2}.$$

Satz 2.35
Seien $(d_k)_{k\in\mathbb{N}}$ und $(s_k)_{k\in\mathbb{N}}$ die Suchrichtungen und Schrittweiten eines allgemeinen Abstiegsverfahrens, das die Minimalforderung

$$f(x_k + s_k d_k) < f(x_k) \quad \text{für alle } k \in \mathbb{N}$$

erfüllt. Sind die Schrittweiten effizient, dann sind sie auch zulässig.

Beweis: Existiert eine konvergente Teilfolge der Iterierten, so folgt aus der Effizient der Schrittweiten und Bemerkung 2.30,

$$\theta \frac{|\nabla f(x_k)^T d_k|^2}{\|d_k\|^2} \leq f(x_k) - f(x_{k+1}) \to 0$$

und damit $\frac{|\nabla f(x_k)^T d_k|^2}{\|d_k\|^2} \to 0$. □

2.3.3 Zulässigkeit der Armijo-Schrittweitenregel

Die Armijo-Schrittweitenregel erzeugt nicht immer zulässige Schrittweiten wie das folgende Beispiel zeigt.

Beispiel 2.36
Wir betrachten ein Abstiegsverfahren mit Suchrichtungen

$$d_k := -2^{-k} \nabla f(x_k)$$

und Armijo-Schrittweiten. Die Suchrichtungen sind offenbar zulässig, da sie die Winkelbedingung erfüllen. Für die lineare Funktion $f(x) = -x$ akzeptiert die Armijo-Schrittweitenregel (für jede Wahl der Armijo-Parameter β und γ in Algorithmus 1) immer die Schrittweite 1, da die aus der linearen Vorhersage erwartete Abnahme der Zielfunktion perfekt mit der tatsächlichen Abnahme übereinstimmt. Wir erhalten also die Iterationsvorschrift

$$x_{k+1} = x_k + 2^{-k}$$

und es folgt $x_k \to x_0 + 2$.

Die Armijo-Schrittweiten können daher nicht zulässig sein, da sonst nach Satz 2.29 der Grenzwert der x_k ein stationärer Punkt von f sein müsste.

Die Armijo-Regel versagt in Beispiel 2.36 weil die Armijo-Schrittweite maximal 1 beträgt und zu kurze Suchrichtungen nicht durch längere Schrittweiten ausgeglichen werden können. Unter recht schwachen Bedingungen an die Länge der Suchrichtungen, kann die Zulässigkeit gezeigt werden.

Satz 2.37
Seien $(d_k)_{k\in\mathbb{N}}$ Abstiegsrichtungen. Gilt für jede konvergente Teilfolge $(x_l)_{l\in L}$ der Iterierten, dass

$$d_l \to 0 \quad \Longrightarrow \quad \frac{\nabla f(x_l)^T d_l}{\|d_l\|} \to 0, \tag{2.8}$$

dann sind die durch die Armijo-Regel erzeugten Schrittweiten zulässig.

Beweis: In Satz 2.22 haben wir gezeigt, dass die Armijo-Regel zu jeder Abstiegsrichtung d_k eine Schrittweite s_k liefert mit

$$f(x_k + s_k d_k) \le f(x_k) + \gamma s_k \nabla f(x_k)^T d_k < f(x_k)$$

liefert. Die erste Bedingung in Definition 2.28 ist also erfüllt.

Zum Beweis der zweiten Zulässigkeitsbedingung (2.5) sei $(x_l)_{l\in L}$ eine konvergente Teilfolge der Iterierten $(x_k)_{k\in\mathbb{N}}$. Angenommen die zweite Bedingung in Definition 2.28 gilt nicht. Es sei also

$$\frac{\nabla f(x_l)^T d_l}{\|d_l\|} \not\to 0 \quad \text{für } L \ni l \to \infty \tag{2.9}$$

und damit nach Voraussetzung auch

$$d_l \not\to 0 \quad \text{für } L \ni l \to \infty.$$

Dann gibt es eine Teilfolge $L' \subseteq L$ und ein $\epsilon > 0$ mit (beachte $\nabla f(x_l)^T d_l < 0$)

$$\|d_l\| \ge \epsilon \quad \text{und} \quad -\frac{\nabla f(x_l)^T d_l}{\|d_l\|} \ge \epsilon \quad \forall l \in L'. \tag{2.10}$$

Aus der Armijo-Bedingung und Bemerkung 2.30 erhalten wir, dass

$$0 \leftarrow f(x_k) - f(x_{k+1}) \ge -\gamma s_k \nabla f(x_k)^T d_k = -\gamma \frac{\nabla f(x_k)^T d_k}{\|d_k\|} s_k \|d_k\|.$$

Mit (2.10) folgt

$$\lim_{L'\ni l\to\infty} s_l \|d_l\| = 0 \quad \text{und} \quad \lim_{L'\ni l\to\infty} s_l = 0.$$

Wir gehen nun wie am Ende vom Beweis von Satz 2.25 vor. Da $(s_l)_{l\in L'}$ eine Nullfolge ist, muss für alle bis auf endlich viele (o.B.d.a. für alle) $l \in L'$ eine Verkürzung der Schrittweite bei der Armijo-Regel stattgefunden haben, d.h. die Schrittweite s_l/β erfüllte die Bedingung noch nicht:

$$f\left(x_l + \frac{s_l}{\beta} d_l\right) > f(x_l) + \gamma \frac{s_l}{\beta} \nabla f(x_l)^T d_l \quad \forall l \in L'.$$

Nach Satz 2.3 existiert für jedes $l \in L'$ ein $\sigma_l \in [0, \frac{s_l}{\beta}]$ mit

$$f(x_l) + \gamma \frac{s_l}{\beta} \nabla f(x_l)^T d_l < f\left(x_l + \frac{s_l}{\beta} d_l\right) = f(x_l) + \frac{s_l}{\beta} \nabla f\left(x_l + \sigma_l d_l\right)^T d_l.$$

also

$$\gamma \nabla f(x_l)^T d_l < \nabla f\left(x_l + \sigma_l d_l\right)^T d_l.$$

und mit $\gamma < 1$ und $\nabla f(x_l)^T d_l < 0$ folgt, dass

$$\begin{aligned} 0 \leq -(1-\gamma) \frac{\nabla f(x_l)^T d_l}{\|d_l\|} &< \frac{\left(\nabla f(x_l + \sigma_l d_l) - \nabla f(x_l)\right)^T d_l}{\|d_l\|} \\ &\leq \|\nabla f(x_l + \sigma_l d_l) - \nabla f(x_l)\|. \end{aligned}$$

Aus $s_l \|d_l\| \to 0$ folgt $\sigma_l d_l \to 0$. Da auch die Folge $(x_l)_{l\in L'}$ konvergiert, existiert eine kompakte Menge, die alle x_l und $x_l + \sigma_l d_l$ enthält. Auf dieser kompakten Menge ist ∇f gleichmäßig stetig, und es folgt $\nabla f(x_l + \sigma_l d_l) - \nabla f(x_l) \to 0$. Insgesamt ist also

$$\frac{\nabla f(x_l)^T d_l}{\|d_l\|} \to 0 \quad \text{für } L' \ni l \to \infty.$$

Dies widerspricht (2.10) und widerlegt damit unsere Annahme, dass die zweite Zulässigkeitsbedingung (2.5) nicht gilt. □

Bemerkung 2.38
Die Vorausetzung (2.8) in Satz 2.37 ist erfüllt, wenn für jede konvergente Teilfolge $(x_l)_{l\in L}$ der Iterierten eine streng monoton wachsende Funktion $\varphi : [0,\infty) \to [0,\infty)$ existiert mit

$$\|d_l\| \geq \varphi\left(\frac{-\nabla f(x_l)^T d_l}{\|d_l\|}\right) \quad \forall l \in L, \tag{2.11}$$

denn im Falle $\frac{\nabla f(x_l)^T d_l}{\|d_l\|} \not\to 0$ existiert eine Teilfolge, für die $\frac{-\nabla f(x_l)^T d_l}{\|d_l\|} \geq \epsilon$ und damit

$$\|d_l\| \geq \varphi\left(-\frac{\nabla f(x_l)^T d_l}{\|d_l\|}\right) \geq \varphi(\epsilon) > \varphi(0) \geq 0$$

gilt.

Beispiel 2.39
Für das Gradientenverfahren mit $d_k = -\nabla f(x_l)$ ist (2.8) erfüllt, da

$$\|d_k\| = \|\nabla f(x_k)\| = \frac{-\nabla f(x_k)^T d_k}{d_k}.$$

Für das Gradientenverfahren liefert die Armijo-Regel also zulässige Schrittweiten. Da die Suchrichtungen d_k ebenfalls zulässig sind folgt aus unserem allgemeinen Präkonvergenzsatz 2.29 also noch einmal die schon in Satz 2.25 gezeigte Präkonvergenz des Gradientenverfahrens.

2.3.4 Die Powell-Wolfe-Schrittweitenregel

Das Beispiel 2.36 zeigt, dass die Armijo Schrittweiten zu kurz sein können. Ein wesentlicher Teil der Beweise der Sätze 2.25 und 2.37 bestand darin zu zeigen, dass eine Verkürzung der Schrittweite bei der Armijo-Regel stattgefunden hat, die verwendete Schrittweite also in einem gewissen Sinne die längst mögliche, noch die Armijo-Bedingung erfülltende war. In den Beweisen der Sätze 2.25 und 2.37 verwendeten wir dies, um ein Zwischenresultat der Form

$$\gamma \nabla f(x_l)^T d_l < \nabla f\left(x_l + \sigma_l d_l\right)^T d_l \tag{2.12}$$

mit einer Zwischenstelle σ_l zu zeigen.

Dies motiviert die Armijo-Regel so abzuändern, dass zusätzlich zur Armijo-Bedingung (2.3) auch eine Bedingung der Form (2.12) gefordert wird und die Schrittweite nötigenfalls über 1 hinaus verlängert wird. Konkret suchen wir zu Parametern $0 < \gamma < 1$ und $\gamma < \eta < 1$ und der aktuellen Iterierten x und Suchrichtung d eine Schrittweite s, die die folgenden beiden *Powell-Wolfe-Bedingungen* erfüllt:

$$f(x + sd) - f(x) \leq \gamma s \nabla f(x)^T d, \tag{2.13}$$

$$\nabla f(x + sd)^T d \geq \eta \nabla f(x)^T d. \tag{2.14}$$

Die erste Bedingung ist die schon bekannte Armijo-Bedingung, die den tatsächlich erreichten Abstieg mit dem durch die lineare Näherung vorhergesagten vergleicht. Sie sorgt dafür, dass der Schritt *klein genug* gewählt wird, so dass die lineare Approximation noch eine gute Näherung an die wahre Funktion darstellt. Die zweite Bedingung sorgt dafür, dass der Schritt *groß genug* gewählt wird, so dass die Steigung im neuen Punkt in diese Richtung hinreichend weit abgenommen hat, vgl. die in der Vorlesung gemalten Skizzen.

Algorithm 4 Powell-Wolfe Schrittweitenregel

Gegeben: Parameter $0 < \gamma < 1$ und $\gamma < \eta < 1$ (z.B. $\gamma = 10^{-2}$ und $\eta = 0.9$)
Gegeben: aktuelle Iterierte x, Richtung $d \in \mathbb{R}^n$

$s^- := 2$
repeat
  $s^- := s^-/2$
until s^- erfüllt Armijo-Bedingung (2.13)

$s^+ := s^-$
repeat
  $s^+ := 2s^+$
until s^+ verletzt Armijo-Bedingung (2.13)

while s^- verletzt zweite Powell-Wolfe Bedingung (2.14) **do**
  $s^{(0)} := \frac{s^+ + s^-}{2}$
  if $s_{(0)}$ erfüllt Armijo-Bedingung (2.13) **then**
    $s^- := s^{(0)}$
  else
    $s^+ := s^{(0)}$
  end if
end while

return $s := s^-$

Zur konstruktiven Bestimmung einer Schrittweite, die die Powell-Wolfe-Bedingungen erfüllt, bemerken wir zunächst, dass solange die Funktion in der Suchrichtung hinreichend rasch abfällt, die Armijo-Bedingung auch erfüllt sein wird. Die Armijo-Bedingung kann nicht für beliebig große Schrittweiten erfüllt sein (zumindest nicht, wenn die Funktion in Richtung der Suchrichtung nicht nach unten unbeschränkt ist). Deshalb muss die Funktion in Richtung der Suchrichtung irgendwann abflachen und die zweite Powell-Wolfe-Bedingung erfüllt sein.

Wir bestimmen deshalb zuerst eine Schrittweite s^-, an der die Armijo-Bedingung erfüllt ist und dann eine Schrittweite s^+, an der die Armijo-Bedingung nicht mehr erfüllt ist. Durch Intervallhalbierung bestimmen wir dann eine Schrittweite, an der die Armijo-Bedingung noch gilt, aber die groß genug ist, so dass auch die zweite Powell-Wolfe-Bedingung erfüllt ist, siehe Algorithmus 4.

Satz 2.40
Seien $0 < \gamma < 1$ und $\gamma < \eta < 1$. Sei d eine Abstiegsrichtung für f im Punkt x. f sei in Richtung d nach unten beschränkt, d.h.

$$\inf_{t \geq 0} f(x + td) > -\infty.$$

Dann terminiert Algorithmus 4 nach endlich vielen Schritten und liefert eine Schrittweite s die (2.13) und (2.14) erfüllt.

Beweis: Aus Satz 2.22 folgt, dass die Repeat-Schleife in Algorithmus 4 nach endlich vielen Schritten mit einem die Armijo-Bedingung (2.13) erfüllenden s^- terminiert.

Da

$$f(x) + \gamma s^- \nabla f(x)^T d \to -\infty \quad \text{für } s^- \to \infty,$$

können wegen der Beschränktheit von f in die Suchrichtung nicht beliebig große s^+ existieren, mit

$$f(x + s^- d) \leq f(x) + \gamma s^- \nabla f(x)^T d.$$

Die zweite Repeat-Schleife in Algorithmus 4 muss also nach endlich vielen Schritten mit einem die Armijo-Bedingung (2.13) verletzenden s^+ terminieren.

Die abschließende While-Schleife des Algorithmus implementiert ein Intervallhalbierungsverfahren für die stetige Funktion

$$g(s) := f(x + sd) - f(x) - \gamma s \nabla f(x)^T d.$$

Die While-Schleife erzeugt also insbesondere eine Folge von s^- mit $g(s^-) \leq 0$ die monoton steigend gegen eine Nullstelle s^* von g konvergieren und eine Folge von s^+ mit $g(s^+) > 0$ die monoton fallend gegen die Nullstelle s^* von g konvergieren. An der Nullstelle muss daher gelten

$$0 \leq g'(s^*) = \nabla f(x + s^* d)^T d - \gamma \nabla f(x)^T d.$$

Zu jedem $\epsilon > 0$ liegt für hinreichend viele Schleifendurchgänge s^- so nah an der Nullstelle s^*, dass

$$g'(s^-) > -\epsilon$$

Mit $\epsilon := (\gamma - \eta) \nabla f(x)^T d$ folgt, dass nach endlichen vielen While-Schleifendurchgängen

$$\nabla f(x + s^- d)^T d > \eta \nabla f(x)^T d$$

gilt, also s^- die zweite Powell-Wolfe Bedingung (2.14) erfüllt und damit die While-Schleife terminiert. □

Satz 2.41
Sei f nach unten beschränkt. $(d_k)_{k\in\mathbb{N}}$ seien Abstiegsrichtungen. Dann sind die durch die Powell-Wolfe-Regel erzeugten Schrittweiten zulässig.

Beweis: Da f nach unten beschränkt ist, erzeugt Algorithmus 4 nach Satz 2.40 für jedes $k \in \mathbb{N}$ eine Schrittweite $s_k > 0$, die die beiden Powell-Wolfe Bedingungen (2.13) und (2.14) erfüllt. Insbesondere erfüllen die Schrittweiten die Armijo-Bedingung (2.13),

$$f(x_k + s_k d_k) \leq f(x_k) + \gamma s_k \nabla f(x_k)^T d_k < f(x_k).$$

Die erste Zulässigkeitsbedingung in Definition 2.28(a) ist also erfüllt.

Um die zweite Zulässigkeitsbedingung (2.5) in Definition 2.28(a) zu zeigen, sei $(x_l)_{l\in L}$ sei eine konvergente Teilfolge der Iterierten $(x_k)_{k\in\mathbb{N}}$. Nach Bemerkung 2.30 gilt

$$\lim_{k\to\infty} (f(x_k) - f(x_{k+1})) = 0.$$

Angenommen die zweite Bedingung (2.5) gilt nicht. Dann erhalten wir wie im Beweis von Satz 2.37 eine Teilfolge $L' \subseteq L$ und ein $\epsilon > 0$ mit

$$-\frac{\nabla f(x_l)^T d_l}{\|d_l\|} \geq \epsilon \quad \forall l \in L'. \tag{2.15}$$

und

$$\lim_{L'\ni l\to\infty} s_l d_l = 0.$$

Aufgrund der zweiten Powell-Wolfe-Bedingung (2.14) gilt

$$\nabla f(x_l + s_l d_l)^T d_l \geq \eta \nabla f(x_l)^T d_l$$

und damit (beachte $\eta < 1$ und $\nabla f(x_l)^T d_l < 0$)

$$\begin{aligned} 0 \leq -(1-\eta)\frac{\nabla f(x_l)^T d_l}{\|d_l\|} &\leq \frac{\left(\nabla f(x_l + s_l d_l) - \nabla f(x_l)\right)^T d_l}{\|d_l\|} \\ &\leq \|\nabla f(x_l + s_l d_l) - \nabla f(x_l)\| \end{aligned}$$

Wie im Beweis von Satz 2.37 existiert eine kompakte Menge, die alle x_l und $x_l + s_l d_l$ enthält, und auf der ∇f gleichmäßig stetig ist. Somit erhalten wir

$$\nabla f(x_l + s_l d_l) - \nabla f(x_l) \to 0$$

und damit

$$\frac{\nabla f(x_l)^T d_l}{\|d_l\|} \to 0 \quad \text{für } L' \ni l \to \infty.$$

im Widerspruch zur Annahme (2.15), womit die zweite Zulässigkeitsbedingung (2.5) bewiesen ist. □

2.4 Das Newton-Verfahren

In diesem Abschnitt setzen wir $f: \mathbb{R}^n \to \mathbb{R}$ stets als zweimal stetig differenzierbar voraus. Das Newton-Verfahren ist eine iterative Methode zur Lösung eines nichtlinearen Gleichungssystems

$$F(x) = 0$$

mit $F: \mathbb{R}^n \to \mathbb{R}^n$. Mit $F := \nabla f$ lässt sich das Verfahren auf Optimierungsprobleme anwenden.

2.4.1 Das Newton-Verfahren für Gleichungssysteme

Wir beschreiben zunächst das Newton-Verfahren für Gleichungssysteme. Dafür sei $F: \mathbb{R}^n \to \mathbb{R}^n$ eine stetig differenzierbare Funktion. Beginnend mit einer Startnäherung $x_0 \in \mathbb{R}^n$ an die Nullstelle von F verbessert das Newton-Verfahren iterativ die k-te Näherung x_k, indem die nichtlineare Funktion $F(x)$ durch ihre lineare Näherung um den Entwicklungspunkt x_k ersetzt und die Nullstelle dieser linearen Näherung bestimmt wird. Mit

$$0 \stackrel{!}{=} F(x) \approx F(x_k) + F'(x_k)(x - x_k)$$

ergibt sich also (falls $F'(x_k)$ invertierbar ist)

$$x_{k+1} := x_k + d_k, \quad \text{wobei} \quad d_k = -F'(x_k)^{-1} F(x_k).$$

Wir fassen dies in Algorithmus 5 zusammen.

Algorithm 5 Newton-Verfahren für nichtlineare Gleichungssysteme

Gegeben: Startwert $x_0 \in \mathbb{R}^n$
for $k = 0, 1, 2, \ldots$ **do**
  if $F(x_k) = 0$ **then**
    STOP
  else
    $d_k = -F'(x_k)^{-1} F(x_k)$
    $x_{k+1} := x_k + d_k$
  end if
end for
return $x_0, x_1, x_2, \ldots$

Im Eindimensionalen kann das Newton-Verfahren zeichnerisch interpretiert werden, indem in der aktuellen Iterierten x_k die Tangente im Punkt $(x_k, f(x_k)$

an den Funktionsgraphen eingezeichnet wird. Die Nullstelle der Tangente liefert dann die nächste Iterierte x_{k+1}, worauf die Tangente im nächsten Punkt $(x_{k+1}, f(x_{k+1}))$ eingezeichnet werden kann, als deren Nullstelle sich x_{k+2} ergibt, vgl. die in der Vorlesung gemalten Skizze.

Das Newton-Verfahren ist nur dann durchführbar, wenn $F'(x_k)$ in jeder Iterierten invertierbar ist. Aus der zeichnerischen Anschauung heraus ist aber offensichtlich, dass auch in diesem Fall das Newton-Verfahren nicht konvergieren muss, vgl. die in der Vorlesung gemalten Skizzen.

Wir werden zeigen, dass das Newton-Vefahren *lokal konvergiert*, d.h. für Startwerte die schon hinreichend nahe an einer Nullstelle liegen. Dafür benötigen wir die folgenden Lemmata, die zeigen, dass jede Nullstelle $\hat{x}$ mit invertierbarem $F'(\hat{x})$ eine Umgebung besitzt, in der die Jacobi-Matrix $F'(x)$ invertierbar und $\hat{x}$ die einzige Nullstelle ist.

Lemma 2.42 (Neumannsche Reihe)
Für alle $A \in \mathbb{R}^{n\times n}$ mit $\|A\| < 1$ konvergiert die Neumannsche Reihe

$$\sum_{k=0}^{\infty} A^k$$

(wobei $A^0 := I$ die Einheitsmatrix bezeichne). $I - A$ ist invertierbar und es gilt

$$\sum_{k=0}^{\infty} A^k = (I-A)^{-1} \quad \text{sowie} \quad \left\|(I-A)^{-1}\right\| \le \frac{1}{1-\|A\|}.$$

Beweis: Sei $S_l := \sum_{k=0}^{l} A^k$. Dann gilt für $m \ge l$

$$\begin{aligned}\|S_m - S_l\| = \left\|\sum_{k=l+1}^{m} A^k\right\| &\le \sum_{k=l+1}^{m} \|A\|^k = \|A\|^{l+1} \sum_{k=0}^{m-l-1} \|A\|^k \\ &\le \|A\|^{l+1} \sum_{k=0}^{\infty} \|A\|^k = \frac{\|A\|^{l+1}}{1-\|A\|} \to 0 \quad \text{für } l \to \infty.\end{aligned}$$

Die Folge der Partialsummen S_l ist also eine Cauchy-Folge, womit die Konvergenz der Neumannschen Reihe gezeigt ist. Aus

$$(I-A)\sum_{k=0}^{\infty} A^k = \sum_{k=0}^{\infty} A^k - \sum_{k=0}^{\infty} A^{k+1} = I$$

und

$$\left\|\sum_{k=0}^{\infty} A^k\right\| \le \sum_{k=0}^{\infty} \|A\|^k = \frac{1}{1-\|A\|}$$

folgen die restlichen Behauptungen. □

Lemma 2.43 (Lemma von Banach)
Sei $A \in \mathbb{R}^{n\times n}$ invertierbar und $B \in \mathbb{R}^{n\times n}$ erfülle $\|A^{-1}B\| < 1$.

Dann ist $A + B$ invertierbar und es gilt

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

Beweis: Wir wenden Lemma 2.42 auf $-A^{-1}B$ an. Damit folgt, dass $I+A^{-1}B$ und damit auch $A + B = A(I + A^{-1}B)$ invertierbar sind, und dass

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

Mit

$$\begin{aligned}(A+B)^{-1} - A^{-1} &= (I+A^{-1}B)^{-1}A^{-1} - A^{-1} = \left((I+A^{-1}B)^{-1} - I\right)A^{-1}\\ &= \left(\sum_{k=0}^{\infty}\left(-A^{-1}B\right)^k - I\right)A^{-1} = \sum_{k=1}^{\infty}\left(-A^{-1}B\right)^k A^{-1}\\ &= -A^{-1}B\sum_{k=0}^{\infty}\left(-A^{-1}B\right)^k A^{-1} = -A^{-1}B(I+A^{-1}B)^{-1}A^{-1}\end{aligned}$$

folgt die zweite Behauptung. □

Folgerung 2.44
Lemma 2.43 zeigt, dass in $\mathbb{R}^{n\times n}$ jedes invertierbare A eine Umgebung invertierbarer Matrizen besitzt

$$U_\delta := \left\{\tilde{A} \in \mathbb{R}^{n\times n} : \left\|\tilde{A} - A\right\| < \delta\right\}, \quad \delta := \frac{1}{\|A^{-1}\|},$$

und dass für jede konvergente Folge $A_k \to A$ invertierbarer Matrizen $A_k \in \mathbb{R}^{n\times n}$ mit invertierbarem Limes $A \in \mathbb{R}^{n\times n}$ gilt

$$A_k^{-1} \to A^{-1}.$$

Die Menge der invertierbaren Matrizen ist also offen in $\mathbb{R}^{n\times n}$ und die Inversion ist eine stetige Abbildung auf dieser Menge.

Lemma 2.45
Sei $F: \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\hat{x} \in \mathbb{R}^n$ sei eine Nullstelle von F mit invertierbarer Jacobi-Matrix $F'(\hat{x})$. Dann existieren $\delta, \gamma > 0$ mit

$$\|F(x)\| \geq \gamma \|x - \hat{x}\| \quad \forall x \in B_\delta(\hat{x}).$$

Insbesondere ist $\hat{x}$ also die einzige Nullstelle von F in $B_\delta(\hat{x})$.

Beweis: Es ist

$$\|x - \hat{x}\| \leq \left\|F'(\hat{x})^{-1}\right\| \|F'(\hat{x})(x - \hat{x})\|$$

also gilt

$$2\gamma \|x - \hat{x}\| \leq \|F'(\hat{x})(x - \hat{x})\| \quad \text{mit} \quad \gamma := \frac{1}{2\|F'(\hat{x})^{-1}\|}.$$

Für hinreichend nah an $\hat{x}$ liegende x gilt (z.B. nach Lemma 2.4)

$$\|F(x) - F(\hat{x}) - F'(\hat{x})(x - \hat{x})\| \leq \gamma \|x - \hat{x}\|.$$

Mit $F(\hat{x}) = 0$ erhalten wir

$$\begin{aligned} 2\gamma \|x - \hat{x}\| &\leq \|F'(\hat{x})(x - \hat{x})\| \leq \|F(x) - F'(\hat{x})(x - \hat{x})\| + \|F(x)\| \\ &\leq \gamma \|x - \hat{x}\| + \|F(x)\| \end{aligned}$$

und damit die Behauptung. □

Wir werden nicht nur zeigen, dass das Newton-Verfahren (lokal) konvergiert, sondern auch seine *Konvergenzgeschwindigkeit* abschätzen.

Definition 2.46
Eine Folge $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ konvergiert

(a) ***linear*** *mit Rate $0 < \gamma < 1$ gegen $x \in \mathbb{R}^n$, falls*

$$\|x_{k+1} - x\| \leq \gamma \|x_k - x\| \quad \textit{für fast alle } k \in \mathbb{N}$$

(b) ***superlinear*** *gegen $x \in \mathbb{R}^n$, falls $x_k \to x$ und $x_{k+1} - x = o(\|x_k - x\|)$, d.h.*

$$\frac{\|x_{k+1} - x\|}{\|x_k - x\|} \to 0 \quad \textit{für } k \to \infty.$$

(c) ***quadratisch*** *gegen $x \in \mathbb{R}^n$, falls $x_k \to x$ und $x_{k+1} - x = O(\|x_k - x\|^2)$, d.h.*

$$\exists C > 0 : \ \|x_{k+1} - x\| \leq C \|x_k - x\|^2 \quad \text{für alle } k \in \mathbb{N}.$$

Bemerkung 2.47
Konvergiert eine durch ein Iterationsverfahren erzeugte Folge linear, so erwarten wir, dass eine konstante Anzahl von Iterationen notwendig ist, um eine weitere richtige Nachkommastelle zu erhalten (z.B. 2 Nachkommastellen pro Iteration für $q = \frac{1}{100}$, 1 Nachkommastelle pro Iteration für $q = \frac{1}{10}$, oder 1 Nachkommastelle pro zwei Iterationen für $q = \frac{1}{\sqrt{10}}$). Bei quadratischer Konvergenz können wir erwarten, dass sich die Anzahl der richtigen Nachkommastellen verdoppelt.

Satz 2.48
Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\hat{x} \in \mathbb{R}^n$ sei eine Nullstelle von F mit invertierbarer Jacobi-Matrix $F'(\hat{x})$. Dann existieren $\delta, C > 0$, so dass gilt:

(a) *$\hat{x}$ ist die einzige Nullstelle von F auf $B_\delta(\hat{x})$*

(b) *$F'(x)$ ist invertierbar und $\|F'(x)^{-1}\| \leq C$ für alle $x \in B_\delta(\hat{x})$.*

(c) *Für jeden Startwert $x_0 \in B_\delta(\hat{x})$ liegen alle von Algorithmus 5 erzeugten Iterierten in $B_\delta(\hat{x})$. Insbesondere ist also die Jacobi-Matrix $F'(x_k)$ an allen Iterierten x_k invertierbar und der Algorithmus damit durchführbar.*

(d) *Für jeden Startwert $x_0 \in B_\delta(\hat{x})$ terminiert der Algorithmus entweder an der Nullstelle $\hat{x}$ oder er erzeugt eine Folge, die superlinear gegen $\hat{x}$ konvergiert.*

(e) *Ist F' zusätzlich lokal Lipschitz-stetig mit Konstante L, d.h.*

$$\|F'(x) - F'(y)\| \leq L \|x - y\| \quad \forall x, y \in B_\delta(\hat{x}),$$

dann ist die Konvergenz in (d) sogar quadratisch und es gilt

$$\|x_{k+1} - \hat{x}\| \leq \frac{CL}{2} \|x_k - \hat{x}\|^2$$

Beweis: (a),(b) Die Existenz einer Kugel $B_\delta(\hat{x})$, so dass (a) und (b) gelten, folgt direkt aus Lemma 2.45 sowie der Stetigkeit von F' und Lemma 2.43.

(c),(d) Unter Verwendung von $F(\hat{x}) = 0$ und Lemma 2.4 erhalten wir für jedes $x_k \in B_\delta(\hat{x})$, in dem $F'(x_k)$ invertierbar ist

$$\begin{aligned} x_{k+1} - \hat{x} &= x_k - F'(x_k)^{-1} F(x_k) - \hat{x} \\ &= F'(x_k)^{-1} \left(F(\hat{x}) - F(x_k) + F'(x_k)(x_k - \hat{x}) \right) \\ &= F'(x_k)^{-1} \int_0^1 \left(F'(x_k) - F'(\hat{x} + t(x_k - \hat{x})) \right) \mathrm{d}t \, (x_k - \hat{x}). \end{aligned}$$

Für jedes $x_k \in B_\delta(\hat{x})$ gilt also

$$\|x_{k+1} - \hat{x}\| \leq \|x_k - \hat{x}\| \, C \sup_{t \in [0,1]} \|F'(x_k) - F'(\hat{x} + t(x_k - \hat{x}))\| .$$

Da F' auf $B_\delta(\hat{x})$ gleichmäßig stetig ist, gilt

$$\sup_{t \in [0,1]} \|F'(x_k) - F'(\hat{x} + t(x_k - \hat{x}))\| \to 0 \quad \text{für } x_k \to \hat{x}$$

Durch Verkleinerung von $B_\delta(\hat{x})$ erreichen wir deshalb, dass

$$\sup_{t \in [0,1]} \|F'(x_k) - F'(\hat{x} + t(x_k - \hat{x}))\| \leq C/2 \quad \forall x_k \in B_\delta(\hat{x})$$

und damit

$$\|x_{k+1} - \hat{x}\| \leq \frac{1}{2} \|x_k - \hat{x}\| .$$

Durch triviale Induktion folgt damit, dass für jeden Startwert $x_0 \in B_\delta(\hat{x})$ alle Iterierten in $B_\delta(\hat{x})$ liegen und $(x_k)_{k \in \mathbb{N}}$ gegen $\hat{x}$ konvergiert.

Die superlineare Konvergenz folgt (falls der Algorithmus nicht an der einzigen Nullstelle $\hat{x}$ in $B_\delta(\hat{x})$ terminiert) aus

$$\frac{\|x_{k+1} - \hat{x}\|}{\|x_k - \hat{x}\|} \leq C \sup_{t \in [0,1]} \|F'(\hat{x} + t(x_k - \hat{x})) - F'(x_k)\| \to 0.$$

(e) Für lokal Lipschitz-stetige F' mit Lipschitz-Konstante $L > 0$ in $B_\delta(\hat{x})$ erhalten wir sogar

$$\begin{aligned} \|x_{k+1} - \hat{x}\| &\leq C \|x_k - \hat{x}\| \int_0^1 \|F'(x_k) - F'(\hat{x} + t(x_k - \hat{x}))\| \, \mathrm{d}t \\ &\leq C \|x_k - \hat{x}\| \int_0^1 L(1-t) \|x_k - \hat{x}\| \, \mathrm{d}t = \frac{CL}{2} \|x_k - \hat{x}\|^2 \end{aligned}$$

und damit quadratische Konvergenzgeschwindigkeit. □

2.4.2 Newton-Verfahren für Optimierungsprobleme

Zur Minimierung einer zweimal stetig differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$ können wir das Newton-Verfahren einsetzen, indem wir es auf die Nullstellenaufgabe

$$F(x) := \nabla f(x) = 0$$

anwenden, siehe Algorithmus 6.

Algorithm 6 Newton-Verfahren für Optimierungsprobleme

Gegeben: Startwert $x_0 \in \mathbb{R}^n$
for $k = 0, 1, 2, \ldots$ **do**
  if $\nabla f(x_k) = 0$ **then**
    STOP
  else
    $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$
    $x_{k+1} := x_k + d_k$
  end if
end for
return $x_0, x_1, x_2, \ldots$

Aus Satz 2.48 folgt die lokale Konvergenz von Algorithmus 6 gegen stationäre Punkte $\hat{x} \in \mathbb{R}^n$ mit invertierbarer Hesse-Matrix. Minima mit invertierbarer Hesse-Matrix sind genau die, für die die alle Eigenwerte der Hesse-Matrix positiv sind, die Hesse-Matrix also positiv definit ist (vgl. Satz 2.11 und Satz 2.6) Das Newton-Verfahren konvergiert also lokal in der Umgebung von Minima, in denen die hinreichenden Optimalitätsbedingungen 2. Ordnung aus Satz 2.12 gelten.

Wir erweitern Satz 2.48 noch um die Tatsache, dass die Eigenwerte in einer Umgebung von $\hat{x}$ positiv bleiben. Dies folgt leicht aus dem folgenden Lemma.

Lemma 2.49
Für alle symmetrischen Matrizen $A, B \in \mathbb{R}^{n\times n}$ gilt

$$\lambda_{min}(A + B) \geq \lambda_{min}(A) - \|B\|$$

Insbesondere besitzt also jede symmetrische und positiv definite Matrix eine Umgebung in $\mathbb{R}^{n\times n}$, in der alle symmetrischen Matrizen positiv definit ist.

Beweis: Für alle $0 \neq x \in \mathbb{R}^n$ ist

$$-\frac{x^T B x}{\|x\|^2} \leq \left|\frac{x^T B x}{\|x\|^2}\right| \leq \|B\| .$$

Aus

$$\frac{x^T Ax}{\|x\|^2} = \frac{x^T(A+B)x}{\|x\|^2} - \frac{x^T Bx}{\|x\|^2} \leq \frac{x^T(A+B)x}{\|x\|^2} + \|B\|$$

folgt also (beachte Satz 2.6)

$$\lambda_{\min}(A+B) = \min_{0\neq x\in\mathbb{R}^n} \frac{x^T(A+B)x}{\|x\|^2} \geq \min_{0\neq x\in\mathbb{R}^n} \frac{x^T Ax}{\|x\|^2} - \|B\| = \lambda_{\min}(A) - \|B\|$$

und damit die Behauptung. □

Satz 2.50
Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $\hat{x} \in \mathbb{R}^n$ sei ein lokales Minimum von f mit positiv definiter Hesse-Matrix $\nabla^2 f(\hat{x})$. Dann existieren $\delta, \mu > 0$, so dass gilt:

(a) $\hat{x}$ ist der einzige stationäre Punkt von f auf $B_\delta(\hat{x})$

(b) Für alle $x \in B_\delta(\hat{x})$ ist

$$\lambda_{min}(\nabla^2 f(x)) \geq \mu,$$

also ist insbesondere $\nabla^2 f(x)$ invertierbar und $\|\nabla^2 f(x)^{-1}\| \leq \frac{1}{\mu}$.

(c) Für jeden Startwert $x_0 \in B_\delta(\hat{x})$ liegen alle von Algorithmus 5 erzeugten Iterierten in $B_\delta(\hat{x})$. Insbesondere ist also $\nabla^2 f(x_k)$ an allen Iterierten x_k invertierbar und der Algorithmus damit durchführbar.

(d) Für jeden Startwert $x_0 \in B_\delta(\hat{x})$ terminiert der Algorithmus terminiert entweder an der Nullstelle $\hat{x}$ oder er erzeugt eine Folge, die superlinear gegen $\hat{x}$ konvergiert.

(e) Ist $\nabla^2 f$ zusätzlich lokal Lipschitz-stetig mit Konstante L in $B_\delta(\hat{x})$, dann ist die Konvergenz in (d) sogar quadratisch und es gilt

$$\|x_{k+1} - \hat{x}\| \leq \frac{L}{2\mu} \|x_k - \hat{x}\|^2 .$$

Beweis: Für jedes ?? und der Stetigkeit von $\nabla^2 f$, dass in einer Umgebung von $\hat{x}$

$$\lambda_{\min}(\nabla^2 f(x)) \geq \mu,$$

gilt. Der Rest der Behauptung folgt aus Satz 2.48. □

Beispiel 2.51
Betrachte

$$f: \mathbb{R} \to \mathbb{R}, \quad f(x) = \sqrt{x^2+1}.$$

Offenbar ist $x = 0$ das einzige Minimum von f und

$$\nabla f(x) = f'(x) = \frac{x}{\sqrt{x^2+1}},$$
$$\nabla^2 f(x) = f''(x) = \frac{1}{\sqrt{x^2+1}} - \frac{x^2}{(x^2+1)^{3/2}} = \frac{1}{(x^2+1)^{3/2}},$$

so dass f die Voraussetzungen von Satz 2.50 erfüllt.

Die Newton-Iteration lautet

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k) = x_k - x_k(x_k^2+1) = -x_k^3.$$

Für jeden Startwert $x_0 \in (-1, 1)$ konvergiert das Newton-Verfahren also (sogar kubisch) gegen das Minimum $\hat{x} = 0$.

Für $|x_0| \geq 1$ divergieren die Iterierten jedoch. Für $|x_0| > 1$ gehen sie betragsmäßig gegen unendlich und für $|x_0| = 1$ alternieren sie zwischen -1 und 1.

Bemerkung 2.52
Das Newton-Verfahren für Optimierungsprobleme können wir auch ohne den Umweg über die Lösung der stationäre Punkte-Gleichung motivieren. Zur Minimierung von f ersetzen wir dazu im aktuellen Iterationspunkt x_k die Funktion f durch ihre quadratische Näherung entsprechend Satz 2.3

$$f(x) \approx f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x)(x - x_k)$$

und wählen die nächste Iterierte als Minimum dieser quadratischen Näherung. Nach Übungsaufgabe 4.1 lautet das Minimum

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k).$$

2.4.3 Newton-artige Verfahren

Die Implementierung des Newton-Verfahrens ist deutlich aufwendiger als die des Gradientenverfahrens. In jedem Iterationsschritt des Newton-Verfahrens wird (zusätzlich zu einer Gradientenauswertung $\nabla f(x_k)$) eine Auswertung der Hesse-Matrix $\nabla^2 f(x_k)$ sowie die Lösung eines linearen Gleichungssystems

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k) \tag{2.16}$$

benötigt. Asymptotisch gesehen lohnt sich der zusätzlich Aufwand. Für jedes $N \in \mathbb{N}$ ist ein superlinear konvergentes Verfahren mit dem N-fachen Aufwand eines linear konvergenten letzlich schneller als das linear konvergente. In der praktischen Anwendung, wo nur endlich viele Iterationsschritte durchgeführt werden, ist es jedoch äußerst erstrebenswert diesen Zusatzaufwand zu umgehen oder zumindest zu verringern.

Mit den Lösungen d_k von (2.16) werden beim Newton-Verfahren ja lediglich Näherungen $x_{k+1} = x_k + d_k$ an die wahre Lösung $\hat{x}$ berechnet. Es liegt also nahe, dass es gar nicht nötig ist, die exakte Lösung von (2.16) zu verwenden. Dies motiviert:

- **Inexakte Newton-Verfahren**: Das Gleichungssystem (2.16) wird nur näherungsweise, z.B. durch Anwendung einiger Schritte eines iterativen Verfahrens wie dem CG-Verfahren gelöst.

- **Newton-artige Verfahren**: Die Hesse-Matrix $\nabla^2 f(x_k)$ wird durch eine invertierbare Matrix $H_k \in \mathbb{R}^{n\times n}$ (die einfacher invertierbar oder einfacher berechenbar ist) ersetzt und das Gleichungssystem

  $$H_k d_k = -\nabla f(x_k)$$

  gelöst.

Wir untersuchen nun, wie genau (2.16) gelöst werden muss, damit die Iterierten noch superlinear konvergieren.

Lemma 2.53
Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $K \subset \mathbb{R}^n$ kompakt und konvex. Dann ist F auf K Lipschitz-stetig,

$$\|F(y) - F(x)\| \leq L \|y - x\| \quad \forall x, y \in K,$$

mit Lipschitz-Konstante $L = \max_{x\in K} \|F'(x)\|$.

Beweis: Gemäß dem mehrdimensionalen Mittelwertsatzes aus Lemma 2.4 gilt

$$F(y) - F(x) = \int_0^1 F'(x + t(y - x))\, \mathrm{d}t(y - x),$$

und es folgt die Behauptung (beachte $x + t(y - x) \in K$, da K konvex). □

Satz 2.54 (Dennis-Moré-Bedingungen)
Sei $f: \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $\hat{x} \in \mathbb{R}^n$ ein Punkt mit invertierbarer Hesse-Matrix $\nabla^2 f(\hat{x}) \in \mathbb{R}^{n\times n}$. Sei $(x_k)_{k\in\mathbb{N}}$ eine Folge mit

$$x_k \neq \hat{x} \quad \forall k \in \mathbb{N} \qquad \textit{und} \qquad x_k \to \hat{x}.$$

Dann sind äquivalent:

(a) $(x_k)_{k\in\mathbb{N}}$ konvergiert superlinear gegen $\hat{x}$ und $\hat{x}$ ist ein stationärer Punkt von f.

(b) Es gilt $\nabla f(x_{k+1}) = o(x_{k+1} - x_k)$.

(c) Es gilt $\nabla f(x_k) + \nabla^2 f(\hat{x})(x_{k+1} - x_k) = o(x_{k+1} - x_k)$.

(d) Es gilt $\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = o(x_{k+1} - x_k)$.

Beweis: „(a) $\Longrightarrow$ (b)": Da $(x_k)_{k\in\mathbb{N}}$ superlinear konvergiert, gilt

$$\|x_{k+1} - \hat{x}\| \leq \frac{1}{2}\|x_k - \hat{x}\| \qquad \text{für hinreichend große } k \in \mathbb{N}.$$

Mit

$$\|x_k - \hat{x}\| \leq \|x_{k+1} - x_k\| + \|x_{k+1} - \hat{x}\| \leq \|x_{k+1} - x_k\| + \frac{1}{2}\|x_k - \hat{x}\|$$

folgt deshalb (für hinreichend große $k \in \mathbb{N}$)

$$\frac{1}{2}\|x_k - \hat{x}\| \leq \|x_{k+1} - x_k\|.$$

Da $(x_k)_{k\in\mathbb{N}}$ beschränkt ist, existiert eine abgeschlossene Kugel K, die alle x_k (und damit auch den Grenzwert $\hat{x}$) enthält. Nach Lemma 2.53 existiert ein $L > 0$ so dass

$$\begin{aligned}\frac{\|\nabla f(x_{k+1})\|}{\|x_{k+1} - x_k\|} = \frac{\|\nabla f(x_{k+1}) - \nabla f(\hat{x})\|}{\|x_{k+1} - x_k\|} &\leq L\frac{\|x_{k+1} - \hat{x}\|}{\|x_{k+1} - x_k\|} \\ &\leq 2L\frac{\|x_{k+1} - \hat{x}\|}{\|x_k - \hat{x}\|} \to 0.\end{aligned}$$

„(b) $\Longrightarrow$ (a)": Es sei

$$\nabla f(x_{k+1}) = o(x_{k+1} - x_k), \qquad \text{also} \qquad \epsilon_k := \frac{\|\nabla f(x_{k+1})\|}{\|x_{k+1} - x_k\|} \to 0.$$

Insbesondere gilt $\nabla f(\hat{x}) = \lim_{k\to\infty} \nabla f(x_{k+1}) = 0$. $\hat{x}$ ist also ein stationärer Punkt.

Da außerdem $\nabla^2 f(\hat{x})$ invertierbar ist, existiert nach Lemma 2.45 ein $\gamma > 0$, so dass für hinreichend große $k \in \mathbb{N}$ gilt

$$\|\nabla f(x_{k+1})\| \geq \gamma \|x_{k+1} - \hat{x}\|.$$

und damit

$$\begin{aligned}\|x_{k+1} - \hat{x}\| &\leq \frac{1}{\gamma} \|\nabla f(x_{k+1})\| = \frac{\epsilon_k}{\gamma} \|x_{k+1} - x_k\| \\ &\leq \frac{\epsilon_k}{\gamma} \|x_{k+1} - \hat{x}\| + \frac{\epsilon_k}{\gamma} \|x_k - \hat{x}\|.\end{aligned}$$

Für hinreichend große k ist $\frac{\epsilon_k}{\gamma} < \frac{1}{2}$ und wir erhalten

$$\|x_{k+1} - \hat{x}\| \leq 2\frac{\epsilon_k}{\gamma} \|x_k - \hat{x}\| = o(x_k - \hat{x}).$$

„(b) $\Longleftrightarrow$ (c)": Der mehrdimensionale Mittelwertsatzes aus Lemma 2.4 für $F(x) := \nabla f(x)$ liefert

$$\begin{aligned}\nabla f(x_{k+1}) &= \nabla f(x_k) + \int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k))(x_{k+1} - x_k)\,\mathrm{d}t \\ &= \left(\int_0^1 \left(\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(\hat{x})\right)\,\mathrm{d}t\right)(x_{k+1} - x_k) \\ &\quad + \nabla f(x_k) + \nabla^2 f(\hat{x})(x_{k+1} - x_k)\end{aligned}$$

Wegen $x_k \to \hat{x}$ gilt, dass

$$\begin{aligned}\delta_k &:= \sup_{t\in[0,1]} \|x_k + t(x_{k+1} - x_k) - \hat{x}\| \\ &\leq \sup_{t\in[0,1]} \left((1-t)\|x_k - \hat{x}\| + t\|x_{k+1} - \hat{x}\|\right) \to 0.\end{aligned}$$

Mit der Stetigkeit von $\nabla^2 f$ folgt, dass (für $k \to \infty$)

$$\begin{aligned}&\sup_{t\in[0,1]} \left\|\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(\hat{x})\right\| \\ &\leq \sup_{\xi\in B_{\delta_k}(\hat{x})} \left\|\nabla^2 f(\xi) - \nabla^2 f(\hat{x})\right\| \to 0.\end{aligned}$$

und damit

$$\int_0^1 \left(\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(\hat{x})\right)\,\mathrm{d}t \to 0.$$

Es gilt also

$$\nabla f(x_{k+1}) = \nabla f(x_k) + \nabla^2 f(\hat{x})(x_{k+1} - x_k) + o(x_{k+1} - x_k).$$

„(c) $\iff$ (d)": Wegen $\nabla^2 f(x_k) \to \nabla^2 f(\hat{x})$ gilt

$$\begin{aligned}\big\|\nabla f(x_k) + \nabla^2 f(\hat{x})(x_{k+1} - x_k) - \big(\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k)\big)\big\| \\ \leq \big\|\nabla^2 f(\hat{x}) - \nabla^2 f(x_k)\big\| \, \|x_{k+1} - x_k\| = o(x_{k+1} - x_k),\end{aligned}$$

und damit die Äquivalenz von (c) und (d). □

Satz 2.54 zeigt uns, wie genau wir bei inexakten Newton-Verfahren das Gleichungssystem lösen müssen bzw. wie gut der in einem Newton-artigen Verfahren verwendete Ersatz für die Hesse-Matrix mit dieser übereinstimmen muss.

Satz 2.55
Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Die Folge $(x_k)_{k\in\mathbb{N}}$ sei erzeugt durch

$$x_{k+1} = x_k + d_k$$

und konvergiere gegen ein $\hat{x} \in \mathbb{R}^n$ mit positiv definiter Hesse-Matrix $\nabla^2 f(\hat{x})$. Außerdem gelte $x_k \neq \hat{x}$ für alle $k \in \mathbb{N}$.

(a) Newton-Verfahren: Sind die d_k exakte Lösungen von

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k), \tag{2.17}$$

dann ist $\hat{x}$ ein Minimum von f und $(x_k)_{k\in\mathbb{N}}$ konvergiert superlinear gegen $\hat{x}$. Ist $\nabla^2 f$ lokal Lipschitz stetig, so ist die Konvergenz sogar quadratisch.

(b) Inexakte Newton-Verfahren: Sind die d_k approximative Lösungen von (2.17) die

$$\big\|\nabla f(x_k) + \nabla^2 f(x_k) d_k\big\| \leq \eta_k \, \|\nabla f(x_k)\|$$

mit einer Nullfolge positiver Zahlen $(\eta_k)_{k\in\mathbb{N}}$ erfüllen, so ist $\hat{x}$ ein Minimum von f und $(x_k)_{k\in\mathbb{N}}$ konvergiert superlinear gegen $\hat{x}$.

(c) Newton-artige Verfahren: Ist d_k eine Lösung von (2.17), wobei die Hesse-Matrix durch eine invertierbare Matrix $H_k \in \mathbb{R}^{n\times n}$ ersetzt wurde, also

$$H_k d_k = -\nabla f(x_k),$$

und gilt $H_k \to \nabla^2 f(\hat{x})$, so ist $\hat{x}$ ein Minimum von f und $(x_k)_{k\in\mathbb{N}}$ konvergiert superlinear gegen $\hat{x}$.

Beweis: Das $\hat{x}$ ein Minimum ist, und $(x_k)_{k\in\mathbb{N}}$ superlinear konvergiert, folgt in allen drei Fällen die Behauptung durch Überprüfen der Dennis-Moré-Bedingungen in Satz 2.54.

In (a) ist dies trivial (und die quadratische Konvergenz folgt aus Satz 2.50).

In (c) folgt die Dennis-Moré-Bedingung (c) aus Satz 2.54 sofort aus

$$\begin{aligned} -\nabla f(x_k) - \nabla^2 f(\hat{x})(x_{k+1} - x_k) &= \left(H_k - \nabla^2 f(\hat{x})\right)(x_{k+1} - x_k) \\ &= o(x_{k+1} - x_k) \end{aligned}$$

Im Fall (b) erhalten wir, dass

$$\begin{aligned} \|\nabla f(x_k)\| &\leq \left\|\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k)\right\| + \left\|\nabla^2 f(x_k)(x_{k+1} - x_k)\right\| \\ &\leq \eta_k \|\nabla f(x_k)\| + \left\|\nabla^2 f(x_k)(x_{k+1} - x_k)\right\|. \end{aligned}$$

Es gilt also (für hinreichend große $k \in \mathbb{N}$)

$$\|\nabla f(x_k)\| \leq \frac{1}{1-\eta_k} \left\|\nabla^2 f(x_k)(x_{k+1} - x_k)\right\|$$

und damit

$$\begin{aligned} &\left\|\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k)\right\| \\ &\leq \eta_k \|\nabla f(x_k)\| \leq \frac{\eta_k}{1-\eta_k} \left\|\nabla^2 f(x_k)(x_{k+1} - x_k)\right\| \\ &\leq \frac{\eta_k}{1-\eta_k} \left\|\nabla^2 f(x_k)\right\| \|x_{k+1} - x_k\| = o(x_{k+1} - x_k), \end{aligned}$$

womit die Dennis-Moré-Bedingung (d) aus Satz 2.54 gezeigt ist. □

Bemerkung
Satz 2.55(b) zeigt, dass ein konvergentes inexaktes Newton-Verfahren sogar superlinear konvergiert, wenn in jedem Newton-Schritt das lineare Gleichungssystem

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k)$$

(z.B. mit einem iterativen Lösungsverfahren) approximativ bis auf einen residualen Fehler von $\eta_k \|\nabla f(x_k)\|$ gelöst wird. Satz 2.55 lässt jedoch offen, unter welchen Bedingungen ein solches Verfahren konvergiert. Tatsächlich lässt sich ohne weitere Voraussetzungen analog zu Satz 2.50 die lokale Konvergenz zeigen, siehe Übungsaufgabe 7.3.

2.4.4 Quasi Newton Verfahren

Wir führen nun ein besonders effizientes Newton-artiges Verfahren ein, das weder die Hesse-Matrix noch Lösungen linearer Gleichungssystem benötigt.

Zur Motivation betrachten wir das Newton-Verfahren für eindimensionale Nullstellenprobleme $F: \mathbb{R} \to \mathbb{R}$, $F(\hat{x}) = 0$,

$$x_{k+1} := x_k - F'(x_k)^{-1} F(x_k).$$

Beim *Sekantenverfahren* wird (ausgehend von zwei Startwerten $x_0, x_1 \in \mathbb{R}$) die Ableitung, also die Tangentensteigung an der Stelle x_k durch die Steigung der Sekante zwischen der aktuellen und der letzten Iterierten angenähert

$$F'(x_k) \approx \frac{F(x_k) - F(x_{k-1})}{x_k - x_{k-1}} =: H_k$$

und es ergibt sich die Iterationsvorschrift

$$x_{k+1} = x_k - \left(\frac{F(x_k) - F(x_{k-1})}{x_k - x_{k-1}} \right)^{-1} F(x_k),$$

vgl. die in der Vorlesung gemalten Skizzen.

Genauso könnnen wir für eindimensionale Optimierungsprobleme

$$\min_{x \in \mathbb{R}} f(x)$$

die zweite Ableitung approximieren durch

$$f''(x_k) \approx \frac{f'(x_k) - f'(x_{k-1})}{x_k - x_{k-1}} =: H_k$$

und erhalten so dass (wiederum zwei Startwerten $x_0, x_1 \in \mathbb{R}$ benötigende) Newton-artige Verfahren

$$x_{k+1} = x_k - H_k^{-1} f'(x_k).$$

Im Eindimensionalen ist die Approximation H_k eindeutig bestimmt durch die Gleichung

$$H_k(x_k - x_{k-1}) = f'(x_k) - f'(x_{k-1}).$$

Das mehrdimensionale Analogon für $f: \mathbb{R}^n \to \mathbb{R}$ lautet

$$H_k(x_k - x_{k-1}) = \nabla f(x_k) - \nabla f(x_{k-1}). \tag{2.18}$$

und bestimmt H_k für $n \geq 2$ nicht mehr eindeutig. Eine Newton-artiges Verfahren

$$x_{k+1} := x_k - H_k^{-1} \nabla f(x_k), \tag{2.19}$$

bei dem alle verwendeten Matrizen H_k, $k \in \mathbb{N}$, die Bedingung (2.18) erfüllen heißt **Quasi-Newton-Verfahren**.

Bemerkung 2.56

(a) Die exakte Hesse-Matrix $H_k := \nabla^2 f(x_k)$ erfüllt die Quasi-Newton-Gleichung (2.18) im Allgemeinen nicht.

(b) Wegen

$$\nabla f(x_k) - \nabla f(x_{k-1}) = \left(\int_0^1 \nabla^2 f(x_{k-1} + t(x_k - x_{k-1}) \, \mathrm{d}t \right) (x_k - x_{k-1})$$

erfüllt die Wahl $H_k := \int_0^1 \nabla^2 f(x_{k-1} + t(x_k - x_{k-1}) \, \mathrm{d}t$ die Quasi-Newton-Gleichung. Das ist aber noch aufwendiger zu berechnen, als die exakte Hesse Matrix $\nabla^2 f(x_k)$, so dass dies keine sinnvolle Wahl wäre.

(c) Die Iterationsvorschrift (2.19) zusammen mit der Quasi-Newton-Bedingung (2.18) ergibt

$$\begin{aligned} \nabla f(x_{k+1}) &= \nabla f(x_k) + (\nabla f(x_{k+1}) - \nabla f(x_k)) \\ &= -H_k(x_{k+1} - x_k) + H_{k+1}(x_{k+1} - x_k) \\ &= (H_{k+1} - H_k)(x_{k+1} - x_k). \end{aligned}$$

Aufgrund der Dennis-Moré-Bedingung (b) in Satz 2.54 konvergiert also ein Quasi-Newton-Verfahrens bereits dann superlinear, wenn

$$H_{k+1} - H_k \to 0.$$

Die Quasi-Newton-Bedingung (2.18) legt H_k noch nicht eindeutig fest. Bemerkung 2.56 motiviert, dass sich H_{k+1} möglichst wenig von H_k unterscheiden sollte. Da im $k+1$-ten Schritt ein lineares Gleichungssystem mit der Matrix H_{k+1} gelöst werden muss, wäre es außerdem wünschenswert, dass sich die Inverse von H_{k+1} möglichst einfach aus der Inverse von H_k berechnen lässt. Diese Eigenschaft besitzen die sogenannten *Niedrig-Rang-Modifikationen.*

Bezeichne $e_k \in \mathbb{R}^n$ den k-ten Einheitsvektor. Die Matrix $e_j e_k^T \in \mathbb{R}^{n \times n}$ besitzt an (j,k)-ter Position den Eintrag 1 und besteht ansonsten aus Nullen. Entsprechend wird für eine Matrix $A \in \mathbb{R}^{n \times n}$ durch Addition zu $A + e_j e_k^T$ nur (j,k)-te Eintrag geändert. Entsprechend ergibt sich für zwei beliebige Vektoren $u, v \in \mathbb{R}^n$ die Matrix $A + uv^T$ durch Abänderung eines einzelnen

Eintrags in der Matrix A, wenn diese auf eine Basis transformiert wurde die u und v enthält. Solche Änderungen heißen Rang-1-Modifikationen.

Das folgende Lemma zeigt, dass die Inversen von Rang-1-Modifikationen wiederum Rang-1-Modifikationen sind und explizit angegeben werden können. Durch triviale Induktion überträgt sich das offenbar auf Modifikationen höheren Ranges.

Lemma 2.57 (Sherman-Morrison-Woodbury-Formel)
Sei $A \in \mathbb{R}^{n\times n}$ invertierbar. Seien $u, v \in \mathbb{R}^n$.

Die Matrix $A+uv^T \in \mathbb{R}^{n\times n}$ ist genau dann invertierbar, wenn $1+v^TA^{-1}u \neq 0$ und in diesem Fall gilt

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

Beweis: Ist $1 + v^TA^{-1}u \neq 0$ dann folgt durch simples Ausmultiplizieren

$$(A + uv^T)\left(A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}\right) = I,$$

also ist $A + uv^T$ invertierbar und die Inverse besitzt die angebene Form.
Ist $1 + v^TA^{-1}u = 0$, dann ist $A + uv^T$ nicht injektiv, da

$$(A + uv^T)(A^{-1}uv^Tv) = uv^Tv + u\,(v^TA^{-1}u)\,v^Tv = 0,$$

aber $(A^{-1}uv^Tv) = \|v\|^2\,A^{-1}u \neq 0$. □

Der einfachste Ansatz zur Bestimmung der Matrizen H_k ist deshalb, beginnend mit einer symmetrischen Startmatrix H_0, die weiteren Matrizen durch *Aufdatierung* als symmetrische Rang-1-Modifikationen

$$H_{k+1} = H_k + \gamma_k u_k u_k^T$$

zu wählen, wobei $\gamma_k \in \mathbb{R}$ und $u_k \in \mathbb{R}^n$ so zu bestimmen sind, dass H_{k+1} die Quasi-Newton-Bedingung (2.18) für $k + 1$ erfüllt, d.h.

$$H_{k+1}d_k = y_k,$$

wobei wir hier und im Folgenden die Abkürzungen

$$d_k := x_{k+1} - x_k \quad \text{und} \quad y_k := \nabla f(x_{k+1}) - \nabla f(x_k)$$

verwenden.

Auf Übungsblatt 8 zeigen wir, dass sich unter der Bedingung

$$(y_k - H_k d_k)^T d_k \neq 0$$

eindeutig die folgende Aufdatierunsformel (die sogenannte **Symmetrische-Rang-1-Formel** (SR1)) ergibt

$$H_{k+1} = H_k + \frac{(y_k - H_k d_k)(y_k - H_k d_k)^T}{(y_k - H_k d_k)^T d_k}.$$

Für die Implementierung eines Quasi-Newton-Verfahrens werden nur die Inversen $B_k := H_k^{-1}$ benötigt. Mit Lemma 2.57 lassen sich leicht Aufdatierungsformeln zur Berechnung der Inverse im $(k+1)$-ten Schritt aus der Inversen im k-ten Schritt herleiten, so dass sich das SR1-Verfahren äußerst effizient (ohne Auswertung der Hesse-Matrix und ohne Lösungen linearer Gleichungsysteme) implementieren lässt. Die SR1-Aufdatierungsformeln besitzen jedoch Nachteil, dass die Bedingung

$$(y_k - H_k d_k)^T d_k \neq 0$$

nicht sichergestellt ist.

In der Praxis haben sich deshalb Rang-2-Korrekturen durchgesetzt. Die wohl populärste ist die **Broyden-Fletcher-Goldfarb-Shanno-Formel**

$$H_{k+1} := H_k + \frac{y_k y_k^T}{y_k^T d_k} - \frac{H_k d_k d_k^T H_k^T}{d_k^T H_k d_k}.$$

Offenbar ist H_k symmetrisch und man kann zeigen (siehe Übungsblatt 8), dass (für $y_k^T d_k \neq 0$ und $d_k^T H_k d_k \neq 0$) die BFGS-Matrizen H_k die Quasi-Newton-Bedingung (2.18) erfüllen.

Die Aufdatierungsformeln für die Inversen $B_k := H_k^{-1}$ ergeben sich aus Lemma 2.57 zu

$$B_{k+1} := B_k + \frac{(d_k - B_k y_k) d_k^T + d_k (d_k - B_k y_k)^T}{d_k^T y_k} - \frac{(d_k - B_k y_k)^T y_k}{(d_k^T y_k)^2} d_k d_k^T$$

(vgl. Übungsblatt 8).

Das dazugehörige Verfahren ist in Algorithmus 7 zusammengefasst.

Algorithm 7 Quasi-Newton-Verfahren mit BFGS-Aufdatierung

Gegeben: Startwert $x_0 \in \mathbb{R}^n$
Gegeben: Symmetrische, positiv definite Startmatrix $B_0 \in \mathbb{R}^{n \times n}$
for $k = 0, 1, 2, \ldots$ **do**
  if $\nabla f(x_k) = 0$ **then**
    STOP
  else
    $d_k := -B_k \nabla f(x_k)$
    $x_{k+1} := x_k + d_k$
  end if
  $B_{k+1} := B_k + \frac{(d_k - B_k y_k) d_k^T + d_k (d_k - B_k y_k)^T}{d_k^T y_k} - \frac{(d_k - B_k y_k)^T y_k}{(d_k^T y_k)^2} d_k d_k^T$
end for
return $x_0, x_1, x_2, \ldots$

Satz 2.58
Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $\nabla^2 f$ lokal Lipschitz stetig. Sei $\hat{x}$ ein lokales Minimum von f mit positiv definiter Hesse-Matrix $\nabla^2 f(\hat{x})$.

Dann existieren $\delta, \epsilon > 0$, so dass für

(a) jeden Startwert mit $\|x_0 - \hat{x}\| < \delta$

(b) jede symmetrische positiv definite Startmatrix mit $\|B_0 - \nabla^2 f(\hat{x})^{-1}\| < \epsilon$

der Algorithmus 7 durchführbar ist und entweder nach endlich vielen Schritten $x_k = \hat{x}$ liefert oder eine Folge $(x_k)_{k \in \mathbb{N}}$, die superlinear gegen $\hat{x}$ konvergiert.

Beweis: Für den Beweis verweisen wir auf [GeigerKanzow1, Satz 11.33]. □

2.4.5 Das globalisierte Newton-Verfahren

Das Newton-Verfahren und seine vorgestellten Varianten konvergieren zwar sehr schnell (superlinear) aber wie Beispiel 2.51 zeigt nur lokal. Das Gradientenverfahren mit Armijo-Regel konvergiert bereits unter recht milden Zusatzvoraussetzungen global (siehe z.B. Folgerung 2.27 für die globale Konvergenz im Beispiel 2.51) aber im Allgemeinen nur linear (vgl. Übungsblatt 4).

Es wäre wünschenswert die globale Konvergenz des Gradientenverfahrens mit der schnellen Konvergenz des Newton-Verfahrens zu verbinden. Dazu müssten die ersten Iterationsschritte mit dem Gradientenverfahren durchgeführt werden, und dann wenn die Näherung schon hinreichend gut ist,

auf das Newton-Verfahren umgeschaltet werden. Anders ausgedrückt soll der Newton-Schritt erst dann ausgeführt werden, wenn auch er zu Konvergenz führt.

Wir erreichen dieses Ziel durch die folgende Strategie: Wir interpretieren das Newton-Verfahren als allgemeines Abstiegsverfahren gemäß Abschnitt 2.3 mit Suchrichtung und Schrittweite

$$d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k) \quad \text{und} \quad s_k = 1.$$

Wenn der Newton-Schritt die Voraussetzungen unseres allgemeinen Konvergenzresultats in Satz 2.29 erfüllt, so führen wir den Newton-Schritt aus. Ansonsten führen wir einen Schritt des Gradientenverfahrens durch.

Wir überprüfen dazu, ob die Newton-Suchrichtungen eine verallgemeinerte Winkelbedingung (siehe Bemerkung 2.32) erfüllen,

$$\frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \, \|d_k\|} \geq \alpha \, \|\nabla f(x_k)\|^p$$

mit $\alpha > 0$ und $p > 0$.

Das Gradientenverfahren liefert uns eine Folge mit $\nabla f(x_k) \to 0$, so dass wir erwarten können, dass diese Bedingung irgendwann erfüllt sein wird. Dann können wir die Suchrichtungen des Gradientenverfahrens durch die des Newton-Verfahrens ersetzen ohne dass die Konvergenz beeinträchtigt wird (zumindest wenn wir die Schrittweiten weiterhin mit der Armijo-Regel bestimmen und die Voraussetzungen von Satz 2.37 gelten). Wenn die Armijo-Regel dann auch noch hinreichend nahe am Minimum die Anfangsschrittweite 1 akzeptiert, dann hätten wir tatsächlich ein Verfahren mit den globalen Konvergenzeigenschaften des Gradientenverfahrens konstruiert, das nach an endlich vielen Schritten in das schnell konvergente Newton-Verfahren übergeht. Wir werden im Beweis von Satz 2.61 sehen, dass dies der Fall ist, wenn wir den Armijo-Parameter $\gamma < 1/2$ wählen.

Wir fassen das so motivierte Verfahren in Algorithmus 8 zusammen.

Nach Konstruktion lässt sich leicht zeigen, dass das globalisierte Newton-Verfahren die globale (Prä-)Konvergenz des Gradientenverfahrens besitzt:

Satz 2.59 (Präkonvergenz des globalisierten Newton-Verfahrens)
Sei $f : \mathbb{R}^n \to \mathbb{R}^n$ zweimal stetig differenzierbar. Algorithmus 8 terminiert entweder nach endlich vielen Schritten an einem stationären Punkt x_k oder er erzeugt eine Folge $(x_k)_{k \in \mathbb{N}}$ von Iterierten mit den Eigenschaften

(a) $f(x_{k+1}) < f(x_k)$

Algorithm 8 Globalisiertes Newton-Verfahren

Gegeben: Startwert $x_0 \in \mathbb{R}^n$
Gegeben: Armijo-Parameter $\beta \in (0,1)$ und $\gamma \in (0, 1/2)$ (z.B. $\beta = 0.5$, $\gamma = 0.01$)
Gegeben: Winkelbedingung-Parameter: $\alpha > 0$, $p > 0$ (z.B. $\alpha = 10^{-6}$, $p = 1$)
for $k := 0, 1, 2, \ldots$ **do**
  if $\nabla f(x_k) = 0$ **then**
    STOP
  else
    if $\nabla^2 f(x_k)$ invertierbar **then**
      $d_k := -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$
      if $-\nabla f(x_k)^T d_k < \alpha \|\nabla f(x_k)\|^{1+p} \|d_k\|$ **then**
        $d_k := -\nabla f(x_k)$
      end if
    else
      $d_k := -\nabla f(x_k)$
    end if
    $s_k := 1/\beta$
    repeat
      $s_k := \beta s_k$
      $x_{k+1} := x_k + s_k d_k$
    until $f(x_{k+1}) \leq f(x_k) - \gamma s_k \nabla f(x_k)^T d_k$
  end if
end for
return $x_0, x_1, x_2, \ldots$

(b) Jeder Häufungspunkt von x_k ist ein stationärer Punkt von f.

Beweis: Wir zeigen, dass die Voraussetzungen unseres allgemeinen Präkonvergenzsatzes 2.29 erfüllt sind. Die Newton-Suchrichung

$$d_k := -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

wird gewählt falls $-\nabla^2 f(x_k)$ invertierbar ist und

$$-\nabla f(x_k)^T d_k \geq \alpha \|\nabla f(x_k)\|^{1+p} \|d_k\|$$

ansonsten nimmt das Verfahren die Gradientensuchrichtung $d_k := -\nabla f(x_k)$. Es gilt also

$$-\frac{\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \, \|d_k\|} \geq \min\{1, \alpha \|\nabla f(x_k)\|^p\} \quad \forall k \in \mathbb{N}.$$

Die Suchrichtungen erfüllen also die verallgemeinerte Winkelbedingung aus unserer Bemerkung 2.32 und sind somit zulässig.

Für die Schrittweiten wird immer die Armijo-Regel verwendet, so dass wir ihre Zulässigkeit mit Satz 2.37 zeigen können. Sei x_l eine konvergente Folge. Da f zweimal stetig differenzierbar ist, existiert eine Konstante $C > 0$ mit $\|\nabla^2 f(x_l)\| \leq C$ für alle x_l. Die Newton-Suchrichtungen erfüllen deshalb

$$0 \leq \frac{-\nabla f(x_l)^T d_l}{\|d_l\|} \leq \|\nabla f(x_l)\| = \left\|\nabla^2 f(x_l) d_l\right\| \leq C \|d_l\|$$

und die Gradientensuchrichtungen erfüllen $\|\nabla f(x_l)\| = \|d_l\| = \frac{-\nabla f(x_l)^T d_l}{\|d_l\|}$. Es gilt also stets

$$0 \leq \frac{-\nabla f(x_l)^T d_l}{\|d_l\|} \leq \max\{1, C\} \|d_l\| .$$

Aus Satz 2.37 folgt die Zulässigkeit der Schrittweiten und damit insgesamt aus Satz 2.29 die Behauptung. □

Wir zeigen nun noch, dass das globalisierte Newton-Verfahren nach endlich vielen Schritten in das Newton-Verfahren übergeht, und damit superlinear bzw. sogar quadratisch konvergiert.

Hierzu verwenden wir die folgende nützliche Ergänzung von Lemma 2.26.

Lemma 2.60
Ist $\hat{x} \in \mathbb{R}^n$ ein isolierter Häufungspunkt einer Folge $(x_k)_{k\in\mathbb{N}}$ und gilt für jede gegen $\hat{x}$ konvergente Teilfolge $(x_l)_{l\in L}$

$$x_l - x_{l+1} \to 0,$$

so konvergiert die gesamte Folge $(x_k)_{k\in\mathbb{N}}$ gegen $\hat{x}$.

Beweis: Angenommen $(x_k)_{k\in\mathbb{N}}$ konvergiert nicht gegen $\hat{x}$, dann existiert eine Umgebung $B_\epsilon(\hat{x})$ von $\hat{x}$ außerhalb derer unendlich viele Folgenglieder liegen. Sei $\epsilon > 0$ so klein gewählt, dass $\hat{x}$ der einzige Häufungspunkt von $(x_k)_{k\in\mathbb{N}}$ in $B_\epsilon(\hat{x})$ ist. Nach Konstruktion liegen jeweils unendlich viele Folgenglieder innerhalb und außerhalb von $B_\epsilon(\hat{x})$. Wir können daher eine Teilfolge konstruieren

$$(x_l)_{l\in L} \quad \text{mit} \quad x_l \in B_\epsilon(\hat{x}), \quad x_{l+1} \notin B_\epsilon(\hat{x}).$$

Da $(x_l)_{l\in L}$ beschränkt und $\hat{x}$ ihr einziger Häufungspunkt ist, folgt aus (b), dass $\lim_{L\ni l\to\infty} x_l = \hat{x}$. Die nachfolgenden Folgenglieder x_{l+1}, $l \in L$ besitzen nach Konstruktion aber den Mindestabstand ϵ von $\hat{x}$, so dass $x_l - x_{l+1} \to 0$ nicht gelten kann. □

Satz 2.61 (Schnelle Konvergenz des globalisierten Newton-Verfahrens) *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Falls Algorithmus 8 nicht terminiert, die von ihm erzeugte Folge $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ einen Häufungspunkt $\hat{x}$ besitzt, und in diesem die Hesse-Matrix positiv definit ist, so gilt:*

(a) $\hat{x}$ ist ein isoliertes lokales Minimum von f.

(b) Die (gesamte) Folge $(x_k)_{k\in\mathbb{N}}$ konvergiert gegen $\hat{x}$.

(c) Nach endlich vielen Iterationsschritten wird nur noch die Newton-Suchrichtung mit Schrittweite 1 verwendet, also das Newton-Verfahren durchgeführt. Insbesondere konvergiert die Folge also superlinear und falls $\nabla^2 f$ lokal Lipschitz-stetig ist sogar quadratisch gegen $\hat{x}$

Beweis: Nach Satz 2.59 ist $\hat{x}$ ein stationärer Punkt von f, und da in $\hat{x}$ die hinreichenden Optimalitätsbedingungen 2. Ordnung erfüllt sind, ist $\hat{x}$ ein lokales Minimum.

Aus Satz 2.50(a),(b) folgt die Existenz von $\delta, \mu > 0$, so dass $\hat{x} \in \mathbb{R}^n$ der einzige stationäre Punkt in $B_\delta(\hat{x})$ (und damit das einzige Minimum) ist und

$$\lambda_{\min}(\nabla^2 f(x)) \geq \mu \quad \text{für alle } x \in B_\delta(\hat{x}). \tag{2.20}$$

Damit ist insbesondere (a) gezeigt.

Um (b) zu zeigen, wenden wir Lemma 2.60 an. Aus Satz 2.59 folgt, dass $\hat{x}$ der einzige Häufungspunkt von $(x_k)_{k\in\mathbb{N}}$ in $B_\delta(\hat{x})$ ist, da jeder anderer Häufungspunkt ebenfalls ein Minimum wäre. Sei $(x_l)_{l\in L}$ eine gegen $\hat{x}$ konvergent Teilfolge. Wegen der Stetigkeit von ∇f gilt $\nabla f(x_l) \to 0$. Für hinreichend große $l \in L$ ist $x_l \in B_\delta(\hat{x})$ und es gilt

$$\begin{aligned}\|x_l - x_{l+1}\| = s_l \|d_l\| &\leq \max\{\|\nabla f(x_l)\|, \left\|(\nabla^2 f(x_l))^{-1}\nabla f(x_l)\right\|\} \\ &\leq \max\left\{1, \frac{1}{\mu}\right\} \|\nabla f(x_l)\| \to 0.\end{aligned}$$

Aus Lemma 2.60 folgt also dass die (gesamte) Folge $(x_k)_{k\in\mathbb{N}}$ gegen $\hat{x}$ konvergiert.

Es bleibt zu zeigen, dass nach endlich vielen Iterationsschritten nur noch die Newton-Suchrichtung mit Schrittweite 1 verwendet wird. Dafür müssen wir zeigen, dass für hinreichend große $k \in \mathbb{N}$

(i) $\nabla^2 f(x_k)$ invertierbar ist,

(ii) die Newton-Suchrichtungen $d_k := -(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$ verwendet werden, da sie die verallgemeinerte Winkelbedingung erfüllen, d.h.

$$-\nabla f(x_k)^T d_k \geq \alpha \|\nabla f(x_k)\|^{1+p} \|d_k\|$$

(iii) die Armijo-Schrittweitenregel die Schrittweite $s_k = 1$ akzeptiert, d.h.

$$f(x_k + d_k) \leq f(x_k) - \gamma \nabla f(x_k)^T d_k.$$

Da $x_k \to \hat{x}$, gilt für hinreichend große $k \in \mathbb{N}$, dass $x_k \in B_\delta(\hat{x})$. (i) folgt daher aus (2.20).

Um (ii) zu zeigen beachten wir, dass wegen der Stetigkeit von $\nabla^2 f$ ein $C > 0$ existiert mit

$$\left\|\nabla^2 f(x)\right\| \leq C \quad \text{für alle } x \in B_\delta(\hat{x}).$$

Da $\nabla f(x_k) \to 0$, ist für hinreichend große $k \in \mathbb{N}$

$$\alpha \|\nabla f(x_k)\|^p \leq \frac{\mu}{C}$$

und die Newton-Suchrichtungen $d_k := -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$ erfüllen dann

$$\begin{aligned} -\nabla f(x_k)^T d_k = d_k^T \nabla^2 f(x_k) d_k &\geq \mu \|d_k\|^2 = \mu \left\|\nabla^2 f(x_k)^{-1}\nabla f(x_k)\right\| \|d_k\| \\ &\geq \frac{\mu}{C} \|\nabla f(x_k)\| \|d_k\| \geq \alpha \|\nabla f(x_k)\|^{1+p} \|d_k\|. \end{aligned}$$

Zum Beweis von (iii) erhalten wir aus der Taylor-Formel in Satz 2.3 ein $s \in [0,1]$mit

$$f(x_k + d_k) = f(x_k) + \nabla f(x_k)^T d_k + \frac{1}{2} d_k^T \nabla^2 f(x_k + s d_k) d_k. \tag{2.21}$$

Für hinreichend große $k \in \mathbb{N}$ ist

$$\lambda_{\min}(\nabla^2 f(x_k)) \geq \mu,$$

und es wird wegen (i) und (ii) die Newton-Richtung als Suchrichtung verwendet, d.h.

$$\nabla f(x_k)^T d_k = -d_k^T \nabla^2 f(x_k) d_k.$$

Damit erhalten wir aus (2.21), dass

$$\begin{aligned} &f(x_k + d_k) - f(x_k) - \gamma \nabla f(x_k)^T d_k \\ &= (1-\gamma)\nabla f(x_k)^T d_k + \frac{1}{2} d_k^T \nabla^2 f(x_k + s d_k) d_k \\ &= -\left(\frac{1}{2} - \gamma\right) d_k^T \nabla^2 f(x_k) d_k + \frac{1}{2} d_k^T \left(\nabla^2 f(x_k + s d_k) - \nabla^2 f(x_k)\right) d_k \\ &\leq -\left(\frac{1}{2} - \gamma\right) \mu \|d_k\|^2 + \frac{1}{2} \left\|\nabla^2 f(x_k + s d_k) - \nabla^2 f(x_k)\right\| \|d_k\|^2. \end{aligned}$$

Da x_k konvergiert und $\|d_k\| \leq \frac{1}{\mu} \|\nabla f_k\| \to 0$ folgt mit dem gleichen Kompaktheitsargument wie im Beweis von Satz 2.37, dass

$$\left\|\nabla^2 f(x_k + s d_k) - \nabla^2 f(x_k)\right\| \to 0.$$

Für hinreichend große $k \in \mathbb{N}$ gilt also (beachte $\gamma < 1/2$)

$$f(x_k + d_k) - f(x_k) - \gamma \nabla f(x_k)^T d_k \leq 0,$$

womit (iii) und damit der letzte Teil der Behauptung bewiesen ist. □

2.5 Nichtlineare Ausgleichsprobleme

Die vorgestellten Ideen können wir auch verwenden, um nichtlineare inverse Probleme zu lösen. Betrachte die nichtlineare Nullstellenaufgabe

$$F(x) = 0, \tag{2.22}$$

wobei $F : \mathbb{R}^n \to \mathbb{R}^m$ eine hinreichend oft stetig differenzierbare Funktion sei. Für $n = m$ können wir das in Abschnitt 2.4.1 vorgestellte Newton-Verfahren für nichtlineare Gleichungssysteme verwenden.

In der Praxis sind die vorkommenden Gleichungssysteme häufig überbestimmt und aufgrund von Messfehlern ist keine exakte Lösbarkeit zu erwarten. Statt einer exakten Lösung von (2.22) begnügt man sich deshalb mit der Suche nach einem Minimum von

$$\Phi(x) := \frac{1}{2} \|F(x)\|^2 = \frac{1}{2} \sum_{i=1}^{m} F_i(x)^2.$$

Zur Minimierung von Φ können wir die in den letzten Abschnitten behandelten Gradienten- und Newton-Verfahren anwenden. Die dazu notwendigen Ableitungen von Φ ergeben sich aus dem folgenden Lemma.

Lemma 2.62
Ist F einmal bzw. zweimal stetig differenzierbar, so auch Φ und es gilt

$$\nabla \Phi(x) = F'(x)^T F(x).$$

$$\nabla^2 \Phi(x) = F'(x)^T F'(x) + \sum_{i=1}^{m} F_i(x) \nabla^2 F_i(x).$$

Beweis: Übungsaufgabe □

Wir stellen ohne Konvergenzbeweise noch zwei populäre Verfahren vor, die direkt auf die Lösung des nichtlinearen Ausgleichsproblems zugeschnitten sind.

2.5.1 Das Gauß-Newton-Verfahren

Die unmittelbare Anwendung des Newton-Verfahrens aus Abschnitt 2.4.2 führt zu der Iterationsvorschrift

$$\begin{aligned} x_{k+1} &:= x_k - \nabla\Phi^2(x_k)^{-1}\nabla\Phi(x_k) \\ &= x_k - \left(F'(x_k)^T F'(x_k) + \sum_{i=1}^{m} F_i(x_k)\nabla^2 F_i(x_k) \right)^{-1} F'(x_k)^T F(x_k), \end{aligned}$$

wodurch gemäß Bemerkung 2.52 x_{k+1} das Minimum der quadratischen Approximation

$$\Phi(x) \approx \Phi(x_k) + \nabla\Phi(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)\nabla^2\Phi(x)(x - x_k)$$

bestimmt wird.

Eine einfachere Approximation ohne $\nabla^2\Phi$ erhalten wir, indem wir die spezielle Gestalt von Φ ausnutzen und

$$\Phi(x) = \frac{1}{2}\|F(x)\|^2 \approx \frac{1}{2}\|F(x_k) + F'(x_k)(x - x_k)\|^2$$

minimieren. Die Lösung dieses linearen Ausgleichsproblems ist (falls die Matrix $F'(x_k)^T F'(x_k)$ invertierbar ist) gegeben durch

$$x_{k+1} := x_k - \left(F'(x_k)^T F'(x_k)\right)^{-1} F'(x_k)^T F(x_k),$$

vgl. die Gaußschen Normalengleichungen in Übungsaufgabe 1.2.

Das auf dieser Iteration beruhende Verfahren heißt *Gauß-Newton-Verfahren* und ist in Algorithmus 9 zusammengefasst.

Algorithm 9 Gauß-Newton-Verfahren

Gegeben: Startwert $x_0 \in \mathbb{R}^n$
for $k = 0, 1, 2, \ldots$ **do**
  if $F'(x_k)^T F(x_k) = 0$ **then**
    STOP
  else
    $d_k = -\left(F'(x_k)^T F'(x_k)\right)^{-1} F'(x_k)^T F(x_k)$
    $x_{k+1} := x_k + d_k$
  end if
end for
return $x_0, x_1, x_2, \ldots$

Man kann (unter geeigneten Voraussetzungen) zeigen, dass das Gauß-Newton Verfahren lokal (jedoch für nicht exakt lösbare Ausgleichsprobleme im Allgemeinen nur mit linearer Geschwindigkeit) gegen ein Minimum von Φ konvergiert.

2.5.2 Levenberg-Marquardt-Verfahren

Die beim Gauß-Newton-Verfahren verwendete lineare Approximation von F in

$$\Phi(x) = \frac{1}{2}\|F(x)\|^2 \approx \frac{1}{2}\|F(x_k) + F'(x_k)(x - x_k)\|^2$$

ist nur für kleine $x - x_k$ gerechtfertigt.

Die Idee des *Levenberg-Marquardt-Verfahrens* besteht darin, x_{k+1} so zu wählen, dass (die Approximation an) $\Phi(x_{k+1})$ möglichst klein wird und gleichzeitig $d_k := x_{k+1} - x_k$ nicht zu groß wird. Man sucht dafür das Minimum d_k von

$$\|F(x_k) + F'(x_k)d_k\|^2 + \mu\|d_k\|^2 \to \min! \tag{2.23}$$

wobei der Parameter $\mu > 0$ steuert, wie wichtig die Minimierung des linearisierten Residuums $F(x_k) + F'(x_k)d_k$ gegenüber der Beschränkung der Größe von d_k ist.

Das Minimierungsproblem (2.23) können wir als lineares Ausgleichsproblem schreiben

$$\left\| \begin{pmatrix} F'(x_k) \\ \sqrt{\mu} I \end{pmatrix} d_k + \begin{pmatrix} F(x_k) \\ 0 \end{pmatrix} \right\| \to \min!$$

und durch Anwendung der Gaußschen Normalengleichungen (Übungsaufgabe 1.2)

$$\begin{pmatrix} F'(x_k)^T & \sqrt{\mu} I \end{pmatrix} \begin{pmatrix} F'(x_k) \\ \sqrt{\mu} I \end{pmatrix} d_k = -\begin{pmatrix} F'(x_k)^T & \sqrt{\mu} I \end{pmatrix} \begin{pmatrix} F(x_k) \\ 0 \end{pmatrix},$$

also

$$d_k = -\left(F'(x_k)^T F'(x_k) + \mu I\right)^{-1} F'(x_k)^T F(x_k).$$

Im Gegensatz zum Gauß-Newton-Verfahren ist dieser Iterationsschritt immer wohldefiniert:

Lemma 2.63
Sei $x_k \in \mathbb{R}^n$.

(a) Für jedes $\mu > 0$ ist $F'(x_k)^T F'(x_k) + \mu I$ positiv definit und damit invertierbar.

(b) Die Abbildung

$$r: \ (0,\infty) \to \mathbb{R}, \quad \mu \mapsto \|d_k\|,$$

mit

$$d_k := -\left(F'(x_k)^T F'(x_k) + \mu I\right)^{-1} F'(x_k)^T F(x_k)$$

ist stetig und monoton fallend.

Für $\mu \to \infty$ konvergiert d_k gegen den Nullvektor. Falls $F'(x_k)^T F'(x_k)$ invertierbar ist, so konvergiert d_k für $\mu \to 0$ gegen den Gauß-Newton-Schritt

$$d_k = -\left(F'(x_k)^T F'(x_k)\right)^{-1} F'(x_k)^T F(x_k).$$

(c) $d_k := -\left(F'(x_k)^T F'(x_k) + \mu I\right)^{-1} F'(x_k)^T F(x_k)$ ist ein globales Minimum von

$$\|F(x_k) + F'(x_k)d\| \to \min! \quad \textit{u.d.N.} \ \|d\| \le r(\mu).$$

Beweis: (a) folgt sofort aus der Symmetrie und der positiven Semidefinitheit der Matrix $F'(x_k)^T F'(x_k)$.

Zum Beweis von (b) seien $(v_l)_{l=1}^n \subset \mathbb{R}^n$ eine ONB aus Eigenvektoren von $F'(x_k)^T F'(x_k)$ mit zugehörigen Eigenwerten $\lambda_l \ge 0$. Sei

$$F'(x_k)^T F(x_k) = \sum_{l=1}^n a_l v_l, \quad a_l := v_l^T F'(x_k)^T F(x_k) \in \mathbb{R}$$

die Entwicklung von $F'(x_k)^T F(x_k) \in \mathbb{R}^n$ in dieser Basis, dann ist

$$d_k = -\sum_{l=1}^n \frac{a_l}{\lambda_l + \mu} v_l, \quad \|d_k\|^2 = \sum_{l=1}^n \left(\frac{a_l}{\lambda_l + \mu}\right)^2$$

und es folgt die Behauptung (b).

(c) folgt sofort daraus, dass d_k ein globales Minimum von (2.23) ist. □

Lemma 2.63 zeigt, dass wir die Wahl von μ als Wahl einer Trust-Region interpretieren können. Wir trauen der Näherung

$$\Phi(x) = \frac{1}{2}\|F(x)\|^2 \approx \frac{1}{2}\|F(x_k) + F'(x_k)(x - x_k)\|^2$$

nur in der Umgebung $\|x - x_k\| \le r(\mu)$ und erhalten d_k als Lösung des restringierten Minimierungsproblems (bzgl. $x - x_k$)

$$\|F(x_k) + F'(x_k)(x - x_k)\| \to \min! \quad \text{u.d.N.} \ \|x - x_k\| \le r(\mu). \tag{2.24}$$

Um zu entscheiden, ob die Wahl von μ (also die Wahl der Trust-Region) geeignet war, vergleichen wir die vorhergesagte Abnahme der Zielfunktion (*predicted reduction*) mit der tatsächlichen Abnahme (*actual reduction*) und berechnen

$$\epsilon := \frac{\text{ared}_k}{\text{pred}_k} = \frac{\|F(x_k)\|^2 - \|F(x_{k+1})\|^2}{\|F(x_k)\|^2 - \|F(x_k) + F'(x_k)(x_{k+1} - x_k)\|^2}.$$

Ist ϵ sehr klein, so war unsere Trust-Region zu groß, wir verwerfen den Schritt und wiederholen ihn mit kleinerer Trust Region (also größerem μ). Ansonsten akzeptieren wir den Schritt und vergrößern vielleicht sogar die Trust-Region, wenn ϵ hinreichend groß ist. Eine einfache Implementierung ist in Algorithmus 10 zusammengefasst.

Algorithm 10 Levenberg-Marquardt-Verfahren

Gegeben: Startwert $x_0 \in \mathbb{R}^n$
Gegeben: Parameter $0 < \beta_0 < \beta_1 < 1$ und $\mu > 0$ (z.B. $\beta_0 := 0.3$, $\beta_1 := 0.9$, $\mu := 1$)
for $k = 0, 1, 2, \ldots$ **do**
  if $F'(x_k)^T F(x_k) = 0$ **then**
    STOP
  end if
  repeat
    $d_k := -\left(F'(x_k)^T F'(x_k) + \mu I\right)^{-1} F'(x_k)^T F(x_k)$.
    $\epsilon := \frac{\|F(x_k)\|^2 - \|F(x_{k+1})\|^2}{\|F(x_k)\|^2 - \|F(x_k) + F'(x_k)(x - x_k)\|^2}$
    if $\epsilon \leq \beta_0$ **then**
      $\mu := 2\mu$
    end if
  until $\epsilon > \beta_0$
  $x_{k+1} := x_k + d_k$
  if $\epsilon > \beta_1$ **then**
    $\mu := \frac{1}{2}\mu$
  end if
end for
return $x_0, x_1, x_2, \ldots$

Moderne Implementierungen kontrollieren explizit den Trust-Region-Radius durch Lösung des restringierten Minimierungsproblems (2.24), vgl. Übungs-

aufgabe 9.3. Für eine solche Variante, bei der noch zusätzlich im Prediktor

$$\begin{aligned}\text{pred}_k &= \|F(x_k)\|^2 - \|F(x_k) + F'(x_k)(x_{k+1} - x_k)\|^2 \\ &= -2F(x_k)^T F'(x_k)(x_{k+1} - x_k) - \|F'(x_k)(x_{k+1} - x_k)\|^2\end{aligned}$$

nur der lineare Term $-2F(x_k)^T F'(x_k)(x_{k+1} - x_k)$ verwendet wird, wird die globale Konvergenz (gegen einen stationären Punkt von Φ) z.B. in [Hanke, Satz 21.3] gezeigt.

Kapitel 3

Restringierte Optimierung

Anmerkung zur Notation: Zur Behandlung von Nebenbedingungen müssen wir häufig auf die Einträge von Vektoren im $\mathbb{R}^n$ zugreifen müssen. Abweichend zu der bisherigen Schreibweise verwenden wir in diesem Kapitel deshalb hochgestellte Indizes (ohne Klammern) zur Nummerierung von Vektoren und tiefgestellte Indizes für ihre Einträge.

3.1 Lineare Optimierung

3.1.1 Motivation und Normalform

Wir wenden uns nun der restringierten Optimierung zu und beginnen mit dem Spezialfall der *linearen Optimierung* (auf englisch auch: *linear programming*), bei der die Zielfunktion linear ist und die Nebenbedingungen durch lineare Gleichungen oder Ungleichungen gegeben sind.

Solche lineare Optimierungsaufgaben werden bereits in der Mittelstufe behandelt und dort graphisch gelöst. Ein typische Beispielaufgabe ist, dass ein Landwirt zur Fütterung seiner Ferkel das Futter zusammenrühren muss. Dazu stehen ihm (das Beispiel und alle Zahlen sind fiktiv) Sojabohnen und Kartoffeln zur Verfügung.

- Eine Einheit Sojabohnen koste 1EUR, enthalte 2 Einheiten Proteine und einen Fettgehalt von 4.
- Eine Einheit Kartoffeln koste 2EUR, enthalte auch 2 Einheit Proteine, aber nur einen Fettgehalt von 2.

Das Futter für die Schweine solle mindestens 10 Einheiten Proteine beinhalten und der Gesamtfettgehalt soll unter 12 liegen. Wie lässt sich diese Mischung am preisgünstigsten erreichen?

Mit den Bezeichnungen x für die Anzahl Einheiten Sojabohnen und y für die Anzahl der Einheiten Kartoffeln ergibt sich das Problem, eine *lineare Zielfunktion* zu minimieren

$$x + 2y \to \min!$$

unter den *linearen* Nebenbedingungen.

$$\begin{aligned} 2x + 2y &\geq 10 \\ 4x + 2y &\leq 12 \\ x &\geq 0 \\ y &\geq 0 \end{aligned}$$

Für die graphische Lösung vergleiche die in der Volesung gemalte Skizze.

Offenbar lässt sich jede Maximierungsaufgabe mit linearer Zielfunktion durch Multiplikation mit -1 in eine Minimierungsaufgabe mit linearer Zielfunktion umformen. Genauso kann jede in der Nebenbedingung auftretende lineare Ungleichung in die Form

$$\sum a_{ij} x_j \leq b_i$$

gebracht werden. Alle Ungleichungen können durch Einführung einer sogenannten Schlupfvariable ξ in die äquivalente Form

$$\sum a_{ij} x_j + \xi = b_i \quad \xi \geq 0$$

gebracht werden. Schließlich kann noch jede Variable ersetzt werden durch

$$x_i = x_i' - x_i'', \quad x_i', x_i'' \geq 0.$$

Wir können also jede Optimierungsaufgabe mit linearer Zielfunktion unter linearen Gleichungs- und Ungleichungsrestriktionen stets in eine äquivalente Minimierungsaufgabe mit linearen Gleichungsrestriktionen und Nichtnegativitätsbedingungen an alle Lösungskomponenten überführen.

Definition 3.1
Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$. Die Aufgabe, einen Minimierer $x \in \mathbb{R}^n$ der Funktion

$$f : \; x \mapsto c^T x$$

unter den Nebenbedingungen $Ax = b$ *und* $x \geq 0$ *zu finden, heißt* lineare Optimierungsaufgabe (auch: lineares Programm) in Normalform. $x \geq 0$ *ist dabei komponentenweise zu verstehen.*

Wie zuvor nennen wir die zu minimierende Funktion $f(x) = c^T x$ *das* Zielfunktional. *Die Menge*

$$\{x \in \mathbb{R}^n \ : \ Ax = b, \ x \geq 0\}$$

heißt zulässiger Bereich.

Bemerkung 3.2
Offenbar kann der zulässige Bereich einelementig sein (dann ist dieses Element die Lösung) oder leer sein (dann existiert keine Lösung). Auch wenn der zulässige Bereich nicht leer ist, existiert möglicherweise keine Lösung, z.B. ist für $n = 1 = m$, $A = 0$, $b = 0$, $c = -1$ *die Zielfunktion* $x \mapsto c^T x = -x$ *auf dem zulässigen Bereich* $[0, \infty)$ *offenbar nach unten unbeschränkt.*

Der zulässige Bereich ist offenbar abgeschlossen und die Zielfunktion stetig. Falls der zulässige Bereich beschränkt und nicht-leer ist, so muss (mindestens) eine Lösung existieren, da stetige Funktionen auf kompakten Mengen ihr Minimum annehmen.

3.1.2 Geometrische Interpretation

Definition 3.3
(a) Zu gegebenen Punkten $y^1, \ldots, y^m \in \mathbb{R}^n$ *heißt*

$$\sum_{i=1}^{m} \lambda_i y^i, \qquad \text{mit } \lambda_i \in [0,1], \ \sum_{i=1}^{n} \lambda_i = 1$$

Konvexkombination. *Die Menge aller Konvexkombinationen*

$$S(y^1, \ldots, y^m) := \{\sum_{i=1}^{m} \lambda_i y^i \ : \ \lambda_i \in [0,1], \ \sum_{i=1}^{n} \lambda_i = 1\} \subset \mathbb{R}^n$$

heißt (der von $y^1, \ldots, y^m$ *aufgespannte)* Simplex.
Für $m = 2$ *ist*

$$[y^1, y^2] := S(y^1, y^2) = \{(1-\lambda)y^1 + \lambda y^2 \ : \ \lambda \in [0,1]\}$$

die aus Definition 2.1 bekannte Verbindungsstrecke *zwischen* y^1 *und* y^2.

(b) Für eine nicht-leere Teilmenge $K \subset \mathbb{R}^n$ nennen wir $x \in K$ eine Ecke von K*, falls keine* durch x *verlaufende Strecke in K existiert, d.h. für alle $y^1, y^2 \in \mathbb{R}^n$ mit $x \in [y^1, y^2] \subseteq K$ (also für jede durch x verlaufende Strecke in K) gilt dass $x = y^1$ oder $x = y^2$.*

Von nun an betrachten wir immer eine durch $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ gegebene lineare Optimierungsaufgabe und die dazugehörige Menge

$$K := \{x \in \mathbb{R}^n \ : \ Ax = b, \ x \geq 0\}.$$

Offenbar ist K *konvex*, für je zwei Punkte $y^1, y^2 \in K$ liegt stets auch die Verbindungsstrecke $[y^1, y^2] \in K$.

Beispiel 3.4
Offenbar ist $0 \in K$ genau dann, wenn $b = 0$. 0 ist dann eine Ecke von K, denn für alle $y^1, y^2 \in K$ mit $0 = (1-\lambda)y^1 + \lambda y^2$, $\lambda \in [0,1]$ gilt wegen $y_1, y_2 \geq 0$, dass $y_1 = y_2 = 0$.

Definition 3.5
Teilmengen $I \subseteq \{1, \ldots, n\}$ nennen wir auch Indexmengen *und bezeichnen die Anzahl ihrer Elemente mit $|I|$. Für $I \neq \emptyset$ bilden wir die aus den dazugehörigen Spalten bestehende Teilmatrix $A = (a_{ij})_{i,j=1,\ldots,n} \in \mathbb{R}^{m\times n}$*

$$A_I = (a_{ij})_{i=1,\ldots,n, \ j\in I} \in \mathbb{R}^{m\times|I|}$$

und zu $v \in \mathbb{R}^n$ den dazugehörigen Teilvektor $v_I = (v_j)_{j\in I} \in \mathbb{R}^{|I|}$.

Jedem $x = (x_1, \ldots, x_n) \in K$ ordnen wir die Indexmenge *seiner von Null verschiedenen Komponenten zu*

$$I_x := \{j \in \{1, \ldots, n\} \ : \ x_j > 0\} \subseteq \{1, \ldots, n\}.$$

Für $x \neq 0$ schreiben wir auch kurz A_x statt A_{I_x} und v_x statt v_{I_x}.

Offenbar gilt $A_x x_x = Ax = b$ für alle $0 \neq x \in K$ und allgemeiner

$$A_x v_x = Av \quad \forall v \in \mathbb{R}^n \text{ mit } I_v \subseteq I_x.$$

Lemma 3.6
(a) Ist $0 \neq x \in K$ keine Ecke von K so existieren paarweise und von x verschiedene $y^1, y^2 \in K$, so dass $x \in [y^1, y^2]$, $I_{y^1}, I_{y^2} \subseteq I_x$.

(b) Ein Punkt der zulässigen Menge $0 \neq x \in K$ ist genau dann eine Ecke von K, wenn A_x injektiv ist (also vollen Spaltenrang $\operatorname{rang}(A_x) = |I_x|$ *besitzt).*

Beweis: Ist $x \in K$ keine Ecke, dann existiert eine Strecke in K, die durch x geht, also $y^1, y^2 \in K$, $y^1 \neq y^2$ und $\lambda \in (0,1)$ mit

$$x = (1-\lambda)y^1 + \lambda y^2.$$

Für jedes $j \notin I_x$ folgt aus $x_j = 0$ auch $y^1_j = 0 = y^2_j$ (beachte $y^1, y^2 \geq 0$) und damit $I_{y^1}, I_{y^2} \subseteq I_x$. Dies zeigt (a).

Außerdem ist

$$A_x(y^1_x - y^2_x) = A(y^1 - y^2) = b - b = 0.$$

Da aus $y^1 \neq y^2$ folgt, dass $y^1_x \neq y^2_x$ ist, kann (wenn $x \in K$ keine Ecke ist) A_x nicht injektiv sein, was mit Kontraposition die Rückrichtung von (b) zeigt.

Auch den Beweis der Hinrichtung führen wir mit Kontraposition. Sei A_x nicht injektiv, d.h. es existiere ein $0 \neq y_x \in \mathbb{R}^{|I_x|}$ mit $A_x y_x = 0$, den wir durch Nullen zu einem Vektor $y \in \mathbb{R}^n$ fortsetzen (womit dann auch die Notation y_x gerechtfertigt ist).

Wähle $\epsilon > 0$ so klein, dass $\epsilon|y_j| < x_j$ für alle $j \in |I_x|$ und definiere

$$y^1 := x - \epsilon y \quad \text{und} \quad y^2 := x + \epsilon y.$$

Dann gilt offenbar $x \in [y^1, y^2]$ und (wegen $y_x \neq 0$) $x \neq y^1$ und $x \neq y^2$. Außerdem folgt aus der Wahl von ϵ und aus $Ay = A_x y_x = 0$, dass

$$\eta \geq 0 \quad \text{und} \quad A\eta = Ax = b \quad \forall \eta \in [y_1, y_2]$$

also $[y^1, y^2] \subseteq K$. □

Folgerung 3.7

(a) $x \mapsto I_x$ ist eine injektive Abbildung von der Menge aller Ecken von K in die Menge aller Indexmengen. Insbesondere existieren also nur endlich viele Ecken.

(b) Zu einer Indexmenge $\emptyset \neq I \subseteq \{1, \ldots, n\}$ existiert genau dann eine Ecke $x \in K$ mit $I_x = I$, falls das lineare Gleichungssystem $A_I \tilde{x} = b$ genau eine Lösung $\tilde{x} \in \mathbb{R}^{|I|}$ besitzt und alle Komponenten von $\tilde{x}$ positiv sind. In diesem Fall ist die Nullfortsetzung von $\tilde{x} \in \mathbb{R}^{|I|}$ die (nach (a) eindeutig bestimmte) Ecke $x \in K$ mit $I_x = I$.

Beweis: (a) Offenbar ist $x = 0$ genau dann wenn $I_x = \emptyset$. Seien also $x, y \in K$ zwei Ecken mit gleicher Indexmenge $\emptyset \neq I_x = I_y =: I$. Dann ist $A_I x_I = b = A_I y_I$ und mit der in Lemma 3.6 gezeigten Injektivität von A_I folgt $x_I = y_I$. Da x und y durch Nullfortsetzung aus x_I und y_I hervorgehen, folgt $x = y$.

(b) Sei $\emptyset \neq I \subseteq \{1, \ldots, n\}$ und $x \in K$ eine Ecke mit $I_x = I$. Dann ist $A_I = A_x$ nach Lemma 3.6 injektiv. Außerdem ist $A_x x_x = b$, so dass $\tilde{x} := x_x \in \mathbb{R}^{|I|}$ die eindeutige Lösung von $A_I \tilde{x} = b$ ist. Offenbar ist auch $\tilde{x} > 0$ und x ist die Nullfortsetzung von $\tilde{x}$.

Sei nun $\emptyset \neq I \subseteq \{1, \ldots, n\}$ und $\tilde{x} \in \mathbb{R}^{|I|}$ sei die eindeutige Lösung von $A_I \tilde{x} = b$. Außerdem gelte $\tilde{x} > 0$. Sei $x \in \mathbb{R}^n$ die Nullfortsetzung von $\tilde{x}$, dann gilt $I_x = I$ und $x_x = \tilde{x}$. Aus $Ax = A_x x_x = A_I \tilde{x} = b$ und $x \geq 0$ folgt $x \in K$. Nach Lemma 3.6 ist x also eine Ecke, da $A_x = A_I$ injektiv ist. □

Lemma 3.8
(a) Jeder Punkt $x \in K$ mit minimaler Indexmenge, also

$$|I_x| \leq |I_y| \quad \forall y \in K$$

ist eine Ecke. Insbesondere besitzt daher jeder nicht-leere zulässige Bereich Ecken.

(b) Existiert auf dem zulässigen Bereich ein Minimum der Zielfunktion, so ist (mindestens) einer der Minimierer eine Ecke.

Beweis: Ist $I_x = \emptyset$, so ist $0 = x \in K$ und 0 ist nach Beispiel 3.4 eine Ecke. Sei also $x \neq 0$.

Angenommen x ist keine Ecke. Nach Lemma 3.6 existieren dann paarweise und von x verschiedene $y^1, y^2 \in K$, so dass $x \in [y^1, y^2]$, $I_{y^1}, I_{y^2} \subseteq I_x$.

Betrachte $v := y^2 - y^1$. Für jedes $\lambda \in \mathbb{R}$ ist

$$A(x + \lambda v) = b + \lambda(b - b) = b.$$

Wir werden zeigen, dass ein λ existiert, so dass $z := x + \lambda v \geq 0$ (und damit zulässig) ist und $|I_z| < |I_x|$ gilt. Dazu beachten wir zunächst dass offenbar $\emptyset \neq I_v \subseteq I_x$. Für jedes $j \in I_v$ setzen wir jetzt

$$\lambda_j := -x_j / v_j = -x_j / (y_j^2 - y_j^1) \neq 0$$

und wählen dann $k \in I_v$, so dass $|\lambda_k| \leq |\lambda_j|$ für alle $j \in I_v$. Das damit definierte $z := x + \lambda_k v$ erfüllt $z_k = x_k + \lambda_k v_k = 0$ und außerdem

$$z_j = x_j + \lambda_k v_j \geq 0,$$

denn wäre z_j negativ, so müsste die Funktion $\lambda \mapsto x_j + \lambda v_j$ eine betragskleinere Nullstelle λ_j als λ_k besitzen. Dies zeigt $z \in K$. Außerdem gilt offenbar $x \in [y^1, z]$ für $\lambda_k > 0$ und $x \in [y^2, z]$ für $\lambda_k < 0$.

Insgesamt müssten also, falls x keine Ecke ist, $y, z \in K$, $\lambda \in (0,1)$ existieren mit $x = (1-\lambda)y + \lambda z$ und $|I_z| < |I_x|$. Dies zeigt (a).

Zum Beweis von (b) sei nun x ein Minimierer mit (unter den Minimierern) minimaler Indexmenge, also

$$x \in \arg\min_{\xi \in K} c^T \xi \quad \text{und} \quad |I_x| \leq |I_{x'}| \quad \forall x' \in \arg\min_{\xi \in K} c^T \xi.$$

Falls x keine Ecke ist, existieren $y, z \in K$, $\lambda \in (0,1)$ mit $x = (1-\lambda)y + \lambda z$ und $|I_z| < |I_x|$. Da x ein Minimierer ist, ist $c^T y \geq c^T x$ und $c^T z \geq c^T x$. Für das Zielfunktional gilt aber auch

$$c^T x = (1-\lambda) c^T y + \lambda c^T z,$$

so dass (insbesondere) $c^T z > c^T x$ nicht möglich ist. z wäre also auch ein Minimierer im Widerspruch zur Minimialität von $|I_x|$ unter allen Minimierern. Damit ist auch (b) gezeigt. □

Bemerkung 3.9
Das betrachtete lineare Optimierungsproblem ist damit gelöst. Es genügt für jede (der endlich vielen, höchstens m-elementigen) nichtleeren Teilmengen von $\{1, \ldots, n\}$ gemäß Korollar 3.7 zu prüfen, ob es eine dazugehörige Ecke $x \in K$ gibt. So erhält man alle Ecken von K (wobei für $b = 0$ noch die Null hinzuzunehmen ist). Gibt es einen Minimierer, so ist dies eine der Ecken, und es müssen nur noch die Wert des Zielfunktionals für die endlich vielen Ecken berechnet und verglichen werden. Falls das Minimierungsproblem lösbar ist, so erhalten wir auf diese Weise also nach endlich vielen Schritten die Lösung.

Ein Nachteil dieses Algorithmus ist, dass er uns nicht zeigt, ob das Zielfunktional nach unten unbeschränkt ist. Noch schwerwiegender ist aber sein großer Aufwand, da es immer nötig ist alle Ecken, also alle (bis zu m-elementige) Teilmengen von $\{1, \ldots, n\}$ zu überprüfen. Nach einem kurzen Einschub über eine Anwendung aus der Spieltheorie im nächsten Abschnitt werden wir einen systematischeren Algorithmus kennenlernen, der sich von überprüfter Ecke zu überprüfter Ecke verbessert, und deshalb in der Praxis meist deutlich schneller ist.

3.1.3 Eine Anwendung aus der Spieltheorie

Wir betrachten ein Spiel zwischen einem Spieler S und seinem Gegner G, die jeweils eine von m bzw. n möglichen Entscheidungen treffen. a_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$ sei die Wahrscheinlichkeit, dass S das Spiel gewinne, falls er

sich für die j-te und der Gegner G sich für die i-te Möglichkeit entschieden habe. In der Vorlesung geben wir eine weihnachtliche Motivation für solche Situationen.

Die Ausgänge des Spiels fassen wir in der *Gewinnmatrix* $A = (a_{ij}) \in \mathbb{R}^{n\times m}$ zusammen. (Im analogen Fall, bei dem a_{ij} die Höhe des Gewinns für S bezeichnet, heißt A auch *Auszahlungsmatrix*.)

Die Strategien des Spielers S und seines Gegners G seien durch Vektoren $x = (x_j) \in \mathbb{R}^m$, $y = (y_i) \in \mathbb{R}^n$ beschreiben. Dabei sei x_j die Wahrscheinlichkeit, dass S die j-te Wahlmöglichkeit wählt und y_i die Wahrscheinlichkeit, dass G die j-te Wahl trifft.

Bezeichne W das Ereignis, dass der Spieler S gewinnt. X_j und Y_i seien das Ereignis, dass S die j-te Wahl bzw. G die i-te Wahl trifft. Dann ist also

$$P(X_i) = x_i, \quad P(Y_j) = y_j, \quad P(W|X_i \cap Y_j) = a_{ij}$$

Mit den üblichen Notationen und Regeln der Wahrscheinlichkeitstheorie gilt

$$\begin{aligned} P(W) &= P(W \cap Y_1) + \ldots + P(W \cap Y_n) \\ &= P(W|Y_1)P(Y_1) + \ldots + P(W|Y_n)P(Y_n) \end{aligned}$$

und

$$\begin{aligned} P(W|Y_i) &= P(W \cap X_1|Y_i) + \ldots + P(W \cap X_m|Y_i) \\ &= P(W|X_1 \cap Y_i)P(X_1) + \ldots + P(W|X_m \cap Y_i)P(X_n) = \sum_{j=1}^{m} a_{ij}x_j, \end{aligned}$$

also

$$P(W) = \sum_{i=1}^{n}\sum_{j=1}^{m} y_i a_{ij} x_j = y^T A x.$$

Der Spieler S versucht nun seine Strategie so zu wählen, dass sie selbst bei schlimmstmöglicher Wahl des Gegners G noch bestmöglich ist, er sucht also $x \in \mathbb{R}^m$ mit

$$f(x) \to \max! \quad \text{u.d.N. } \mathbb{1}^T x = 1, x \geq 0,$$

wobei

$$f(x) := \min\{y^T A x : \ y \in \mathbb{R}^n, \ y \geq 0, \ \mathbb{1}^T y = 1\} \quad \text{und} \quad \mathbb{1} = \begin{pmatrix} 1 & \ldots & 1 \end{pmatrix}^T.$$

Die Suche nach einer optimalen Strategie führt demnach auf ein kompliziertes restringiertes Maximierungsproblem, bei dem jede Auswertung der Zielfunktion $x \mapsto f(x)$ selbst die Lösung eines linearen Minimierungsproblems

erfordert. Überraschenderweise lässt sich die Maximierung von f jedoch in ein lineares Optimierungsproblem überführen. Dazu betrachten wir für festes $x \in \mathbb{R}^m$ mit $x \geq 0$ die Funktion

$$y \mapsto y^T Ax = \sum_{i=1}^{n} y_i (Ax)_i \qquad (y \in \mathbb{R}^n,\ y \geq 0,\ \mathbb{1}^T y = 1)$$

wobei $(Ax)_i$ die i-te Komponente von $Ax \in \mathbb{R}^n$ sei. Offenbar wird diese Summe minimal für einen Einheitsvektor $y = e_i$, wenn $i \in \{1, \ldots, n\}$ ein Index ist, für den der i-te Eintrag von Ax minimal ist. Es gilt also

$$f(x) = \min\{y^T Ax :\ y \in \mathbb{R}^n,\ y \geq 0,\ \mathbb{1}^T y = 1\} = \min\{(Ax)_i :\ i = 1, \ldots, n\}$$

und dies können wir umformen zu einem Maximumsausdruck

$$f(x) = \min\{(Ax)_i :\ i = 1, \ldots, n\} = \max\{\lambda :\ \lambda \in \mathbb{R},\ \mathbb{1}\lambda \leq Ax\}$$

Das Problem

$$f(x) \to \max! \quad \text{u.d.N. } x \in \mathbb{R}^m,\ x \geq 0,\ \mathbb{1}^T x = 1$$

ist also äquivalent zum linearen Optimierungsproblem

$$g(x, \lambda) := \lambda \to \max! \quad \text{u.d.N. } (x, \lambda) \in \mathbb{R}^{m+1},\ x \geq 0,\ \mathbb{1}^T x = 1,\ \mathbb{1}\lambda \leq Ax.$$

Nach Umformulierung dieses Problems in Normalform können wir die optimale Strategie also mit dem in Bemerkung 3.9 skizzierten Algorithmus oder dem im nächsten Abschnitt eingeführten Simplex-Verfahren bestimmen, siehe Übungsaufgabe ?.

3.1.4 Das Simplexverfahren

Wir stellen in dieser Vorlesung das Simplex-Verfahren nur für den Fall vor, dass $n > m$ und dass die Indexmengen aller Ecken von K genau m-elementig sind (diese Ecken heißen auch *nicht entartet*). Mit Lemma 3.8(a) folgt, dass es keine kleineren Indexmengen von zulässigen Punkten gibt, und damit wiederum mit Lemma 3.8(a), dass ein zulässiger Punkt genau dann eine Ecke ist, wenn er eine m-elementige Indexmenge besitzt.

Das Verfahren operiert direkt auf den Indexmengen. Entsprechend bezeichnen wir von nun an die Indexmenge zu einer Ecke direkt mit $B \subseteq \{1, \ldots, n\}$ (*Basisvariablen*) und ihr Komplement mit $N := \{1, \ldots, n\} \setminus B$ (*Nichtbasisvariablen*). Die aus den zugehörigen Spalten bzw. Komponenten bestehenden

Teilmatrizen bzw. Teilvektoren von A bzw. einem Vektor v bezeichnen wir entsprechend mit A_B, A_N, v_B und v_N.

Das Grundprinzip des Algorithmus besteht darin, so von Ecke zu Ecke zu gehen, dass sich der Wert des Zielfunktionals in jedem Schritt verringert. Aus einer (zu einer Ecke gehörigen) Indexmenge B wird also solange in jedem Schritt eine neue (zu einer *besseren* Ecke gehörigen) Indexmenge B' zu konstruiert, bis die Indexmenge zu einer *optimalen* Ecke gehört.

Wir beschreiben das Verfahren von hinten nach vorne. Unter der Annahme, dass B zu einer Ecke des zulässigen Bereichs gehört, zeigen wir

1. wie wir erkennen können, ob B zu einer optimalen Ecke gehört (das Stoppkriterium für das Simplex-Verfahren).

2. wie ein zu einer nicht optimalen Ecke gehörendes B zu einem zu einer besseren Ecke gehörendes B' verbessert werden kann (der Pivotschritt).

Dann zeigen wir,

3. wir wir für jedes Problem (mit nicht-leerem zulässigen Bereich), eine Ecke des zulässigen Bereichs finden können (Bestimmung einer Startecke).

3.1.4.1 Das Stoppkriterium

$B \subseteq \{1, \ldots, n\}$ gehöre zu einer Ecke. Nach Lemma 3.6 ist dann $A_B \in \mathbb{R}^{m \times |B|}$ injektiv. Aufgrund unserer Annahme, dass alle Ecken nicht entartet sind, ist $|B| = m$, also $A_B \in \mathbb{R}^{m \times m}$ bijektiv. Jeder zulässige Punkt $x \in K \subseteq \mathbb{R}^n$ ist durch $x_B \in \mathbb{R}^m$ und $x_N \in \mathbb{R}^{n-m}$ eindeutig festgelegt und da

$$b = Ax = A_B x_B + A_N x_N, \quad \text{also} \quad x_B = A_B^{-1}(b - A_N x_N).$$

ist jeder zulässige Punkt (bei festem B) sogar schon durch Kenntnis von $x_N \geq 0$ eindeutig festgelegt. Die zu B gehörige Ecke ist in diesem Sinne der zu $x_N = 0$ gehörige Punkt. (Beachte aber, dass nicht jede Wahl von $x_N \geq 0$ zu einem zulässigen Punkt, d.h. zu $x_B \geq 0$ führen muss.)

Wie ändert sich das Zielfunktional, wenn wir statt der zu B gehörigen Ecke einen anderen Punkt wählen? In der zu B gehörigen Ecke ist

$$c^T x = c_B^T x_B + c_N^T x_N = c_B^T A_B^{-1} b,$$

und in jedem anderen zulässigen Punkt ist

$$\begin{aligned} c^T x &= c_B^T x_B + c_N^T x_N = c_B^T A_B^{-1}(b - A_N x_N) + c_N^T x_N \\ &= c_B^T A_B^{-1} b + (c_N^T - c_B^T A_B^{-1} A_N) x_N. \end{aligned}$$

Wenn es einen Punkt mit niedrigerem Zielfunktionalwert als die zu B gehörige Ecke gibt, so gibt es ein dazugehöriges $x_N \geq 0$ mit $(c_N^T - c_B^T A_B^{-1} A_N) x_N \leq 0$. Ist die Ecke nicht optimal, so muss also der sogenannte *Vektor der reduzierten Kosten*

$$c_N^T - c_B^T A_B^{-1} A_N$$

mindestens einen negativen Eintrag besitzen. Damit haben wir das

Stoppkriterium für das Simplex-Verfahren: Gilt $c_N^T - c_B^T A_B^{-1} A_N \geq 0$, so gehört B zu einer optimalen Ecke und wir beenden das Verfahren.

3.1.4.2 Der Pivotschritt

Wie finden wir, wenn das Stoppkriterium an der aktuellen Ecke noch nicht erfüllt ist, eine zu einer besseren Ecke gehörende Indexmenge?

Es gelte $c_N^T - c_B^T A_B^{-1} A_N \ngeq 0$. Sei etwa der j-te Eintrag ($j \in N$) dieses Vektors negativ.

Jedes $x_N = (x_k)_{k \in N}$ mit $x_k = 0$ für $k \in N \setminus \{j\}$ und $x_j > 0$ führt dann also (mit dem dazugehörigen $x_B := A_B^{-1}(b - A_N x_N)$ wie im letzten Abschnitt) zu einem $x \in \mathbb{R}^n$ mit geringerem Zielfunktionalswert und der Wert ist um so geringer, umso größer x_j gewählt wurde. Das so erhaltene neue x ist offenbar genau dann zulässig, wenn $x_B \geq 0$ gilt.

Wir betrachten nun x_B in Abhängigkeit von x_j. Für $x_j = 0$ ist $x_N = 0$ und x_B enthält die von Null verschiedenen Koordinaten der aktuellen Ecke. Offenbar hängt x_B stetig von x_j ab, d.h. es existiert ein $\epsilon > 0$, so dass alle Einträge von x_B positiv bleiben für $x_j \in [0, \epsilon)$.

Nun kann entweder

(a) x_B **positiv bleiben für alle** $x_j > 0$**.** In dem Fall lässt sich das Zielfunktional beliebig klein machen und wir brechen das Verfahren ab mit dem Ergebnis, dass das Zielfunktional auf dem zulässigen Bereich nach unten unbeschränkt ist.

(b) x_B **aufhören, positiv zu bleiben.** Nach dem Zwischenwertsatz existiert dann ein kleinster Wert $\hat{x}_j$, für den das erste Mal ein Eintrag von

x_B Null wird. Wir ersetzen dann den dazugehörigen Index in B durch j und definieren so B'.

Aus unserer Annahme, dass alle Ecken m-elementige Indexmengen besitzen, folgt mit Lemma 3.8(a), dass es keine kleineren Indexmengen gibt, und damit wiederum mit Lemma 3.8(a), dass ein zulässiger Punkt genau dann eine Ecke ist, wenn er eine m-elementige Indexmenge besitzt. Da $|B| = |B'| = m$ ist, und der zu B' gehörige Punkt nach Konstruktion zulässig ist, gehört B' zu einer Ecke mit (wiederum nach Konstruktion) geringerem Zielfunktionalwert.

Zur Implementierung drücken wir (a) und (b) noch in Formeln aus. Dazu bezeichne $\hat{a}_{kl}$ die Einträge von $\hat{A} := A_B^{-1}A_N \in \mathbb{R}^{m\times(n-m)}$ und $\hat{b}_k$ die Einträge von $\hat{b} := A_B^{-1}b \in \mathbb{R}^m$. Da für die zu B gehörige Ecke $x_N = 0$ und $x_B = A_B^{-1}b = \hat{b}$ gilt, sind alle Einträge von $\hat{b}$ positiv.

O.B.d.A. seien die Indizes so sortiert, dass

$$N = \{1, \dots, n-m\} \quad \text{und} \quad B = \{n-m+1, \dots, n\}. \tag{3.1}$$

Dann ist $c_N = (c_k)_{k=1}^{n-m}$, $c_B = (c_{n-m+k})_{k=1}^{m}$ und der j-te Eintrag von $c_N^T - c_B^T A_B^{-1} A_N$ ist $(j = 1, \dots, n-m)$

$$c_j - \sum_{k=1}^{m} c_{n-m+k}\hat{a}_{kj}$$

also

$$c_N^T - c_B^T A_B^{-1} A_N \not\geq 0 \iff \exists j \in \{1, \dots, n-m\} : \; c_j - \sum_{k=1}^{m} c_{n-m+k}\hat{a}_{kj} < 0.$$

Aus $x_B = A_B^{-1}(b - A_N x_N)$ folgt (mit dieser Nummerierung)

$$x_{n-m+k} = \hat{b}_k - \sum_{l=1}^{n-m} \hat{a}_{kl} x_l \quad \forall k \in \{1, \dots, m\}.$$

Ist wie in der obigen Konstruktion $x_l = 0$ für all $l \in N \setminus \{j\}$ so ergibt sich

$$x_{n-m+k} = \hat{b}_k - \hat{a}_{kj} x_j \quad \forall k \in \{1, \dots, m\}.$$

Fall (a) tritt ein, wenn alle $x_{n-m+k} > 0$ positiv bleiben für $x_j > 0$, also (da $\hat{b} > 0$) genau dann wenn $\hat{a}_{kj} \leq 0$ für alle $k \in \{1, \dots, m\}$.

Ist $\hat{a}_{kj} > 0$ für mindestens ein k, dann tritt (b) ein und es ist

$$0 = x_{n-m+k} = \hat{b}_k - \hat{a}_{kj} x_j \quad \Longleftrightarrow \quad x_j = \frac{\hat{b}_k}{\hat{a}_{kj}}.$$

Den durch j zu ersetzenden Index $n - m + \hat{k}$ erhalten wir gemäß der Beschreibung von (b) durch

$$\hat{k} := \arg \min_{\substack{k=1,\dots,m \\ \hat{a}_{kj}>0}} \frac{\hat{b}_k}{\hat{a}_{kj}}.$$

Wir ersetzen also den in N enthaltenen Index j durch den Index $n - m + \hat{k}$, d.h. den j-ten Eintrag von N durch den $\hat{k}$-ten Eintrag von B.

Bemerkung 3.10
Ohne die Sortierungsannahme (3.1) ist der j-te Eintrag von $c_N^T - c_B^T A_B^{-1} A_N \in \mathbb{R}^{n-m}$ gegeben durch

$$c_{N(j)} - \sum_{k=1}^{m} c_{B(k)} \hat{a}_{kj}$$

wobei $N(j)$ das j-te Element von N und $B(k)$ das k-te Element von B bezeichne. Die Elemente von $x_B = A_B^{-1}(b - A_N x_N) \in \mathbb{R}^m$ sind gegeben durch

$$x_{B(k)} = \hat{b}_k - \sum_{l=1}^{n-m} \hat{a}_{kl} x_{N(l)} \quad \forall k \in \{1, \dots, m\}.$$

Mit $x_{N(l)} = 0$ für alle $l \neq j$ folgt

$$x_{B(k)} = \hat{b}_k - \hat{a}_{kj} x_{N(j)}.$$

Wir wählen also wie oben (falls Elemente $\hat{a}_{kj} > 0$ existieren)

$$\hat{k} := \arg \min_{\substack{k=1,\dots,m \\ \hat{a}_{kj}>0}} \frac{\hat{b}_k}{\hat{a}_{kj}}$$

und ersetzen das $\hat{k}$-te Element von B durch das j-te Element von N.

Zusammengefasst erhalten wir so das Simplex-Verfahren für lineare Optimierungsaufgaben in Normalform ohne entartete Ecken in Algorithmus 11. Bei der Formulierung schließen wir den trivialen Fall $m = n$ aus.

Algorithm 11 Simplex-Verfahren

Gegeben lineare Opt.aufgabe in Normalform ohne entartete Ecken
$A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}^m$, $n > m$

$\emptyset \neq B \subset \{1,\dots,n\}$ sei Indexmenge eine zulässigen Ecke.
$N := \{1,\dots,n\} \setminus B$. $\hat{A} := A_B^{-1} A_N \in \mathbb{R}^{m\times(n-m)}$, $\hat{b} := A_B^{-1} b \in \mathbb{R}^m$

while $\exists j \in \{1,\dots,n-m\}:\ c_{N(j)} - \sum_{k=1}^{m} c_{B(k)} \hat{a}_{kj} < 0$ **do**
  if $\exists k:\ \hat{a}_{kj} > 0$ **then**
    Bestimme $\hat{k} \in \{1,\dots,m\}$, so dass $\frac{\hat{b}_{\hat{k}}}{\hat{a}_{\hat{k}j}} \leq \frac{\hat{b}_k}{\hat{a}_{kj}}$ für alle k mit $\hat{a}_{kj} > 0$.
    $B := (B \setminus \{B(\hat{k})\}) \cup \{N(j)\}$.
    $N := \{1,\dots,n\} \setminus B$. $\hat{A} := A_B^{-1} A_N$, $\hat{b} := A_B^{-1} b$
  else
    return Zielfunktional ist nach unten unbeschränkt.
  end if
end while
return Minimierer x ist gegeben durch $x_B := \hat{b}$, $x_N := 0$.

3.1.4.3 Bestimmung einer Startecke

Wir zeigen noch, wie sich für jedes Problem eine Startecke konstruieren lässt (bzw. gezeigt werden kann, dass der zulässige Bereich leer ist). Durch Negation von Zeilen von A können wir ohne Einschränkung $b \geq 0$ annehmen. Dann betrachten wir das folgende *Ersatzproblem*:

$$\text{Minimiere} \quad \begin{pmatrix} 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

unter den Nebenbedingungen

$$\begin{pmatrix} A & I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = b \quad \text{und} \quad \begin{pmatrix} x \\ y \end{pmatrix} \geq 0,$$

wobei $y \in \mathbb{R}^m$ und $I \in \mathbb{R}^{m\times m}$ die Einheitsmatrix ist.

Wieder betrachten wir zur Vereinfachung nur den Fall, dass das Ersatzproblem keine entarteten Ecken besitzt. Dann ist $\begin{pmatrix} 0 & b^T \end{pmatrix}^T$ eine Ecke des zulässigen Bereichs des Ersatzproblems, da die dazugehörige Indexmenge höchstens m-elementig ist.

Ist der zulässige Bereich des ursprünglichen Problems nicht leer, so sind die Minimierer für das Ersatzproblem offensichtlich genau die Vektoren $\begin{pmatrix} x \\ 0 \end{pmatrix}$ für die x im zulässigen Bereich liegt. Außerdem ist dann (da für beide Probleme

die Ecken genau die Punkte mit m-elementiger Indexmenge sind) eine optimale Ecke des Ersatzproblems eine Ecke des ursprünglichen Problems. Ist der ursprüngliche zulässige Bereich nicht leer, so lässt sich also durch Anwendung des Simplex-Verfahrens auf das Ersatzproblem eine zulässige Ecke des ursprünglichen Problems finden.

Ist der zulässige Bereich des ursprünglichen Problems dagegen leer, so haben die Minimierer des Ersatzproblem nicht die Form $\begin{pmatrix} x \\ 0 \end{pmatrix}$ (oder es existieren keine). Die Anwendung des Simplex-Verfahrens auf das Ersatzproblem liefert also auch die Aussage, ob der ursprüngliche zulässige Bereich leer ist oder nicht.

3.2 Restringierte nichtlineare Optimierung

Dieser Abschnitt folgt wieder dem Lehrbuch von Michael und Stefan Ulbrich [Ulbrich].

Wir betrachten nun nicht-lineare restringierte Optimierungsprobleme der Form

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{u.d.N. } g(x) \leq 0,\ h(x) = 0 \tag{3.2}$$

mit

- *Zielfunktion* $f:\ \mathbb{R}^n \to \mathbb{R}$,
- *Ungleichungsrestriktionen* $g:\ \mathbb{R}^n \to \mathbb{R}^m$ und
- *Gleichungsrestriktionen* $h:\ \mathbb{R}^n \to \mathbb{R}^p$.

Wir setzen in diesem Abschnitt über restringierte nichtlineare Optimierung stets voraus, dass f, g und h stetig differenzierbar sind.

Analog zu Definition 3.5 definieren wir:

Definition 3.11
Die Menge

$$X = \{x \in \mathbb{R}^n :\ g(x) \leq 0,\ h(x) = 0\}$$

heißt ***zulässiger Bereich*** *des Optimierungsproblems (3.2).*

Für zulässige *Punkte* $x \in X$ *heißt*

$$I_x := \{j \in \{1, \ldots, m\}\ :\ g_j(x) < 0\} \subseteq \{1, \ldots, m\}$$

die ***Indexmenge inaktiver Ungleichungsnebenbedingungen****.*

3.2.1 Optimalitätsbedingungen

Die in Kapitel 2 vorgestellten Verfahren zur lokalen unrestringierten Optimierung einer Funktion

$$f(x) \to \min! \quad \text{u.d.N. } x \in \mathbb{R}^n$$

beruhten darauf, dass in einem lokalen Minimum $\hat{x} \in \mathbb{R}^n$ stets die notwendige Bedingung 1. Ordnung aus Definition und Satz 2.10 gelten muss

$$\nabla f(\hat{x}) = 0. \tag{3.3}$$

Jede Richtung $d \in \mathbb{R}^n$ mit $\nabla f(\hat{x})^T d < 0$ ist eine Abstiegsrichtung für f. Für hinreichend kleine $s > 0$ ist $f(\hat{x} + sd) < f(\hat{x})$ (siehe Definition 2.21 und Satz 2.22). Die für ein lokales Minimum im unrestringierten Fall notwendige Bedingung 1. Ordnung (3.3) können wir also so interpretieren, dass es in einem lokalen Minimum keine Abstiegsrichtungen geben kann.

Tangentialkegel. Im restringierten Fall gilt die notwendige Bedingung (3.3) nicht mehr.

Beispiel 3.12
Das Minimierungsproblem

$$f(x) := x \to \min! \quad \textit{u.d.N. } x \in \mathbb{R}, x \geq 0$$

besitzt offensichtlich das Minimum $\hat{x} = 0$, aber $f'(\hat{x}) = 1$. $d = -1$ ist eine Abstiegsrichtung für f, jedoch gehört $\hat{x} + sd$ für kein $s > 0$ zum zulässigen Bereich.

In einer lokalen Lösung $\hat{x}$ eines restringierten Minimierungsproblems können also durchaus Abstiegsrichtung für f existieren, so lange lokal in dieser Richtung kein zulässiger Punkt liegt. Aus dieser Anschauung heraus, erhalten wir eine erste notwendiges Optimalitätbedingung für restringierte Optimierungsprobleme:

Definition 3.13
*(a) Eine Menge $K \subseteq \mathbb{R}^n$ heißt **Kegel**, falls*

$$\lambda x \in K \quad \textit{für alle } \lambda \geq 0,\ x \in K.$$

(b) Für eine nicht-leere Menge $X \subseteq \mathbb{R}^n$ heißt ein normierter Vektor $d \in \mathbb{R}^n$, $\|d\| = 1$ ***Tangentialrichtung im Punkt*** *$x \in X$, falls eine Folge $(x^k)_{k\in\mathbb{N}} \subseteq X$ existiert mit*

$$x^k \to x, \quad x^k \neq x, \qquad \text{und} \qquad \lim_{k\to\infty} \frac{x^k - x}{\|x^k - x\|} = d.$$

Bezeichnet X den zulässigen Bereich eines Optimierungsproblems, so sprechen wir auch von ***zulässigen Richtungen****.*

Der von den Tangentialrichtungen aufgespannte Kegel heißt ***Tangentialkegel***

$$T(X, x) := \{\lambda d : \ \lambda \geq 0, \ d \text{ Tangentialrichtung}\} \subseteq \mathbb{R}^n.$$

Beispiel 3.14

(a) Ist X offen, dann gilt für alle $x \in X$

$$T(X, x) = \mathbb{R}^n.$$

(b) Ist x ein isolierter Punkt von X, so ist $T(X, x) = \emptyset$.

(c) Für Häufungspunkte x einer Menge $X \subseteq \mathbb{R}^n$ gilt stets $T(X, x) \neq \emptyset$ (Übungsaufgabe 13.1).

Satz 3.15

Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und $X \subseteq \mathbb{R}^n$ eine beliebige Menge. In jedem lokalen Minimum $\hat{x} \in X$ von f in X gilt

$$\nabla f(\hat{x})^T d \geq 0 \quad \text{für alle } d \in T(X, \hat{x}).$$

Beweis: Offenbar genügt es die Behauptung für alle Tangentialrichtungen zu zeigen. Sei also $d \in \mathbb{R}^n$ mit $\|d\| = 1$ und $(x^k)_{k\in\mathbb{N}} \subseteq X$ eine Folge mit

$$x^k \to \hat{x}, \quad x^k \neq \hat{x} \quad \text{und} \quad \lim_{k\to\infty} \frac{x^k - \hat{x}}{\|x^k - \hat{x}\|} = d.$$

Ist $\hat{x}$ ein lokales Minimum von f in X, dann gilt $f(x^k) \geq f(\hat{x})$ für hinreichend große k, und aus der Taylor-Formel aus Satz 2.3 folgt

$$0 \leq \frac{f(x^k) - f(\hat{x})}{\|x^k - \hat{x}\|} = \frac{\nabla f(\hat{x})^T (x^k - \hat{x}) + \rho(x^k - \hat{x})}{\|x^k - \hat{x}\|} \to \nabla f(\hat{x})^T d,$$

und damit die Behauptung. □

Satz 3.15 gilt für allgemeine zulässige Bereiche $\emptyset \neq X \subseteq \mathbb{R}^n$, spielt jedoch in der Praxis jedoch üblicherweise keine Rolle, da der Tangentialkegel nicht leicht zu bestimmen ist.

Der linearisierte Tangentialkegel. Wir betrachten von nun an das in der Praxis meist relevante Problem eines Gleichungs- und Ungleichungsrestringierten Optimierungsproblem

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{u.d.N. } g(x) \le 0,\ h(x) = 0 \tag{3.4}$$

mit zulässigem Bereich X gemäß Definition 3.11.

Es liegt nahe, den Tangentialkegel in einem Punkt $x \in X$ durch Linearisierung der in x relevanten Nebenbedingungen zu beschreiben, vgl. die in der Vorlesung gemalten Skizzen:

Definition 3.16
*Zu $x \in X$ definieren wir den **linearisierten Tangentialkegel** durch*

$$T_l(g,h,x) := \{d \in \mathbb{R}^n :\ \nabla h_i(x)^T d = 0,\ \nabla g_j(x)^T d \le 0 \quad \forall i \in \{1,\dots,p\},\ j \notin I_x\}.$$

Lemma 3.17
Für alle $x \in X$ ist

$$T(X,x) \subseteq T_l(g,h,x).$$

Beweis: Offenbar ist T_l ein Kegel, so dass es genügt zu zeigen, dass alle Tangentialrichtungen in $T_l(g,h,x)$ enthalten sind. Sei also $d \in \mathbb{R}^n$ mit $\|d\| = 1$ und $(x^k)_{k\in\mathbb{N}} \subseteq X$ eine Folge mit

$$x^k \to x, \quad x^k \neq x \quad \text{und} \quad \lim_{k\to\infty} \frac{x^k - x}{\|x^k - x\|} = d.$$

Aus der Taylor-Formel aus Satz 2.3 folgt, dass für alle $i \in \{1,\dots,p\}$ und $j \notin I_x$

$$0 = \frac{h_i(x^k) - h_i(x)}{\|x^k - x\|} = \frac{\nabla h_i(x)^T(x^k - x) + \rho_{h_i}(x^k - x)}{\|x^k - x\|} \to \nabla h_i(x)^T d,$$
$$0 \ge \frac{g_j(x^k) - g_j(x)}{\|x^k - x\|} = \frac{\nabla g_j(x)^T(x^k - x) + \rho_{g_j}(x^k - x)}{\|x^k - x\|} \to \nabla g_j(x)^T d,$$

wobei wir in der zweiten Zeile ausgenutzt haben, dass $g_j(x^k) \le 0 = g_j(x)$ gilt, da x^k zulässig und die j-te Ungleichungsnebenbedingung in x aktiv ist.□

Beispiel 3.18
Im Allgemeinen stimmt der linearisierte Tangentialkegel nicht mit dem Tangentialkegel überein. Mehr noch, der linearisierte Tangentialkegel hängt von der speziellen Wahl von g und h ab. Die Menge

$$\begin{aligned} X &= \{x \in \mathbb{R}^2 :\ -1 \le x_1 \le 1,\ x_2 = 0\} \\ &= \{x \in \mathbb{R}^2 :\ (x_1+1)^3 \ge x_2,\ x_1 \le 1,\ x_2 = 0\} \end{aligned}$$

lässt sich beschreiben durch

$$g(x) = \begin{pmatrix} -x_1 - 1 \\ x_1 - 1 \end{pmatrix}, \quad h(x) = x_2,$$

aber auch durch

$$\tilde{g}(x) = \begin{pmatrix} x_2 - (x_1 + 1)^3 \\ x_1 - 1 \end{pmatrix}, \quad h(x) = x_2.$$

In Übungsaufgabe ? zeigen wir, dass für $x = (-1, 0)^T \in X$ *gilt*

$$T(X, x) = T_l(g, h, x) \subsetneq T_l(\tilde{g}, \tilde{h}, x).$$

Constraint Qualifications. Wir suchen nun nach Bedingungen, die uns erlauben die notwendige Optimalitätsbedingung in Satz 3.15 mit dem linearisierten Tangentialkegel zu formulieren. Die einfachste solche Bedingung ist natürlich die folgende:

Definition 3.19
Die Bedingung $T_l(g, h, x) = T(X, x)$ *heißt* ***Abadie Constraint Qualification*** *(ACQ) für* $x \in X$*.*

Gilt (ACQ) in einem lokalen Minimum $\hat{x} \in X$, so ist nach Satz 3.15

$$\nabla f(\hat{x})^T d \geq 0 \quad \text{für alle } d \in T_l(g, h, x). \tag{3.5}$$

Die Aussage in (3.5) gilt jedoch auch schon unter schwächeren Voraussetzungen:

Definition 3.20
(a) Zu einem nichtleeren Kegel K *definieren wir den* ***Polarkegel*** *durch*

$$K^\circ = \{v \in \mathbb{R}^n : v^T d \leq 0 \quad \forall d \in K\}.$$

K° besteht also aus den Vektoren die zu allen Kegelvektoren in einem Winkel von mindestens $\pi/2$ stehen.

(b) Die Bedingung $T_l(g, h, x)^\circ = T(X, x)^\circ$ *heißt* ***Guignard Constraint Qualification*** *(GCQ) für* $x \in X$*.*

(c) Jede Bedingung, die (GCQ) impliziert, heißt ***Constraint Qualification*** *(CQ).*

Beispiel 3.21
(a) Offenbar ist die (ACQ) eine Constraint Qualification.

(b) Die Forderung

$$g_i \text{ konkav für alle } i \notin I_x, \quad h \text{ affin linear.}$$

ist eine Constraint Qualification, vgl. Übungsaufgabe ?.

(c) Ein Punkt $x \in X$ heißt ***regulär****, wenn die Vektoren*

$$\{\nabla h_i(x), \nabla g_j(x) : \ i \in \{1, \dots, p\}, \ j \notin I_x\} \subset \mathbb{R}^n$$

linear unabhängig sind. Die Forderung

$$x \text{ ist regulär}$$

ist eine Constraint Qualification, siehe z.B. [Ulbrich, 16.2].

Satz 3.22
In jedem lokalen Minimum $\hat{x} \in X$ von f in X, das eine Constraint Qualification erfüllt, gilt

$$\nabla f(\hat{x})^T d \geq 0 \quad \text{für alle } d \in T_l(g, h, \hat{x}).$$

Beweis: Nach Satz 3.15 gilt

$$\nabla f(\hat{x})^T d \geq 0 \quad \text{für alle } d \in T(X, \hat{x}),$$

also $-\nabla f(\hat{x}) \in T(X, \hat{x})^\circ$. (CQ) impliziert $-\nabla f(\hat{x}) \in T_l(g, h, \hat{x})^\circ$ und damit die Behauptung. □

3.2.2 Die Karush-Kuhn-Tucker-Bedingungen

Um Satz 3.22 in die übliche Form der notwendigen Optimalitätsbedingungen 1. Ordnung umzuformulieren, benötigen wir noch einige Vorbereitungen.

Das folgende Lemma zeigt, dass wir einen Punkt und eine abgeschlossene konvexe Menge stets durch eine Ebene voneinander trennen können (und wir im Fall eines abgeschlossenen konvexen Kegel diese Ebene durch den Nullpunkt gehen lassen können), vgl. die in der Vorlesung gemalten Skizzen.

Lemma 3.23 (Trennungssatz von Hahn-Banach)
Sei $\emptyset \neq K \subset \mathbb{R}^n$ abgeschlossen und konvex und sei $x \notin K$. Dann existieren $\nu \in \mathbb{R}^n$ und $r \in \mathbb{R}$ mit

$$\nu^T x > r \quad \text{aber} \quad \nu^T y \leq r \quad \text{für alle } y \in K.$$

Ist K ein abgeschlossener konvexer Kegel, dann gilt die Aussage mit $r = 0$.

Beweis: Die Ebene lässt sich anschaulich leicht dadurch konstruieren, dass wir den Punkt x auf die Menge K projezieren und eine Ebene durch den Projektionspunkt senkrecht zur Projektionsrichtung wählen, vgl. die in der Vorlesung gemalten Skizzen.

Wir führen diese Konstruktion nun rigoros durch. Nach Übungsaufgabe ? wird das Minimum zwischen einer kompakten Menge und einer abgeschlossenen Menge angenommen. Es existiert also ein $\eta \in K$ mit $\|x-\eta\| = \inf_{y\in K} \|x-y\|$. Entsprechend unserer anschaulichen Motivation setzen wir

$$\nu := x - \eta \quad \text{und} \quad r := \nu^T \eta.$$

Dann ist

$$\nu^T x := \nu^T (x-\eta) + r = \|\nu\|^2 + r \geq r,$$

und die erste Bedingung ist erfüllt.

Um die zweite Bedingung zu zeigen, sei $y \in K$. Da K konvex ist, liegt die Verbindungsstrecke $[y, \eta]$ in K. Aufgrund der Wahl von η gilt also für alle $t \in [0,1]$

$$\|t(y-\eta) - \nu\|^2 = \|ty + (1-t)\eta - x\|^2 \geq \|\eta - x\|^2 = \|\nu\|^2$$

also

$$t^2 \|y-\eta\|^2 - 2t(y-\eta)^T \nu \geq 0$$

und mit $t \to 0$ folgt $\nu^T y \leq \eta^T \nu = r$.

Sei nun K zusätzlich ein Kegel. Für das oben konstruierte $\nu \in \mathbb{R}^n$ und $r \in \mathbb{R}$ gilt dann

$$\nu^T y \leq r$$

für alle $y \in K$. Speziell für $0 \in K$ folgt $r \geq 0$. Mit

$$\nu^T x \geq r \geq 0$$

gilt also die erste Bedingung auch mit rechter Seite 0.

Außerdem gilt wegen der Kegeleigenschaft für alle $y \in K$ auch $\lambda y \in K$ und damit

$$\lambda \nu^T y \geq r \quad \text{für alle } \lambda \geq 0$$

mit Grenzübergang $\lambda \to \infty$ folgt

$$\nu^T y \geq 0,$$

so dass auch die zweite Bedingung mit rechter Seite 0 erfüllt ist. □

Lemma 3.24
Sei $A \in \mathbb{R}^{n\times m}$. A ist genau dann injektiv, wenn A^TA bijektiv ist.

Beweis: Für $v \in \mathbb{R}^m$ ist

$$A^TAv = 0 \quad \Longleftrightarrow \quad w^TA^TAv = 0 \quad \forall w \in \mathbb{R}^m \quad \Longleftrightarrow \quad Av = 0,$$

wobei die Hinrichtung im letzten Schritt aus der Wahl $w := v$ folgt. Injektivität von A ist also äquivalent zur Injektivität von A^TA und dieses ist (da A^TA eine quadratische Matrix ist) äquivalent zur Bijektivität von A^TA. □

Definition und Satz 3.25
Sei $A \in \mathbb{R}^{n\times m}$. Die Menge

$$K := \{Ax : \ x \in \mathbb{R}^m, \ x \geq 0\} \subseteq \mathbb{R}^n$$

*heißt (von den Spaltenvektoren von A) **endlich erzeugter Kegel**. K ist ein konvexer, abgeschlossener Kegel.*

Beweis: Die Konvexität und die Kegeleigenschaft ist trivial. Die Abgeschlossenheit zeigen wir mit vollständiger Induktion über $m \in \mathbb{N}$.

Wir beginnen mit dem Fall $m = 1$: Für $A = 0$ ist $K = \{0\}$ offensichtlich abgeschlossen. Für $0 \neq A \in \mathbb{R}^{n\times 1}$ ist A injektiv und und damit nach Lemma 3.24 A^TA invertierbar (d.h. $A^TA \neq 0 \in \mathbb{R}$) Für jede konvergente Folge $Ax^{(l)} \in K$ mit $x^{(l)} \geq 0$ ist wegen

$$x^{(l)} = (A^TA)^{-1}A^TAx^{(l)}$$

auch $(x^{(l)})_{l\in\mathbb{N}}$ eine konvergente Folge. $x := \lim_{l\to\infty} x^{(l)}$ erfüllt offenbar $x \geq 0$ und

$$\lim_{l\to\infty} Ax^{(l)} = Ax \in K.$$

Damit ist die Abgeschlossenheit von K für $m = 1$ gezeigt.

Für den Induktionsschritt sei die Aussage für $m \in \mathbb{N}$ bewiesen und es sei $A \in \mathbb{R}^{n,m+1}$. Ist A injektiv, so folgt die Abgeschlossenheit von $K = \{Ax : x \in \mathbb{R}^{m+1}, \ x \geq 0\}$ wie im Fall $m = 1$ aus der Invertierbarkeit von A^TA.

Sei also A nicht injektiv. Wir müssen zeigen, dass für jede konvergente Folge $Ax^{(l)} \in K$ mit $0 \leq x^{(l)} \in \mathbb{R}^{m+1}$ ein $0 \leq x \in \mathbb{R}^{m+1}$ existiert mit $Ax^{(l)} \to Ax$.

Da A nicht injektiv ist, existiert $0 \neq \xi \in \mathbb{R}^{m+1}$ mit $A\xi = 0$. Wie im Beweis von Lemma 3.8 können wir zu jedem $x^{(l)}$ ein $\alpha_l \in \mathbb{R}$ (möglicherweise $\alpha_l = 0$) finden, so dass $x^{(l)} + \alpha_l\xi \geq 0$ gilt und $x^{(l)} + \alpha_l\xi$ eine Nullkomponente

besitzt. Durch Übergang zu einer Teilfolge können wir o.B.d.A. annehmen, dass immer die selbe Komponente von $x^{(l)} + \alpha_l \xi$ gleich Null ist.

Sei $i \in \mathbb{N}$ der Index dieser Komponente und $I := \{1, \ldots, m+1\} \setminus \{i\}$. Mit der Notation aus Definition 3.5 ist

$$A_I(x_I^{(l)} + \alpha_l \xi_I) = A(x^{(l)} + \alpha_l \xi) = Ax^{(l)}.$$

Aufgrund der Induktionsannahme für $A_I \in \mathbb{R}^{n \times m}$ existiert ein $0 \leq x_I \in \mathbb{R}^m$ mit

$$A_I(x_I^{(l)} + \alpha_l \xi_I) \to A_I x_I$$

und durch Nullfortsetzung von x_I erhalten wir ein $0 \leq x \in \mathbb{R}^{m+1}$ mit $Ax^{(l)} \to Ax$, womit die Abgeschlossenheit gezeigt ist. □

Lemma 3.26 (Lemma von Farkas)
Seien $A \in \mathbb{R}^{n \times m}$ und $b \in \mathbb{R}^n$. Es gilt stets genau eine der beiden folgenden Aussagen:

(a) Es existiert ein $x \in \mathbb{R}^m$ mit $x \geq 0$ mit $Ax = b$.

(b) Es existiert ein $y \in \mathbb{R}^n$ mit $A^T y \geq 0$ und $b^T y < 0$.

Beweis: Offensichtlich können (a) und (b) nicht gleichzeitig gelten, da sonst

$$0 > y^T b = y^T Ax = (A^T y)^T x \geq 0.$$

Falls (a) nicht gilt, dann ist der endlich erzeugte Kegel

$$b \notin K := \{Ax : \ x \geq 0\} \subseteq \mathbb{R}^n$$

gemäß Definition und Satz 3.25 konvex und abgeschlossen. Nach dem Trennungssatz von Hahn-Banach existiert also ein $\nu \in \mathbb{R}^n$ mit

$$\nu^T b > 0 \quad \text{und} \quad \nu^T Ax \leq 0 \quad \forall x \geq 0.$$

Aus $\nu^T Ax \leq 0$ für alle $x \geq 0$ folgt $\nu^T A \leq 0$ und damit erfüllt $y := -\nu$ die Aussage in (b). □

Folgerung 3.27
Seien $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{n \times p}$, $c \in \mathbb{R}^n$. Dann sind äquivalent:

(a) Für alle $d \in \mathbb{R}^n$ mit $A^T d \leq 0$ und $B^T d = 0$ gilt $c^T d \leq 0$.

(b) Es gibt $u \in \mathbb{R}^m$ mit $u \geq 0$ und $v \in \mathbb{R}^p$ mit $c = Au + Bv$.

Beweis: Dies folgt durch Einführung von Schlupfvariablen sofort aus dem Lemma von Farkas, vgl. Übungaufgabe 14.2. □

Jetzt können wir die übliche Form der notwendigen Optimalitätsbedingung 1. Ordnung für ein restringiertes Optimierungsproblem angeben:

Satz 3.28
Seien $f: \mathbb{R}^n \to \mathbb{R}$, $g: \mathbb{R}^n \to \mathbb{R}^m$ *und* $h: \mathbb{R}^n \to \mathbb{R}^p$ *stetig differenzierbar. Ist* $\hat{x} \in \mathbb{R}^n$ *eine lokale Lösung von*

$$f(x) \to \min! \quad u.d.N. \;\; g(x) \leq 0, \; h(x) = 0,$$

in der eine Constraint Qualification gilt, dann existieren $\hat{\lambda} \in \mathbb{R}^m$ *und* $\hat{\mu} \in \mathbb{R}^p$*, so dass die folgenden* ***Karush-Kuhn-Tucker-Bedingungen*** *(KKT) erfüllt sind:* [1]

(a) $\nabla f(\hat{x}) + \nabla g(\hat{x})\hat{\lambda} + \nabla h(\hat{x})\hat{\mu} = 0$

(b) $h(\hat{x}) = 0$

(c) $\hat{\lambda} \geq 0$, $g(\hat{x}) \leq 0$, $\hat{\lambda}^T g(\hat{x}) = 0$.

Beweis: Nach Satz 3.22 gilt $\nabla f(\hat{x})^T d \geq 0$ für alle

$$d \in T_l(g,h,\hat{x}) = \{d \in \mathbb{R}^n : \; \nabla h_i(\hat{x})^T d = 0, \; \nabla g_j(\hat{x})^T d \leq 0 \quad \forall i \in \{1,\ldots,p\}, \; j \notin I_{\hat{x}}\}.$$

Mit

$$c := -\nabla f(\hat{x}) \in \mathbb{R}^n, \quad A = \left(\frac{\partial}{\partial_{x_k}} g_j(\hat{x})\right)_{\substack{k=1,\ldots,n \\ j \notin I_{\hat{x}}}} \in \mathbb{R}^{n \times (m - |I_{\hat{x}}|)},$$

$$B := \nabla h(\hat{x}) \in \mathbb{R}^{n \times p}$$

gilt also

$$c^T d \geq 0 \quad \text{für alle } d \in \mathbb{R}^n \text{ mit } B^T d = 0 \text{ und } A^T d \leq 0.$$

Durch Anwendung des Farkas-Lemmas in Form von Folgerung 3.27 ist das äquivalent zur Existenz von $\hat{\lambda} \in \mathbb{R}^{m-|I_{\hat{x}}|}$, $\hat{\lambda} \geq 0$ und $\hat{\mu} \in \mathbb{R}^p$ mit $c = A\hat{\lambda} + B\hat{\mu}$ und durch Nullfortsetzung von $\hat{\lambda}$ zu einem Vektor in $\mathbb{R}^m$ folgt die Behauptung. □

[1] Wir verwenden dabei auch für die möglicherweise nicht-skalaren Funktionen g und h die Notation $\nabla g(x) := g'(x)^T \in \mathbb{R}^{n \times m}$, $\nabla h(x) := h'(x)^T \in \mathbb{R}^{n \times p}$.

Im Falle eines lediglich Gleichungs-restringierten Optimierungsproblems ergibt sich aus Satz 3.28 (zusammen mit der Constraint Qualificiation aus Beispiel 3.21(c)) die aus der Analysis bekannte *Lagrange-Multiplikatorenregel*. Wir verwenden die entsprechenden Begriffe auch im allgemeinen Fall:

Definition 3.29
Die Vektoren λ und μ in Satz 3.28 heißen ***Lagrange-Multiplikatoren****. Die Funktion*

$$L: \ \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}, \quad L(x, \lambda, \mu) := f(x) + \lambda^T g(x) + \mu^T h(x)$$

heißt ***Lagrange-Funktion****. Offensichtlich ist die Bedingung (a) in Satz 3.28 äquivalent zu*

$$\nabla_x L(\hat{x}, \hat{\lambda}, \hat{\mu}) = 0.$$

Bemerkung 3.30
Zur Veranschaulichung der Lagrange-Funktion bemerken wir, dass

$$\sup_{0 \le \lambda \in \mathbb{R}^m, \ \mu \in \mathbb{R}^p} L(x, \lambda, \mu) = \begin{cases} f(x) & \text{für } x \in X, \\ \infty & \text{sonst.} \end{cases}$$

Das restringierte Minimierungsproblem

$$f(x) \to \min! \quad \text{u.d.N. } x \in X$$

ist also äquivalent zur Lösung von

$$\min_{x \in \mathbb{R}^n} \sup_{0 \le \lambda \in \mathbb{R}^m, \ \mu \in \mathbb{R}^p} L(x, \lambda, \mu).$$

Dies können wir geometrisch als Sattelpunktproblem interpretieren (vgl. die in der Vorlesung gemalten Skizzen).

3.2.3 Optimalitätsbedingungen zweiter Ordnung

Die KKT-Bedingungen (a)-(c) aus Satz 3.28

$$\nabla_x L(\hat{x}, \hat{\lambda}, \hat{\mu}) = 0, \quad h(\hat{x}) = 0, \quad \hat{\lambda} \ge 0, \quad g(\hat{x}) \le 0, \quad \hat{\lambda}^T g(\hat{x}) = 0 \qquad (3.6)$$

bilden (analog zu $\nabla f(\hat{x}) = 0$ im unrestringierten Fall) eine notwendige Optimalitätsbedingung erster Ordnung. Analog zur positiven Definitheit von $\nabla^2 f(\hat{x})$ im unrestringierten Fall zeigt der folgende Satz eine *hinreichende Optimalitätsbedingung zweiter Ordnung* für restringierte Optimierungsprobleme.

Satz 3.31
Seien $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ und $h : \mathbb{R}^n \to \mathbb{R}^p$ zweimal stetig differenzierbar. Erfüllt $\hat{x} \in \mathbb{R}^n$ die KKT-Bedingungen (3.6) mit Lagrange-Multiplikatoren $\hat{\lambda} \in \mathbb{R}^m$ und $\hat{\mu} \in \mathbb{R}^p$ und ist

$$d^T \nabla^2_{xx} L(\hat{x}, \hat{\lambda}, \hat{\mu}) d > 0 \quad \text{für alle } 0 \neq d \in T_l(g, h, \hat{x}), \tag{3.7}$$

dann ist $\hat{x}$ eine isolierte lokale Lösung von

$$f(x) \to \min! \quad \text{u.d.N. } g(x) \leq 0,\ h(x) = 0. \tag{3.8}$$

Beweis: Angenommen $\hat{x} \in \mathbb{R}^n$, $\hat{\lambda} \in \mathbb{R}^m$ und $\hat{\mu} \in \mathbb{R}^p$ erfüllen (3.6) und (3.7), aber $\hat{x}$ ist kein isoliertes lokales Minimum des restringierten Optimierungsproblems (3.8). Dann existiert in jeder Umgebung von $\hat{x}$ ein zulässiger Punkt mit mindestens genauso geringem Zielfunktionswert, also eine Folge $(x_k)_{k\in\mathbb{N}} \subseteq \mathbb{R}^n$ mit $x_k \to \hat{x}$ und

$$f(x_k) \leq f(\hat{x}), \quad x_k \neq \hat{x}, \quad g(x_k) \leq 0, \quad h(x_k) = 0 \quad \forall k \in \mathbb{N}. \tag{3.9}$$

Setze $d_k := x_k - \hat{x}$. Die normierte Folge $\frac{d_k}{\|d_k\|}$ besitzt einen Häufungspunkt $0 \neq d \in \mathbb{R}^n$. Nach Übergang auf eine Teilfolge können wir o.B.d.A. annehmen, dass $\frac{d_k}{\|d_k\|} \to d$. Offenbar gilt nach Konstruktion, dass

$$0 \neq d \in T(X, \hat{x}) \subseteq T_l(g, h, \hat{x})$$

(siehe Definition 3.13 und Lemma 3.17).

Aus den Eigenschaften (3.9) der Folge $(x_k)_{k\in\mathbb{N}}$ und den KKT-Bedingungen (3.6) folgt außerdem, dass

$$\begin{aligned} L(x_k, \hat{\lambda}, \hat{\mu}) &= f(x_k) + \hat{\lambda}^T g(x_k) + \hat{\mu}^T h(x_x) \\ &\leq f(\hat{x}) = f(\hat{x}) + \hat{\lambda}^T g(\hat{x}) + \hat{\mu}^T h(\hat{x}) = L(\hat{x}, \hat{\lambda}, \hat{\mu}). \end{aligned}$$

Aus der Taylor-Formel aus Satz 2.3 und $\nabla_x L(\hat{x}, \hat{\lambda}, \hat{\mu}) = 0$ erhalten wir also

$$\begin{aligned} 0 &\geq \frac{L(x_k, \hat{\lambda}, \hat{\mu}) - L(\hat{x}, \hat{\lambda}, \hat{\mu})}{\|d_k\|^2} \\ &= \frac{1}{2} \frac{d_k^T \nabla^2_{xx} L(\hat{x}, \hat{\lambda}, \hat{\mu}) d_k}{\|d_k\|^2} + \frac{o(\|d_k\|^2)}{\|d_k\|^2} \to \frac{1}{2} d^T L(\hat{x}, \hat{\lambda}, \hat{\mu}) d \end{aligned}$$

und damit

$$d^T \nabla^2_{xx} L(\hat{x}, \hat{\lambda}, \hat{\mu}) d \leq 0, \tag{3.10}$$

was im Widerspruch zu (3.7) steht. □

Eine Verschärfung von Satz 3.31 beweisen wir in Übungsaufgabe 14.4.

3.2.4 Sequential Quadratic Programming

Das Newton-Verfahren für unrestringierte Optimierungsprobleme erhielten wir durch Anwendung des Newton-Verfahrens für Gleichungssysteme auf die notwendige Optimalitätsbedingung 1. Ordnung, wobei wir jeden Schritt gemäß Bemerkung 2.52 auch als Minimierung der quadratischen Taylor-Näherung in der aktuellen Iterierten interpretieren konnten. Analog versuchen wir nun restringierte Optimierungsprobleme zu behandeln.

Das Lagrange-Newton-Verfahren. Wir betrachten zunächst lediglich gleichungsrestringierte Problem der Form

$$f(x) \to \min! \quad \text{u.d.N. } h(x) = 0,$$

wobei $f: \mathbb{R}^n \to \mathbb{R}$ und $h: \mathbb{R}^n \to \mathbb{R}^p$ in diesem Abschnitt als zweimal stetig differenzierbar vorausgesetzt werden.

In jedem lokalen Minimum $\hat{x}$, in dem eine (CQ) gilt, gelten die KKT-Bedingungen aus Satz 3.28, d.h.

$$\exists \hat{\mu} \in \mathbb{R}^p: \ \nabla_x L(\hat{x}, \hat{\mu}) = 0, \quad h(\hat{x}) = 0,$$

mit $L(x, \mu) = f(x) + \mu^T h(x)$, $\nabla_x L(x, \mu) = \nabla f(x) + \nabla h(x)\mu$.

Zur Bestimmung eines lokalen Minimums benutzen wir daher das Newton-Verfahren zur Bestimmung einer Lösung $(\hat{x}, \hat{\mu}) \in \mathbb{R}^n \times \mathbb{R}^p$ des Gleichungssystems

$$F(x, \mu) := \begin{pmatrix} \nabla_x L(x, \mu) \\ h(x) \end{pmatrix} = 0.$$

F ist stetig differenzierbar mit (vgl. Übungsaufgabe 14.3)

$$F'(x, \mu) = \begin{pmatrix} \nabla^2_{xx} L(x, \mu) & \nabla h(x) \\ \nabla h(x)^T & 0 \end{pmatrix} \in \mathbb{R}^{(n+p)\times(n+p)}$$

Wir fassen die Implementierung der Newton-Iterationsvorschrift

$$x^{k+1} := x^k - F'(x^k)^{-1} F(x^k)$$

in Algorithmus 12 zusammen. Entsprechend der Anwendung des Newton-Verfahrens auf die Lagrange-Gleichungen heißt das Verfahren **Lagrange-Newton-Verfahren**.

Algorithm 12 Lagrange-Newton-Verfahren

Gegeben: Startwert $x_0 \in \mathbb{R}^n$ und $\mu^0 \in \mathbb{R}^p$
for $k = 0, 1, 2, \ldots$ **do**
  if $h(x^k) = 0$ und $\nabla_x L(x^k, \mu^k) = 0$ **then**
    STOP
  else

$$\begin{pmatrix} d_x^k \\ d_\mu^k \end{pmatrix} := - \begin{pmatrix} \nabla_{xx}^2 L(x^k, \mu^k) & \nabla h(x^k) \\ \nabla h(x^k)^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} \nabla_x L(x^k, \mu^k) \\ h(x^k) \end{pmatrix}$$

    $x^{k+1} := x^k + d_x^k$, $\mu^{k+1} := \mu^k + d_\mu^k$
  end if
end for
return $x_0, x_1, x_2, \ldots$

Ist $F'(\hat{x}, \hat{\mu})$ invertierbar (und f und h hinreichend glatt), so folgt die lokale quadratische Konvergenz des Lagrange-Newton-Verfahrens aus der des Newton-Verfahrens für Gleichungssysteme in 2.48. Wir zeigen noch, dass dies erfüllt ist für (im Sinne von Beispiel 3.21(c)) reguläre Punkte, in denen die Optimalitätsbedingung 2. Ordnung aus 3.31 gilt.

Lemma 3.32
Seien $x \in \mathbb{R}^n$ und $\mu \in \mathbb{R}^p$, so dass

(a) die Vektoren $\nabla h_i(x)$, $i = 1, \ldots, p$ linear unabhängig sind

(b) die Matrix $\nabla_{xx}^2 L(x, \mu)$ positiv definit ist auf $\mathcal{N}(h'(x))$, d.h.

$$d^T \nabla_{xx}^2 L(x, \mu) d > 0 \quad \text{für alle } 0 \neq d \in \mathbb{R}^n \text{ mit } \nabla h(x)^T d = 0.$$

Dann ist $F'(x, \mu)$ invertierbar.

Beweis: Da $F'(x, \mu)$ quadratisch ist, genügt es, Injektivität zu zeigen. Sei dazu

$$\begin{pmatrix} v \\ w \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}^p \quad \text{mit } F'(x, \mu) \begin{pmatrix} v \\ w \end{pmatrix} = 0,$$

also

$$\begin{aligned} \nabla_{xx}^2 L(x, \mu) v + \nabla h(x) w &= 0, \\ \nabla h(x)^T v &= 0. \end{aligned}$$

Es folgt dass

$$v^T \nabla_{xx}^2 L(x, \mu) v = v^T \nabla_{xx}^2 L(x, \mu) v + v^T \nabla h(x) w = 0$$

und aus (b) folgt $v = 0$. Damit ist aber auch $\nabla h(x)w = 0$ und mit (a) folgt $w = 0$. $F'(x, \mu)$ ist also injektiv und damit invertierbar. □

Die lokale Konvergenz des Lagrange-Newton-Verfahrens folgt aus der Konvergenz des Newton-Verfahrens für Gleichungssysteme:

Satz 3.33
Seien $f : \mathbb{R}^n \to \mathbb{R}$ und $h : \mathbb{R}^n \to \mathbb{R}^p$ zweimal stetig differenzierbar. Die Minimierungsaufgabe

$$f(x) \to \min! \quad u.d.N.\ h(x) = 0$$

besitze eine lokale Lösung $\hat{x} \in \mathbb{R}^n$ mit zugehörigem Lagrange-Multiplikator $\hat{\mu} \in \mathbb{R}^p$, so dass $\hat{x}$ regulär und die Optimalitätsbedingung 2. Ordnung aus 3.31 erfüllt ist.

Dann gibt es eine Umgebung von $(\hat{x}, \hat{\mu}) \in \mathbb{R}^n \times \mathbb{R}^p$, so dass für jeden Startwert aus dieser Umgebung, der Algorithmus 12 durchführbar ist und entweder nach endlich vielen Schritten an $(\hat{x}, \hat{\mu})$ terminiert, oder superlinear gegen $(\hat{x}, \hat{\mu})$ konvergiert.

Sind $\nabla^2 f$ und $\nabla^2 h_i$ lokal Lipschitz-stetig, so ist die Konvergenzrate quadratisch.

Beweis: Alle Aussagen folgen sofort aus Satz 2.48. Für die quadratische Konvergenzrate sieht man dabei leicht, dass aus der lokalen Lipschitz-Stetigkeit von $\nabla^2 f$ und $\nabla^2 h_i$ die von F' folgt. □

Sequential Quadratic Programming. Das Newton-Verfahren für unrestringierte Optimierungsprobleme erhielten wir gemäß Bemerkung 2.52 auch durch Minimierung der quadratischen Taylor-Näherung in der aktuellen Iterierten.

Mit Satz 3.28 kann man zeigen, dass der Iterationsschritt des Lagrange-Newton-Verfahren äquivalent dazu ist, in der k-ten Iterierten $(x^k, \mu^k) \in \mathbb{R}^n \times \mathbb{R}^d$ das quadratische Problem

$$\nabla f(x^k)^T d + \frac{1}{2} d^T H_k d \to \min! \quad \text{u.d.N. } h(x^k) + \nabla h(x^k)^T d = 0$$

zu lösen und mit der Lösung d und dem zugehörigen Lagrange-Parameter ν die nächste Iterierte zu definieren

$$x^{k+1} := x^k + d, \quad \mu^{k+1} := \nu.$$

Dabei ist $H_k := \nabla^2_{xx} L(x^k, \mu^k)$. Dieses Vorgehen heißt auch *Sequential Quadratic Programming* (SQP).

In dieser Form lässt sich das Verfahren auch auf Ungleichleichungsrestriktionen erweitern. Zur Lösung von Problemen der Form

$$f(x) \to \min! \quad \text{u.d.N. } g(x) \leq 0,\ h(x) = 0,$$

wir im aktuellen Iterationsschritt (x^k, λ^k, μ^k) das quadratische Problem

$$\nabla f(x^k)^T d + \frac{1}{2} d^T H_k d \to \min! \quad \text{u.d.N. } \begin{cases} g(x^k) + \nabla g(x^k)^T d \leq 0, \\ h(x^k) + \nabla h(x^k)^T d = 0. \end{cases}$$

gelöst, wobei $H_k := \nabla^2_{xx} L(x^k, \lambda^k, \mu^k)$. Mit der Lösung $d \in \mathbb{R}^n$ dieses quadratischen Problem und den dazugehörigen Lagrange-Parametern $\tilde{\lambda}$ und $\tilde{\mu}$ definiert man dann die nächste Iterierte durch

$$x^{k+1} := x^k + d, \quad \lambda^{k+1} := \tilde{\lambda}, \quad \mu^{k+1} := \tilde{\mu}.$$

Für Konvergenzaussagen für solche allgemeine SQP-Verfahren und für Verfahren zur Lösung der ungleichungsrestringierten quadratischen Teilprobleme verweisen wir wieder auf [Ulbrich].

3.2.5 Penalty- und Barrier-Verfahren

Zum Abschluss dieses Kapitels über restringierte Optimierung skizzieren wir noch zwei Verfahren, die auf der Idee beruhen, ein restringiertes Optimierungsproblem durch ein unrestringiertes Abzuändern.

Penalty-Verfahren. Penalty-Verfahren ersetzen das restringierte Problem

$$f(x) \to \min! \quad \text{u.d.N. } x \in X$$

durch das unrestringierte Problem

$$f(x) + r(x) \to \min! \quad \text{u.d.N. } x \in \mathbb{R}^n$$

wobei die *Straffunktion* $r(x)$ so gewählt ist, dass $r(x) = 0$ für alle $x \in X$ gilt und $r(x)$ außerhalb X möglichst groß ist.

Das gleichungs- und ungleichungsrestringierte Problem

$$f(x) \to \min! \quad \text{u.d.N. } g(x) \leq 0,\ h(x) = 0,$$

ersetzt man z.B. durch das unrestringierte Problem

$$f_\alpha(x) := f(x) + \frac{\alpha}{2} \sum_{i=1}^{m} \max\{0, g_i(x)\}^2 + \frac{\alpha}{2} \sum_{i=1}^{p} h_i(x)^2 \to \min!$$

mit *Strafparameter* $\alpha > 0$. Typischerweise wird f_α iterativ für eine aufsteigende Folge von $\alpha_k \to \infty$ gelöst wobei als Startwert zur Minimierung von $f_{\alpha_{k+1}}$ das Minimum von f_{α_k} gewählt wird.

Penalty-Verfahren ändern die Zielfunktion erst außerhalb des zulässigen Bereichs ab. Liegt das wahre Minimum des restringierten Problems auf dem Rand, so konvergieren die Iterierten typischerweise von außerhalb des zulässigen Bereichs gegen das Minimum, vgl. die in der Vorlesung gemalten Skizzen.

Barriere-Verfahren. Analog können wir das restringierte Problem auch durch ein unrestringiertes ersetzen, bei dem bereits innerhalb des zulässigen Bereichs eine *Barriere* errichtet wird, die verhindert, dass wir dem Rand zu nahe kommen.

Das ungleichungsrestringierte Problem

$$f(x) \to \min! \quad \text{u.d.N.} \quad g(x) \leq 0$$

können wir gemäß dieser Idee z.B. durch das unrestringierte Problem

$$f_\alpha(x) := f(x) - \alpha \sum_{i=1}^{m} \ln(-g_i(x))$$

ersetzen und für eine absteigende Folge $\alpha_k \to 0$ iterativ f_{α_k} minimieren, wobei wir wie bei den Penalty-Methoden als Startwert für die Minimierung von $f_{\alpha_{k+1}}$ das Minimum von f_{α_k} wählen.

Liegt das wahre Minimum des restringierten Problems auf dem Rand, so konvergieren die Iterierten eines Barriere-Verfahren von innerhalb des zulässigen Bereichs gegen das Minimum, vgl. die in der Vorlesung gemalten Skizzen. Barriere-Verfahren nennt man deshalb auch *Innere-Punkte-Methoden*.

Kapitel 4

Globale Optimierung

Wir beenden die Vorlesung mit einigen elementaren Überlegungen zur globalen Optimierung. Zunächst bemerken wir, wie leicht sich ein konvergenter Algorithmus zur Bestimmung eines globalen Minimums einer stetigen Funktion aufstellen lässt, siehe Algorithmus 13.

Algorithm 13 Triviale globale Optimierung

Gegeben: Stetige Funktion $f : \mathbb{R}^n \to \mathbb{R}$.
Wähle abzählbare dichte Teilmenge $(q_m)_{m\in\mathbb{N}} \subseteq \mathbb{R}^n$
$x_1 := q_1$
for $k := 1, 2, 3, \ldots$ **do**
  if $f(q_{k+1}) < f(x_k)$ **then**
    $x_{k+1} := q_{k+1}$
  else
    $x_{k+1} := x_k$
  end if
end for
return $x_1, x_2, \ldots$

Lemma 4.1
Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig. Die von Algorithmus 13 erzeugte Folge $(x_k)_{k\in\mathbb{N}}$ besitzt die Eigenschaften

(a) $f(x_{k+1}) < f(x_k)$

(b) Jeder Häufungspunkt von x_k ist ein globales Minimum von f.

Insbesondere konvergiert die von Algorithmus 13 erzeugte Folge gegen das globale Minimum von f, falls dieses eindeutig ist und die Niveauflächen von f beschränkt sind.

Beweis: (a) ist trivial. Um (b) zu zeigen sei $(x_k)_{k\in K}$ eine konvergente Teilfolge und $x = \lim_{K\ni k\to\infty} x_k$. Angenommen es existiert ein y mit $f(y) < f(x)$. Dann können wir wegen der Stetigkeit von f ein $\epsilon > 0$ so wählen, dass

$$f(\eta) < f(x) \quad \forall \eta \in B_\epsilon(y).$$

Da $(q_m)_{m\in\mathbb{N}}$ dicht in $\mathbb{R}^n$ liegt, existiert ein $q_m \in B_\epsilon(y)$, also

$$f(q_m) < f(x),$$

Es gilt aber nach Konstruktion

$$f(x_l) \le f(q_m) \quad \text{für alle } l \ge m$$

und damit $f(x) \le \inf_{l\in\mathbb{N}} f(q_l)$. □

Das Durchprobieren einer dichten Teilmenge in Algorithmus 13 ist natürlich überaus aufwändig. Asymptotisch können wir den Algorithmus natürlich versuchen dadurch zu beschleunigen, dass wir die Iterierten zusätzlich als Startwerte einer Newton-Verfahren verwenden. Das fundamentale Problem bleibt aber bestehen, ohne das Durchprobieren einer dichten Teilmenge gibt es keine rigoros konvergenten Algorithmen:

Definition 4.2
*Ein **Optimierungsalgorithmus** $\mathcal{A}$ ist eine Abbildung von $C(\mathbb{R}^n)$ in den Raum der reellen Folgen. Wir sagen, dass ein Optimierungsalgorithmus **auf Funktionsauswertungen basiert**, wenn zu jeder Funktion f eine abzählbare Menge von Auswertungspunkten $(q_m)_{m\in\mathbb{N}}$ existiert, so dass*

$$\mathcal{A}(f) = \mathcal{A}(g) \quad \forall g \in C(\mathbb{R}^n) \text{ mit } g(q_m) = f(q_m)\ \forall m \in \mathbb{N}.$$

Lemma 4.3
Sei $\mathcal{A}$ ein auf Funktionsauswertungen beruhender Algorithmus, der für jede stetige Funktion f eine Folge mit den Eigenschaften (a) und (b) aus Lemma 4.1 erzeugt.

Dann muss $\mathcal{A}$ jede Funktion, für die die von ihm erzeugte Folge einen Häufungspunkt besitzt, auf einer dichten Teilmenge auswerten.

Beweis: Sei f eine Funktion, die von $\mathcal{A}$ auf einer nicht dichten Teilmenge $(q_m)_{m\in\mathbb{N}} \subseteq \mathbb{R}^n$ ausgewertet wird und für die von $\mathcal{A}$ erzeugte Folge $(x_k)_{k\in\mathbb{N}}$ einen Häufungspunkt $\hat{x}$ besitzt.

Dann existieren $y \in \mathbb{R}^n$ und $\epsilon > 0$, so dass

$$\hat{x} \notin B_\epsilon(y) \quad \text{und} \quad q_m \notin B_\epsilon(y)\ \forall m \in \mathbb{N}.$$

Wir ändern jetzt f auf $B_\epsilon(y)$ ab. Dazu sei $\rho : \mathbb{R}^n \to \mathbb{R}$ eine nichtnegative stetige Funktion mit[1]

$$\emptyset \neq \operatorname{supp}(\rho) \subseteq B_\epsilon(y).$$

Dann ist für alle $C > 0$

$$\mathcal{A}(f) = \mathcal{A}(f - C\rho)$$

aber für hinreichend große C ist $\hat{x}$ kein globales Minimum mehr von $f(x) - C\rho(x)$. □

Bemerkung 4.4
Offenbar gilt Lemma 4.3 auch dann noch, wenn wir höhere Differenzierbarkeitsforderungen an die betrachteten Funktionen stellen und der Algorithmus Ableitungen von f auswerten darf.

Analoges gilt natürlich auch in der restringierten Optimierung. Auch hier ist durch Durchprobieren einer dichten Teilmenge des Zulässigkeitsbereiches ein trivialer aber global konvergenter Optimierungsalgorithmus gegeben und ein konvergenter Algorithmus benötigt in jeder offenen Teilmenge des Zulässigkeitsbereichs eine Funktionsauswertung.

Unter Zusatzvoraussetzungen lässt sich die Zahl der benötigten Auswertungen jedoch deutlich reduzieren. Eine realistische Annahme ist, dass eine Schranke für die Ableitungen von f bekannt ist. Damit können für Teilbereiche des betrachteten Zulässigkeitsbereichs untere Schranken an f aufgestellt werden. Liegen diese unteren Schranken schon über dem niedrigsten bekannten Zielfunktionswert, dann kann der Teilbereich in der weiteren Betrachtung ausgeschlossen werden.

Wir beschreiben einen einfachen auf dieser *branch-and-bound*-Strategie beruhender Algorithmus zur globalen Minimierung einer eindimensionalen Funktion $f \in C^1([0,1])$ bei Kenntnis einer Schranke $C \geq \max_{[0,1]} f'(x)$. Dazu werden in baumartiger Verfeinerung die (hier zur besseren Darstellung wie Vektoren geschriebenen) Mengen

$$\underbrace{\begin{pmatrix}1/2\end{pmatrix}}_{=:v_1} \rightsquigarrow \underbrace{\begin{pmatrix}1/4\\3/4\end{pmatrix}}_{=:v_2} \rightsquigarrow \underbrace{\begin{pmatrix}1/8\\3/8\\5/8\\7/8\end{pmatrix}}_{=:v_3}$$

[1] z.B.

$$\rho(x) := h\left(\frac{x-y}{\epsilon}\right), \quad h(x) := \begin{cases} 0 & \text{für } |x| \geq 1, \\ \exp\left(-\frac{1}{1-|x|^2}\right) & \text{für } |x| < 1, \end{cases}$$

erzeugt, wobei beim Übergang von v_k zu v_{k+1} jedes Element x von v_k durch die Elemente $x - h/2$, $x + h/2$ mit $h = 2^{-k}$ ersetzt wird. So würde sich insgesamt eine dichte Teilmenge von $[0, 1]$ ergeben, und wir könnten den trivialen Algorithmus 13 implementieren, indem wir in jedem Schritt das Minimum x_{best} von f über alle bisher erzeugten Werte bilden.

Der Eintrag x in v_k wird mit dieser Vorschrift im weiteren Verlauf eine dichte Teilmenge von $[x - h, x + h]$ erzeugen. Da aber

$$f(\xi) \geq f(x) - Ch \quad \forall \xi \in [x - h, x + h]$$

können wir dieses Teilintervall von der Betrachtung ausschließen, wenn für den besten bisher bekannten Wert x_{best} schon gilt dass $f(x_{\text{best}}) \leq f(x) - Ch$. Wir ignorieren daher solche Werte bei jedem Übergang von v_k auf v_{k+1}, vgl. Algorithmus 14.

Algorithm 14 Ein einfacher branch-and-bound-Algorithmus

Gegeben: Funktion $f \in C^1([0, 1])$.
Gegeben: Schranke $C > \max_{x\in[0,1]} |f'(x)|$.
$x_{\text{best}} = 0.5$. $v_1 := \{0.5\}$.
for $k := 2, 3, \ldots$ **do**
  $h = 2^{-k}$.
  $v_{k+1} := \{x - h/2, x + h/2 : \ x \in v_k : \ f(x_{\text{best}}) > f(x) - Ch\}$
  if $f(x_{\text{best}}) > \min\{f(x) : \ x \in v_{k+1}\}$ **then**
    $x_{\text{best}} := \arg\min\{f(x) : \ x \in v_{k+1}\}$
  end if
end for
return x_{best}

Literaturverzeichnis

[Ulbrich] M. Ulbrich, S. Ulbrich: *Nichtlineare Optimierung*, Birkhäuser Basel 2012.

[GeigerKanzow1] C. Geiger, C. Kanzow: *Numerische Verfahren zur Losung unrestringierter Optimierungsaufgaben*, Springer 1999.

[GeigerKanzow2] C. Geiger, C. Kanzow: *Theorie und Numerik Restringierter Optimierungsaugaben*, Springer 2002.

[Hanke] M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner Verlag, Wiesbaden, 2009.