Vorlesungsskript
Grundlagen der Optimierung

Wintersemester 2021

Roland Herzog*

2022-01-31

*Interdisciplinary Center for Scientific Computing, Heidelberg University, 69120 Heidelberg, Germany (roland.herzog@iwr.uni-heidelberg.de, https://scoop.iwr.uni-heidelberg.de/team/roland-herzog).

Material für 14 Wochen.

Inhaltsverzeichnis

Kapitel 0 Einführung

§ 1 Grundbegriffe und Klassifikation von Optimierungsaufgaben

Die mathematische Optimierung beschäftigt sich mit Aufgaben der Form

$$\left.\begin{array}{rlll} \text{Minimiere} & f(x) & \text{über } x \in \Omega & \text{(\textbf{Zielfunktion})} \\ \text{sodass} & g_i(x) \leq 0 & \text{für } i \in \mathcal{I} & \text{(\textbf{Ungleichungsnebenbedingungen})} \\ \text{und} & h_j(x) = 0 & \text{für } j \in \mathcal{E}. & \text{(\textbf{Gleichungsnebenbedingungen})} \end{array}\right\} \tag{1.1}$$

$\Omega \subseteq \mathbb{R}^n$ heißt die **Grundmenge** und x die **Optimierungsvariable** oder einfach die **Variable** der Aufgabe. Oft sind dabei

- die Funktionen $f, g_i, h_j \colon \mathbb{R}^n \to \mathbb{R}$ hinreichend glatt (C^2-Funktionen),
- $\mathcal{I}$ und $\mathcal{E}$ endliche (evtl. leere) Indexmengen.

Im Fall $\Omega = \mathbb{R}^n$ spricht man von **kontinuierlicher Optimierung**. Im Fall $\Omega = \mathbb{Z}^n$ handelt es sich um **diskrete (ganzzahlige) Optimierungsaufgaben**, die in dieser Lehrveranstaltung nur am Rande behandelt werden.

Definition 1.1 (Grundbegriffe)**.**

(*i*) *Für eine Optimierungsaufgabe* (1.1) *heißt*

$$F := \left\{ x \in \Omega \,\middle|\, g_i(x) \leq 0 \text{ für alle } i \in \mathcal{I},\ h_j(x) = 0 \text{ für alle } j \in \mathcal{E} \right\}$$

die ***zulässige Menge*** *(englisch:* feasible set*). Jedes $x \in F$ heißt* ***zulässiger Punkt****.*

(*ii*) *Die Ungleichung $g_i(x) \leq 0$ heißt an der Stelle x* ***aktiv****, wenn $g_i(x) = 0$ gilt. Sie heißt* ***inaktiv****, wenn $g_i(x) < 0$ ist. Sie heißt* ***verletzt****, wenn $g_i(x) > 0$ ist.*

(*iii*) *Der Wert*

$$f^* := \inf \{ f(x) \mid x \in F \}$$

heißt der ***Optimalwert*** *(englisch:* optimal value*) der Aufgabe* (1.1)*.*

(*iv*) *Im Fall $F = \emptyset$ nennt man die Aufgabe* (1.1) ***unzulässig*** *(englisch:* infeasible*). Es gilt dann $f^* = +\infty$. Im Fall $f^* = -\infty$ heißt das Problem* ***unbeschränkt*** *(englisch:* unbounded*).*

(v) *Ein Punkt $x^* \in F$ heißt ein* ***globaler Minimierer***, ***globale Minimalstelle*** *oder* ***global optimale Lösung***, *wenn gilt:*

$$f(x^*) \leq f(x) \text{ für alle } x \in F.$$

Äquivalent dazu ist: $f(x^*) = f^*$. In diesem Fall heißt die Zahl f^* dann auch das ***globale Minimum*** *oder der* ***globale Minimalwert*** *von* (1.1).

(vi) *Ein globaler Minimierer x^* heißt* ***strikt***, *wenn gilt:*

$$f(x^*) < f(x) \text{ für alle } x \in F,\ x \neq x^*.$$

(vii) *Ein Punkt $x^* \in F$ heißt ein* ***lokaler Minimierer***, ***lokale Minimalstelle*** *oder* ***lokal optimale Lösung***, *wenn es eine Umgebung $U(x^*)$ gibt, sodass gilt:*

$$f(x^*) \leq f(x) \text{ für alle } x \in F \cap U(x^*).$$

In diesem Fall heißt $f(x^*)$ dann auch ein ***lokales Minimum*** *oder ein* ***lokaler Minimalwert*** *von* (1.1).

(viii) *Ein lokaler Minimierer x^* heißt* ***strikt***, *wenn gilt:*

$$f(x^*) < f(x) \text{ für alle } x \in F \cap U(x^*), \quad x \neq x^*.$$

(ix) *Eine Optimierungsaufgabe* (1.1) *heißt* ***lösbar***, *wenn sie mindestens einen globalen Minimierer besitzt, also einen zulässigen Punkt, an dem der Optimalwert angenommen wird. Ansonsten heißt die Aufgabe* ***unlösbar***.

Quizfrage: Welche Eigenschaften haben die in Abbildung 1.1 markierten Punkte?

Quizfrage: Was ist der Unterschied zwischen einem lokalen und einem globalen Minimierer?

Quizfrage: Ist jeder globale Minimierer auch ein lokaler Minimierer? Ist jeder lokale Minimierer auch ein globaler Minimierer?

Quizfrage: Gibt es Optimierungsaufgaben, die einen lokalen Minimierer besitzen, aber keinen globalen?

Beachte: Eine Maximierungsaufgabe „Maximiere $f(x)$ über $x \in F$“ kann durch Übergang zu „Minimiere $-f(x)$ über $x \in F$“ immer in eine Minimierungsaufgabe umgeschrieben werden.

Neben der Frage, welche verschiedenen Klassen von Optimierungsaufgaben es gibt, sind folgende Fragestellungen in der mathematischen Optimierung von Bedeutung:

(1) Wann existieren Optimallösungen?

(2) Wie erkennt man sie? ($\rightsquigarrow$ Optimalitätsbedingungen)

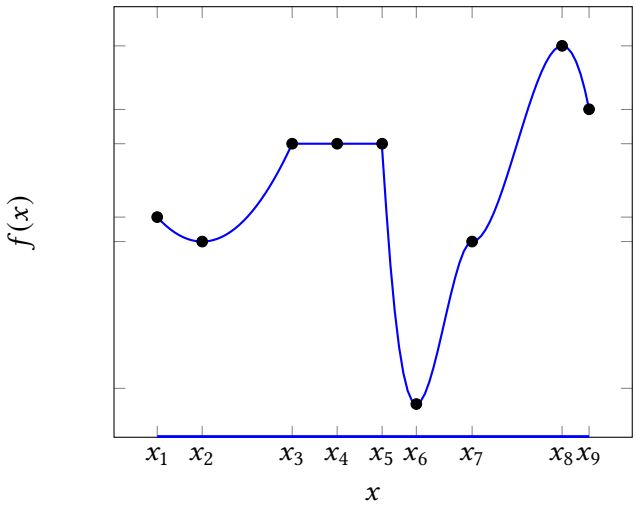


Abbildung 1.1: Illustration der Begriffe aus Definition 1.1 anhand einer Zielfunktion $f : \mathbb{R} \to \mathbb{R}$. Die zulässige Menge ist das auf der x-Achse markierte Intervall.

(3) Wie kann man Lösungen algorithmisch berechnen?

In dieser Lehrveranstaltung werden wir diese Fragen für einige wichtige Typen von Optimierungsaufgaben (1.1) beantworten. Aufgaben der allgemeinen Form (1.1) mit nichtlinearer Zielfunktion und/oder nichtlinearen Nebenbedingungen werden in der Lehrveranstaltung *Nichtlineare Optimierung* behandelt. Später schließen sich Veranstaltungen beispielsweise zu unendlich-dimensionalen Optimierungsaufgaben, insbesondere Aufgaben der Optimierung mit partiellen Differentialgleichungen, an.

Solange nichts anderes gesagt wird, gehen wir ab jetzt immer von $\Omega = \mathbb{R}^n$ aus.

Definition 1.2 (Klassifikation von Optimierungsaufgaben)**.**

(*i*) *Eine Optimierungsaufgabe* (1.1) *heißt* ***frei*** *oder* ***unrestringiert*** *(englisch:* unconstrained*), wenn* $\mathcal{I} = \mathcal{E} = \emptyset$ *ist, andernfalls* ***gleichungs-*** *und/oder* ***ungleichungs-beschränkt*** *oder* ***-restringiert*** *(englisch:* equality constrained, inequality constrained*).*[1]

(*ii*) *Ungleichungsbeschränkungen der besonders einfachen Art*

$$\ell_i \leq x_i \leq u_i, \quad i = 1, \ldots, n$$

mit $\ell_i \in \mathbb{R} \cup \{-\infty\}$ *und* $u_i \in \mathbb{R} \cup \{\infty\}$ *heißen* ***Box-Beschränkungen*** *mit* ***oberer Schranke*** u *und* ***unterer Schranke*** ℓ*.*

(*iii*) *Sind* f*,* g *und* h *(affin-)lineare Funktionen von* x*, so sprechen wir von* ***linearer Optimierung****.*[2] *Eine lineare Optimierungsaufgabe heißt auch* ***lineares Programm*** *(englisch:* linear program*,* ***LP****), also z. B.*

$$\textit{Minimiere} \quad c^\mathsf{T} x \quad \textit{sodass} \quad A x = b \quad \textit{und} \quad x \geq 0.$$

[1] Wir behandeln unrestringierte Aufgaben in Kapitel 1.
[2] Diese werden in Kapitel 2 behandelt.

(*iv*) *Sind allgemeiner f und alle g_i konvexe Funktionen und sind alle h_j wieder (affin-)linear, so sprechen wir von* ***konvexer Optimierung****. Hierbei darf außerdem noch $\Omega \subseteq \mathbb{R}^n$ eine konvexe Teilmenge sein.*[3]

(*v*) *Ist f ein quadratisches Polynom und sind g und h (affin-)linear, so sprechen wir von* ***quadratischer Optimierung****. Eine quadratisches Optimierungsaufgabe heißt auch* ***quadratisches Programm*** *(**QP**).*

(*vi*) *Im allgemeinen Fall spricht man von* ***nichtlinearer Optimierung*** *und von einem* ***nichtlinearen Programm*** *(**NLP**). Nichtlineare Optimierungsaufgaben und zugehörige Lösungsalgorithmen werden in der Lehrveranstaltung* Nichtlineare Optimierung *behandelt.*

Bemerkung 1.3. *Die Grundsteine der linearen Optimierung wurden in den 1940er Jahren von einer Projektgruppe SCOOP (Scientific Computation of Optimum Programs) um George Dantzig (1914–2005) bei der U.S. Air Force gelegt. Im militärischen Sprachgebrauch wurde die Ressourcenplanung als die Erstellung eines Programms bezeichnet, und diese Bezeichnung hat sich erhalten. George Dantzig entwickelte 1947 das Simplex-Verfahren (siehe Kapitel 2). Mehr zur Historie findet man in Gass, Assad, 2005.*

Nicht jede Optimierungsaufgabe ist lösbar. Man kann aber unter recht allgemeinen Annahmen die Existenz eines globalen Minimierers beweisen, wie der folgende Existenzsatz zeigt:

Satz 1.4 (Existenz eines globalen Minimierers)**.**
Die zulässige Menge $F \subseteq \mathbb{R}^n$ sei nichtleer. Weiter sei $f : F \to \mathbb{R}$ ***unterhalbstetig*** *(auch:* ***halbstetig von unten*** *englisch: lower semicontinuous) auf F, d. h.,*

$$\left(x^{(k)}\right) \subseteq F, \quad x^{(k)} \to x^* \in F \quad \Rightarrow \quad \liminf_{k\to\infty} f(x^{(k)}) \geq f(x^*).$$

Für irgendein $m \in \mathbb{R}$ sei die Sub-Levelmenge

$$L := \{x \in F \mid f(x) \leq m\}$$

kompakt in $\mathbb{R}^n$ und nichtleer. Dann besitzt die Aufgabe

$$\textit{Minimiere} \quad f(x) \quad \textit{über} \quad x \in F$$

mindestens einen globalen Minimierer.

Beweis. Wir zeigen zuerst, dass f auf F nach unten beschränkt sein muss. Andernfalls gibt es eine Folge $\left(x^{(k)}\right) \subseteq F$ mit der Eigenschaft $f(x^{(k)}) \leq -k$. Für hinreichend große $k \in \mathbb{N}$ liegen die Glieder dieser Folge in der Menge L. Da aber L kompakt ist, existiert eine konvergente Teilfolge $\left(x^{(k^{(\ell)})}\right) \subseteq L$ mit der Eigenschaft $x^{(k^{(\ell)})} \to x^* \in L$ für $\ell \to \infty$. Aufgrund der Unterhalbstetigkeit von f folgt $f(x^*) \leq \liminf_{\ell\to\infty} f(x^{(k^{(\ell)})}) = -\infty$, Widerspruch. ~~Aufgrund der Unterhalbstetigkeit von f folgt $f(x^*) \leq \liminf_{\ell\to\infty} f(x^{(k^{(\ell)})})$. Demnach wäre $\left(f(x^{(k^{(\ell)})})\right)$ nach unten durch $f(x^*)$ beschränkt, Widerspruch.~~

[3] Diese Aufgaben werden in Kapitel 3 besprochen.

Es sei nun $f^* := \inf_{x \in F} f(x) \in \mathbb{R}$ der Optimalwert. Dann gibt es eine Folge $\left(x^{(k)}\right) \subseteq F$ mit der Eigenschaft[4] $f(x^{(k)}) \searrow f^*$. Für hinreichend große $k \in \mathbb{N}$ gehört die Folge zur Sub-Levelmenge L, und aufgrund der Kompaktheit existiert eine konvergente Teilfolge $x^{(k^{(\ell)})} \to x^*$, deren Grenzwert x^* in L liegt und insbesondere zulässig ist. Wegen der Unterhalbstetigkeit von f gilt $\lim_{\ell \to \infty} f(x^{(k^{(\ell)})}) \geq f(x^*)$, aber auch $\lim_{\ell \to \infty} f(x^{(k^{(\ell)})}) = f^*$. Dies zeigt, dass x^* ein globaler Minimierer ist. □

Bemerkung 1.5.
Wenn f sogar stetig auf F ist, dann folgt die Aussage von Satz 1.4 aus dem Satz von Weierstraß: Stetige reellwertige Funktionen nehmen auf kompakten Mengen ihr Minimum (und Maximum) an.

§ 2 Notation und Wiederholung von Differenzierbarkeitsbegriffen

In diesem Skript verwenden wir farbige Kennzeichnungen für **Definitionen** und Hervorhebungen .

- Die natürlichen Zahlen sind $\mathbb{N} = \{1, 2, \ldots\}$. Wir schreiben $\mathbb{N}_0$ für $\mathbb{N} \cup \{0\}$.
- Wir bezeichnen offene Intervalle mit (a, b) und abgeschlossene Intervalle mit $[a, b]$.
- Matrizen werden üblicherweise mit lateinischen Großbuchstaben bezeichnet, Vektoren mit lateinischen Kleinbuchstaben und Skalare mit griechischen oder lateinischen Kleinbuchstaben. Die Einheitsmatrix wird mit Id bezeichnet. Wir unterscheiden den Vektorraum der Spaltenvektoren $\mathbb{R}^n$ vom Vektorraum der Zeilenvektoren $\mathbb{R}_n$.
- Unendliche Folgen $\mathbb{N} \to \mathbb{R}^n$ bezeichnen wir mit $\left(x^{(k)}\right)$ und nicht mit (x_k) etc., um einen Konflikt mit der Bezeichnung der Komponenten eines Vektors $x = (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n$ zu vermeiden. Endlich viele Vektoren werden dennoch auch mit x_1, x_2 etc. bezeichnet.
- Die durch die streng monoton wachsende Folge $\mathbb{N} \ni \ell \mapsto k^{(\ell)} \in \mathbb{N}$ gebildete **Teilfolge** einer Folge $\left(x^{(k)}\right)$ wird mit $\left(x^{(k^{(\ell)})}\right)$ bezeichnet.
- Für Vektoren $x, y \in \mathbb{R}^n$ bezeichnet $x^\mathsf{T} y$ das Euklidische Skalarprodukt (Innenprodukt) und $\|x\|$ die euklidische Norm:
$$\|x\| = \sqrt{x^\mathsf{T} x}.$$
Wir schreiben also nicht $\langle x\,, y \rangle$ oder $x \cdot y$ für das Euklidische Skalarprodukt.
- Ist $M \in \mathbb{R}^{n \times n}$ eine symmetrische, positiv definite Matrix, so erzeugt sie ein Skalarprodukt $x^\mathsf{T} M y$ und eine Norm $\|x\|_M = \sqrt{x^\mathsf{T} M x}$ auf $\mathbb{R}^n$. Es gilt $\|x\| = \|x\|_{\mathrm{Id}}$.
- Für $\varepsilon > 0$ und $x^* \in \mathbb{R}^n$ ist
$$B_\varepsilon(x^*) := \{x \in \mathbb{R}^n \mid \|x - x^*\| < \varepsilon\}$$

[4] Für eine reelle Zahlenfolge $\left(y^{(k)}\right)$ bedeutet $y^{(k)} \searrow y$, dass $y^{(k)} \geq y$ gilt und $y^{(k)} \to y$. Die Monotonie der Folge ist damit nicht gemeint.

die **offene ε-Umgebung** von x^* oder auch die **offene ε-Kugel** um x^*. Die **abgeschlossene ε-Umgebung** von x^* oder auch die **abgeschlossene ε-Kugel** notieren wir als

$$\overline{B_\varepsilon(x^*)} := \{x \in \mathbb{R}^n \mid \|x - x^*\| \le \varepsilon\}.$$

- Das **Innere** einer Menge $M \subseteq \mathbb{R}^n$ bezeichnen wir mit $\operatorname{int} M$ und den **Abschluss** mit $\overline{M}$.

- Für eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ und gegebenes $x \in \mathbb{R}^n$ heißt die Ableitung der partiellen Funktion $t \mapsto f(x + t\,e_i)$ an der Stelle $t = 0$ die i-te **partielle Ableitung** von f an der Stelle x, kurz: $\frac{\partial}{\partial x_i} f(x)$. Dabei ist $e_i = (0, \dots, 0, 1, 0, \dots, 0)^\mathsf{T}$ einer der Vektoren der Standardbasis von $\mathbb{R}^n$. Mit anderen Worten:
$$\frac{\partial}{\partial x_i} f(x) = \lim_{t \to 0} \frac{f(x + t\,e_i) - f(x)}{t}.$$

- Allgemeiner heißt die Ableitung der Funktion $t \mapsto f(x + t\,d)$ an der Stelle $t = 0$ die **(beidseitige) Richtungsableitung** von f an der Stelle x in Richtung $d \neq 0$, kurz:
$$\frac{\partial}{\partial d} f(x) = \lim_{t \to 0} \frac{f(x + t\,d) - f(x)}{t}.$$

- Die rechtsseitige Ableitung der Funktion $t \mapsto f(x + t\,d)$ an der Stelle $t = 0$ heißt die **(einseitige) Richtungsableitung** von f an der Stelle x in Richtung $d \neq 0$, kurz:
$$f'(x; d) = \lim_{t \searrow 0} \frac{f(x + t\,d) - f(x)}{t}.$$

- Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt **differenzierbar** (kurz: **diffbar**) an der Stelle $x \in \mathbb{R}^n$, falls ein Vektor $v \in \mathbb{R}_n$ (Zeilenvektor) existiert, sodass gilt:
$$\frac{f(x + d) - f(x) - v\,d}{\|d\|} \to 0 \quad \text{für } d \to 0.$$
Der Vektor v heißt in dem Fall die **(totale) Ableitung** von f an der Stelle x und wird mit $f'(x)$ bezeichnet.

- Für diffbare Funktionen $f\colon \mathbb{R}^n \to \mathbb{R}$ gilt
$$f'(x) = \left(\frac{\partial f(x)}{\partial x_1}, \quad \dots, \quad \frac{\partial f(x)}{\partial x_n}\right) \in \mathbb{R}_n.$$
Den transponierten Vektor (Spaltenvektor)
$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = f'(x)^\mathsf{T} \in \mathbb{R}^n$$
bezeichnen wir als den **Gradienten** bzgl. des Euklidischen Skalarprodukts von f an der Stelle x.

- Für diffbare Funktionen gilt:

$$f'(x;d) = \frac{\partial}{\partial d} f(x) = f'(x)\, d = \nabla f(x)^\mathsf{T} d.$$

- Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt **stetig partiell diffbar** oder kurz: $C^1(\mathbb{R}^n, \mathbb{R})$, wenn alle partiellen Ableitungen $\frac{\partial f(x)}{\partial x_i}$ als Funktionen von der Stelle x stetig sind. C^1-Funktionen sind überall diffbar.

- Die Matrix

$$f''(x) = \left(\frac{\partial^2 f(x)}{\partial x_i\, \partial x_j}\right)_{i,j=1}^{n} = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}$$

  bestehend aus den zweiten partiellen Ableitungen der Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ an der Stelle x heißt die **Hessematrix**.

- Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt **zweimal stetig partiell differenzierbar** oder kurz: $C^2(\mathbb{R}^n, \mathbb{R})$, wenn alle Einträge in $f''(x)$ als Funktionen von der Stelle x stetig sind. In diesem Fall ist $f''(x)$ nach dem Satz von Schwarz symmetrisch.

Schließlich benötigen wir häufig den Satz von Taylor:

Satz 2.1 (Taylor, siehe Geiger, Kanzow, 1999, Satz A.2 oder auch Heuser, 2002, Satz 168.1)**.**
Es sei $G \subseteq \mathbb{R}^n$ offen, $k \in \mathbb{N}_0$ und $f\colon G \to \mathbb{R}$ $(k+1)$-mal stetig partiell diffbar, kurz: $C^{k+1}(G, \mathbb{R})$. Falls x_0 und $x_0 + d$ und die gesamte Verbindungsstrecke in G liegen, dann existiert $\xi \in (0,1)$, sodass gilt:

$$\begin{aligned} &\text{im Fall } k = 0: \quad f(x_0 + d) = f(x_0) + f'(x_0 + \xi\, d)\, d \quad (\textbf{\textit{Mittelwertsatz}}), \\ &\text{im Fall } k = 1: \quad f(x_0 + d) = f(x_0) + f'(x_0)\, d + \frac{1}{2} d^\mathsf{T} f''(x_0 + \xi\, d)\, d. \end{aligned}$$

Für vektorwertige Funktionen $F\colon \mathbb{R}^n \to \mathbb{R}^m$ heißt die Matrix der partiellen Ableitungen aller Komponentenfunktionen $F_1, \ldots, F_m$, also

$$\begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \cdots & \frac{\partial F_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_m(x)}{\partial x_1} & \cdots & \frac{\partial F_m(x)}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m \times n},$$

die **Jacobimatrix** von F an der Stelle x. F heißt **diffbar**, wenn alle Komponentenfunktionen diffbar sind. F heißt **stetig partiell diffbar**, wenn alle Einträge der Jacobimatrix als Funktionen von der Stelle x stetig sind. C^1-Funktionen sind überall diffbar.

Kapitel 1 Unrestringierte Optimierung

Wir betrachten in diesem Kapitel das unrestringierte (freie) Optimierungsproblem (1.1) mit $\Omega = \mathbb{R}^n$ und $\mathcal{I} = \mathcal{E} = \emptyset$, also

$$\text{Minimiere} \quad f(x) \text{ über } x \in \mathbb{R}^n.$$

Wir beschränken uns auf das Auffinden *lokaler* Minimalstellen. Globale Minimierer zu bestimmen ist sehr schwierig und nur unter zusätzlichen Voraussetzungen an die Funktion f überhaupt algorithmisch möglich.

Im gesamten Kapitel 1 sei $f\colon \mathbb{R}^n \to \mathbb{R}$ mindestens einmal stetig partiell diffbar, kurz: C^1. Es gilt also für die (beidseitige) Richtungsableitung

$$\frac{\partial}{\partial d} f(x) := \lim_{t \to 0} \frac{f(x + t\,d) - f(x)}{t} = f'(x)\,d.$$

§ 3 Optimalitätsbedingungen der unrestringierten Optimierung

Literatur: Geiger, Kanzow, 1999, Kapitel 2

Satz 3.1 (Notwendige Bedingung 1. Ordnung)**.**
Es sei x^* ein lokaler Minimierer, und f sei C^1 in einer Umgebung $U(x^*)$. Dann ist die Ableitung $f'(x^*) = 0$.

Beweis. Es sei $d \in \mathbb{R}^n$ beliebig. Dann gilt für hinreichend kleine $t > 0$ aufgrund der lokalen Optimalität von x^* die Ungleichung $f(x^*) \le f(x^* + t\,d)$. Nach dem Satz von Taylor 2.1 existiert weiter für jedes hinreichend kleine $t > 0$ (sodass $x^* + t\,d$ in $U(x^*)$ bleibt) jeweils ein $\xi_t \in (0, 1)$, sodass gilt:

$$f(x^* + t\,d) = f(x^*) + f'(x^* + \xi_t\,t\,d)\,(t\,d).$$

Aus beiden Aussagen zusammen folgt:

$$0 \le \frac{f(x^* + t\,d) - f(x^*)}{t} = \frac{1}{t} f'(x^* + \xi_t\,t\,d)\,(t\,d) = f'(x^* + \xi_t\,t\,d)\,d$$

für hinreichend kleine $t > 0$. Der Grenzübergang[1] $t \searrow 0$ liefert wegen der C^1-Eigenschaft von f nun $f'(x^*)\,d \ge 0$. Analog erhält man unter Verwendung von $-d$ die Folgerung $f'(x^*)\,d \le 0$, zusammen also $f'(x^*)\,d = 0$ für alle $d \in \mathbb{R}^n$. □

[1] Wie üblich bedeutet dies: $t > 0$ und $t \to 0$, nicht notwendigerweise monoton.

Ein Punkt $x \in \mathbb{R}^n$ mit der Eigenschaft $f'(x) = 0$ heißt **stationärer Punkt** von f.

Quizfrage: Wie kann man sich die Eigenschaft „$f'(x) = 0$“ beispielsweise für eine Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ vorstellen?

Beachte: Die Bedingung „$f'(x) = 0$“ ist keinesfalls hinreichend dafür, dass x ein lokaler Minimierer von f ist. Mit Hilfe von Bedingungen 2. Ordnung kann man stationäre Punkte genauer unterscheiden.

Satz 3.2 (Notwendige Bedingung 2. Ordnung)**.**
Es sei x^* ein lokaler Minimierer, und f sei C^2 in einer Umgebung $U(x^*)$. Dann ist die Hessematrix $f''(x^*)$ positiv semidefinit.[2]

Beweis. Wir nehmen an, dass $f''(x^*)$ *nicht* positiv semidefinit ist. Dann existiert $d \in \mathbb{R}^n$ mit

$$d^\mathsf{T} f''(x^*)\, d < 0.$$

Nach dem Satz von Taylor 2.1 existiert für alle hinreichend kleinen $t > 0$ (sodass $x^* + t\,d$ in $U(x^*)$ bleibt) jeweils ein $\xi_t \in (0,1)$, sodass gilt:

$$f(x^* + t\,d) = f(x^*) + t \underbrace{f'(x^*)\,d}_{=0 \text{ nach Satz 3.1}} + \frac{1}{2}\, t^2 d^\mathsf{T} f''(x^* + \xi_t\, t\,d)\, d.$$

Der Term $d^\mathsf{T} f''(x)\, d$ hängt nach Voraussetzung in der Umgebung $U(x^*)$ stetig vom Punkt x ab. Nach Annahme ist also $d^\mathsf{T} f''(x)\, d < 0$ für alle x hinreichend nahe bei x^*. Folglich gibt es ein $t_0 > 0$, sodass $d^\mathsf{T} f''(x^* + \xi_t\, t\,d)\, d < 0$ für alle $t \in (0, t_0)$ gilt. Daraus folgt

$$f(x^* + t\,d) < f(x^*) \quad \text{für alle } t \in (0, t_0),$$

im Widerspruch zur Voraussetzung, dass x^* ein lokaler Minimierer von f ist. □

Quizfrage: Wie kann man sich die Eigenschaft „$f''(x)$ ist positiv semidefinit“ beispielsweise für eine Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ vorstellen?

Quizfrage: Kann man den Satz 3.2 auch konstruktiv, also ohne Widerspruchsbeweis, zeigen?

Beachte: Auch die Bedingungen „$f'(x) = 0$“ und „$f''(x)$ ist positiv semidefinit“ gemeinsam sind nicht hinreichend dafür, dass x ein lokaler Minimierer von f ist.

Satz 3.3 (Hinreichende Bedingung 2. Ordnung)**.**
Es sei f eine C^2-Funktion in einer Umgebung $U(x^*)$, und es gelte

(*i*) $f'(x^*) = 0$ *und*

(*ii*) $f''(x^*)$ *ist positiv definit.*

[2] Aufgrund der Symmetrie von $f''(x^*)$ ist dies äquivalent dazu, dass alle Eigenwerte von $f''(x^*)$ nicht-negativ sind.

Dann gilt: Zu jedem $\beta \in (0, \mu)$, wobei $\mu > 0$ der kleinste Eigenwert von $f''(x^*)$ ist, gibt es eine Umgebung $U_\beta(x^*)$ von x^* mit der Eigenschaft

$$f(x) \geq f(x^*) + \frac{\beta}{2}\|x - x^*\|^2 \quad \text{für alle } x \in U_\beta(x^*). \tag{3.1}$$

Insbesondere ist x^* ein strikter lokaler Minimierer von f.

Zu der Eigenschaft (3.1) sagt man auch, die Funktion f habe mindestens **quadratisches Wachstum** in der Nähe von x^* bzw. f verhalte sich lokal **stark konvex** (siehe Kapitel 3).

Quizfrage: Wie kann man sich die Eigenschaft (3.1) beispielsweise für eine Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ vorstellen?

Quizfrage: Welche Eigenschaft der Funktion f beschreibt der kleinste Eigenwert μ von $f''(x^*)$?

Beweis. Aus der linearen Algebra ist bekannt, dass die Werte des Rayleigh-Quotienten für die symmetrische Matrix $f''(x^*)$ nach oben bzw. unten durch den größten bzw. den kleinsten Eigenwert beschränkt sind, dass also insbesondere gilt:

$$d^\mathsf{T} f''(x^*)\, d \geq \mu\, \|d\|^2 \quad \text{für alle } d \in \mathbb{R}^n.$$

Nach dem Satz von Taylor 2.1 existiert für jedes $d \in \mathbb{R}^n$ mit $\|d\|$ hinreichend klein (sodass $x^* + d \in U(x^*)$ liegt) jeweils ein $\xi_d \in (0, 1)$, sodass gilt:

$$f(x^* + d) = f(x^*) + \underbrace{f'(x^*)\, d}_{=0 \text{ nach Annahme } (i)} + \frac{1}{2} d^\mathsf{T} f''(x^* + \xi_d\, d)\, d. \tag{3.2}$$

Da die Hessematrix $f''(x)$ und damit auch ihre Eigenwerte nach Voraussetzung stetig von x abhängen (**Quizfrage:** Warum ist das eigentlich so?), gibt es für jedes $\beta \in (0, \mu)$ eine Umgebung $U_\beta(x^*)$, sodass der kleinste Eigenwert von $f''(x)$ für jedes $x \in U_\beta(x^*)$ nicht kleiner als β ist. Wie oben folgt daraus:

$$\frac{1}{2} d^\mathsf{T} f''(x)\, d \geq \frac{\beta}{2}\|d\|^2$$

für alle $x \in U_\beta(x^*) \subseteq U(x^*)$. Für ein solches x und $d := x - x^*$ erhalten wir also aus (3.2):

$$\begin{aligned} f(x) = f(x^* + d) &= f(x^*) + \frac{1}{2} d^\mathsf{T} f''(\underbrace{x^* + \xi_d\, d}_{\in U_\beta(x^*)})\, d \\ &\geq f(x^*) + \frac{\beta}{2}\|d\|^2. \end{aligned}$$

□

Erfüllt f an einem stationären Punkt x^* die notwendige, aber nicht die hinreichende Bedingung 2. Ordnung, so ist keine Aussage über das Vorliegen eines lokalen Minimierers möglich. Es gibt also eine „unentscheidbare Lücke“ zwischen diesen Bedingungen.

Ende der Woche 1

§ 4 Das Gradientenverfahren

Literatur: Geiger, Kanzow, 1999, Kapitel 8

Das Gradientenverfahren ist der einfachste Vertreter in der Klasse der Abstiegsverfahren. Bei Abstiegsverfahren entsteht eine Folge von Iterierten $(x^{(k)}) \subseteq \mathbb{R}^n$. In jeder Iteration werden folgende Schritte ausgeführt:

(1) Bestimmen einer Abstiegsrichtung $d^{(k)}$ für f am aktuellen Punkt $x^{(k)}$.

(2) Bestimmen einer Schrittlänge $t^{(k)} > 0$, sodass $f(x^{(k)} + t^{(k)}\, d^{(k)}) \leq f(x^{(k)})$ gilt.[3]

(3) Aufdatieren durch $x^{(k+1)} := x^{(k)} + t^{(k)}\, d^{(k)}$.

(4) Erhöhen des Iterationszählers $k \rightsquigarrow k+1$.

Definition 4.1 (Abstiegsrichtung).
*Ein Vektor $d \in \mathbb{R}^n$ heißt **Abstiegsrichtung** für f im Punkt $x \in \mathbb{R}^n$, wenn gilt:*

$$f'(x)\, d < 0. \tag{4.1}$$

Der negative Gradient $-\nabla f(x)$ ist die **Richtung des steilsten Abstiegs** von f im Punkt x. Er ist immer eine Abstiegsrichtung, außer in einem stationären Punkt. Wir können (4.1) auch schreiben als $\nabla f(x)^\top d < 0$. Anschaulich bedeutet dies, dass der Winkel zwischen der Richtung d und dem negativen Gradienten $-\nabla f(x)$ kleiner als 90° ist, siehe Abbildung 4.1.

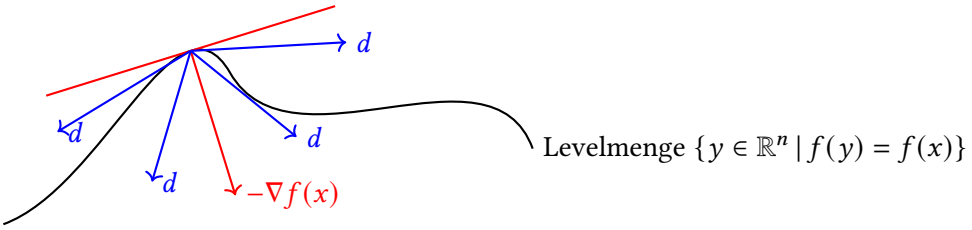


Abbildung 4.1: Verschiedene Abstiegsrichtungen d für f im Punkt x.

Quizfrage: Mit welchem Begriff könnte man die Menge aller Abstiegsrichtungen einer Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ in einem Punkt x geometrisch beschreiben?

§ 4.1 Vorstellung des Verfahrens

Beim (einfachen) **Gradientenverfahren** wird als Abstiegsrichtung $d^{(k)} = -\nabla f(x^{(k)})$ gewählt. Es heißt deshalb auch das **Verfahren des steilsten Abstiegs** (englisch: *steepest descent method*). Es

[3] Der neue Funktionswert ist also geringer oder wenigstens nicht größer als der aktuelle, daher der Name „Abstiegsverfahren".

orientiert sich nur an den Funktionswerten von f, nicht an den Optimalitätsbedingungen aus § 3.

Bei der Wahl der Schrittweiten $t^{(k)}$ verwendet das Verfahren einen Algorithmus zur **Liniensuche**, bei der f entlang einer Richtung d nach einer geeigneten Schrittweite „durchsucht“ wird. Wie das folgende Beispiel zeigt, reicht es dabei nicht aus, dass $(f(x^{(k)}))$ von Iteration zu Iteration streng monoton fällt, um Konvergenz gegen einen Minimierer oder wenigstens gegen einen stationären Punkt zu erzielen:

Beispiel 4.2. *Es seien $f(x) = x^2$, $x^{(0)} = 1$ und $d^{(k)} = -1$ sowie als Schrittweiten $t^{(k)} = \left(\frac{1}{2}\right)^{k+2}$ gewählt. Dann ist die Folge der Iterierten gegeben durch*

$$x^{(k+1)} = x^{(k)} + t^{(k)}\,(-1) = x^{(0)} - \sum_{i=0}^{k} \left(\tfrac{1}{2}\right)^{i+2} = \tfrac{1}{2} + \left(\tfrac{1}{2}\right)^{k+2}.$$

Daraus folgt $x^{(k+1)} < x^{(k)}$ und $f(x^{(k+1)}) < f(x^{(k)})$. Die Folge der Funktionswerte fällt also streng monoton. Jedoch konvergiert $x^{(k)} \searrow x^* = 1/2$, also gegen einen „uninteressanten“ Punkt und nicht gegen den strikten globalen Minimierer von f bei $x = 0$.

Quizfrage: Was ist das „Problem“ mit den in Beispiel 4.2 gewählten Schrittweiten?

Angesichts des Beispiels 4.2 sollten wir uns also fragen, welche Bedingung man an die Schrittweiten stellen sollte, um Konvergenz des Gradientenverfahrens gegen einen stationären Punkt ($f'(x) = 0$) zu erhalten.

Die **exakte Liniensuche**

$$\text{„Bestimme } t^{(k)} := t_{\min} \text{ so, dass } f(x^{(k)} + t_{\min} d^{(k)}) = \min_{t \geq 0} f(x^{(k)} + t\, d^{(k)}) \text{ gilt“} \tag{4.2}$$

ist wegen ihres Aufwands außer in Sonderfällen für besonders einfache Zielfunktionen f nicht praktikabel.

Quizfrage: Welche weitere Schwierigkeit kann sich beim Versuch, die Schrittweite nach (4.2) zu wählen, außerdem noch ergeben?

Daher greift man zu einer besser realisierbaren Schrittweitenstrategie: Zu einer gegebenen Abstiegsrichtung d für die Funktion f im Punkt x bestimmt man eine Schrittweite $t > 0$, sodass die **Armijo-Bedingung**[4] erfüllt ist:

$$f(x + t\,d) \leq f(x) + \sigma\, t\, f'(x)\, d. \tag{4.3}$$

Dabei ist $\sigma \in (0,1)$ der **Armijo-Parameter**. **Quizfrage:** Welche anschauliche Bedeutung hat der Parameter σ?

Zur Veranschaulichung der Bedingung (4.3) führen wir die **Liniensuchfunktion**

$$\varphi(t) := f(x + t\,d) \tag{4.4}$$

[4] Armijo, 1966

zur **Suchrichtung** d ein. Man nennt φ auch den **Schnitt** durch die Funktion f am Punkt x in Richtung d. Die Funktion φ erbt die Differenzierbarkeitseigenschaften von f, ist also auf $\mathbb{R}$ stetig diffbar, und es gilt

$$\varphi'(t) = f'(x + t\,d)\,d.$$

Also lautet die Armijo-Bedingung (4.3) alternativ

$$\varphi(t) \leq \varphi(0) + \sigma\,t\,\varphi'(0). \tag{4.5}$$

Diese Bedingung wird in Abbildung 4.2 illustriert. **Beachte:** Beim Gradientenverfahren gilt $\varphi'(0) = f'(x)\,d = -\|\nabla f(x)\|^2$.

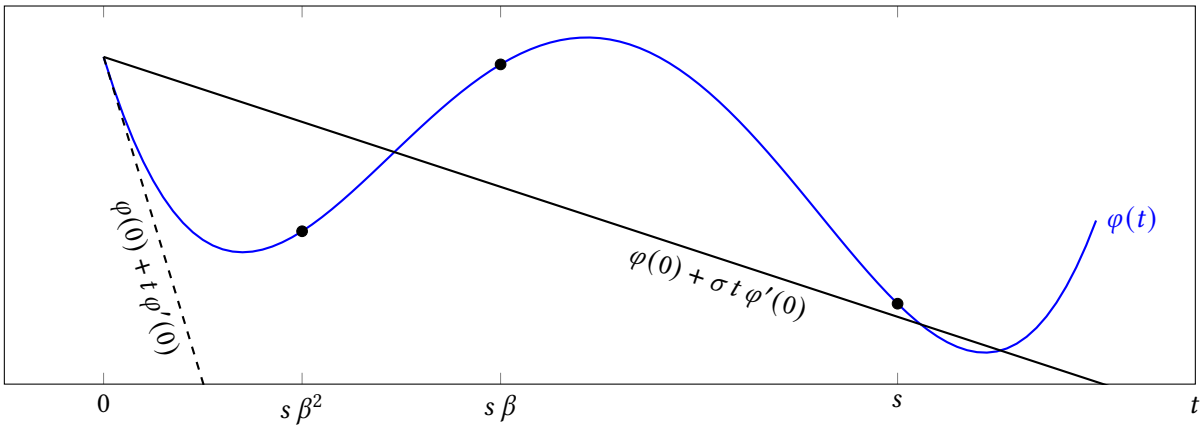


Abbildung 4.2: Darstellung der Armijo-Bedingung (4.5) und einigen Test-Schrittweiten beim Backtracking. Der Armijo-Parameter ist hier als $\sigma = 0.1$ und der Backtracking-Parameter als $\beta = 0.5$ gewählt.

In der praktischen Durchführung wird eine Schrittweite, die (4.5) erfüllt, über eine **Backtracking-Strategie** gefunden: Man beginnt mit einer Startschrittweite $s > 0$ und testet nacheinander die (kleiner werdenden) Schrittweiten $t = s, s\,\beta, s\,\beta^2$ etc., bis zum ersten Mal (4.5) erfüllt ist. Dabei ist $\beta \in (0,1)$ der **Backtracking-Parameter**.

Satz 4.3 (Wohldefiniertheit der Armijo-Backtracking-Strategie)**.**
Es sei $\sigma \in (0,1)$ beliebig. Zu jedem Paar $(x,d) \in \mathbb{R}^n \times \mathbb{R}^n$ mit $f'(x)\,d < 0$ existiert ein $T > 0$, sodass die Armijo-Bedingung (4.5) für alle $t \in [0,T]$ gilt.

Beachte: Aus diesem Satz folgt, dass die Armijo-Backtracking-Strategie wohldefiniert ist, da Schrittweiten der Form $t = s\,\beta^\ell$ für endliches $\ell \in \mathbb{N}_0$ immer im Interval $[0,T]$ landen.

Beweis. Angenommen, die Aussage sei falsch, dann existiert eine Folge $t^{(k)} \searrow 0$ mit der Eigenschaft

$$f(x + t^{(k)} d) > f(x) + \sigma\,t^{(k)} f'(x)\,d$$

für alle $k \in \mathbb{N}$, also auch

$$\frac{f(x + t^{(k)} d) - f(x)}{t^{(k)}} > \sigma\,f'(x)\,d.$$

Im Grenzübergang $k \to \infty$ folgt

$$f'(x)\,d \geq \sigma\, f'(x)\,d,$$

was im Widerspruch zur Voraussetzung $f'(x)\,d < 0$ steht. □

Wir geben nun das Gradientenverfahren mit Armijo-Liniensuche an:

Algorithmus 4.4 (Gradientenverfahren mit Armijo-Schrittweitensuche)**.**
Eingabe: *Startschätzung* $x^{(0)} \in \mathbb{R}^n$
Eingabe: *Armijo-Parameter* $\sigma \in (0,1)$, *Backtracking-Parameter* $\beta \in (0,1)$, *Startschrittweite* $s > 0$

1: *Setze* $k := 0$
2: ***while*** *Abbruchkriterium nicht erfüllt* ***do***
3: *Setze* $d^{(k)} := -\nabla f(x^{(k)})$
4: *Bestimme eine Schrittweite* $t^{(k)} > 0$ *mit der Armijo-Backtracking-Strategie zur Startschrittweite* s, *sodass* (4.3) *erfüllt ist, also:*

$$f(x^{(k)} + t^{(k)}\,d^{(k)}) \leq f(x^{(k)}) + \sigma\, t^{(k)}\, f'(x^{(k)})\,d^{(k)}$$

5: *Setze* $x^{(k+1)} := x^{(k)} + t^{(k)}\,d^{(k)}$
6: *Setze* $k := k+1$
7: ***end while***

Zur Durchführung des Gradientenverfahrens werden folgende problemspezifische Routinen benötigt:

(1) Routine zur Auswertung der Zielfunktion $f(x)$.

(2) Routine zur Auswertung der Ableitung $f'(x)$ bzw. zur Auswertung von Richtungsableitungen $f'(x)\,d$.

Quizfrage: Angenommen, für die Funktion f liegt (neben der Routine für die Auswertung der Funktionswerte) eine Routine vor, die zu einer gegegenen Stelle x und einer gegebenen Richtung d die Richtungsableitung $f'(x)\,d$ bestimmt. Wieso reicht das für die Durchführung von Algorithmus 4.4 aus? Wie bestimmt man insbesondere den negativen Gradienten in Zeile 3?

Für den Beweis eines Konvergenzsatzes für das Gradientenverfahrens benötigen wir folgendes Resultat:

Lemma 4.5 (Konvergenz des Differenzenquotienten bei variabler Stelle und Richtung)**.**
Es seien $x, d \in \mathbb{R}^n$, $(x^{(k)}), (d^{(k)}) \subseteq \mathbb{R}^n$ *mit* $x^{(k)} \to x$ *und* $d^{(k)} \to d$ *sowie* $t^{(k)} \searrow 0$. *Dann gilt*

$$\lim_{k\to\infty} \frac{f(x^{(k)} + t^{(k)}\,d^{(k)}) - f(x^{(k)})}{t^{(k)}} = f'(x)\,d.$$

Beweis. Wegen des Mittelwertsatzes 2.1 existiert zu jedem $k \in \mathbb{N}$ ein $\xi^{(k)} \in (0,1)$ mit

$$\begin{aligned} & f(x^{(k)} + t^{(k)}\, d^{(k)}) - f(x^{(k)}) = t^{(k)} f'(x^{(k)} + \xi^{(k)}\, t^{(k)}\, d^{(k)})\, d^{(k)} \\ \Rightarrow \quad & \lim_{k\to\infty} \frac{f(x^{(k)} + t^{(k)}\, d^{(k)}) - f(x^{(k)})}{t^{(k)}} = \lim_{k\to\infty} f'(\underbrace{x^{(k)} + \xi^{(k)}\, t^{(k)}\, d^{(k)}}_{\to x})\, d^{(k)} = f'(x)\, d. \end{aligned}$$

□

Wir analysieren jetzt Algorithmus 4.4 ohne Abbruchbedingung, sodass eine unendliche Folge $(x^{(k)})$ entsteht. Insbesondere nehmen wir an, dass kein Punkt $x^{(k)}$ stationär ist.

Satz 4.6 (Ein globaler Konvergenzsatz für das Gradientenverfahren)**.**
Jeder Häufungspunkt x^* einer durch Algorithmus 4.4 erzeugten Folge $(x^{(k)})$ ist ein stationärer Punkt von f, erfüllt also $f'(x^*) = 0$.

Beweis. Es sei $x^* \in \mathbb{R}^n$ ein Häufungspunkt von $(x^{(k)})$. Es gibt also eine Teilfolge $(x^{(k^{(\ell)})})$ mit $x^{(k^{(\ell)})} \to x^*$, und wegen der Stetigkeit von f gilt $f(x^{(k^{(\ell)})}) \to f(x^*)$. Da $(f(x^{(k)}))$ aber monoton fällt, konvergiert die gesamte Folge $f(x^{(k)}) \to f(x^*)$. Somit gilt also auch $f(x^{(k+1)}) - f(x^{(k)}) \to 0$.

Angenommen, es sei $f'(x^*) \neq 0$. Aus Zeilen 3 bis 5 des Algorithmus 4.4 folgt

$$\underbrace{f(x^{(k+1)}) - f(x^{(k)})}_{\to 0} \leq \sigma\, t^{(k)}\, f'(x^{(k)})\, d^{(k)} = -\sigma\, t^{(k)}\, \|\nabla f(x^{(k)})\|^2 \leq 0,$$

also

$$t^{(k)}\, \|\nabla f(x^{(k)})\|^2 \to 0.$$

Auf der Teilfolge gilt aber auch $\nabla f(x^{(k^{(\ell)})}) \to \nabla f(x^*) \neq 0$, also muss $t^{(k^{(\ell)})} \to 0$ gelten.

Nötigenfalls durch Einschränkung auf eine weitere Teilfolge (sodass $t^{(k^{(\ell)})} \leq \beta\, s$ gilt, was wegen $t^{(k^{(\ell)})} \to 0$ immer geht) können wir davon ausgehen, dass in der Armijo-Backtracking-Suche die Schrittweite $\beta^{-1}\, t^{(k^{(\ell)})}$ probiert, aber nicht akzeptiert wurde:

$$\begin{aligned} & f(x^{(k^{(\ell)})} + \beta^{-1}\, t^{(k^{(\ell)})}\, d^{(k^{(\ell)})}) > f(x^{(k^{(\ell)})}) + \sigma\, \beta^{-1}\, t^{(k^{(\ell)})}\, \nabla f(x^{(k^{(\ell)})})^{\mathsf{T}} d^{(k^{(\ell)})} \\ \Rightarrow \quad & \frac{f(x^{(k^{(\ell)})} + \beta^{-1}\, t^{(k^{(\ell)})}\, d^{(k^{(\ell)})}) - f(x^{(k^{(\ell)})})}{\beta^{-1}\, t^{(k^{(\ell)})}} > \sigma\, \nabla f(x^{(k^{(\ell)})})^{\mathsf{T}} d^{(k^{(\ell)})} = -\sigma\, \|\nabla f(x^{(k^{(\ell)})})\|^2. \end{aligned}$$

Die Grenzübergänge $x^{(k^{(\ell)})} \to x^*$, $d^{(k^{(\ell)})} = -\nabla f(x^{(k^{(\ell)})}) \to -\nabla f(x^*)$ und $t^{(k^{(\ell)})} \to 0$ für $\ell \to \infty$ ergeben mit Lemma 4.5:

$$-\|\nabla f(x^*)\|^2 \geq -\sigma\, \|\nabla f(x^*)\|^2,$$

was wegen $\sigma \in (0,1)$ zum Widerspruch führt. Es gilt also $\nabla f(x^*) = 0$ und damit $f'(x^*) = 0$. □

Bemerkung 4.7 (Zur praktischen Implementierung des Gradientenverfahrens).
Typische Abbruchkriterien beim Gradientenverfahren[5] *sind:*

(i) $f(x^{(k-1)}) - f(x^{(k)}) \leq ATOL_f + RTOL_f\,|f(x^{(k-1)})|,$

(ii) $\|x^{(k-1)} - x^{(k)}\| \leq ATOL_x + RTOL_x\,\|x^{(k-1)}\|.$

Gefordert werden beide Bedingungen gleichzeitig. Dabei wird oft $RTOL_f = RTOL_x^2$ *gewählt. Als „Notbremsen" dienen zusätzlich die Abfragen*

(iii) $\|\nabla f(x^{(k)})\| \leq ATOL_{\nabla f(x)} + RTOL_{\nabla f(x)}\|\nabla f(x^{(0)})\|,$

(iv) $k \leq k_{\max}$

Als Parameter der Armijo-Liniensuche wählt man z. B. $\sigma = 10^{-2}$ *und* $\beta = 1/2$.

Quizfrage: Welche Bedeutung haben die Bedingungen (i) bis (iii)?

Quizfrage: Wie setzt man ATOL und RTOL, wenn man in Bedingungen (i) bis (iii) entweder nur eine absolute oder nur eine relative Abbruchbedingung verwenden möchte?

Bemerkung 4.8 (Alternative Startschrittweite bei der Armijo-Liniensuche).
In der praktischen Durchführung verwendet man beim Gradientenverfahren oft eine iterationsabhängige Startschrittweite $s^{(k)} > 0$. *Man geht davon aus, dass der durch* $s^{(k)}$ *erreichbare Abstieg im aktuellen Schritt in erster Näherung gleich groß sein wird wie der im letzten Schritt realisierte Abstieg:*

$$\begin{aligned} s^{(k)}\, f'(x^{(k)})\, d^{(k)} &= f(x^{(k)}) - f(x^{(k-1)}) \\ \Rightarrow \quad s^{(k)} &= \frac{f(x^{(k)}) - f(x^{(k-1)})}{f'(x^{(k)})\, d^{(k)}} > 0. \end{aligned} \tag{4.6}$$

Speziell beim Gradientenverfahren ergibt sich dann also

$$s^{(k)} = -\frac{f(x^{(k)}) - f(x^{(k-1)})}{\|\nabla f(x^{(k)})\|^2} \tag{4.7}$$

als Vorschlag für die Startschrittweite ab Iteration $k = 1$. *Ersetzt man auch die rechte Seite in* (4.6) *durch die lineare Näherung* $f(x^{(k)}) - f(x^{(k-1)}) \approx t^{(k-1)} f'(x^{(k-1)})\, d^{(k-1)}$, *so erhalten wir an Stelle von* (4.7) *den Vorschlag*

$$s^{(k)} = t^{(k-1)}\, \frac{\|\nabla f(x^{(k-1)})\|^2}{\|\nabla f(x^{(k)})\|^2} \tag{4.8}$$

für die Startschrittweite.

Auch unter Verwendung dieser Startschrittweiten kann man Satz 4.6 beweisen.

[5] Mehr dazu findet man etwa in Gill, Murray, Wright, 1981. ATOL steht für „absolute Toleranz" und RTOL für „relative Toleranz".

§ 4.2 Das Gradientenverfahren in einem alternativen Skalarprodukt

Bei der Herleitung des Gradientenverfahrens/Verfahrens des steilsten Abstiegs haben wir stillschweigend die Eigenschaft benutzt, dass der Gradient

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$

die Richtung des steilsten Anstiegs und $-\nabla f(x)$ die des steilsten Abstiegs der Funktion $f : \mathbb{R}^n \to \mathbb{R}$ im Punkt x darstellt, die wir als Suchrichtung verwendet haben. Dies ist aber nur dann richtig, wenn der Raum der Optimierungsvariablen $\mathbb{R}^n$ mit dem üblichen (Euklidischen) Skalarprodukt $(x, y) := x^\mathsf{T} y$ ausgestattet ist.

Wir wollen untersuchen, wie sich das Verfahren ändert, wenn man als Skalarprodukt

$$(x, y)_M := x^\mathsf{T} M\, y$$

mit einer symmetrischen, positiv definiten Matrix (s. p. d.) M wählt. Dementsprechend ändert sich auch die Norm zur Längen- und Abstandsmessung in

$$\|x\|_M := \left(x^\mathsf{T} M\, x\right)^{1/2}.$$

Per Definition maximiert die Richtung des steilsten Anstiegs den Ausdruck $f'(x)\, d$ über alle Vektoren $d \in \mathbb{R}^n$ konstanter Länge:

$$\begin{aligned} \text{Maximiere} \quad & f'(x)\, d \quad \text{über } d \in \mathbb{R}^n \\ \text{unter} \quad & \|d\|_M = 1. \end{aligned} \tag{4.9}$$

Die Normierung auf die Länge 1 ist willkürlich gewählt.

Aufgabe (4.9) ist eine restringierte Optimierungsaufgabe, die wir jedoch ohne Kenntnisse der Theorie lösen können: Wir schreiben die Zielfunktion als M-Skalarprodukt um:[6]

$$f'(x)\, d = \nabla f(x)^\mathsf{T} d = \nabla f(x)^\mathsf{T} M^{-1} M\, d = \left(M^{-1} \nabla f(x)\right)^\mathsf{T} M\, d,$$

wobei die Symmetrie $M = M^\mathsf{T}$ benutzt wurde. Die Cauchy-Schwarzsche Ungleichung zeigt, dass dieser Ausdruck genau dann maximal wird, wenn d parallel zu $M^{-1}\nabla f(x)$ liegt. Er wird dagegen minimal, wenn d antiparallel zu $M^{-1}\nabla f(x)$ liegt. Wir fassen zusammen:

Lemma 4.9 (Richtung des steilsten Abstiegs im M-Skalarprodukt)**.**
Die eindeutige Lösung d^* von (4.9) ist, falls $f'(x) \neq 0$ gilt, gegeben durch

$$d^* = M^{-1}\nabla f(x) =: \nabla_M f(x). \tag{4.10}$$

(Die ohnehin willkürliche Normierung $\|d\|_M = 1$ in (4.9) wurde dabei fallengelassen.)

[6] Das heißt, wir bestimmen hier den Riesz-Repräsentanten von $f'(x)$.

Daher ist $d^* = -\nabla_M f(x)$ die **Richtung des steilsten Abstiegs bzgl. des M-Skalarprodukts**. Wir berechnen diese durch Lösung des linearen Gleichungssystems

$$M\, d^* = -\nabla f(x). \tag{4.11}$$

Bei Verwendung des Euklidischen Skalarprodukts ($M = \mathrm{Id}$) schreiben wir weiter $\nabla f(x)$ statt $\nabla_{\mathrm{Id}} f(x)$. Manchmal wird die Verwendung von $\nabla_M f(x)$ an Stelle der Euklidischen Gradientenrichtung $\nabla f(x)$ als **Vorkonditionierung** bezeichnet.

Nach Konstruktion ist für jede beliebige s. p. d. Matrix M die Lösung d^* von (4.11) eine Abstiegsrichtung für f im Punkt x. Dies können wir auch nochmals durch direkte Rechnung bestätigen, vgl. (4.1):

$$f'(x)\, d^* = -\nabla f(x)^\top M^{-1} \nabla f(x) = -\|\nabla f(x)\|_{M^{-1}}^2 = -\|\nabla_M f(x)\|_M^2 < 0, \tag{4.12}$$

falls nicht x bereits ein stationärer Punkt ist.

Algorithmisch ergeben sich durch Verwendung des M-Skalarprodukts an Stelle des Euklidischen Skalarprodukts folgende Änderungen: In Algorithmus 4.4 lautet Zeile 3 nun $d^{(k)} := -\nabla_M f(x^{(k)})$, er wird durch Lösung des linearen Gleichungssystems

$$M\, d^{(k)} = -\nabla f(x^{(k)})$$

ausgeführt. Die übrigen Schritte, insbesondere die Armijo-Bedingung

$$f(x^{(k)} + t^{(k)}\, d^{(k)}) \le f(x^{(k)}) + \sigma\, t^{(k)}\, f'(x^{(k)})\, d^{(k)}$$

bleiben unverändert. Der globale Konvergenz-Satz 4.6 gilt weiter. Als Abbruchbedingung (*ii*) in Bemerkung 4.7 dient nun $\|x^{(k)} - x^{(k-1)}\|_M \le \mathrm{ATOL}_x + \mathrm{RTOL}_x \|x^{(k-1)}\|_M$ und als Bedingung (*iii*) $\|\nabla_M f(x^{(k)})\|_M \le \mathrm{ATOL}_{\nabla f(x)} + \mathrm{RTOL}_{\nabla_M f(x)} \|\nabla f(x^{(0)})\|_M$.

Quizfrage: Warum ändert sich Abbruchbedingung (*i*) nicht?

Als Startschrittweite analog (4.7) bzw. (4.8) wählt man

$$s^{(k)} = -\frac{f(x^{(k)}) - f(x^{(k-1)})}{\|\nabla_M f(x^{(k)})\|_M^2} \quad \text{bzw.} \quad s^{(k)} = t^{(k-1)} \frac{\|\nabla_M f(x^{(k-1)})\|_M^2}{\|\nabla_M f(x^{(k)})\|_M^2}. \tag{4.13}$$

Zur Unterscheidung vom Euklidischen Fall heißt das Verfahren dann auch das **vorkonditionierte Gradientenverfahren**. Wir geben es der Vollständigkeit halber nochmal an:

Algorithmus 4.10 (Vorkonditioniertes Gradientenverfahren mit Armijo-Schrittweitensuche).
Eingabe: *Startschätzung* $x^{(0)} \in \mathbb{R}^n$
Eingabe: *Armijo-Parameter* $\sigma \in (0,1)$, *Backtracking-Parameter* $\beta \in (0,1)$, *Startschrittweite* $s > 0$
Eingabe: *s. p. d. Matrix* $M \in \mathbb{R}^{n\times n}$
1: *Setze* $k := 0$
2: ***while*** *Abbruchkriterium nicht erfüllt* ***do***
3: *Bestimme* $d^{(k)}$ *durch Lösung des linearen Gleichungssystems* $M\, d^{(k)} = -\nabla f(x^{(k)})$

4: *Bestimme eine Schrittweite* $t^{(k)} > 0$ *mit der Armijo-Backtracking-Strategie zur Startschrittweite* s, *sodass* (4.3) *erfüllt ist, also:*

$$f(x^{(k)} + t^{(k)}\, d^{(k)}) \leq f(x^{(k)}) + \sigma\, t^{(k)}\, f'(x^{(k)})\, d^{(k)}$$

5: *Setze* $x^{(k+1)} := x^{(k)} + t^{(k)}\, d^{(k)}$
6: *Setze* $k := k + 1$
7: ***end while***

Beachte: Das Verfahren verallgemeinert das unvorkonditionierte Gradientenverfahren (Algorithmus 4.4), das sich im Fall $M = \mathrm{Id}$ ergibt.

§ 4.3 Konvergenz bei quadratischer Zielfunktion und exakter Liniensuche

Literatur: Geiger, Kanzow, 1999, Kapitel 8.2

Um die Konvergenzgeschwindigkeit des (vorkonditionierten) Gradientenverfahrens zu untersuchen, wenden wir es auf die einfachsten sinnvollen unrestringierten Optimierungsaufgaben an. Bei diesen ist die Zielfunktion ein stark konvexes quadratisches Polynom:

$$f(x) = \frac{1}{2} x^{\mathsf{T}} Q\, x + c^{\mathsf{T}} x + \gamma \tag{4.14}$$

mit einer s. p. d. Matrix $Q \in \mathbb{R}^{n \times n}$, $c \in \mathbb{R}^n$ und $\gamma \in \mathbb{R}$. Der globale Minimierer von f ist eindeutig und charakterisiert durch $f'(x^*) = 0$, also durch das lineare Gleichungssystem

$$Q\, x^* = -c \quad \text{oder äquivalent} \quad x^* = -Q^{-1} c, \tag{4.15}$$

denn dies ist die einzige Lösung der notwendigen Bedingungen (Satz 3.1), und die hinreichenden Bedingungen (Satz 3.3) sind dort erfüllt.

Quizfrage: Welche Rolle spielt die Symmetrie der Matrix Q in (4.14)?

Natürlich wird man das Gradientenverfahren zur Lösung von (4.14) überhaupt nur dann in Erwägung ziehen, wenn

(1) die direkte Lösung des linearen Gleichungssystems (4.15) mit dem Gauss-Verfahren etwa aufgrund der Dimension von Q zu aufwändig ist

(2) oder wenn die Matrix Q nicht explizit vorliegt.

Beachte: Das Gradientenverfahren (Algorithmus 4.10) kommt bereits mit Matrix-Vektor-Produkten $Q\, x$ aus. Diese werden bei der Berechnung des Gradienten $\nabla f(x) = Q\, x + c$ in Zeile 3 benötigt.

Im Fall der quadratischen Zielfunktion lässt sich sogar die exakte Schrittweite (4.2)

$$t_{\min} = \operatorname*{arg\,min}_{t \geq 0} f(x^{(k)} + t\, d^{(k)})$$

im k-ten Schritt berechnen:

$$t^{(k)} := t_{\min} = \frac{(d^{(k)})^\mathsf{T} M\, d^{(k)}}{(d^{(k)})^\mathsf{T} Q\, d^{(k)}}. \tag{4.16}$$

In diesem Abschnitt wählen wir statt der Armijo-Strategie in Algorithmus 4.4 stets die exakte Schrittweite $t_{\min}$.

Im Folgenden seien $\lambda_{\min}(Q;M) > 0$ und $\lambda_{\max}(Q;M) > 0$ der kleinste und größte Eigenwert des **verallgemeinerten Eigenwertproblems**

$$Q\,x = \lambda\, M\, x \quad \text{oder äquivalent} \quad M^{-1}Q\,x = \lambda\, x$$

mit den s. p. d. Matrizen Q und M. Weiter sei

$$\kappa := \operatorname{cond}_2(Q;M) = \frac{\lambda_{\max}(Q;M)}{\lambda_{\min}(Q;M)}$$

die verallgemeinerte (spektrale) **Konditionszahl** von Q bzgl. M.

Satz 4.11 (Globaler Konvergenzsatz für quadratische Zielfunktionen)**.**
Es seien Q und M s. p. d. Matrizen. Das Gradientenverfahren im M-Skalarprodukt (Algorithmus 4.10) mit exakter Schrittweite $t_{\min}$, angewendet zur Minimierung der Zielfunktion (4.14), konvergiert für jeden Startvektor $x^{(0)} \in \mathbb{R}^n$ gegen den eindeutigen globalen Minimierer x^*, und es gelten die Abschätzungen

$$f(x^{(k+1)}) - f(x^*) \le \left(\frac{\kappa-1}{\kappa+1}\right)^2 \big(f(x^{(k)}) - f(x^*)\big) \tag{4.17}$$

und deswegen auch

$$\|x^{(k+1)} - x^*\|_Q \le \left(\frac{\kappa-1}{\kappa+1}\right) \|x^{(k)} - x^*\|_Q, \tag{4.18a}$$

$$\|x^{(k)} - x^*\|_Q \le \left(\frac{\kappa-1}{\kappa+1}\right)^k \|x^{(0)} - x^*\|_Q. \tag{4.18b}$$

Beachte: Damit können wir das Gradientenverfahren auch als ein iteratives Verfahren zur Lösung linearer Gleichungssysteme mit s. p. d. Koeffizientenmatrizen verstehen.

Bemerkung 4.12 (Zum Konvergenzverhaltens des Gradientenverfahrens)**.**

(*i*) *Für große Konditionszahlen κ ist die Konvergenz sehr langsam. Es zeigt sich ein Zick-Zack-Verlauf bei den Iterierten.*

(*ii*) *Im gegenteiligen Extremfall ist $\kappa = 1$, d. h., $M = Q$ (oder ein Vielfaches davon), konvergiert das Gradientenverfahren in einem Schritt: $x^{(1)} = x^*$. Allerdings bedeutet dies, dass bei der Berechnung der Suchrichtung $d^{(0)} = \nabla_M f(x^{(0)})$ ein lineares Gleichungssystem mit $M = Q$ als Koeffizientenmatrix zu lösen ist. Wenn man dies kann, so kann man natürlich auch direkt die Optimalitätsbedingungen $Q\,x^* = -c$ lösen.*

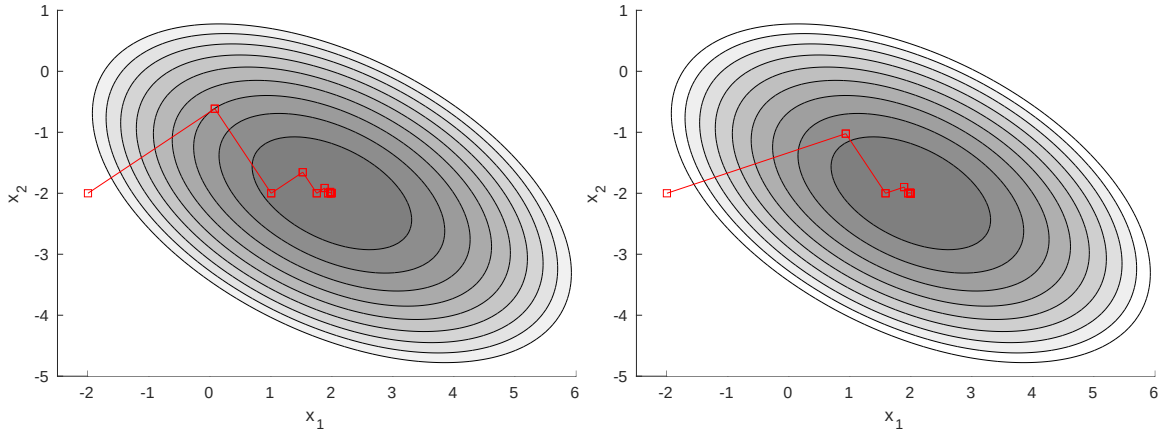


Abbildung 4.3: Illustration des Gradientenverfahrens (Algorithmus 4.10) mit Startpunkt $x^{(0)} = (-2,-2)^\mathsf{T}$ und exakter Schrittweite (4.16) für die Minimierung von (4.14) mit $Q = \left(\begin{smallmatrix} 3 & 2 \\ 2 & 6 \end{smallmatrix}\right)$ und $c = \left(\begin{smallmatrix} -2 \\ 8 \end{smallmatrix}\right)$. Die exakte Lösung ist $x^* = (2,-2)^\mathsf{T}$. Verlauf bei Verwendung des Skalarprodukts $M = \mathrm{Id}$ (links) und $M = \mathrm{diag}(Q)$ (rechts).

(*iii*) *Für allgemeine C^2-Funktionen f ist die Konvergenzgeschwindigkeit in der Nähe eines lokalen Optimums x^*, an dem $f''(x^*)$ s. p. d. ist, wegen*

$$f(x) = f(x^*) + \nabla f(x^*)^\mathsf{T}(x - x^*) + \frac{1}{2}(x - x^*)^\mathsf{T} f''(x^* + \xi\,(x - x^*))(x - x^*)$$

durch die verallgemeinerte Konditionszahl der Hessematrix $f''(x^*)$ bzgl. M bestimmt.

(*iv*) *In der Praxis sucht man einen Kompromiss bei der Wahl von M, sodass die Konditionszahl κ möglichst klein, lineare Gleichungssysteme mit M als Koeffizientenmatrix aber noch leicht zu lösen sind. Manchmal ist bereits die Wahl*

$$M = \mathrm{diag}\big(f''(x^{(0)})\big)$$

konvergenzbeschleunigend.

Das in vielerlei Hinsicht beste Abstiegsverfahren zur Minimierung von (4.14) bzw. zur Lösung linearer Gleichungssysteme (4.15) mit s. p. d. Matrix Q ist das **Verfahren der konjugierten Gradienten** (**CG-Verfahren**), siehe Vorlesung *Nichtlineare Optimierung* oder *Numerische Lineare Algebra*. Beim CG-Verfahren erhält man mit i. W. demselben Aufwand pro Iteration an Stelle von (4.18b) die Konvergenzabschätzung

$$\|x^{(k)} - x^*\|_Q \le 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \|x^{(0)} - x^*\|_Q.$$

Es gibt auch nichtlineare Varianten des CG-Verfahrens für allgemeine Zielfunktionen, siehe Lehrveranstaltung *Nichtlineare Optimierung*.

Ende der Woche 2

§ 5 Das Newton-Verfahren

Wir untersuchen in diesem Abschnitt das Newton-Verfahren zur Lösung der (nichtlinearen) Gleichung $F(x) = 0$. Dabei wird $F\colon \mathbb{R}^n \to \mathbb{R}^n$ im gesamten Abschnitt als stetig partiell diffbar (C^1-Funktion) angenommen. Später wenden wir das Verfahren auf die notwendige Bedingung 1. Ordnung der Aufgabe „Minimiere $f(x)$ über $x \in \mathbb{R}^n$“ an, also zur Lösung von $F(x) = \nabla f(x) = 0$.

Idee: Es sei $x^{(0)}$ die Schätzung einer Nullstelle von F. Wir legen im Punkt $x^{(0)}$ die Tangente (ein **lineares Modell**) an die Funktion und bestimmen *deren* Nullstelle:

$$F(x^{(0)}) + F'(x^{(0)})(x - x^{(0)}) = 0 \quad \Leftrightarrow \quad x = x^{(0)} - F'(x^{(0)})^{-1} F(x^{(0)}).$$

Diese Nullstelle dient als nächste Iterierte usw.

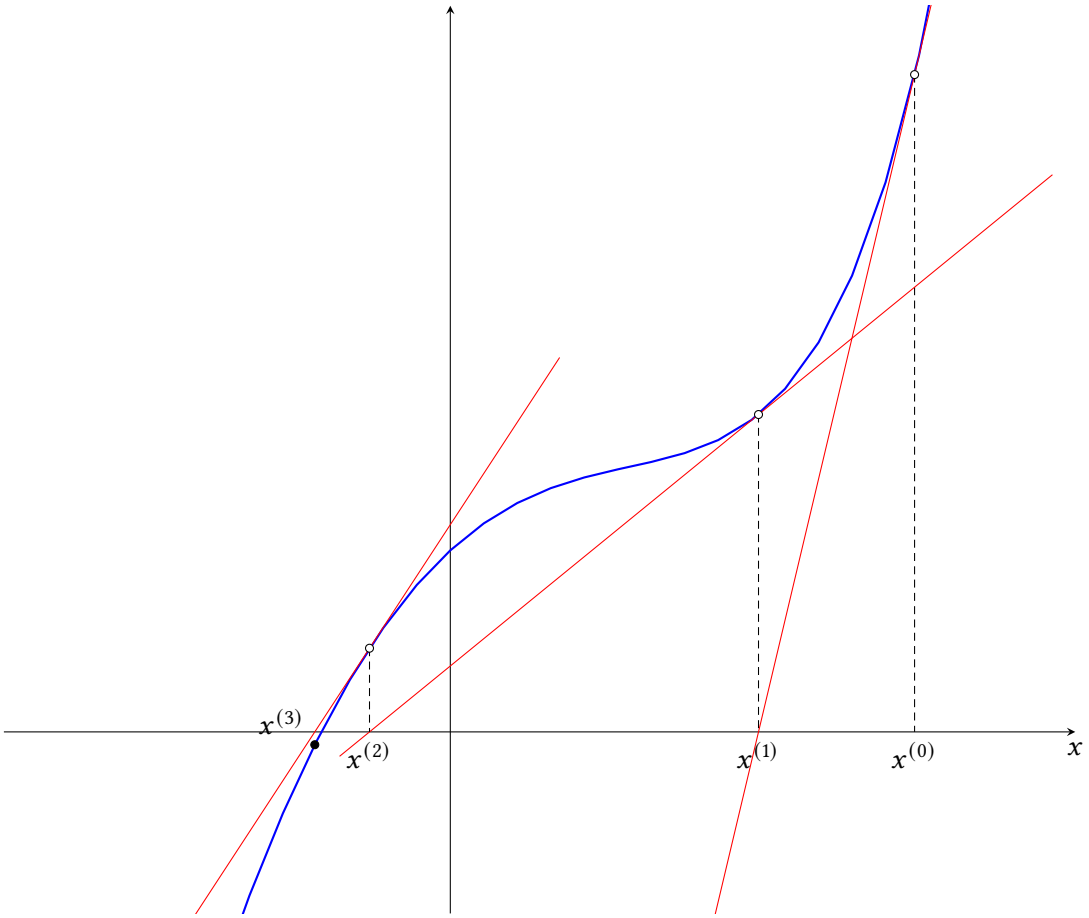


Abbildung 5.1: Illustration des Newton-Verfahrens zur Suche einer Nullstelle der Funktion $F(x) = \exp(0.9\,x) - x^2$.

Der Vektor $F(x^{(k)})$ heißt dabei das **Residuum** zur Iterierten $x^{(k)}$, und $F'(x^{(k)})$ ist die zugehörige

Jacobimatrix:

$$F'(x) = \begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \cdots & \frac{\partial F_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_n(x)}{\partial x_1} & \cdots & \frac{\partial F_n(x)}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Algorithmus 5.1 (Lokales Newton-Verfahren).
Eingabe: *Startschätzung* $x^{(0)} \in \mathbb{R}^n$

1: *Setze* $k := 0$
2: ***while*** *Abbruchkriterium nicht erfüllt* ***do***
3: *Löse das lineare Gleichungssystem* $F'(x^{(k)})\, d^{(k)} := -F(x^{(k)})$ *für die* ***Newton-Richtung*** $d^{(k)}$
4: *Setze* $x^{(k+1)} := x^{(k)} + d^{(k)}$
5: *Setze* $k := k + 1$
6: ***end while***

§ 5.1 Einige Hilfsresultate

Literatur: Geiger, Kanzow, 1999, Kapitel 7, Lemma B.7 und B.8

Definition 5.2 (Matrixnorm).
Es sei $A \in \mathbb{R}^{m\times n}$. *Wir definieren die durch die Euklidischen Normen im* $\mathbb{R}^n$ *und* $\mathbb{R}^m$ *induzierte* ***Matrixnorm***

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$

$\|A\|$ wird auch als **Spektralnorm** von A bezeichnet, und es gilt der Zusammenhang

$$\|A\| = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\top A)}$$

mit dem größten Singulärwert $\sigma_{\max}$ von A und dem größten Eigenwert $\lambda_{\max}$ von $A^\top A$. Weiter gilt $\|Ax\| \leq \|A\|\|x\|$ und $\|AB\| \leq \|A\|\|B\|$ für alle Matrizen A, B und Vektoren x passender Größe.

Lemma 5.3 (Banach-Lemma).

(*i*) *Es sei* $M \in \mathbb{R}^{n\times n}$ *mit* $\|M\| < 1$. *Dann ist* $\mathrm{Id} - M$ *regulär (invertierbar), und es gilt*

$$\|(\mathrm{Id} - M)^{-1}\| \leq \frac{1}{1 - \|M\|}.$$

(*ii*) *Es seien* $A, B \in \mathbb{R}^{n\times n}$ *mit* $\|\mathrm{Id} - BA\| < 1$. *Dann sind* A *und* B *regulär, und es gilt*

$$\|B^{-1}\| \leq \frac{\|A\|}{1 - \|\mathrm{Id} - BA\|} \quad \textit{und} \quad \|A^{-1}\| \leq \frac{\|B\|}{1 - \|\mathrm{Id} - BA\|}.$$

Aussage (i) besagt, Matrizen „in der Nähe“ der Einheitsmatrix invertierbar sind. Aussage (ii) besagt, dass wenn $\mathrm{Id} - BA$ klein ist, also $B \approx A^{-1}$ gilt, notwendig A und B invertierbar sind.

Beweis. Aussage (i): Für $x \in \mathbb{R}^n$ gilt

$$\|(\mathrm{Id} - M)\, x\| = \|x - Mx\| \geq \|x\| - \|Mx\| \geq \underbrace{\big(1 - \|M\|\big)}_{>0} \|x\|.$$

Es folgt $(\mathrm{Id} - M)\, x \neq 0$ für $x \neq 0$, d. h., $\mathrm{Id} - M$ ist injektiv und damit regulär.

Es sei nun $y \in \mathbb{R}^n$ beliebig und $x := (\mathrm{Id} - M)^{-1} y$. Für eine Abschätzung der Norm von $(\mathrm{Id} - M)^{-1}$ müssen wir $\|x\|$ durch $\|y\|$ abschätzen. Die Abschätzung oben zeigt

$$\begin{aligned} & \|y\| \geq (1 - \|M\|)\, \|x\| \\ \Rightarrow \quad & \|(\mathrm{Id} - M)^{-1}\| = \max_{y \neq 0} \frac{\|(\mathrm{Id} - M)^{-1} y\|}{\|y\|} \leq \frac{1}{1 - \|M\|}. \end{aligned}$$

Aussage (ii): Es sei $M = \mathrm{Id} - BA$, also $\|M\| < 1$. Wegen Aussage (i) ist $\mathrm{Id} - M = \mathrm{Id} - (\mathrm{Id} - BA) = BA$ regulär, d. h., A und B sind beide regulär. Weiter gilt

$$\begin{aligned} & (\mathrm{Id} - M)^{-1} = (BA)^{-1} = A^{-1} B^{-1} \\ \Rightarrow \quad & B^{-1} = A(\mathrm{Id} - M)^{-1} \\ \Rightarrow \quad & \|B^{-1}\| \leq \|A\| \|(\mathrm{Id} - M)^{-1}\| \overset{(i)}{\leq} \frac{\|A\|}{1 - \|M\|} = \frac{\|A\|}{1 - \|\mathrm{Id} - BA\|}. \end{aligned}$$

Die andere Ungleichung folgt analog. □

Lemma 5.4. *Es sei F eine C^1-Funktion, $x^* \in \mathbb{R}^n$ und die Jacobimatrix $F'(x^*)$ regulär. Dann existieren eine Umgebung $B_\delta(x^*)$ und eine Konstante $c > 0$, sodass $F'(x)$ für alle $x \in B_\delta(x^*)$ regulär ist, und es gilt:*

$$\|F'(x)^{-1}\| \leq c := 2\, \|F'(x^*)^{-1}\| \quad \textit{für alle } x \in B_\delta(x^*).$$

Beweis. Da F' im Punkt x^* stetig ist, existiert ein $\delta > 0$ mit

$$\|F'(x^*) - F'(x)\| \leq \varepsilon = \frac{1}{2\, \|F'(x^*)^{-1}\|}$$

für alle $x \in B_\delta(x^*)$, also auch

$$\begin{aligned} \|\mathrm{Id} - F'(x^*)^{-1} F'(x)\| &= \|F'(x^*)^{-1} \big(F'(x^*) - F'(x)\big)\| \\ &\leq \|F'(x^*)^{-1}\| \|F'(x^*) - F'(x)\| \\ &\leq 1/2 < 1. \end{aligned}$$

Nach dem Banach-Lemma 5.3, Aussage (ii) [mit $A = F'(x)$ und $B = F'(x^*)^{-1}$] folgt, dass $F'(x)$ für $x \in B_\delta(x^*)$ regulär ist, und es gilt

$$\|F'(x)^{-1}\| \leq \frac{\|F'(x^*)^{-1}\|}{1 - \|\mathrm{Id} - F'(x^*)^{-1} F'(x)\|} \leq 2\, \|F'(x^*)^{-1}\| =: c.$$

□

Bemerkung 5.5 (Einordnung von Lemma 5.4)**.**
Lemma 5.4 korrespondiert zu einem allgemeineren Ergebnis der Funktionalanalysis: Die Menge aller stetig invertierbaren linearen Operatoren zwischen Banachräumen ist offen.

Lemma 5.6. *Es sei F eine C^1-Funktion und $x^* \in \mathbb{R}^n$. Für alle $\varepsilon > 0$ existiert $\delta > 0$ mit*

$$\|F(x) - F(x^*) - F'(x)(x - x^*)\| \leq \varepsilon \|x - x^*\|$$

für alle $\|x - x^*\| \leq \delta$.

Quizfrage: Was würde die Aussage des Satzes bedeuten, wenn an Stelle von x der Punkt x^* stehen würde?

Beweis. Es sei $\varepsilon > 0$ gegeben. Aus der Dreiecksungleichung ergibt sich

$$\begin{aligned}&\|F(x) - F(x^*) - F'(x)(x - x^*)\| \\ &\quad \leq \|F(x) - F(x^*) - F'(x^*)(x - x^*)\| + \|F'(x^*) - F'(x)\|\|x - x^*\|.\end{aligned}$$

Da F nach Voraussetzung in x^* diffbar ist, existiert $\delta_1 > 0$ mit

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

für alle $\|x - x^*\| < \delta_1$. Andererseits ist F' stetig in x^*, sodass $\delta_2 > 0$ existiert mit

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

für alle $\|x - x^*\| < \delta_2$. Mit $\delta := \min\{\delta_1, \delta_2\}$ folgt die Behauptung. □

Zur Charakterisierung der Konvergenzgeschwindigkeit von Algorithmen führen wir folgende Begriffe ein:

Definition 5.7 (Q-Konvergenzraten)**.**
Es sei $(x^{(k)}) \subseteq \mathbb{R}^n$ eine Folge und $x^* \in \mathbb{R}^n$.

(*i*) *$(x^{(k)})$ konvergiert gegen x^* (mindestens)* ***q-linear****, falls ein $c \in (0, 1)$ existiert mit*

$$\|x^{(k+1)} - x^*\| \leq c \|x^{(k)} - x^*\| \quad \text{für alle } k \in \mathbb{N} \text{ hinreichend groß.}$$

(*ii*) *$(x^{(k)})$ konvergiert gegen x^* (mindestens)* ***q-superlinear****, falls es eine Nullfolge $(\varepsilon^{(k)})$ gibt mit*

$$\|x^{(k+1)} - x^*\| \leq \varepsilon^{(k)} \|x^{(k)} - x^*\| \quad \text{für alle } k \in \mathbb{N}.$$

(*iii*) *Es gelte $x^{(k)} \to x^*$. $(x^{(k)})$ konvergiert gegen x^* (mindestens)* ***q-quadratisch****, falls ein $C > 0$ existiert mit*

$$\|x^{(k+1)} - x^*\| \leq C \|x^{(k)} - x^*\|^2 \quad \text{für alle } k \in \mathbb{N}.$$

Abschätzung (4.18a) zeigt beispielsweise die q-lineare Konvergenz des Gradientenverfahrens bei quadratischer Zielfunktion.

Quizfrage: Angenommen, eine Folge konvergiere q-superlinear wie oben definiert. Konvergiert sie dann auch noch q-superlinear, wenn man die in der Definition verwendete Euklidische Norm durch die Norm $\|x\|_M$ mit einer s. p. d. Matrix M austauscht? Wie ist das bei q-quadratischer Konvergenz? Und bei q-linearer Konvergenz?

§ 5.2 Das lokale Newton-Verfahren für $F(x) = 0$

Wir können nun einen lokalen Konvergenzsatz für Algorithmus 5.1 (ohne Abbruchbedingung) beweisen:

Satz 5.8 (Lokaler Konvergenzsatz für Newton-Verfahren)**.**
Es sei F eine C^1-Funktion und $x^* \in \mathbb{R}^n$ ein Punkt mit $F(x^*) = 0$ und $F'(x^*)$ regulär. Dann existiert eine Umgebung $B_\delta(x^*)$ von x^*, sodass für jedes $x^{(0)} \in B_\delta(x^*)$ gilt:

(*i*) *Das lokale Newton-Verfahren ist wohldefiniert und erzeugt eine Folge $(x^{(k)})$, die gegen x^* konvergiert.*

(*ii*) *Die Konvergenzrate ist q-superlinear.*

(*iii*) *Ist F' Lipschitz-stetig in $B_\delta(x^*)$, so ist die Konvergenzrate sogar q-quadratisch.*

Beweis. Aussage (*i*): Nach Lemma 5.4 existieren $\delta_1 > 0$ und $c > 0$, sodass $F'(x)$ für alle $x \in B_{\delta_1}(x^*)$ regulär ist mit

$$\|F'(x)^{-1}\| \le c = 2\,\|F(x^*)^{-1}\|. \tag{5.1}$$

Nach Lemma 5.6 existiert zu $\varepsilon = 1/(2c)$ ein $\delta_2 > 0$ mit

$$\|F(x) - F(x^*) - F'(x)(x - x^*)\| \le \frac{1}{2c}\|x - x^*\| \tag{5.2}$$

für alle $x \in B_{\delta_2}(x^*)$. Setze $\delta := \min\{\delta_1, \delta_2\}$ und wähle $x^{(0)} \in B_\delta(x^*)$. Dann ist der Schritt $x^{(1)} := x^{(0)} - F'(x^{(0)})^{-1}F(x^{(0)})$ wohldefiniert, und es gilt

$$\begin{aligned}
\|x^{(1)} - x^*\| &= \|x^{(0)} - x^* - F'(x^{(0)})^{-1}F(x^{(0)})\| \\
&= \|F'(x^{(0)})^{-1}\big[F'(x^{(0)})(x^{(0)} - x^*) - F(x^{(0)}) + \overbrace{F(x^*)}^{=0}\big]\| \\
&\le \|F'(x^{(0)})^{-1}\|\,\|F(x^{(0)}) - F(x^*) - F'(x^{(0)})(x^{(0)} - x^*)\| \\
&\le c\,\frac{1}{2c}\|x^{(0)} - x^*\| = \frac{1}{2}\|x^{(0)} - x^*\|,
\end{aligned}$$

also liegt auch $x^{(1)}$ wieder in $B_\delta(x^*)$. Per Induktion ist $x^{(k)}$ wohldefiniert, gehört zu $B_\delta(x^*)$, und $x^{(k)} \to x^*$ q-linear.

Aussage *(ii)*: Wir stellen zunächst eine Gleichung für den Fehler auf:[7]

$$\begin{aligned} x^{(k+1)} - x^* &= x^{(k)} - x^* - F'(x^{(k)})^{-1}\big(F(x^{(k)}) - F(x^*)\big) \\ &= F'(x^{(k)})^{-1}\big[F'(x^{(k)})(x^{(k)} - x^*) - \big(F(x^{(k)}) - F(x^*)\big)\big] \\ &= F'(x^{(k)})^{-1}\Big[F'(x^{(k)})(x^{(k)} - x^*) - \int_0^1 F'(x^{(k)} + t\,(x^* - x^{(k)}))(x^{(k)} - x^*)\,\mathrm{d}t\Big] \\ &= F'(x^{(k)})^{-1}\Big[\int_0^1 F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\,\mathrm{d}t\Big](x^{(k)} - x^*). \end{aligned}$$

Daraus erhalten wir folgende wichtige Abschätzung:

$$\|x^{(k+1)} - x^*\| \le \big\|F'(x^{(k)})^{-1}\big\| \int_0^1 \overbrace{\big\|F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\big\|}^{=:D^{(k)}(t)}\,\mathrm{d}t\,\|x^{(k)} - x^*\|. \tag{5.3}$$

Wegen $x^{(k)} \to x^*$ gilt $x^{(k)} + t\,(x^* - x^{(k)}) \to x^*$ gleichmäßig auf $t \in [0,1]$. Außerdem ist F' stetig. Zu jedem $\varepsilon > 0$ existiert also ein Index $k_0 \in \mathbb{N}$ mit

$$\begin{aligned} &\|D^{(k)}(t)\| \le \varepsilon \quad \text{für alle } k \ge k_0 \text{ und alle } t \in [0,1]. \\ \Rightarrow \quad &0 \le \int_0^1 \|D^{(k)}(t)\|\,\mathrm{d}t \le \varepsilon \quad \text{für alle } k \ge k_0. \end{aligned}$$

Das bedeutet aber: $\int_0^1 \|D^{(k)}(t)\|\,\mathrm{d}t \to 0$. Jetzt liefern (5.1) und (5.3):

$$\|x^{(k+1)} - x^*\| \le c \int_0^1 \|D^{(k)}(t)\|\,\mathrm{d}t\,\|x^{(k)} - x^*\|,$$

also die q-superlineare Konvergenz.

Quizfrage: Welche Norm ist im Ausdruck $\|D^{(k)}(t)\|$ eigentlich gemeint?

Aussage *(iii)*: Da $x^{(k)}$ und $x^{(k)} + t\,(x^* - x^{(k)})$ für alle $t \in [0,1]$ in $B_\delta(x^*)$ liegen, können wir das Integral unter den stärkeren Voraussetzungen besser abschätzen:

$$\int_0^1 \big\|F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\big\|\,\mathrm{d}t \le \int_0^1 L\,t\,\|x^* - x^{(k)}\|\,\mathrm{d}t = \frac{L}{2}\|x^{(k)} - x^*\|.$$

Aus (5.3) erhalten wir nun:

$$\|x^{(k+1)} - x^*\| \le c\,\frac{L}{2}\|x^{(k)} - x^*\|^2.$$

□

Bemerkung 5.9 (Zum lokalen Newton-Verfahren)**.**

(*i*) *Das lokale Newton-Verfahren (Algorithmus 5.1) kann scheitern, denn $F'(x^{(k)})$ muss nicht regulär sein, falls man außerhalb der (unbekannten) garantierten Konvergenzumgebung $B_\delta(x^*)$ startet.*

(*ii*) *Das sogenannte vereinfachte Newton-Verfahren, bei dem in Zeile 3 von Algorithmus 5.1 statt $F'(x^{(k)})$ die feste (invertierbare) Matrix $F'(x^{(0)})$ verwendet wird, konvergiert noch lokal q-linear.*

[7] **Beachte:** Unter dem Integral stehen Matrizen.

§ 5.3 Das lokale Newton-Verfahren in der Optimierung

Literatur: Geiger, Kanzow, 1999, Kapitel 9

Für den Rest von § 5 wird f als zweimal stetig partiell diffbar (C^2-Funktion) angenommen. Wir betrachten wieder die unrestringierte Aufgabe

$$\text{Minimiere } f(x) \text{ über } x \in \mathbb{R}^n. \tag{5.4}$$

Das Newton-Verfahren in der Optimierung lässt sich auf zwei verschiedene Weisen motivieren:

(*i*) Die notwendige Optimalitätsbedingung 1. Ordnung für (5.4) lautet

$$f'(x) = 0 \quad \text{oder äquivalent} \quad \nabla f(x) = 0,$$

siehe Satz 3.1. Wenden wir zur Lösung dieser i. A. nichtlinearen Gleichung (Nullstellensuche) das Newton-Verfahren mit $F(x) = \nabla f(x)$ und $F'(x) = f''(x)$ an, so erhalten wir die Iterationsvorschrift

$$x^{(k+1)} = x^{(k)} - f''(x^{(k)})^{-1} \nabla f(x^{(k)}). \tag{5.5}$$

(*ii*) Im aktuellen Iterationspunkt $x^{(k)}$ ersetzen wir (5.4) durch die Minimierung des **quadratischen Ersatzmodells** (Taylorpolynoms)

$$m^{(k)}(x) := f(x^{(k)}) + \nabla f(x^{(k)})^\mathsf{T}(x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^\mathsf{T} f''(x^{(k)})(x - x^{(k)}). \tag{5.6}$$

Falls die Hessematrix $f''(x^{(k)})$ positiv definit ist, so ist der eindeutige Minimierer durch

$$0 = \nabla m^{(k)}(x) = \nabla f(x^{(k)}) + f''(x^{(k)})(x - x^{(k)})$$

charakterisiert, vgl. (4.15). Wir wählen die Lösung dieses lineare Gleichungssystem als nächste Iterierte $x^{(k+1)}$ und erhalten wiederum die Iterationsvorschrift

$$x^{(k+1)} = x^{(k)} - f''(x^{(k)})^{-1} \nabla f(x^{(k)}).$$

Bemerkung 5.10 (Zum lokalen Newton-Verfahren)**.**

(*i*) *Satz 5.8 liefert die lokal q-superlineare bzw. q-quadratische Konvergenz von Algorithmus 5.1 mit $F(x) = \nabla f(x)$ gegen einen stationären Punkt x^* von f. Dieser kann auch ein lokaler Maximierer oder ein Sattelpunkt von f sein, da wir $f''(x^*)$ nur als regulär und nichts über die Definitheit voraussetzen.*

(*ii*) *Ist $f''(x^{(k)})$ s. p. d., so ist die aus dem linearen Gleichungssystem*

$$f''(x^{(k)})\, d^{(k)} = -\nabla f(x^{(k)})$$

erhaltene Newton-Richtung $d^{(k)}$ eine Abstiegsrichtung für f, vergleiche (4.12):

$$f'(x^{(k)})\, d^{(k)} = \nabla f(x^{(k)})^\mathsf{T} d^{(k)} = -\nabla f(x^{(k)})^\mathsf{T} \underbrace{f''(x^{(k)})^{-1}}_{\text{positiv definit}} \nabla f(x^{(k)}) < 0, \quad \text{falls } \nabla f(x^{(k)}) \neq 0.$$

Wegen der festen Schrittweite $t^{(k)} = 1$ im lokalen Newton-Verfahren ist jedoch i. A. kein Abstieg in f garantiert, wenn $x^{(k)}$ „weit" von einem lokalen Minimierer x^* entfernt ist.

(*iii*) *Das Newton-Verfahren ist invariant gegenüber affin-linearen Transformationen in der Grundmenge und in der Wertemenge. Das bedeutet, dass das Verfahren, angewendet auf die Aufgaben*

$$\textit{Minimiere} \quad f(x) \quad \textit{über } x \in \mathbb{R}^n \qquad \textit{und} \qquad \textit{Minimiere} \quad c\, f(Ay + b) + d \quad \textit{über } y \in \mathbb{R}^n$$

mit regulärer Matrix $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $c > 0$ und $d \in \mathbb{R}$ folgende Eigenschaft besitzt: Gelten für die Startschätzungen $x^{(0)} = Ay^{(0)} + b$, dann gilt auch $x^{(k)} = Ay^{(k)} + b$ für alle $k \in \mathbb{N}$.

***Quizfrage:** Gilt diese Eigenschaft auch für Gradientenverfahren?*

§ 5.4 Ein globalisiertes Newton-Verfahren in der Optimierung

Idee: Kombiniere die globalen Konvergenzeigenschaften des Gradientenverfahrens (Algorithmus 4.10) mit der schnellen lokalen Konvergenz des Newton-Verfahrens (Algorithmus 5.1).

Algorithmus 5.11 (Globalisiertes Newton-Verfahren in der Optimierung)**.**
Eingabe: *Startschätzung $x^{(0)} \in \mathbb{R}^n$*
Eingabe: *Armijo-Parameter $\sigma \in (0, 1/2)$, Backtracking-Parameter $\beta \in (0, 1)$*
Eingabe: *Globalisierungs-Parameter $\varrho_1 > 0$, $\varrho_2 > 0$ und $p > 0$*
Eingabe: *s. p. d. Matrix $M \in \mathbb{R}^{n\times n}$*

1: *Setze $k := 0$*
2: ***while** Abbruchkriterium nicht erfüllt **do***
3: *Löse, wenn möglich, das lineare Gleichungssystem $f''(x^{(k)})\, d^{(k)} := -\nabla f(x^{(k)})$ nach der Newton-Richtung $d^{(k)}$*
4: *Ist dieses System nicht oder nicht eindeutig lösbar oder gilt*

$$-f'(x^{(k)})\, d^{(k)} \le \min\{\varrho_1,\ \varrho_2 \|d^{(k)}\|_M^p\}\, \|d^{(k)}\|_M^2, \tag{5.7}$$

so setze $d^{(k)} := -\nabla_M f(x^{(k)})$ *// Fallback auf Gradientenrichtung*
5: *Bestimme eine Schrittweite $t^{(k)}$ mit der Armijo-Backtracking-Strategie zur Startschrittweite $s = 1$, sodass (4.3) erfüllt ist, also:*

$$f(x^{(k)} + t^{(k)}\, d^{(k)}) \le f(x^{(k)}) + \sigma\, t^{(k)}\, f'(x^{(k)})\, d^{(k)}$$

6: *Setze $x^{(k+1)} := x^{(k)} + t^{(k)}\, d^{(k)}$*
7: *Setze $k := k + 1$*
8: ***end while***

Zur Durchführung des globalisierten Newton-Verfahrens werden folgende problemspezifische Routinen benötigt:

(1) Routine zur Auswertung der Zielfunktion $f(x)$.

(2) Routine zur Auswertung der Ableitung $f'(x)$ bzw. des Gradienten $\nabla f(x)$.

(3) Routine zur Auswertung der Hessematrix $f''(x)$.

Bemerkung 5.12 (Zum globalisierten Newton-Verfahren)**.**

(*i*) *Bei unbrauchbarer Newton-Richtung weichen wir also auf einen Gradientenschritt aus. Entweder die nicht erfüllte Bedingung (5.7) oder aber die Wahl $d^{(k)} = -\nabla_M f(x^{(k)})$ sichert $f'(x^{(k)})\, d^{(k)} < 0$. Die Armijo-Backtracking-Strategie liefert also immer eine Schrittweite (Satz 4.3), und der Algorithmus ist wohldefiniert.*

(*ii*) *Als Abbruchbedingungen kommen wiederum diejenigen aus Bemerkung 4.7 zum Einsatz.*

(*iii*) *Die Vorgabe $\sigma < 1/2$ und die Wahl der Startschrittweite $s = 1$ sind wesentlich, damit für hinreichend große $k \in \mathbb{N}$ tatsächlich volle Newton-Schritte ($t^{(k)} = 1$) gegangen werden können.*

(*iv*) *Im praktischen Einsatz kommt in Algorithmus 5.11 auch die nicht-monotone Armijo-Regel zum Einsatz, bei der hinreichender Abstieg nur im Vergleich zum Maximum der letzten Funktionswerte gefordert wird, siehe Geiger, Kanzow, 1999, Ende Abschnitt 9.3, S.96.*

Wir geben Konvergenzaussagen für Algorithmus 5.11 ohne Abbruchbedingung, sodass eine unendliche Folge $(x^{(k)})$ entsteht, an.

Satz 5.13 (Globaler Konvergenzsatz für das globalisierte Newton-Verfahren)**.**
Es sei $(x^{(k)})$ eine durch Algorithmus 5.11 erzeugte Folge. Dann gilt:

(*i*) *Jeder Häufungspunkt x^* von $(x^{(k)})$ ist ein stationärer Punkt von f, erfüllt also $f'(x^*) = 0$.*

(*ii*) *Ist x^* ein isolierter Häufungspunkt von $(x^{(k)})$, dann konvergiert bereits die gesamte Folge $x^{(k)} \to x^*$.*

Beweis. siehe Geiger, Kanzow, 1999, Satz 9.5 und Satz 9.7 □

Satz 5.14 (Lokaler Konvergenzsatz für das globalisierte Newton-Verfahren)**.**
Es seien $(x^{(k)})$, $(d^{(k)})$ durch den Algorithmus 5.11 erzeugte Folgen. Ist x^* ein Häufungspunkt von $(x^{(k)})$ und ist $f''(x^*)$ s. p. d., so gilt:

(*i*) *Die gesamte Folge $(x^{(k)})$ konvergiert gegen den strikten lokalen Minimierer x^*.*

(*ii*) *Für alle hinreichend großen $k \in \mathbb{N}$ ist die verwendete Suchrichtung $d^{(k)}$ immer die Newton-Richtung, und es wird die volle Schrittweite $t^{(k)} = 1$ akzeptiert.*

(*iii*) *$(x^{(k)})$ konvergiert q-superlinear gegen x^*.*

(*iv*) *Ist f'' Lipschitz-stetig in einer Umgebung von x^*, so konvergiert $(x^{(k)})$ q-quadratisch gegen x^*.*

Beweis. siehe Geiger, Kanzow, 1999, Satz 9.10 □

Alle hier besprochenen Basis-Algorithmen zur Lösung freier Optimierungsaufgaben sind **Liniensuchverfahren** (englisch: *line search methods*), die in jeder Iteration

(1) eine Suchrichtung $d^{(k)}$

(2) und anschließend eine geeignete Schrittlänge $t^{(k)}$

bestimmen. Als Alternative sind auch **Trust-Region-Verfahren** (englisch: *trust-region methods*) etabliert, die beide Schritte gemeinsam durchführen, siehe Vorlesung *Nichtlineare Optimierung* und Geiger, Kanzow, 1999, Abschnitt 14.

Allen hier besprochenen Verfahren ist gemeinsam, dass sie die Suchrichtung $d^{(k)}$ durch Minimierung eines lokalen quadratischen Ersatzmodells

$$q^{(k)}(d) := f(x^{(k)}) + f'(x^{(k)})\, d + \frac{1}{2} d^{\mathsf{T}} B^{(k)}\, d$$

gewinnen, d. h. (bei s. p. d. Matrix $B^{(k)}$) aus dem linearen Gleichungssystem

$$B^{(k)}\, d^{(k)} = -\nabla f(x^{(k)}).$$

Folgende Tabelle fasst typische Eigenschaften dieser Verfahren zusammen:

Gradientenverfahren	$B^{(k)} = \mathrm{Id}$	q-linear, einfaches Verfahren
vork. Gradientenverf.	$B^{(k)} = M$	q-linear, einfaches Verfahren
Quasi-Newton-Verf.	$B^{(k)}$ variiert	bis q-superlinear, oft guter Kompromiss
Newton-Verfahren	$B^{(k)} = f''(x^{(k)})$	q-superlinear oder besser, aber aufwändig

Mehr insbesondere zu Quasi-Newton-Verfahren in der Vorlesung *Nichtlineare Optimierung*.

Ende der Woche 3

Kapitel 2 Lineare Optimierung

§ 6 Einführung

Literatur: Geiger, Kanzow, 2002, Kapitel 3.1

Lineare Optimierungsaufgaben (**lineare Programme**, **LP**) sind insbesondere in den Wirtschaftswissenschaften von großer Bedeutung. Sie umfassen u. a. Transport- und Logistikprobleme, Kürzeste-Wege-Aufgaben usw. Es sind im gesamten Kapitel 2 stets f, g und h aus der allgemeinen Aufgabenstellung (1.1) (affin-)lineare Funktionen von x, und die Grundmenge ist $\Omega = \mathbb{R}^n$.

Eine lineare Optimierungsaufgabe kann also immer in folgender Form geschrieben werden:

$$\left.\begin{aligned} \text{Minimiere} \quad & c^\mathsf{T} x + \gamma \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} \quad & A_{\text{ineq}} x \leq b_{\text{ineq}} \\ \text{und} \quad & A_{\text{eq}} x = b_{\text{eq}}. \end{aligned}\right\} \tag{6.1}$$

Dabei heißt $c \in \mathbb{R}^n$ der **Kostenvektor** der Aufgabe. Weiter sind $A_{\text{ineq}} \in \mathbb{R}^{m \times n}$ und $b_{\text{ineq}} \in \mathbb{R}^m$ sowie $A_{\text{eq}} \in \mathbb{R}^{p \times n}$ und $b_{\text{eq}} \in \mathbb{R}^p$. Die Ungleichungen sind komponentenweise zu verstehen. Es ist erlaubt, dass $m = 0$ (keine Ungleichungen) oder $p = 0$ (keine Gleichungen) gilt, sodass die Beschränkungen des jeweiligen Typs nicht vertreten sind.

Quizfrage: In der Regel setzt man den konstanten Term γ in der Zielfunktion gleich null. Warum stellt das keine Einschränkung in der Aufgabenstellung dar?

Quizfrage: Warum stellt der Verzicht auf Ungleichungen der Form $A_{\text{ineq}} x \geq b_{\text{ineq}}$ ebenfalls keine Einschränkung in der Aufgabenstellung dar?

Lineare Programme sind Spezialfälle konvexer Optimierungsaufgaben (Kapitel 3), daher brauchen wir nicht zwischen lokalen und globalen Lösungen zu unterscheiden (Satz 14.1). Wir wollen das hier aber schon einmal direkt nachweisen:

Satz 6.1. *Jeder lokale Minimierer von (6.1) ist bereits ein globaler Minimierer.*

Beweis. Wir bezeichnen mit

$$F := \{x \in \mathbb{R}^n \mid A_{\text{ineq}} x \leq b_{\text{ineq}} \text{ und } A_{\text{eq}} x = b_{\text{eq}}\}$$

die zulässige Menge und mit $f(x) := c^\mathsf{T} x + \gamma$ die Zielfunktion. Es sei nun x^* ein lokaler Minimierer von (6.1), d. h., es existiert eine Umgebung $U(x^*)$ mit $f(x^*) \leq f(x)$ für alle $x \in F \cap U(x^*)$, vgl. Definition 1.1.

Wir führen einen Widerspruchsbeweis. Angenommen, es gäbe ein $\widehat{x} \in F$ mit $f(\widehat{x}) < f(x^*)$. Wir betrachten Punkte x_α entlang der Verbindungsstrecke zwischen x^* und $\widehat{x}$, also

$$x_\alpha = \alpha\,\widehat{x} + (1-\alpha)\,x^* \quad \text{mit } \alpha \in [0,1].$$

Alle diese Punkte sind zulässig, denn:

$$\begin{aligned} A_{\text{ineq}} x_\alpha &= A_{\text{ineq}}(\alpha\,\widehat{x} + (1-\alpha)\,x^*) = \alpha\,A_{\text{ineq}}\widehat{x} + (1-\alpha)\,A_{\text{ineq}}x^* \leq \alpha\,b_{\text{ineq}} + (1-\alpha)\,b_{\text{ineq}} = b_{\text{ineq}}, \\ A_{\text{eq}} x_\alpha &= A_{\text{eq}}(\alpha\,\widehat{x} + (1-\alpha)\,x^*) \;\; = \alpha\,A_{\text{eq}}\widehat{x} + (1-\alpha)\,A_{\text{eq}}x^* \;\;\;\; = \alpha\,b_{\text{eq}} + (1-\alpha)\,b_{\text{eq}} \;\;\;\; = b_{\text{eq}}. \end{aligned}$$

Für die Werte der Zielfunktion gilt

$$\begin{aligned} f(x_\alpha) &= c^\mathsf{T} x_\alpha + \gamma \\ &= c^\mathsf{T}(\alpha\,\widehat{x} + (1-\alpha)\,x^*) + \gamma \\ &= \alpha\,(c^\mathsf{T}\widehat{x} + \gamma) + (1-\alpha)\,(c^\mathsf{T}x^* + \gamma) \\ &= \alpha\,f(\widehat{x}) + (1-\alpha)\,f(x^*). \end{aligned}$$

Für $\alpha \in (0,1]$ folgt daher wegen $f(\widehat{x}) < f(x^*)$:

$$f(x_\alpha) < \alpha\,f(x^*) + (1-\alpha)\,f(x^*) = f(x^*).$$

Nun liegt aber für hinreichend kleines $\alpha > 0$ der Punkt x_α in der Umgebung $U(x^*)$ und damit in $U(x^*) \cap F$. Dies steht im Widerspruch zur lokalen Optimalität von x^*. Also kann ein $\widehat{x}$ wie oben angenommen nicht existieren, d. h., es gilt

$$f(x^*) \leq f(x) \quad \text{für alle } x \in F.$$

Mit anderen Worten: Jeder lokale Minimierer von (6.1) ist bereits ein globaler Minimierer. □

Beispiel 6.2 (Mozartproblem)**.**
Eine Firma stellt Mozartkugeln und Mozarttaler her und benötigt dafür folgende Zutaten pro Einheit des hergestellten Produkts:

	Marzipan	*Nougat*	*Schokolade*	*Gewinn pro Einheit*
Mozartkugeln	1	2	1	9
Mozarttaler	1	1	2	8
Lagerbestand	6	11	9	

Wir möchten bestimmen, wieviele Einheiten an Mozartkugeln und -talern produziert werden sollten, um den Gewinn zu maximieren.

Es sei

$$\begin{aligned} x_1 &= \textit{Menge an Mozartkugeln}, \\ x_2 &= \textit{Menge an Mozarttalern}. \end{aligned}$$

Wir erhalten folgende lineare Optimierungsaufgabe:

$$\left.\begin{aligned} \textit{Maximiere} \quad & 9\,x_1 + 8\,x_2 \quad \textit{über } x \in \mathbb{R}^2 \\ \textit{sodass} \quad & x_1 + x_2 \leq 6 \\ & 2\,x_1 + x_2 \leq 11 \\ & x_1 + 2\,x_2 \leq 9 \\ \textit{und} \quad & x_1 \geq 0 \\ & x_2 \geq 0. \end{aligned}\right\} \tag{6.2}$$

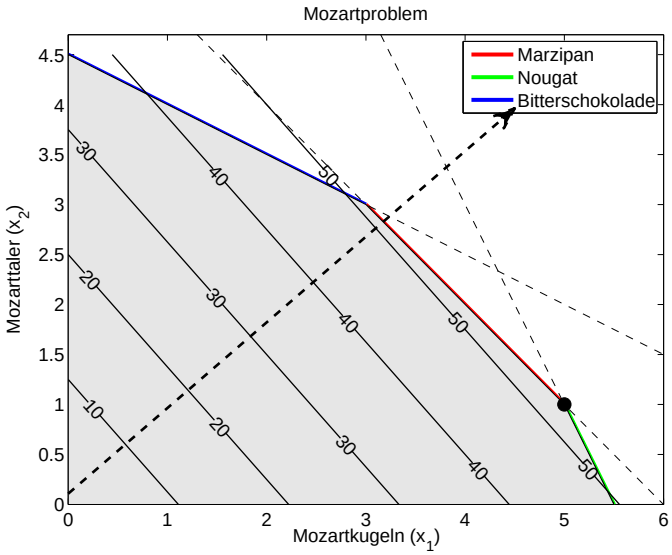


Abbildung 6.1: Zulässige Menge (Fünfeck), Niveaulinien der Zielfunktion und globaler Maximierer beim Mozartproblem (Beispiel 6.2).

Aus der Modellierung von Beispiel 6.2 ergibt sich folgender häufig vorkommender Spezialfall der allgemeinen linearen Optimierungsaufgabe (6.1):

$$\left.\begin{aligned} \text{Maximiere} \quad & c^\mathsf{T} x \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} \quad & Ax \leq b \\ \text{und} \quad & x \geq 0. \end{aligned}\right\} \tag{6.3}$$

Dabei sind der **Kostenvektor** $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ mit $m \in \mathbb{N}$. Ein LP der Gestalt (6.3) heißt in **kanonischer Form**. Beim Mozartproblem (6.2) ist z. B.

$$c = \begin{pmatrix} 9 \\ 8 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 6 \\ 11 \\ 9 \end{pmatrix}.$$

Quizfrage: Können wir jedes LP in kanonischer Form schreiben?

Enthält ein gegebenes LP …

(*i*) eine Beschränkung der Form $a^\mathsf{T}x \geq \beta$, so können wir sie mit (-1) multiplizieren: $-a^\mathsf{T}x \leq -\beta$.

(*ii*) eine Gleichungsbeschränkung $a^\mathsf{T}x = \beta$, so können wir diese in Form zweier Ungleichungen, $a^\mathsf{T}x \leq \beta$ und $a^\mathsf{T}x \geq \beta$ bzw. $-a^\mathsf{T}x \leq -\beta$, schreiben.

(*iii*) für eine Variable x_i keine Beschränkung der Form $x_i \geq 0$ (**freie Variable**), so ersetzen wir $x_i := x_i^+ - x_i^-$ und fordern $x_i^+ \geq 0$ und $x_i^- \geq 0$.

Mit Hilfe dieser Transformationen kann gezeigt werden:

Lemma 6.3 (Transformierbarkeit in kanonische Form)**.**
Jedes LP ist äquivalent zu einem LP in kanonischer Form.

Quizfrage: Was bedeutet diese Äquivalenz genau?

Definition 6.4 (Hyperebene, Halbraum, Polyeder)**.**

(*i*) *Es sei $a \in \mathbb{R}^n$, $a \neq 0$ und $\beta \in \mathbb{R}$. Die Menge*

$$H(a,\beta) := \{x \in \mathbb{R}^n \mid a^\mathsf{T}x = \beta\} \tag{6.4}$$

*heißt **Hyperebene** (englisch:* hyperplane*) im $\mathbb{R}^n$ mit **Normalenvektor** a.*

(*ii*) *Eine Hyperebene teilt den Raum $\mathbb{R}^n$ in zwei abgeschlossene **Halbräume** (englisch:* half-spaces*)*

$$H^-(a,\beta) := \{x \in \mathbb{R}^n \mid a^\mathsf{T}x \leq \beta\} \quad \text{und} \quad H^+(a,\beta) := \{x \in \mathbb{R}^n \mid a^\mathsf{T}x \geq \beta\}. \tag{6.5}$$

(*iii*) *Der Durchschnitt endlich vieler Halbräume wird als **(konvexes) Polyeder** (griechisch für Vielflächner, englisch:* polyhedron*) bezeichnet.*

Ein Polyeder kann also durch endlich viele affin-lineare Gleichungs- und Ungleichungsrestriktionen beschrieben werden. Insbesondere ist die zulässige Menge

$$\{x \in \mathbb{R}^n \mid Ax \leq b,\ x \geq 0\}$$

der Aufgabe (6.3) ein Polyeder. Ein Beispiel war bereits in Abbildung 6.1 zu sehen.

Für die algorithmische Behandlung von LPs sind Gleichungen allerdings geeigneter als Ungleichungen. Daher führen wir jetzt die für uns wichtigste Form linearer Optimierungsaufgaben ein.

Definition 6.5 (LP in Normalform)**.**
Ein LP der Gestalt

$$\left.\begin{aligned} \textit{Minimiere} \quad & c^\mathsf{T}x \quad \textit{über } x \in \mathbb{R}^n \\ \textit{sodass} \quad & Ax = b \\ \textit{und} \quad & x \geq 0 \end{aligned}\right\} \tag{6.6}$$

*mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ heißt in **Normalform** bzw. **Standardform**.*

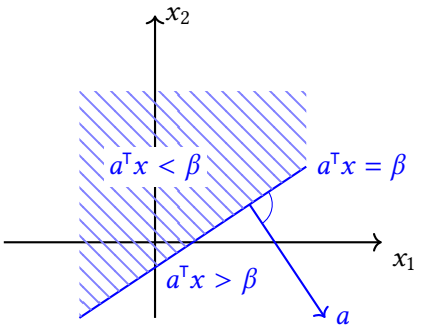


Abbildung 6.2: Darstellung einer Hyperebene mit Normalenvektor a und der beiden Halbräume.

Enthält ein gegebenes LP …

(*i*) eine Ungleichungsbeschränkung $a^\mathsf{T}x \leq \beta$, so führen wir eine zusätzliche, sogenannte **Schlupfvariable** (**Überschussvariable**, englisch: *slack variable*) s ein und ersetzen die Ungleichung durch

$$a^\mathsf{T}x + s = \beta, \quad s \geq 0.$$

Quizfrage: Wie geht man bei $a^\mathsf{T}x \geq \beta$ vor?

(*ii*) freie Variablen x_i, so setzen wir wie bereits bei der Umwandlung einer Aufgabe in kanonische Form $x_i := x_i^+ - x_i^-$ und fordern $x_i^+ \geq 0$ und $x_i^- \geq 0$.

Beachte: Eine Schlupfvariable gibt den Abstand (englisch: *slack*) zur Gleichheit an.

Mit obigen Umformungen kann man zeigen:

Lemma 6.6 (Transformierbarkeit in Normalform)**.**
Jedes LP ist äquivalent zu einem LP in Normalform.

Beispiel 6.7 (Mozartproblem in Normalform)**.**
Wir führen drei Schlupfvariablen s_1, s_2, s_3 ein:

$$\left.\begin{aligned} \textit{Minimiere} \quad & -9\,x_1 - 8\,x_2 \\ \textit{sodass} \quad & x_1 + x_2 + s_1 = 6 \\ & 2\,x_1 + x_2 + s_2 = 11 \\ & x_1 + 2\,x_2 + s_3 = 9 \\ \textit{und} \quad & x_1, x_2 \geq 0 \\ & s_1, s_2, s_3 \geq 0. \end{aligned}\right\} \tag{6.7}$$

Die Aufgabe hat nun fünf Variablen $(x, s) \in \mathbb{R}^2 \times \mathbb{R}^3$!

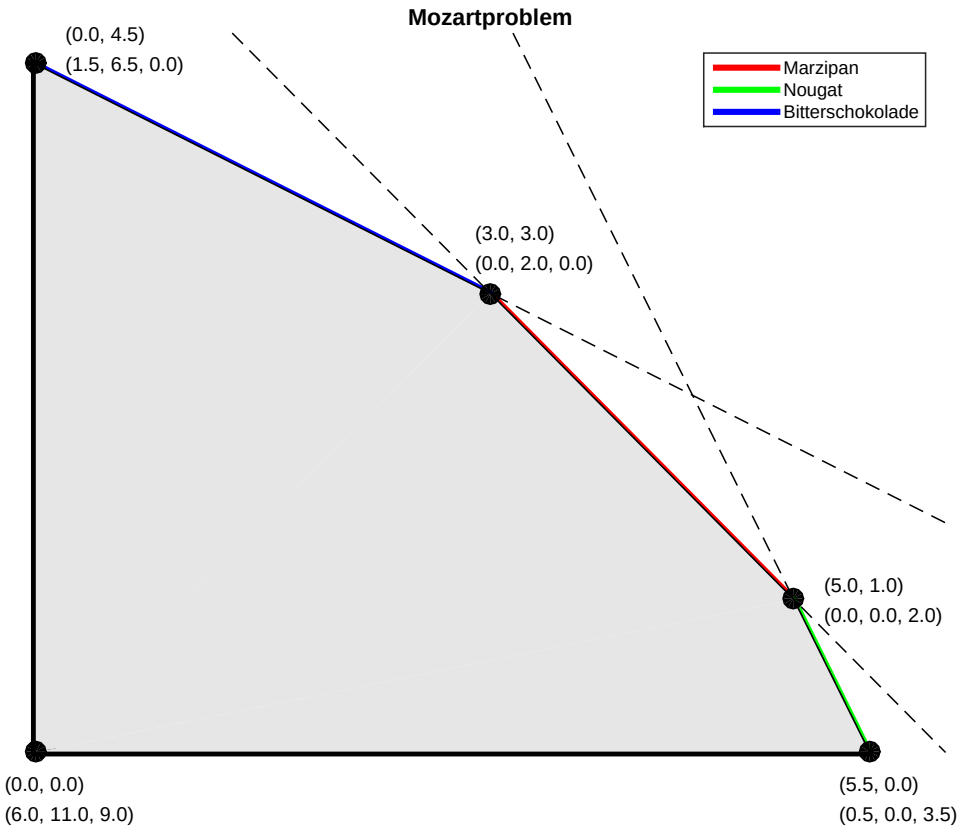


Abbildung 6.3: Darstellung der Ecken der zulässigen Menge beim Mozartproblem in Normalform (6.7) mitsamt den jeweiligen Werten der Schlupfvariablen.

Sofern nichts anderes gesagt wird, gehen wir jetzt immer davon aus, dass ein LP in Normalform vorliegt. Die zulässige Menge

$$P := \{x \in \mathbb{R}^n \mid A\,x = b,\ x \geq 0\} \tag{6.8a}$$

heißt dann ein **in Normalform beschriebes Polyeder**, kurz: **Polyeder in Normalform**. Für die Dimension der Matrix $A \in \mathbb{R}^{m\times n}$ nehmen wir dabei

$$1 \leq m \leq n \tag{6.8b}$$

an.[1]

§ 6.1 Existenz von Lösungen

In diesem Abschnitt gehen wir der Frage nach, wann die lineare Optimierungsaufgabe (6.6) eine Lösung besitzt.

[1] Für $m = 0$ ist die Aufgabe entweder unbeschränkt (wenn ein $c_i < 0$ ist), oder $x^* = 0$ ist eine Lösung (wenn $c \geq 0$ gilt). Im Fall $m > n$ können entweder solange redundante Gleichungen gestrichen werden, bis $m \leq n$ wird, oder $A\,x = b$ ist unlösbar, d. h. (6.6) ist unzulässig.

Vorüberlegung: Die zulässige Menge $P \subseteq \mathbb{R}^n$ von (6.6) ist immer abgeschlossen. (**Quizfrage:** Warum eigentlich?) *Falls* sie auch nichtleer und beschränkt (also kompakt) ist, dann besitzt die stetige Zielfunktion $c^\mathsf{T}x$ über P nach dem Satz von Weierstraß bzw. Satz 1.4 einen Minimierer. Allerdings ist P im Allgemeinen nicht beschränkt, siehe Abbildung 6.4. Mit Satz 1.4 können wir dann nicht argumentieren, da die Sub-Levelmengen $L := \{x \in P \mid c^\mathsf{T}x \le m\}$ möglicherweise nicht beschränkt (kompakt) sind.

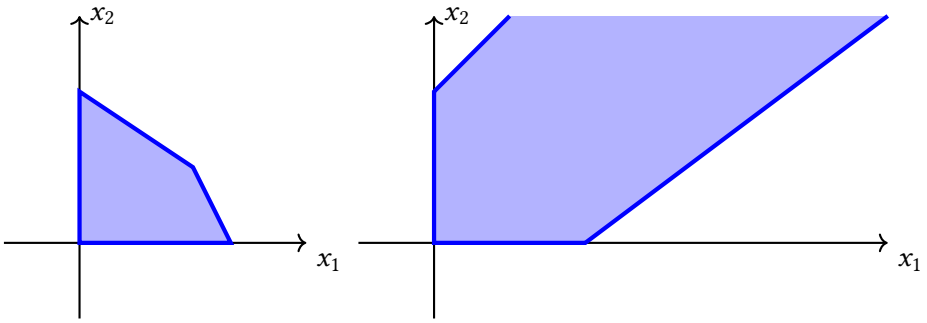


Abbildung 6.4: Kompaktes (abgeschlossenes und beschränktes) Polyeder (links) und unbeschränktes Polyeder (rechts).

Intuitiv sollte klar sein, dass ein LP unbeschränkt ist, falls die Zielfunktion entlang eines Strahles $t \mapsto x + t\,d$ abfällt, der für alle $t \ge 0$ in der zulässigen Menge bleibt. Wir formulieren dies als Resultat:

Lemma 6.8. *Wir betrachten ein LP in Normalform (6.6) mit zulässiger Menge P wie in (6.8). Weiter sei $P \neq \emptyset$. Ist f^* endlich, dann gilt $c^\mathsf{T}d \ge 0$ für alle Richtungen in der Menge*

$$\{d \in \mathbb{R}^n \mid A\,d = 0,\ d \ge 0\}. \tag{6.9}$$

Die Menge in (6.9) wird **Rezessionskegel** (englisch: *recession cone*) des Polyeders (6.8) genannt. **Quizfrage:** Welche Bedeutung hat die Menge in (6.9)?

Beweis. Wir führen einen Widerspruchsbeweis. Angenommen, es gebe eine Richtung d aus der Menge (6.9) mit der Eigenschaft $c^\mathsf{T}d < 0$. (**Quizfrage:** Was bedeutet $c^\mathsf{T}d < 0$?) Dann ist $x + t\,d$ für alle $t \ge 0$ zulässig, und es gilt

$$c^\mathsf{T}(x + t\,d) = c^\mathsf{T}x + t\,c^\mathsf{T}d \to -\infty \text{ für } t \to \infty.$$

Daraus folgt $f^* = -\infty$, im Widerspruch zur Endlichkeit von f^*. □

Satz 6.9 (Existenzsatz für LPs)**.**
Wir betrachten ein LP in allgemeiner Form (6.1) mit zulässiger Menge F. Ist der Optimalwert

$$f^* = \inf\{c^\mathsf{T}x \mid x \in F\}$$

endlich, also die Aufgabe (6.1) weder unzulässig ($f^* = +\infty$) noch unbeschränkt ($f^* = -\infty$), so besitzt (6.6) mindestens einen Minimierer.

Zum Beweis des Satzes benötigen wir ein Hilfsresultat. Dieses sagt aus, dass die Menge der nichtnegativen Linearkombinationen einer gegebenen Menge von Vektoren abgeschlossen ist.

Lemma 6.10 (Abgeschlossenheit der Menge nicht-negativer Linearkombinationen[2])**.**
Es sei $B \in \mathbb{R}^{m\times n}$ eine Matrix (ohne Einschränkungen an $n, m \in \mathbb{N}$). Die Menge

$$K := \{B\,d \mid d \in \mathbb{R}^n,\ d \geq 0\} \tag{6.10}$$

der nichtnegativen Linearkombinationen der Spalten von B ist abgeschlossen.

Quizfrage: Wie kann man sich die Menge in (6.10) grafisch vorstellen?

Beweis. Wir bezeichnen die Spalten von B mit $b_1, \ldots, b_n \in \mathbb{R}^m$ und die Komponenten des Vektors d, also die Koeffizienten der Linearkombination, mit $\delta_1, \ldots, \delta_n$. Wir benutzen Induktion nach der Spaltenanzahl n. Es sei zunächst $n = 1$, dann ist

$$K = \{B\,d \mid d \in \mathbb{R}^n,\ d \geq 0\} = \{\delta_1\, b_1 \mid \delta_1 \geq 0\}$$

eine abgeschlossene Halbgerade oder $\{0\}$.

Induktionsschluss von $n-1$ auf n: Es sei bereits gezeigt, dass die Menge der nicht-negativen Linearkombinationen von höchstens $n-1$ Vektoren abgeschlossen ist. Es sei nun K von den n Vektoren $b_1, \ldots, b_n$ erzeugt. Wir müssen zeigen, dass K abgeschlossen ist.

Wir machen eine Fallunterscheidung:

Fall 1: Falls diese n Vektoren linear unabhängig sind, dann hat $B \in \mathbb{R}^{m\times n}$ vollen (Spalten-)Rang, also $\operatorname{Rang}(B) = n$. Also ist $B^\mathsf{T}B \in \mathbb{R}^{n\times n}$ positiv definit, insbesondere invertierbar. Wir betrachten eine Folge $(d^{(k)}) \subseteq \mathbb{R}^n$ nichtnegativer Vektoren, sodass $(B\,d^{(k)}) \subseteq K$ gilt und $B\,d^{(k)} =: y^{(k)} \to y$ in $\mathbb{R}^m$. Zu zeigen ist $y \in K$. Aufgrund des vollen Rangs von B können wir die Koeffizienten $d^{(k)}$ aus $y^{(k)}$ eindeutig rekonstruieren:

$$d^{(k)} = (B^\mathsf{T}B)^{-1}B^\mathsf{T}y^{(k)} \geq 0.$$

Der Grenzübergang $k \to \infty$ zeigt die Konvergenz der Folge $(d^{(k)})$ gegen einen Grenzwert $d \geq 0$, und es gilt $B\,d = y$. Damit gehört der Grenzwert y zu K. **Beachte:** In diesem Fall wird die Induktionsannahme gar nicht verwendet.

Fall 2: Falls die Vektoren $b_1, \ldots, b_n$ linear abhängig sind, dann existieren Zahlen $\gamma_1, \ldots, \gamma_n$ mit der Eigenschaft

$$\sum_{i=1}^{n} \gamma_i\, b_i = 0, \tag{6.11}$$

wobei nicht alle $\gamma_i = 0$ sind. Sagen wir o. B. d. A., mindestens ein γ_i ist < 0. Es sei nun $y \in K$ ein beliebiges Element in K mit der Darstellung

$$y = \sum_{i=1}^{n} \delta_i\, b_i, \quad \delta_i \geq 0. \tag{6.12}$$

[2] Beweis aus Werner, 2007, Lemma 1.5

Wir werden gleich zeigen, dass sich y bereits als nichtnegative Linearkombination von nur $n-1$ der Vektoren $b_1, \dots, b_n$ darstellen lässt. Da $y \in K$ beliebig war, folgt dann

$$K = \bigcup_{s=1}^{n} \{B_{-s}\, d \,|\, d \in \mathbb{R}^{n-1},\ d \geq 0\},$$

wobei B_{-s} die Matrix B ohne die Spalte b_s bezeichnet. Nach Induktionsvoraussetzung ist jede der Mengen in der Vereinigung abgeschlossen, also auch deren endliche Vereinigung.

In der Linearkombination (6.12) seien alle $\delta_i > 0$, ansonsten sind wir fertig. Es sei $1 \leq s \leq n$ einer derjenigen Indizes, für die gilt:

$$-\frac{\delta_s}{\gamma_s} = \min\left\{-\frac{\delta_i}{\gamma_i} \,\middle|\, \gamma_i < 0,\ i = 1, \dots, n\right\} =: t > 0.$$

Wir geben nun eine neue Linearkombination für y an mit Koeffizienten

$$\widehat{\delta}_i := \delta_i + t\,\gamma_i, \quad i = 1, \dots, n.$$

Nach Konstruktion sind alle $\widehat{\delta}_i \geq 0$, und es gilt $\widehat{\delta}_s = 0$. In der Tat gilt

$$y = \sum_{i=1}^{n} \delta_i\, b_i \overset{(6.11)}{=} \sum_{i=1}^{n} (\delta_i + t\,\gamma_i)\, b_i = \sum_{\substack{i=1 \\ i \neq s}}^{n} \widehat{\delta}_i\, b_i.$$

Damit ist gezeigt: $y \in \{B_{-s}\, d \,|\, d \in \mathbb{R}^{n-1},\ d \geq 0\}$. □

Die Menge (6.10) heißt auch die **konische Hülle** (englisch: *conic hull*) der Vektoren $b_1, \dots, b_n$, kurz:

$$K = \{B\, d \,|\, d \in \mathbb{R}^n,\ d \geq 0\} = \operatorname{cone}\{b_1, \dots, b_n\}. \tag{6.13}$$

Wir können nun Satz 6.9 beweisen.

Beweis von Satz 6.9. Wir können o. B. d. A. annehmen, dass das betreffende LP in Normalform (6.6) mit zulässiger Menge wie in (6.8) gegeben ist. Es sei $f^* = \inf\{c^\mathsf{T} x \,|\, x \in P\}$ der endliche Optimalwert von (6.6). Es existiert also eine sogenannte Minimalfolge $(x^{(k)}) \subseteq P$ mit der Eigenschaft $c^\mathsf{T} x^{(k)} \searrow f^*$.

Betrachte die Folge

$$\begin{pmatrix} c^\mathsf{T} x^{(k)} \\ 0 \end{pmatrix} = \begin{pmatrix} c^\mathsf{T} x^{(k)} \\ Ax^{(k)} - b \end{pmatrix}.$$

Diese konvergiert gegen $(f^*, 0)^\mathsf{T} \in \mathbb{R} \times \mathbb{R}^m$. Andererseits gehören die Elemente der Folge zu der Menge

$$\widetilde{K} := \left\{ \begin{pmatrix} c^\mathsf{T} z \\ Az \end{pmatrix} - \begin{pmatrix} 0 \\ b \end{pmatrix} \in \mathbb{R} \times \mathbb{R}^m \,\middle|\, z \geq 0 \right\}.$$

Diese Menge ist, abgesehen von der Verschiebung um den konstanten Vektor $\left(\begin{smallmatrix} 0 \\ b \end{smallmatrix}\right)$, von der Bauart (6.10) mit der Matrix $B = \left(\begin{smallmatrix} c^\mathsf{T} \\ A \end{smallmatrix}\right) \in \mathbb{R}^{(1+m)\times n}$. Nach Lemma 6.10 ist $\widetilde{K}$ abgeschlossen. Daraus folgt, dass der Grenzwert $\left(\begin{smallmatrix} f^* \\ 0 \end{smallmatrix}\right)$ in $\widetilde{K}$ liegt. Das heißt, es existiert ein $x^* \geq 0$ mit der Eigenschaft $c^\mathsf{T} x^* = f^*$ und $Ax^* - b = 0$. Damit ist x^* eine Lösung des LP (6.6). □

Bemerkung 6.11. *Es ist eine durchaus bemerkenswerte Eigenschaft linearer Optimierungsaufgaben, dass sie bereits dann eine optimale Lösung besitzen, wenn der Optimalwert endlich ist. Wie wir aus Beispielen wie „Minimiere $1/x$ unter der Nebenbedingung $x \geq 1$“ wissen, ist das für nichtlineare Aufgaben i. A. nicht der Fall.*

§ 6.2 Die Bedeutung der Ecken

Definition 6.12 (Extremalpunkt bzw. Ecke eines Polyeders)**.**
*Ein Vektor $x \in P$ heißt **Extremalpunkt** oder **Ecke** eines Polyeders P (nicht notwendig in Normalform gegeben), wenn aus*

$$x = \alpha\, y + (1-\alpha)\, z$$

für $y, z \in P$ und $\alpha \in (0,1)$ bereits $y = z$ folgt.

Eine Ecke ist also dadurch gekennzeichnet, dass sie nicht auf der Verbindungsstrecke zweier anderer Punkte y, z von P liegt. Man sagt auch: Eine Ecke ist keine echte Konvexkombination (Definition 13.4) zweier anderer Punkte von P.

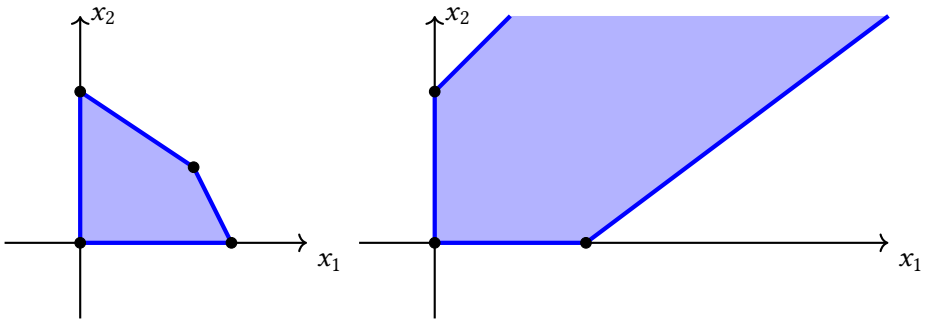


Abbildung 6.5: Polyeder (nicht in Normalform) und ihre Ecken: Viereck im $\mathbb{R}^2$ und unbeschränktes Polyeder mit drei Ecken im $\mathbb{R}^2$.

Ecken eines Polyeders in *Normalform* können wie folgt charakterisiert werden:

Satz 6.13 (Charakterisierung der Ecken eines Polyeders in Normalform)**.**
*Es sei P ein Polyeder in Normalform wie in (6.8) und $x \in P$ gegeben. Ferner sei $\mathcal{I}(x) = \{1 \leq i \leq n \mid x_i > 0\}$ die Menge der **inaktiven Indizes** (bzgl. der Ungleichungen $x \geq 0$). Dann gilt:*

$$\begin{aligned} & x \text{ ist eine Ecke von } P \\ \Leftrightarrow\ & \text{die Menge der Spalten } (a_i)_{i \in \mathcal{I}(x)} \text{ von } A \text{ ist linear } \underline{\text{un}}\text{abhängig.} \end{aligned}$$

Beachte: Insbesondere ist $x = 0$, sofern zu P gehörig, immer eine Ecke.

Beweis. „⇒“: Es sei x eine Ecke von P. Wir nehmen an, dass die Spalten $(a_i)_{i \in \mathcal{I}(x)}$ linear <u>ab</u>hängig sind. Damit muss natürlich notwendigerweise $I(x) \neq \emptyset$ sein. Wegen der linearen Abhängigkeit gibt es

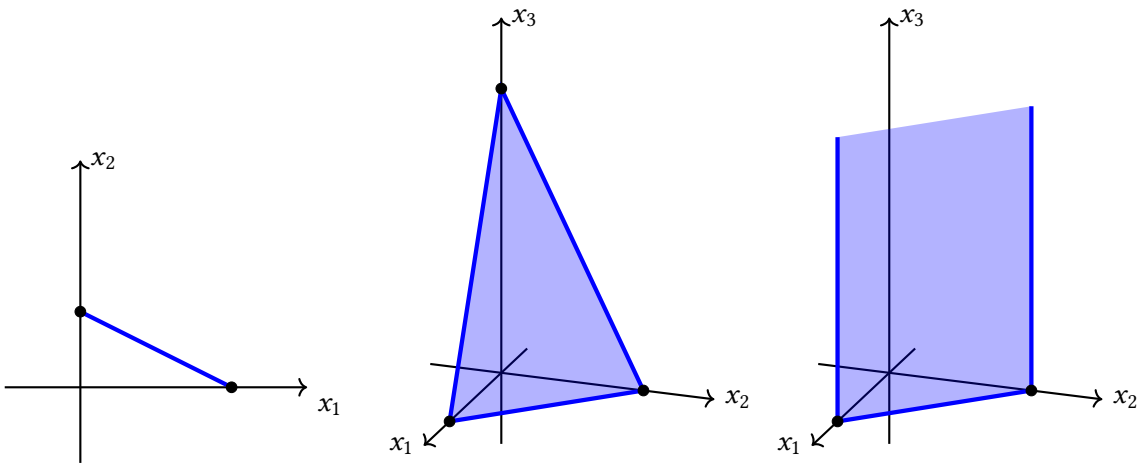


Abbildung 6.6: Polyeder in Normalform und ihre Ecken: Strecke im $\mathbb{R}^2$ ($n = 2$, $m = 1$) und Flächen im $\mathbb{R}^3$ ($n = 3$, $m = 1$).

Koeffizienten γ_i, sodass gilt:

$$\sum_{i \in \mathcal{I}(x)} \gamma_i \, a_i = 0,$$

und mindestens ein γ_i ist $\neq 0$. Wegen $x_i > 0$ für alle $i \in \mathcal{I}(x)$ existiert $\delta > 0$, sodass $x_i \pm \delta \, \gamma_i \geq 0$ bleibt für alle $i \in \mathcal{I}(x)$. Wir definieren nun Punkte $y, z \in \mathbb{R}^n$ durch

$$y_i = \begin{Bmatrix} x_i + \delta \, \gamma_i, & \text{falls } i \in \mathcal{I}(x) \\ 0 & \text{sonst} \end{Bmatrix} \quad \text{und} \quad z_i = \begin{Bmatrix} x_i - \delta \, \gamma_i, & \text{falls } i \in \mathcal{I}(x) \\ 0 & \text{sonst.} \end{Bmatrix}$$

Damit ist $y \neq z$, und es gilt $y, z \geq 0$ sowie

$$Ay = \sum_{i=1}^{n} y_i \, a_i = \sum_{i \in \mathcal{I}(x)} (x_i + \delta \, \gamma_i) \, a_i = b + \delta \sum_{i \in \mathcal{I}(x)} \gamma_i \, a_i = b,$$

also liegt $y \in P$. Ganz analog folgt auch $z \in P$. Dies ist aber ein Widerspruch zur Definition 6.12 einer Ecke, denn es gilt $x = \frac{y+z}{2}$ mit $y \neq z$.

„⇐“: Umgekehrt seien nun die Spaltenvektoren $(a_i)_{i \in \mathcal{I}(x)}$ linear unabhängig. (Möglicherweise ist $I(x) = \emptyset$.) Für zwei Vektoren $y, z \in P$ gelte $x = \alpha \, y + (1 - \alpha) \, z$ mit einem $\alpha \in (0, 1)$. Wir müssen $y = z$ zeigen. Für alle $j \notin \mathcal{I}(x)$ gilt $x_j = y_j = z_j = 0$ wegen $y, z \geq 0$. Also ist

$$0 = b - b = A(y - z) = \sum_{i \in \mathcal{I}(x)} (y_i - z_i) \, a_i,$$

und aus der linearen Unabhängigkeit der $(a_i)_{i \in \mathcal{I}(x)}$ folgt $y_i = z_i$ auch für $i \in \mathcal{I}(x)$. Insgesamt gilt also $y = z$, d. h., nach Definition 6.12 ist x eine Ecke von P. □

Beachte: Der Koordinatenvektor einer Ecke eines Polyeders in Normalform (6.8) muss mindestens $n - m$ Nulleinträge haben, da jeweils höchstens m Spalten von $A \in \mathbb{R}^{m \times n}$ linear unabhängig sind (siehe auch Abbildung 6.6).

Aus Satz 6.13 ergibt sich folgende Idee zur Generierung potentieller Ecken:

- Jeder Vektor $x \in P \subseteq \mathbb{R}^n$ wird durch n (linear unabhängige) Bedingungen an seine Koordinaten festgelegt.
- Wähle eine Indexmenge $N \subseteq \{1, 2, \ldots, n\}$ mit $|N| = n - m$ und setze $x_i = 0$ für $i \in N$.
- Die restlichen Indizes bilden die Menge $B = \{1, 2, \ldots, n\} \setminus N$ mit $|B| = m$.

Die Wahl von N erfolge so, dass der Punkt x durch die Bedingungen $A\,x = b$ und $x_i = 0$ für $i \in N$ eindeutig bestimmt ist. (Damit das möglich ist, muss man voraussetzen, dass $\operatorname{Rang}(A) = m$ gilt.) Die Spalten von A und die Komponenten von x werden so umsortiert und partitioniert, dass wir

$$A = [A_B\ A_N] \quad \text{und} \quad A\,x = A_B\,x_B + \underbrace{A_N x_N}_{=0} = A_B\,x_B = b$$

erhalten. Nun soll also $A_B\,x_B = b$ eindeutig lösbar sein, also muss A_B invertierbar (regulär) sein.

Definition 6.14 (Basisvektor, Basis)**.**
Es sei P wie in (6.8) ein Polyeder in Normalform. Weiter sei $B \subseteq \{1, \ldots, n\}$ mit $|B| = m$ eine (geordnete) Indexmenge (ein m-Tupel) von Spaltenindizes und $N = \{1, \ldots, n\} \setminus B$.

(*i*) *Ist die mit den Spaltenindizes B gebildete Untermatrix A_B regulär, so heißt die Indexmenge B eine* ***Basis*** *und A_B die zugehörige* ***Basismatrix****. N heißt dann* ***Nichtbasis*** *und A_N die zugehörige* ***Nichtbasismatrix****.*

(*ii*) *Es sei A_B eine Basismatrix. Ein Punkt $x \in \mathbb{R}^n$ heißt ein* ***Basisvektor*** *von P zur Basis B, wenn $A_B\,x_B = b$ und $x_N = 0$ gilt.*

(*iii*) *In der Literatur wird ein Basisvektor auch häufig als* ***Basislösung*** *bezeichnet. Der Begriff „-lösung“ weist darauf hin, dass der Vektor das lineare Gleichungssystem $A\,x = b$ löst. Dieser Praxis folgen wir hier nicht, um Verwechslungen mit Optimallösungen zu vermeiden.*

(*iv*) *Ein Basisvektor heißt* ***zulässig****, wenn $x_B \geq 0$ gilt.*

(*v*) *Ist x ein Basisvektor zur Basis B, dann heißen die Komponenten von x_B* ***abhängige Variable*** *und die Komponenten von x_N* ***unabhängige Variable****.*

Beachte: Damit überhaupt eine Basis existiert, muss notwendig A vollen Rang haben, also $\operatorname{Rang}(A) = m$ gelten. Dies kann zumindest theoretisch immer durch Streichen von Zeilen erreicht werden, wobei numerisch die Bestimmung des Ranges schwierig sein kann.

Beispiel 6.15 (Basisvektoren, vgl. Geiger, Kanzow, 2002, Beispiel 3.17 auf S.97)**.**

Es seien

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \textit{und} \quad b = \begin{pmatrix} 4 \\ 6 \\ 2 \\ 3 \end{pmatrix}$$

gegeben. Der Vektor $x = (2, 0, 0, 2, 6, 0, 3)^\mathsf{T}$ *ist zulässiger Basisvektor zur Basis* $B = \{1, 4, 5, 7\}$*, denn: Die Untermatrix*

$$A_B = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

ist regulär (d. h., B ist Basis), und es gilt $x_B = (2, 2, 6, 3)^\mathsf{T} \geq 0$, $x_N = (0, 0, 0)^\mathsf{T}$ *sowie* $A_B\, x_B = b$*. Ein anderer zulässiger Basisvektor, dieses Mal zur Basis* $B = \{4, 5, 6, 7\}$*, ist* $x_B = b$*, da* $b \geq 0$ *ist.*

Satz 6.16 (Zusammenhang zwischen Ecken und zulässigen Basisvektoren)**.**
Es sei P wie in (6.8) *ein Polyeder in Normalform, und es gelte* $\operatorname{Rang}(A) = m$*. Dann sind äquivalent:*

(*i*) $x \in \mathbb{R}^n$ *ist eine Ecke von P.*

(*ii*) $x \in \mathbb{R}^n$ *ist zulässiger Basisvektor von P zu einer geeigneten Basis.*

Beachte: Eine Ecke kann mehrere Darstellungen als zulässiger Basisvektor zu verschiedenen Basen besitzen.

Satz 6.17 (Hauptsatz der linearen Optimierung, vgl. Geiger, Kanzow, 2002, Satz 3.6)**.**
Es sei P wie in (6.8) *ein Polyeder in Normalform, und es gelte* $\operatorname{Rang}(A) = m$*. Dann gilt:*

(*i*) *Ist* $P \neq \emptyset$*, dann besitzt P mindestens einen zulässigen Basisvektor (eine Ecke).*

(*ii*) *P hat nur endlich viele zulässige Basisvektoren (Ecken).*

(*iii*) *Besitzt das Problem*

$$\textit{Minimiere} \quad c^\mathsf{T} x \quad \textit{sodass}\, x \in P$$

eine Lösung, so ist auch einer der zulässigen Basisvektoren von P eine Lösung.

Die Aussage (*iii*) bedeutet, dass die Lösungsmenge eines LPs unter den obigen Voraussetzungen entweder leer ist oder mindestens eine Ecke enthält.

Beweis. Aussage (*i*): Zunächst stellen wir fest, dass es mindestens eine Basis gibt, da $\operatorname{Rang}(A) = m$ gilt. Gehört der Nullvektor zu P, dann ist er ein zulässiger Basisvektor zu jeder Basis. Andernfalls wählen wir ein $x^* \in P$ mit der minimalen Anzahl positiver Komponenten. Die Indexmenge $\mathcal{I}(x^*) = \{1 \leq i \leq n \mid x_i^* > 0\}$ ist nicht leer. Wir zeigen, dass die Spaltenvektoren (a_i), $i \in \mathcal{I}(x^*)$ linear unabhängig sind.

Nach Satz 6.13 ist dann x^* eine Ecke, und nach Satz 6.16 auch ein zulässiger Basisvektor von P (zu einer geeigneten Basis, die durch Auffüllen von $\mathcal{I}(x^*)$ entsteht).

Wir führen einen Widerspruchsbeweis und nehmen an, die Spaltenvektoren (a_i), $i \in \mathcal{I}(x^*)$ seien linear abhängig, also gilt

$$\sum_{i \in \mathcal{I}(x^*)} \gamma_i a_i = 0,$$

und o. B. d. A. ist mindestens ein $\gamma_i < 0$. Wegen $x_i^* > 0$ für alle $i \in \mathcal{I}(x^*)$ können wir wie im Beweis von Lemma 6.10 $\delta = \min \left\{ -\frac{x_i^*}{\gamma_i} \,\middle|\, \gamma_i < 0,\ i \in \mathcal{I}(x^*) \right\} > 0$ wählen. Daraus folgt, dass

$$x_i^* + \delta \gamma_i \geq 0 \quad \text{für alle } i \in \mathcal{I}(x^*)$$

ist und mindestens einmal Gleichheit gilt. Der Vektor

$$\bar{x} = \begin{cases} x_i^* + \delta \gamma_i, & \text{falls } i \in \mathcal{I}(x^*) \\ 0 & \text{sonst} \end{cases}$$

gehört dann zu P (Beweis wie in Satz 6.13), hat aber weniger positive Komponenten als x^*, im Widerspruch zur Voraussetzung.

Aussage *(ii)*: Es gibt nur endlich viele, nämlich höchstens $\binom{n}{m}$ Möglichkeiten, eine Basis, d. h. m linear unabhängige Spalten von A auszuwählen.[3] Zu jeder Basis gehört nur genau ein Basisvektor (der auch unzulässig sein kann).

Aussage *(iii)*: Nach Voraussetzung ist der Optimalwert

$$f^* = \inf\{c^\mathsf{T} x \mid x \in P\}$$

endlich und wird auch angenommen. Wir betrachten nun das LP mit der modifizierten zulässigen Menge

$$\begin{aligned} \text{Minimiere} \quad & c^\mathsf{T} x \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} \quad & x \in \widehat{P} = \{x \in \mathbb{R}^n \mid A x = b,\ x \geq 0,\ c^\mathsf{T} x = f^*\}. \end{aligned}$$

Nach Voraussetzung ist auch $\widehat{P} \neq \emptyset$, und $\widehat{P}$ ist wieder ein Polyeder in Normalform (es ist einfach eine Zeile in A und b hinzugekommen). Ist nun $\widehat{P} = P$, also die Zielfunktion konstant auf P, so sind insbesondere alle zulässigen Basisvektoren von P Lösung.

Ist dagegen $\widehat{P} \subsetneq P$, so gilt Rang $\binom{A}{c^\mathsf{T}} = m + 1$.[4] Nach Aussage *(i)* besitzt $\widehat{P}$ mindestens einen zulässigen Basisvektor x^*, der nach Satz 6.16 eine Ecke von $\widehat{P}$ ist. Es bleibt noch zu zeigen, dass x^* auch Ecke von P ist. Es seien also $y, z \in P$ und $\alpha \in (0,1)$, sodass $x^* = \alpha y + (1-\alpha) z$ gilt.

$$f^* \overset{x^* \in \widehat{P}}{=} c^\mathsf{T} x^* = \underbrace{\alpha c^\mathsf{T} y}_{\geq f^*} + \underbrace{(1-\alpha) c^\mathsf{T} z}_{\geq f^*} \geq \alpha f^* + (1-\alpha) f^* = f^*.$$

Also gilt $c^\mathsf{T} y = c^\mathsf{T} z = f^*$, d. h., $y, z \in \widehat{P}$. Da x^* eine Ecke von $\widehat{P}$ ist, muss $y = z$ gelten. Damit ist x^* eine Ecke von P und nach Satz 6.16 auch zulässiger Basisvektor von P, und wegen $x^* \in \widehat{P}$ ist x^* eine Lösung des LP. □

[3] Hierbei ignorieren wir die Anordnung der Basiselemente, da sie keinen Einfluss auf den zugehörigen Basisvektor hat.
[4] Zu den Gleichungen $A x = b$ ist eine neue Zeile dazukommen, die wesentlich ist.

Ende der Woche 4

§ 7 Simplex-Algorithmus

Literatur: Geiger, Kanzow, 2002, Kapitel 3.2–3.4

Idee des Simplex-Algorithmus': Laufe von einem zulässigen Basisvektor (Ecke) zu einem benachbarten, bis es keinen besseren (mit kleinerem Funktionswert) mehr gibt. Dabei heißen zwei Basisvektoren **benachbart**, wenn sich die zugehörigen Basen in genau einem Index unterscheiden.

Im gesamten § 7 sei P wie in (6.8) ein Polyeder in Normalform, und es gelte $\operatorname{Rang}(A) = m$.

§ 7.1 Der Simplex-Schritt

Literatur: Geiger, Kanzow, 2002, Kapitel 3.2

Es sei x irgendein (zulässiger) Basisvektor von P (zur Konstruktion siehe § 7.2) zur Basis B, und es sei $N = \{1, \dots, n\} \setminus B$. Die Spalten von A und die Komponenten von x und c seien entsprechend partitioniert. Um zu einer benachbarten Ecke zu gelangen, müssen wir einem Index $r \in N$ erlauben, sich von der Null zu lösen, während die anderen Nichtbasis-Einträge bei Null verbleiben. Wir machen also den Ansatz

$$x_r(t) := t \geq 0, \quad x_j(t) := 0 \text{ für alle } j \in N \setminus \{r\}$$

oder kurz: $x_N(t) = t\, e_r$ mit einem Standard-Basisvektor $e_r \in \mathbb{R}^{n-m}$, $t \geq 0$. Die Basis-Einträge $x_B(t)$ berechnen wir in Abhängigkeit von $x_N(t)$ aus dem linearen Gleichungssystem

$$A_B x_B(t) + A_N x_N(t) = b \quad \Leftrightarrow \quad x_B(t) = A_B^{-1}(b - t\, A_N e_r) = x_B + t \underbrace{(-A_B^{-1} a_r)}_{=:\Delta x_B}. \tag{7.1}$$

Hierbei ist a_r die r-te Spalte von A, und Δx_B bezeichnet die Richtung der Änderungen der x_B-Komponenten.

Durch Einsetzen von (7.1) erhalten wir folgende Darstellungen der Werte der Zielfunktion in Abhängigkeit von $t \geq 0$:

$$\begin{aligned} c^\mathsf{T} x(t) &= c_B^\mathsf{T} x_B(t) + c_N^\mathsf{T} x_N(t) \\ &= c_B^\mathsf{T} x_B + t\, c_B^\mathsf{T} \Delta x_B + t\, c_N^\mathsf{T} e_r \\ &= c^\mathsf{T} x + t \begin{pmatrix} c_B \\ c_N \end{pmatrix}^\mathsf{T} \begin{pmatrix} \Delta x_B \\ e_r \end{pmatrix} \end{aligned} \tag{7.2}$$

und

$$\begin{aligned} c^\mathsf{T} x(t) &= c_B^\mathsf{T} x_B(t) + c_N^\mathsf{T} x_N(t) \\ &= c_B^\mathsf{T} A_B^{-1}(b - t\, A_N e_r) + t\, c_N^\mathsf{T} e_r \\ &= c^\mathsf{T} x + t \underbrace{\left(c_N - A_N^\mathsf{T} A_B^{-\mathsf{T}} c_B\right)^\mathsf{T}}_{=:\widetilde{c}_N} e_r \\ &= c^\mathsf{T} x + t\, \widetilde{c}_r. \end{aligned} \tag{7.3}$$

Die Größe $\widetilde{c}_N$ bezeichnet man als den Vektor der **reduzierten Kosten**. Er ist durch die Daten der Aufgabe sowie durch die aktuelle Basis eindeutig bestimmt. Er erlaubt es uns, zu erkennen, wenn der gegenwärtige Basisvektor bereits ein Minimierer ist.

Lemma 7.1 (Erkennen einer Lösung)**.**
Es sei x ein zulässiger Basisvektor zur Basis B. Wenn für die reduzierten Kosten

$$\widetilde{c}_N := c_N - A_N^\mathsf{T} A_B^{-\mathsf{T}} c_B \geq 0 \tag{7.4}$$

gilt, dann ist x eine Lösung des LP (6.6).

Beweis. Es sei z ein beliebiger für (6.6) zulässiger Vektor (nicht notwendig ein Basisvektor). Dennoch partitionieren wir z ebenso wie x. Wir vergleichen die Funktionswerte $c^\mathsf{T} x$ und $c^\mathsf{T} z$ mit einer Rechnung ähnlich wie in (7.3):

$$\begin{aligned} c^\mathsf{T} z &= c_B^\mathsf{T} z_B + c_N^\mathsf{T} z_N \\ &= c_B^\mathsf{T} A_B^{-1}(b - A_N z_N) + c_N^\mathsf{T} z_N \\ &= c^\mathsf{T} x + \left(c_N - A_N^\mathsf{T} A_B^{-\mathsf{T}} c_B\right)^\mathsf{T} z_N \\ &= c^\mathsf{T} x + \widetilde{c}_N^\mathsf{T} z_N. \end{aligned}$$

Da $z_N \geq 0$ ist, gilt $c^\mathsf{T} z \geq c^\mathsf{T} x$, d. h., x ist ein Minimierer der Aufgabe (6.6). □

Quizfrage: Was vermuten Sie, gilt auch die Umkehrung von Lemma 7.1?

Wir gehen für die weitere Herleitung des Simplex-Schrittes also jetzt davon aus, dass $\widetilde{c}_N$ noch nicht in allen Einträgen ≥ 0 ist. Welche benachbarte Ecke soll das Verfahren dann wählen? Auch darüber gibt der Vektor der reduzierten Kosten Aufschluss. Damit die Zielfunktion fällt, wählen wir einen Index $r \in N$ aus, für den $\widetilde{c}_r < 0$ ist, denn wegen (7.2) fallen dann die Werte proportional zu $t \geq 0$. Diese Auswahlentscheidung nennt man auch „**pricing**“.

Es ergibt sich die Frage, wie groß t werden darf, sodass $x_B(t)$ noch zulässig, also $x_B(t) \geq 0$ bleibt. Die Darstellung (7.1)

$$x_B(t) = x_B + t\, \Delta x_B$$

liefert darüber Aufschluss.

Lemma 7.2 (Erkennen eines unbeschränkten LPs)**.**
Gilt $\Delta x_B \geq 0$, so ist das LP (6.6) unbeschränkt, also nicht lösbar.

Beweis. Nach Konstruktion erfüllt $x(t)$ für alle $t \in \mathbb{R}$ die Bedingung $Ax(t) = b$. Nach Voraussetzung gilt außerdem $x_B(t) \geq 0$ für alle $t \geq 0$, d. h., $x(t)$ ist für alle $t \geq 0$ zulässig für (6.6).

Es gilt nach (7.1) und (7.2):

$$c^\mathsf{T} x(t) = c^\mathsf{T} x + t \begin{pmatrix} c_B \\ c_N \end{pmatrix}^\mathsf{T} \begin{pmatrix} \Delta x_B \\ e_r \end{pmatrix} = c^\mathsf{T} x + \underbrace{t\, \widetilde{c}_r}_{<0} \to -\infty \quad \text{für } t \to \infty.$$

□

Beachte: Das ist genau die Situation, die in Lemma 6.8 beschrieben wird: Die Richtung $d = \left(\begin{smallmatrix} \Delta x_B \\ e_r \end{smallmatrix}\right)$ ist im Rezessionskegel der zulässigen Menge von (6.6) und ist eine Abstiegsrichtung für die Zielfunktion.

Wir gehen für die weitere Beschreibung des Simplex-Schrittes also jetzt davon aus, dass $\Delta x_i < 0$ für mindestens ein $i \in B$ ist. Die Zulässigkeitsbedingung für $x(t)$ ist genau dann erfüllt, wenn

$$t \geq 0 \quad \text{und} \quad x_B(t) = x_B + t\, \Delta x_B \geq 0$$

gilt oder äquivalent dazu:

$$0 \leq t \leq -\frac{x_i}{\Delta x_i} \text{ für alle } i \in B \text{ mit } \Delta x_i < 0.$$

Um mit $x_B(t)$ einen neuen zulässigen *Basisvektor* zu erhalten, muss eine Komponente von B nach N wechseln, denn r wechselt ja von N nach B. Wir wählen deshalb die größtmögliche Schrittlänge:

$$\hat{t} := \min \left\{ -\frac{x_i}{\Delta x_i} \,\middle|\, i \in B,\ \Delta x_i < 0 \right\} = -\frac{x_\ell}{\Delta x_\ell} \quad \text{„**Quotiententest**“ (englisch: *ratio test*).} \tag{7.5}$$

Es ist also ℓ der Index bzw. einer der Indizes, an denen das Minimum angenommen wird. Damit wird dann $x_\ell(\hat{t}) = 0$ sein, und wir nehmen den Index ℓ in die neue Nichtbasis auf.

Wir fassen zusammen: Als **Simplex-Schritt** bezeichnet man, ausgehend von der gegebenen Basis B und dem zugehörigen zulässigen Basisvektor x:

(*i*) die Berechnung der reduzierten Kosten $\widetilde{c}_N$ nach (7.4) (lineares Gleichungssystem mit A_B^T lösen),

(*ii*) die Auswahl eines Index' $r \in N$ mit $\widetilde{c}_r < 0$,

(*iii*) die Bestimmung des Vektors Δx_B nach (7.1) (lineares Gleichungssystem mit A_B lösen) und der Schrittlänge $\hat{t}$ nach (7.5)

(*iv*) und die Bestimmung des neuen zulässigen Basisvektors $x^+ := x(\hat{t})$ und der geänderten Basis $B^+ := \big(B \cup \{r\}\big) \setminus \{\ell\}$ und Nichtbasis $N^+ := \big(N \cup \{\ell\}\big) \setminus \{r\}$.

Satz 7.3 (Simplex-Schritt)**.**
Es sei x ein zulässiger Basisvektor von P zur Basis B, und es sei $N = \{1, \ldots, n\} \setminus B$. Es gelte $\widetilde{c}_r < 0$ für mindestens ein $r \in N$, und es sei $\Delta x_B := -A_B^{-1} a_r$. Es gelte weiter $\Delta x_i < 0$ für mindestens ein $i \in B$. Wird

dann $\hat{t} \geq 0$ nach (7.5) bestimmt und wird das Minimum für den Index $\ell \in B$ angenommen, so gelten für den Vektor x^+ mit

$$x_i^+ := \begin{cases} x_i + \hat{t}\,\Delta x_i & \text{für } i \in B,\ i \neq \ell, \\ \hat{t} & \text{für } i = r, \\ 0 & \text{sonst} \end{cases}$$

die folgenden Aussagen:

(*i*) *Der Vektor x^+ ist zulässiger Basisvektor von P zur neuen Basis*

$$B^+ := \big(B \cup \{r\}\big) \setminus \{\ell\}.$$

(*ii*) *Für die Zielfunktionswerte gilt*

$$c^\mathsf{T} x^+ = c^\mathsf{T} x + \hat{t}\,\widetilde{c_r} \leq c^\mathsf{T} x.$$

Beweis. Für Aussage (*i*) müssen wir zeigen:

(a) $A_{B^+} x^+_{B^+} = b$,

(b) $x^+_{N^+} = 0$,

(c) $x^+_{B^+} \geq 0$ und

(d) A_{B^+} ist regulär.

Die Punkte (a) bis (c) folgen aus der Konstruktion von x^+. Wir weisen noch nach, dass die Spalten $(a_i)_{i \in B^+}$ linear unabhängig sind und machen dafür den Ansatz:

$$\begin{aligned} 0 &= \sum_{i \in B,\ i \neq \ell} \gamma_i\, a_i + \gamma_r\, a_r \\ &= \sum_{i \in B,\ i \neq \ell} \gamma_i\, a_i - \gamma_r A_B\, \Delta x_B \\ &= \sum_{i \in B,\ i \neq \ell} \gamma_i\, a_i - \gamma_r \Big(\sum_{i \in B} \Delta x_i a_i \Big) \\ &= \sum_{i \in B,\ i \neq \ell} (\gamma_i - \gamma_r\, \Delta x_i)\, a_i - \gamma_r\, \Delta x_\ell a_\ell. \end{aligned}$$

Nach Voraussetzung waren die Spalten $(a_i)_{i \in B}$ linear unabhängig, also folgt

$$\gamma_i - \gamma_r\, \Delta x_i = 0 \quad \text{für alle } i \in B,\ i \neq \ell \quad \text{und} \quad \gamma_r\, \Delta x_\ell = 0.$$

Wegen $\Delta x_\ell < 0$ gilt $\gamma_r = 0$ und damit $\gamma_i = 0$ für alle $i \in B$, $i \neq \ell$.

Die Aussage (*ii*) folgt aus (7.3). □

Bemerkung 7.4 (Der Fall $c^\mathsf{T} x^+ = c^\mathsf{T} x$).
Es gelten die Voraussetzungen von Satz 7.3.

(*i*) *Der Fall $c^\mathsf{T} x^+ = c^\mathsf{T} x$ tritt genau dann auf, wenn sich im Quotiententest (7.5) $\hat{t} = 0$ ergibt, also auch $x^+ = x$ gilt. Es ändern sich also nur die Indexmenge $B \rightsquigarrow B^+$ und $N \rightsquigarrow N^+$. Dieselbe Ecke hat also eine Darstellung als Basisvektor zu verschiedenen Basen. Dazu muss allerdings notwendig*

$$x_i = 0 \quad \text{für mindestens ein } i \in B \tag{7.6}$$

gelten. Ein Basisvektor x, für den (7.6) zutrifft, heißt **entartet**.[5]

(*ii*) *Ist x dagegen ein nicht entarteter Basisvektor, so gilt unter den Voraussetzungen von Satz 7.3 immer $\hat{t} > 0$ und daher*

$$c^\mathsf{T} x^+ = c^\mathsf{T} x + \hat{t}\, \widetilde{c}_r < c^\mathsf{T} x.$$

Der Zielfunktionswert nimmt dann also strikt ab.

Beispiel 7.5 (Nochmal Beispiel 6.15).
Wir führen einen Simplex-Schritt für Beispiel 6.15 durch, ausgehend vom (zulässigen) Basisvektor $x = (2,0,0,2,6,0,3)^\mathsf{T}$ zur Basis $B = \{1,4,5,7\}$. Der Zielfunktionswert ist $c^\mathsf{T} x = (-2,-3,-4,0,0,0,0)\, x = -4$.

(*i*) *Die reduzierten Kosten sind*

$$\begin{aligned}
\widetilde{c}_N &= c_N - A_N^\mathsf{T} A_B^{-\mathsf{T}} c_B \\
&= \begin{pmatrix} -3 \\ -4 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & 3 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} -2 \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{=(0,0,-2,0)^\mathsf{T}} \\
&= \begin{pmatrix} -3 \\ -4 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & 3 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ -2 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} -3 \\ -4 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix} = \begin{pmatrix} -3 \\ -4 \\ 2 \end{pmatrix} =: \begin{pmatrix} \widetilde{c}_2 \\ \widetilde{c}_3 \\ \widetilde{c}_6 \end{pmatrix}
\end{aligned}$$

(*ii*) *Wir wählen einen Index $r \in N = \{2,3,6\}$ mit $\widetilde{c}_r < 0$ aus, hier $r = 3$ (Alternative: $r = 2$).*

(*iii*) *Wir berechnen*

$$\Delta x_B = -A_B^{-1} a_r = -\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ -1 \\ -1 \end{pmatrix} =: \begin{pmatrix} \Delta x_1 \\ \Delta x_4 \\ \Delta x_5 \\ \Delta x_7 \end{pmatrix}$$

[5] Da bei der Bestimmung von $\hat{t}$ jedoch nicht alle Basis-Indizes mitspielen, sondern nur diejenigen mit $\Delta x_i < 0$, ist auch bei einem entarteten Basisvektor durchaus $\hat{t} > 0$ möglich.

und führen den Quotiententest durch:

$$\hat{t} := \min\left\{-\frac{x_i}{\Delta x_i} \,\middle|\, i \in B,\ \Delta x_i < 0\right\} = \min\Big\{\underbrace{\frac{2}{1}}_{i=4}, \underbrace{\frac{6}{1}}_{i=5}, \underbrace{\frac{3}{1}}_{i=7}\Big\}.$$

Beachte: *$\Delta x_1 = 0$ nimmt an der Minimumbildung nicht teil! Das Minimum $\hat{t} = 2$ wird eindeutig beim Index $\ell = 4$ angenommen.*

(*iv*) *Die neue Basis ist also $B^+ = \big(B \cup \{r\}\big) \setminus \{\ell\} = \{1, 3, 5, 7\}$ und die Nichtbasis $N^+ = \{2, 4, 6\}$. Neuer Basisvektor ist*

$$x^+ = \begin{pmatrix} 2 + \hat{t}\,\Delta x_1 & \text{bleibt in } B^+ \\ 0 & \text{bleibt in } N^+ \\ 0 + \hat{t} & \text{wechselt in } B^+ \\ 2 + \hat{t}\,\Delta x_4 = 0 & \text{wechselt in } N^+ \\ 6 + \hat{t}\,\Delta x_5 & \text{bleibt in } B^+ \\ 0 & \text{bleibt in } N^+ \\ 3 + \hat{t}\,\Delta x_7 & \text{bleibt in } B^+ \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 2 \\ 0 \\ 4 \\ 0 \\ 1 \end{pmatrix}$$

mit neuem Funktionswert $c^\mathsf{T} x^+ = -12$. Wie erwartet hat sich der Funktionswert also um $\hat{t}\,\widetilde{c}_r = 2\,(-4) = -8$ verändert.

Würden wir in Schritt (ii) stattdessen den Index $r = 2$ wählen, so erhielten wir $\Delta x_B = (0, -1, -3, 0)^\mathsf{T}$ und dann in Schritt (iii) im Quotiententest $\hat{t} = 2$ und $\ell = 4$ oder $\ell = 5$. Dies würde dazu führen, dass in jedem Fall beide Koordinaten $x_4^+ = x_5^+ = 0$ werden, d. h., x^+ ist dann ein entarteter Basisvektor. Wir erhielten dann in Schritt (iv) $B^+ = \{1, 2, 5, 7\}$ und $N^+ = \{3, 4, 6\}$ oder $B^+ = \{1, 2, 4, 7\}$ und $N^+ = \{3, 5, 6\}$ und in beiden Fällen $x^+ = (2, 2, 0, 0, 0, 0, 3)^\mathsf{T}$ mit neuem Funktionswert $c^\mathsf{T} x^+ = -10$.

§ 7.2 Der Simplex-Algorithmus

Literatur: Geiger, Kanzow, 2002, Kapitel 3.3–3.4

Wir geben jetzt den kompletten Simplex-Algorithmus zur Lösung des LP (6.6) in Normalform mit $\operatorname{Rang}(A) = m$ an. Der leichteren Lesbarkeit wegen verzichten wir darauf, die Iterierten nach dem Iterationszähler k zu benennen.

Algorithmus 7.6 (Simplex-Algorithmus (Dantzig 1947))**.**
Eingabe: *Aufgabenbeschreibung durch A, b und c*
Eingabe: *zulässiger Basisvektor x von P mit zugehöriger Basis B und Nichtbasis N*
Ausgabe: *ein optimaler Basisvektor von (6.6) oder die Aussage, dass (6.6) unbeschränkt ist*

1: *Setze $k := 0$*
2: *Berechne die reduzierten Kosten*

$$\widetilde{c}_N := c_N - A_N^\mathsf{T} A_B^{-\mathsf{T}} c_B$$

3: ***if*** $\widetilde{c}_N \geq 0$ ***then***

4: x *ist eine Lösung von* (6.6), ***STOP***
5: ***else***
6: *Wähle einen Index* $r \in N$ *mit* $\widetilde{c_r} < 0$
7: *Berechne* $\Delta x_B := -A_B^{-1} a_r$
8: ***if*** $\Delta x_B \geq 0$ ***then***
9: *Aufgabe* (6.6) *ist unbeschränkt,* ***STOP***
10: ***else***
11: *Bestimme* $\hat{t} \geq 0$ *und* $\ell \in B$ *gemäß*

$$\hat{t} := \min\left\{ -\frac{x_i}{\Delta x_i} \,\middle|\, i \in B,\ \Delta x_i < 0 \right\} = -\frac{x_\ell}{\Delta x_\ell}$$

12: *Setze*

$$x_i^+ := \begin{cases} x_i + \hat{t}\,\Delta x_i & \text{für } i \in B,\ i \neq \ell, \\ \hat{t} & \text{für } i = r, \\ 0 & \text{sonst} \end{cases}$$

13: *Setze* $B^+ := \big(B \cup \{r\}\big) \setminus \{\ell\}$
14: *Setze* $N^+ := \big(N \cup \{\ell\}\big) \setminus \{r\}$
15: *Setze* $x := x^+$
16: *Setze* $B := B^+$ *und* $N := N^+$
17: *Setze* $k := k + 1$
18: ***end if***
19: ***end if***
20: *Gehe zu Zeile 2*

Quizfrage: Bei der Herleitung des Simplex-Verfahrens bedeuteten A_B und A_N sowie x_B und x_N immer eine Auswahl von Spalten von A bzw. von Einträgen in x. Es sind also $x_B \in \mathbb{R}^m$ und $x_N \in \mathbb{R}^{n-m}$ „kurze“ Vektoren und $A_B \in \mathbb{R}^{m\times m}$ und $A_N \in \mathbb{R}^{m\times(n-m)}$ „schmale“ Matrizen. Würde man das auch in dieser Form z. B. in Python implementieren? Wo könnte ein Nachteil liegen?

Wir können einen vorläufigen Konvergenzsatz für das Simplex-Verfahren angeben, der allerdings die nicht vorab überprüfbare Voraussetzung verwendet, dass im Verlauf keine entarteten Basisvektoren auftreten.

Satz 7.7 (Endlichkeit des Simplex-Verfahrens)**.**
Sind alle im Simplex-Verfahren auftretenden Basisvektoren nicht entartet, so bricht das Verfahren nach endlich vielen Iterationen ab, und zwar entweder mit einem optimalen Basisvektor (Ecke) von (6.6) *oder mit der Feststellung, dass* (6.6) *unbeschränkt ist.*

Beweis. Nach Bemerkung 7.4 Punkt (*ii*) gilt $c^\mathsf{T} x^+ < c^\mathsf{T} x$ für alle Iterierten. Daher kann kein Basisvektor mehrfach im Verfahren auftreten. Da es nach Satz 6.17 nur endlich viele zulässige Basisvektoren gibt, muss das Verfahren in Zeile 4 oder in Zeile 9 abbrechen. □

Der Simplex-Algorithmus 7.6 lässt noch Freiheiten

- bei der Wahl der Austausch-Indizes r in Zeile 6
- und evtl. bei der Wahl von ℓ in Zeile 11,

vgl. Beispiel 7.5. Durch geeignete Zusatzregeln kann man erreichen, dass das Verfahren auch bei Vorkommen entarteter Basisvektoren immer terminiert.

Dabei geht es um die Vermeidung von Zyklen, d. h. Situationen, in denen

$$x^{(k)} = x^{(k+1)} = \ldots = x^{(k+p)}$$

und

$$B^{(k)} \rightsquigarrow B^{(k+1)} \rightsquigarrow \ldots \rightsquigarrow B^{(k+p)} = B^{(k)}$$

gilt.

Satz 7.8 (Regel von Bland)**.**
Wählt man in Zeile 6 den Index r und in Zeile 11 den Index ℓ als den jeweils kleinsten *in Frage kommenden Index, dann bricht der Simplex-Algorithmus 7.6 stets nach endlich vielen Iterationen ab, und zwar entweder mit einer Lösung von (6.6) oder mit der Feststellung, dass (6.6) unbeschränkt ist.*

Beweis. Mit der Zusatzregel von Bland kann man zeigen, dass keine Zyklen mehr auftreten, siehe Geiger, Kanzow, 2002, Satz 3.27. □

Bemerkung 7.9 (Alternativer Beweis von Satz 6.9)**.**
Der Simplex-Algorithmus in Verbindung mit der Regel von Bland bietet eine konstruktive Möglichkeit, den Existenzsatz 6.9 zu beweisen.

Quizfrage: Angenommen, das Simplex-Verfahren hat einen optimalen Basisvektor x^* gefunden, es gibt aber noch weitere optimale Basisvektoren. Wie kann man das Simplex-Verfahren dazu benutzen, ausgehend von x^* einen weiteren optimalen Basisvektor zu bestimmen?

Quizfrage: Was könnte der Grund sein, warum man im Simplex-Verfahren mit benachbarten Ecken arbeitet? Man könnte doch auch größere Änderungen in den Basis-Indizes zulassen?

Quizfrage: Ist das Simplex-Verfahren ein Abstiegsverfahren? Wenn ja, können Sie die Schritte Schritte (1) bis (3) eines allgemeinen Abstiegsverfahrens (siehe Anfang von § 4 auf Seite 15) im Simplexverfahren (Algorithmus 7.6) wiederfinden?

Finden der ersten Ecke

Beachte: Für den Start des Simplex-Algorithmus 7.6 muss ein zulässiger Basisvektor von P bekannt sein.

Beobachtung: War das LP ursprünglich in kanonischer Form (6.3) gegeben (etwa beim Mozartproblem, Beispiel 6.2), also in der Form

$$\left.\begin{aligned} \text{Maximiere} \quad & c^\mathsf{T} x \\ \text{sodass} \quad & Ax \le b \\ \text{und} \quad & x \ge 0 \end{aligned}\right\}$$

mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$, und führen wir Schlupfvariablen $s \in \mathbb{R}^m$ ein, so erhalten wir das äquivalente Problem in Normalform mit den Variablen $(x, s)^\mathsf{T} \in \mathbb{R}^{n+m}$:

$$\begin{aligned} \text{Minimiere} \quad & -c^\mathsf{T} x \\ \text{sodass} \quad & Ax + s = b \\ \text{und} \quad & x \ge 0, \quad s \ge 0. \end{aligned}$$

Falls $b \ge 0$ ist, dann ist $(0, b)^\mathsf{T}$ ein zulässiger Basisvektor zur Basis $B = \{n+1, \dots, n+m\}$, mit dem man das Verfahren starten kann.

Im Allgemeinen kann man einen zulässigen Basisvektor für (6.6) durch Lösen eines Hilfsproblems („**Phase I**“) bestimmen:

Satz 7.10 (Phase-I-Problem)**.**
In dem LP in Normalform (6.6) sei (o. B. d. A.) $b \ge 0$.[6] Dann gelten für das Hilfsproblem

$$\left.\begin{aligned} \textit{Minimiere} \quad & \mathbf{1}^\mathsf{T} z \quad \textit{über } (x, z) \in \mathbb{R}^n \times \mathbb{R}^m \\ \textit{sodass} \quad & Ax + z = b \\ \textit{und} \quad & x \ge 0, \quad z \ge 0 \end{aligned}\right\} \tag{7.7}$$

mit $\mathbf{1} = (1, 1, \dots, 1)^\mathsf{T} \in \mathbb{R}^m$ folgende Aussagen:

(*i*) *Der Vektor $(x, z)^\mathsf{T} = (0, b)^\mathsf{T}$ ist ein zulässiger Basisvektor für (7.7) zur Basis $B = \{n+1, \dots, n+m\}$.*

(*ii*) *Das LP (7.7) besitzt eine Lösung.*

(*iii*) *Es sei $(x^*, z^*)^\mathsf{T}$ ein optimaler Basisvektor für (7.7). Ist $z^* \ne 0$, so besitzt das LP (6.6) keinen zulässigen Punkt. Ist dagegen $z^* = 0$ und gilt $\operatorname{Rang}(A) = m$, so ist x^* ein (zulässiger) Basisvektor für (6.6) zu einer geeigneten Basis.*

Die Voraussetzung $b \ge 0$ ist keine Einschränkung, ggf. multiplizieren wir betreffende Zeilen mit -1.

Beweis. Aussage (*i*) folgt sofort aus der Definition eines Basisvektors, da die zugehörigen Spalten von $[A, \mathrm{Id}]$ gerade die Einheitsmatrix Id bilden. Damit ist das Hilfsproblem (7.7) nicht unzulässig.

Aussage (*ii*): Wegen $z \ge 0$ ist die Zielfunktion $\mathbf{1}^\mathsf{T} z = \sum_{i=1}^m z_i$ über der zulässigen Menge selbst ≥ 0, d. h., (7.7) ist nicht unbeschränkt. Aus Satz 6.9 folgt die Existenz einer Lösung.

[6] Über den Rang von A muss hier nichts vorausgesetzt werden. Der Rang von $[A, \mathrm{Id}]$ ist immer gleich m.

Aussage (*iii*): Es sei $(x^*, z^*)^\mathsf{T}$ ein optimaler Basisvektor für (7.7) und zunächst $z^* \neq 0$. Der Optimalwert von (7.7) ist daher $\mathbf{1}^\mathsf{T} z^* > 0$. Gäbe es einen zulässigen Punkt $\bar{x}$ von (6.6), so wäre $(\bar{x}, 0)^\mathsf{T}$ zulässig für (7.7) mit Funktionswert 0, im Widerspruch zur Optimalität von $(x^*, z^*)^\mathsf{T}$.

Wir betrachten nun den Fall $z^* = 0$. Es sei B^* mit $|B^*| = m$ eine zu $(x^*, z^*)^\mathsf{T}$ gehörige Basis. Es ist also $[A, \mathrm{Id}]_{B^*}$ regulär. Nach Definition gehören positive Komponenten von x^* notwendig zu B^*, sodass die zugehörigen Spalten von A linear unabhängig sind. Falls erforderlich, können diese Spalten durch weitere Spalten von A zu m linear unabhängigen Spalten ergänzt werden, da $\mathrm{Rang}(A) = m$ vorausgesetzt wurde. Mit $Ax^* = b$ folgt hieraus, dass x^* ein (möglicherweise entarteter) zulässiger Basisvektor für (6.6) ist. □

Bemerkung 7.11 (Zu Phase I und II)**.**

(*i*) *Das Hilfsproblem* (7.7) *ist wiederum ein LP in Normalform, dessen Matrix* $[A, \mathrm{Id}]$ *stets vollen Rang m hat.*

(*ii*) *Wir können das Simplex-Verfahren (Algorithmus 7.6) in der* ***Phase I*** *auf das Hilfsproblem* (7.7) *anwenden. Ein erster zulässiger Basisvektor ist nach Satz 7.10 (i) bekannt. Dann erhalten wir (wenn wir Zyklen mit der Regel von Bland vermeiden) im Fall* $\mathrm{Rang}(A) = m$ *nach endlich vielen Schritten entweder einen zulässigen Basisvektor für das eigentliche LP* (6.6) *oder die Information, dass* (6.6) *unzulässig ist (keinen zulässigen Punkt besitzt).*

(*iii*) *Ist der in Phase I berechnete Basisvektor* (x^*, z^*) *entartet, so enthält die Basis* B^* *möglicherweise noch Indizes in* $\{n+1, \ldots, n+m\}$*, die man vor dem Start des eigentlichen Simplex-Algorithmus („****Phase II****“) für* (6.6) *in zusätzlichen Schritten noch austauschen muss. Mehr Informationen dazu findet man zum Beispiel in Geiger, Kanzow, 2002, Aufgabe 3.22.*

(*iv*) *Für Phase I haben wir nicht benötigt, dass A vollen Rang hat, da* $[A, \mathrm{Id}]$ *in jedem Fall vollen Rang hat. Erhält man dann* $z^* = 0$ *und ist der Rang von A nicht maximal, so kann* x^* *unmöglich ein Basisvektor für* (6.6) *mit einer Basis in* $\{1, \ldots, n\}$ *sein. Aus Phase I erhält man dann aber Informationen darüber, welche Zeile(n) von* $Ax = b$ *gestrichen werden können. Details dazu können Sie zum Beispiel in Geiger, Kanzow, 2002, Aufgabe 3.23 finden.*

Quizfrage: Wieviele Iterationen benötigt das Simplex-Verfahren in Phase I mindestens, um einen zulässigen Basisvektor der Aufgabe (6.6) zu einer Basis in $\{1, \ldots, n\}$ zu finden?

Bemerkung 7.12 (Zur Komplexität des Simplex-Verfahrens)**.**
Das Simplexverfahren hat sich in der Praxis bei vielen Aufgabenstellungen bewährt. Es gibt jedoch ein konstruiertes Beispiel von Klee und Minty (1972)[7]*, bei dem* alle *Ecken eines Polyeders besucht werden, und zwar in jeder Problemdimension (Anzahl der Variablen) n. Da die Anzahl der Ecken exponentiell mit n wächst, ist das Simplex-Verfahren im schlechtesten Fall von der Laufzeit nicht polynomial in n. Dies ist eine Motivation für Innere-Punkte-Verfahren.*

Ende der Woche 5

[7] siehe z. B. Hamacher, Klamroth, 2006, S.81

§ 8 Optimalitätsbedingungen der linearen Optimierung (Dualität)

Literatur: Geiger, Kanzow, 2002, Kapitel 3.1.2

Wir betrachten weiterhin ein LP in Normalform, also

$$\left.\begin{array}{rll} \text{Minimiere} & c^\mathsf{T} x & \text{über } x \in \mathbb{R}^n \\ \text{sodass} & A\,x = b & \\ \text{und} & x \geq 0 & \end{array}\right\} \tag{8.1}$$

mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$.

Beachte: Über den Rang von A sowie die Dimensionen von $m, n \in \mathbb{N}$ wird in diesem Abschnitt nichts vorausgesetzt.

Eng verwandt mit (8.1) ist das folgende LP, das dieselben Daten (A, b, c) verwendet:

$$\left.\begin{array}{rll} \text{Maximiere} & b^\mathsf{T}\lambda & \text{über } \lambda \in \mathbb{R}^m \\ \text{sodass} & A^\mathsf{T}\lambda \leq c. & \end{array}\right\} \tag{8.2}$$

Führen wir in (8.2) die **dualen Schlupfvariablen** $s \in \mathbb{R}^n$ ein, so erhalten wir die zu (8.2) äquivalente und von uns bevorzugte Darstellung:

$$\left.\begin{array}{rll} \text{Maximiere} & b^\mathsf{T}\lambda & \text{über } (\lambda, s) \in \mathbb{R}^m \times \mathbb{R}^n \\ \text{sodass} & A^\mathsf{T}\lambda + s = c & \\ \text{und} & s \geq 0. & \end{array}\right\} \tag{8.3}$$

Achtung: Das LP (8.3) liegt nicht in Normalform vor, da die Bedingung $\lambda \geq 0$ fehlt (und die Zielfunktion maximiert wird).

Definition 8.1 (Duales LP)**.**
Das LP (8.2) bzw. (8.3) heißt das zu (8.1) gehörige ***duale LP****. In diesem Zusammenhang heißt (8.1) das* ***primale LP****. Man spricht auch von* ***primal-dualen Paaren****.*

Quizfrage: Was ist die duale Aufgabe der dualen Aufgabe (8.3)?

Ziel: Verständnis des Zusammenhangs von (8.1) und (8.3)

Wir bezeichnen wie bisher auch den Optimalwert von (8.1) mit f^* und den Optimalwert von (8.2) bzw. von (8.3) mit d^*:

$$\begin{aligned} f^* &:= \inf\{c^\mathsf{T}x \mid A\,x = b,\ x \geq 0\} \\ d^* &:= \sup\{b^\mathsf{T}\lambda \mid A^\mathsf{T}\lambda \leq c\} = \sup\{b^\mathsf{T}\lambda \mid A^\mathsf{T}\lambda + s = c,\ s \geq 0\}. \end{aligned}$$

Quizfrage: Welchen Wert hat d^*, wenn die duale Aufgabe unbeschränkt bzw. unzulässig ist?

Satz 8.2 (**Schwache Dualität**).
Es sei $x \in \mathbb{R}^n$ zulässig für das primale LP (8.1), und es sei $(\lambda, s) \in \mathbb{R}^m \times \mathbb{R}^n$ zulässig für das duale LP (8.3). Dann gilt für die Funktionswerte

$$b^\mathsf{T}\lambda \leq c^\mathsf{T}x.$$

Beachte: Schwache Dualität bedeutet also gerade: $d^* \leq f^*$.

Beweis. Aus der Zulässigkeit ergibt sich

$$b^\mathsf{T}\lambda = (Ax)^\mathsf{T}\lambda = x^\mathsf{T}(A^\mathsf{T}\lambda) = x^\mathsf{T}(c - s) = c^\mathsf{T}x - x^\mathsf{T}s \leq c^\mathsf{T}x, \tag{8.4}$$

denn wegen $x \geq 0$ und $s \geq 0$ gilt $x^\mathsf{T}s \geq 0$. □

Bemerkung 8.3 (Veranschaulichung des dualen LPs).
Jedes zulässige Paar (λ, s) des dualen LPs (8.3) liefert mit $b^\mathsf{T}\lambda$ eine untere Schranke für den optimalen Zielfunktionswert des primalen LPs, also

$$b^\mathsf{T}\lambda \leq f^* := \inf\{c^\mathsf{T}x \mid Ax = b,\ x \geq 0\}.$$

Wegen $Ax = b$ im primalen LP gilt auch $\lambda^\mathsf{T}Ax = \lambda^\mathsf{T}b$ für alle $\lambda \in \mathbb{R}^m$ (Linearkombination der Gleichungen). Im dualen LP suchen wir also eine Linearkombination der Gleichungsnebenbedingungen $Ax = b$ (repräsentiert durch den Koeffizientenvektor $\lambda \in \mathbb{R}^m$), die den Wert der primalen Zielfunktion am stärksten einschränkt.[8]

Folgerung 8.4 (Erkennen primal-dualer Lösungen).
Es sei $x^* \in \mathbb{R}^n$ zulässig für das primale LP (8.1), und es sei $(\lambda^*, s^*) \in \mathbb{R}^m \times \mathbb{R}^n$ zulässig für das duale LP (8.3). Wenn

$$c^\mathsf{T}x^* = b^\mathsf{T}\lambda^* \tag{8.5}$$

gilt, dann ist x^* eine Lösung des primalen LP, und (λ^*, s^*) ist eine Lösung des dualen LP.

Beweis. Es seien x und (λ, s) irgendwelche zulässigen Punkte für das primale bzw. das duale LP. Aus der schwachen Dualität (Satz 8.2) folgt

$$\underbrace{b^\mathsf{T}\lambda \leq \overbrace{c^\mathsf{T}x^* = b^\mathsf{T}\lambda^*}^{\text{primale Optimalität}}}_{\text{duale Optimalität}} \leq c^\mathsf{T}x,$$

d. h., x^* ist eine Lösung von (8.1), und (λ^*, s^*) ist Lösung von (8.3). □

[8] Eine analoge Aussage gilt natürlich auch für das primale LP. Die Sprechweise von primal-dualen Paaren ist daher gerechtfertigt.

Der folgende Satz zeigt, dass das System

$$\left.\begin{aligned} A^\top \lambda + s &= c, \quad & s &\geq 0 && \text{duale Zulässigkeit} \\ Ax &= b, & x &\geq 0 && \text{primale Zulässigkeit} \\ x_i\, s_i &= 0, & i &= 1, \dots, n && \text{Komplementarität} \end{aligned}\right\} \tag{8.6}$$

notwendige und hinreichende Optimalitätsbedingungen sind, und zwar gleichzeitig für das primale wie auch für das duale LP.

Beachte: Die Komplementaritätsbedingungen (englisch: *complementary slackness*) $x_i\, s_i = 0$ können äquivalent auch summiert formuliert werden:

$$x^\top s = \sum_{i=1}^{n} x_i\, s_i = 0.$$

Satz 8.5 (Notwendige und hinreichende Optimalitätsbedingungen)**.**

(*i*) *Ist x^* eine Lösung für das primale LP (8.1), dann existieren (λ^*, s^*), sodass (x^*, λ^*, s^*) das System (8.6) erfüllt.*

(*ii*) *Ist (λ^*, s^*) eine Lösung für das duale LP (8.3), dann existiert x^*, sodass (x^*, λ^*, s^*) das System (8.6) erfüllt.*

(*iii*) *Erfüllt (x^*, λ^*, s^*) das System (8.6), dann ist x^* eine Lösung von (8.1), und (λ^*, s^*) ist eine Lösung von (8.3).*

In jedem Fall sind die Optimalwerte gleich: $f^* = d^*$.

Für den Beweis der Aussagen (*i*) und (*ii*) benötigen wir folgendes Hilfsresultat.

Lemma 8.6 (**Farkas-Lemma** (1902))**.**
Es seien $B \in \mathbb{R}^{m\times n}$ und $c \in \mathbb{R}^n$. Dann sind äquivalent:

(*i*) *Das System $B^\top \xi = c$ besitzt eine Lösung $\xi \geq 0$.*

(*ii*) *Es gilt $c^\top d \geq 0$ für alle Elemente der Menge $\{d \in \mathbb{R}^n \mid B\, d \geq 0\}$.*

Aussage (*i*) bedeutet, dass c in der abgeschlossenen Menge

$$K := \{B^\top \xi \mid \xi \in \mathbb{R}^m,\ \xi \geq 0\}$$

liegt, vgl. Lemma 6.10. Um Aussage (*ii*) zu veranschaulichen, machen wir folgende Überlegung:

$$\begin{aligned} B\, d \geq 0 \quad &\Leftrightarrow \quad \xi^\top B\, d \geq 0 \quad \text{für alle } \xi \geq 0 \\ &\Leftrightarrow \quad (B^\top \xi)^\top d \geq 0 \quad \text{für alle } \xi \geq 0 \\ &\Leftrightarrow \quad K \text{ gehört zum Halbraum } H^+(d, 0). \end{aligned}$$

Die Aussage (ii) können wir also lesen als: „Wann immer der Halbraum $H^+(d,0)$ die Menge K enthält, enthält er auch den Punkt c." Die Negation von Aussage (ii) bedeutet dagegen, dass es eine Hyperebene $H(d,0)$ gibt, sodass K im Halbraum $H^+(d,0)$ enthalten ist, c aber nicht. Man nennt dann $H(d,0)$ eine **trennende Hyperebene**.

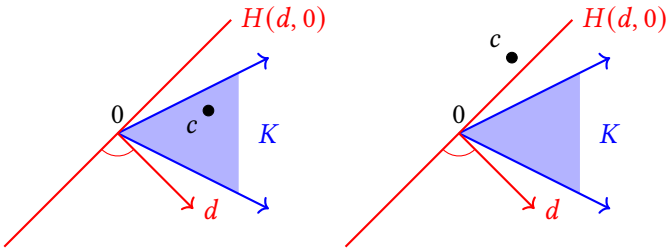


Abbildung 8.1: Illustration der beiden Fälle (links: Aussagen (i) und (ii) sind beide erfüllt und rechts: beide nicht erfüllt) im Farkas-Lemma 8.6.

Beweis von Lemma 8.6. Wir zeigen zunächst Aussage (i) $\Rightarrow$ Aussage (ii): Es sei dazu $\xi \geq 0$ mit $B^\top \xi = c$ gegeben. Weiter sei $d \in \mathbb{R}^n$ so, dass $B\,d \geq 0$ gilt. Dann folgt

$$c^\top d = (B^\top \xi)^\top d = \xi^\top (B\,d) \geq 0.$$

Um Aussage (ii) $\Rightarrow$ Aussage (i) zu zeigen, führen wir einen Widerspruchsbeweis. Wir nehmen also an, dass $c \notin K$ liegt. Wegen $0 \in K$ gilt insbesondere $c \neq 0$. Es sei $\overline{B_R(c)}$ die abgeschlossene Kugel mit Radius $R = \|c\|$. Wir betrachten die Aufgabe der orthogonalen Projektion von c auf die Menge $K \cap \overline{B_R(c)}$, also

$$\begin{aligned} \text{Minimiere} \quad & \|x - c\| \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & x \in K \cap \overline{B_R(c)}. \end{aligned} \tag{8.7}$$

Da K nach Lemma 6.10 abgeschlossen und $\overline{B_R(c)}$ kompakt ist, ist auch $K \cap \overline{B_R(c)}$ kompakt. Nach dem Satz von Weierstraß bzw. Satz 1.4 besitzt (8.7) daher einen globalen Minimierer w. Der Punkt w ist ebenfalls ein globaler Minimierer der relaxierten Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & \|x - c\| \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & x \in K, \end{aligned} \tag{8.8}$$

weil Punkte außerhalb von $\overline{B_R(c)}$ als globale Minimierer von (8.8) nicht in Betracht kommen. (**Quizfrage:** Warum?)

Behauptung: Der Vektor $d = w - c$ dient als Normalenvektor einer Hyperebene, die K vom Punkt c trennt. Die Konstruktion wird in Abbildung 8.2 veranschaulicht. Beachte, dass $K \ni w \neq c \notin K$ gilt, also $d \neq 0$.

Es sei y ein beliebiger Punkt in K. Wir betrachten Punkte auf der Verbindungsstrecke von y und w, also $\alpha\, y + (1-\alpha)\, w$ für $\alpha \in [0,1]$. Diese gehören ebenfalls zu K (**Quizfrage:** Warum?). Wir erhalten

$$\begin{aligned} \|w - c\|^2 &\leq \|\alpha\, y + (1-\alpha)\, w - c\|^2 \quad \text{(denn } w \text{ ist optimal für (8.8))} \\ &= \|\alpha\,(y - w) + (w - c)\|^2 \\ &= \alpha^2 \|y - w\|^2 + 2\,\alpha (y - w)^\top (w - c) + \|w - c\|^2. \end{aligned}$$

Daraus folgt

$$2\,(y-w)^\mathsf{T}(\underbrace{w-c}_{=d}) \geq -\alpha\,\|y-w\|^2$$

für alle $\alpha \in [0,1]$. Der Grenzübergang $\alpha \searrow 0$ zeigt

$$(y-w)^\mathsf{T} d \geq 0 \text{ für alle } y \in K. \tag{8.9}$$

Durch Einsetzen von $y = 2\,w$ und $y = 0$ (beide gehören zu K) folgt daraus $w^\mathsf{T} d \geq 0$ und gleichzeitig $w^\mathsf{T} d \leq 0$, also

$$w^\mathsf{T} d = 0. \tag{8.10}$$

Außerdem erhalten wir

$$c^\mathsf{T} d = (c-w)^\mathsf{T} d + w^\mathsf{T} d = -\|\underbrace{w-c}_{=d\neq 0}\|^2 + \underbrace{w^\mathsf{T} d}_{=0} < 0. \tag{8.11}$$

Insgesamt folgt

$$y^\mathsf{T} d \overset{(8.9)}{\geq} w^\mathsf{T} d = 0 \overset{(8.11)}{>} c^\top d \quad \text{for all } y \in K.$$

Diese Ungleichung zeigt, dass tatsächlich wie behauptet $K \subseteq H^+(d,0)$ ist, aber $c \notin H^+(d,0)$. Die Aussage (ii) gilt also nicht, was zu zeigen war. □

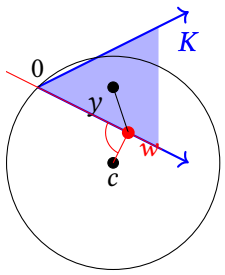


Abbildung 8.2: Illustration der Konstruktion des Normalenvektors $d = w - c$ der trennenden Hyperebene (rot) im Beweis des Farkas-Lemmas 8.6.

Wir können nun Satz 8.5 beweisen.

Beweis von Satz 8.5. Wir zeigen zunächst die hinreichenden Bedingungen.
Aussage (iii): Aus (8.6) folgt insbesondere, dass x^* und (λ^*, s^*) zulässig sind für (8.1) und (8.3). Wegen (8.4) gilt

$$b^\mathsf{T}\lambda^* = c^\mathsf{T} x^* - \underbrace{(x^*)^\mathsf{T} s^*}_{=0} = c^\mathsf{T} x^*. \tag{8.12}$$

Folgerung 8.4 zeigt nun, dass x^* und (λ^*, s^*) bereits Lösungen von (8.1) bzw. (8.3) sind.

Aussage (i): Um die notwendigen Bedingungen zu zeigen, benötigen wir das Farkas-Lemma 8.6.[9] Es sei also x eine Lösung des primalen LP (8.1). Insbesondere ist f^* endlich, und aus Lemma 6.8 folgt, dass $c^\mathsf{T} d \geq 0$ für alle Richtungen im Rezessionskegel

$$\{d \in \mathbb{R}^n \mid A\,d = 0,\ d \geq 0\}$$

[9] Ein direkter Beweis ohne Rückgriff auf das Farkas-Lemma oder den Simplex-Algorithmus findet sich bei Forsgren, 2008.

gilt. Setzen wir

$$B := \begin{bmatrix} A \\ -A \\ \mathrm{Id} \end{bmatrix},$$

so ist $Bd \geq 0$ äquivalent zu $Ad = 0$ und $d \geq 0$. Es ist also gerade die Aussage (ii) des Farkas-Lemma 8.6 erfüllt. Daraus folgt, dass ein Vektor $\xi =: (\lambda^+, \lambda^-, s) \geq 0$ existiert mit $B^\mathsf{T}\xi = c$. Setzen wir noch $\lambda := \lambda^+ - \lambda^-$, dann folgt $A^\mathsf{T}\lambda + s = c$ und $s \geq 0$. Das heißt, das duale LP ist zulässig.

Wegen der schwachen Dualität (Satz 8.2) $d^* \leq f^*$ ist der duale Optimalwert d^* endlich. Aus Satz 6.9 (in der Variante für das duale LP) folgt, dass die duale Aufgabe (8.3) lösbar ist. Es existiert also ein für die duale Aufgabe zulässiges Paar (λ^*, s^*), sodass $b^\mathsf{T}\lambda^* = d^*$ gilt. Wir müssen noch zeigen, dass $d^* = f^*$ gilt und nicht etwa $d^* < f^*$. Dann folgt aus (8.12) die noch fehlende Komplementaritätsbedingung $(x^*)^\mathsf{T}s^* = 0$, die den Beweis von (8.6) vervollständigt.

Um $d^* = f^*$ zu bestätigen, wenden wir nochmals das Farkas-Lemma 8.6 an, dieses Mal in der Form ¬ Aussage (i) ⇒ ¬ Aussage (ii). Es sei dazu $\varepsilon > 0$ beliebig. Wir wissen, dass das System

$$\underbrace{\begin{bmatrix} A & 0 \\ c^\mathsf{T} & 1 \end{bmatrix}}_{=:B^\mathsf{T}} \underbrace{\begin{pmatrix} x \\ \alpha \end{pmatrix}}_{=:\xi} = \begin{pmatrix} b \\ f^* - \varepsilon \end{pmatrix}$$

für $\binom{x}{\alpha} \geq 0$ *nicht* lösbar ist, denn das würde bedeuten: $Ax = b$, $x \geq 0$ und $c^\mathsf{T}x \leq c^\mathsf{T}x + \alpha = f^* - \varepsilon$; es wäre also x ein primal zulässiger Punkt mit kleinerem Funktionswert als der Optimalwert. Aus dem Farkas-Lemma 8.6 folgt jetzt, dass es einen Vektor d geben muss, für den $Bd \geq 0$ gilt sowie $\binom{b}{f^*-\varepsilon}^\mathsf{T} d < 0$. Wir partitionieren $d =: \binom{-\lambda}{\alpha}$ und erhalten

$$\begin{bmatrix} A^\mathsf{T} & c \\ 0 & 1 \end{bmatrix} \begin{pmatrix} -\lambda \\ \alpha \end{pmatrix} \geq 0 \quad \text{und} \quad \begin{pmatrix} b \\ f^* - \varepsilon \end{pmatrix}^\mathsf{T} \begin{pmatrix} -\lambda \\ \alpha \end{pmatrix} < 0,$$

also

$$A^\mathsf{T}\lambda \leq \alpha c, \quad \alpha \geq 0 \quad \text{und} \quad b^\mathsf{T}\lambda > \alpha (f^* - \varepsilon). \tag{8.13}$$

Der Fall $\alpha = 0$ führt schnell zum Widerspruch, denn dann wäre

$$0 \geq \underbrace{x^\mathsf{T}(A^\mathsf{T}\lambda)}_{\geq 0 \quad \leq 0} = \lambda^\mathsf{T}(Ax) = b^\mathsf{T}\lambda > 0.$$

Es muss also $\alpha > 0$ sein, und wir können durch Skalierung $\alpha = 1$ in (8.13) erreichen.[10] Damit gilt also nun

$$A^\mathsf{T}\lambda \leq c \quad \text{und} \quad b^\mathsf{T}\lambda > f^* - \varepsilon.$$

Damit ist λ dual zulässig, und aufgrund der Optimalität von x^* und des schwachen Dualitätssatzes 8.2 gilt $f^* = c^\mathsf{T}x^* \geq b^\mathsf{T}\lambda > f^* - \varepsilon$. Da $\varepsilon > 0$ beliebig war, muss $d^* := \sup\{b^\mathsf{T}\lambda \mid A^\mathsf{T}\lambda \leq c\} = \sup\{b^\mathsf{T}\lambda \mid A^\mathsf{T}\lambda + s = c,\ s \geq 0\}$ gelten.

Der Beweis von Aussage (ii) folgt ganz analog zum Beweis von Aussage (i). □

[10] Wir ersetzen dazu α durch $\alpha/\alpha = 1$ und λ durch λ/α.

Der Satz 8.5 sagt im Prinzip aus, dass wir das primale LP nicht lösen können, ohne auch das duale LP gleichzeitig zu lösen. Es ist daher nicht verwunderlich, dass im Simplex-Algorithmus, den wir in § 7 besprochen haben, auch die dualen Optimierungsvariablen (λ, s) implizit vorkommen. Bei der Berechnung der reduzierten Kosten treten die Größen

$$\begin{aligned} \lambda &:= A_B^{-\top} c_B, \\ \widetilde{c}_N := s_N &:= c_N - A_N^\top \lambda, \\ s_B &:= 0 \end{aligned} \tag{8.14}$$

auf. In jedem Simplex-Schritt sind alle Bedingungen im Optimalitätssystem (8.6) erfüllt mit Ausnahme von $s_N \geq 0$. Die Iterierten sind also primal zulässig und *dual unzulässig*, bis eine optimale Ecke gefunden wurde.

Genauer heißt Algorithmus 7.6 auch **primaler Simplex-Algorithmus**. Es gibt auch eine **duale Simplex-Variante**, in welcher in jedem Schritt alle Bedingungen im Optimalitätssystem (8.6) erfüllt sind mit Ausnahme von $x_B \geq 0$. Die Iterierten des dualen Simplex-Verfahrens sind also dual zulässig und *primal unzulässig*. Wir besprechen das duale Simplex-Verfahren in § 9.

Satz 8.7 (Mögliche primal-duale Situationen)**.**
Für jedes primal-duale Paar von LP können folgende Situationen auftreten:

		duales LP (8.2) bzw. (8.3)		
		lösbar $d^* \in \mathbb{R}$	*unbeschränkt* $d^* = \infty$	*unzulässig* $d^* = -\infty$
primales LP (8.1)	*lösbar* $f^* \in \mathbb{R}$	*(I)* $d^* = f^*$	—	—
	unbeschränkt $f^* = -\infty$	—	—	*(III)*
	unzulässig $f^* = \infty$	—	*(III)*	*(II)*

Beweis. Zu Zeile 1 und Spalte 1:

$$\begin{aligned} & f^* \in \mathbb{R} \\ \overset{\text{(Satz 6.9)}}{\Leftrightarrow} \quad & \text{das primale Problem (8.1) besitzt eine Lösung} \\ \overset{\text{(Satz 8.5)}}{\Leftrightarrow} \quad & \text{die Optimalitätsbedingungen (8.6) besitzen eine Lösung} \\ \overset{\text{(Satz 8.5)}}{\Leftrightarrow} \quad & \text{das duale Problem (8.3) besitzt eine Lösung} \\ \overset{\text{(Satz 6.9)}}{\Leftrightarrow} \quad & d^* \in \mathbb{R}. \end{aligned}$$

Für „⇐“ in der letzten Aussage: Bringe (8.3) in Normalform und benutze Satz 6.9. Aus dem schwachen Dualitätssatz 8.2 und (8.12) folgt außerdem, dass dann $d^* = f^*$ gelten muss.

Zu Zeile 2: Es sei $P \neq \emptyset$ und $f^* = -\infty$. Falls $D \neq \emptyset$ wäre, so würde nach dem Satz 8.2schwachen Dualitätssatz $d^* \leq f^* = -\infty$ gelten, Widerspruch, also muss $D = \emptyset$ und $d^* = -\infty$ gelten. Analoges gilt für die 2. Spalte (Fall (III)).

Fall (II) kann auftreten. (**Quizfrage:** Beispiel?) □

Bemerkung 8.8 (Starke Dualität)**.**
*Zu der Erkenntnis $d^* = f^*$ im Fall (I) sagt man auch: „Es tritt **keine Dualitätslücke** auf" (zwischen den Optimalwerten) oder „Es herrscht **starke Dualität**".*

§ 9 Duales Simplex-Verfahren

Literatur: Nocedal, Wright, 2006, Kapitel 13.6, Vanderbei, 2008, Kapitel 6.4

In diesem Abschnitt geben wir eine zweite Variante des Simplex-Verfahrens an, das sogenannte **duale Simplex-Verfahren**. Bei dieser tauschen primale und duale Variablen praktisch ihre Rollen. Eine Motivation dafür, beide Varianten zu betrachten, sind die unterschiedlichen Warmstart-Eigenschaften der beiden Varianten. Darunter versteht man die Fähigkeit eines Verfahrens, bei einer Änderung der Aufgabe die neue Lösung kostengünstig, ausgehend von der bisherigen Lösung, aufzudatieren. Wir gehen auf die Warmstart-Fähigkeiten später noch genauer ein.

Wir verwenden weiter den Begriff **Basis** wie in Definition 6.14, also als eine Auswahl von m Indizes aus $\{1, \dots, n\}$, sodass die Untermatrix A_B regulär ist.

Beachte: Eine Basis B legt gemäß

$$\begin{aligned} & & \lambda &:= A_B^{-\top} c_B, \\ x_B &:= A_B^{-1} b, & s_B &:= 0, \\ x_N &:= 0, & s_N &:= c_N - A_N^{\top} \lambda \end{aligned} \tag{9.1}$$

sowohl die primalen wie auch die dualen Variablen eindeutig fest.

Eine Basis B heißt **primal zulässig**, wenn der durch (9.1) beschriebene Vektor x primal zulässig ist, also die Bedingung $x_B \geq 0$ erfüllt. Eine Basis B heißt **dual zulässig**, wenn das durch (9.1) beschriebene Paar von Vektoren (λ, s) dual zulässig ist, also die Bedingung $s_N \geq 0$ erfüllt. Im Unterschied zum primalen Simplex-Verfahren werden wir mit primal unzulässigen Basisvektoren arbeiten. Dafür sind die Größen (λ, s) stets dual zulässig, siehe Tabelle 9.1.

Wir leiten jetzt einen Schritt des dualen Simplex-Verfahrens analog zu § 7.1 her. Es sei dazu als Ausgangspunkt eine dual zulässige Basis B gegeben und (λ, s) die dazugehörigen dualen Variablen gemäß (9.1). Zur Motivation des *pricing*-Schritts untersuchen wir, was passiert, wenn wir einem der Indizes in $s_B = 0$ erlauben, sich von der Null zu lösen. Wir machen also den Ansatz $s_B(t) := t\, e_\ell$ mit einem Standard-Basisvektor $e_\ell \in \mathbb{R}^m$, $t \geq 0$. In Abhängigkeit von t ergibt sich der Wert von λ nun

	Eigenschaft	primales Simplex-Verfahren	duales Simplex-Verfahren
primale Zulässigkeit	$x_B \geq 0$	✓	erst in der Lösung
	$x_N = 0$	✓	✓
	$A\,x = b$	✓	✓
duale Zulässigkeit	$s_B = 0$	✓	✓
	$s_N \geq 0$	erst in der Lösung	✓
	$A^\mathsf{T}\lambda + s = c$	✓	✓
Komplementarität	$x^\mathsf{T} s = 0$	✓	✓

Tabelle 9.1: Unterschiede zwischen primalem und dualem Simplex-Verfahren.

aus

$$A_B^\mathsf{T}\lambda(t) + s_B(t) = c_B,$$

also

$$\lambda(t) = A_B^{-\mathsf{T}}(c_B - t\,e_\ell) = \lambda + t\,\underbrace{(-A_B^{-\mathsf{T}}e_\ell)}_{=:\Delta\lambda}.$$

Welchen Index ℓ wählen wir? Dazu betrachten wir die Werte der dualen Zielfunktion:

$$b^\mathsf{T}\lambda(t) = b^\mathsf{T}\lambda - t\,b^\mathsf{T}A_B^{-\mathsf{T}}e_\ell = b^\mathsf{T}\lambda - t\,e_\ell^\mathsf{T}x_B = b^\mathsf{T}\lambda - t\,x_\ell.$$

Hier übernimmt also $x_B := A_B^{-1}b$ die Rolle der reduzierten Kosten. Da wir die duale Zielfunktion maximieren wollen, wählen wir $\ell \in B$ so, dass $x_\ell < 0$ ist. Falls bereits $x_B \geq 0$ gilt, so haben wir eine primal und dual optimale Lösung gefunden. (**Quizfrage:** Begründung?)

Nach diesem **pricing**-Schritt berechnen wir $\Delta\lambda := -A_B^{-\mathsf{T}}e_\ell$. Die Aufdatierung von s_N erhalten wir aus

$$s_N(t) = c_N - A_N^\mathsf{T}\lambda(t) = c_N - A_N^\mathsf{T}(\lambda + t\,\Delta\lambda) = s_N + t\,\underbrace{(-A_N^\mathsf{T}\Delta\lambda)}_{=:\Delta s_N}.$$

Die Wahl der Schrittweite ergibt sich aus der Bedingung der dualen Zulässigkeit, also $s_N(t) \geq 0$. Wir erhalten ähnlich zum primalen **Quotiententest**

$$\hat{t} := \min\left\{-\frac{s_i}{\Delta s_i}\,\middle|\, i \in N,\ \Delta s_i < 0\right\} = -\frac{s_r}{\Delta s_r}.$$

Falls $\Delta s_N \geq 0$ ist, so ist die duale Aufgabe unbeschränkt und damit auch die primale Aufgabe nicht lösbar. (Die primale Aufgabe ist dann notwendigerweise unzulässig, siehe Satz 8.7).

Schließlich datieren wir zur Vorbereitung des nächsten Schrittes die dualen Variablen gemäß

$$\lambda^+ := \lambda + \hat{t}\,\Delta\lambda \quad \text{und} \quad s_i^+ := \begin{cases} s_i + \hat{t}\,\Delta s_i & \text{für } i \in N,\ i \neq r, \\ \hat{t} & \text{für } i = \ell, \\ 0 & \text{sonst} \end{cases}$$

und die Basis/Nichtbasis auf:

$$B^+ := \big(B \cup \{r\}\big) \setminus \{\ell\}$$
$$N^+ := \big(N \cup \{\ell\}\big) \setminus \{r\}.$$

Der Vollständigkeit halber geben wir das duale Simplex-Verfahren nochmal komplett an und stellen es dem primalen Verfahren gegenüber (Algorithmen 9.1 und 9.2). Nachdem wir in (8.14) gesehen haben, in welcher Beziehung die reduzierten Kosten zu den dualen Variablen stehen, nutzen wir die Gelegenheit, die dualen Variablen im primalen Verfahren nochmal mit den üblichen Bezeichnungen (λ, s) umzubenennen.

Eine erste dual zulässige Ecke (sofern existent) kann mit Hilfe eines dualen Phase-I-Problems gefunden werden.

Wir gehen jetzt auf die eingangs erwähnten Warmstart-Fähigkeiten des primalen und dualen Simplex-Verfahrens ein und betrachten dazu zwei Situationen. In beiden Fällen gehen wir davon aus, dass wir mit Hilfe des (primalen oder dualen) Simplex-Verfahrens bereits eine optimale Lösung $x \in \mathbb{R}^n$ des primalen Problems (8.1) mit Basis B und gleichzeitig eine optimale Lösung $(\lambda, s) \in \mathbb{R}^m \times \mathbb{R}^n$ des dualen Problems bestimmt haben.

Hinzufügen einer Variablen

Zunächst betrachten wir die Situation, dass wir der primalen Aufgabe eine neue Variable $\bar{x}$ hinzufügen, also die Aufgabe zu

$$\begin{aligned} \text{Minimize} \quad & \begin{pmatrix} c \\ \bar{c} \end{pmatrix}^\mathsf{T} \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \quad \text{über } \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \in \mathbb{R}^{n+1} \\ \text{sodass} \quad & \begin{bmatrix} A & \bar{a} \end{bmatrix} \begin{pmatrix} x \\ \bar{x} \end{pmatrix} = b \\ \text{und} \quad & \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \geq 0 \end{aligned} \tag{9.2}$$

erweitern wollen. Wir können die neue Variable mit $\bar{x} = 0$ initialisieren und erhalten einen weiterhin primal zulässigen Basisvektor zur bisherigen Basis B. Die neue Nichtbasis ist $N \cup \{n+1\}$. Die Bedingungen der dualen Zulässigkeit für das neue Problem lauten

$$\begin{bmatrix} A^\mathsf{T} \\ \bar{a}^\mathsf{T} \end{bmatrix} \lambda + \begin{pmatrix} s \\ \bar{s} \end{pmatrix} = \begin{pmatrix} c \\ \bar{c} \end{pmatrix}, \quad \begin{pmatrix} s \\ \bar{s} \end{pmatrix} \geq 0. \tag{9.3}$$

Wir können die neue duale Schlupfvariable $\bar{s}$ mit $\bar{c} - \bar{a}^\mathsf{T}\lambda$ initialisieren, aber sie wird i. A. nicht $\bar{s} \geq 0$ erfüllen. Die Komplementaritätsbedingung $x^\mathsf{T} s + \bar{x}^\mathsf{T}\bar{s} = 0$ gilt aber weiterhin.

Diese Situation ist prädestiniert für das primale Simplex-Verfahren. Wir können es mit dem primal zulässigen Basisvektor warmstarten. Eine erneute Phase I ist nicht erforderlich. Das duale Simplex-Verfahren dagegen würde in Ermangelung eines dual zulässigen Basisvektors mit einem Phase-I-Vorlauf starten müssen und könnte von der zuvor bestimmten Lösung nicht profitieren.

Algorithmus 9.1 (Primaler Simplex-Algorithmus (Dantzig 1947)).

Eingabe: *Aufgabenbeschreibung durch A, b, c*
Eingabe: *primal zulässiger Basisvektor x von P mit zugehöriger Basis B und Nichtbasis N*
Ausgabe: *ein optimaler Basisvektor von (8.1) (und ein optimaler Basisvektor (8.3)) oder die Aussage, dass (8.1) unbeschränkt ist*

1: *Setze* $k := 0$
2: *Berechne die primalen reduzierten Kosten (duale Variablen)*

$$\lambda := A_B^{-\mathsf{T}} c_B$$
$$s_N := c_N - A_N^{\mathsf{T}} \lambda$$
$$s_B := 0$$

3: ***if*** $s_N \geq 0$ ***then***
4: *x ist eine Lösung von (8.1), und (λ, s) ist eine Lösung von (8.3),* ***STOP***
5: ***else***
6: *Wähle einen Index $r \in N$ mit $s_r < 0$*
7: *Berechne*

$$\Delta x_B := -A_B^{-1} a_r$$

8: ***if*** $\Delta x_B \geq 0$ ***then***
9: *Aufgabe (8.1) ist unbeschränkt,* ***STOP***
10: ***else***
11: *Bestimme $\hat{t} \geq 0$ und $\ell \in B$ gemäß*

$$\hat{t} := \min\left\{ -\frac{x_i}{\Delta x_i} \,\middle|\, i \in B,\ \Delta x_i < 0 \right\} = -\frac{x_\ell}{\Delta x_\ell}$$

12: *Setze*

$$x_i^+ := \begin{cases} x_i + \hat{t}\,\Delta x_i & \text{für } i \in B,\ i \neq \ell, \\ \hat{t} & \text{für } i = r, \\ 0 & \text{sonst} \end{cases}$$

13: *Setze $B^+ := \big(B \cup \{r\}\big) \setminus \{\ell\}$*
14: *Setze $N^+ := \{1, \ldots, n\} \setminus B^+$*
15: *Setze $x := x^+$*
16: *Setze $B := B^+$ und $N := N^+$*
17: *Setze $k := k + 1$*
18: ***end if***
19: ***end if***
20: *Gehe zu Zeile 2*

Algorithmus 9.2 (Dualer Simplex-Algorithmus (Lemke, 1954)).

Eingabe: *Aufgabenbeschreibung durch A, b, c*
Eingabe: *dual zulässiger Basisvektor (λ, s) von P mit zugehöriger Basis B und Nichtbasis N*
Ausgabe: *ein optimaler Basisvektor von (8.3) (und ein optimaler Basisvektor von (8.1)) oder die Aussage, dass (8.3) unbeschränkt ist*

1: *Setze* $k := 0$
2: *Berechne die dualen reduzierten Kosten (primale Variablen)*

$$x_B := A_B^{-1} b$$
$$x_N := 0$$

3: ***if*** $x_B \geq 0$ ***then***
4: *(λ, s) ist eine Lösung von (8.3), und x ist eine Lösung von (8.1),* ***STOP***
5: ***else***
6: *Wähle einen Index $\ell \in B$ mit $x_\ell < 0$*
7: *Berechne*

$$\Delta\lambda := -A_B^{-\mathsf{T}} e_\ell$$
$$\Delta s_N := -A_N^{\mathsf{T}} \Delta\lambda$$

8: ***if*** $\Delta s_N \geq 0$ ***then***
9: *Aufgabe (8.3) ist unbeschränkt,* ***STOP***
10: ***else***
11: *Bestimme $\hat{t} \geq 0$ und $r \in N$ gemäß*

$$\hat{t} := \min\left\{ -\frac{s_i}{\Delta s_i} \,\middle|\, i \in N,\ \Delta s_i < 0 \right\} = -\frac{s_r}{\Delta s_r}$$

12: *Setze $\lambda^+ := \lambda + \hat{t}\,\Delta\lambda$ und*

$$s_i^+ := \begin{cases} s_i + \hat{t}\,\Delta s_i & \text{für } i \in N,\ i \neq r\cancel{\ell}, \\ \hat{t} & \text{für } i = \cancel{\ell}r, \\ 0 & \text{sonst} \end{cases}$$

13: *Setze $B^+ := \big(B \cup \{r\}\big) \setminus \{\ell\}$*
14: *Setze $N^+ := \{1, \ldots, n\} \setminus B^+$*
15: *Setze $\lambda := \lambda^+$ und $s := s^+$*
16: *Setze $B := B^+$ und $N := N^+$*
17: *Setze $k := k + 1$*
18: ***end if***
19: ***end if***
20: *Gehe zu Zeile 2*

Hinzufügen einer Nebenbedingung

Wir betrachten jetzt eine andere Veränderung der primalen Aufgabe (8.1) und fügen ihr eine neue Ungleichungsnebenbedingung $\bar{a}^\mathsf{T} x \leq \bar{b}$ bzw. $\bar{a}^\mathsf{T} x + \bar{x} = \bar{b}$ mit zugehöriger Schlupfvariable $\bar{x}$ hinzu:

$$\begin{aligned} \text{Minimiere} \quad & \begin{pmatrix} c \\ 0 \end{pmatrix}^\mathsf{T} \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \quad \text{über} \ \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \\ \text{sodass} \quad & \begin{bmatrix} A & 0 \\ \bar{a}^\mathsf{T} & 1 \end{bmatrix} x = \begin{pmatrix} b \\ \bar{b} \end{pmatrix} \\ \text{und} \quad & \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \geq 0. \end{aligned} \tag{9.4}$$

Die Bedingungen der dualen Zulässigkeit für die neue Aufgabe lauten

$$\begin{bmatrix} A^\mathsf{T} & \bar{a} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \lambda \\ \bar{\lambda} \end{pmatrix} + \begin{pmatrix} s \\ \bar{s} \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix}, \quad \begin{pmatrix} s \\ \bar{s} \end{pmatrix} \geq 0. \tag{9.5}$$

Die bisherige Lösung x wird i. A. nicht länger primal zulässig sein, da $\bar{a}^\mathsf{T} x \leq \bar{b}$ verletzt ist. Wir können aber die bisherige dual optimale Lösung durch $\bar{\lambda} = \bar{s} = 0$ erweitern und sind weiterhin dual zulässig. Genauer erweitern wir die bisherige Basis zu $B \cup \{n+1\}$. Die neue Basismatrix ist daher

$$\begin{bmatrix} A_B & 0 \\ \bar{a}^\mathsf{T} & 1 \end{bmatrix}.$$

Diese ist weiterhin regulär (**Quizfrage:** Warum?), sodass wir tatsächlich von einer Basis sprechen können.

Diese Situation ist nun wie geschaffen für das duale Simplex-Verfahren. Wir können es mit dem dual zulässigen Basisvektor warmstarten. Eine erneute duale Phase I ist nicht erforderlich. Das primale Simplex-Verfahren dagegen würde angesichts eines fehlenden primal zulässigen Basisvektors mit einem Phase-I-Vorlauf starten müssen und könnte von der zuvor bestimmten Lösung nicht profitieren.

Bemerkung 9.3 (Das duale Simplex-Verfahren in der ganzzahligen linearen Optimierung)**.** *Die in* (9.4) *beschriebene Situation, dass wir einem bereits gelösten LP eine Ungleichungsnebenbedingung hinzufügen wollen, kommt vor allem bei der Lösung sogenannter* ***(gemischt-)ganzzahliger linearer Optimierungsaufgaben*** *(****(gemischt-)ganzzahliges lineares Programm****, englisch:* mixed-integer linear program*,* ***MILP****) vor. Das sind lineare Optimierungsaufgaben, bei denen einige oder alle der Optimierungsvariablen x_i ganzzahlig sein müssen, also $x_i \in \mathbb{Z}$ an Stelle von $x_i \in \mathbb{R}$. Bei Verwendung der Normalform geht es z. B. um Aufgaben der Form*

$$\left.\begin{aligned} \textit{Minimiere} \quad & c^\mathsf{T} x \quad \textit{über } x \in \mathbb{Z}^n \\ \textit{sodass} \quad & A\,x = b \\ \textit{und} \quad & x \geq 0. \end{aligned}\right\} \tag{9.6}$$

Solche Aufgaben fallen in den Bereich der ganzzahligen Optimierung. In einem gängigen Lösungsansatz, den man ***branch and bound*** *nennt, wird zunächst ein relaxiertes LP gelöst, bei dem die die Ganzzahligkeitsbedingungen vernachlässigt werden, also* (8.1)*. Dessen Lösung bezeichnen wir jetzt mit x^*. Dann wird*

eine Variable x_i^* ausgewählt, die die Ganzzahligkeitsbedingung verletzt, und es werden die zwei LPs

$$\left.\begin{array}{rll} \textit{Minimiere} & c^\mathsf{T}x & \textit{über } x \in \mathbb{R}^n \\ \textit{sodass} & Ax = b & \\ \textit{und} & x \geq 0 & \\ \textit{sowie} & x_i \geq \lceil x_i^* \rceil & \end{array}\right\} \quad \left\{\begin{array}{rll} \textit{Minimiere} & c^\mathsf{T}x & \textit{über } x \in \mathbb{R}^n \\ \textit{sodass} & Ax = b & \\ \textit{und} & x \geq 0 & \\ \textit{sowie} & x_i \leq \lfloor x_i^* \rfloor & \end{array}\right. \tag{9.7}$$

gelöst. Dabei sind $\lceil\cdot\rceil$ und $\lfloor\cdot\rfloor$ die ***obere*** *bzw.* ***untere Gaußklammer****, d. h.,*

$$\lceil z \rceil := \min\{y \in \mathbb{Z} \mid y \geq z\} \quad \textit{(kleinste ganze Zahl oberhalb von } z\textit{)},$$
$$\lfloor z \rfloor := \max\{y \in \mathbb{Z} \mid y \leq z\} \quad \textit{(kleinste ganze Zahl unterhalb von } z\textit{)}.$$

Für die Lösung der beiden Aufgaben in (9.7) bietet sich das duale Simplex-Verfahren besonders an, weil die Lösung ohne die hinzugefügten Ungleichungsnebenbedingungen bereits bekannt ist.

Ende der Woche 6

§ 10 Sensitivitätsanalyse

In diesem Abschnitt gehen wir der Frage nach, wie empfindlich (sensitiv) der Optimalwert (also der Zielfunktionswert an einer optimalen Lösung) eines LPs in Normalform (8.1) gegenüber Änderungen im Kostenvektor c und in der rechten Seite b abhängen.

Motivation: Was wäre etwa beim Mozartproblem (Beispiel 6.7), wenn wir den Gewinn c pro produzierter Einheit Mozartkugeln/-taler ändern, indem wir die Verkaufspreise abändern? Und was passiert, wenn wir eine Änderung in den nutzbaren Ressourcen (dem Lagerbestand b) feststellen, z. B. durch den unerwarteten Verfall von Zutaten?

Quizfrage: Was sind weitere Beispiele linearer Optimierungsaufgaben, bei denen es von Interesse sein könnte, Änderungen von b und/oder c zu untersuchen? Durch welche Ereignisse könnten diese Änderungen ausgelöst worden sein?

Quizfrage: Was sind Beispiele von Veränderungen in der Aufgabenstellungen, die *nicht* durch Änderungen in b und/oder c dargestellt werden können?

Natürlich könnten wir die Aufgabe mit den modifizierten Daten b oder c einfach erneut lösen und die Änderung in der Zielfunktion ablesen. Es wird sich jedoch zeigen, dass wir in vielen Fällen eine Vorhersage bereits auf Basis der Lösung des unveränderten Problems treffen können.

Wir machen in diesem Abschnitt folgende **Voraussetzung**: Es seien x^* und (λ^*, s^*) Lösungen der primalen Aufgabe (8.1) bzw. der dualen Aufgabe (8.3) zu einer Basis B, also optimale Ecken, wie sie mit dem primalen oder dem dualen Simplex-Verfahren berechnet werden.

Änderungen im Kostenvektor

Wir bezeichnen mit $\Delta c \in \mathbb{R}^n$ eine Änderungsrichtung im Kostenvektor c der primalen Aufgabe und betrachten folgende Familie primal-dualer Aufgaben mit Parameter $t \in \mathbb{R}$:

$$\left.\begin{array}{rl} \text{Minimiere} & (c + t\,\Delta c)^\mathsf{T} x \\ \text{sodass} & A\,x = b \\ \text{und} & x \geq 0 \end{array}\right\} \quad \left\{\begin{array}{rl} \text{Maximiere} & b^\mathsf{T}\lambda \\ \text{sodass} & A^\mathsf{T}\lambda + s = c + t\,\Delta c \\ \text{und} & s \geq 0. \end{array}\right. \tag{10.1}$$

Welche Aussagekraft besitzen die Lösungen x^* und (λ^*, s^*) des „ungestörten“ Aufgabenpaares ($t = 0$) noch für (10.1)?

Da (10.1) dieselbe primal zulässige Menge besitzt wie die ungestörte Aufgabe (8.1), ist x^* weiterhin primal zulässig. Die dual zulässige Menge hat sich jedoch gegenüber (8.3) geändert. Wir können aber den Versuch unternehmen, die duale Lösung aufzudatieren. Dazu gehen wir wie in § 9 bei der Herleitung des dualen Simplex-Verfahrens vor. Durch die Basis B sind die dualen Variablen wie folgt festgelegt, vgl. (9.1):

$$\lambda(t) = \lambda^* + t\,\Delta\lambda \qquad \text{mit } \Delta\lambda := A_B^{-\mathsf{T}}\Delta c_B \tag{10.2a}$$

$$s_N(t) = s_N^* + t\,\Delta s_N \qquad \text{mit } \Delta s_N := \Delta c_N - A_N^\mathsf{T}\Delta\lambda \tag{10.2b}$$

$$s_B(t) \equiv s_B^* = 0. \tag{10.2c}$$

Wann sind die auf diese Art und Weise erhaltenen Vektoren x^* und $(\lambda(t), s(t))$ optimal für (10.1)? Wir überprüfen dazu die Optimalitätsbedingungen (8.6). Die primale Zulässigkeit

$$Ax^* = b, \quad x^* \geq 0$$

ist erfüllt, ebenso die Komplementaritätsbedingung:

$$x_B^* \underbrace{s_B(t)}_{=0} + x_N^* \underbrace{s_N}(t) = 0.$$

Bzgl. der dualen Zulässigkeit ist die erste Bedingung

$$\begin{aligned} A^\mathsf{T}\lambda(t) + s(t) &= \begin{pmatrix} A_B^\mathsf{T}\lambda^* + t\,A_B^\mathsf{T}A_B^{-\mathsf{T}}\Delta c_B + s_B(t) \\ A_N^\mathsf{T}\lambda^* + t\,A_N^\mathsf{T}A_B^{-\mathsf{T}}\Delta c_B + s_N(t) \end{pmatrix} = \begin{pmatrix} c_B + t\,\Delta c_B + 0 \\ A_N^\mathsf{T}\lambda^* + t\,A_N^\mathsf{T}A_B^{-\mathsf{T}}\Delta c_B + s_N^* + t\,\Delta c_N - t\,A_N^\mathsf{T}A_B^{-\mathsf{T}}\Delta c_B \end{pmatrix} \\ &= \begin{pmatrix} c_B + t\,\Delta c_B \\ c_N + t\,\Delta c_N \end{pmatrix} = c + t\,\Delta c \end{aligned}$$

nach Konstruktion von $\lambda(t)$ und $s(t)$ erfüllt. Die Vorzeichenbedingung $s_N(t) \geq 0$ jedoch gilt nicht automatisch, sondern genau dann, wenn der Störungsparameter t der Bedingung[11]

$$\sup_{\substack{i \in N \\ \Delta s_i > 0}} \underbrace{\left\{-\frac{s_i^*}{\Delta s_i}\right\}}_{\leq 0} \leq t \leq \inf_{\substack{i \in N \\ \Delta s_i < 0}} \underbrace{\left\{-\frac{s_i^*}{\Delta s_i}\right\}}_{\geq 0}. \tag{10.3}$$

genügt.

[11] Wir schreiben hier sup und inf, da die betreffenden Indexmengen durchaus leer sein können.

Beachte: Die durch (10.3) beschriebene Menge ist ein abgeschlossenes (möglicherweise unbeschränktes) Intervall $I(\Delta c)$, das die 0 enthält. Im Extremfall ist $I(\Delta c) = \{0\}$.

Für $t \in I(\Delta c)$ ist also tatsächlich x^* auch für die gestörten Probleme (10.1) weiterhin eine optimale Ecke. Der zugehörige Optimalwert lässt sich daher bequem aus der primalen Aufgabe ablesen:

$$f^*(t) = (c + t\,\Delta c)^\mathsf{T} x^* = f^* + t\,\Delta c^\mathsf{T} x^*. \tag{10.4}$$

Wir fassen unsere Erkenntnisse zusammen:

Satz 10.1 (Sensitivitätssatz bei LP bei Änderungen im Kostenvektor)**.**
Es seien x^* und (λ^*, s^*) Lösungen der primalen Aufgabe (8.1) bzw. der dualen Aufgabe (8.3) zu einer Basis B. Dann gilt:

(*i*) *Für beliebiges $\Delta c \in \mathbb{R}^n$ und zugehörige t gemäß (10.3) ist x^* für $(10.1)_{\text{primal}}$ weiterhin ein optimaler Basisvektor, und $(\lambda(t), s(t))$ aus (10.2) ist ein optimaler Basisvektor für $(10.1)_{\text{dual}}$. Der gemeinsame Optimalwert beider Aufgaben ist $c^\mathsf{T} x^* + t\,(\Delta c)^\mathsf{T} x^*$.*

(*ii*) *Ist die rechte Grenze des Intervalls (10.3) echt positiv, dann ist die Optimalwertfunktion*

$$c \mapsto \Phi := \textit{gemeinsamer Optimalwert von (8.1) und (8.3)}$$

an der Stelle c in Richtung Δc (einseitig) richtungsdiffbar, und die Richtungsableitung ist gegeben durch

$$\Phi'(c; \Delta c) = (\Delta c)^\mathsf{T} x^*.$$

(*iii*) *Ist s^* nicht entartet, gilt also $s^*_N > 0$, dann ist die Optimalwertfunktion in einer offenen Kugel $B_r(c)$ von c linear mit*

$$\Phi(c + \Delta c) = (c + \Delta c)^\mathsf{T} x^* \quad \textit{für } \Delta c \in B_r(0).$$

Damit ist Φ überall in dieser Kugel differenzierbar, und es gilt

$$\Phi'(c + \Delta c) \equiv (x^*)^\mathsf{T} \quad \textit{für } \Delta c \in B_r(0).$$

Beweis. Aussage (*i*): Diese Aussage haben wir durch Bestätigung der Optimalitätsbedingungen (8.6) bereits bewiesen.

Aussage (*ii*): Unter der genannten Voraussetzung ist $\Phi(c + t\,\Delta c)$ für hinreichend kleine $t > 0$ durch (10.4) gegeben. Für die Richtungsdiffbarkeit von Φ betrachten wir den Differenzenquotienten für solche t:

$$\frac{\Phi(c + t\,\Delta c) - \Phi(c)}{t} = \frac{c^\mathsf{T} x^* + t\,(\Delta c)^\mathsf{T} x^* - c^\mathsf{T} x^*}{t} = (\Delta c)^\mathsf{T} x^*,$$

also ist das auch der Wert im Grenzwert $t \searrow 0$, der Richtungsableitung $\Phi'(c; \Delta c)$.

Aussage (*iii*): Wenn s^* nicht entartet ist, dann enthält das zulässige Intervall (10.3) für jede beliebige Richtung Δc immer ein offenes Intervall um die 0. Wir müssen aber zeigen, dass die Länge dieses

Intervalls über alle Richtungen Δc konstanter Norm gleichmäßig von 0 weg beschränkt bleibt. Wir zeigen dazu, dass die Funktion, die die obere Intervallgrenze angibt,

$$\Delta c \mapsto \inf_{\substack{i \in N \\ \Delta s_i < 0}} \underbrace{\left\{ -\frac{s_i^*}{\Delta s_i} \right\}}_{\geq 0}, \tag{10.5}$$

auf der Einheitssphäre $\{\Delta c \in \mathbb{R}^n \mid \|\Delta c\| = 1\}$ gleichmäßig von 0 weg beschränkt ist. Das ist ausreichend, weil sich die untere Intervallgrenze durch den Übergang $\Delta c \rightsquigarrow -\Delta c$ ergibt.

Wegen (10.2) hängt Δs_N linear (und damit stetig) von Δc ab:

$$\Delta s_N = \Delta c_N - A_N^\mathsf{T} A_B^{-\mathsf{T}} \Delta c_B.$$

Da die Sphäre kompakt ist, existiert für jede Komponente $i \in N$ von Δs_N ein endliches

$$\beta_i := \max\left\{ \Delta s_i = [\Delta c_N - A_N^\mathsf{T} A_B^{-\mathsf{T}} \Delta c_B]_i \,\middle|\, \|\Delta c\| = 1 \right\}.$$

Es gilt $\beta_i > 0$ für alle $i \in N$. (**Quizfrage:** Warum?) Wir setzen nun $\beta := \max\{\beta_i \mid i \in N\} > 0$ und $\alpha := \min\{s_i^* \mid i \in N\} > 0$.

Für beliebiges Δc aus der Einheitssphäre und das zugehörige Δs gilt: Falls $\Delta s_N \geq 0$ ist, dann erhalten wir

$$\inf_{\substack{i \in N \\ \Delta s_i < 0}} \left\{ -\frac{s_i^*}{\Delta s_i} \right\} = \inf \emptyset = \infty.$$

Andernfalls gilt

$$\inf_{\substack{i \in N \\ \Delta s_i < 0}} \left\{ -\frac{s_i^*}{\Delta s_i} \right\} = \min_{\substack{i \in N \\ \Delta s_i < 0}} \left\{ -\frac{s_i^*}{\Delta s_i} \right\} \geq \frac{\min\limits_{i \in N} s_i^*}{\max\limits_{\substack{i \in N \\ \Delta s_i < 0}} \{-\Delta s_i\}} \geq \frac{\alpha}{\max\limits_{\substack{i \in N \\ \Delta s_i < 0}} \{-\Delta s_i\}} \geq \frac{\alpha}{\beta} > 0.$$

Die letzte Ungleichung gilt, da wir jeden der im Nenner vorkommenden Werte mit $0 < -\Delta s_i \leq \beta_i \leq \beta$ abschätzen können und daher auch $\max\{-\Delta s_i \mid i \in N,\ \Delta s_i < 0\} \leq \beta$ gilt. Zusammenfassend bekommen wir also die gewünschte Aussage

$$\inf_{\|\Delta c\|=1} \inf_{\substack{i \in N \\ \Delta s_i < 0}} \left\{ -\frac{s_i^*}{\Delta s_i} \right\} \geq \frac{\alpha}{\beta} =: r > 0.$$

Daraus folgt, dass die Vereinigung der Menge aller zulässigen Störungen $t\,\Delta c$ die offene Kugel $B_r(c)$ enthält:

$$\bigcup_{\|\Delta c\|=1} \left\{ t\,\Delta c \,\middle|\, t \in I(\Delta c) \right\} \supseteq \bigcup_{\|\Delta c\|=1} \left\{ t\,\Delta c \,\middle|\, t \in (-r, r) \right\} = B_r(c).$$

Weiter folgt in Verbindung mit (10.4), dass für alle Kostenvektoren $c + \Delta c$ mit $\Delta c \in B_r(0)$ die Optimalwertfunktion die Gestalt $\Phi(c + \Delta c) = \Phi(c) + \Delta c^\mathsf{T} x^* = (c + \Delta c)^\mathsf{T} x^*$ hat. Die Differenzierbarkeit von Φ in $B_r(c)$ mit Ableitung $(x^*)^\mathsf{T}$ ist eine unmittelbare Konsequenz. □

Änderungen in der rechten Seite

Wir betrachten jetzt Änderungen in der rechten Seite b und bezeichnen mit $\Delta b \in \mathbb{R}^m$ eine entsprechende Änderungsrichtung. Das primal-duale Paar von Aufgaben hat nun die Gestalt

$$\left.\begin{array}{rl} \text{Minimiere} & c^\mathsf{T} x \\ \text{sodass} & A\,x = b + t\,\Delta b \\ \text{und} & x \geq 0 \end{array}\right\} \quad \left\{\begin{array}{rl} \text{Maximiere} & (b + t\,\Delta b)^\mathsf{T}\lambda \\ \text{sodass} & A^\mathsf{T}\lambda + s = c \\ \text{und} & s \geq 0. \end{array}\right. \tag{10.6}$$

Dieses Mal ist (λ^*, s^*) weiterhin dual zulässig, die primal zulässige Menge hat sich jedoch geändert. Wir unternehmen daher jetzt den Versuch, die primale Lösung aufzudatieren. Das Vorgehen ähnelt dem bei der Herleitung des primalen Simplex-Verfahrens in § 7. Durch die Basis B ist die primale Variable wie folgt festgelegt:

$$x_B(t) = x_B + t\,\Delta x_B \quad \text{mit } \Delta x_B := A_B^{-1}\Delta b \tag{10.7a}$$

$$x_N(t) \equiv x_N^* = 0. \tag{10.7b}$$

Wann sind die auf diese Art und Weise erhaltenen Vektoren $x(t)$ und (λ^*, s^*) optimal für (10.6)? Wir überprüfen dazu die Optimalitätsbedingungen (8.6). Die duale Zulässigkeit

$$A^\mathsf{T}\lambda^* + s^* = c, \quad s^* \geq 0$$

ist erfüllt, ebenso die Komplementaritätsbedingung:

$$\underbrace{x_B(t)}_{=0}\,s_B^* + \underbrace{x_N(t)}_{=0}\,s_N^* = 0.$$

Bzgl. der primalen Zulässigkeit ist die erste Bedingung

$$Ax(t) = A_B x_B(t) + A_N x_N(t) = A_B x_B^* + t\,A_B\Delta x_B + A_N x_N = A_B x_B^* + t\,A_B A_B^{-1}\Delta b + 0 = b + t\,\Delta b$$

nach Konstruktion von $x(t)$ erfüllt. Die Vorzeichenbedingung $x_B(t) \geq 0$ jedoch gilt nicht automatisch, sondern genau dann, wenn der Störungsparameter t der Bedingung

$$\sup_{\substack{i\in B\\ \Delta x_i>0}} \underbrace{\left\{-\frac{x_i^*}{\Delta x_i}\right\}}_{\leq 0} \leq t \leq \inf_{\substack{i\in B\\ \Delta x_i<0}} \underbrace{\left\{-\frac{x_i^*}{\Delta x_i}\right\}}_{\geq 0}. \tag{10.8}$$

genügt.

Für diese t ist also tatsächlich (λ^*, s^*) auch für die gestörten Probleme (10.6) weiterhin eine optimale Ecke. Der zugehörige Optimalwert lässt sich daher dieses Mal bequem aus der dualen Aufgabe ablesen:

$$d^*(t) = (b + t\,\Delta b)^\mathsf{T}\lambda^* = d^* + t\,\Delta b^\mathsf{T}\lambda^*. \tag{10.9}$$

Die Erkenntnisse fassen wir wie folgt zusammen:

Satz 10.2 (Sensitivitätssatz bei LP bei Änderungen in der rechten Seite).
Es seien x^* und (λ^*, s^*) Lösungen der primalen Aufgabe (8.1) bzw. der dualen Aufgabe (8.3) zu einer Basis B. Dann gilt:

(*i*) *Für beliebiges $\Delta b \in \mathbb{R}^n$ und zugehörige t gemäß (10.8) ist $(\lambda^* s^*)$ für $(10.6)_{\text{dual}}$ weiterhin ein optimaler Basisvektor, und $x(t)$ aus (10.7) ist ein optimaler Basisvektor für $(10.6)_{\text{primal}}$. Der gemeinsame Optimalwert beider Aufgaben ist $b^\mathsf{T}\lambda^* + t\,(\Delta b)^\mathsf{T}\lambda^*$.*

(*ii*) *Ist die rechte Grenze des Intervalls (10.8) echt positiv, dann ist die Optimalwertfunktion*

$$b \mapsto \Psi := \textit{gemeinsamer Optimalwert von (8.1) und (8.3)}$$

an der Stelle b in Richtung Δb (einseitig) richtungsdiffbar, und die Richtungsableitung ist gegeben durch

$$\Psi'(b; \Delta b) = (\Delta b)^\mathsf{T}\lambda^*.$$

(*iii*) *Ist x^* nicht entartet, gilt also $x^*_B > 0$, dann ist die Optimalwertfunktion in einer offenen Kugel $B_r(b)$ von b linear mit*

$$\Psi(b + \Delta b) = (b + \Delta b)^\mathsf{T}\lambda^* \quad \textit{für } \Delta b \in B_r(0).$$

Damit ist Ψ überall in dieser Kugel differenzierbar, und es gilt

$$\Psi'(b + \Delta b) \equiv (\lambda^*)^\mathsf{T} \quad \textit{für } \Delta b \in B_r(0).$$

Der Beweis erfolgt analog zum Beweis von Satz 10.1.

Beispiel 10.3 (Sensitivitäten beim Mozartproblem). *Das Mozartproblem in Normalform (Beispiel 6.7) ist durch die Daten*

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 1 \end{pmatrix}, \quad c = \begin{pmatrix} -9 \\ -8 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad b = \begin{pmatrix} 6 \\ 11 \\ 9 \end{pmatrix} \begin{matrix} \textit{Marzipan} \\ \textit{Nougat} \\ \textit{Schokolade} \end{matrix}$$

gegeben. Die eindeutige primal optimale Lösung ist $x^* = (5, 1, 0, 0, 2)^\mathsf{T}$ zur Basis $B = \{1, 2, 5\}$. Auch die duale Lösung (λ^*, s^*) ist eindeutig, und zwar $\lambda^* = (-7, -1, 0)^\mathsf{T}$ und $s^* = (0, 0, 7, 1, 0)^\mathsf{T}$. Beide Basisvektoren sind nicht entartet, denn x^* hat Nulleinträge nur in der Nichtbasis $N = \{3, 4\}$, und s^* hat Nulleinträge nur in der Basis B. Wir erhalten also folgende Darstellung des Optimalwerts als Funktion des Ressourcenvektors b

$$\Psi(b + \Delta b) = (b + \Delta b)^\mathsf{T}\lambda^* = -53 + (\Delta b)^\mathsf{T} \begin{pmatrix} -7 \\ -1 \\ 0 \end{pmatrix}$$

für Δb mit hinreichend kleiner Norm. Das bedeutet beispielsweise, dass wir pro Einheit an Marzipan, das wir zusätzlich zur Verfügung haben ($\Delta b = (1, 0, 0)^\mathsf{T}$), sieben Einheiten zusätzlichen Gewinn machen können. Wenn wir also die Gelegenheit hätten, Marzipan am Markt zuzukaufen, dann wären sieben

Geldeinheiten pro Einheit Marzipan der Preis, den wir höchstens bezahlen sollten, damit sich der Zukauf noch lohnt. Der so ermittelte Preis von sieben Geldeinheiten pro Einheit Marzipan ist kein realer Preis, sondern er dient uns als Vergleichspreis. Man bezeichnet ihn deshalb auch als ***Schattenpreis****. Er ergibt sich aus der dualen Lösung der ungestörten Aufgabe, also letztlich aus den Problemdaten A, b und c, die i. d. R. nur uns als Unternehmen bekannt sind.*

Bis zu welcher Größenordnung ist es für uns sinnvoll, Marzipan zuzukaufen (falls dessen Preis unter sieben Geldeinheiten pro Einheit Marzipan liegt)? Dazu bestimmen wir aus (10.8) den zulässigen Bereich für t im Ausdruck $t\,\Delta b$. Dafür benötigen wir die Änderungsrichtung $\Delta x_B = A_B^{-1}\Delta b = (-1, 2, -3)$. Aus (10.8) ergibt sich die erlaubte Störungsgröße:

$$\begin{aligned} & \sup_{\substack{i\in B\\ \Delta x_i>0}} \underbrace{\left\{-\frac{x_i^*}{\Delta x_i}\right\}}_{\leq 0} \leq t \leq \inf_{\substack{i\in B\\ \Delta x_i<0}} \underbrace{\left\{-\frac{x_i^*}{\Delta x_i}\right\}}_{\geq 0}. \\ \Leftrightarrow \quad & \max\underbrace{\left\{-\frac{1}{2}\right\}}_{i=2} \leq t \leq \min\left\{\underbrace{-\frac{5}{-1}}_{i=1}, \underbrace{-\frac{2}{-3}}_{i=5}\right\} \\ \Leftrightarrow \quad & -\frac{1}{2} \leq t \leq \frac{2}{3}. \end{aligned}$$

Bis zu $2/3$ Einheiten Marzipan können also zugekauft werden, ohne dass sich die Lösungsstruktur ändert.

Quizfrage: *Was passiert an der Stelle $t = 2/3$? Wie sieht die Rechnung aus, wenn man stattdessen $\Delta b = (0, 1, 0)^\mathsf{T}$ (Nougat) verwendet? Und bei $\Delta b = (0, 0, 1)^\mathsf{T}$ (Schokolade)?*

Ende der Woche 7

§ 11 Lineare Optimierungsaufgaben auf Graphen

In diesem Abschnitt behandeln wir eine prominente Klasse linearer Optimierungsaufgaben. Wir beginnen mit einem einführenden Beispiel.

Beispiel 11.1 (Kostenminimaler Transport)**.** *Ein Unternehmen verfügt über das in Abbildung 11.1 dargestellte* ***Transportnetzwerk****. Dabei entsprechen die* ***Knoten*** *1–3 den Produktionsstätten, 4–5 den Zwischenlagern und 6–9 den Verkaufsstätten. Die* ***Kanten*** *zwischen den Knoten entsprechen den möglichen Transportwegen. Die* ***Produktionsmengen*** *der Produktionsstätten sowie die* ***Bedarfe*** *der Verkaufsstätten (für einen festen Zeitraum, z. B. einen Monat) seien bekannt.*

Es geht darum, den Transport der produzierten Waren von den Produktionsstätten über die Zwischenlager zu den Verkaufsstätten zu planen. Jeder Transportweg (Kante) ist dabei mit Transportkosten belegt, die proportional zu der Warenmenge sind, die über diesen Weg transportiert wird. Außerdem wird es üblicherweise ***Kapazitätsbeschränkungen*** *auf jedem Transportweg geben. Gesucht ist nun eine optimale Belegung der Kanten mit den darüber zu transportierenden Warenmengen, sodass (unter Beachtung aller Restriktionen) die Gesamttransportkosten minimiert werden.*

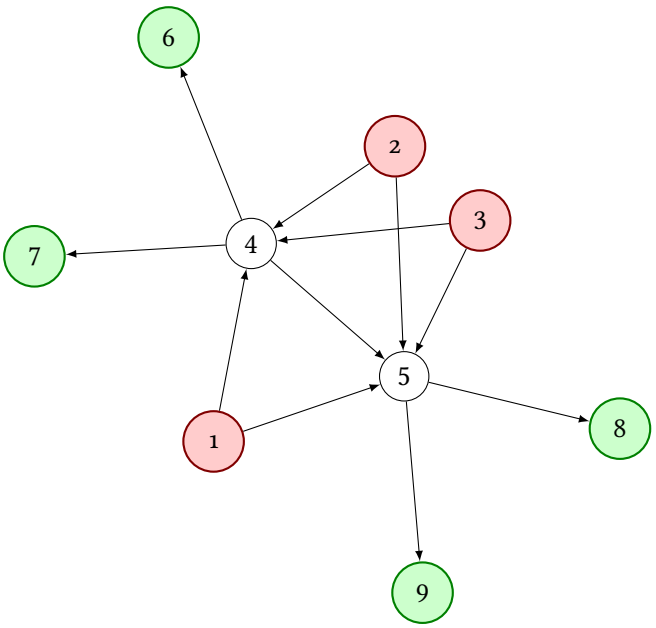


Abbildung 11.1: Transportnetzwerk eines Unternehmens (siehe Beispiel 11.1) mit Produktionsstätten 1–3 (rot), Zwischenlagern 4–5 und Verkaufsstätten 6–9 (grün).

Definition 11.2 (Graphen)**.**

(*i*) *Ein* ***gerichteter Graph*** *(kurz:* ***Digraph****, englisch:* directed graph, digraph*)* (V, E) *besteht aus einer endlichen Menge* V *von Knoten (englisch:* vertices, nodes*) und einer endlichen Menge* E *von* ***gerichteten Kanten*** *(englisch:* directed edges, directed arcs*) zwischen Knoten.*

(*ii*) *Eine gerichtete Kante ist ein Paar* $e = (x, y) \in V \times V$*. Dabei heißt* $x \in V$ *der* ***Anfangsknoten*** *(englisch:* tail vertex*) und* $y \in V$ *der* ***Endknoten*** *(englisch:* head vertex*).*

(*iii*) *Eine gerichtete Kante* $e = (x, y)$ *heißt* ***Schleife****, wenn* $x = y$ *ist. Ein Digraph heißt* ***einfach*** *(englisch:* simple digraph*), wenn keine der Kanten eine Schleife ist.*

Quizfrage: Es gibt Situationen, bei denen zwischen zwei Knoten mehrere Kanten betrachtet werden soll, die beispielsweise verschiedenen Transportwegen entsprechen. Das ist aber in der Definition eines Digraphen nicht vorgesehen. Wie kann man das Problem lösen?

Der Graph aus Abbildung 11.1 wird beispielsweise beschrieben durch

$$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9\} \tag{11.1a}$$

$$E = \big\{(1, 4), (1, 5), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5), (4, 6), (4, 7), (5, 8), (5, 9)\big\}. \tag{11.1b}$$

Diese Beschreibung ist allerdings für die Formulierung von Optimierungsaufgaben ungeeignet.

Definition 11.3 (Inzidenzmatrix)**.** *Es sei* (V, E) *ein einfacher Digraph. Die Knotenmenge sei* $V = \{v_1, v_2, \ldots, v_m\}$ *und die Kantenmenge* $E = \{e_1, e_2, \ldots, e_n\}$*. Die zu diesem Digraphen gehörende* ***Knoten-Kanten-Inzidenzmatrix*** *(englisch:* node-edge incidence matrix*)* $A = (a_{ij})$ *hat die Dimension* $m \times n$

und ist wie folgt definiert:

$$a_{ij} = \begin{cases} -1, & \text{falls die Kante } e_j \text{ im Knoten } v_i \text{ startet,} \\ 1, & \text{falls die Kante } e_j \text{ im Knoten } v_i \text{ endet,} \\ 0, & \text{sonst.} \end{cases}$$

Wir sprechen auch kurz von der ***Inzidenzmatrix*** *des Digraphen* (V, E).

Nummerieren wir die Kanten wie sie in (11.1) aufgezählt werden, so ist die Inzidenzmatrix des Digraphen in Abbildung 11.1 gegeben durch

$$A = \begin{bmatrix} -1 & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -1 & -1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & 1 & \cdot & 1 & \cdot & -1 & -1 & -1 & \cdot & \cdot \\ \cdot & 1 & \cdot & 1 & \cdot & 1 & 1 & \cdot & \cdot & -1 & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{bmatrix} \in \mathbb{R}^{9\times 11}. \tag{11.2}$$

Der besseren Lesbarkeit wegen wurden die Nullen durch „·“ ersetzt.

Quizfrage: Können Sie eine Vermutung anstellen, was die Einträge der Matrix AA^T aussagen? Diese Matrix wird die **Laplacematrix** des Digraphen genannt. Sie hat viele interessante Eigenschaften und Anwendungen in der Graphentheorie, die wir in dieser Vorlesung aber nicht weiter betrachten.

Es sollte klar sein, dass jeder einfache Digraph durch seine Inzidenzmatrix eindeutig (bis auf Umordnung der Knoten und Kanten) beschrieben wird.

Beachte: Da jede Kante (Spalte) genau einen Anfang (Eintrag −1) und ein Ende (Eintrag +1) hat, sind alle Spaltensummen gleich null, also $\mathbf{1}^\mathsf{T}A = 0$.

Wir überlegen uns jetzt an diesem Beispiel, was das Matrix-Vektor-Produkt Ax bedeutet. Der Vektor $x \in \mathbb{R}^{11}$ steht dabei für die Warenmengen, die über die Kanten fließen. Wir betrachten die Zeile 5 des Matrix-Vektor-Produkts, also

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & -1 & -1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{11} \end{pmatrix} = x_2 + x_4 + x_6 + x_7 - x_{10} - x_{11}.$$

Im Ergebnis spielen also nur die Warenströme der an den Knoten 5 (eines der Zwischenlager) angrenzenden Kanten (mit den Nummern 2, 4, 6, 7, 10, 11) eine Rolle. Dabei werden die über die Kanten 2, 4, 6 und 7 *eingehenden* Warenströme positiv gezählt und die über die Kanten 10 und 11 *ausgehenden* Warenströme negativ.

Das Matrix-Vektor-Produkt $A\,x$ gibt also offenbar den Vektor der **Knotenbilanzen** an, der sich bei der Belegung der Kanten (Transportwege) mit den Transportmengen ergibt, die im Vektor x eingetragen sind. Mit dieser Erkenntnis können wir unser Beispiel 11.1 des **kostenminimalen Transports kostenminimalen Flusses** nun als lineare Optimierungsaufgabe formulieren. Die zu minimierenden Gesamtkosten aller Transportströme setzen sich als Summe der Kosten über die einzelnen Kanten zusammen:

$$\sum_{j=1}^{n} c_j\,x_j = c^\mathsf{T} x.$$

Dabei sind c_j die gegebenen Transportkosten pro Wareneinheit über die Kante j. Weiter sind die geforderten Bilanzen b_i aller Knoten $i = 1, \ldots, m$ gegeben. Man unterscheidet

- **Bedarfsknoten** oder **Senken** ($b_i > 0$), englisch: *demand nodes, sinks*,
- **Angebotsknoten** oder **Quellen** ($b_i < 0$), englisch: *supply nodes, sources*,
- **Durchfluss**- oder **Umladeknoten** ($b_i = 0$), englisch: *transshipment nodes*.

Die Erfüllung aller Knotenbilanzen wird durch das lineare Gleichungssystem

$$A\,x = b$$

ausgedrückt. Dieses heißen auch **Flusserhaltungsgleichungen**. Zusätzlich ist zu beachten, dass die Transportmengen über die Kanten nicht negativ sein dürfen; dies würde einer Umkehrung der Flussrichtung entsprechen. Schließlich sind eventuelle Kapazitätsbeschränkungen der einzelnen Transportwege (Kanten) einzuhalten:

$$0 \le x_i \le u_i \quad \text{für alle } i = 1, \ldots, n.$$

Definition 11.4 (Flussnetzwerk, kostenminimaler Fluss)**.**

(*i*) *Ein einfacher gerichteter Digraph mit Inzidenzmatrix $A \in \mathbb{R}^{m\times n}$, **Kantenkapazitäten** $u \in \mathbb{R}^n$ und **Knotenbilanzen** $b \in \mathbb{R}^m$ wird als **Transportnetzwerk** oder **Flussnetzwerk** bezeichnet.*

(*ii*) *Eine Kantenbelegungsvektor $x \in \mathbb{R}^n$, der die **Erhaltungsbedingung** $A\,x = b$ erfüllt, heißt ein **Fluss** oder **Flussvektor** auf diesem Netzwerk. Ein Fluss heißt **zulässig**, wenn zusätzlich die **Kapazitätsbeschränkungen** $0 \le x \le u$ erfüllt sind.*

(*iii*) *Eine lineare Optimierungsaufgabe der Form*

$$\begin{aligned} &\textit{Minimiere} && c^\mathsf{T} x \quad \textit{über } x \in \mathbb{R}^n \\ &\textit{unter} && A\,x = b \\ &\textit{sowie} && 0 \le x \le u \end{aligned} \tag{11.3}$$

*mit gegebenem **Kantenkostenvektor** $c \in \mathbb{R}^n$ heißt eine Aufgabe des **kostenminimalen Transports** oder des **kostenminimalen Flusses**. Einige oder alle Komponenten der oberen Schranke u dürfen dabei $+\infty$ sein, was den Fall „ohne Beschränkung“ repräsentiert.*

Beispiel 11.5. *Für den durch die Inzidenzmatrix (11.2) dargerstellten Digraphen aus Beispiel 11.1 und die Beispieldaten*

$$b = (-100,\ -200,\ -300,\ 0,\ 0,\ 150,\ 150,\ 150,\ 150)^\mathsf{T},$$
$$c = (0.8,\ 2.0,\ 2.5,\ 1.0,\ 1.2,\ 2.0,\ 1.0,\ 1.0,\ 1.0,\ 1.0,\ 1.0)^\mathsf{T},$$
$$u = (\infty,\ \infty,\ \infty,\ \infty,\ \infty,\ \infty,\ \infty,\ \infty,\ \infty,\ \infty,\ \infty)^\mathsf{T},$$

erhalten wir den Fluss

$$x^* = (100,\ 0,\ 0,\ 200,\ 200,\ 100,\ 0,\ 150,\ 150,\ 150,\ 150)^\mathsf{T} \tag{11.4}$$

als optimale Lösung der Aufgabe (11.3) des kostenminimalen Transports. Die Lösung ist in Abbildung 11.2 dargestellt. Die zugehörigen Transportkosten betragen $c^\mathsf{T}x^* = 1320$. Die Lösung wurde unter Verwendung des Simplex-Verfahrens in `linprog` *aus dem Modul* `scipy.optimize` *bestimmt, siehe Abbildungen 11.3 und 11.4 für den* Python*-Code.*

Beachte: Die Matrix A der Nebenbedingung $Ax = b$ hat hier $m = 9$ Zeilen, effektiv jedoch nur $m = 8$, da $\operatorname{Rang}(A) = 8$ beträgt. Da A außerdem $n = 11$ Spalten besitzt, hat jede Nichtbasis im Simplex-Verfahren die Mächtigkeit $|N| = 3$. Jede Ecke und damit auch die von `linprog` gefundene Lösung x^* besitzt damit mindestens drei Nulleinträge, d. h. Kanten, über die nichts transportiert wird. Wie erwartet trifft das insbesondere auf die optimale Ecke x^* zu.

Damit eine Aufgabe der Form (11.3) überhaupt zulässige Punkte (Flüsse) besitzt, muss notwendig $\mathbf{1}^\mathsf{T}b = 0$ gelten, denn

$$Ax = b \quad \Rightarrow \underbrace{\mathbf{1}^\mathsf{T}A}_{=0}\, x = \mathbf{1}^\mathsf{T}b. \tag{11.5}$$

Die Bedarfe und Angebote in einem Transportnetzwerk müssen sich also ausgleichen. Sollte in einem Transportnetzwerk $\mathbf{1}^\mathsf{T}b < 0$ gelten, dann liegt ein **Überangebot** des zu transportierenden Gutes vor, wodurch die Aufgabe (11.3) unzulässig wird. Um Abhilfe zu schaffen, wird ein zusätzlicher Knoten eingeführt, der einem künstlichen Abnehmer entspricht, dessen Bedarf gerade das Überangebot kompensiert. Diese **künstliche Senke** wird mit dann z. B. mit allen Angebotsknoten durch neue Kanten verbunden, und es werden Kosten für diese Kanten gesetzt.

Quizfrage: Was bedeutet es in Beispiel 11.1, wenn die künstliche Senke mit den drei Produktionsstätten verbunden wird? Und was bedeutet es, wenn sie mit den vier Verkaufsstellen verbunden wird? Wofür könnten dabei z. B. Kosten anfallen? Was bedeutet es, mehrere künstliche Senken in den Digraphen aufzunehmen?

Im Fall $\mathbf{1}^\mathsf{T}b > 0$ liegt dagegen ein **Mangel** an dem zu transportierenden Gut vor. Durch Schaffung eines **zusätzlichen Angebotsknotens**, den wir mit geeigneten Knoten im Netzwerk verbinden, können wir den Mangel kompensieren.

Quizfrage: Was bedeutet es in Beispiel 11.1, wenn der zusätzliche Angebotsknoten mit den drei Produktionsstätten verbunden wird? Und was bedeutet es, wenn er mit den vier Verkaufsstellen verbunden wird? Wofür stehen die dabei anfallenden Kosten? Was bedeutet es, mehrere Angebotsknoten zusätzlich in den Digraphen aufzunehmen?

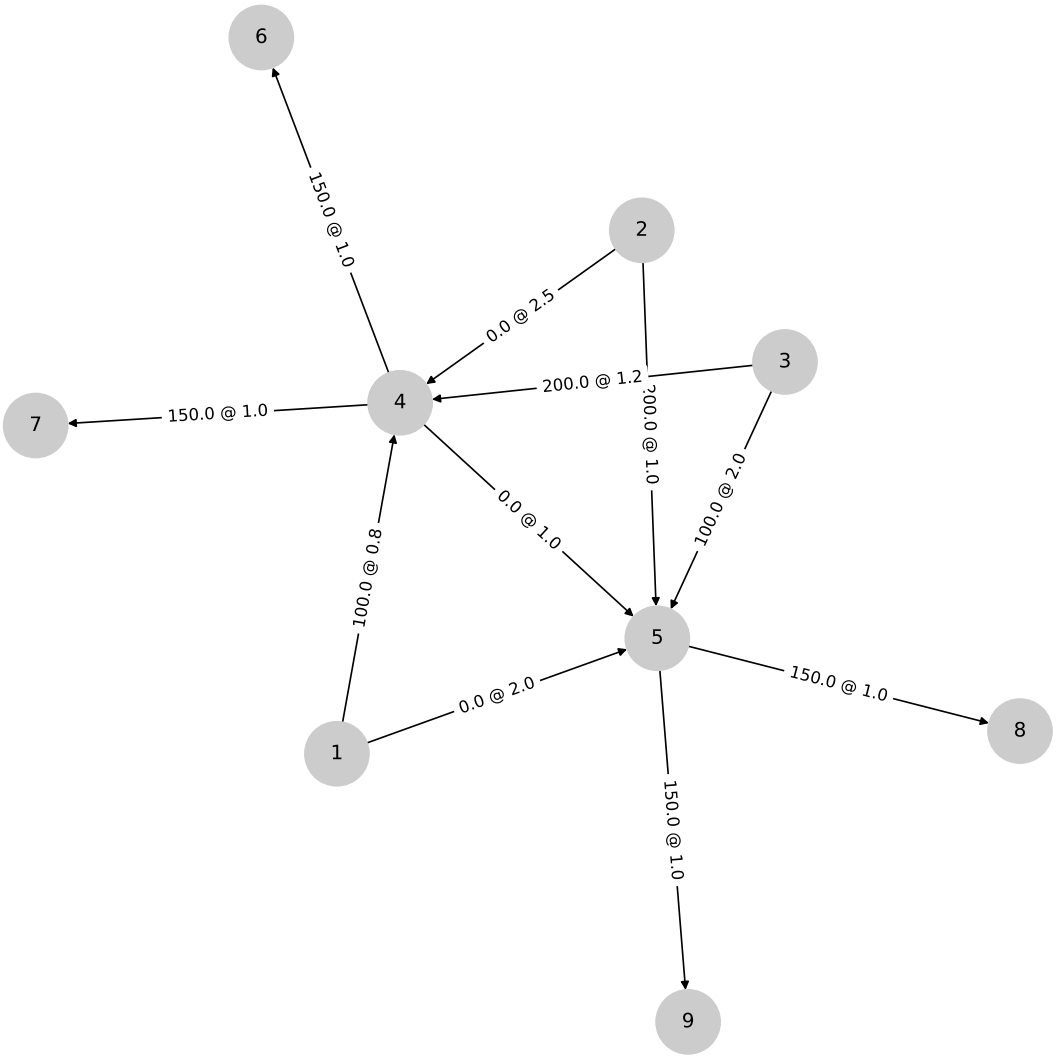


Abbildung 11.2: Eine optimale Lösung von Beispiel 11.5.

Bemerkung 11.6. *Das Simplex-Verfahren ist nicht die effizienteste Lösungsmöglichkeit für Aufgaben kostenminimaler Flüsse auf Transportnetzwerken. Es gibt dafür eine spezielle Variante, das* ***Netzwerk-Simplex-Verfahren****, siehe etwa Gerdts, Lempio, 2011, Abschnitt 4.2 oder Vanderbei, 2008, Kapitel 14. Diese nutzt aus, dass die Matrix A eine Inzidenzmatrix ist, die nur aus Einträgen $\{0, \pm 1\}$ besteht. Die beiden aufwändigsten Schritte, die Lösung der linearen Gleichungssysteme in Zeile 2 und Zeile 7 von Algorithmus 9.1 bzw. Algorithmus 9.2, erfordern dabei nur Additionen und Subtraktionen von Vektoren.*

```
# This code solves a minimal cost network flow problem.

# Resolve the dependencies.
from scipy.optimize import linprog
import numpy as np
import networkx as nx

# Construct the digraph using vertices and edges. Notice that networkx may
# shuffle the edges, so we attach the cost to the edges.
vertices = range(1,10)
edges = [(1,4), (1,5), (2,4), (2,5), (3,4), (3,5), (4,5), (4,6), (4,7), (5,8), (5,9)]
costs = np.array([0.8, 2.0, 2.5, 1.0, 1.2, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0])
edgesWithCosts = [(v1, v2, {"costs": f'{c}'}) for ((v1, v2), c) in zip(edges, costs)]
G = nx.DiGraph()
G.add_nodes_from(vertices)
G.add_edges_from(edgesWithCosts)

# Setup the incidence matrix.
A = nx.incidence_matrix(G, oriented = True)
A = A.toarray()

# Retrieve the cost vector from the edge data in the order stored in the digraph.
c = list(zip(*G.edges.data('costs')))[2]

# Setup the vector of vertex balances.
b = np.array([-100, -200, -300, 0, 0, 150, 150, 150, 150])

# Setup the lower and upper bounds.
bounds = [(0, None) for j in edges]

# Call linprog to solve the problem.
result = linprog(c, A_eq = A, b_eq = b, bounds = bounds, method = 'simplex')
```

Abbildung 11.3: Lösung einer Aufgabe des kostenminimalen Transports.

§ 12 Ganzzahlige Lösungen

Im obigen Beispiel 11.1 hat sich der optimale Fluss x^* über jede Kante als ganzzahlig herausgestellt, siehe (11.4). Dies ist bei vielen Aufgabenstellungen auf Transportnetzwerken erwünscht oder sogar erforderlich, weil sich die verwendeten Transporteinheiten (Paletten, LKW etc.) nicht teilen lassen. Es stellt sich die Frage, wie man die Ganzzahligkeit der Lösung einer linearen Optimierungsaufgabe garantieren kann, ohne sie explizit zu fordern. Da das Simplex-Verfahren auf den Ecken der zulässigen Menge

$$\{x \in \mathbb{R}^n \mid A\,x = b,\ 0 \le x \le u\} \tag{12.1}$$

```
# Attach attributes to the graph's edges.
edgeFlow = dict(zip(G.edges, result.x))
edgeFlowAtCosts = dict([((v1,v2), f'{x} @ {c}')
  for (v1, v2, c), x in zip(G.edges.data('costs'), result.x)])

# Layout the digraph (assign vertex positions).
positions = nx.nx_agraph.graphviz_layout(G, prog = "neato")

# Resolve further dependencies.
import matplotlib.pyplot as plt
import tikzplotlib

# Show and export the digraph showing the optimal flow.
plt.figure(figsize = (10,10))
nx.draw(G, positions, with_labels = True, node_color = [[0.8] * 3], node_size = 1500)
nx.draw_networkx_edge_labels(G, positions, edge_labels = edgeFlowAtCosts)
#  plt.savefig("../graphs/solveOptimalTransport.pdf")
#  tikzplotlib.save("../graphs/solveOptimalTransport.tex")
plt.show()
```

Abbildung 11.4: Darstellung der Lösung einer Aufgabe des kostenminimalen Transports.

arbeitet und (Lösbarkeit der Aufgabe vorausgesetzt) eine der Ecken als Lösung zurückgibt, geht es um die Frage, wann die Ecken dieses Polyeders *alle* ausschließlich ganzzahlige Koordinaten haben. Beim Mozartproblem zum Beispiel hatte die zulässige Menge diese Eigenschaft nicht, siehe Abbildung 6.3.

Die nachfolgende Definition wurde allgemeiner gefasst.

Definition 12.1 (Unimodularität und totale Unimodularität)**.**

(*i*) *Eine Matrix* $A \in \mathbb{Z}^{m\times n}$ *heißt* ***unimodular****, wenn jede ihrer quadratischen Untermatrizen* $\widehat{A}$ *maximaler Dimension* $r = \min\{m, n\}$ *die Eigenschaft* $\det(\widehat{A}) \in \{0, \pm 1\}$ *besitzt.*

(*ii*) *Eine Matrix* $A \in \mathbb{Z}^{m\times n}$ *heißt* ***total unimodular****, wenn jede ihrer quadratischen Untermatrizen* $\widehat{A}$ *der Dimension* $1 \leq r \leq \min\{m, n\}$ *die Eigenschaft* $\det(\widehat{A}) \in \{0, \pm 1\}$ *besitzt.*

Eine Matrix $\widehat{A}$ heißt dabei eine **Untermatrix** von A, wenn sie durch eine Auswahl gewisser Zeilen und Spalten von A gebildet wird.

Beachte: A total unimodular $\Rightarrow$ A unimodular.

Der nachfolgende Satz wurde allgemeiner gefasst.

Satz 12.2 (Charakterisierung unimodularer Matrizen)**.** *Eine Matrix* $A \in \mathbb{Z}^{m\times n}$ *ist genau dann unimodular, wenn die Inverse jeder regulären Untermatrix* $\widehat{A}$ *der Dimension* $r = \min\{m, n\}$ *nur ganzzahlige Einträge besitzt.*

Beweis. Es sei $A \in \mathbb{Z}^{m\times n}$. Es sei zunächst A unimodular und $\widehat{A}$ eine reguläre Untermatrix der Dimension $r = \min\{m, n\}$. Es gilt also $\det(\widehat{A}) = 1$ oder $\det(\widehat{A}) = -1$. Wir betrachten den Fall $r = m \leq n$. Die Einträge von $(\widehat{A})^{-1}$ können mit Hilfe der Cramerschen Regel wie folgt dargestellt werden:

$$\left((\widehat{A})^{-1}\right)_{ij} = \frac{1}{\det(\widehat{A})} \det\begin{bmatrix} \widehat{a}_1 & \cdots & \widehat{a}_{i-1} & e_j & \widehat{a}_{i+1} & \cdots & \widehat{a}_m \end{bmatrix}.$$

Die Matrix im Zähler hat nur ganzzahlige Einträge, also ist auch ihre Determinante ganzzahlig. (**Quizfrage:** Warum eigentlich?) Damit ist auch $\left((\widehat{A})^{-1}\right)_{ij}$ ganzzahlig. Im Fall $r = n \leq m$ argumentiert man ähnlich. (**Quizfrage:** Wie genau?)

Umgekehrt habe nun A die Eigenschaft, dass jede reguläre Untermatrix $\widehat{A}$ der Dimension $r = \min\{m, n\}$ eine ganzzahlige Inverse besitzt. Für jede solche Untermatrix sind $\det(\widehat{A})$ und $\det\left((\widehat{A})^{-1}\right)$ beide ganzzahlig. Wegen

$$\det(\widehat{A})\ \det\left((\widehat{A})^{-1}\right) = \det\left(\widehat{A}\,(\widehat{A})^{-1}\right) = \det(\mathrm{Id}) = 1$$

bleiben nur die Möglichkeiten $\det(\widehat{A}) = \det\left((\widehat{A})^{-1}\right) = 1$ oder $\det(\widehat{A}) = \det\left((\widehat{A})^{-1}\right) = -1$. Andererseits erfüllt jede Untermatrix $\widehat{A}$ der Dimension r, die nicht regulär ist, $\det(\widehat{A}) = 0$. Also ist A unimodular. □

Der folgende Satz wurde eingefügt.

Satz 12.3 (Bedeutung unimodularer Matrizen)**.** *Es sei $A \in \mathbb{Z}^{m\times n}$ mit $\mathrm{Rang}(A) = m$. Dann sind die folgenden Aussagen äquivalent:*

(*i*) *Die Matrix A ist unimodular.*

(*ii*) *Für jeden Vektor $b \in \mathbb{Z}^m$ besitzt das Polyeder in Normalform*

$$P = \{x \in \mathbb{R}^n \mid A\,x = b,\ x \geq 0\}$$

nur ganzzahlige Ecken.

Beweis. Der Beweis findet sich als Übungsaufgabe 2 auf Übungsblatt 8. □

Wir wenden uns nun der Bedeutung der totalen Unimodularität zu.

Lemma 12.4 (Totale Unimodularität verwandter Matrizen)**.** *Es sei $A \in \mathbb{Z}^{m\times n}$. Dann sind die folgenden Aussagen äquivalent:*

(*i*) *A ist total unimodular.*

(*ii*) *$-A$ ist total unimodular.*

(*iii*) *A^T ist total unimodular.*

(*iv*) $\begin{bmatrix} A, & -A \end{bmatrix}$ *ist total unimodular.*

(*v*) $\begin{bmatrix} A, & \mathrm{Id} \end{bmatrix}$ *ist total unimodular.*

Beweis. Der Beweis findet sich als Übungsaufgabe 1 auf Übungsblatt 8. □

Die Aussage des folgenden Satzes wurde korrigiert.

Satz 12.5 (Bedeutung total unimodularer Matrizen)**.** *Es sei $A \in \mathbb{Z}^{m \times n}$ mit* $\mathrm{Rang}(A) = m$*. Dann sind die folgenden Aussagen äquivalent:*

(*i*) *Die Matrix A ist total unimodular.*

(*ii*) *Für jeden Vektor $b \in \mathbb{Z}^m$ besitzt das Polyeder in kanonischer Form*

$$P = \{x \in \mathbb{R}^n \mid A\,x \leq b,\ x \geq 0\}$$

nur ganzzahlige Ecken.

Beweis. Der Beweis findet sich als Übungsaufgabe 2 auf Übungsblatt 8. □

Man kann zeigen, dass Inzidenzmatrizen A für einfache Digraphen total unimodular sind, siehe zum Beispiel Schrijver, 2003, Theorem 13.9. Aus Satz 12.3 folgt daher, dass alle Ecken im zulässigen Polyeder (11.3) ganzzahlig sind, solange nur die Knotenbilanzen $b \in \mathbb{Z}^m$ und die Kantenkapazitäten $u \in \mathbb{Z}^n$ jeweils ganzzahlig sind. Da das Simplex-Verfahren auf den Ecken arbeitet, erhält man dann (im Falle der Lösbarkeit) automatisch eine ganzzahlige Lösung für Aufgaben des kostenminimalen Flusses, wie wir es z. B. mit (11.4) beobachtet hatten.

Beachte: Auf den Kostenvektor c kommt es dabei nicht an!

Quizfrage: Wenn man statt mit ganzzahligen Lösungen mit „Halben“ (beispielsweise mit halben Paletten, halben Litern etc.) arbeiten will, also mit Lösungen in $\mathbb{Z}^n/2 = \{z/2 \mid z \in \mathbb{Z}^n\}$, wie kann man die totale Unimodularität der Matrix dann nutzen?

Bemerkung 12.6. *Die totale Unimodularität von Inzidenzmatrizen einfacher Digraphen führt auch bei zu Aufgaben des kostenminimalen Flusses verwandten linearen Optimierungsaufgaben zu ganzzahligen Lösungen, darunter Aufgaben des maximalen Flusses, Kürzeste-Wege-Aufgaben. Auch für Zuordnungsprobleme, die mit ungerichteten Graphen arbeiten, ist die Inzidenzmatrix total unimodular. Bei dieser wichtigen Klasse linearer Optimierungsaufgaben erhält man also quasi ganzzahlige Lösungen „umsonst“, ohne weiteres Zutun.*

Beispiele für Transportprobleme, bei denen die Matrix A, die die Gleichungsnebenbedingung beschreibt, *nicht* total unimodular ist, sind beispielsweise **Mehrgütertransportprobleme** (**Mehrgüterflussprobleme**, englisch: *multi-commodity flow problems*). Bei diesen müssen *verschiedene* Güter über ein *gemeinsames* Netzwerk transportiert werden, wobei sich die Güter die *Kantenkapazitäten* jeweils *teilen* müssen. Die Transportkosten für jede Kante sind wie in Beispiel 11.1 proportional zu der darüber transportierten Warenmenge und können für verschiedene Güter unterschiedlich sein. Beispielsweise für zwei Güter erhält man die Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & (c^{(1)})^\mathsf{T} x^{(1)} + (c^{(2)})^\mathsf{T} x^{(2)} && \text{über } \left(x^{(1)}, x^{(2)}\right) \in \mathbb{R}^n \times \mathbb{R}^n \\ \text{unter} \quad & \begin{cases} A_0\, x^{(1)} & = b^{(1)} & \text{(Knotenbilanzen Transportgut 1)} \\ A_0\, x^{(2)} & = b^{(2)} & \text{(Knotenbilanzen Transportgut 2)} \\ x^{(1)} + x^{(2)} & \le u & \text{(Kapazitätsbeschränkung)} \end{cases} \\ \text{sowie} \quad & x^{(1)} \ge 0, \quad x^{(2)} \ge 0. \end{aligned} \tag{12.2}$$

Obwohl A_0 total unimodular ist, trifft das auf die Matrix

$$\begin{bmatrix} A_0 & 0 \\ 0 & A_0 \\ \mathrm{Id} & \mathrm{Id} \end{bmatrix}$$

nicht mehr zu!

Benötigt man in solchen Aufgaben die Ganzzahligkeit $x \in \mathbb{Z}^n$ der Lösung, so muss man sie extra fordern und Verfahren der diskreten Optimierung wie **branch and bound** anwenden, vgl. Bemerkung 9.3.

Ende der Woche 8

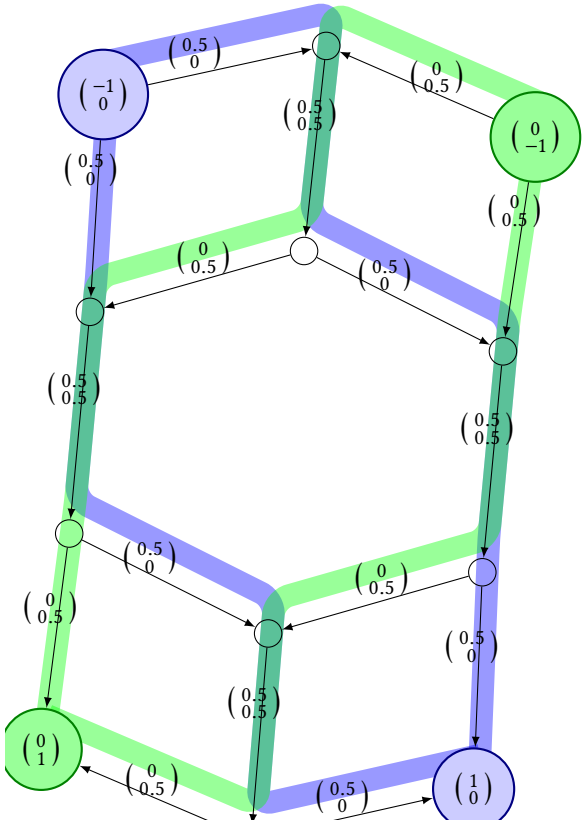


Abbildung 12.1: Darstellung eines Transportnetzwerks für ein Zweigüterflussproblem. Die Menge eins des ersten Gutes (blau) soll von der Quelle oben links zur Senke unten rechts transportiert werden. Dieselbe Menge des zweiten Gutes (grün) soll von der Quelle oben rechts zur Senke unten links transportiert werden. Die Kantenkapazitäten sind alle gleich eins. Der einzige zulässige Fluss ist an den jeweiligen Kanten eingezeichnet. Er ist nicht ganzzahlig.

Kapitel 3 Konvexe Optimierung

§ 13 Einführung

Unser Ziel ist es auch in diesem Kapitel wieder, notwendige und hinreichende Optimalitätsbedingungen herzuleiten und grundlegende Verfahren für konvexe Optimierungsaufgaben kennenzulernen. Charakteristisch für die konvexe Optimierung wird das Zusammenspiel zwischen Eigenschaften konvexer Mengen und konvexer Funktionen sein.

§ 13.1 Konvexe Mengen

Literatur: Geiger, Kanzow, 2002, Kapitel 2.1.1

Definition 13.1 (Konvexe Menge).
*Eine Menge $C \subseteq \mathbb{R}^n$ heißt **konvex**, wenn mit $x, y \in C$ und $\alpha \in [0,1]$ auch $\alpha x + (1-\alpha) y \in C$ ist, also die gesamte Verbindungsstrecke von x und y.*

Abbildung 13.1: Konvexe Mengen (blau) und eine nichtkonvexe Menge (rot) im $\mathbb{R}^2$.

Beispiel 13.2 (Konvexe Mengen).
Wichtige konvexe Mengen sind:

(*i*) *offene Kugeln* $B_\varepsilon(y) = \{x \in \mathbb{R}^n \mid \|x - y\| < \varepsilon\}$

(*ii*) *abgeschlossene Kugeln* $\overline{B_\varepsilon(y)} = \{x \in \mathbb{R}^n \mid \|x - y\| \leq \varepsilon\}$

(*iii*) *Hyperebenen* $H(a, \beta) = \{x \in \mathbb{R}^n \mid a^\top x = \beta\}$ *mit* $a \in \mathbb{R}^n$, $a \neq 0$ *und* $\beta \in \mathbb{R}$

(*iv*) *offene Halbräume* $\{x \in \mathbb{R}^n \mid a^\top x < \beta\}$

(*v*) *abgeschlossene Halbräume* $H^-(a,\beta) = \{x \in \mathbb{R}^n \mid a^\top x \leq \beta\}$ *und* $H^+(a,\beta) = \{x \in \mathbb{R}^n \mid a^\top x \geq \beta\}$

(*vi*) *das Einheitssimplex im* $\mathbb{R}^n$ *(siehe Abbildung 13.2)*

$$\Delta_n = \left\{ x \in \mathbb{R}^n \;\middle|\; \sum_{i=1}^{n} x_i \leq 1,\ x_i \geq 0,\ i = 1, \ldots, n \right\}.$$

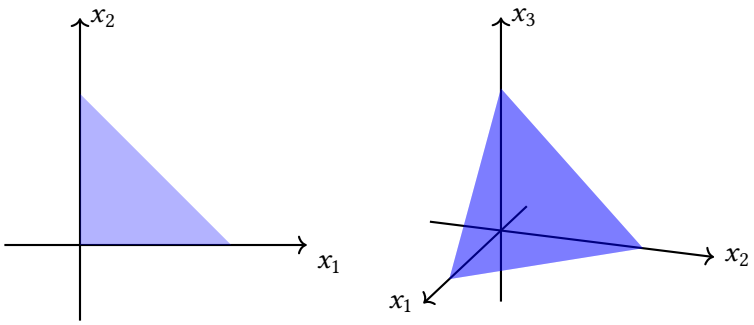


Abbildung 13.2: Einheitssimplizes im $\mathbb{R}^2$ und $\mathbb{R}^3$.

Quizfrage: Was sind die konvexen Teilmengen von $\mathbb{R}$?

Satz 13.3 (Operationen auf konvexen Mengen)**.**

(*i*) *Es sei* $\{C_j\}_{j \in J}$ *eine beliebige Familie konvexer Mengen in* $\mathbb{R}^n$*. Dann ist* $\bigcap_{j \in J} C_j$ *konvex.*

(*ii*) *Es seien* $C_i \subseteq \mathbb{R}^{n_i}$ *konvex,* $i = 1, \ldots, k$*. Dann ist das kartesische Produkt* $C_1 \times \cdots \times C_k$ *konvex in* $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$*.*

(*iii*) *Es sei* $f \colon \mathbb{R}^n \to \mathbb{R}^m$ *eine (affin-)lineare Abbildung, also* $f(x) = Ax + b$*, und* $C \subseteq \mathbb{R}^n$ *und* $D \subseteq \mathbb{R}^m$ *konvexe Mengen. Dann sind das Bild* $f(C) \subseteq \mathbb{R}^m$ *und das Urbild* $f^{-1}(D) \subseteq \mathbb{R}^n$ *konvex.*

(*iv*) *Sind* $C_1, C_2 \subseteq \mathbb{R}^n$ *konvex, dann sind*

$$\beta C_1 = \{\beta x_1 \mid x_1 \in C_1\} \quad \text{für } \beta \in \mathbb{R}$$

sowie die ***Minkowski-Summe***

$$C_1 + C_2 = \{x_1 + x_2 \mid x_1 \in C_1,\ x_2 \in C_2\}$$

konvex. Insbesondere sind Verschiebungen (Translationen) konvexer Mengen, die sich ergeben, wenn man in der Minkowski-Summe $C_2 = \{x\}$ *setzt, konvex. Man schreibt dann auch* $C_1 + x_2$ *statt* $C_1 + \{x_2\}$*.*

Beachte: Eine Menge $C \subseteq \mathbb{R}^n$ ist genau dann konvex, wenn $\alpha C + (1-\alpha) C \subseteq C$ für alle $\alpha \in [0,1]$ gilt.

Beweis. Der Beweis findet sich als Übungsaufgabe 1 auf Übungsblatt 9. □

Definition 13.4 (Konvexkombination)**.**

(*i*) $x \in \mathbb{R}^n$ *heißt eine* ***Konvexkombination*** *von* $x_1, \dots, x_m \in \mathbb{R}^n$*, falls* $x = \sum_{i=1}^m \alpha_i x_i$ *gilt mit Koeffizienten* $\alpha_i \geq 0$ *und* $\sum_{i=1}^m \alpha_i = 1$*. Eine solche Konvexkombination heißt* ***echt****, wenn alle* $\alpha_i > 0$ *sind.*

(*ii*) *Ist* $M \subseteq \mathbb{R}^n$ *irgendeine (nicht notwendigerweise endliche) Menge, so heißt* x *eine Konvexkombination von* M*, wenn* x *eine Konvexkombination von endlich vielen Vektoren* $x_0, \dots, x_m \in M$ *ist.*

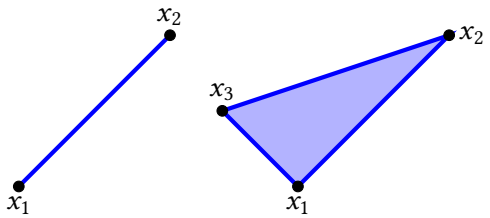


Abbildung 13.3: Konvexkombinationen von zwei und drei Punkten in $\mathbb{R}^2$. Bei zwei Punkten kann man auch $\alpha\, x_1 + (1-\alpha)\, x_2$ mit $\alpha \in [0,1]$ schreiben.

Lemma 13.5 (Charakterisierung konvexer Mengen)**.**
Eine Menge $M \subseteq \mathbb{R}^n$ *ist genau dann konvex, wenn sie alle Konvexkombinationen von* M *enthält.*

Beweis. „⇒“: Es sei M konvex. Für $m \in \mathbb{N}$ und $x_1, \dots, x_m \in M$ sowie $\alpha_1, \dots, \alpha_m \geq 0$ mit $\sum_{i=1}^m \alpha_i = 1$ ist zu zeigen: $x = \sum_{i=1}^m \alpha_i x_i \in M$.

Induktion nach m: Für $m = 1, 2$ ist die Behauptung erfüllt. Es sei bereits gezeigt, dass M alle Konvexkombinationen von höchstens m Elementen enthält.

Schluss auf $m+1$: Es seien $\alpha_i \geq 0$, $\sum_{i=1}^{m+1} \alpha_i = 1$ und $x = \sum_{i=1}^{m+1} \alpha_i x_i$. O. B. d. A. gilt $\alpha_{m+1} < 1$. (Ansonsten ist $x = x_{m+1}$ und nichts zu zeigen.) Setze $\beta_i := \frac{\alpha_i}{1-\alpha_{m+1}}$ für $i = 1, \dots, m$. Dann ist $\beta_i \geq 0$ und $\sum_{i=1}^m \beta_i = 1$. Der Vektor $y = \sum_{i=1}^m \beta_i x_i$ gehört zu M, also auch $x = (1-\alpha_{m+1})\, y + \alpha_{m+1}\, x_{m+1}$.

„⇐“: Es seien $x_1, x_2 \in M$. Nach Voraussetzung enthält M alle Konvexkombinationen $\alpha\, x_1 + (1-\alpha)\, x_2$ mit $\alpha \in [0,1]$, d. h., M ist konvex. □

Definition 13.6 (Konvexe Hülle)**.**
Es sei $M \subseteq \mathbb{R}^n$*. Der Durchschnitt aller konvexen Teilmengen von* $\mathbb{R}^n$*, die* M *enthalten, also*

$$\operatorname{conv}(M) = \bigcap \big\{ C \subseteq \mathbb{R}^n \,\big|\, C \text{ ist konvex und } M \subseteq C \big\}, \tag{13.1}$$

heißt die ***konvexe Hülle*** *von* M*.* $\operatorname{conv}(M)$ *ist also die kleinste konvexe Menge, die* M *enthält.*

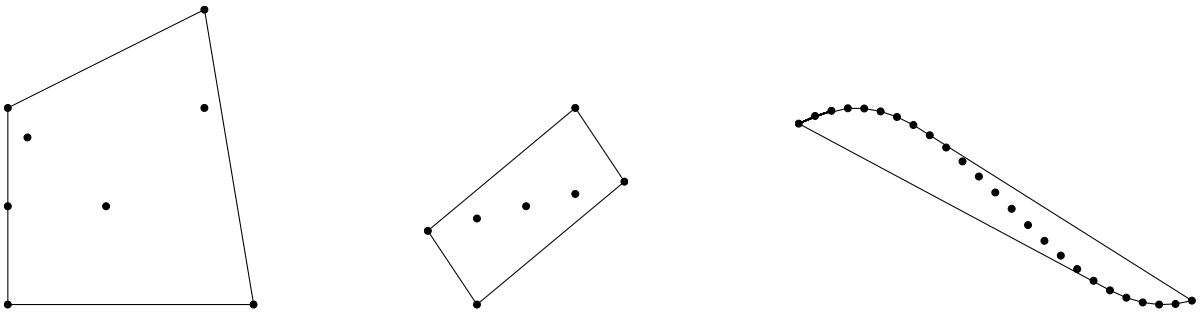


Abbildung 13.4: Konvexe Hüllen einiger Punktmengen in $\mathbb{R}^2$.

Beachte: Es gilt $\underbrace{M \subseteq \operatorname{conv}(M)}_{\text{„Hülle“ konvex}}$, daher der Name **konvexe Hülle**.

Lemma 13.7 (Charakterisierung der konvexen Hülle)**.**
Es sei $M \subseteq \mathbb{R}^n$. Dann gilt:

$$\operatorname{conv}(M) = \textit{Menge aller Konvexkombinationen von } M.$$

Beweis. Es sei C die Menge aller Konvexkombinationen von M. Natürlich gilt dann $M \subseteq C$. Im Fall $M = \emptyset$ ist nichts zu zeigen, weil dann auch $C = \emptyset$ ist. Wir gehen also jetzt von $M \neq \emptyset$ aus.

„$\operatorname{conv}(M) \subseteq C$“: Wir zeigen: C ist konvex. Damit kommt diese Menge im Durchschnitt (13.1) vor, also gilt $\operatorname{conv}(M) \subseteq C$.

Es seien $x, y \in C$, also gibt es Zahlen $m, \ell \in \mathbb{N}$ und $\beta_1, \ldots, \beta_m \geq 0$ sowie $\gamma_1, \ldots, \gamma_\ell \geq 0$ mit $\sum_{i=1}^m \beta_i = 1$ und $\sum_{j=1}^\ell \gamma_j = 1$, sodass $x = \sum_{i=1}^m \beta_i x_i$ und $y = \sum_{j=1}^\ell \gamma_j y_j$ gelten mit irgendwelchen $x_1, \ldots, x_m \in M$ und $y_1, \ldots, y_\ell \in M$. Es sei $\alpha \in [0, 1]$. Dann gilt

$$\alpha x + (1 - \alpha) y = \alpha \sum_{i=1}^m \beta_i x_i + (1 - \alpha) \sum_{j=1}^\ell \gamma_j y_j,$$

d. h., $\alpha x + (1 - \alpha) y$ ist Linearkombination der $\{x_i\}_{i=1}^m \cup \{y_j\}_{j=1}^\ell$. Die Koeffizienten sind ≥ 0 und ergeben in der Summe 1. Damit ist $\alpha x + (1 - \alpha) y \in C$, also C konvex.

„$\operatorname{conv}(M) \supseteq C$“: Es sei $x \in C$, also eine Konvexkombination von M. Wegen $M \subseteq \operatorname{conv}(M)$ ist x auch eine Konvexkombination von $\operatorname{conv}(M)$. $\operatorname{conv}(M)$ ist konvex, stimmt also nach Lemma 13.5 mit der Menge seiner Konvexkombinationen überein. Also ist $x \in \operatorname{conv}(M)$. □

Beachte: M ist konvex $\Leftrightarrow M = \operatorname{conv}(M)$.

§ 13.2 Konvexe Funktionen

Literatur: Geiger, Kanzow, 2002, Kapitel 2.1.2

Definition 13.8 (Konvexe Funktion).
Es sei $C \subseteq \mathbb{R}^n$ konvex. Eine Funktion $f : C \to \mathbb{R}$ heißt

(*i*) ***konvex*** *(englisch:* convex*) auf C, falls*

$$f(\alpha x + (1-\alpha)\, y) \leq \alpha f(x) + (1-\alpha) f(y) \tag{13.2}$$

für alle $x, y \in C$ und $\alpha \in [0,1]$ gilt.

(*ii*) ***strikt konvex*** *(englisch:* strictly convex*) auf C, falls*

$$f(\alpha x + (1-\alpha)\, y) < \alpha f(x) + (1-\alpha) f(y) \tag{13.3}$$

für alle $x, y \in C$ mit $x \neq y$ und $\alpha \in (0,1)$ gilt.

(*iii*) ***μ-stark konvex*** *(englisch: μ-*strongly convex*) oder* ***stark konvex*** *mit Parameter $\mu > 0$ auf C, falls*

$$f(\alpha x + (1-\alpha)\, y) + \frac{\mu}{2}\, \alpha\, (1-\alpha) \|x - y\|^2 \leq \alpha f(x) + (1-\alpha) f(y) \tag{13.4}$$

für alle $x, y \in C$ und $\alpha \in [0,1]$ gilt.

(*iv*) ***konkav*** *(englisch:* concave*) bzw.* ***strikt konkav*** *bzw.* ***stark konkav*** *auf C, wenn $-f$ konvex bzw. strikt konvex bzw. stark konvex auf C ist.*

Die Bedingung (13.2) können wir so lesen, dass der Funktionswert an einer Konvexkombination immer kleiner oder gleich der Konvexkombination der Funktionswerte ist. Anschaulich bedeutet (13.2) damit, dass der Funktionsgraph von f unterhalb aller Sehnen verläuft, siehe Abbildung 13.5.

Beachte: Zur Definition einer konvexen Funktion gehört notwendigerweise auch eine *konvexe Menge.*

Beachte: In Geiger, Kanzow, 1999, Definition 3.2 wird die Bedingung (*iii*) als **gleichmäßige Konvexität** (englisch: *uniform convexity*) bezeichnet. Das ist in der Literatur leider nicht einheitlich.

Quizfrage: Was hat die starke Konvexität von f mit der Konvexität von $f(\cdot) - \frac{\mu}{2}\|\cdot\|^2$ zu tun?

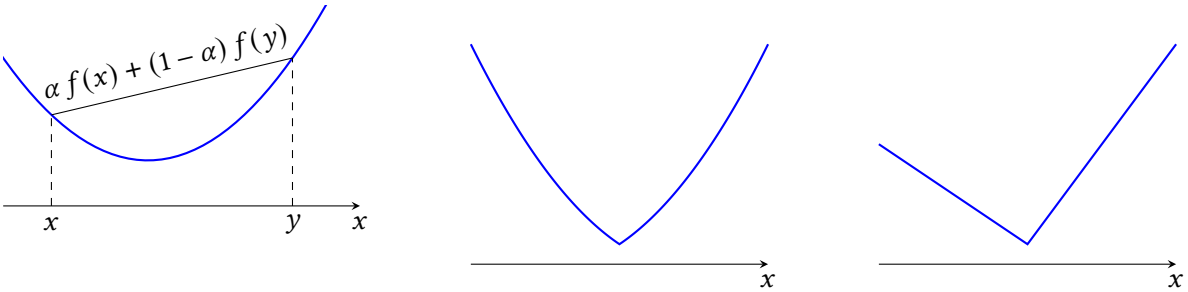


Abbildung 13.5: Beispiele strikt konvexer Funktionen (links und Mitte) und konvexe, aber nicht strikt konvexe Funktion (rechts) auf Intervallen in $\mathbb{R}$.

Beachte: f stark konvex $\Rightarrow$ f strikt konvex $\Rightarrow$ f konvex.

Beispiel 13.9 (Beispiele konvexer Funktionen)**.**

(*i*) *Die affin-lineare Funktion* $f(x) = a^\top x + \beta$ *ist gleichzeitig konvex und konkav auf* $\mathbb{R}^n$.

(*ii*) *Die quadratische Funktion* $f(x) = \frac{1}{2}x^\top Q\, x + c^\top x + \gamma$ *mit symmetrischer Matrix* $Q \in \mathbb{R}^{n\times n}$ *ist*

- *konvex* $\quad\Leftrightarrow\quad$ Q *ist positiv semidefinit,*
- μ*-stark konvex* $\quad\Leftrightarrow\quad$ Q *ist positiv definit mit kleinstem Eigenwert* $\lambda_{\min}(Q) \geq \mu > 0$.

(*iii*) $f(x) = \|x - z\|$ *mit* $z \in \mathbb{R}^n$ *ist konvex auf* $\mathbb{R}^n$*, aber nicht strikt konvex.*

(*iv*) $f(x) = \|x - z\|^2$ *mit* $z \in \mathbb{R}^n$ *ist stark konvex auf* $\mathbb{R}^n$ *mit* $\mu = 2$.

(*v*) $f(x) = \|x - z\|^4$ *mit* $z \in \mathbb{R}^n$ *ist strikt konvex auf* $\mathbb{R}^n$*, aber nicht stark konvex.*

(*vi*) $f(x) = \exp(x)$ *ist strikt konvex auf* $\mathbb{R}$*, aber nicht stark konvex.*

(*vii*) $f(x) = \exp(x)$ *ist stark konvex auf jedem Intervall* $[c, \infty)$ *mit* $c \in \mathbb{R}$.

(*viii*) $f(x) = \ln(x)$ *ist strikt konkav auf* $(0, \infty)$*, aber nicht stark konkav.*

Quizfrage: Gibt es außer affin-linearen Funktionen noch weitere Funktionen auf $\mathbb{R}^n$, die gleichzeitig konvex und konkav sind?

Quizfrage: Welche Konvexitätseigenschaften haben $f(x) = \|x - z\|_1$ und $f(x) = \|x - z\|_\infty$ im $\mathbb{R}^n$?

In der konvexen Optimierung ist es hilfreich, auch Funktionen zuzulassen, deren Funktionswerte in $\mathbb{R} \cup \{\infty\}$ liegen. Man spricht dann von **erweitert reellwertigen Funktionen**. Für den Umgang mit ∞ gelten in $\mathbb{R} \cup \{\infty\}$ folgende Regeln:

(*i*) $a + \infty = \infty + a = \infty$ für alle $a \in \mathbb{R}$ sowie für $a = \infty$,

(*ii*) $a\,\infty = \infty\, a = \infty$ für alle $a > 0$ sowie für $a = \infty$,

(*iii*) $0\,\infty = \infty\, 0 = 0$,

(*iv*) $a < \infty$ für alle $a \in \mathbb{R}$.

(*v*) $\infty \leq \infty$.

Die Kommutativität und Assoziativität der Addition und der Multiplikation sowie das Distributivgesetz gelten auch in $\mathbb{R} \cup \{\infty\}$ weiter, sofern darin jeweils alle Terme definiert sind. Beispielsweise gilt $(2+1)\infty = 2\,\infty + 1\,\infty = \infty + \infty = \infty$. Der Ausdruck $(-2+1)\infty$ ist in $\mathbb{R} \cup \{\infty\}$ dagegegen nicht erklärt.

Der Mehrwert erweitert reellwertiger Funktionen liegt in folgenden Überlegungen begründet:

(1) Wir können jede Funktion $f\colon M \to \mathbb{R}$, die auf einer Teilmenge $M \subseteq \mathbb{R}^n$ definiert ist, auf ganz $\mathbb{R}^n$ fortsetzen, in dem wir $\overline{f}\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ definieren als

$$\overline{f}(x) := \begin{cases} f(x) & \text{falls } x \in M, \\ \infty & \text{falls } x \notin M. \end{cases} \tag{13.5}$$

(2) Wir können die zulässige Menge einer Optimierungsaufgabe einfach dadurch in die Aufgabe einbauen, dass wir den Wert der Zielfunktion außerhalb der zulässigen Menge auf ∞ setzen. Dies gelingt einfach durch Addition einer Indikatorfunktion.

Definition 13.10 (Indikatorfunktion). *Es sei $M \subseteq \mathbb{R}^n$ irgendeine Menge. Die Funktion $I_M\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, definiert durch*

$$I_M(x) = \begin{cases} 0, & \text{falls } x \in M, \\ \infty, & \text{falls } x \notin M, \end{cases} \tag{13.6}$$

*heißt die **Indikatorfunktion** (englisch: indicator function) von M.*

Definition 13.11 (Eigentlicher Definitionsbereich). *Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ eine Funktion. Dann heißt*

$$\operatorname{dom} f := \{x \in \mathbb{R}^n \mid f(x) < \infty\} \tag{13.7}$$

*der **eigentliche Definitionsbereich** oder (englisch: (effective) domain) von f.*

Wir erweitern jetzt die Definition 13.8 auf erweitert reellwertige Funktionen.

Definition 13.12 (Erweitert reellwertige konvexe Funktion).
Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ heißt

(i) ***konvex*** *(englisch: convex) auf $\mathbb{R}^n$, falls*

$$f(\alpha x + (1-\alpha)\, y) \le \alpha f(x) + (1-\alpha) f(y) \tag{13.8}$$

gilt für alle $x, y \in \mathbb{R}^n$ und $\alpha \in [0,1]$.

(ii) ***strikt konvex*** *auf $\mathbb{R}^n$, falls*

$$f(\alpha x + (1-\alpha)\, y) < \alpha f(x) + (1-\alpha) f(y) \tag{13.9}$$

gilt für alle $x, y \in \operatorname{dom} f$ mit $x \neq y$ und $\alpha \in (0,1)$.

(iii) ***μ-stark konvex*** *oder* ***stark konvex*** *mit Parameter $\mu > 0$ auf C, falls*

$$f(\alpha x + (1-\alpha)\, y) + \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|^2 \le \alpha f(x) + (1-\alpha) f(y) \tag{13.10}$$

gilt für alle $x, y \in \mathbb{R}^n$ und $\alpha \in [0,1]$.

Satz 13.13 (Konvexität reellwertiger Funktionen und erweitert reellwertiger Funktionen)**.**

(*i*) *Es sei $C \subseteq \mathbb{R}^n$ konvex und $f\colon C \to \mathbb{R}$ konvex bzw. strikt konvex bzw. stark konvex. Dann ist die Fortsetzung $\overline{f}\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ auf $\mathbb{R}^n$ konvex bzw. strikt konvex bzw. stark konvex.*

(*ii*) *Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex bzw. strikt konvex bzw. stark konvex. Dann ist $\operatorname{dom} f \subseteq \mathbb{R}^n$ konvex, und die Einschränkung $f|_{\operatorname{dom} f}\colon \operatorname{dom} f \to \mathbb{R}$ ist auf $\operatorname{dom} f$ konvex bzw. strikt konvex bzw. stark konvex.*

Beachte: Dieses Resultat zeigt, dass wir im Folgenden immer davon ausgehen können, dass eine konvexe Funktion auf ganz $\mathbb{R}^n$ definiert ist. Wir werden daher auch nicht mehr zwischen f und $\overline{f}$ unterscheiden.

Beweis. Wir führen den Beweis hier nur für die Aussage zur „gewöhnlichen“ Konvexität. Die Fälle „strikte Konvexität“ und „starke Konvexität“ können als eigene Übung ergänzt werden.

Aussage (*i*): Es sei $C \subseteq \mathbb{R}^n$ konvex und $f\colon C \to \mathbb{R}$ konvex. Wir müssen zeigen:

$$\overline{f}(\alpha x + (1-\alpha)\, y) \leq \alpha \overline{f}(x) + (1-\alpha)\overline{f}(y) \quad \text{für alle } x, y \in \mathbb{R}^n \text{ und } \alpha \in [0,1]. \tag{13.11}$$

Falls $x, y \in C$ liegen, dann auch $\alpha x + (1-\alpha)\, y$ für alle $\alpha \in [0,1]$. In diesem Fall lautet die zu zeigende Ungleichung (13.11) einfach

$$f(\alpha x + (1-\alpha)\, y) \leq \alpha f(x) + (1-\alpha) f(y),$$

die wegen der vorausgesetzten Konvexität von f auf C erfüllt ist. Ist andererseits etwa $x \notin C$, dann lautet die zu zeigende Ungleichung (13.11) im Falle von $\alpha \in (0,1]$

$$\overline{f}(\alpha x + (1-\alpha)\, y) \leq \infty + (1-\alpha)\overline{f}(y) = \infty,$$

ist also erfüllt. Im Fall von $\alpha = 0$ lautet (13.11) dagegen $\overline{f}(y) \leq \overline{f}(y)$, was natürlich ebenfalls gilt. Der Beweis für den Fall $y \notin C$ ist analog.

Aussage (*ii*): Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex, es gilt also

$$f(\alpha x + (1-\alpha)\, y) \leq \alpha f(x) + (1-\alpha) f(y) \quad \text{für alle } x, y \in \mathbb{R}^n \text{ und } \alpha \in [0,1] \tag{13.12}$$

Wir zeigen zunächst, dass $\operatorname{dom} f \subseteq \mathbb{R}^n$ konvex ist. Es seien dazu $x, y \in \operatorname{dom} f$ und $\alpha \in [0,1]$. Dann folgt aus (13.12)

$$f(\alpha x + (1-\alpha)\, y) \leq \underbrace{\alpha f(x)}_{<\infty} + \underbrace{(1-\alpha) f(y)}_{<\infty} < \infty,$$

sodass die linke Seite endlich ist. Also gehört $\alpha x + (1-\alpha)\, y$ zu $\operatorname{dom} f$. Dieselbe Ungleichung zeigt auch bereits, dass $f|_{\operatorname{dom} f}\colon \operatorname{dom} f \to \mathbb{R}$ wie behauptet konvex ist. □

Folgerung 13.14 (Konvexität der Indikatorfunktion)**.** *Die Indikatorfunktion $I_M\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ einer Menge $M \subseteq \mathbb{R}^n$ ist genau dann konvex, wenn M konvex ist.*

Es gilt folgende wichtige Charakterisierung konvexer Funktionen:

Satz 13.15 (Epigraph-Charakterisierung konvexer Funktionen)**.**
Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ ist genau dann konvex, wenn ihr ***Epigraph***

$$\operatorname{epi} f := \{(x, \gamma) \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \geq f(x)\} \tag{13.13}$$

eine konvexe Menge ist.

Beweis. Es sei zunächst f konvex, und es seien (x, γ) und (y, δ) Punkte in $\operatorname{epi} f$. Für die Konvexkombination mit $\alpha \in [0,1]$ gilt:

$$f(\alpha x + (1-\alpha)\, y) \leq \alpha f(x) + (1-\alpha) f(y) \leq \alpha \gamma + (1-\alpha)\, \delta.$$

Das bedeutet aber $\big(\alpha x + (1-\alpha)\, y, \alpha \gamma + (1-\alpha)\, \delta\big) = \alpha(x, \gamma) + (1-\alpha)(y, \delta) \in \operatorname{epi} f$. Also ist $\operatorname{epi} f$ konvex.

Umgekehrt sei nun $\operatorname{epi} f$ konvex und $x, y \in \mathbb{R}^n$ sowie $\alpha \in [0,1]$. Wir müssen

$$f(\alpha x + (1-\alpha)\, y) \leq \alpha f(x) + (1-\alpha) f(y)$$

nachweisen. Im Fall $f(x) = \infty$ oder $f(y) = \infty$ ist diese Aussage klar. Es seien also $x, y \in \operatorname{dom} f$. Dann gehören $(x, f(x))$ und $(y, f(y))$ zu $\operatorname{epi} f$. Die Konvexität von $\operatorname{epi} f$ zeigt, dass auch $\alpha\,(x, f(x)) + (1-\alpha)\,(y, f(y))$ zu $\operatorname{epi} f$ gehört, also gilt

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha)\, y),$$

was zu zeigen war. □

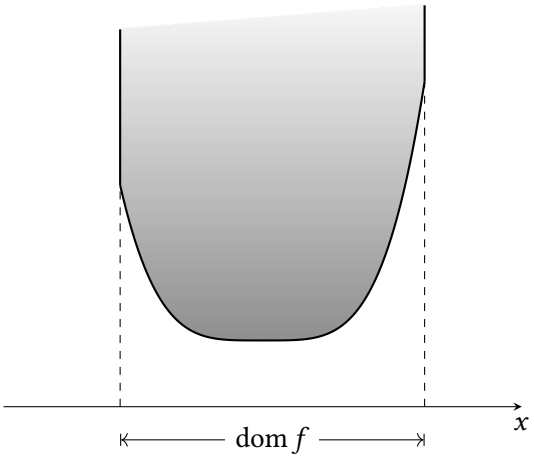


Abbildung 13.6: Epigraph einer Funktion $f\colon \mathbb{R} \to \mathbb{R} \cup \{\infty\}$.

Satz 13.16 (Operationen auf konvexen Funktionen)**.**

(*i*) *Sind* $f_i \colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *konvex auf* $\mathbb{R}^n$ *und* $\beta_i \geq 0$ *für* $i = 1, \ldots, m$*, dann ist die durch*

$$f(x) \coloneqq \sum_{i=1}^{m} \beta_i f_i(x)$$

definierte Funktion $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *konvex auf* $\mathbb{R}^n$*. Ist ein* f_i *strikt bzw. stark konvex und das zugehörige* $\beta_i > 0$*, so ist* f *strikt bzw. stark konvex auf* $\mathbb{R}^n$*.*

(*ii*) *Sind die Funktionen* $f_i \colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *konvex auf* $\mathbb{R}^n$ *für alle* i *aus irgendeiner Indexmenge* I*, dann ist die durch das punktweise Supremum*

$$f(x) \coloneqq \sup\{f_i(x) \mid i \in I\}$$

definierte Funktion $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *konvex auf* $\mathbb{R}^n$*.*

(*iii*) *Ist* $g \colon \mathbb{R}^n \to \mathbb{R}^m$ *affin-linear und* $f \colon \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ *konvex, so ist* $(f \circ g)$ *konvex auf* $\mathbb{R}^n$*.*

(*iv*) *Ist* $g \colon \mathbb{R}^n \to \mathbb{R}\cancel{\cup\{\infty\}}$ *konvex und ist* $f \colon \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ *konvex und monoton wachsend, so ist* $(f \circ g) \colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *konvex.*

Beachte: Aus Aussage (*iv*) folgt insbesondere, dass die Funktion g^2 konvex auf $\mathbb{R}^n$ ist, wenn $g \colon \mathbb{R}^n \to \mathbb{R}\cancel{\cup\{\infty\}}$ konvex ist und $g(x) \geq 0$ gilt für alle $x \in \mathbb{R}^n$. (**Quizfrage:** Genaue Begründung?)

Beweis. Der Beweis findet sich als Übungsaufgabe 3 auf Übungsblatt 9. □

Im Folgenden wollen wir die Konvexität diffbarer Funktionen mit Hilfe der ersten und zweiten Ableitung charakterisieren. Natürlich impliziert die Diffbarkeit, dass die Funktionswerte endlich sind, daher gelten diese Charakterisierungen in Satz 13.17 und Satz 13.18 nur für reellwertige konvexe Funktionen.

Satz 13.17 (Charakterisierung konvexer Funktionen mittels erster Ableitung)**.**
Es sei $C \subseteq \mathbb{R}^n$ *offen und konvex sowie* $f \colon C \to \mathbb{R}$ *diffbar. Dann gelten:*

(*a*) *Es sind äquivalent:*

(*i*) f *ist konvex auf* C*.*

(*ii*) *Für alle* $x, y \in C$ *gilt:*

$$f(x) - f(y) \geq f'(y)(x - y). \tag{13.14}$$

(*iii*) *Für alle* $x, y \in C$ *gilt:*

$$\big(f'(x) - f'(y)\big)(x - y) \geq 0. \tag{13.15}$$

Man sagt zu (13.15)*, die Ableitung* f' *sei auf* C *ein* ***monotoner Operator****.*

(b) Es sind äquivalent:

(*i*) *f ist strikt konvex auf C.*

(*ii*) *Für alle $x, y \in C$ mit $x \neq y$ gilt:*

$$f(x) - f(y) > f'(y)(x - y). \tag{13.16}$$

(*iii*) *Für alle $x, y \in C$ mit $x \neq y$ gilt:*

$$\big(f'(x) - f'(y)\big)(x - y) > 0. \tag{13.17}$$

Man sagt zu (13.17), die Ableitung f' sei auf C ein ***strikt monotoner Operator****.*

(c) Es sind äquivalent:

(*i*) *f ist stark konvex auf C.*

(*ii*) *Es existiert $\mu > 0$, sodass für alle $x, y \in C$ gilt:*

$$f(x) - f(y) \geq f'(y)(x - y) + \frac{\mu}{2} \|x - y\|^2. \tag{13.18}$$

(*iii*) *Es existiert $\mu > 0$, sodass für alle $x, y \in C$ gilt:*

$$\big(f'(x) - f'(y)\big)(x - y) \geq \mu \|x - y\|^2. \tag{13.19}$$

Man sagt zu (13.19), die Ableitung f' sei auf C ein ***stark monotoner Operator****.*

Beachte: Nach Aussage (a) ist eine diffbare Funktion $f : C \to \mathbb{R}$ genau dann konvex, wenn der Graph oberhalb aller seiner Tangentialebenen

$$T(x; y) := f(y) + f'(y)(x - y) \quad \text{(Tangentialebene an } f \text{ im Punkt } y \in C)$$

verläuft, siehe Abbildung 13.7. Anders ausgedrückt: Eine diffbare Funktion ist genau dann konvex, wenn alle Taylormodelle erster Ordnung (Tangenten) die Funktion unterschätzen.

Beweis. Wir zeigen nur die Aussage (c) über die Charakterisierung der starken Konvexität. Die Aussagen (a) und (b) lassen sich analog beweisen.

(*i*) $\Rightarrow$ (*ii*): Es sei f stark konvex auf C und $x, y \in C$, $\alpha \in (0, 1)$. Dann gilt

$$\begin{aligned}
& f(y + \alpha\,(x - y)) = f(\alpha\, x + (1 - \alpha)\, y) \leq \alpha f(x) + (1 - \alpha) f(y) - \frac{\mu}{2}\, \alpha\, (1 - \alpha) \|x - y\|^2 \\
\Rightarrow \quad & \frac{f(y + \alpha\,(x - y)) - f(y)}{\alpha} \leq f(x) - f(y) - \frac{\mu}{2}\, (1 - \alpha)\, \|x - y\|^2 \\
\Rightarrow \quad & f'(y)(x - y) \leq f(x) - f(y) - \frac{\mu}{2} \|x - y\|^2 \quad \text{(Grenzübergang } \alpha \searrow 0\text{)},
\end{aligned}$$

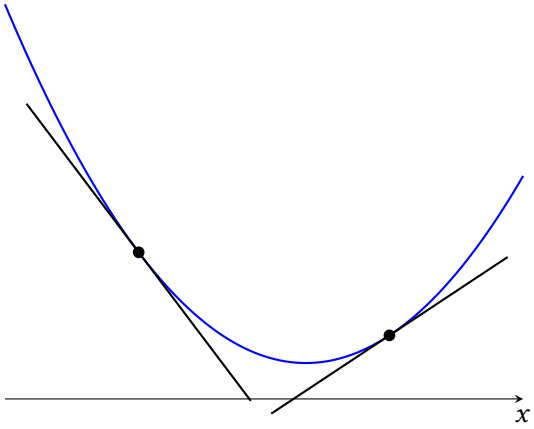


Abbildung 13.7: Charakterisierung der Konvexität diffbarer Funktionen über ihre Tangenten.

d. h., es gilt (13.18).

(ii) $\Rightarrow$ *(i)*: Es gelte (13.18), und es seien $x, y \in C$ und $\alpha \in (0,1)$. Setze $z := \alpha x + (1-\alpha) y$. Eine zweimalige Anwendung von (13.18) ergibt

$$f(x) - f(z) \geq f'(z)(x-z) + \frac{\mu}{2}\|x-z\|^2,$$
$$f(y) - f(z) \geq f'(z)(y-z) + \frac{\mu}{2}\|y-z\|^2.$$

Wir multiplizieren die erste Ungleichung mit α, die zweite mit $(1-\alpha)$ und addieren:

$$\alpha f(x) + (1-\alpha) f(y) - f(z) \geq \underbrace{\alpha f'(z)x + (1-\alpha) f'(z) y - f'(z) z}_{=0 \text{ nach Definition von } z} + \underbrace{\frac{\mu}{2}\alpha\|x-z\|^2 + \frac{\mu}{2}(1-\alpha)\|y-z\|^2}_{=\frac{\mu}{2}\alpha(1-\alpha)\|x-y\|^2 \text{ (nachrechnen)}}.$$

Durch Einsetzen von $z = \alpha x + (1-\alpha) y$ folgt schließlich

$$\alpha f(x) + (1-\alpha) f(y) - f(\alpha x + (1-\alpha) y) \geq \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|^2,$$

d. h., f ist stark konvex.

(ii) $\Rightarrow$ *(iii)*: Es seien $x, y \in C$. Eine zweimalige Anwendung von (13.18) ergibt

$$f(x) - f(y) \geq f'(y)(x-y) + \frac{\mu}{2}\|x-y\|^2,$$
$$f(y) - f(x) \geq f'(x)(y-x) + \frac{\mu}{2}\|x-y\|^2,$$

und aus der Addition der Ungleichungen folgt (13.19).

(iii) $\Rightarrow$ *(ii)*: Es seien $x, y \in C$. Wir betrachten die Funktion $t \mapsto D(t) := f(y + t(x-y))$ und deren Ableitung $t \mapsto d(t) := f'(y + t(x-y))^\mathsf{T}(x-y)$ auf $[0,1]$. Wir zeigen zunächst, dass d auf $[0,1]$ stark

monoton ist. Es seien dazu $s, t \in [0,1]$ beliebig. Dann ist

$$\begin{aligned}(d(t) - d(s))(t-s) &= \big[f'(y + t\,(x-y))^\mathsf{T}(x-y) - f'(y + s\,(x-y))^\mathsf{T}(x-y)\big](t-s)\\ &= \big[f'(y + t\,(x-y)) - f'(y + s\,(x-y))\big]^\mathsf{T}(t-s)(x-y)\\ &\geq \mu\,\|(t-s)(x-y)\|^2 \quad \text{wegen (13.19)}\\ &= \mu\,\|x-y\|^2|t-s|^2.\end{aligned}$$

Da monotone Funktionen Riemann-integrierbar sind (Heuser, 2003, Satz 83.3), ist der Hauptsatz der Differential- und Integralrechnung (siehe etwa Heuser, 2003, Satz 79.1) anwendbar, und es folgt

$$D(1) - D(0) = \int_0^1 d(t)\,\mathrm{d}t.$$

Daher gilt weiter

$$\begin{aligned}D(1) - D(0) - d(0) &= \int_0^1 [d(t) - d(0)]\,\mathrm{d}t\\ &= \int_0^1 \frac{1}{t}\,[d(t) - d(0)]\,(t-0)\,\mathrm{d}t\\ &\geq \mu\,\|x-y\|^2 \int_0^1 t\,\mathrm{d}t\\ &= \frac{\mu}{2}\|x-y\|^2.\end{aligned}$$

Das Einsetzen der Definitionen von D und d ergibt schließlich

$$f(x) - f(y) - f'(y)(x-y) \geq \frac{\mu}{2}\,\|x-y\|^2,$$

also (13.18). □

Satz 13.18 (Charakterisierung konvexer Funktionen mittels zweiter Ableitungen)**.**
Es sei $C \subseteq \mathbb{R}^n$ offen und konvex sowie $f : C \to \mathbb{R}$ zweimal stetig partiell diffbar (eine C^2-Funktion). Dann gelten:

(a) Es sind äquivalent:

(i) f ist konvex auf C.

(ii) $f''(x)$ ist positiv semidefinit (hat nur nicht-negative Eigenwerte) für alle $x \in C$.

(b) Ist $f''(x)$ positiv definit für alle $x \in C$, so ist f strikt konvex auf C.

(c) Es sind äquivalent:

(i) f ist stark konvex auf C mit Konstante $\mu > 0$.

(ii) Der kleinste Eigenwert von $f''(x)$ erfüllt $\lambda_{\min}(f''(x)) \geq \mu > 0$ für alle $x \in C$.

Beachte: Die Umkehrung von Aussage (b) gilt nicht, wie das Beispiel $f(x) = x^4$ zeigt. Diese Funktion ist strikt konvex auf $\mathbb{R}$, aber $f''(0) = 0$ ist nur semidefinit.

Beweis. Wir beweisen zuerst Aussage (c).

$(i) \Rightarrow (ii)$: Es sei f stark konvex auf C mit Konstante $\mu > 0$. Nach Satz 13.17 (c) ist f' dann stark monoton auf C, erfüllt also (13.19). Für beliebiges $x \in C$ und $d \in \mathbb{R}^n$ ist daher

$$\begin{aligned} d^\mathsf{T} f''(x)\, d &= \left[\lim_{t\to 0} \frac{f'(x+td) - f'(x)}{t}\right]^\mathsf{T} d \\ &= \left[\lim_{t\to 0} \frac{f'(x+td) - f'(x)}{t^2}\right]^\mathsf{T} (td) \\ &\geq \lim_{t\to 0} \frac{1}{t^2}\, \mu \, \|td\|^2 \quad \text{wegen (13.19)} \\ &= \mu \, \|d\|^2. \end{aligned}$$

Daraus folgt, dass $\lambda_{\min}(f''(x)) \geq \mu$ ist.

$(ii) \Rightarrow (i)$: Es gelte nun umgekehrt $\lambda_{\min}(f''(x)) \geq \mu > 0$ für alle $x \in C$. Es seien $x, y \in C$ beliebig. Aus dem Hauptsatz der Differential- und Integralrechnung, angewendet auf die C^1-Funktion $D(t) = f'(y + t\,(x-y))\,(x-y)$ und ihre Ableitung $d(t) = (x-y)^\mathsf{T} f''(y + t\,(x-y))\,(x-y)$, folgt

$$D(1) - D(0) = \int_0^1 d(t)\,\mathrm{d}t,$$

also

$$\begin{aligned} [f'(x) - f'(y)]\,(x-y) &= \int_0^1 (x-y)^\mathsf{T} f''(y + t\,(x-y))\,(x-y)\,\mathrm{d}t \\ &\geq \mu \int_0^1 \|x-y\|^2\,\mathrm{d}t \\ &= \mu\,\|x-y\|^2. \end{aligned}$$

Das heißt, f' ist stark monoton, und nach Satz 13.17 (c) ist f stark konvex.

Der Beweis von Aussage (a) erfolgt genau auf die gleiche Weise mit $\mu = 0$. Zum Beweis von Aussage (b) nehmen wir an, dass $f''(x)$ für alle $x \in C$ positiv definit ist. Es seien $x, y \in C$, $x \neq y$, und wir setzen D und d wie oben. Dann ist $d(t) > 0$ für alle $t \in [0,1]$. Wieder mit dem Hauptsatz der Differential- und Integralrechnung folgt

$$[f'(x) - f'(y)]\,(x-y) = \int_0^1 (x-y)^\mathsf{T} f''(y + t\,(x-y))\,(x-y)\,\mathrm{d}t > 0.$$

Das heißt, f' ist strikt monoton, und aus Satz 13.17 (b) folgt die strikte Konvexität von f auf C. □

Ende der Woche 9

§ 14 Konvexe Optimierungsaufgaben

Wir betrachten die **konvexe Optimierungsaufgabe**

$$\text{Minimiere} \quad f(x) \quad \text{über } x \in \mathbb{R}^n \tag{14.1}$$

mit konvexer Zielfunktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Diese Problemklasse enthält insbesondere Aufgaben der Form

$$\text{Minimiere} \quad g(x) \quad \text{über } x \in C, \tag{14.2}$$

wobei die zulässige Menge $C \subseteq \mathbb{R}^n$ konvex und $g\colon C \to \mathbb{R}$ eine konvexe reellwertige Funktion ist. Setzen wir dann $f := g + \delta_C$ mit der Indikatorfunktion δ_C von C, setzen also g durch den Wert ∞ außerhalb von C fort, so ergibt sich eine Aufgabe der Form (14.1). Formal handelt es sich bei (14.1) um eine unrestringierte Optimierungsaufgabe, die jedoch implizit die Nebenbedingung $x \in C = \operatorname{dom} f$ enthält.

Die Grundbegriffe aus Definition 1.1 gelten auch für Aufgaben mit erweitert reellwertigen Zielfunktionen weiter. Der Optimalwert f^* der Aufgabe (14.1) ist wieder entweder $f^* = -\infty$, oder f^* ist endlich, oder es gilt $f^* = \infty$. Letzteres ist genau dann der Fall, wenn $f \equiv \infty$ ist, also $\operatorname{dom} f = \emptyset$ gilt. Diesen Fall werden wir aber im Folgenden oft ausschließen. Wir sagen dazu: Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ heißt **eigentlich** (englisch: *proper function*), wenn f nicht identisch ∞ ist.

Beispiele für (14.1) sind sämtliche linearen Optimierungsaufgaben aus Kapitel 2, nicht notwendigerweise in Normalform. Weitere Beispiele folgen.

Die fundamentale Bedeutung der Konvexität in der Optimierung erläutert der folgende Satz.

Satz 14.1 (Hauptsatz der konvexen Optimierung)**.**

(*i*) *Jeder lokale Minimierer von* (14.1) *ist bereits ein globaler Minimierer.*[1]

(*ii*) *Die Lösungsmenge von* (14.1) *ist konvex (evtl. leer).*

(*iii*) *Ist* f strikt *konvex auf* $\mathbb{R}^n$ *und eigentlich, so besitzt* (14.1) *höchstens eine Lösung.*

Beweis. Aussage (*i*): Es sei x^* lokaler Minimierer, d. h., es existiert eine Umgebung $U(x^*)$ mit $f(x^*) \le f(x)$ für alle $x \in U(x^*)$, vgl. Definition 1.1. Angenommen, es gäbe ein $\widehat{x} \in \mathbb{R}^n$ mit $f(\widehat{x}) < f(x^*)$. Für $\alpha \in (0, 1]$ gilt dann

$$f(\alpha\,\widehat{x} + (1-\alpha)\,x^*) \le \alpha f(\widehat{x}) + (1-\alpha) f(x^*) < f(x^*).$$

Für α hinreichend klein liegt aber $\alpha\,\widehat{x} + (1-\alpha)\,x^* \in U(x^*)$, im Widerspruch zur lokalen Optimalität von x^*. Also kann ein solches $\widehat{x}$ nicht existieren, d. h., es ist

$$f(x^*) \le f(x) \quad \text{für alle } x \in \mathbb{R}^n.$$

[1] Wir brauchen also nicht zwischen lokalen und globalen Minimierern zu unterscheiden!

Aussage *(ii)*: Es seien x^* und x^{**} Lösungen von (14.1), also $f(x^*) = f(x^{**}) = f^*$ (Optimalwert). Für $\alpha \in [0,1]$ gilt

$$f(\alpha\, x^* + (1-\alpha)\, x^{**}) \le \alpha f(x^*) + (1-\alpha) f(x^{**}) = f^*.$$

Das zeigt, dass auch $\alpha\, x^* + (1-\alpha)\, x^{**}$ ein Minimierer von (14.1) ist.

Aussage *(iii)*: Es seien x^* und x^{**} zwei verschiedene Minimierer von (14.1). Da f eigentlich ist, gilt $x^*, x^{**} \in \operatorname{dom} f$ und $f(x^*) = f(x^{**}) = f^* \in \mathbb{R}$. Die Konvexkombination $(x^* + x^{**})/2$ liegt ebenfalls in $\operatorname{dom} f$, daher folgt aus der strikten Konvexität (13.9)

$$f\Big(\frac{x^* + x^{**}}{2}\Big) < \frac{1}{2} f(x^*) + \frac{1}{2} f(x^{**}) = f^*,$$

im Widerspruch zur Optimalität von x^* und x^{**}. □

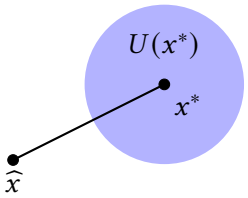


Abbildung 14.1: Illustration des Beweises von Satz 14.1 *(i)* (lokale Minimierer sind globale Minimierer).

§ 15 Trennungssätze für konvexe Mengen

§ 15.1 Die Aufgabe der orthogonalen Projektion

Literatur: Geiger, Kanzow, 2002, Kapitel 2.1.3

Das folgende Beispiel führt eine der wichtigsten konvexen Optimierungsaufgaben ein.

Beispiel 15.1 (Projekionsaufgabe). *Zu einer nichtleeren, abgeschlossenen, konvexen Menge $C \subseteq \mathbb{R}^n$ und einem gegebenen Punkt $p \in \mathbb{R}^n$ suchen wir denjenigen Punkt $x \in C$, der p am nächsten liegt, also die* ***orthogonale Projektion*** *von p auf C (bzgl. des Euklidischen Skalarprodukts), kurz:* $\operatorname{proj}_C(p)$. *Als Optimierungsaufgabe können wir dies in der Form*

$$\text{Minimiere} \quad f(x) := \|x - p\| + \delta_C(x) \quad \text{über } x \in \mathbb{R}^n \tag{15.1}$$

oder auch als

$$\text{Minimiere} \quad g(x) := \frac{1}{2}\|x - p\|^2 + \delta_C(x) \quad \text{über } x \in \mathbb{R}^n \tag{15.2}$$

schreiben.

Lemma 15.2 (Projektionsaufgabe: Existenz und Eindeutigkeit).
Es sei $C \subseteq \mathbb{R}^n$ nichtleer, abgeschlossen und konvex. Für jedes $p \in \mathbb{R}^n$ besitzen (15.1) und (15.2) dieselbe eindeutige Lösung x^*. Diese heißt die ***orthogonale Projektion*** *von p auf C, kurz:* $\operatorname{proj}_C(p)$.

Beweis. Es sei $p \in \mathbb{R}^n$. Wir betrachten zunächst (15.2). Um die Existenz einer Lösung zu zeigen, führen wie eine Hilfsaufgabe ein. Dazu wählen wir ein $w \in C$ beliebig und definieren B als die kompakte Kugel

$$B := \overline{B_r(p)} \quad \text{mit } r := \|p - w\|.$$

Die Hilfsaufgabe lautet

$$\begin{aligned} \text{Minimiere} \quad & \frac{1}{2}\|x - p\|^2 \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & x \in C \cap B. \end{aligned} \tag{15.3}$$

Ein Punkt $x^* \in C$ ist genau dann ein globaler Minimierer von (15.2), wenn er ein globaler Minimierer von (15.3) ist (**Quizfrage:** Warum?). Die zulässige Menge von (15.3) ist nicht leer (denn sie enthält den Punkt w) und als Schnitt der abgeschlossenen Menge C mit der kompakten Menge B wieder kompakt. Nach dem Satz von Weierstraß bzw. Satz 1.4 besitzt (15.3) und damit (15.2) einen globalen Minimierer x^*. Die Eindeutigkeit des globalen Minimierers von (15.2) folgt aus der strikten Konvexität von g mit Satz 14.1 (*iii*).

Mit Hilfe von Hausaufgabe 3 der Übung 01 und der strikten Monotonie der Wurzelfunktion auf $\mathbb{R}_{\geq 0}$ kann gezeigt werden, dass jeder lokale Minimierer von (15.2) auch ein lokaler Minimierer von (15.1) ist und umgekehrt. Da beide Aufgaben konvex sind, sind lokale Minimierer bereits globale Minimierer. Damit besitzen (15.1) und (15.2) denselben eindeutigen globalen Minimierer. □

Satz 15.3 (Projektionssatz: notwendige und hinreichende Bedingungen für (15.2))**.**
Es sei $C \subseteq \mathbb{R}^n$ nichtleer, abgeschlossen und konvex und $p \in \mathbb{R}^n$. Es gilt $x^* = \operatorname{proj}_C(p)$ genau dann, wenn $x^* \in C$ ist und gilt:

$$(x^* - p)^\mathsf{T}(x - x^*) \geq 0 \quad \textit{für alle } x \in C. \tag{15.4}$$

Beachte: (15.4) ist eine **Variationsungleichung**. Sie besagt, dass der Winkel zwischen $x^* - p$ und $x - x^*$ 90° nicht übersteigen darf. Anders ausgedrückt: C ist enthalten im Halbraum $H^+(a, \beta) = \{x \in \mathbb{R}^n \mid a^\mathsf{T}x \geq \beta\}$ mit Normalenvektor $a = x^* - p$ und $\beta = a^\mathsf{T}x^*$, vgl. Abbildung 15.1.

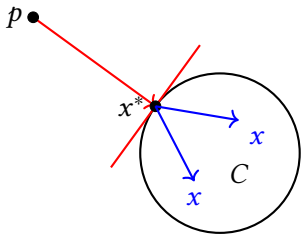


Abbildung 15.1: Der Winkel zwischen $x^* - p$ und $x - x^*$ darf 90° nicht übersteigen.

Beweis. Definiere wie in (15.1) und (15.2) die Funktionen $f(x) := \|x - y\| + \delta_C(x)$ und $g(x) := \frac{1}{2}\|x - y\|^2 + \delta_C(x)$.

„⇒“: Es sei $x^* = \operatorname{proj}_C(p)$, also insbesondere $x^* \in C$. Dann gilt $x^* + \alpha\,(x - x^*) \in C$ für alle $x \in C$ und

$\alpha \in (0,1)$. Aus der Optimalität von x^* folgt

$$\frac{1}{2}\|x^* - p\|^2 = g(x^*) \leq g(x^* + \alpha\,(x - x^*)) = \frac{1}{2}\|(x^* - p) + \alpha\,(x - x^*)\|^2$$
$$\Rightarrow \quad 0 \leq \alpha\,(x^* - p)^\mathsf{T}(x - x^*) + \frac{\alpha^2}{2}\|x - x^*\|^2.$$

Division durch α und Grenzübergang $\alpha \searrow 0$ liefern die Behauptung (15.4).

„⇐“: Es gelte $x^* \in C$ und (15.4). Daraus folgt

$$\begin{aligned} 0 &\geq (p - x^*)^\mathsf{T}(x - x^*) \quad \text{für alle } x \in C \\ &= (p - x^*)^\mathsf{T}(x - p + p - x^*) \\ &= (p - x^*)^\mathsf{T}(x - p) + \|p - x^*\|^2 \\ &\geq -\|p - x^*\|\|x - p\| + \|p - x^*\|^2 \quad \text{(Cauchy-Schwarz)}. \end{aligned}$$

Daraus folgt weiter $f(x) = \|x - p\| \geq \|p - x^*\| = f(x^*)$ für alle $x \in C$, d. h., $x^* = \mathrm{proj}_C(p)$. □

§ 15.2 Affine Unterräume

Definition 15.4 (Affiner Unterraum). *Eine Menge $A \subseteq \mathbb{R}^n$ heißt ein **affiner Unterraum** (englisch: affine subspace) von $\mathbb{R}^n$, wenn mit $x, y \in A$ und $\alpha \in \mathbb{R}$ auch $\alpha\,x + (1 - \alpha)\,y \in A$ liegt, also die gesamte Verbindungsgerade durch x und y.*

Lemma 15.5 (Struktur affiner Unterräume). *Eine Menge $A \subseteq \mathbb{R}^n$ ist genau dann ein affiner Unterraum von $\mathbb{R}^n$, wenn es einen Unterraum $U \subseteq \mathbb{R}^n$ und einen Vektor $x_0 \in \mathbb{R}^n$ gibt, sodass gilt:*

$$A = U + x_0. \tag{15.5}$$

In diesem Fall gilt (15.5) für jedes $x_0 \in A$, und U ist unabhängig von der Wahl von x_0.

Der zu einem affinen Unterraum A gehörende Unterraum U heißt auch der **Richtungsraum** von A.

Beweis. „⇐“: Es sei A eine Menge von der Form (15.5). Weiter seien $x_1, x_2 \in A$, also $x_1 = u_1 + x_0$ und $x_2 = u_2 + x_0$ mit $u_1, u_2 \in U$. Schließlich sei $\alpha \in \mathbb{R}$. Dann ist auch

$$\alpha\,x_1 + (1 - \alpha)\,x_2 = \alpha\,(u_1 + x_0) + (1 - \alpha)\,(u_2 + x_0) = \underbrace{\alpha\,u_1 + (1 - \alpha)\,u_2}_{\in U} + x_0$$

von der Form (15.5).

„⇒“: Es sei A ein affiner Unterraum und $x_0 \in A$ beliebig, aber fest. Definiere $U := \{x - x_0 \mid x \in A\}$. Nach Konstruktion gilt dann $A = U + x_0$. Wir müssen zeigen: U ist ein Unterraum von $\mathbb{R}^n$. Es seien

dazu $u_1 = x_1 - x_0$ und $u_2 = x_2 - x_0$ Elemente von U mit irgendwelchen $x_1, x_2 \in A$. Dann sind $2x_1 + (1-2)x_0 = 2x_1 - x_0 \in A$ und $2x_2 + (1-2)x_0 = 2x_2 - x_0 \in A$ und daher

$$\begin{aligned} u_1 + u_2 &= x_1 - x_0 + x_2 - x_0 \\ &= \overbrace{\frac{1}{2}\underbrace{(2x_1 - x_0)}_{\in A} + \left(1 - \frac{1}{2}\right)\underbrace{(2x_2 - x_0)}_{\in A}}^{\in A} - x_0 \in A - x_0. \end{aligned} \tag{15.6}$$

Das zeigt $u_1 + u_2 \in U$, also $U + U \subseteq U$. Weiterhin sei $\beta \in \mathbb{R}$, dann gilt

$$\begin{aligned} \beta u_1 &= \beta (x_1 - x_0) \\ &= \underbrace{\beta x_1 + (1-\beta) x_0}_{\in A} - x_0. \end{aligned}$$

Das zeigt $\beta u_1 \in U$, also $\beta U \subseteq U$. Damit ist U ein Unterraum.

Der Beweis der letzten Implikation zeigt auch, dass (15.5) für jedes beliebige $x_0 \in A$ mit einem Unterraum der Form $\{x - x_0 \mid x \in A\}$ gilt. Es bleibt also lediglich zu zeigen, dass die Mengen $\{x - x_0 \mid x \in A\}$ und $\{x - \tilde{x}_0 \mid x \in A\}$ für beliebige $x_0, \tilde{x}_0 \in A$ übereinstimmen. Es seien dafür $x, x_0, \tilde{x}_0 \in A$. Wir schreiben $x - x_0 = x - x_0 + \tilde{x}_0 - \tilde{x}_0$, und weil

$$x - x_0 + \tilde{x}_0 = \underbrace{x - x_0 + \tilde{x}_0 - x_0}_{\in \{x - x_0 \mid x \in A\}} + x_0 \in A,$$

ist der Unterraum U unabhängig vom gewählten Aufpunkt x_0. □

Beachte: Ein affiner Unterraum ist nach (15.5) also ein „verschobener“ Unterraum.

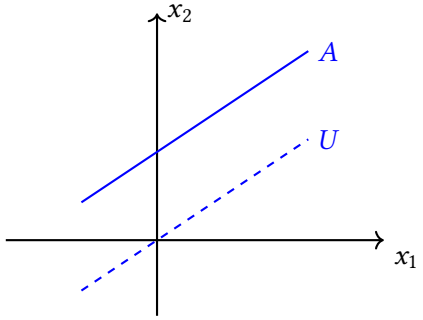


Abbildung 15.2: Ein 1-dimensionaler affiner Unterraum A von $\mathbb{R}^2$. Der zugehörige Richtungsraum U ist gestrichelt gezeichnet.

Wir ordnen einem affinen Unterraum A die **Dimension** $\dim A := \dim U$ zu. Aus technischen Gründen ist es günstig, den Fall zuzulassen, bei dem $A = \emptyset$ ist damit auch $U = \emptyset$ ist. In diesem Fall setzen wir $\dim A := \dim U = -1$.

Definition 15.6 (Affine Unabhängigkeit, Affinkombination)**.**

(*i*) *Eine Menge von Vektoren* $\{x_0, x_1, \ldots, x_k\}$ *in* $\mathbb{R}^n$, $k \in \mathbb{N}_0$, *heißt* ***affin unabhängig*** *(englisch:* affine independent*), wenn die Vektoren* $x_1 - x_0, \ldots, x_k - x_0$ *linear unabhängig sind.*[2]

(*ii*) *Eine maximale Menge affin unabhängiger Vektoren eines affinen Unterraumes* A *von* $\mathbb{R}^n$ *heißt eine* ***affine Basis*** *von* A.

(*iii*) *Ein Vektor* $x \in \mathbb{R}^n$ *heißt eine* ***Affinkombination*** *von* $x_0, \ldots, x_m \in \mathbb{R}^n$, $m \in \mathbb{N}_0$, *wenn gilt:*

$$x = \sum_{i=0}^{m} \alpha_i \, x_i$$

mit Koeffizienten $\alpha_i \in \mathbb{R}$, *die* $\sum_{i=0}^{m} \alpha_i = 1$ *erfüllen. Ist* $M \subseteq \mathbb{R}^n$ *irgendeine (nicht notwendig endliche) Menge, so heißt* x *eine Affinkombination von* M, *wenn* x *eine Affinkombination von endlich vielen Vektoren* $x_0, \ldots, x_m \in M$ *ist.*

Lemma 15.7 (Dimension eines affinen Unterraums)**.**
Es sei $A \subset \mathbb{R}^n$ *ein affiner Unterraum.*

(*i*) *Ist* $\{x_0, x_1, \ldots, x_k\}$ *eine affine Basis von* A, *dann lässt sich jedes Element von* A *auf eindeutige Art und Weise aus* $\{x_0, x_1, \ldots, x_k\}$ *affinkombinieren. Genauer hat jedes* $x \in A$ *die Darstellung*

$$x = \sum_{i=0}^{k} \alpha_i \, x_i \tag{15.7}$$

mit Koeffizienten $\alpha = (\alpha_0, \ldots, \alpha_k)^\mathsf{T}$, *die sich aus der eindeutigen Lösung des linearen Gleichungssystems*

$$\underbrace{\begin{bmatrix} 1 & \cdots & 1 \\ | & & | \\ x_0 & \cdots & x_k \\ | & & | \end{bmatrix}}_{=:B} \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_k \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ | \\ x \\ | \end{pmatrix}}_{=:b} \tag{15.8}$$

ergeben. Die Matrix $B \in \mathbb{R}^{(n+1)\times(k+1)}$ *hat Rang* $k+1$. *Daher ist* $B^\mathsf{T}B$ *regulär, und (15.8) kann äquivalent als*

$$B^\mathsf{T}B\,\alpha = B^\mathsf{T}b \tag{15.9}$$

geschrieben werden.

(*ii*) A *besitzt genau dann eine affine Basis* $\{x_0, x_1, \ldots, x_k\}$ *mit* $k+1$ *Elementen, wenn* $\dim A = k$ *ist.*

Beachte: Im Fall von $A = \emptyset$ ist $k = -1$ und die affine Basis die leere Menge.

Beweis. Der Beweis findet sich als Übungsaufgabe 3 auf Übungsblatt 10. □

[2] Ein einzelner Vektor $x_0 \in \mathbb{R}^n$ ist also immer affin unabhängig.

Lemma 15.8 (Operationen auf affinen Unterräumen)**.**
Es sei $A = U + x_0$ und $B = V + y_0$ zwei affine Unterräume von $\mathbb{R}^n$ wie in (15.5).

(*i*) *Die folgenden Aussagen sind äquivalent:*

(a) $A = B$.

(b) $U = V$ *und* $x_0 - y_0 \in U$.

(c) $U = V$ *und* $0 \in A - B$.

(d) $U = V$ *und* $A \cap B \neq \emptyset$.

Beachte: *Aussage (b) besagt, dass wir jeden beliebigen Aufpunkt $x_0 \in A$ wählen können, um einen affinen Unterraum $A = U + x_0$ zu repräsentieren.*

(*ii*) *Die Menge $\alpha A + \beta B$ für beliebige $\alpha, \beta \in \mathbb{R}$ ist ein affiner Unterraum von $\mathbb{R}^n$.*

(*iii*) *$A \cap B$ ist ein affiner Unterraum von $\mathbb{R}^n$.*

Beweis. Aussage (*i*): Wir zeigen zunächst (a) $\Rightarrow$ (c). Es sei zunächst $A = B$. Nach Aussage (*i*) gilt $U = A - A = A - B$ und andererseits $V = B - B = A - B$, also $U = V$. Die Folgerung $0 \in A - B$ ist klar.

Wir zeigen nun (c) $\Rightarrow$ (a). Es seien dazu $U = V$ und $0 \in A - B$. Aus letzerer Beziehung folgt, dass es $u \in U$ und $v \in V$ gibt, sodass $u + x_0 = v + y_0$ gilt. Es sei nun $x_1 = u_1 + x_0$ irgendein Punkt in A mit irgendeinem $u_1 \in U$. Dann ist $x_1 = u_1 - u + u + x_0 = u_1 - u + v + y_0$. Da $u_1 - u + v \in U = V$ gilt, liegt $x_1 \in B$, also $A \subseteq B$. Analog zeigt man $B \subseteq A$.

Die Äquivalenz (c) $\Leftrightarrow$ (d) ist klar. Wir zeigen nun noch (b) $\Leftrightarrow$ (d). Dazu sei $U = V$. Wir müssen überprüfen, dass die Bedingungen $x_0 - y_0 \in U$ und $A \cap B \neq \emptyset$ äquivalent sind. Es sei zunächst $x_0 - y_0 \in U$, also $x_0 = u + y_0$ für ein $u \in U$. Damit ist $x_0 = 0 + x_0 \in A$ und ebenfalls $x_0 = u + y_0 \in B$, also liegt $x_0 \in A \cap B$. Umgekehrt gebe es ein $x \in A \cap B$, also gilt $x = u + x_0 = v + y_0$ mit irgendwelchen $u, v \in U$. Dann ist $x_0 - y_0 = u - v \in U$.

Aussage (*ii*): Für $\alpha, \beta \in \mathbb{R}$ besteht $\alpha A + \beta B$ aus Elementen der Form $x = \alpha (u + x_0) + \beta (v + y_0)$ mit irgendwelchen $u \in U$ und $v \in V$. Da $\alpha u \in U$ und $\beta v \in V$ gilt und $U + V$ ein Unterraum von $\mathbb{R}^n$ ist, ist gezeigt, dass $\alpha A + \beta B$ ein affiner Unterraum ist.

Aussage (*iii*): Im Fall $A \cap B = \emptyset$ ist nichts zu zeigen. Es sei also $x \in A \cap B$. Wir können also gemäß Aussage (*i*) (b) $A = U + x$ und $B = V + x$ darstellen. Folglich gilt $A \cap B = (U \cap V) + x$, und da $U \cap V$ ein Unterraum von $\mathbb{R}^n$ ist, ist $A \cap B$ ein affiner Unterraum. □

Definition 15.9 (Affine Hülle)**.** *Es sei $M \subseteq \mathbb{R}^n$. Der Durchschnitt aller affinen Unterräume von $\mathbb{R}^n$, die M enthalten, also*

$$\operatorname{aff}(M) = \bigcap \big\{ A \subseteq \mathbb{R}^n \,\big|\, A \text{ ist affiner Unterraum von } \mathbb{R}^n \text{ und } M \subseteq A \big\}, \tag{15.10}$$

*heißt die **affine Hülle** von M. $\operatorname{aff}(M)$ ist also der kleinste affine Unterraum, der M enthält.*

Beachte: Es gilt $M \subseteq \operatorname{aff}(M)$, daher der Name **affine Hülle**.
„Hülle“ affin

Definition 15.10 (Dimension einer Menge)**.** *Es sei $M \subseteq \mathbb{R}^n$. Die **Dimension** von M ist* $\dim M := \dim \operatorname{aff}(M)$.

Quizfrage: Ist diese Definition konsistent mit den bereits bekannten Definitionen der Dimension von Unterräumen von $\mathbb{R}^n$ und von affinen Unterräumen von $\mathbb{R}^n$?

Analog zu Lemma 13.5 und Lemma 13.7 gilt:

Lemma 15.11 (Charakterisierung affiner Unterräume und der affinen Hülle)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge.

(*i*) *Eine Menge $M \subseteq \mathbb{R}^n$ ist genau dann ein affiner Unterraum, wenn sie alle Affinkombinationen ihrer Elemente enthält.*

(*ii*) $\operatorname{aff}(M)$ *ist gleich der Menge aller Affinkombinationen von M.*

(*iii*) *Es gilt* $M \subseteq \operatorname{conv}(M) \subseteq \operatorname{aff}(M)$.

Beachte: A ist affiner Unterraum $\Leftrightarrow A = \operatorname{aff}(A)$.

Beweis. Aussage (*i*): „$\Rightarrow$“: Es sei M ein affiner Unterraum von $\mathbb{R}^n$. Für $m \in \mathbb{N}$ und $x_1, \dots, x_m \in M$ sowie $\alpha_1, \dots, \alpha_m \in \mathbb{R}$ mit $\sum_{i=1}^m \alpha_i = 1$ ist zu zeigen: $x = \sum_{i=1}^m \alpha_i x_i \in M$.

Induktion nach m: Für $m = 1, 2$ ist die Behauptung erfüllt. Es sei bereits gezeigt, dass M alle Affinkombinationen von höchstens m Elementen enthält.

Schluss auf $m+1$ für $m \geq 2$[GM]: Es seien $\alpha_i \in \mathbb{R}$, $\sum_{i=1}^{m+1} \alpha_i = 1$ und $x = \sum_{i=1}^{m+1} \alpha_i x_i$. O. B. d. A. gilt $\alpha_{m+1} \neq 1$. (Alle Koeffizienten α_i können nur gleich 1 sein, wenn $m = 0$ ist. ~~Ansonsten ist $x = x_{m+1}$ und nichts zu zeigen.~~[GM]) Setze $\beta_i := \frac{\alpha_i}{1-\alpha_{m+1}}$ für $i = 1, \dots, m$. Dann ist $\sum_{i=1}^m \beta_i = 1$. Der Vektor $y = \sum_{i=1}^m \beta_i x_i$ gehört zu M, also auch $x = (1 - \alpha_{m+1})\, y + \alpha_{m+1} x_{m+1}$.

„$\Leftarrow$“: Es seien $x_1, x_2 \in M$. Nach Voraussetzung enthält M alle Affinkombinationen $\alpha x_1 + (1-\alpha) x_2$, d. h., M ist affiner Unterraum.

Aussage (*ii*): Es sei A die Menge aller Affinkombinationen von M. Natürlich gilt dann $M \subseteq A$. Im Fall $M = \emptyset$ ist nichts zu zeigen, weil dann auch $A = \emptyset$ ist. Wir gehen also jetzt von $M \neq \emptyset$ aus.

„$\operatorname{aff}(M) \subseteq A$“: Wir zeigen: A ist ein affiner Unterraum. Damit kommt diese Menge im Durchschnitt (15.10) vor, also gilt $\operatorname{aff}(M) \subseteq A$.

Es seien $x, y \in A$, also gibt es Zahlen $m, \ell \in \mathbb{N}$ und $\beta_1, \dots, \beta_m \in \mathbb{R}$ sowie $\gamma_1, \dots, \gamma_\ell \in \mathbb{R}$ mit $\sum_{i=1}^m \beta_i = 1$ und $\sum_{j=1}^\ell \gamma_j = 1$, sodass $x = \sum_{i=1}^m \beta_i x_i$ und $y = \sum_{j=1}^\ell \gamma_j y_j$ gelten mit irgendwelchen $x_1, \dots, x_m \in M$ und $y_1, \dots, y_\ell \in M$. Es sei $\alpha \in \mathbb{R}$. Dann gilt

$$\alpha x + (1-\alpha) y = \alpha \sum_{i=1}^m \beta_i x_i + (1-\alpha) \sum_{j=1}^\ell \gamma_j y_j,$$

d. h., $\alpha x + (1-\alpha) y$ ist Linearkombination der $\{x_i\}_{i=1}^m \cup \{y_j\}_{j=1}^\ell$. Die Koeffizienten ergeben in der Summe 1. Damit ist $\alpha x + (1-\alpha) y \in A$, also A ist affiner Unterraum.

„$\operatorname{aff}(M) \supseteq A$“: Es sei $x \in A$, also eine Affinkombination von M. Wegen $M \subseteq \operatorname{aff}(M)$ ist x auch eine Affinkombination von $\operatorname{aff}(M)$. $\operatorname{aff}(M)$ ist ein affiner Unterraum, stimmt also nach Aussage *(i)* mit der Menge seiner Affinkombinationen überein. Also ist $x \in \operatorname{aff}(M)$.

Aussage *(iii)*: Nach Lemma 13.7 gilt $M \subseteq \operatorname{conv}(M)$, und $\operatorname{conv}(M)$ sind gerade die Konvexkombinationen von M. Da jede Konvexkombination auch eine Affinkombination ist, gilt weiter $\operatorname{conv}(M) \subseteq \operatorname{aff}(M)$. □

Quizfrage: Welche Beziehung besteht zwischen $\operatorname{aff}(M_1 + M_2)$ und $\operatorname{aff}(M_1) + \operatorname{aff}(M_2)$ für beliebige Mengen $M_1, M_2 \subseteq \mathbb{R}^n$?

Satz 15.12. *Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge der Dimension k. Dann existieren $k+1$ affin unabhängige Punkte $x_0, \dots, x_k \in M$, die eine affine Basis von $\operatorname{aff}(M)$ bilden.*

Beweis. Die affine Hülle $\operatorname{aff}(M)$ besteht aus den Affinkombinationen von M (Lemma 15.11). Da die Dimension einer Menge von Affinkombinationen gleich der maximalen Anzahl affin unabhängiger Punkte ist, folgt die Behauptung. □

§ 15.3 Topologische Eigenschaften konvexer Mengen

Wir geben jetzt einen wichtigen Satz der Konvexgeometrie an, der die Arbeit mit konvexen Hüllen erheblich vereinfacht:

Satz 15.13 (Carathéodory). *Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge der Dimension k und $x \in \operatorname{conv}(M)$ eine Konvexkombination von Punkten $x_0, \dots, x_m \in M$ mit $m \geq 0$. Dann ist x bereits eine Konvexkombination von höchstens $k+1$ dieser Punkte.*

Beachte: Dass jedes $x \in \operatorname{conv}(M)$ eine Konvexkombination von Punkten $x_0, \dots, x_m \in M$ mit $m \geq 0$ ist, ist durch Lemma 13.7 gesichert. Der Satz besagt, dass bereits eine Teilmenge von (höchstens) $k+1$ dieser Punkte ausreicht, aus denen man x konvexkombinieren kann. Die Auswahl dieser Punkte hängt

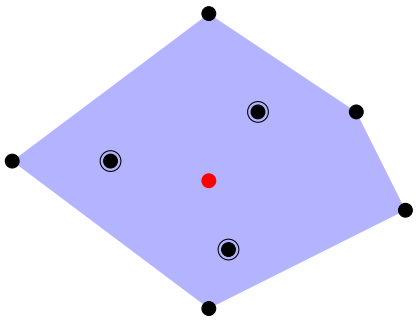


Abbildung 15.3: Illustration des Satzes von Carathéodory 15.13 für eine Menge der Dimension $k = 2$. Die schwarzen Punkte bilden die Menge M. Der rote Punkt $x \in \operatorname{conv}(M)$ ist bereits eine Konvexkombination der drei hervorgehobenen Punkte.

von x ab. Im Einzelfall können auch weniger als $k+1$ Punkte ausreichen. Das ist z. B. dann der Fall, wenn $x \in M$ ist.

Beweis. Der Beweis findet sich als Übungsaufgabe 3 auf Übungsblatt 10. □

Eine konvexe Menge $C \subseteq \mathbb{R}^n$, deren Dimension $\dim C < n$ ist, besitzt keine inneren Punkte in $\mathbb{R}^n$. Das ändert sich, wenn man zur Relativtopologie in $\operatorname{aff}(C)$ übergeht. Dies führt zu folgenden Begriffen:

Definition 15.14 (Relatives Inneres). *Es sei $C \subseteq \mathbb{R}^n$ eine konvexe Menge.*

(*i*) *Ein Punkt $x \in C$ heißt ein **relativ innerer Punkt** von C, wenn es ein $\varepsilon > 0$ gibt, sodass $B_\varepsilon(x) \cap \operatorname{aff}(C) \subseteq C$ liegt. Die Menge aller relativ inneren Punkte von C heißt das **relative Innere** (englisch:* relative interior*) und wird mit $\operatorname{rel\,int}(C)$ bezeichnet.*

(*ii*) *Ein Punkt $x \in \mathbb{R}^n$ heißt ein **relativer Randpunkt** von C, wenn $x \in \overline{C} \setminus \operatorname{rel\,int}(C)$ liegt. Die Menge aller relativen Randpunkte von C heißt der **relative Rand** (englisch:* relative boundary*) und wird mit $\operatorname{rel}\partial(C)$ bezeichnet.*

Quizfrage: Was ist $\operatorname{rel\,int}(C)$, wenn die konvexe Menge $C \subseteq \mathbb{R}^n$ die volle Dimension n hat?

Quizfrage: Was ist $\operatorname{rel\,int}(A)$ für einen affinen Unterraum $A \subseteq \mathbb{R}^n$?

Quizfrage: Wenn $C_1 \subseteq C_2$ ist, gilt dann auch immer $\operatorname{rel\,int}(C_1) \subseteq \operatorname{rel\,int}(C_2)$?

Quizfrage: Wodurch ist das relative Innere des Epigraphen $\operatorname{epi} f$ einer konvexen Funktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ charakterisiert?

Quizfrage: Warum benötigen wir keinen Begriff des relativen Abschlusses?

Satz 15.15 (vgl. Jarre, Stoer, 2004, Satz 7.2.5). *Jede nichtleere konvexe Menge $C \subseteq \mathbb{R}^n$ besitzt ein nichtleeres relatives Inneres. Es gilt:* $\dim \operatorname{rel\,int}(C) = \dim C$.

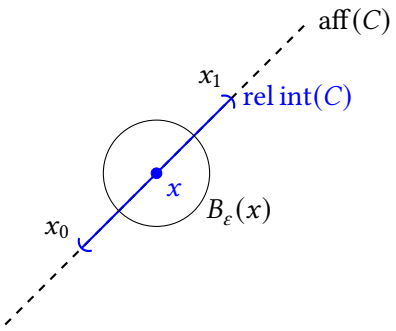


Abbildung 15.4: Illustration einer 1-dimensionalen konvexen Menge $C = \operatorname{conv}(\{x_0, x_1\}) = \{\alpha\, x_0 + (1-\alpha)\, x_1 \mid \alpha \in [0,1]\}$ und ihres relativen Inneren $\operatorname{rel\,int}(C) = \{\alpha\, x_0 + (1-\alpha)\, x_1 \mid \alpha \in (0,1)\}$ in $\mathbb{R}^2$. Insbesondere der Punkt x ist ein relativ innerer Punkt.

Beweis. Es sei $k := \dim C$. Da C nichtleer ist, gilt, $k \geq 0$. Es gibt also nach Satz 15.12 affin unabhängige Punkte $x_0, \ldots, x_k \in C$. Nach Lemma 15.7 lässt sich jeder Punkt $x \in \operatorname{aff}(C)$ eindeutig als Affinkombination $x = \sum_{i=0}^k \alpha_i\, x_i$ schreiben, wobei sich die Koeffizienten aus der eindeutigen Lösung des linearen Gleichungssystems (15.9) ergeben:

$$B^\mathsf{T} B\, \alpha = B^\mathsf{T} b.$$

Wir zeigen jetzt, dass der Mittelpunkt der Punkte $\{x_0, \ldots, x_k\}$

$$\overline{x} := \frac{1}{k+1} \sum_{i=0}^k x_i$$

ein relativ innerer Punkt von C ist. Dazu konstruieren wir eine abgeschlossene Kugel

$$\overline{B_\varepsilon^{\|\cdot\|_\infty}(\overline{x})} = \left\{ x \in \mathbb{R}^n \,\middle|\, \|x - \overline{x}\|_\infty \leq \varepsilon \right\}$$

mit der Eigenschaft $\overline{B_\varepsilon^{\|\cdot\|_\infty}(\overline{x})} \cap \operatorname{aff}(C) \subseteq \operatorname{conv}(\{x_0, \ldots, x_k\}) \subseteq \operatorname{conv}(C) = C$. (**Quizfrage:** Warum gilt die letzte Inklusion?)

Wir setzen dazu $\varepsilon := 1/\big((k+1)\|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty\big)$. Dabei ist $\|\cdot\|_\infty$ die durch die ∞-Norm im Definitions- und Bildbereich induzierte Matrixnorm, also[3]

$$\|A\|_\infty = \max_{x \neq 0} \frac{\|A\,x\|_\infty}{\|x\|_\infty} = \max_{\|x\|_\infty = 1} \|A\,x\|_\infty. \tag{15.11}$$

Es sei nun $x \in \overline{B_\varepsilon^{\|\cdot\|_\infty}(\overline{x})} \cap \operatorname{aff}(C)$. Wir bezeichnen mit α die Koeffizienten in der Affinkombination $x = \sum_{i=0}^k \alpha_i\, x_i$. Weiterhin sind $\overline{\alpha}$ die Koeffizienten von $\overline{x}$, also $\overline{\alpha} = \frac{1}{k+1}\mathbf{1}$. Um zu zeigen, dass tatsächlich $x \in \operatorname{conv}(\{x_0, \ldots, x_k\})$ liegt, müssen wir für die Koeffizienten zeigen: $\alpha \geq 0$. Es gilt

$$\begin{aligned}
\|\alpha - \overline{\alpha}\|_\infty &= \left\| (B^\mathsf{T}B)^{-1}B^\mathsf{T}\left[\begin{pmatrix}\frac{1}{x}\end{pmatrix} - \begin{pmatrix}\frac{1}{\overline{x}}\end{pmatrix}\right]\right\|_\infty \leq \|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty \left\|\begin{pmatrix}\frac{1}{x}\end{pmatrix} - \begin{pmatrix}\frac{1}{\overline{x}}\end{pmatrix}\right\|_\infty \\
&= \|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty \|x - \overline{x}\|_\infty \leq \|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty\, \varepsilon = \frac{1}{k+1}.
\end{aligned}$$

[3] Diese ist auch als **Zeilensummennorm** bekannt, da für $A \in \mathbb{R}^{m\times n}$ die Beziehung

$$\|A\|_\infty = \max\left\{ \sum_{j=1}^n |a_{ij}| \,\middle|\, i = 1, \ldots, m \right\}$$

gilt.

Daraus folgt wie gewünscht $\alpha \geq 0$ mit Hilfe der Dreiecksungleichung.

Wir zeigen jetzt noch, dass die Dimension von $\operatorname{rel\,int}(C)$ ebenfalls gleich k ist. Dazu geben wir $k+1$ affin unabhängige Punkte in $\operatorname{rel\,int}(C)$ an. Damit ist $\dim \operatorname{rel\,int}(C) \geq k$, und da außerdem $\operatorname{rel\,int}(C) \subseteq C$ die Beziehung $\dim \operatorname{rel\,int}(C) \leq \dim C = k$ impliziert, ist dann die Behauptung gezeigt. Zur Konstruktion der Punkte machen wir den Ansatz

$$\overline{x_i} \coloneqq \beta\,\overline{x} + (1-\beta)\,x_i$$

und wählen $\beta \in (0,1)$ so klein, dass $\|\overline{x_i} - \overline{x}\|_\infty \leq \varepsilon$ bleibt für alle $i = 0, \ldots, k$. Damit liegen alle $\overline{x_i} \in \overline{B_\varepsilon^{\|\cdot\|_\infty}(\overline{x})}$ und außerdem in $\operatorname{aff}(C)$ (**Quizfrage:** Begründung?).

Um die affine Unabhängigkeit der Punkte $\{\overline{x_0}, \ldots, \overline{x_k}\}$ zu zeigen, machen wir den Ansatz

$$\begin{aligned} 0 = \sum_{i=1}^{k} \gamma_i \big(\overline{x_i} - \overline{x_0}\big) &= \sum_{i=1}^{k} \gamma_i \big(\beta\,\overline{x} + (1-\beta)\,x_i - \beta\,\overline{x} - (1-\beta)\,x_0\big) \\ &= (1-\beta) \sum_{i=1}^{k} \gamma_i (x_i - x_0). \end{aligned}$$

Da die Punkte $\{x_0, \ldots, x_k\}$ aber affin unabhängig sind, folgt $\gamma_i = 0$ für alle $i = 1, \ldots, k$. Also sind auch die Punkte $\{\overline{x_0}, \ldots, \overline{x_k}\}$ affin unabhängig. □

Lemma 15.16 (**Accessibility lemma**, vgl. Jarre, Stoer, 2004, Lemma 7.2.6)**.**
Es seien $C \subseteq \mathbb{R}^n$ konvex, $x_1 \in \overline{C}$ und $x_2 \in \operatorname{rel\,int}(C)$. Dann gilt $\alpha\,x_1 + (1-\alpha)\,x_2 \in \operatorname{rel\,int}(C)$ für alle $\alpha \in [0,1)$, d. h., die gesamte Verbindungsstrecke (evtl. mit Ausnahme von x_1 selbst) gehört zum relativen Inneren von C.

Beweis. Aufgrund der Voraussetzung $x_1 \in \overline{C}$ gilt $x_1 \in C + B_\varepsilon(0)$ für alle $\varepsilon > 0$. Wegen $x_2 \in \operatorname{rel\,int}(C)$ gibt es ein $r > 0$, sodass $B_r(x_2) \cap \operatorname{aff}(C) \subseteq C$ liegt.

Es sei nun $\alpha \in [0,1)$ beliebig und $x \coloneqq \alpha\,x_1 + (1-\alpha)\,x_2$. Wir zeigen: Mit $r(\alpha) \coloneqq \frac{1-\alpha}{1+\alpha}\,r$ gilt $B_{r(\alpha)}(x) \cap \operatorname{aff}(C) \subseteq C$, d. h., $x \in \operatorname{rel\,int}(C)$. Dazu halten wir zunächst fest:

$$\begin{aligned} B_{r(\alpha)}(x) &= B_{r(\alpha)}\big(\alpha\,x_1 + (1-\alpha)\,x_2\big) \\ &= \alpha\,x_1 + (1-\alpha)\,x_2 + B_{r(\alpha)}(0) \\ &\subseteq \alpha\,\big(C + B_{r(\alpha)}(0)\big) + (1-\alpha)\,x_2 + B_{r(\alpha)}(0) \\ &= \alpha\,C + (1-\alpha)\,x_2 + (1+\alpha)\,B_{r(\alpha)}(0) \\ &= \alpha\,C + (1-\alpha)\big[x_2 + B_r(0)\big] \\ &= \alpha\,C + (1-\alpha)\,B_r(x_2). \end{aligned}$$

Durch den Schnitt mit $\operatorname{aff}(C)$ folgt

$$B_{r(\alpha)}(x) \cap \operatorname{aff}(C) \subseteq \big[\alpha\,C + (1-\alpha)\,B_r(x_2)\big] \cap \operatorname{aff}(C).$$

Wegen $\operatorname{aff}(C) = \alpha \operatorname{aff}(C) + (1-\alpha)\operatorname{aff}(C)$ (**Quizfrage:** Begründung?) gilt weiter

$$B_{r(\alpha)}(x) \cap \operatorname{aff}(C) \subseteq \alpha \underbrace{\big[C \cap \operatorname{aff}(C)\big]}_{\subseteq C} + (1-\alpha) \underbrace{\big[B_r(x_2) \cap \operatorname{aff}(C)\big]}_{\subseteq C} \subseteq C.$$

Die letzte Inklusion folgt aus der Konvexität von C. □

Wir geben noch eine nützliche Charakterisierung des relativen Inneren einer konvexen Menge an, die es uns erlaubt, mit einzelnen Richtungen zu argumentieren statt gleichzeitig mit allen Richtungen in einer Kugel:

Lemma 15.17 (Charakterisierung des relativen Inneren). *Es sei $C \subseteq \mathbb{R}^n$ konvex und nichtleer. Für einen Punkt $x \in \mathbb{R}^n$ sind folgende Aussagen äquivalent:*

(*i*) $x \in \operatorname{rel\,int}(C)$.

(*ii*) *Zu jedem $y \in \operatorname{aff}(C)$ existiert ein $\varepsilon > 0$, sodass $x + \varepsilon(y-x) \in C$ und $x - \varepsilon(y-x) \in C$ liegen.*

(*iii*) *Zu jedem $y \in C$ existiert ein $\varepsilon > 0$, sodass $x + \varepsilon(y-x) \in C$ und $x - \varepsilon(y-x) \in C$ liegen.*

Quizfrage: Was bedeutet die Bedingung aus Aussage (*iii*) anschaulich?

Beweis. Aussage (*i*) ⇒ Aussage (*ii*): Es sei $x \in \operatorname{rel\,int}(C)$ und $y \in \operatorname{aff}(C)$. Wir können $y \neq x$ annehmen, sonst ist die Aussage klar. Wegen $x \in \operatorname{rel\,int}(C)$ gibt es ein $\tilde{\varepsilon} > 0$, sodass $B_{\tilde{\varepsilon}}(x) \cap \operatorname{aff}(C) \subseteq C$. Für $\varepsilon := \tilde{\varepsilon}/\|y-x\|$ sind dann

$$\underbrace{x + \varepsilon(y-x)}_{=\varepsilon y + (1-\varepsilon)x \in \operatorname{aff}(C)} \in B_{\tilde{\varepsilon}}(x) \cap \operatorname{aff}(C) \subseteq C \quad \text{und} \quad \underbrace{x - \varepsilon(y-x)}_{=-\varepsilon y + (1+\varepsilon)x \in \operatorname{aff}(C)} \in B_{\tilde{\varepsilon}}(x) \cap \operatorname{aff}(C) \subseteq C.$$

Aussage (*ii*) ⇒ Aussage (*iii*): Klar, da $C \subseteq \operatorname{aff}(C)$ ist.

Aussage (*iii*) ⇒ Aussage (*i*): Nach Satz 15.15 können wir ein $y \in \operatorname{rel\,int}(C)$ wählen, insbesondere gilt $y \in C$. Nach Voraussetzung existiert $\varepsilon > 0$, sodass $z := x - \varepsilon(y-x) \in C$ liegt. Das heißt aber auch

$$x = \frac{1}{1+\varepsilon} z + \frac{\varepsilon}{1+\varepsilon} y,$$

d. h., x ist eine echte Konvexkombination von $z \in C$ und $y \in \operatorname{rel\,int}(C)$. Nach dem Accessibility lemma 15.16 gehört also x zu $\operatorname{rel\,int}(C)$. □

Lemma 15.18 (Relatives Inneres des Epigraphen, vgl. Rockafellar, 1970, Lemma 7.3). *Es sei $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex. Dann gilt*

$$\operatorname{rel\,int} \operatorname{epi} f = \{(x,\gamma) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \operatorname{rel\,int} \operatorname{dom} f,\ \gamma > f(x)\}. \tag{15.12}$$

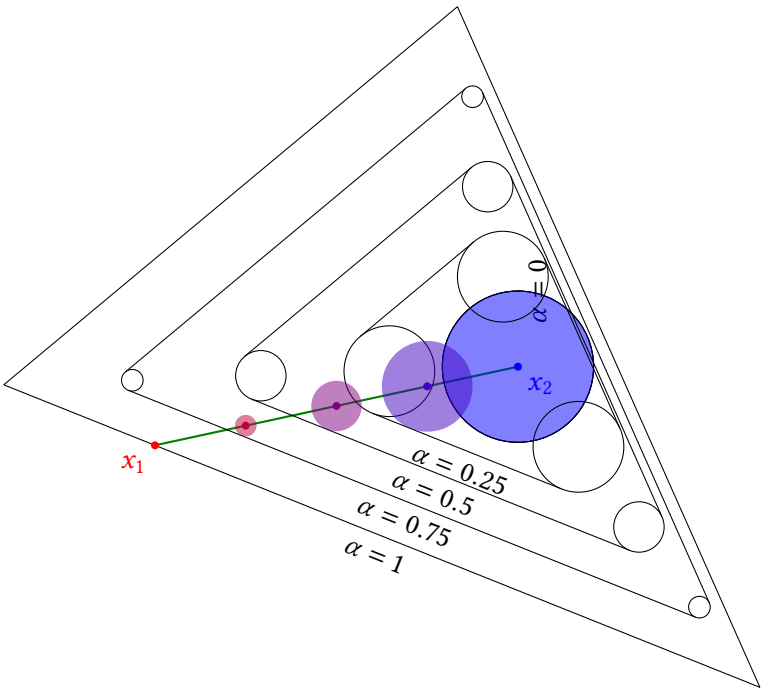


Abbildung 15.5: Illustration der Inklusion $B_{r(\alpha)}(x) \subseteq \alpha C + (1-\alpha) B_r(x_2)$ aus dem Beweis von Lemma 15.16 für $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$.

Beweis. □

Lemma 15.19 (Rockafellar, 1970, Theorem 6.2)**.** *Ist $C \subseteq \mathbb{R}^n$ konvex, dann sind das relative Innere* $\operatorname{rel\,int}(C)$ *und der Abschluss $\overline{C}$ konvex. Die Mengen C,* $\operatorname{rel\,int}(C)$ *und $\overline{C}$ haben alle dieselbe affine Hülle, also auch dieselbe Dimension.*

Beweis. Es sei $C \subseteq \mathbb{R}^n$ konvex. Wir zeigen zuerst, dass $\operatorname{rel\,int}(C)$ konvex ist. Es seien dazu $x_1, x_2 \in \operatorname{rel\,int}(C)$ und $\alpha \in [0,1]$. Dann gehört nach Lemma 15.16 auch die Konvexkombination $\alpha x_1 + (1-\alpha) x_2$ zu $\operatorname{rel\,int}(C)$, also ist $\operatorname{rel\,int}(C)$ konvex.

Der Abschluss von C erfüllt die Beziehung

$$\overline{C} = \bigcap \left\{ C + \overline{B_\varepsilon(0)} \,\middle|\, \varepsilon > 0 \right\}.$$

Die Mengen, über die der Durchschnitt gebildet wird, sind nach Beispiel 13.2 und Satz 13.3 konvex, also ist auch $\overline{C}$ konvex.

Wir zeigen jetzt noch $\operatorname{aff} \operatorname{rel\,int}(C) = \operatorname{aff} C = \operatorname{aff} \overline{C}$. Aufgrund von $\operatorname{rel\,int}(C) \subseteq C \subseteq \overline{C}$ folgt

$$\operatorname{aff} \operatorname{rel\,int}(C) \subseteq \operatorname{aff} C \subseteq \operatorname{aff} \overline{C}.$$

Wegen $\overline{C} \subseteq \operatorname{aff}(C)$ gilt auch $\operatorname{aff} \overline{C} \subseteq \operatorname{aff} \operatorname{aff}(C) = \operatorname{aff}(C)$. Also bleibt noch $\operatorname{aff} \operatorname{rel\,int}(C) = \operatorname{aff} C$ zu zeigen. Wir wissen aber bereits aus Satz 15.15, dass $\dim \operatorname{rel\,int}(C) = \dim C$ gilt, also muss $\operatorname{aff} \operatorname{rel\,int}(C) = \operatorname{aff} C$ sein. □

Satz 15.20 (Rockafellar, 1970, Theorem 6.3)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex. Dann gelten:

(*i*) $\overline{\operatorname{rel\,int}(C)} = \overline{C}$, *d. h.,* $\operatorname{rel\,int}(C)$ *and* C *haben denselben Abschluss.*

(*ii*) $\operatorname{rel\,int}(C) = \operatorname{rel\,int}(\overline{C})$, *d. h., C und $\overline{C}$ haben dasselbe relative Innere.*

(*iii*) $\operatorname{rel}\partial(C) = \operatorname{rel}\partial(\overline{C})$, *d. h., C und $\overline{C}$ haben denselben relativen Rand.*

Beweis. Wir können $C \neq \emptyset$ annehmen, ansonsten sind alle Mengen leer.

Aussage (*i*): Die Inklusion $\overline{\operatorname{rel\,int}(C)} \subseteq \overline{C}$ folgt unmittelbar aus $\operatorname{rel\,int}(C) \subseteq C$. Für die umgekehrte Aussage sei nun $x_1 \in \overline{C}$ und $x_2 \in \operatorname{rel\,int}(C)$. Dann ist nach Lemma 15.16 $\alpha x_1 + (1-\alpha) x_2 \in \operatorname{rel\,int}(C)$. Der Grenzübergang $\alpha \nearrow 1$ zeigt $x_1 \in \overline{\operatorname{rel\,int}(C)}$.

Aussage (*ii*): Aus $C \subseteq \overline{C}$ und der Tatsache, dass beide Mengen dieselbe affine Hülle haben (Lemma 15.19) folgt $\operatorname{rel\,int}(C) \subseteq \operatorname{rel\,int}(\overline{C})$. Für die umgekehrte Aussage sei nun $x \in \operatorname{rel\,int}(\overline{C})$, also existiert $\varepsilon > 0$ mit $B_\varepsilon(x) \cap \operatorname{aff}(C) \subseteq \overline{C}$. Es sei außerdem $y \in \operatorname{rel\,int}(C)$ (Satz 15.15). Ist $y = x$, so sind wir fertig. Es sei also jetzt $y \neq x$. Ziel ist die Konstruktion eines Punktes $z \in \overline{C}$, sodass x als echte Konvexkombination von $y \in \operatorname{rel\,int}(C)$ und $z \in \overline{C}$ geschrieben werden kann. Denn dann folgt aus Lemma 15.16, dass $x \in \operatorname{rel\,int}(C)$ liegt.

Wir definieren

$$z := x + \delta\,(x - y) \quad \text{mit } \delta := \frac{\varepsilon}{\|x - y\|}.$$

Dann ist $\|z - x\| = \delta\,\|x - y\| = \varepsilon$, also gilt $z \in \overline{B_\varepsilon(x)}$. Wegen $\overline{B_\varepsilon(x)} \subseteq \overline{C}$ (**Quizfrage:** Begründung?) folgt auch $z \in \overline{C}$. Wir können nun x schreiben als echte Konvexkombination

$$x = \frac{1}{1+\delta}\, z + \left(1 - \frac{1}{1+\delta}\right) y$$

mit $z \in \overline{C}$ und $y \in \operatorname{rel\,int}(C)$. Nach Lemma 15.16 gilt $x \in \operatorname{rel\,int}(C)$, was zu zeigen war.

Aussage (*iii*): Nach Definition des relativen Randes gilt

$$\operatorname{rel}\partial(C) = \overline{C} \setminus \operatorname{rel\,int}(C)$$

und weiter

$$\operatorname{rel}\partial(\overline{C}) = \overline{C} \setminus \operatorname{rel\,int}(\overline{C}) = \overline{C} \setminus \operatorname{rel\,int}(C),$$

wobei die letzte Gleichheit aus Aussage (*i*) folgt. □

Folgerung 15.21 (Gleichheit der Abschlüsse und der relativen Inneren konvexer Mengen)**.**
Es sei $C_1, C_2 \subseteq \mathbb{R}^n$ konvex. Die folgenden Aussagen sind äquivalent:

(*i*) $\overline{C_1} = \overline{C_2}$.

(*ii*) $\operatorname{rel\,int}(C_1) = \operatorname{rel\,int}(C_2)$.

Beweis. Aussage (*i*) $\Rightarrow$ Aussage (*ii*): Aus $\overline{C_1} = \overline{C_2}$ folgt $\operatorname{rel\,int}(\overline{C_1}) = \operatorname{rel\,int}(\overline{C_2})$. Satz 15.20 (*ii*) impliziert $\operatorname{rel\,int}(C_1) = \operatorname{rel\,int}(C_2)$.

Aussage (*ii*) $\Rightarrow$ Aussage (*i*): Aus $\operatorname{rel\,int}(C_1) = \operatorname{rel\,int}(C_2)$ folgt $\overline{\operatorname{rel\,int}(C_1)} = \overline{\operatorname{rel\,int}(C_2)}$. Satz 15.20 (*i*) impliziert $\overline{C_1} = \overline{C_2}$. □

Definition 15.22 (Algebraisches Inneres). *Es sei $M \subseteq \mathbb{R}^n$. Ein Punkt $x_0 \in M$ heißt ein* ***algebraisch innerer Punkt*** *von M, wenn gilt: Für jedes $d \in \mathbb{R}^n$ existiert ein $\varepsilon_d > 0$, sodass*

$$\{x_0 + t\,d \mid 0 \leq t < \varepsilon_d\} \subseteq M$$

liegt. Die Menge aller algebraisch inneren Punkte von M heißt das ***algebraische Innere*** *(englisch:* algebraic interior *oder englisch:* core*) von M, kurz:* $\operatorname{core} M$.

Es ist leicht zu sehen, dass $\operatorname{int} M \subseteq \operatorname{core} M$ gilt. Die Umkehrung ist jedoch i. A. falsch. Es gilt aber:

Lemma 15.23 (Algebraisches Inneres konvexer Mengen). *Es sei $C \subseteq \mathbb{R}^n$ konvex. Dann gilt* $\operatorname{int} C = \operatorname{core} C$.

Beweis. Es ist nur $\operatorname{core} C \subseteq \operatorname{int} C$ zu zeigen. Im Fall $\operatorname{core} C = \emptyset$ ist nichts zu zeigen. Es sei also $x_0 \in \operatorname{core} C$. Dann ist $\dim C = n$ und damit $\operatorname{aff} C = \mathbb{R}^n$. (**Quizfrage:** Begründung?) Nach Definition des algebraischen Inneren ist die Bedingung aus Lemma 15.17 (*ii*) erfüllt. Damit gehört x_0 zu $\operatorname{rel\,int} C = \operatorname{int} C$. □

Ende der Woche 10

§ 15.4 Trennungssätze

Literatur: Geiger, Kanzow, 2002, Kapitel 2.1.4

Ziel: Trennung konvexer Mengen durch eine Hyperebene, sodass jede der Mengen in einem anderen Halbraum liegt

Definition 15.24 (Trennende Hyperebene). *Es seien $A, B \subseteq \mathbb{R}^n$ zwei nichtleere Mengen und $H(a, \beta) = \{x \in \mathbb{R}^n \mid a^\mathsf{T} x = \beta\}$ eine Hyperebene.*

(*i*) *Wir sagen, $H(a, \beta)$ sei eine* ***trennende Hyperebene*** *(englisch:* separating hyperplane*) für die Mengen A und B, falls eine der Mengen in $H^-(a, \beta)$ und die andere in $H^+(a, \beta)$ enthalten ist, wenn also gilt*

$$a^\mathsf{T} x \leq \beta \leq a^\mathsf{T} y \quad \text{für alle } x \in A \text{ und alle } y \in B. \tag{15.13}$$

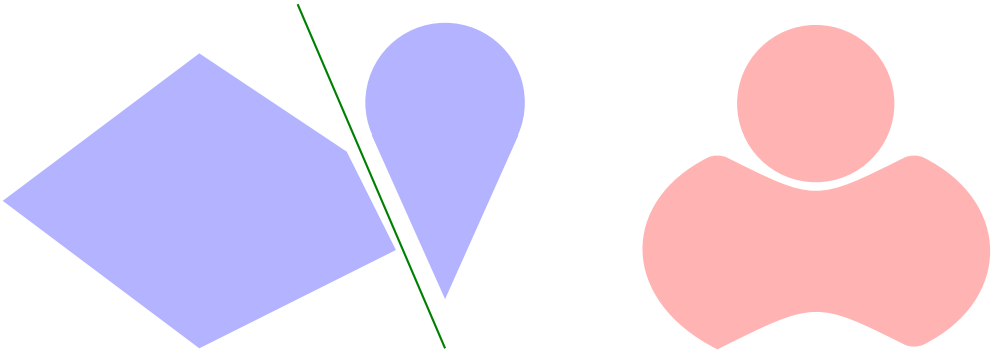


Abbildung 15.6: Zwei disjunkte konvexe Mengen (links) sind durch eine Hyperebene trennbar. Ist eine der Mengen nichtkonvex, stimmt diese Aussage nicht mehr (rechts).

(*ii*) *Wir sagen, die Hyperebene $H(a,\beta)$ sei eine* ***eigentlich trennende Hyperebene*** *(englisch:* properly separating hyperplane*) für die Mengen A und B, falls $H(a,\beta)$ die Mengen A und B trennt, aber nicht beide Mengen A und B enthält, wenn also gilt:*

$$a^\mathsf{T}x \le \beta \le a^\mathsf{T}y \quad \text{für alle } x \in A \text{ und alle } y \in B \tag{15.14a}$$

und

$$a^\mathsf{T}\overline{x} < a^\mathsf{T}\overline{y} \quad \text{für ein } \overline{x} \in A \text{ und ein } \overline{y} \in B. \tag{15.14b}$$

(*iii*) *Wir sagen, die Hyperebene $H(a,\beta)$ sei eine* ***strikt trennende Hyperebene*** *(englisch:* strictly separating hyperplane*) für die Mengen A und B, falls genau eine der Mengen im offenen Halbraum $\operatorname{int} H^-(a,\beta)$ und die andere in $\operatorname{int} H^+(a,\beta)$ enthalten ist, wenn also gilt*

$$a^\mathsf{T}x < \beta < a^\mathsf{T}y \quad \text{für alle } x \in A \text{ und alle } y \in B. \tag{15.15}$$

Beachte: Der „Offset“ β ist nicht die entscheidende Größe in der Definition 15.24. Gilt beispielsweise $a^\mathsf{T}x \le a^\mathsf{T}y$ für alle $x \in A$ und alle $y \in B$, dann lässt sich β in (15.13) immer nachträglich passend definieren.

Quizfrage: Wie kann man in den einzelnen Fällen der Definition 15.24 das β jeweils passend definieren?

Wir beginnen mit einer Aussage zur Trennung eines Punktes und einer konvexen Menge.

Lemma 15.25 (Trennung von Punkt und konvexer Menge)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex und nichtleer und $\overline{x} \in \mathbb{R}^n$ ein Punkt, sodass $\overline{x} \notin \operatorname{int}(C)$. Dann existiert eine Hyperebene $H(a,\beta)$, die $\overline{x}$ von C trennt, also

$$a^\mathsf{T}x \ge \beta \ge a^\mathsf{T}\overline{x} \quad \text{für alle } x \in C. \tag{15.16}$$

Beweis. Wir wählen eine Folge $(x^{(k)})$ mit der Eigenschaft $x^{(k)} \notin \overline{C}$ und $x^{(k)} \to \overline{x}$ gilt. **Quizfrage:** Wieso ist das möglich? Die Menge $\overline{C}$ ist nichtleer, abgeschlossen und konvex. Nach Lemma 15.2

existiert die orthogonale Projektion $\widehat{x}^{(k)} := \operatorname{proj}_{\overline{C}}(x^{(k)})$, und nach Satz 15.3 ist diese charakterisiert durch

$$\begin{aligned}(\widehat{x}^{(k)} - x^{(k)})^\mathsf{T} x &\geq (\widehat{x}^{(k)} - x^{(k)})^\mathsf{T}\widehat{x}^{(k)} \quad \text{für alle } k \in \mathbb{N} \text{ und alle } x \in \overline{C} \\ &= (\widehat{x}^{(k)} - x^{(k)})^\mathsf{T}(\widehat{x}^{(k)} - x^{(k)} + x^{(k)}) \\ &= \|\widehat{x}^{(k)} - x^{(k)}\|^2 + (\widehat{x}^{(k)} - x^{(k)})^\mathsf{T} x^{(k)} \\ &\geq (\widehat{x}^{(k)} - x^{(k)})^\mathsf{T} x^{(k)}. \end{aligned} \tag{15.17}$$

Wir setzen

$$a^{(k)} := \frac{\widehat{x}^{(k)} - x^{(k)}}{\|\widehat{x}^{(k)} - x^{(k)}\|}.$$

Beachte: Der Nenner ist $\neq 0$, da $\widehat{x}^{(k)} \in \overline{C}$ und $x^{(k)} \notin \overline{C}$.

Damit erhalten wir aus (15.17)

$$a^{(k)^\mathsf{T}} x \geq a^{(k)^\mathsf{T}} x^{(k)} \quad \text{für alle } k \in \mathbb{N} \text{ und alle } x \in \overline{C},$$

d. h., mit dem Normalenvektor $a^{(k)}$ können wir jeweils den Punkt $x^{(k)}$ von $\overline{C}$ trennen. Wegen $\|a^{(k)}\| = 1$ und der Kompaktheit der Einheitssphäre in $\mathbb{R}^n$ existiert eine konvergente Teilfolge $a^{(k^{(\ell)})} \to a$ mit $\|a\| = 1$, und der Grenzübergang $\ell \to \infty$ zeigt die Behauptung

$$a^\mathsf{T} x \geq a^\mathsf{T} \overline{x} \quad \text{für alle } x \in \overline{C}. \qquad \square$$

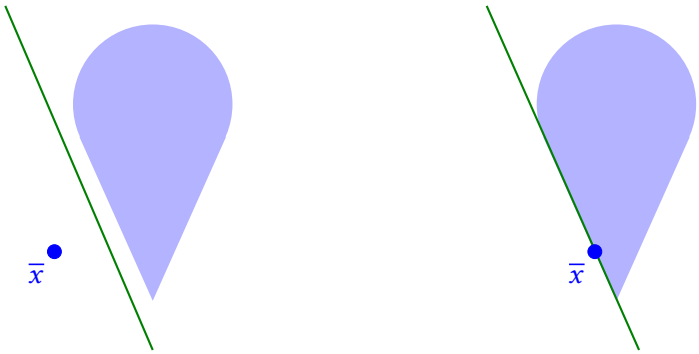


Abbildung 15.7: Illustration von Lemma 15.25 in zwei Fällen.

Satz 15.26 (Trennungssatz).
Es seien $C_1, C_2 \subseteq \mathbb{R}^n$ konvex und nichtleer sowie $C_1 \cap C_2 = \emptyset$. Dann existiert eine Hyperebene $H(a, \beta)$, die C_1 und C_2 trennt, also

$$a^\mathsf{T} x_1 \leq \beta \leq a^\mathsf{T} x_2 \quad \textit{für alle } x_1 \in C_1 \textit{ und alle } x_2 \in C_2. \tag{15.18}$$

Beweis. Wir betrachten

$$C := C_2 - C_1 = \{x_2 - x_1 \,|\, x_1 \in C_1,\ x_2 \in C_2\},$$

also die Minkowski-Summe von C_2 und $-C_1$. Nach Satz 13.3 ist C konvex, und wegen $C_1 \cap C_2 = \emptyset$ gilt $0 \notin C$, also erst recht $0 \notin \operatorname{int}(C)$. Aus Lemma 15.25 bekommen wir die Existenz einer Hyperebene $H(a, \beta)$, sodass $0 \leq a^\mathsf{T} x$ gilt für alle $x \in C$. Das heißt aber

$$a^\mathsf{T} x_1 \leq a^\mathsf{T} x_2 \quad \text{für alle } x_1 \in C_1 \text{ und alle } x_2 \in C_2.$$

□

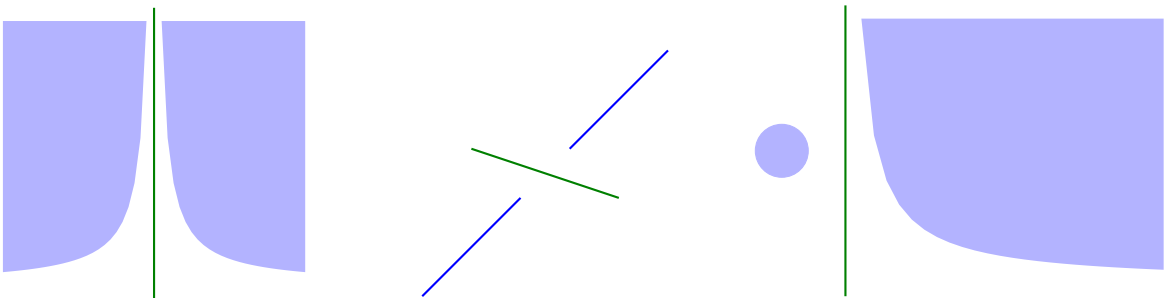


Abbildung 15.8: Illustration der Aussagen von Satz 15.26 (Trennung disjunkter konvexer Mengen), Satz 15.30 (eigentliche Trennung) und Satz 15.32 (strikte Trennung).

Für den folgenden, sogenannten eigentlichen Trennungssatz 15.30 benötigen wir zwei vorbereitende Lemmas.

Lemma 15.27 (Abschluss und relatives Inneres unter linearen Transformationen, vgl. Rockafellar, 1970, Theorem 6.6)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex und $A \in \mathbb{R}^{m \times n}$. Dann ist $AC \subseteq \mathbb{R}^m$ konvex, und es gilt:

(*i*) $A\overline{C} \subseteq \overline{AC}$.

(*ii*) $A \operatorname{rel\,int}(C) = \operatorname{rel\,int}(AC)$.

Beweis. Im Fall $C = \emptyset$ ist nichts zu zeigen. Wir gehen also ab jetzt von $C \neq \emptyset$ aus. Die Konvexität von AC ist offensichtlich.

Aussage (*i*): Es sei $\overline{x} \in \overline{C}$, dann existiert eine Folge $(x^{(k)}) \subseteq C$ mit der Eigenschaft $x^{(k)} \to \overline{x}$. Das impliziert $Ax^{(k)} \in AC$ und $Ax^{(k)} \to A\overline{x}$, also gilt $A\overline{x} \in \overline{AC}$.

Aussage (*ii*): Es gilt

$$\begin{aligned} A \operatorname{rel\,int}(C) \subseteq AC \subseteq A\overline{C} &= A\,\overline{\operatorname{rel\,int}(C)} \quad \text{nach Satz 15.20} \\ &\subseteq \overline{A \operatorname{rel\,int}(C)} \quad \text{nach Aussage (i).} \end{aligned}$$

Die Bildung des Abschlusses in allen Termen dieser Ungleichung zeigt $\overline{AC} = \overline{A \operatorname{rel\,int}(C)}$. Aus Folgerung 15.21 folgt damit:

$$\operatorname{rel\,int}(AC) = \operatorname{rel\,int}(A \operatorname{rel\,int}(C)) \subseteq A \operatorname{rel\,int}(C).$$

Um die umgekehrte Ungleichung zu zeigen, sei $z \in A \operatorname{rel\,int}(C)$, d. h., $z = Az'$ für ein $z' \in \operatorname{rel\,int}(C)$. Weiter sei x irgendein Element von AC, d. h., $x = Ax'$ für ein $x' \in C$. Aus Lemma 15.17 *(i)* $\Rightarrow$ *(iii)* folgt, dass $y' := z' \pm \varepsilon\,(x' - z')$ für ein geeignetes $\varepsilon > 0$ in C liegt, also

$$y := Ay' = z \pm \varepsilon\,(x - z) \in AC.$$

Das heißt aber, dass die Voraussetzung aus Lemma 15.17 *(iii)* für die Menge AC erfüllt ist, also gehört z zu $\operatorname{rel\,int}(AC)$, was $A \operatorname{rel\,int}(C) \subseteq \operatorname{rel\,int}(AC)$ zeigt. □

Beachte: Die Aussage *(i)* verwendet nur die Stetigkeit linearer Abbildungen und gilt auch für nichtkonvexe Mengen.

Folgerung 15.28 (Relatives Inneres der Minskowski-Summe, vgl. Jarre, Stoer, 2004, Satz 7.2.8)**.** *Es seien $C_1, C_2 \subseteq \mathbb{R}^n$ konvex. Dann gilt*

$$\operatorname{rel\,int}(C_1) + \operatorname{rel\,int}(C_2) = \operatorname{rel\,int}(C_1 + C_2). \tag{15.19}$$

Beweis. Setze $C := C_1 \times C_2 \subseteq \mathbb{R}^n \times \mathbb{R}^n \cong \mathbb{R}^{2n}$ und $A := \begin{bmatrix} \mathrm{Id} & \mathrm{Id} \end{bmatrix}$, sodass also $A\left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right) = x_1 + x_2$ gilt. Dann ist C nach Satz 13.3 konvex, und aus Lemma 15.27 folgt

$$\operatorname{rel\,int}(C_1) + \operatorname{rel\,int}(C_2) = A \operatorname{rel\,int}(C) = \operatorname{rel\,int}(AC) = \operatorname{rel\,int}(C_1 + C_2).$$ □

Lemma 15.29 (Eigentliche Trennung von Punkt und konvexer Menge)**.** *Es sei $C \subseteq \mathbb{R}^n$ konvex und nichtleer. Falls $0 \notin \operatorname{rel\,int}(C)$ liegt, dann lassen sich 0 und C durch eine Hyperebene $H(a, \beta)$ eigentlich trennen.*

Beweis. Im Fall $0 \notin \overline{C}$, wählen wir $a := \operatorname{proj}_{\overline{C}}(0)$. Dann ist $a \neq 0$. Der Projektionssatz 15.3 impliziert

$$(a - 0)^\mathsf{T}(x - a) \geq 0 \quad \text{für alle } x \in C,$$

also gilt $a^\mathsf{T}x \geq \|a\|^2 > 0 = a^\mathsf{T}0$ für alle $x \in C$. Dies zeigt die eigentliche (und sogar die starke) Trennung von 0 und C in diesem Fall.

Andernfalls gilt $0 \in \overline{C} \setminus \operatorname{rel\,int}(C) = \operatorname{rel}\partial(C)$. Wegen $0 \in \overline{C} \subseteq \operatorname{aff}(C)$ ist $U := \operatorname{aff}(C)$ ein Unterraum von $\mathbb{R}^n$. Da 0 ein relativer Randpunkt von C ist, schneidet jede Kugel $B_\varepsilon(0)$ sowohl U als auch $\overline{C}$ als auch das Komplement $\mathbb{R}^n \setminus \overline{C}$. Durch Setzen von $\varepsilon := 1/k$ können wir also eine Folge $(x^{(k)}) \subseteq U$ konstruieren mit der Eigenschaft $x^{(k)} \notin \overline{C}$ und $x^{(k)} \to 0$. Wir setzen

$$\widehat{x}^{(k)} := \operatorname{proj}_{\overline{C}}(x^{(k)}).$$

Diese Folge erfüllt $\widehat{x}^{(k)} \neq x^{(k)}$ und $\widehat{x}^{(k)} \to 0$, und der Projektionssatz 15.3 impliziert

$$(\widehat{x}^{(k)} - x^{(k)})^\mathsf{T}(x - \widehat{x}^{(k)}) \geq 0 \quad \text{für alle } x \in C. \tag{15.20}$$

Wir setzen

$$a^{(k)} := \frac{\widehat{x}^{(k)} - x^{(k)}}{\|\widehat{x}^{(k)} - x^{(k)}\|}$$

und schreiben (15.20) als

$$(a^{(k)})^\mathsf{T} x \geq (a^{(k)})^\mathsf{T} \widehat{x}^{(k)} \quad \text{für alle } x \in C. \tag{15.21}$$

Wegen $\|a^{(k)}\| = 1$ gibt es eine konvergente Teilfolge $a^{(k^{(\ell)})} \to a$ mit $\|a\| = 1$ und insbesondere $a \neq 0$. Weiter gehört wegen $x^{(k)} \in U$ und $\widehat{x}^{(k)} \in \overline{C} \subseteq U$ auch $a^{(k)}$ zu U, und damit gilt auch $a \in U$. Der Grenzübergang auf der Teilfolge in (15.21) zeigt schließlich

$$a^\mathsf{T} x \geq a^\mathsf{T} 0 = 0 \quad \text{für alle } x \in C$$

Es bleibt zu zeigen, dass es ein $\overline{x} \in C$ gibt, sodass $a^\mathsf{T}\overline{x} > 0$ gilt. Nehmen wir an, dies sei nicht der Fall, d. h., es gelte $a^\mathsf{T} x = 0$ für alle $x \in C$. Weil $a \in U = \operatorname{aff}(C)$ liegt, gibt es Punkte $x_i \in C$ und Koeffzienten α_i, sodass

$$a = \sum_{i=0}^{m} \alpha_i x_i$$

gilt sowie $\sum_{i=0}^{m} \alpha_i = 1$. Das impliziert aber

$$\|a\|^2 = a^\top \sum_{i=0}^{m} \alpha_i x_i = \sum_{i=0}^{m} \alpha_i \underbrace{a^\top x_i}_{=0} = 0,$$

im Widerspruch zu $\|a\| = 1$. Folglich muss es ein $\overline{x} \in C$ geben, sodass $a^\mathsf{T}\overline{x} > 0$ gilt. □

Satz 15.30 (Eigentlicher Trennungssatz, vgl. Jarre, Stoer, 2004, Satz 7.2.8)**.**
Es seien $C_1, C_2 \subseteq \mathbb{R}^n$ konvex und nichtleer. Genau dann lassen sich C_1 und C_2 durch eine Hyperebene $H(a, \beta)$ eigentlich trennen, wenn $\operatorname{rel\,int}(C_1) \cap \operatorname{rel\,int}(C_2) = \emptyset$ *gilt.*

Beweis. Wir nehmen zunächst an, dass C_1 und C_2 durch $H(a, \beta)$ eigentlich getrennt werden. Es gilt also $a^\mathsf{T} x \leq \beta \leq a^\mathsf{T} y$ für alle $x \in C_1$ und alle $y \in C_2$, und es gibt $\overline{x} \in C_1$ und $\overline{y} \in C_2$, für die $a^\mathsf{T}\overline{x} < a^\mathsf{T}\overline{y}$ gilt. Wir zeigen nun:

$$a^\mathsf{T} x < a^\mathsf{T} y \quad \text{für alle } x \in \operatorname{rel\,int}(C_1) \text{ und alle } y \in \operatorname{rel\,int}(C_2), \tag{15.22}$$

woraus dann $\operatorname{rel\,int}(C_1) \cap \operatorname{rel\,int}(C_2) = \emptyset$ folgt. Nehmen wir an, dass (15.22) unwahr ist, dann gibt es ein $x \in \operatorname{rel\,int}(C_1)$ und ein $y \in \operatorname{rel\,int}(C_2)$ mit $a^\mathsf{T} x = a^\mathsf{T} y$. Für hinreichend kleines $\varepsilon > 0$ sind $\widehat{x} := x - \varepsilon\,(\overline{x} - x) \in C_1$ und $\widehat{y} := y - \varepsilon\,(\overline{y} - y) \in C_2$. Dann ist aber

$$a^\mathsf{T}(\widehat{x} - \widehat{y}) = a^\mathsf{T}\big(x - \varepsilon\,(\overline{x} - x) - y + \varepsilon\,(\overline{y} - y)\big) = \varepsilon\, a^\mathsf{T}(\overline{y} - \overline{x}) > 0$$

im Widerspruch zu $a^\mathsf{T} x \leq a^\mathsf{T} y$ für alle $x \in C_1$ und alle $y \in C_2$.

Umgekehrt nehmen wir nun an, dass $\operatorname{rel\,int}(C_1) \cap \operatorname{rel\,int}(C_2) = \emptyset$ gilt. Insbesondere ist dann

$$0 \notin \operatorname{rel\,int}(C_1) - \operatorname{rel\,int}(C_2) = \operatorname{rel\,int}(C_1 - C_2),$$

wobei die Gleichheit der Mengen wie in Folgerung 15.28 folgt. Nach Lemma 15.29 können also 0 und $C := C_1 - C_2$ eigentlich getrennt werden. Das heißt, es gibt ein $a \in \mathbb{R}^n$, $a \neq 0$, sodass $a^\mathsf{T} x \leq 0$ gilt für alle $x \in C$, und es existiert ein $\overline{x} \in C$ mit $a^\mathsf{T}\overline{x} < 0$. Das bedeutet aber $a^\mathsf{T} x_1 \leq a^\mathsf{T} x_2$ für alle $x_1 \in C_1$ und alle $x_2 \in C_2$, und für gewisse $\overline{x}_1 \in C_1$ und $\overline{x}_2 \in C_2$ gilt $a^\mathsf{T}\overline{x}_1 < a^\mathsf{T}\overline{x}_2$. □

Wir bereiten nun den strikten Trennungssatz vor.

Lemma 15.31 (Abgeschlossenheit der Minkowskisumme)**.**
Es seien $M_1, M_2 \subseteq \mathbb{R}^n$, M_1 abgeschlossen und M_2 kompakt. Dann ist die Minkowski-Summe $M_1 + M_2$ abgeschlossen.

Beachte: Die Konvexität der beiden Mengen spielt hier keine Rolle.

Beweis. Wir setzen $M := M_1 + M_2$. Falls M_1 oder M_2 die leere Menge ist, dann ist $M = \emptyset$ und die Aussage klar. Es sei nun also $M_1, M_2 \neq \emptyset$ und damit $M \neq \emptyset$. Weiter sei $(z^{(k)}) \subseteq M$ eine konvergente Folge mit Grenzwert z. Es existieren also Folgen $(x^{(k)}) \subseteq M_1$ und $(y^{(k)}) \subseteq M_2$ mit $z^{(k)} = x^{(k)} + y^{(k)}$.

$$M_2 \text{ ist beschränkt} \quad \Rightarrow \quad (y^{(k)}) \text{ ist beschränkt} \quad \Rightarrow \quad (x^{(k)}) \text{ ist beschränkt.}$$

Es existieren also konvergente Teilfolgen $x^{(k^{(\ell)})} \to x$ und $y^{(k^{(\ell)})} \to y$, sodass auch $z^{(k^{(\ell)})} = x^{(k^{(\ell)})} + y^{(k^{(\ell)})} \to x + y$ für $\ell \to \infty$. M_1 und M_2 sind abgeschlossen, also liegt $x + y \in M_1 + M_2 = M$. Andererseits konvergiert $z^{(k^{(\ell)})}$ auch gegen z, also gilt $z = x + y \in M$, d. h., M ist abgeschlossen. □

Satz 15.32 (Strikter Trennungssatz, vgl. Geiger, Kanzow, 2002, Satz 2.24, S.36)**.**
Es seien $C_1, C_2 \subseteq \mathbb{R}^n$ konvex und nichtleer sowie $C_1 \cap C_2 = \emptyset$. Weiter sei C_1 abgeschlossen und C_2 kompakt. Dann existiert eine Hyperebene $H(a, \beta)$, die C_1 und C_2 strikt trennt, also

$$a^\mathsf{T} x_1 < \beta < a^\mathsf{T} x_2 \quad \text{für alle } x_1 \in C_1 \text{ und alle } x_2 \in C_2. \tag{15.23}$$

Beweis. Wir betrachten die Optimierungsaufgabe

$$\text{Minimiere} \quad \|x_1 - x_2\|, \quad (x_1, x_2) \in C_1 \times C_2. \tag{15.24}$$

Eine Lösung (x_1^*, x_2^*) dieser Aufgabe, sofern existent, realisiert den Abstand zwischen C_1 und C_2, also $\inf\{\|x_1 - x_2\| \mid x_1 \in C_1,\ x_2 \in C_2\}$.

Um die Existenz einer Lösung zu zeigen, halten wir fest, dass die Menge $C := C_1 - C_2$ nichtleer sowie nach Satz 13.3 konvex und nach Lemma 15.31 abgeschlossen ist. (15.24) ist also gleichzeitig eine Projektionsaufgabe (15.1), und zwar projizieren wir den Vektor $0 \in \mathbb{R}^n$ auf C. Es sei nun $(x_1^*, x_2^*) \in C_1 \times C_2$ die nach Lemma 15.2 eindeutige Lösung von (15.24). Wir konstruieren daraus nun die Daten (a, β) der Hyperebene und setzen dazu

$$a := \frac{x_2^* - x_1^*}{2} \neq 0, \quad \widehat{x} := \frac{x_1^* + x_2^*}{2}, \quad \beta := a^\mathsf{T} \widehat{x}.$$

Wir zeigen nun, dass

$$x_1^* = \operatorname{proj}_{C_1}(\widehat{x}), \quad x_2^* = \operatorname{proj}_{C_2}(\widehat{x}). \tag{15.25}$$

gilt. Dazu seien $x_1 \in C_1$ und $x_2 \in C_2$ beliebig, dann gilt

$$\begin{aligned}
\|x_1^* - \widehat{x}\| + \|\widehat{x} - x_2^*\| &= \|x_1^* - \widehat{x} + (\widehat{x} - x_2^*)\|, && \text{denn } x_1^* - \widehat{x} = \widehat{x} - x_2^* \\
&= \|x_1^* - x_2^*\| \\
&\le \|x_1 - x_2\|, && \text{denn } (x_1^*, x_2^*) \text{ ist optimal für (15.24)} \\
&\le \|x_1 - \widehat{x}\| + \|\widehat{x} - x_2\| && \text{wegen der Dreiecksungleichung.}
\end{aligned}$$

Setzen wir speziell $x_2 = x_2^*$ ein, so folgt

$$\|x_1^* - \widehat{x}\| \le \|x_1 - \widehat{x}\| \quad \text{für alle } x_1 \in C_1.$$

Setzen wir dagegen $x_1 = x_1^*$, so folgt

$$\|x_2^* - \widehat{x}\| \le \|x_2 - \widehat{x}\| \quad \text{für alle } x_2 \in C_2.$$

Dies bestätigt (15.25). Aus dem Projektionssatz 15.3 folgt daher

$$-a^\mathsf{T}(x_1 - x_1^*) = (x_1^* - \widehat{x})^\mathsf{T}(x_1 - x_1^*) \ge 0 \quad \text{für alle } x_1 \in C_1.$$

Dies impliziert

$$a^\mathsf{T}x_1 \le a^\mathsf{T}x_1^* = a^\mathsf{T}\widehat{x} + a^\mathsf{T}(x_1^* - \widehat{x}) = \beta - \|a\|^2 < \beta \quad \text{für alle } x_1 \in C_1.$$

Analog zeigt man

$$a^\mathsf{T}x_2 \ge \beta + \|a\|^2 > \beta \quad \text{für alle } x_2 \in C_2. \qquad \square$$

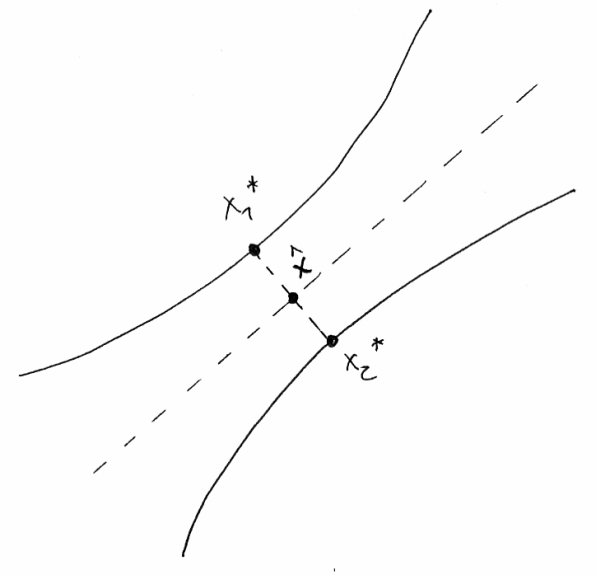


Abbildung 15.9: Illustration der Lage der Punkte x_1^*, x_2^* und $\widehat{x}$ im Beweis von Satz 15.32.

Quizfrage: Wie kann man die „Lücke“ $\inf\{a^\mathsf{T}x_2 \mid x_2 \in C_2\} - \sup\{a^\mathsf{T}x_1 \mid x_1 \in C_1\}$ interpretieren?

Bemerkung 15.33 (Der strikte Trennungssatz und das Farkas-Lemma). *Das Farkas-Lemma 8.6 ist eine spezielle Version des strikten Trennungssatzes 15.32, und zwar für den Fall, dass C_1 ein abgeschlossener, konvexer* Kegel *(siehe Lemma 6.10)*

$$C_1 = \{B^\mathsf{T}\xi \mid \xi \in \mathbb{R}^m,\ \xi \geq 0\}$$

ist und C_2 die kompakte einpunktige *Menge $C_2 = \{c\}$.*

§ 16 Das Subdifferential und die Richtungsableitung konvexer Funktionen

Literatur: Geiger, Kanzow, 2002, Kapitel 6.3, Rockafellar, 1970

Ziel: Verallgemeinerung der Ableitung für nicht-glatte konvexe Funktionen

§ 16.1 Das Subdifferential

Definition 16.1 (Subdifferential).
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ eine konvexe Funktion.

(*i*) *Ein Vektor $s \in \mathbb{R}^n$ heißt ein (Euklidischer)* ***Subgradient*** *von f im Punkt $x_0 \in \mathbb{R}^n$, wenn die* ***Subgradientenungleichung*** *gilt:*

$$f(x) \geq f(x_0) + s^\mathsf{T}(x - x_0) \quad \text{für alle } x \in \mathbb{R}^n. \tag{16.1}$$

Man sagt: Die rechte Seite in (16.1) ist eine ***lineare Minorante*** *(mit Euklidischem Gradienten s), die f in x_0* ***stützt****, kurz: eine* ***lineare Stützfunktion****.*[4]

(*ii*) *Die Menge $\partial f(x_0)$* aller *Subgradienten im Punkt x_0 heißt das* ***Subdifferential*** *von f in x_0.*

(*iii*) *f heißt* ***subdifferenzierbar*** *(kurz:* ***subdiffbar****) im Punkt $x_0 \in \mathbb{R}^n$, wenn $\partial f(x_0) \neq \emptyset$ ist.*

Beachte: (16.1) verallgemeinert die Ungleichung (13.14), die für konvexe *diffbare* Funktionen gilt.

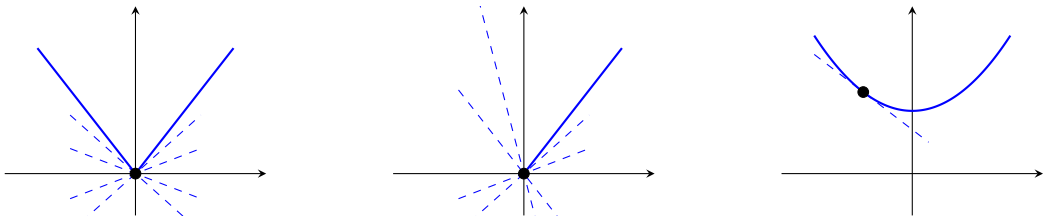
Beachte: Wenn $0 \in \partial f(x_0)$ gilt, dann ist x_0 ein globaler Minimierer von f.

Beachte: Wenn f eigentlich ist, also nicht $f \equiv \infty$, dann gilt $\partial f(x_0) = \emptyset$ in allen Punkten $x_0 \notin \operatorname{dom} f$.
Quizfrage: Warum?

[4] Eine andere Anschauung: $\binom{s}{-1}$ ist Normalenvektor einer Hyperebene durch den Punkt $(x_0, f(x_0))^\mathsf{T}$, die den (Epi-)Graphen in diesem Punkt stützt.

Bemerkung 16.2 (Zum Begriff „Subgradient“). *Die Bezeichnung* ***Subgradient*** *ist eigentlich irreführend, weil das Subdifferential die Ableitung ersetzt und nicht den Gradienten (der ja nichts anderes als eine vom verwendeten Innenprodukt abhängige Darstellung der Ableitung ist). Konzeptionell besser wäre es daher, einen Vektor $s \in \mathbb{R}_n$ (Zeilenvektor) eine* ***Subableitung*** *von f im Punkt $x_0 \in \mathbb{R}^n$ zu nennen, wenn $f(x) \geq f(x_0) + s\,(x - x_0)$ für alle $x \in \mathbb{R}^n$ gilt, und das Subdifferential aus allen Subableitungen bestehen zu lassen. Leider ist diese Bezeichnungsweise überhaupt nicht verbreitet. Wir folgen daher der allgemein üblichen Bezeichnung und sprechen von Subgradienten (hier stets im Sinne des Euklidischen Skalarprodukts).*

Quizfrage: Wie würde die Definition eines Subgradienten von f im Punkt x_0 bzgl. des M-Skalarprodukts aussehen?



(a) $f(x) = |x|$, $\partial f(0) = [-1, 1]$

(b) $f(x) = |x| + \delta_{[0,\infty)}$, $\partial f(0) = (-\infty, 1]$

(c) f an der Stelle x_0 diffbar, $\partial f(x_0) = \{\nabla f(x_0)\}$

Abbildung 16.1: Das Subdifferential $\partial f(x_0)$ besteht aus den *Steigungen* aller linearen Minoranten an f im Punkt x_0.

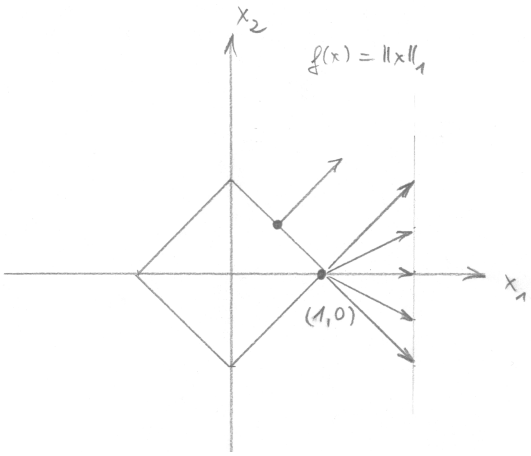


Abbildung 16.2: Das Subdifferential der 1-Norm $\|x\|_1 = |x_1| + |x_2|$ (Beispiel 16.3) in zwei Punkten derselben Levelmenge.

Beispiel 16.3 (Beispiele zum Subdifferential).

Wir betrachten als Beispiel für f verschiedene Normen auf $\mathbb{R}^n$.

(*i*) $f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$. *Dann ist* $s \in \partial f(x)$ *genau dann, wenn gilt:*

$$s_i \in \begin{cases} \{-1\} & \text{falls } x_i < 0, \\ [-1,1] & \text{falls } x_i = 0, \\ \{1\} & \text{falls } x_i > 0, \end{cases}$$

für alle $i = 1, \dots, n$.

(*ii*) $f(x) = \|x\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$. *Dann gilt*

$$\partial f(x) = \begin{cases} \left\{\frac{x}{\|x\|_2}\right\}, & \text{falls } x \neq 0, \\ \{s \in \mathbb{R}^n \mid \|s\|_2 \leq 1\}, & \text{falls } x = 0. \end{cases}$$

(*iii*) $f(x) = \|x\|_\infty = \max_{i=1,\dots,n} |x_i|$. *Dann gilt für* $x \neq 0$

$$\partial f(x) = \left\{ s \in \mathbb{R}^n \;\middle|\; \|s\|_1 \leq 1,\ \operatorname{sgn} s_i = \operatorname{sgn} x_i \text{ für diejenigen } i \text{ mit } |x_i| = \|x\|_\infty \text{ und } s_i = 0 \text{ für diejenigen } i \text{ mit } |x_i| < \|x\|_\infty \right\}$$

sowie

$$\partial f(0) = \{s \in \mathbb{R}^n \mid \|s\|_1 \leq 1\}.$$

Das folgende Beispiel zeigt, dass das Subdifferential eigentlicher konvexer Funktionen auch in Punkten, die zu dom f gehören, leer sein kann.

Beispiel 16.4 (Leeres Subdifferential)**.**

(a) Es sei

$$f(x) := \begin{cases} -\sqrt{1-x^2} & \text{für } x \in [-1,1], \\ \infty & \text{sonst.} \end{cases}$$

Dann ist f *eine eigentliche, konvexe, unterhalbstetige Funktion, aber* $\partial f(1) = \partial f(-1) = \emptyset$.

(b) Es sei

$$f(x) := \begin{cases} 1 & \text{für } x = 0, \\ x & \text{für } x > 0, \\ \infty & \text{für } x < 0. \end{cases}$$

Dann ist f *eine eigentliche, konvexe (aber nicht unterhalbstetige) Funktion, und es gilt* $\partial f(0) = \emptyset$.

Dieses Beispiel deutet schon darauf hin, dass es die Punkte des relativen Randes von dom f sind, in denen das Subdifferential leer sein kann. In der Tat gilt folgender Satz:

Satz 16.5 (Wann ist das Subdifferentials leer bzw. nichtleer?)**.**
Es sei $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *konvex und eigentlich.*

(*i*) *Für alle* $x_0 \notin \operatorname{dom} f$ *ist* $\partial f(x_0) = \emptyset$.

(*ii*) *Für alle* $x_0 \in \operatorname{rel\,int}(\operatorname{dom} f)$ *ist* $\partial f(x_0) \neq \emptyset$.

Beweis. Aussage (*i*): Es sei $x_0 \notin \operatorname{dom} f$, also $f(x_0) = \infty$. Die Subgradientenungleichung (16.1), also

$$f(x) \geq \underbrace{f(x_0)}_{=\infty} + s^\mathsf{T}(x - x_0)$$

kann mit $x \in \operatorname{dom} f$ für kein $s \in \mathbb{R}^n$ erfüllt sein.

Aussage (*ii*): Betrachte den Epigraphen

$$\operatorname{epi} f = \{(x, \gamma) \in \mathbb{R}^n \times \mathbb{R} \,|\, \gamma \geq f(x)\}.$$

Nach Satz 13.15 ist $\operatorname{epi} f \subseteq \mathbb{R}^n \times \mathbb{R}$ konvex. Wir wenden den eigentlichen Trennungssatz 15.30 an, um die Mengen $C_1 = \{(x_0, f(x_0))\}$ und $C_2 = \operatorname{epi} f$ eigentlich zu trennen. Beachte, dass $\operatorname{rel\,int}(C_1) \cap \operatorname{rel\,int}(C_2) = \emptyset$ ist, da der Punkt $(x_0, f(x_0))$ nicht zu $\operatorname{rel\,int}(\operatorname{epi} f)$ gehört (**Quizfrage:** Begründung?). Es existiert also eine Hyperebene $H(a, \beta)$ mit Normalenvektor $a = -(s, \sigma) \in \mathbb{R}^n \times \mathbb{R}$, $(s, \sigma) \neq 0$, die den Punkt $(x_0, f(x_0))$ und $\operatorname{epi} f$ eigentlich trennt, also gilt

$$\begin{pmatrix} s \\ \sigma \end{pmatrix}^\mathsf{T} \begin{pmatrix} x \\ \gamma \end{pmatrix} \leq \begin{pmatrix} s \\ \sigma \end{pmatrix}^\mathsf{T} \begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix} \quad \text{für alle } (x, \gamma) \in \operatorname{epi} f, \tag{16.2}$$

und es existiert ein Punkt $(\overline{x}, \overline{\gamma}) \in \operatorname{epi} f$ mit der Eigenschaft

$$\begin{pmatrix} s \\ \sigma \end{pmatrix}^\mathsf{T} \begin{pmatrix} \overline{x} \\ \overline{\gamma} \end{pmatrix} < \begin{pmatrix} s \\ \sigma \end{pmatrix}^\mathsf{T} \begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix}. \tag{16.3}$$

Wir zeigen $\sigma \leq 0$ und unterscheiden drei Fälle:

(*i*) Ist $f(x_0) = 0$, so folgt aus (16.2) mit der Wahl $x = x_0$ und $\gamma = 1$:

$$s^\mathsf{T} x_0 + \sigma \leq s^\mathsf{T} x_0 + 0.$$

(*ii*) Ist $f(x_0) > 0$, so folgt aus (16.2) mit der Wahl $x = x_0$ und $\gamma = 2\, f(x_0)$:

$$s^\mathsf{T} x_0 + 2\,\sigma\, f(x_0) \leq s^\mathsf{T} x_0 + \sigma\, f(x_0).$$

(*iii*) Ist $f(x_0) < 0$, so folgt aus (16.2) mit der Wahl $x = x_0$ und $\gamma = 0$:

$$s^\mathsf{T} x_0 + 0 \leq s^\mathsf{T} x_0 + \sigma\, f(x_0).$$

In allen drei Fällen folgt aus der jeweiligen Ungleichung $\sigma \leq 0$.

Wir zeigen jetzt, dass sogar $\sigma < 0$ gilt, indem wir die Annahme $\sigma = 0$ zum Widerspruch führen. Aus $\sigma = 0$ folgt mit (16.3) $s^\mathsf{T}(\overline{x} - x_0) < 0$. Da $x_0 \in \operatorname{rel\,int}(\operatorname{dom} f)$ war, existiert nach Lemma 15.17 ein $\varepsilon > 0$, sodass auch $\tilde{x} := x_0 - \varepsilon\,(\overline{x} - x_0)$ noch zu $\operatorname{dom} f$ gehört. Durch Einsetzen von $(\tilde{x}, f(\tilde{x})) \in \operatorname{epi} f$ in (16.2) erhalten wir den Widerspruch

$$s^\mathsf{T} x_0 - \varepsilon \underbrace{s^\mathsf{T}(\overline{x} - x_0)}_{<0} \leq s^\mathsf{T} x_0.$$

Durch positive Skalierung des Normalenvektors $a = -(s, \sigma)$ können wir nun o. B. d. A. $\sigma = -1$ annehmen. Dann ergibt sich aus (16.2) mit der Wahl $\gamma = f(x)$ die Folgerung

$$s^\mathsf{T} x - f(x) \leq s^\mathsf{T} x_0 - f(x_0) \quad \text{für alle } x \in \operatorname{dom} f.$$

Da dieselbe Ungleichung trivialerweise auch für $x \notin \operatorname{dom} f$ gilt, erhalten wir schließlich

$$f(x) \geq f(x_0) + s^\mathsf{T}(x - x_0) \quad \text{für alle } x \in \mathbb{R}^n,$$

d. h., $s \in \partial f(x_0)$. □

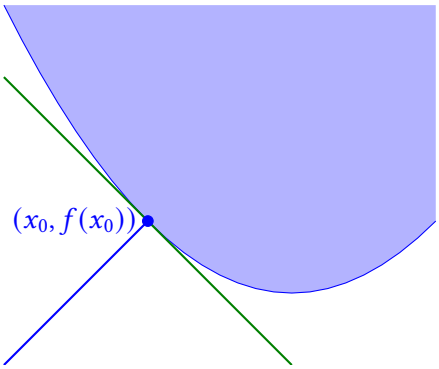


Abbildung 16.3: Konstruktion eines Subgradienten aus dem Beweis von Satz 16.5 mit Hilfe des Normalenvektors einer Hyperebene, die den Punkt $(x_0, f(x_0))$ eigentlich von $\operatorname{epi} f$ trennt.

Satz 16.6 (Elementare Eigenschaften des Subdifferentials). *Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ eine konvexe Funktion und $x_0 \in \mathbb{R}^n$. Dann ist $\partial f(x_0)$ abgeschlossen und konvex.*

Beweis. □

Ein wichtiger Satz über das Subdifferential ist die folgende Summenregel, deren Beweis wiederum den eigentlichen Trennungssatz 15.30 verwendet.

Satz 16.7 (Summenregel für das Subdifferential, vgl. Rockafellar, 1970, Theorem 23.8). *Es seien $f_1, f_2\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ zwei konvexe Funktionen. Dann gilt*

$$\partial f_1(x_0) + \partial f_2(x_0) \subseteq \partial(f_1 + f_2)(x_0) \quad \textit{für alle } x_0 \in \mathbb{R}^n. \tag{16.4}$$

Falls

$$(\operatorname{rel\,int}\operatorname{dom} f_1) \cap (\operatorname{rel\,int}\operatorname{dom} f_2) \neq \emptyset \tag{16.5}$$

erfüllt ist, dann gilt sogar

$$\partial f_1(x_0) + \partial f_2(x_0) = \partial(f_1 + f_2)(x_0) \quad \textit{für alle } x_0 \in \mathbb{R}^n. \tag{16.6}$$

Beweis. Es sei $x_0 \in \mathbb{R}^n$ beliebig, aber fest. Falls $\partial f_1(x_0) = \emptyset$ oder $\partial f_2(x_0) = \emptyset$ ist, dann ist die linke Seite in (16.4) die leere Menge und nichts zu zeigen. Es sei also $s \in \partial f_1(x_0) + \partial f_2(x_0)$, d. h., $s = s_1 + s_2$ mit $s_1 \in \partial f_1(x_0)$ und $s_2 \in \partial f_2(x_0)$. Die Subdifferentialungleichung liefert

$$\begin{aligned} f_1(x) &\geq f_1(x_0) + s_1^\mathsf{T}(x - x_0), \\ f_2(x) &\geq f_2(x_0) + s_2^\mathsf{T}(x - x_0) \end{aligned}$$

für alle $x \in \mathbb{R}^n$. Die Addition der Ungleichungen ergibt

$$(f_1 + f_2)(x) \geq (f_1 + f_2)(x_0) + (s_1 + s_2)^\mathsf{T}(x - x_0),$$

für alle $x \in \mathbb{R}^n$, d. h., wir haben $s = s_1 + s_2 \in \partial(f_1 + f_2)(x_0)$.

Wir nehmen jetzt an, dass $(\operatorname{rel\,int}\operatorname{dom} f_1) \cap (\operatorname{rel\,int}\operatorname{dom} f_2) \neq \emptyset$ und zeigen die umgekehrte Inklusion. Es sei dazu weiterhin $x_0 \in \mathbb{R}^n$ beliebig, aber fest. Wir können von $\partial(f_1 + f_2)(x_0) \neq \emptyset$ ausgehen, sonst ist wegen (16.4) nichts zu zeigen. Es sei also $s \in \partial(f_1 + f_2)(x_0)$. Wir müssen zeigen, dass s zerlegt werden kann in $s = s_1 + s_2$ mit $s_1 \in \partial f_1(x_0)$ und $s_2 \in \partial f_2(x_0)$.

Wir setzen

$$\begin{aligned} g_1(x) &:= f_1(x + x_0) - f_1(x_0) - s^\mathsf{T}x \\ g_2(x) &:= f_2(x + x_0) - f_2(x_0). \end{aligned}$$

Diese Funktionen erfüllen noch immer die Voraussetzung $(\operatorname{rel\,int}\operatorname{dom} g_1) \cap (\operatorname{rel\,int}\operatorname{dom} g_2) \neq \emptyset$, und $s \in \partial(f_1 + f_2)(x_0)$ wird zu $0 \in \partial(g_1 + g_2)(\tilde{x}_0)$ mit $\tilde{x}_0 := 0$. (**Quizfrage:** Details?) Wir müssen jetzt zeigen: Es gibt ein $\tilde{s} \in \partial g_1(0)$, sodass $-\tilde{s} \in \partial g_2(0)$ ist.

Wir betrachten die konvexen Mengen[5]

$$\begin{aligned} C_1 &:= \{(x, \gamma) \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \geq g_1(x)\} = \operatorname{epi} g_1, \\ C_2 &:= \{(x, \gamma) \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \leq -g_2(x)\} = \operatorname{hypo} -g_2. \end{aligned}$$

Quizfrage: Warum sind beide dieser Mengen nichtleer? Lemma 15.18 zeigt, dass

$$\begin{aligned} \operatorname{rel\,int} C_1 &= \{(x, \gamma) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \operatorname{rel\,int}\operatorname{dom} g_1,\ \gamma > g_1(x)\}, \\ \operatorname{rel\,int} C_2 &= \{(x, \gamma) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \operatorname{rel\,int}\operatorname{dom} g_2,\ \gamma < -g_2(x)\}. \end{aligned}$$

Wegen $0 \in \partial(g_1 + g_2)(\tilde{x}_0)$ ist $\tilde{x}_0 = 0$ ein globaler Minimierer für $g_1 + g_2$ mit $(g_1 + g_2)(0) = 0$. Daraus folgt, dass $\operatorname{rel\,int} C_1 \cap \operatorname{rel\,int} C_2 = \emptyset$ ist. (**Quizfrage:** Details?)

[5] Der **Hypograph** einer Funktion $f := \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ ist die Menge

$$\operatorname{hypo} f := \{(x, \gamma) \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \leq f(x)\}.$$

Der Hypograph ist konvex genau dann, wenn f konkav ist.

Nach dem eigentlichen Trennungssatz 15.30 lassen sich C_1 und C_2 durch eine Hyperebene $H(a, \beta)$ eigentlich trennen. Wir schreiben den Normalenvektor als $a = (\tilde{s}, \sigma) \in \mathbb{R}^n \times \mathbb{R}$. Es gilt also

$$\begin{pmatrix} \tilde{s} \\ \sigma \end{pmatrix}^\mathsf{T} \begin{pmatrix} x_1 \\ \gamma_1 \end{pmatrix} \leq \begin{pmatrix} \tilde{s} \\ \sigma \end{pmatrix}^\mathsf{T} \begin{pmatrix} x_2 \\ \gamma_2 \end{pmatrix} \quad \text{für alle } (x_1, \gamma_1) \in C_1 \text{ und } (x_2, \gamma_2) \in C_2, \tag{16.7}$$

und es existieren $(\overline{x}_1, \overline{\gamma}_1) \in C_1$ und $(\overline{x}_2, \overline{\gamma}_2) \in C_2$, für die die Ungleichung strikt ist. Die Komponente σ kann nicht gleich Null sein, denn sonst hätten wir C_1 und C_2 durch eine Hyperebene mit Normalenvektor $\tilde{s}$ strikt getrennt, was aufgrund der Voraussetzung $(\operatorname{rel\,int} \operatorname{dom} g_1) \cap (\operatorname{rel\,int} \operatorname{dom} g_2) \neq \emptyset$ und Satz 15.30 aber nicht möglich ist.

Wegen $(g_1 + g_2)(0) = 0$ liegt $0 \in C_1 \cap C_2$. Aus (16.7) folgt somit $\sigma \leq 0$. Da $\sigma \neq 0$ ist, können wir a so skalieren, dass $\sigma = -1$ wird. Außerdem gilt wegen $0 \in C_1 \cap C_2$, dass die eigentlich trennende Hyperebene $H(a, \beta)$ den Offset $\beta = 0$ hat. Es folgt nun aus (16.7)

$$\begin{pmatrix} \tilde{s} \\ -1 \end{pmatrix}^\mathsf{T} \begin{pmatrix} x_1 \\ \gamma_1 \end{pmatrix} \leq 0 \leq \begin{pmatrix} \tilde{s} \\ -1 \end{pmatrix}^\mathsf{T} \begin{pmatrix} x_2 \\ \gamma_2 \end{pmatrix} \quad \text{für alle } (x_1, \gamma_1) \in C_1 \text{ und } (x_2, \gamma_2) \in C_2,$$

also

$$\begin{aligned} \tilde{s}^\mathsf{T} x &\leq \gamma \quad \text{für alle } (x, \gamma) \in C_1, \\ \tilde{s}^\mathsf{T} x &\geq \gamma \quad \text{für alle } (x, \gamma) \in C_2. \end{aligned}$$

Wegen $g_1(0) = g_2(0) = 0$ und der Definition von C_1 und C_2 heißt das aber

$$\begin{aligned} g_1(x) &\geq g_1(0) + \tilde{s}^\mathsf{T}(x - 0) \quad \text{für alle } x \in \mathbb{R}^n, \\ g_2(x) &\geq g_2(0) + (-\tilde{s})^\mathsf{T}(x - 0) \quad \text{für alle } x \in \mathbb{R}^n. \end{aligned}$$

Also gilt $\tilde{s} \in \partial g_1(\tilde{x}_0)$ und $-\tilde{s} \in \partial g_2(\tilde{x}_0)$, was zu zeigen war. □

Wie das folgende Beispiel zeigt, ist die Gleichheit der Subdifferentiale (16.6) ohne eine Regularitätsbedingung wie $(\operatorname{rel\,int} \operatorname{dom} f_1) \cap (\operatorname{rel\,int} \operatorname{dom} f_2) \neq \emptyset$ im Allgemeinen falsch.

Beispiel 16.8 (Phelps, 1993, Bemerkung nach Theorem 3.16). *Es seien $f_1 := \delta_{C_1}$ und $f_2 := \delta_{C_2}$ zwei Indikatorfunktionen auf $\mathbb{R}^2$, und zwar für die konvexen Mengen*

$$\begin{aligned} C_1 &:= \{(x, \gamma) \mid \gamma \geq x^2\} = \operatorname{epi}(x \mapsto x^2) \\ C_2 &:= \{(x_1, x_2) \mid x_2 = 0\}. \end{aligned}$$

Dann ist $\partial f_1(0,0) = \{(s_1, s_2) \mid s_1 = 0,\ s_2 \leq 0\}$ und $\partial f_2(0,0) = \{(s_1, s_2) \mid s_1 = 0\}$, jedoch $\partial(f_1 + f_2)(0,0) = \mathbb{R}^2$. Die Regularitätsbedingung (16.5) ist nicht erfüllt, denn es gilt $\operatorname{dom} f_1 \cap \operatorname{dom} f_2 = C_1 \cap C_2 = \{(0,0)\}$, und dieser Punkt ist kein relativ innerer Punkt von C_1.

Wir schließen diesen Abschnitt mit der Kettenregel, die wir hier ohne Beweis angeben.

Satz 16.9 (Kettenregel für das Subdifferential, vgl. Rockafellar, 1970, Theorem 23.9). *Es sei $f \colon \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ konvex und $A \in \mathbb{R}^{n \times m}$. Weiter sei $g \colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ definiert durch $g(x) := f(Ax)$. Dann gilt:*

$$A^\mathsf{T} \partial f(Ax_0) \subseteq \partial g(x_0) \quad \textit{für alle } x_0 \in \mathbb{R}^n. \tag{16.8}$$

Falls

$$(\operatorname{Bild} A) \cap (\operatorname{rel\,int} \operatorname{dom} f) \neq \emptyset \tag{16.9}$$

erfüllt ist, dann gilt sogar

$$A^{\mathsf{T}} \partial f(A\,x_0) = \partial g(x_0) \quad \textit{für alle } x_0 \in \mathbb{R}^n. \tag{16.10}$$

Weitere Eigenschaften des Subdifferentials folgen in § 16.3.

Ende der Woche 11

§ 16.2 Die Richtungsableitung

Frage: Gibt es einen Zusammenhang des Subdifferentials mit der Richtungsableitung?

Definition 16.10 ((Einseitige) Richtungsableitung)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ eine (nicht notwendigerweise konvexe) Funktion und $x_0 \in \operatorname{dom} f$. Weiter sei $d \in \mathbb{R}^n$. Dann heißt der Grenzwert (sofern er in $\mathbb{R} \cup \{\pm\infty\}$ existiert)

$$f'(x_0; d) := \lim_{t \searrow 0} \frac{f(x_0 + t\,d) - f(x_0)}{t} \tag{16.11}$$

die ***(einseitige) Richtungsableitung*** *der Funktion f im Punkt x_0 in Richtung d.*

Beispiel 16.11 (Beispiele zur Richtungsableitung)**.**
Wir betrachten als Beispiel für f wie in Beispiel 16.3 verschiedene Normen auf $\mathbb{R}^n$.

(*i*) $f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$. *Dann gilt*

$$f'(x; d) = \sum_{\substack{i=1 \\ x_i > 0}}^n d_i - \sum_{\substack{i=1 \\ x_i < 0}}^n d_i + \sum_{\substack{i=1 \\ x_i = 0}}^n |d_i|.$$

(*ii*) $f(x) = \|x\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$. *Dann gilt*

$$f'(x; d) = \begin{cases} \frac{x^{\mathsf{T}} d}{\|x\|_2}, & \textit{falls } x \neq 0, \\ \|d\|_2, & \textit{falls } x = 0. \end{cases}$$

(*iii*) $f(x) = \|x\|_\infty = \max_{i=1,\dots,n} |x_i|$. *Dann gilt für $x \neq 0$*

$$f'(x; d) = \max\left\{ (\operatorname{sgn} x_i)\, d_i \,\middle|\, i = 1, \dots, n,\ |x_i| = \|x\|_\infty \right\}$$

sowie

$$f'(0; d) = \|d\|_\infty.$$

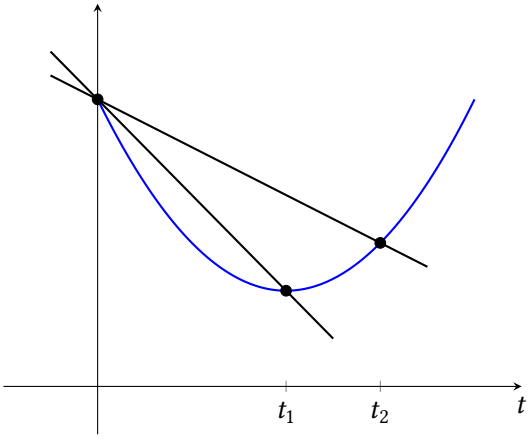


Abbildung 16.4: Illustration der Monotonie des Differenzenquotienten q (Lemma 16.12).

Zur weiteren Untersuchung der Richtungsableitung führen wir die Abkürzung

$$q(t) := \frac{f(x_0 + t\,d) - f(x_0)}{t}, \quad t > 0 \tag{16.12}$$

für den Differenzenquotienten ein.

Lemma 16.12 (Monotonie des Differenzenquotienten)**.**
Es seien $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex, $x_0 \in \operatorname{dom} f$ und $d \in \mathbb{R}^n$. Dann ist $q(t)$ monoton wachsend auf $\mathbb{R}_{\geq 0}$.

Beweis. Es sei $0 \leq t_1 < t_2$. Aus

$$x_0 + t_1\,d = \tfrac{t_1}{t_2}(x_0 + t_2\,d) + (1 - \tfrac{t_1}{t_2})\,x_0$$

und der Konvexität von f folgt

$$f(x_0 + t_1\,d) = f\Big(\frac{t_1}{t_2}(x_0 + t_2\,d) + \big(1 - \frac{t_1}{t_2}\big)\,x_0\Big) \leq \frac{t_1}{t_2} f(x_0 + t_2\,d) + \big(1 - \frac{t_1}{t_2}\big) f(x_0).$$

Daraus folgt

$$f(x_0 + t_1\,d) - f(x_0) \leq \frac{t_1}{t_2} f(x_0 + t_2\,d) - \frac{t_1}{t_2} f(x_0)$$

und damit $q(t_1) \leq q(t_2)$. □

Quizfrage: An welcher Stelle im Beweis geht die Voraussetzung $x_0 \in \operatorname{dom} f$ ein?

Im folgenden Satz erlauben wir ausnahmsweise in Erweiterung von Definition 13.12, dass eine konvexe Funktion $g\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ auch den Wert $-\infty$ annehmen kann. Das macht keine Schwierigkeiten, wenn man *definiert*, dass eine solche Funktion konvex ist, wenn ihr Epigraph eine konvexe Menge ist.
Beachte: Es gilt

$$g(x) = \inf\{\gamma \in \mathbb{R} \mid (x, \gamma) \in \operatorname{epi} g\} \tag{16.13}$$

unabhängig davon, ob $g(x) = -\infty$, $g(x) \in \mathbb{R}$ oder $g(x) = \infty$ ist. Die bisherigen Resultate aus Satz 16.5, Satz 16.6 und Lemma 16.12 gelten weiter.

Satz 16.13 (Existenz und elementare Eigenschaften der Richtungsableitung)**.**
Es seien $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und $x_0 \in \operatorname{dom} f$. Dann existiert die Richtungsableitung $f'(x_0; d)$ (mit Werten in $\mathbb{R} \cup \{\pm\infty\}$) für alle Richtungen $d \in \mathbb{R}^n$. Es gilt

$$f(x) \geq f(x_0) + f'(x_0; x - x_0) \quad \text{für alle } x \in \mathbb{R}^n. \tag{16.14}$$

Die Richtungsableitung hat folgende Eigenschaften:

(*i*) *Die Funktion*

$$\mathbb{R}^n \ni d \mapsto f'(x_0; d) \in \mathbb{R} \cup \{\pm\infty\} \tag{16.15}$$

ist positiv homogen und konvex.

(*ii*) *Folglich ist die Funktion (16.15) subadditiv:*

$$f'(x_0; d_1 + d_2) \leq f'(x_0; d_1) + f'(x_0; d_2) \quad \text{für alle } d_1, d_2 \in \mathbb{R}^n. \tag{16.16}$$

Den auf der rechten Seite möglicherweise vorkommenden Fall $\infty + (-\infty)$ oder $(-\infty) + \infty$ interpretieren wir als ∞.

(*iii*) *Es gilt $f'(x_0; 0) = 0$.*

(*iv*) *Es gilt*

$$-f'(x_0; -d) \leq f'(x_0; d) \quad \text{für alle } d \in \mathbb{R}^n.$$

(*v*) *Es sei U der Richtungsraum von* aff dom f *und $x_0 \in$* rel int dom f*. Dann ist $f'(x_0; d) \in \mathbb{R}$ für alle $d \in U$ und $f'(x_0; d) = \infty$ für alle $d \notin U$.*

Beweis. Aus der Monotonie von q (Lemma 16.12) folgt

$$f'(x_0; d) = \lim_{t \searrow 0} q(t) = \inf_{t > 0} q(t) \in \mathbb{R} \cup \{\pm\infty\}.$$

Mit $x \in \mathbb{R}^n$ beliebig und $d := x - x_0$ folgt außerdem aus der Monotonie von q

$$f'(x_0; d) = \lim_{t \searrow 0} q(t) \leq q(1) = \frac{f(x_0 + 1\, d) - f(x_0)}{1} = f(x) - f(x_0)$$

und damit die Ungleichung (16.14).

Zu Aussage (*i*): Für $\alpha > 0$ und $d \in \mathbb{R}^n$ gilt

$$f'(x_0; \alpha\, d) = \lim_{t \searrow 0} \frac{f(x_0 + t\, \alpha\, d) - f(x_0)}{t} = \alpha \lim_{t \searrow 0} \frac{f(x_0 + t\, \alpha\, d) - f(x_0)}{\alpha\, t} = \alpha\, f'(x_0; d),$$

d. h., $d \mapsto f'(x_0; d)$ ist positiv homogen. Um die Konvexität zu zeigen, seien $d_1, d_2 \in \mathbb{R}^n$ und $\alpha \in [0, 1]$. Für $t > 0$ folgt

$$\begin{aligned} &f\big(x_0 + t\,(\alpha\, d_1 + (1 - \alpha)\, d_2)\big) - f(x_0) \\ &\quad = f\big(\alpha\,(x_0 + t\, d_1) + (1 - \alpha)\,(x_0 + t\, d_2)\big) - f(x_0) \\ &\quad \leq \alpha\,\big[f(x_0 + t\, d_1) - f(x_0)\big] + (1 - \alpha)\big[f(x_0 + t\, d_2) - f(x_0)\big] \end{aligned}$$

aus der Konvexität von f. Die Division durch $t > 0$ und der (monotone) Grenzübergang $\lim_{t \searrow 0}$ auf der linken Seite zeigen weiter

$$\begin{aligned} & f'(x_0; \alpha\, d_1 + (1-\alpha)\, d_2) \\ & \quad \leq \frac{f\big(x_0 + t\,(\alpha\, d_1 + (1-\alpha)\, d_2)\big) - f(x_0)}{t} \\ & \quad \leq \alpha\, \frac{f(x_0 + t\, d_1) - f(x_0)}{t} + (1-\alpha)\, \frac{f(x_0 + t\, d_2) - f(x_0)}{t} \end{aligned}$$

für beliebige $t > 0$. **Beachte:** Wir können den Grenzübergang auf der rechten Seite nicht vollziehen, da dann der undefinierte Ausdruck $\infty - \infty$ entstehen könnte. Wir können aber wie folgt argumentieren: Es seien (d_1, γ_1) und (d_2, γ_2) in $\operatorname{epi} f'(x_0; \cdot)$. Dann gibt es für jedes $\varepsilon > 0$ ein $\delta > 0$, sodass

$$\frac{f(x_0 + t\, d_1) - f(x_0)}{t} \leq \gamma_1 + \varepsilon \quad \text{und} \quad \frac{f(x_0 + t\, d_2) - f(x_0)}{t} \leq \gamma_2 + \varepsilon$$

für alle $t \in (0, \delta)$ gilt. Für diese t haben wir also

$$f'(x_0; \alpha\, d_1 + (1-\alpha)\, d_2) \leq \alpha\, \gamma_1 + (1-\alpha)\, \gamma_2 + \varepsilon,$$

und da $\varepsilon > 0$ beliebig war, gilt dieselbe Ungleichung auch mit $\varepsilon = 0$. Das heißt aber $\alpha\, (d_1, \gamma_1) + (1-\alpha)\, (d_2, \gamma_2) \in \operatorname{epi} f'(x_0; \cdot)$, was zu zeigen war.

Zu Aussage *(ii)*: Es seien $d_1, d_2 \in \mathbb{R}^n$. Wir haben eben mit Aussage *(i)* insbesondere gezeigt, dass

$$\left\{ \frac{\gamma_1 + \gamma_2}{2} \,\middle|\, (d_1, \gamma_1), (d_2, \gamma_2) \in \operatorname{epi} f'(x_0; \cdot) \right\} \subseteq \left\{ \gamma \,\middle|\, \left(\frac{d_1 + d_2}{2}, \gamma \right) \in \operatorname{epi} f'(x_0; \cdot) \right\} \tag{16.17}$$

gilt. Es folgt:

$$\begin{aligned} & f'(x_0; d_1 + d_2) \\ & \quad = 2\, f'(x_0; \tfrac{1}{2}(d_1 + d_2)) && \text{positive Homogenität} \\ & \quad = 2\, \inf \left\{ \gamma \,\middle|\, \left(\frac{d_1 + d_2}{2}, \gamma \right) \in \operatorname{epi} f'(x_0; \cdot) \right\} && \text{wegen (16.13)} \\ & \quad \leq 2\, \inf \left\{ \frac{\gamma_1 + \gamma_2}{2} \,\middle|\, (d_1, \gamma_1), (d_2, \gamma_2) \in \operatorname{epi} f'(x_0; \cdot) \right\} && \text{wegen (16.17)} \\ & \quad = \inf \{ \gamma_1 \,|\, (d_1, \gamma_1) \in \operatorname{epi} f'(x_0; \cdot) \} + \inf \{ \gamma_2 \,|\, (d_2, \gamma_2) \in \operatorname{epi} f'(x_0; \cdot) \} && \text{siehe unten} \\ & \quad = f'(x_0; d_1) + f'(x_0; d_2) && \text{wegen (16.13).} \end{aligned}$$

Wenn kein (d_1, γ_1) in $\operatorname{epi} f'(x_0; \cdot)$ liegt, also $f'(x_0; d_1) = \infty$ ist, dann ist die linke Menge in (16.17) leer, also das Infimum darüber ∞. Damit die nachfolgende Gleichheit auch dann stimmt, wenn $f'(x_0; d_2) = -\infty$ ist, muss hier $\infty + (-\infty)$ als ∞ interpretiert werden. Dasselbe gilt natürlich bei Vertauschen der Rollen von d_1 und d_2.

Zu Aussage *(iii)*: Für $d = 0$ gilt

$$f'(x_0; d) = \lim_{t \searrow 0} \frac{f(x_0 + t\, 0) - f(x_0)}{t} = 0.$$

Zu Aussage (*iv*): Aus Aussage (*iii*) und Aussage (*ii*) folgt

$$0 = f'(x_0; d - d) \leq f'(x_0; d) + f'(x_0; -d).$$

Wenn beide Ausdrücke $f'(x_0; d)$ und $f'(x_0; -d)$ endlich sind, dann folgt die Behauptung $-f'(x_0; -d) \leq f'(x_0; d)$ unmittelbar. Wenn einer oder beide der Ausdrücke $f'(x_0; d)$ und $f'(x_0; -d)$ gleich ∞ sind, dann ist die Aussage $-f'(x_0; -d) \leq f'(x_0; d)$ ebenfalls klar. Der Fall, dass einer oder beide der Ausdrücke $-f'(x_0; -d) \leq f'(x_0; d)$ gleich $-\infty$ sind, kann wegen der obigen Ungleichung nicht vorkommen.

Zu Aussage (*v*): Es sei $x_0 \in \operatorname{rel\,int} \operatorname{dom} f$ und $d \in U$, dem Richtungsraum von $\operatorname{aff} \operatorname{dom} f$. Dann ist $x_0 \pm \varepsilon\, d \in \operatorname{dom} f$ für hinreichend kleines $\varepsilon > 0$, siehe Lemma 15.17. Folglich sind

$$\frac{f(x_0 + t\, d) - f(x_0)}{t} \quad \text{und} \quad \frac{f(x_0 - t\, d) - f(x_0)}{t}$$

für hinreichend kleine $t > 0$ endlich, und aufgrund der Monotonie des Differenzenquotienten folgt $f'(x_0; d) < \infty$ und $f'(x_0; -d) < \infty$. Zusammen mit Aussage (*iv*) ergibt sich also

$$-\infty < -f'(x_0; -d) \leq f'(x_0; d) < \infty,$$

d. h. $f'(x_0; d)$ ist endlich.

Andererseits gehört, falls $d \notin U$ liegt, $x_0 + t\, d$ für *alle* $t > 0$ nicht zu $\operatorname{aff} \operatorname{dom} f$ und damit auch nicht zu $\operatorname{dom} f$. In diesem Fall ist also $f(x_0 + t\, d) = \infty$ für alle $t > 0$ und damit $q(t) \equiv \infty$, folglich $f'(x_0; d) = \infty$. □

Folgerung 16.14 (Endlichkeit der Richtungsableitung)**.** *Ist $x_0 \in \operatorname{int} \operatorname{dom} f$, dann ist die Richtungsableitung $f'(x_0; d)$ für alle $d \in \mathbb{R}^n$ endlich.*

Beweis. Da $\operatorname{int} \operatorname{dom} f$ nichtleer ist, ist $\dim \operatorname{dom} f = n$ und damit der Richtungsraum $U = \mathbb{R}^n$. Die Behauptung folgt nun aus Satz 16.13 (*v*). □

§ 16.3 Zusammenhang zwischen Subdifferential und Richtungsableitung

Das wesentliche Resultat, in dem sich Subdifferential und Richtungsableitung einer konvexen Funktion gegenseitig charakterisieren, ist der folgende Satz.

Satz 16.15 (Zusammenhang zwischen Subdifferential und Richtungsableitung, vgl. Rockafellar, 1970, Theorem 23.2 und 23.4)**.**
Es sei $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex.

(*i*) *Für jedes $x_0 \in \mathbb{R}^n$ gilt:*

$$\partial f(x_0) = \left\{ s \in \mathbb{R}^n \,\middle|\, s^\mathsf{T} d \leq f'(x_0; d) \text{ für alle } d \in \mathbb{R}^n \right\}. \tag{16.18}$$

(Möglicherweise sind beide Mengen leer.)

(*ii*) *Für jedes* $x_0 \in \operatorname{rel\,int} \operatorname{dom} f$ *und* $d \in \mathbb{R}^n$ *gilt:*

$$f'(x_0; d) = \sup\left\{s^\mathsf{T} d \,\middle|\, s \in \partial f(x_0)\right\} \in \mathbb{R} \cup \{\infty\}. \tag{16.19}$$

Genauer: Es sei U *der Richtungsraum von* $\operatorname{aff} \operatorname{dom} f$*. Ist* $d \in U$*, dann sind beide Seiten in* (16.19) *endlich, und das Supremum ist ein Maximum. Ist* $d \notin U$*, dann sind beide Seiten in* (16.19) *gleich* ∞*.*

Die Aussage (16.19) ist in der Literatur auch als **max formula** bekannt.

Beweis. Aussage (*i*): Es sei zunächst $s \in \partial f(x_0)$. Dann ist notwendigerweise $f(x_0)$ endlich (Satz 16.5). Es sei $d \in \mathbb{R}^n$ beliebig, aber fest. Aus der Subgradientenungleichung (16.1) folgt

$$f(x_0 + t\,d) \geq f(x_0) + s^\mathsf{T}(t\,d)$$

für alle $t > 0$. Sortieren der Terme und Division durch t ergibt

$$\frac{f(x_0 + t\,d) - f(x_0)}{t} \geq s^\mathsf{T} d,$$

und durch Grenzübergang folgt $f'(x_0; d) \geq s^\mathsf{T} d$ für $d \in \mathbb{R}^n$ beliebig.

Nun sei $s \in \mathbb{R}^n$ ein Vektor mit der Eigenschaft $s^\mathsf{T} d \leq f'(x_0; d)$ für alle $d \in \mathbb{R}^n$. Weiter sei $x \in \mathbb{R}^n$ beliebig und $d := x - x_0$. Aus der Monotonie des Differenzenquotienten (Lemma 16.12) folgt insbesondere

$$s^\mathsf{T}(x - x_0) = s^\mathsf{T} d \leq f'(x_0; d) = \lim_{t \searrow 0} q(t) \leq q(1) = \frac{f(x_0 + 1\,d) - f(x_0)}{1},$$

d. h., es gilt

$$f(x) \geq f(x_0) + s^\mathsf{T}(x - x_0).$$

Da $x \in \mathbb{R}^n$ beliebig war, folgt $s \in \partial f(x_0)$.

Aussage (*ii*): Es sei $x_0 \in \operatorname{rel\,int} \operatorname{dom} f$. Dann ist $\partial f(x_0) \neq \emptyset$ nach Satz 16.5. Es sei $d \in \mathbb{R}^n$ beliebig, aber fest. Die Richtungsableitung $f'(x_0; d)$ existiert (mit Werten in $\mathbb{R} \cup \{\pm\infty\}$) nach Satz 16.13. Nach Aussage (*i*) gilt $f'(x_0; d) \geq s^\mathsf{T} d$ für alle $s \in \partial f(x_0)$, also auch

$$f'(x_0; d) \geq \sup\left\{s^\mathsf{T} d \,\middle|\, s \in \partial f(x_0)\right\}. \tag{16.20}$$

Die rechte Seite kann nicht gleich $-\infty$ sein. (**Quizfrage:** Warum?)

Wir müssen zeigen, dass in (16.20) Gleichheit gilt. Es sei dazu U der Richtungsraum von $\operatorname{aff} \operatorname{dom} f$. Wir unterscheiden zwei Fälle. Falls $d \notin U$ liegt, dann ist $f'(x_0; d) = \infty$ nach Satz 16.13 (*v*). Wir müssen zeigen, dass dann auch die rechte Seite in (16.20) gleich ∞ ist. Wir können d eindeutig darstellen als $d = d_1 + d_2$ mit $d_1 \in U$ und $d_2 \in U^\perp$. Nach Annahme ist $d_2 \neq 0$. Es sei $\bar{s}$ irgendein festes Element von $\partial f(x_0)$; es gilt nach Aussage (*i*) also $\bar{s}^\mathsf{T} y \leq f'(x_0; y)$ für alle $y \in \mathbb{R}^n$. Wir zeigen, dass $\bar{s} + \alpha\, d_2$ für alle $\alpha \in \mathbb{R}$ in $\partial f(x_0)$ liegt. Dazu zeigen wir:

$$(\bar{s} + \alpha\, d_2)^\mathsf{T} y \leq f'(x_0; y) \tag{16.21}$$

für alle $y \in \mathbb{R}^n$. Falls $y \notin U$ liegt, dann ist die rechte Seite in (16.21) gleich ∞, die linke Seite aber endlich. Falls dagegen $y \in U$ liegt, dann ist $d_2^\mathsf{T} y = 0$, also gilt (16.21) auch in diesem Fall. Für die rechte Seite in (16.20) erhalten wir nun

$$\sup\{s^\mathsf{T} d \,|\, s \in \partial f(x_0)\} \geq \sup\{(\bar{s} + \alpha\, d_2)^\mathsf{T} d \,|\, \alpha \in \mathbb{R}\} = \bar{s}^\mathsf{T} d + \sup\{\alpha \,\|d_2\|^2 \,|\, \alpha \in \mathbb{R}\} = \infty.$$

Damit ist in diesem Fall die Gleichheit in (16.20) gezeigt.

Im anderen Fall ist $d \in U$. Wir definieren die Funktion $g(d) := f'(x_0; d)$. Diese ist nach Satz 16.13 *(i)* konvex mit Werten in $\mathbb{R} \cup \{\infty\}$, da der Wert $-\infty$ oben bereits ausgeschlossen wurde. Ihr eigentlicher Definitionsbereich ist $\operatorname{dom} g = U$, siehe Satz 16.13 *(v)*. Da U ein Unterraum von $\mathbb{R}^n$ ist, gilt $\operatorname{rel\,int} U = U$. Nach Satz 16.5 *(ii)* existiert also ein Element $\bar{s} \in \partial g(d)$, d. h., es gilt

$$f'(x_0; y) \geq f'(x_0; d) + \bar{s}^\mathsf{T}(y - d) \quad \text{für alle } y \in \mathbb{R}^n. \tag{16.22}$$

Setzen wir speziell $y = 0$ ein, so folgt

$$0 \geq f'(x_0; d) - \bar{s}^\mathsf{T} d.$$

Setzen wir andererseits $y = 2\,d$ ein, so ergibt sich

$$f'(x_0; 2\,d) = 2\, f'(x_0; d) \geq f'(x_0; d) + \bar{s}^\mathsf{T}(2d - d),$$

d. h. $f'(x_0; d) \geq \bar{s}^\mathsf{T} d$. Durch beide Ungleichungen zusammen erhalten wir also $f'(x_0; d) = \bar{s}^\mathsf{T} d$. Es bleibt zu zeigen, dass $\bar{s}$ zu $\partial f(x_0)$ gehört. Aus (16.22) folgt nun aber $f'(x_0; y) \geq f'(x_0; d) + \bar{s}^\mathsf{T}(y - d) = \bar{s}^\mathsf{T} y$ für alle $y \in \mathbb{R}^n$. Nach Aussage *(i)* gilt damit tatsächlich $\bar{s} \in \partial f(x_0)$, und wir haben gezeigt:

$$f'(x_0; d) \geq \sup\{s^\mathsf{T} d \,|\, s \in \partial f(x_0)\} \geq \bar{s}^\mathsf{T} d = f'(x_0; d).$$

Insbesondere ist das Supremum ein Maximum. □

Folgerung 16.16 (Unbeschränktheit des Subdifferentials)**.** *Es sei $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex. Weiter sei $x_0 \in \operatorname{dom} f$ und $\partial f(x_0) \neq \emptyset$. Dann gehört mit jedem $s \in \partial f(x_0)$ auch $s + U^\perp$ zu $\partial f(x_0)$, wobei U der Richtungsraum von* $\operatorname{aff} \operatorname{dom} f$ *und $U^\perp$ sein orthogonales Komplement ist. Wenn also* $\dim \operatorname{dom} f < n$ *ist, dann ist $\partial f(x_0)$ unbeschränkt.*

Quizfrage: Wie kann man sich diese Tatsache vorstellen?

Beweis. Es sei $s \in \partial f(x_0)$ und $\overline{d} \in U^\perp$. Nach Satz 16.15 gilt $s^\mathsf{T} d \leq f'(x_0; d)$ für alle $d \in \mathbb{R}^n$. Wir zeigen, dass diese Ungleichung auch für $s + \overline{d}$ gilt. Es sei dazu $d \in \mathbb{R}^n$ beliebig. Wir können d eindeutig darstellen als $d = d_1 + d_2$ mit $d_1 \in U$ und $d_2 \in U^\perp$. Falls $d_2 \neq 0$ gilt, dann ist $f'(x_0; d) = \infty$, und damit ist

$$(s + \overline{d})^\mathsf{T} d \leq f'(x_0; d)$$

gezeigt. Andernfalls ist $d_2 = 0$, also $d \in U$, und daher gilt

$$(s + \overline{d})^\mathsf{T} d = s^\mathsf{T} d \leq f'(x_0; d)$$

nach Voraussetzung. □

§ 16.4 Weitere Eigenschaften konvexer Funktionen

Lemma 16.17 (Lokale Beschränktheit nach oben impliziert lokale Lipschitz-Stetigkeit)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und $x_0 \in \operatorname{dom} f$. Wenn f auf einer Kugel $B_r(x_0)$ nach oben beschränkt ist, dann gilt:

(*i*) *f ist auf $B_r(x_0)$ auch nach unten beschränkt.*

(*ii*) *f ist auf $B_{r/2}(x_0)$ Lipschitz-stetig.*

Beweis. Aussage (*i*): Es gelte $f(x) \leq \widetilde{M}$ auf der Kugel $B_r(x_0)$ für ein $\widetilde{M} \in \mathbb{R}$. Es sei $x \in B_r(x_0)$ ein beliebiges Element und $y \coloneqq 2\,x_0 - x$. Dann ist y ebenfalls in $B_r(x_0)$, und es gilt

$$f(x_0) = f\Big(\frac{x+y}{2}\Big) \leq \frac{1}{2} f(x) + \frac{1}{2} f(y).$$

Durch Umstellen ergibt sich $f(x) \geq 2\,f(x_0) - f(y) \geq 2\,f(x_0) - \widetilde{M}$.

Aussage (*ii*): Aufgrund von Aussage (*i*) ist

$$M \coloneqq \sup\{|f(x)| \,|\, x \in B_r(x_0)\}$$

endlich. Es seien $x, y \in B_{r/2}(x_0)$ und $x \neq y$. Wir definieren $z \coloneqq x + \frac{r}{2}\,\frac{(x-y)}{\|x-y\|}$. Dann ist $z \in x + \frac{r}{2} B_1(0) \subseteq x_0 + r\,B_1(0) = B_r(x_0)$. Wir setzen $\alpha \coloneqq \frac{r}{2\,\|x-y\|}$.[6] Also ist $z = x + \alpha\,(x-y)$ und daher

$$x = \frac{1}{\alpha+1} z + \frac{\alpha}{\alpha+1} y.$$

Die Konvexität von f ergibt

$$f(x) \leq \frac{1}{\alpha+1} f(z) + \frac{\alpha}{\alpha+1} f(y)$$

und damit

$$f(x) - f(y) \leq \frac{1}{\alpha+1}\big[f(z) - f(y)\big] \leq \frac{2\,M}{\alpha+1} = 2\,M\,\frac{2\,\|x-y\|}{r + 2\,\|x-y\|} \leq \frac{4\,M}{r}\,\|x-y\|.$$

Durch Vertauschen den Rollen von x und y erhalten wir ganz analog $f(y) - f(x) \leq \frac{4\,M}{r}\,\|x-y\|$, also zusammen

$$|f(x) - f(y)| \leq \frac{4\,M}{r}\,\|x-y\| \tag{16.23}$$

für alle $x, y \in B_{r/2}(x_0)$. □

Beachte: Es sieht hier zunächst so aus, als würde die Lipschitz-Konstante in (16.23) für $r \searrow 0$ „explodieren“. Jedoch gilt natürlich eine Abschätzung (16.23) mit derselben Lipschitz-Konstanten auch auf allen Kugeln um x_0 mit kleinerem Radius.

[6] Es gilt $\alpha \in (0, 2/3)$, wobei $\alpha \approx 0$ ist für $y \approx x$ und $\alpha \approx 2/3$, falls $\|y - x\| \approx \delta$.

Es stellt sich die Frage, in welchen Punkten $x_0 \in \operatorname{dom} f$ die Voraussetzung von Lemma 16.17 gilt, also dass f in einer Umgebung von x_0 nach oben beschränkt (und damit bereits in einer kleineren Umgebung von x_0 Lipschitz-stetig) ist. Dazu muss natürlich notwendigerweise $x_0 \in \operatorname{int} \operatorname{dom} f$ sein. Das ist aber auch bereits hinreichend, wie das folgende Resultat zeigt.

Satz 16.18 (Stetigkeit konvexer Funktionen). *Es sei $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex. Folgende Aussagen sind für $x_0 \in \mathbb{R}^n$ äquivalent:*

(*i*) $x_0 \in \operatorname{int} \operatorname{dom} f$.

(*ii*) *f ist in einer Umgebung von x_0 nach oben beschränkt.*

(*iii*) *f ist in einer Umgebung von x_0 Lipschitz-stetig.*

(*iv*) *f ist im Punkt x_0 stetig.*

Beachte: Konvexe Funktionen sind also genau im (möglicherweise leeren) Inneren ihres eigentliches Definitionsbereiches **lokal beschränkt** und **lokal Lipschitz-stetig**.

Beweis. Aussage (*ii*) $\Rightarrow$ Aussage (*iii*): Das wurde in Lemma 16.17 gezeigt.

Aussage (*iii*) $\Rightarrow$ Aussage (*iv*): Das ist offensichtlich.

Aussage (*iv*) $\Rightarrow$ Aussage (*i*): Das folgt sofort aus der ε-δ-Definition der Stetigkeit.

Aussage (*i*) $\Rightarrow$ Aussage (*ii*): Es sei $x_0 \in \operatorname{int} \operatorname{dom} f$. Wir können affin unabhängige Vektoren $v_0, \dots, v_n$ und $r > 0$ so wählen, dass mit $\Delta := \operatorname{conv}(\{v_0, \dots, v_n\})$ gilt: $B_r(x_0) \subseteq \Delta \subseteq \operatorname{dom} f$. (Zur Konstruktion siehe nachfolgendes Lemma 16.19.) Jeder Punkt $x \in B_r(x_0) \subseteq \Delta$ hat dann eine eindeutige Darstellung als Konvexkombination

$$x = \sum_{j=0}^{n} \alpha_j v_j$$

mit Koeffizienten $\alpha_j \geq 0$ und $\sum_{j=0}^{n} \alpha_j = 1$. Aus der Jensenschen Ungleichung für die konvexe Funktion f folgt nun

$$f(x) = f\Big(\sum_{j=0}^{n} \alpha_j v_j\Big) \leq \sum_{j=0}^{n} \alpha_j f(v_j) \leq \underbrace{\max_{j=0,\dots,n} f(v_j)}_{=:M} \sum_{j=0}^{n} \alpha_j = M.$$

für alle $x \in \Delta$ und insbesondere für alle $x \in B_r(x_0)$. □

Lemma 16.19 (Zwischen eine Menge und einen inneren Punkt passt immer ein Simplex[7]).
Es sei $M \subseteq \mathbb{R}^n$ eine Menge und $x_0 \in \operatorname{int} M$, sodass $B_R(x_0) \subseteq M$ liegt. Dann existieren affin unabhängige Punkte $v_0, \dots, v_n \in M$ und ein $r > 0$, sodass mit der konvexen Hülle $\Delta := \operatorname{conv}(\{v_0, \dots, v_n\})$ gilt:

$$B_r(x_0) \subseteq \Delta \subseteq B_R(x_0) \subseteq M.$$

[7] Ein **Simplex** (Plural: **Simplizes**) im $\mathbb{R}^n$ ist die konvexe Hülle affin unabhängiger Punkte im $\mathbb{R}^n$.

Quizfrage: Wie kann diese Aussage veranschaulicht werden?

Beweis. Wir gehen zur Vereinfachung der Notation von $x_0 = 0$ und $R = 1$ aus, was wir durch Ersetzen von M durch $(M - x_0)/R$ immer erreichen können.

Die gesuchten Punkte können wir nun beispielsweise wählen als

$$v_0 := -\frac{1}{n+1}\mathbf{1} \quad \text{und} \quad v_j := e_j - \frac{1}{n+1}\mathbf{1}, \quad j = 1, \dots, n.$$

Hierbei ist e_j der j-te Einheitsvektor im $\mathbb{R}^n$. Die affine Unabhängigkeit der v_j sowie die Eigenschaft $\|v_j\| < 1$ für $j = 0, \dots, n$ prüft man leicht nach. Da $B_1(0)$ konvex ist, folgt bereits

$$\Delta \subseteq B_1(0) \subseteq M.$$

Aufgrund der affinen Unabhängigkeit lässt sich jedes $x \in \mathbb{R}^n$ eindeutig als Affinkombination $x = \sum_{j=0}^n \alpha_j v_j$ schreiben mit Koeffizienten, die $\sum_{j=0}^n \alpha_j = 1$ erfüllen. Der Ursprung ist gerade der Mittelpunkt von $v_0, \dots, v_n$, hat also den Koeffizientenvektor $\overline{\alpha} = 1/(n+1)\,\mathbf{1}$, denn es gilt

$$\sum_{j=0}^n \overline{\alpha}_j v_j = \frac{1}{n+1}\sum_{j=0}^n v_j = \frac{1}{n+1}\left[-\frac{n+1}{n+1}\mathbf{1} + \sum_{j=1}^n e_j\right] = \frac{1}{n+1}\left[-\frac{n+1}{n+1}\mathbf{1} + \mathbf{1}\right] = 0.$$

Insbesondere ist der Ursprung also eine echte Konvexkombination der Punkte $v_0, \dots, v_n$. Aufgrund der stetigen (linearen) Abhängigkeit der Koeffizienten α vom Punkt x gilt für eine ganze Umgebung $B_r(0)$ mit geeignetem Radius $r > 0$ die Eigenschaft $\alpha \geq 0$, d. h., $B_r(0) \subseteq \Delta$, was zu zeigen war.

Genauer: Wir können wie im Beweis von Satz 15.15 vorgehen. Das lineare Gleichungssystem, das die Koeffizienten der Darstellung eines beliebigen Punktes x als Affinkombination der Punkte $v_0, \dots, v_n$ ermittelt, vgl. (15.8), lautet

$$\underbrace{\begin{bmatrix} 1 & \cdots & 1 \\ | & & | \\ v_0 & \cdots & v_n \\ | & & | \end{bmatrix}}_{=:B} \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ | \\ x \\ | \end{pmatrix}}_{=:b}.$$

Durch die Wahl $\varrho := 1/\big((n+1)\|B^{-1}\|_\infty\big)$ erreichen wir zunächst $\overline{B_\varrho^{\|\cdot\|_\infty}(\overline{x})} \subseteq \Delta$. (**Quizfrage:** Warum müssen wir hier, anders als im Beweis von Satz 15.15, nicht $\|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty$ nehmen?) Die Äquivalenz der Normen $\|\cdot\|_\infty$ und $\|\cdot\|_2$ im $\mathbb{R}^n$ zeigt schließlich, dass wir $r := \varrho\sqrt{n}$ wählen können, sodass sogar für die abgeschlossene r-Kugel gilt:

$$\overline{B_r(0)} \subseteq \Delta \subseteq B_1(0) \subseteq M. \qquad \square$$

Aus dem Satz 16.18 folgen weitere Konsequenzen für das Subdifferential und die Richtungsableitung konvexer Funktionen. Wir wissen bereits aus Satz 16.13 (*v*), dass in Punkten $x_0 \in \operatorname{int}\operatorname{dom} f$ gilt, dass $f'(x_0; d)$ für alle Richtungen $d \in \mathbb{R}^n$ endlich ist. Es gilt jedoch mehr:

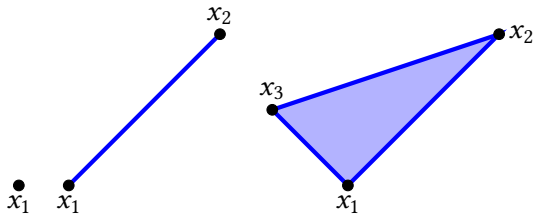


Abbildung 16.5: Simplizes der Dimensionen 0, 1 und 2 im $\mathbb{R}^2$.

Lemma 16.20 (Lipschitz-Stetigkeit der Richtungsableitung). *Es seien $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und $x_0 \in \operatorname{int}\operatorname{dom} f$. Weiter sei L eine Lipschitz-Konstante von f in einer Kugel $B_r(x_0)$, siehe Satz 16.18. Dann gilt*

$$|f'(x_0; d_1) - f'(x_0; d_2)| \le L \, \|d_1 - d_2\| \quad \text{für alle } d_1, d_2 \in \mathbb{R}^n \tag{16.24}$$

und insbesondere $|f'(x_0; d)| \le L \, \|d\|$ für alle $d \in \mathbb{R}^n$.

Beweis. Es sei $d \in \mathbb{R}^n$ beliebig. Für hinreichend kleines $t > 0$ gilt

$$|f(x_0 + t\,d) - f(x_0)| \le L\, t \, \|d\|.$$

Die Division durch t und der Grenzübergang $t \searrow 0$ zeigt $|f'(x_0; d)| \le L \, \|d\|$.

Um die Lipschitz-Stetigkeit zu zeigen, schätzen wir ab:

$$\begin{aligned}
&f'(x_0; d_1) - f'(x_0; d_2) \\
&\quad = f'(x_0; d_1) - f'(x_0; d_1 + (d_2 - d_1)) \\
&\quad \ge f'(x_0; d_1) - f'(x_0; d_1) - f'(x_0; d_2 - d_1) \quad \text{wegen der Subadditivität, siehe (16.16)} \\
&\quad \ge -L \, \|d_1 - d_2\|.
\end{aligned}$$

Analog gilt auch

$$\begin{aligned}
&f'(x_0; d_1) - f'(x_0; d_2) \\
&\quad = f'(x_0; d_2 + (d_1 - d_2)) - f'(x_0; d_2) \\
&\quad \le f'(x_0; d_2) + f'(x_0; d_1 - d_2) - f'(x_0; d_2) \quad \text{wegen der Subadditivität, siehe (16.16)} \\
&\quad \le L \, \|d_1 - d_2\|.
\end{aligned}$$

Das zeigt die Behauptung (16.24). □

Wir wissen bereits aus Satz 16.5 und Satz 16.6, dass das Subdifferential einer eigentlichen, konvexen Funktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ in Punkten $x_0 \in \operatorname{rel\,int}\operatorname{dom} f$ nichtleer, abgeschlossen und konvex ist. Im folgenden Satz wird das Subdifferential noch genauer charakterisiert.

Satz 16.21 (Kompaktheit des Subdifferentials). *Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex. Es sei weiter $x_0 \in \operatorname{int}\operatorname{dom} f$ und L eine Lipschitz-Konstante von f in einer Kugel $B_r(x_0)$, siehe Satz 16.18. Dann ist $\partial f(x_0)$ kompakt, und es gilt $\|s\| \le L$ für alle $s \in \partial f(x_0)$.*

Beweis. Für $s \in \partial f(x_0)$ gilt nach Satz 16.15 und Lemma 16.20:

$$s^\mathsf{T} d \leq f'(x_0; d) \leq L \|d\| \quad \text{für alle } d \in \mathbb{R}^n.$$

Mit der Wahl $d = s$ folgt $\|s\| \leq L$. □

Bemerkung 16.22 (Verallgemeinerung von Satz 16.21)**.** *Wenn $x_0 \in \operatorname{rel\,int} \operatorname{dom} f$ ist, dann gilt*

$$\partial f(x_0) = U^\perp + \partial(f|_{\operatorname{aff}\operatorname{dom} f})(x_0). \tag{16.25}$$

Hier ist U der Richtungsraum von $\operatorname{aff}\operatorname{dom} f$ und $U^\perp$ sein orthogonales Komplement. Der Ausdruck $\partial(f|_{\operatorname{aff}\operatorname{dom} f})(x_0)$ ist das Subdifferential der Einschränkung von f auf $\operatorname{aff}\operatorname{dom} f$. Für diese Funktion ist x_0 ein innerer Punkt, und man kann zeigen, dass $\partial(f|_{\operatorname{aff}\operatorname{dom} f})(x_0)$ kompakt ist.

Satz 16.23 (Beschränktheit des Subdifferentials, vgl. Geiger, Kanzow, 2002, Lemma 6.19)**.** *Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und $B \subseteq \operatorname{int}\operatorname{dom} f$ eine kompakte Menge. Dann ist die Bildmenge*

$$\partial f(B) = \bigcup_{x \in B} \partial f(x)$$

ebenfalls beschränkt.

Beweis. Wir betrachten die Überdeckung von B durch offene Kugeln

$$\bigcup_{x \in B} B_{r_x}(x),$$

wobei die Radien $r_x > 0$ so gewählt werden, dass $\overline{B_{r_x}(x)} \subseteq \operatorname{int}\operatorname{dom} f$ bleibt. Aufgrund der Kompaktheit von B sind endlich viele Kugeln ausreichend, sagen wir

$$B \subseteq \bigcup_{j=1,\dots,m} B_{r_j}(x_j) \subseteq \bigcup_{j=1,\dots,m} \overline{B_{r_j}(x_j)} \subseteq \operatorname{int}\operatorname{dom} f.$$

Nach Satz 16.18 ist f stetig auf $\operatorname{int}\operatorname{dom} f$ und damit auf jeder der endlich vielen kompakten Kugeln $\overline{B_{r_x}(x_j)}$ beschränkt. Somit ist f auch auf der Vereinigung $\bigcup_{j=1,\dots,m} \overline{B_{r_j}(x_j)}$ beschränkt.

Es sei nun $s \in \partial f(x_s)$ für ein $x_s \in B$. Es gilt also

$$f(x) \geq f(x_s) + s^\mathsf{T}(x - x_s) \quad \text{für alle } x \in \mathbb{R}^n.$$

Der Punkt x_s gehöre zur Kugel $B_{r_j}(x_j)$. Im Fall $s \neq 0$ setzen wir $x := x_s + \frac{r_j}{\|s\|} s \in \overline{B_{r_x}(x_j)}$ in die Subgradientenungleichung ein und erhalten

$$f(x) \geq f(x_s) + s^\mathsf{T} \frac{r_j}{\|s\|} s = f(x_s) + r_j \|s\|,$$

also

$$\|s\| \leq \frac{1}{r_j}\big[f(x) - f(x_s)\big].$$

Da x und x_s zu $\bigcup_{j=1,\dots,m} \overline{B_{r_j}(x_j)}$ gehören, wo f beschränkt ist, und $s \in \partial f(B)$ bis auf die Bedingung $s \neq 0$ beliebig war, folgt, dass $\partial f(B) \setminus \{0\}$ beschränkt ist und damit auch $\partial f(B)$. □

Satz 16.24 (Subdifferential einer konvexen *diffbaren* Funktion, vgl. Rockafellar, 1970, Theorem 25.1)**.** *Es sei* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *konvex und* $x_0 \in \operatorname{dom} f$.

(*i*) *Wenn* f *an der Stelle* x_0 *diffbar ist, dann ist* $\partial f(x_0) = \{\nabla f(x_0)\}$, *und es gilt* $x_0 \in \operatorname{int}\operatorname{dom} f$.

(*ii*) *Wenn das Subdifferential* $\partial f(x_0)$ *einelementig ist, dann gilt* $x_0 \in \operatorname{int}\operatorname{dom} f$, *und* f *ist an der Stelle* x_0 *diffbar.*

Beweis. Aussage (*i*): Es sei f an der Stelle x_0 diffbar. Dann sind die Richtungsableitungen gegeben durch $f'(x_0; d) = \nabla f(x_0)^\mathsf{T} d$. Gemäß Satz 16.15 ist $s \in \partial f(x_0)$ genau dann, wenn $s^\mathsf{T} d \le f'(x_0; d)$, also

$$s^\mathsf{T} d \le \nabla f(x_0)^\mathsf{T} d$$

für alle $d \in \mathbb{R}^n$ ist. Diese Bedingung ist genau für $s = \nabla f(x_0)$ erfüllt. (**Quizfrage:** Warum?) Weiter folgt aus der Diffbarkeit von f an der Stelle x_0 auch die Stetigkeit von f an dieser Stelle. Nach Satz 16.18 ist also $x_0 \in \operatorname{int}\operatorname{dom} f$.

Aussage (*ii*): Es sei $\partial f(x_0) = \{s\}$. Wir zeigen zuerst, dass dann $x_0 \in \operatorname{int}\operatorname{dom} f$ folgt. Aus Folgerung 16.16 ergibt sich zunächst, dass $\operatorname{dom} f$ notwendigerweise volle Dimension haben muss, da sonst $\partial f(x_0)$ unbeschränkt ist. Angenommen, $x_0 \notin \operatorname{int}\operatorname{dom} f$, dann gilt auch $x_0 \notin \operatorname{core}\operatorname{dom} f$ nach Lemma 15.23. Es gibt also eine Richtung $\overline{d} \in \mathbb{R}^n$, $\overline{d} \neq 0$, sodass $x_0 \in \operatorname{dom} f$ liegt, aber $x_0 + \varepsilon\,\overline{d} \notin \operatorname{dom} f$ für alle $\varepsilon > 0$. Daraus folgt $f'(x_0; \overline{d}) = \infty$.

Wir zeigen unter Verwendung von Satz 16.15, dass dann auch $s + \overline{d} \in \partial f(x_0)$ gilt, im Widerspruch zu Voraussetzung, dass $\partial f(x_0)$ einelementig ist. Es sei dazu $d \in \mathbb{R}^n$ beliebig. Wir können d eindeutig darstellen als $d = d_1 + d_2$ mit $d_1 \perp \overline{d}$ und $d_2 \in \operatorname{span}\{\overline{d}\}$, also $d = d_1 + \alpha\,\overline{d}$. Wir machen eine Fallunterscheidung nach dem Vorzeichen von α. Wenn $\alpha \ge 0$ ist, dann gilt

$$f'(x_0; d) \le f'(x_0; d_1) + \alpha f'(x_0; \overline{d}) = \infty$$

wegen Satz 16.13. Damit ist in diesem Fall $(s + \overline{d})^\mathsf{T} d \le f'(x_0; d)$ klar. Andernfalls ist $\alpha < 0$ und daher

$$(s + \overline{d})^\mathsf{T} d = s^\mathsf{T} d + \overline{d}^\mathsf{T}(d_1 + \alpha\overline{d}) = s^\mathsf{T} d + \alpha\,\|\overline{d}\|^2 < s^\mathsf{T} d \le f'(x_0; d)$$

nach Satz 16.15. Wir haben also gezeigt, dass

$$(s + \overline{d})^\mathsf{T} d \le f'(x_0; d) \quad \text{für alle } d \in \mathbb{R}^n$$

gilt. Aus Satz 16.15 folgt damit $s + \overline{d} \in \partial f(x_0)$, Widerspruch. Folglich gilt notwendigerweise $x_0 \in \operatorname{int}\operatorname{dom} f$.

Wir definieren die konvexe Funktion

$$g(\delta x) := f(x_0 + \delta x) - f(x_0) - s^\mathsf{T}\delta x.$$

Für diese gilt dann $g(\delta x) \ge 0$ (**Quizfrage:** Warum?) sowie $0 \in \operatorname{int}\operatorname{dom} g$ und $\partial g(0) = \{0\}$. Wir müssen zeigen:

$$\lim_{\delta x \to 0} \frac{g(\delta x)}{\|\delta x\|} = 0. \tag{16.26}$$

Es sei $R > 0$ so, dass $B_R(0) \subseteq \operatorname{dom} g$ liegt. Lemma 16.19 zeigt, dass es affin unabhängige Punkte $v_0, \dots, v_n$ und ein $r > 0$ gibt, sodass mit $\Delta := \operatorname{conv}(\{v_0, \dots, v_n\})$ gilt: $B_r(0) \subseteq \Delta \subseteq B_R(0) \subseteq \operatorname{dom} g$. Folglich gilt $\|v_j\| < R$ für alle $j = 0, \dots, n$.

Jedes Element in Δ und insbesonderes jedes $y \in B_r(0)$ hat eine (eindeutige) Darstellung als Konvexkombination

$$y = \sum_{j=0}^{n} \alpha_j v_j \tag{16.27}$$

mit Koeffizienten $\alpha_j \geq 0$ und $\sum_{j=0}^{n} \alpha_j = 1$.

Es sei nun $\delta x \in \mathbb{R}^n$ beliebig, aber fest, mit $\|\delta x\| < \frac{r^2}{R}$ und $\delta x \neq 0$. Daraus folgt $\frac{\|\delta x\|}{r} \|v_j\| < r$.

Wir schätzen nun ab:

$$\begin{aligned} g(\delta x) &= g\Big(\frac{\|\delta x\|}{r} r \frac{\delta x}{\|\delta x\|}\Big) \\ &= g\Big(\frac{\|\delta x\|}{r} \sum_{j=0}^{n} \alpha_j v_j\Big) \quad \text{wegen (16.27)} \\ &\leq \sum_{j=0}^{n} \alpha_j g\Big(\frac{\|\delta x\|}{r} v_j\Big) \quad \text{Jensensche Ungleichung} \\ &\leq \max_{j=0,\dots,n} g\Big(\frac{\|\delta x\|}{r} v_j\Big). \end{aligned} \tag{16.28}$$

Die Funktionswerte sind alle endlich, da, wie bereits gezeigt, $\frac{\|\delta x\|}{r} v_j \in B_r(0) \subseteq \operatorname{dom} g$ liegt.

Nach Satz 16.15 gilt wegen $\partial g(0) = \{0\}$:

$$g'(0; d) = \max\{s^\mathsf{T} d \mid s \in \partial g(0)\} = 0$$

für alle $d \in \mathbb{R}^n$. Zusammen mit der Monotonie des Differenzenquotienten erhalten wir

$$0 = g'(0; d) = \lim_{t \searrow 0} \frac{g(t\,d) - g(0)}{t} = \lim_{t \searrow 0} \frac{g(t\,d)}{t} \leq \frac{g(t\,d)}{t} \tag{16.29}$$

für alle $t > 0$ und alle $d \in \mathbb{R}^n$. Wir können nun (16.26) zeigen: Wegen (16.28) haben wir

$$0 \leq \frac{g(\delta x)}{\|\delta x\|} \leq \max_{j=0,\dots,n} \frac{g\Big(\frac{\|\delta x\|}{r} v_j\Big)}{\|\delta x\|},$$

und wegen (16.29) gilt

$$\lim_{\|\delta x\| \searrow 0} \frac{g\Big(\frac{\|\delta x\|}{r} v_j\Big)}{\|\delta x\|} = \lim_{t \searrow 0} \frac{g\big(\frac{t}{r} v_j\big)}{t} = 0$$

für alle $j = 0, \dots, n$. □

Ende der Woche 12

§ 17 Kegel

Literatur: Geiger, Kanzow, 2002, Kapitel 2.2.1

Definition 17.1 (Kegel).
*Eine Menge $K \subseteq \mathbb{R}^n$ heißt ein **Kegel** (englisch: cone), wenn $\beta x \in K$ gilt für $x \in K$ und alle $\beta > 0$. Kurz: $\beta K \subseteq K$ für alle $\beta > 0$. Der Kegel K heißt **spitz**, wenn $0 \in K$ ist, ansonsten **stumpf**.*

Beachte: Mit jedem Punkt $x \in \mathbb{R}^n$ enthält ein Kegel bereits die ganze Halbgerade $\{\beta x : \beta > 0\}$.

Beispiel 17.2 (Kegel).
Beispiele für Kegel sind:

(i) *offene Halbgerade $\{\beta a \mid \beta > 0\}$ mit $a \in \mathbb{R}^n$, $a \neq 0$*

(ii) *abgeschlossene Halbgerade $\{\beta a \mid \beta \geq 0\}$ mit $a \in \mathbb{R}^n$, $a \neq 0$*

(iii) *offener Orthant $\{x \in \mathbb{R}^n \mid x > 0\}$*

(iv) *abgeschlossener Orthant $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n \mid x \geq 0\}$*

(v) *der **Lorentzkegel** $K = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} \mid \|x\| \leq t\}$*

(vi) *die Menge der symmetrisch positiv semidefiniten Matrizen S^n_+ in $\mathbb{R}^{n \times n}$.*

Beachte: Kegel sind i. A. nicht konvex.

Satz 17.3 (Operationen auf Kegeln).

(i) *Es sei $\{K_j\}_{j \in J}$ eine beliebige Familie von Kegeln. Dann ist $\bigcap_{j \in J} K_j$ ein Kegel.*

(ii) *Es seien $K_i \subseteq \mathbb{R}^{n_i}$, $i = 1, \ldots, k$ Kegel. Dann ist das kartesische Produkt $K_1 \times \cdots \times K_k$ ein Kegel in $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$.*

(iii) *Es sei $f : \mathbb{R}^n \to \mathbb{R}^m$ eine lineare Abbildung und $K \subseteq \mathbb{R}^n$ und $\widehat{K} \subseteq \mathbb{R}^m$ Kegel. Dann sind das Bild $f(K) \subseteq \mathbb{R}^m$ und das Urbild $f^{-1}(\widehat{K}) \subseteq \mathbb{R}^n$ Kegel. **Quizfrage:** Gilt das auch, wenn f affin-linear ist?*

(iv) *Sind $K_1, K_2 \subseteq \mathbb{R}^n$ Kegel, dann sind*

$$\beta K_1 = \{\beta x_1 \mid x_1 \in K_1\} \quad \text{für } \beta \in \mathbb{R}$$

sowie die Minkowski-Summe

$$K_1 + K_2$$

Kegel.

(v) *Das Komplement* $K^c = \mathbb{R}^n \setminus K$ *eines Kegels* $K \subseteq \mathbb{R}^n$ *ist ein Kegel.*

(vi) *Ist* $K \subseteq \mathbb{R}^n$ *ein Kegel, dann sind das Innere* $\operatorname{int}(K)$*, der Abschluss* $\overline{K}$ *und der Rand* ∂K *wieder Kegel.*

Beweis. Übung □

Lemma 17.4 (Konvexe Kegel)**.**
Es sei $K \subseteq \mathbb{R}^n$*.*

(a) *Folgende Aussagen sind äquivalent:*

(i) K *ist ein konvexer Kegel.*

(ii) *Es gilt* $\alpha_1 x_1 + \alpha_2 x_2 \in K$ *für alle* $x_1, x_2 \in K$ *und* $\alpha_1, \alpha_2 > 0$*.*

(b) *Folgende Aussagen sind äquivalent:*

(i) K *ist ein spitzer konvexer Kegel.*

(ii) *Es gilt* $\alpha_1 x_1 + \alpha_2 x_2 \in K$ *für alle* $x_1, x_2 \in K$ *und* $\alpha_1, \alpha_2 \geq 0$*.*

Beweis. Aussage (a): Es sei zunächst $K \subset \mathbb{R}^n$ ein konvexer Kegel und $x_1, x_2 \in K$ sowie $\alpha_1, \alpha_2 > 0$. Dann sind $\alpha_1 x_1$ und $\alpha_2 x_2$ in K und

$$\alpha_1 x_1 + \alpha_2 x_2 = 2 \overbrace{\left(\underbrace{\tfrac{1}{2}\alpha_1 x_1}_{\in K} + \underbrace{\tfrac{1}{2}\alpha_2 x_2}_{\in K} \right)}^{\in K \text{ (Konvexität)}} \in K.$$

Umgekehrt seien $x_1, x_2 \in K$ und $\alpha \in (0,1)$. Nach Voraussetzung ist $\alpha x_1 + (1-\alpha) x_2 \in K$, also K konvex. Es sei weiter $x \in K$ und $\beta > 0$. Wähle $\alpha_1 = \alpha_2 = \beta/2$ und $x_1 = x_2 = x$. Nach Voraussetzung ist $\alpha x = \alpha x_1 + (1-\alpha) x_2 \in K$, also K ein Kegel.

Aussage (b): analog. □

Beachte: Der Rezessionskegel (6.9) eines Polyeders in Normalform sowie die konische Hülle (6.13) $\operatorname{cone}\{b_1, \ldots, b_n\}$ einer Menge von Vektoren sind konvexe Kegel, die zudem abgeschlossen sind.

Für den Rest dieses Abschnitts werden wir uns mit Kegeln beschäftigen, die in der konvexen Optimierung eine besondere Bedeutung haben.

§ 17.1 Radialkegel und Kegel zulässiger Richtungen

Definition 17.5 (Kegel zulässiger Richtungen). *Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge und $x \in M$. Dann heißt*

$$\mathcal{F}_M(x) = \{d \in \mathbb{R}^n \mid \text{es gibt ein } \varepsilon > 0, \text{ sodass } x + t\,d \in M \text{ liegt für alle } t \in [0, \varepsilon]\} \tag{17.1}$$

*der **Kegel der zulässigen Richtungen** (englisch: cone of feasible directions) der Menge M im Punkt x. Ein Vektor $d \in \mathcal{F}_M(x)$ heißt **zulässige Richtung** von M im Punkt x. Man definiert $\mathcal{F}_M(x) := \emptyset$ für $x \notin M$.*

Definition 17.6 (Radialkegel). *Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge und $x \in M$. Dann heißt*

$$\mathcal{K}_M(x) = \{\beta\,(y - x) \mid y \in M,\ \beta > 0\} = \bigcup_{\beta > 0} \beta\,(M - x) \tag{17.2}$$

*der **von $M - x$ erzeugte Kegel** oder der **Radialkegel** (englisch: radial cone) an die Menge M im Punkt x. Ein Vektor $d \in \mathcal{K}_M(x)$ heißt **radiale Richtung** von M im Punkt x. Man definiert $\mathcal{K}_M(x) := \emptyset$ für $x \notin M$.*

Quizfrage: Warum sind $\mathcal{F}_M(x)$ und $\mathcal{K}_M(x)$ Kegel?

Quizfrage: Was sind $\mathcal{F}_M(x)$ und $\mathcal{K}_M(x)$ für $x \in \operatorname{int}(M)$?

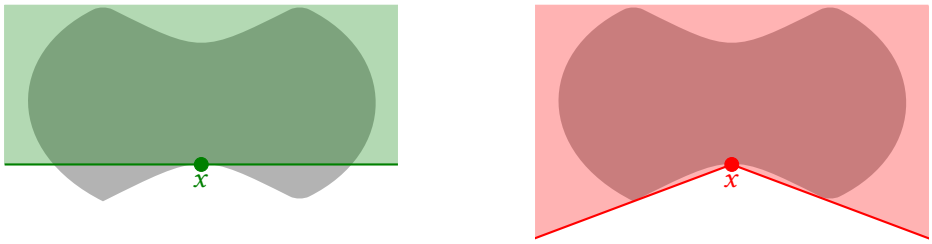


Abbildung 17.1: Kegel der zulässigen Richtungen $\mathcal{F}_M(x)$ (grün) und Radialkegel $\mathcal{K}_M(x)$ (rot) einer nichtkonvexen Menge M in einem Punkt x. Dargestellt ist jeweils der verschobene Kegel $x + \mathcal{F}_M(x)$ bzw. $x + \mathcal{K}_M(x)$.

Satz 17.7 (Eigenschaften von $\mathcal{F}_M(x)$, $\mathcal{K}_M(x)$ und deren Zusammenhang).
Es sei $M \subseteq \mathbb{R}^n$ beliebig und $x \in M$. Dann gilt:

(*i*) *$\mathcal{F}_M(x)$ und $\mathcal{K}_M(x)$ sind spitze Kegel.*

(*ii*) *$\mathcal{F}_M(x) \subseteq \mathcal{K}_M(x)$.*

(*iii*) *$M \subseteq x + \mathcal{K}_M(x)$.*

(*iv*) *Es sei $C \subseteq \mathbb{R}^n$ konvex und $x \in C$. Dann ist $\mathcal{K}_C(x)$ ein spitzer konvexer Kegel, und es gilt $\mathcal{F}_C(x) = \mathcal{K}_C(x)$.*

Beweis. Übung. □

Quizfrage: Beispiel für eine nichtkonvexe Menge M und einen Punkt $x \in M$, sodass dennoch $\mathcal{F}_M(x) = \mathcal{K}_M(x)$ gilt?

Lemma 17.8 (Richtungsableitung der Indikatorfunktion)**.** *Es sei $C \subseteq \mathbb{R}^n$ eine nichtleere konvexe Menge und $x \in C$. Dann gilt:*

$$\delta'_C(x; d) = \begin{cases} 0, & \text{falls } d \in \mathcal{F}_C(x) \text{ oder äquivalent: } d \in \mathcal{K}_C(x), \\ \infty & \text{sonst.} \end{cases} \tag{17.3}$$

Beweis. Der Differenzenquotient für $t > 0$

$$q(t) = \frac{\delta_C(x + t\,d) - \delta_C(x)}{t} = \frac{\delta_C(x + t\,d) - 0}{t}$$

wird entweder gleich null für ein $t_0 > 0$ (und dann wegen der Konvexität von C auch für alle $t \in [0, t_0]$, oder es ist $q(t) = \infty$ für alle $t > 0$. Der erste Fall ist genau der Fall $d \in \mathcal{F}_C(d)$, siehe (17.1). Die Gleichheit $\mathcal{F}_C(x) = \mathcal{K}_C(x)$ wurde in Satz 17.7 gezeigt. □

§ 17.2 Normalenkegel

Definition 17.9 (Normalenkegel)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge und $x \in M$. Dann heißt

$$\mathcal{N}_M(x) := \{s \in \mathbb{R}^n \mid s^\mathsf{T}(y - x) \leq 0 \text{ für alle } y \in M\} \tag{17.4}$$

*der **Normalenkegel** (englisch: normal cone) von M im Punkt x. Ein Vektor $s \in \mathcal{N}_M(x)$ heißt **Normalenrichtung** von M im Punkt x. Man definiert $\mathcal{N}_M(x) := \emptyset$ für $x \notin M$.*

Beachte: s ist genau dann eine Normalenrichtung von M in $x \in M$, wenn M enthalten ist im Halbraum $\{y \in \mathbb{R}^n \mid s^\mathsf{T} y \leq \beta := s^\mathsf{T} x\}$ mit Normalenvektor s. Mit anderen Worten: Die Normalenrichtungen im Punkt x sind (bis auf $s = 0$) gerade die Normalenvektoren von Hyperebenen, die den Punkt x von M trennen.

Quizfrage: Was ist $\mathcal{N}_M(x)$ für $x \in \operatorname{int}(M)$?

Lemma 17.10 (Eigenschaften des Normalenkegels)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge und $x \in M$. Dann gilt:

(*i*) *Der Normalenkegel $\mathcal{N}_M(x)$ ist ein konvexer abgeschlossener Kegel.*

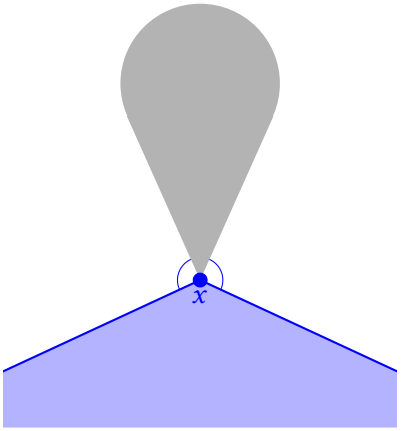


Abbildung 17.2: Normalenkegel $\mathcal{N}_M(x)$ (blau) einer konvexen Menge M in einem Punkt x. Dargestellt ist der verschobene Kegel $x + \mathcal{N}_M(x)$.

(*ii*) *Es gilt*

$$\mathcal{N}_M(x) = \mathcal{K}_M(x)^\circ := \{s \in \mathbb{R}^n \mid s^\mathsf{T} d \leq 0 \text{ für alle } d \in \mathcal{K}_M(x)\}.$$

Man sagt: Der Normalenkegel ist der ***Polarkegel***[8] *des Radialkegels.*

Beweis. Übung. □

Lemma 17.11 (Normalenkegel und Subdifferential der Indikatorfunktion)**.** *Es sei $C \subset \mathbb{R}^n$ eine nichtleere konvexe Menge. Dann gilt:*

$$\partial \delta_C(x) = \mathcal{N}_C(x) \tag{17.5}$$

für alle $x \in \mathbb{R}^n$, d. h., das Subdifferential der Indikatorfunktion einer konvexen Menge ist gerade der Normalenkegel von C im Punkt x.

Beweis. Falls $x \notin C$ ist, dann ist $\partial \delta_C(x) = \emptyset$ (**Quizfrage:** Warum nochmal?) und $\mathcal{N}_C(x) = \emptyset$ per Definition. Im Fall $x \in C$ ist $s \in \partial \delta_C(x)$ genau dann, wenn gilt:

$$\delta_C(y) \geq \underbrace{\delta_C(x)}_{=0} + s^\mathsf{T}(y - x) \quad \text{für alle } y \in \mathbb{R}^n.$$

Da diese Ungleichung für $y \notin C$ trivialerweise erfüllt ist, reicht es, sie für $y \in C$ zu fordern. Es ist also $s \in \partial \delta_C(x)$ genau dann, wenn

$$0 \geq 0 + s^\mathsf{T}(y - x) \quad \text{für alle } y \in C$$

gilt. Das ist aber gerade die Definition dafür, dass s zum Normalenkegel $\mathcal{N}_C(x)$ gehört, siehe (17.4). □

[8] Allgemein ist der **Polarkegel** einer beliebigen Menge $M \subseteq \mathbb{R}^n$ gegeben durch

$$M^\circ = \{s \in \mathbb{R}^n \mid s^\mathsf{T} y \leq 0 \text{ für alle } y \in M\}.$$

Quizfrage: Stimmt die Aussage von Lemma 17.11 auch noch im Fall $C = \emptyset$?

§ 18 Optimalitätsbedingungen der konvexen Optimierung

Literatur: Rockafellar, 1970, Section 27

Wir betrachten wieder die konvexe Optimierungsaufgabe aus (14.1)

$$\text{Minimiere} \quad f(x) \quad \text{über } x \in \mathbb{R}^n \tag{18.1}$$

mit konvexer Zielfunktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Desweiteren nehmen wir in diesem Abschnitt durchgängig f als eigentlich an. Das ist keine Einschränkung, denn wenn $f \equiv \infty$ ist, dann ist die Aufgabe (18.1) nicht interessant.

Satz 18.1 (Notwendige und hinreichende Optimalitätsbedingungen). *Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex. Dann sind die folgenden Aussagen für einen Punkt $x^* \in \mathbb{R}^n$ äquivalent:*

(*i*) *x^* ist ein (globaler) Minimierer für (18.1).*

(*ii*) *Es gilt $f'(x^*; d) \geq 0$ für alle $d \in \mathbb{R}^n$.*

(*iii*) *Es gilt $0 \in \partial f(x^*)$.*

Beweis. Die Bedingung $0 \in \partial f(x^*)$ ist äquivalent dazu, dass

$$f(x) \geq f(x^*) + 0^\mathsf{T}(x - x^*)$$

für alle $x \in \mathbb{R}^n$ gilt. Das ist aber wiederum nach Definition äquivalent dazu, dass x^* ein globaler Minimierer ist. Nach Satz 16.15 ist weiterhin $0 \in \partial f(x^*)$ äquivalent dazu, dass $f'(x^*; d) \geq 0^\mathsf{T} d = 0$ für alle $d \in \mathbb{R}^n$ ist. □

Beachte: Dieses Resultat verallgemeinert die notwendigen Optimalitätsbedingungen der unrestringierten Optimierung aus Satz 3.1.

Folgerung 18.2. *Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex. Dann sind die folgenden Aussagen für einen Punkt $x^* \in \mathbb{R}^n$, an dem f diffbar ist, äquivalent:*

(*i*) *x^* ist ein (globaler) Minimierer für (18.1).*

(*ii*) *Es gilt $f'(x^*; d) = 0$ für alle $d \in \mathbb{R}^n$.*

(*iii*) *Es gilt $\nabla f(x^*) = 0$.*

Beweis. Das Resultat folgt sofort aus Satz 16.24 und der Tatsache, dass wegen der Diffbarkeit von f in x^* gilt: $f'(x^*;d) = \nabla f(x^*)^\mathsf{T} d$. □

Da f eigentlich ist, kann die Aufgabe (18.1) bereits implizite Beschränkungen dadurch beinhalten, dass f möglicherweise nicht überall endlich ist. Jeder globale Minimierer x^* liegt notwendigerweise in $\operatorname{dom} f$.

Möchte man aber weitere Beschränkungen hinzufügen, so kann man dies durch Betrachtung der Aufgabe

$$\text{Minimiere} \quad f(x) + \delta_C(x) \quad \text{über } x \in \mathbb{R}^n \tag{18.2}$$

tun, wobei C eine nichtleere konvexe Menge ist. Wir sprechen hier von „abstrakten Nebenbedingungen“ im Gegensatz zu Nebenbedingungen, die in Form von Gleichungen oder Ungleichungen gegeben sind, vgl. (1.1). Effektiv findet die Minimierung dann über $C \cap \operatorname{dom} f$ statt. Wir geben eine Version von Satz 18.1 für diese Aufgabe an. Zuvor benötigen wir jedoch eine Aussage über die Richtungsableitung von Funktionen der Bauart wie in (18.2).

Lemma 18.3 (Richtungsableitung von $f + \delta_C$)**.** *Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex, C eine nichtleere konvexe Menge sowie $x \in C$. Dann gilt:*

$$(f + \delta_C)'(x) = f'(x;d) + \delta_C'(x;d) = \begin{cases} f'(x;d) & \text{für } d \in \mathcal{K}_C(x), \\ \infty & \text{sonst.} \end{cases} \tag{18.3}$$

Den im mittleren Term möglicherweise vorkommenden Fall $(-\infty) + \infty$ interpretieren wir also als ∞.

Beweis. Unter Beachtung von $\delta_C(x) = 0$ betrachten wir den Differenzenquotienten für $f + \delta_C$ an der Stelle x in Richtung d und mit $t > 0$:

$$\frac{f(x + t\,d) - f(x)}{t} + \frac{\delta_C(x + t\,d)}{t} =: q_1(t) + q_2(t).$$

Wir unterscheiden verschiedene Fälle. Falls $f'(x;d) \in \mathbb{R}$ oder $f'(x;d) = \infty$ ist, so können wir direkt den Grenzübergang $t \searrow 0$ durchführen und erhalten (18.3). Ist $f'(x;d) = -\infty$ und $\delta_C'(x;d) = 0$, dann ist $q_2(t) = 0$ für alle hinreichend kleinen $t > 0$, und wir können ebenfalls den Grenzübergang duurchführen. Der verbleibende Fall $f'(x;d) = -\infty$ und $\delta_C'(x;d) = \infty$ bedeutet, dass $q_1(t)$ für alle hinreichend kleinen $t > 0$ endlich ist, aber $q_2(t) = \infty$. Damit ist $q_1(t) + q_2(t) = \infty$ für alle hinreichend kleinen $t > 0$, also auch der Grenzwert. Damit ist die erste Gleichheit in (18.3) gezeigt. Die zweite folgt sofort aus Lemma 17.8. □

Satz 18.4 (Notwendige und hinreichende Optimalitätsbedingungen unter abstrakten Nebenbedingungen)**.** *Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und $C \subset \mathbb{R}^n$ konvex und nichtleer. Dann sind die folgenden Aussagen für einen Punkt $x^* \in \mathbb{R}^n$ äquivalent:*

(*i*) *x^* ist ein (globaler) Minimierer für (18.2).*

(*ii*) *Es gilt $f'(x^*;d) \geq 0$ für alle $d \in \mathcal{K}_C(x^*)$.*

(*iii*) *Es gilt* $f'(x^*; x - x^*) \geq 0$ *für alle* $x \in C$.

(*iv*) *Es gilt* $0 \in \partial(f + \delta_C)(x^*)$.

Ferner gilt: Die beiden folgenden Aussagen sind untereinander äquivalent, und jede der Aussagen ist hinreichend für jede der Aussagen (i) bis (iv).

(*v*) *Es gilt* $0 \in \partial f(x^*) + \mathcal{N}_C(x^*)$.

(*vi*) *Es gibt ein* $s \in \partial f(x^*)$ *mit der Eigenschaft* $s^\mathsf{T}(x - x^*) \geq 0$ *für alle* $x \in C$.

Falls die Regularitätsbedingung $\operatorname{rel\,int} \operatorname{dom} f \cap \operatorname{rel\,int} C \neq \emptyset$ *gilt, dann ist jede der Aussagen (v) und (vi) auch notwendig für jede der Aussagen (i) bis (iv).*

Beweis. Wir setzen $g := f + \delta_C$. Nach Satz 18.1 sind folgende Aussagen äquivalent: Aussage (*i*); $0 \in \partial g(x^*)$, also Aussage (*iv*); und $(f + \delta_C)'(x; d) \geq 0$ für alle $d \in \mathbb{R}^n$, was nach (18.3) gleichbedeutend ist mit

$$f'(x; d) \geq 0 \quad \text{für } d \in \mathcal{K}_C(x), \tag{18.4}$$

also Aussage (*ii*). Aus Aussage (*ii*) folgt aber sofort Aussage (*iii*), weil

$$x - x^* \in \mathcal{K}_C(x^*) = \{\beta\,(x - x^*) \mid x \in C,\ \beta > 0\}$$

ist. Umgekehrt folgt aus $f'(x^*; x - x^*) \geq 0$ mit der positiven Homogenität auch $f'(x^*; \beta\,(x - x^*)) = \beta\, f'(x^*; x - x^*) \geq 0$ für alle $\beta > 0$, also impliziert Aussage (*iii*) auch Aussage (*ii*). Die Äquivalenzen von Aussagen (*i*) bis (*iv*) sind damit gezeigt.

Aufgrund der Summenregel für das Subdifferential aus Satz 16.7 gilt $\partial f(x^*) + \partial \delta_C(x^*) \subseteq \partial(f + \delta_C)(x^*)$, wobei nach Lemma 17.11 wiederum $\partial \delta_C(x^*) = \mathcal{N}_C(x^*)$ ist. Wir haben also gezeigt, dass Aussage (*v*) hinreichend für Aussage (*iv*) ist.

Aussage (*v*) bedeutet, dass ein $s \in \partial f(x^*)$ existiert mit der Eigenschaft $-s \in \mathcal{N}_C(x^*)$. Die Definition 17.9 zeigt sofort die Äquivalenz mit Aussage (*vi*).

Aus Satz 16.7 folgt sogar $\partial f(x^*) + \partial \delta_C(x^*) = \partial(f + \delta_C)(x^*)$, also die Äquivalenz von Aussage (*iv*) und Aussage (*v*), falls die Regularitätsbedingung $(\operatorname{rel\,int} \operatorname{dom} f) \cap (\operatorname{rel\,int} \operatorname{dom} \delta_C) = (\operatorname{rel\,int} \operatorname{dom} f) \cap (\operatorname{rel\,int} C)$ erfüllt ist. Damit ist alles gezeigt. □

Folgerung 18.5 (Projektionsaufgabe). *Wir betrachten nochmals die Projektionsaufgabe (15.2). Es sei also C eine nichtleere, abgeschlossene, konvexe Menge und* $p \in \mathbb{R}^n$. *Dann gilt nach Satz 18.4 (iii):* $x^* \in C$ *ist genau dann gleich* $\operatorname{proj}_C(p)$, *also der eindeutige Minimierer von*

$$\textit{Minimiere} \quad f(x) := \frac{1}{2}\|x - p\|^2 + \delta_C(x) \quad \textit{über } x \in \mathbb{R}^n,$$

wenn $f'(x^*; x - x^*) \geq 0$ *gilt für alle* $x \in C$, *also*

$$(x^* - p)^\mathsf{T}(x - x^*) \geq 0 \quad \textit{für alle } x \in C.$$

Das ist genau die Bedingung, die wir bereits aus Satz 15.3 als notwendige und hinreichende Bedingung kennen, vgl. (15.4).

Quizfrage: Ist die Regularitätsbedingung erfüllt? Wie lautet die Bedingung aus Satz 18.4 *(iii)* ausgeschrieben?

Quizfrage: Wofür benötigt man denn hier die Abgeschlossenheit von C?

§ 19 Bundle-Verfahren

Literatur: Geiger, Kanzow, 2002, Kapitel 6

Wir besprechen in diesem Abschnitt ein Verfahren für *allgemeine* konvexe Optimierungsaufgaben ohne weitere Struktur. Genauer betrachten wir Aufgaben der Form

$$\text{Minimiere} \quad f(x) \quad \text{über } x \in \mathbb{R}^n \tag{19.1}$$

mit konvexer Zielfunktion $f\colon \mathbb{R}^n \to \mathbb{R}$. Wir nehmen also der Einfachheit halber an, dass f überall endlich ist.

Für Aufgaben mit mehr Struktur, etwa

$$\text{Minimiere} \quad g(Ax) + h(x),$$

wobei $g\colon \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ und $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvexe Funktionen sind und $A \in \mathbb{R}^{m\times n}$ ist, gibt es geeignetere Verfahren, die diese Struktur ausnutzen. Diese verwenden aber i. d. R. eine andere, äquivalente Formulierung als Sattelpunktaufgabe

$$\text{Minimiere} \quad \max_{p\in\mathbb{R}^m} p^\mathsf{T} Ax + h(x) - g^*(p)$$

bzw. die zugehörigen Optimalitätsbedingungen in der Form

$$\begin{aligned} Ax &\in \partial g^*(p), \\ -A^\mathsf{T} p &\in \partial h(x), \end{aligned}$$

wobei g^* die sogenannte **Fenchel-konjugierte** (auch: **konvex konjugierte**) (englisch: *Fenchel conjugate, convex conjugate*) Funktion zu g ist. Mehr zu Fenchel-konjugierten Funktionen, dualen Aufgaben und Sattelpunktaufgaben, deren Optimalitätsbedingungen und darauf aufbauende Lösungsverfahren erfährt man in Vorlesungen zur konvexen Optimierung.

§ 19.1 Die Richtung des steilsten Abstiegs

Warum benötigen wir überhaupt spezielle Verfahren für Optimierungsaufgaben (19.1), in denen die Zielfunktion konvex, aber i. A. nicht diffbar ist? Wir wollen dies an einem Beispiel motivieren. Dazu

führen wir zunächst, analog zu (4.9), die **Richtung des steilsten Abstiegs** von f im Punkt x als Lösung der Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & f'(x;d) \quad \text{über } d \in \mathbb{R}^n \\ \text{unter} \quad & \|d\| \leq 1 \end{aligned} \tag{19.2}$$

ein.[9] Da die Zielfunktion $f'(x;d)$ nach Lemma 16.20 stetig und $\overline{B_1(0)}$ kompakt ist, existiert nach dem Satz von Weierstraß bzw. Satz 1.4 eine globale Lösung von (19.2). Aufgrund der Konvexität der Zielfunktion (Satz 16.13) existieren keine lokalen Minimierer, die nicht gleichzeitig globale Minimierer sind.

Lemma 19.1 (Eindeutigkeit der Richtung des steilsten Abstiegs). *Es sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex und $x_0 \in \mathbb{R}^n$. Falls ein $d \in \mathbb{R}^n$ existiert, sodass $f'(x_0;d) < 0$ ist, dann ist die Lösung von (19.2) eindeutig.*

Beweis. □

Wir werden in Satz 19.3 die Richtung des steilsten Abstiegs charakterisieren. Dazu benötigen wir folgendes Hilfsresultat.

Lemma 19.2 (Minimax-Lemma). *Es seien $M_1, M_2 \subseteq \mathbb{R}^n$ beliebige nichtleere Mengen. Dann gilt*

$$\sup_{x\in M_1} \inf_{y\in M_2} x^\mathsf{T} y \leq \inf_{y\in M_2} \sup_{x\in M_1} x^\mathsf{T} y. \tag{19.3}$$

Beweis.

$$\begin{aligned} & x^\mathsf{T} y \leq \sup_{\overline{x}\in M_1} \overline{x}^\mathsf{T} y && \text{für alle } x \in M_1,\, y \in M_2 \\ \Rightarrow \quad & \inf_{\overline{y}\in M_2} x^\mathsf{T} \overline{y} \leq \sup_{\overline{x}\in M_1} \overline{x}^\mathsf{T} y && \text{für alle } x \in M_1,\, y \in M_2 \\ \Rightarrow \quad & \inf_{\overline{y}\in M_2} x^\mathsf{T} \overline{y} \leq \inf_{\overline{y}\in M_2} \sup_{\overline{x}\in M_1} \overline{x}^\mathsf{T} \overline{y} && \text{für alle } x \in M_1 \\ \Rightarrow \quad & \sup_{\overline{x}\in M_1} \inf_{\overline{y}\in M_2} \overline{x}^\mathsf{T} \overline{y} \leq \inf_{\overline{y}\in M_2} \sup_{\overline{x}\in M_1} \overline{x}^\mathsf{T} \overline{y}. \end{aligned}$$

□

Sogenannte **Minimax-Theoreme** beschäftigen sich mit der Frage, unter welchen Voraussetzungen in (19.3) die Gleichheit gilt.[10]

[9] Statt $\|d\| \leq 1$ könnten wir wie in (4.9) auch $\|d\| = 1$ schreiben. (**Quizfrage:** Begründung?)

[10] Minimax-Theoreme haben Anwendungen u. a. in der Spieltheorie. Ein klassisches Resultat ist das von von Neumann, 1928, das besagt: Wenn $C_1 \subseteq \mathbb{R}^n$ und $C_2 \subseteq \mathbb{R}^m$ beide konvex und kompakt sind und $f\colon C_1 \times C_2 \to \mathbb{R}$ stetig und **konkav-konvex** ist, also

$$\begin{aligned} f(\cdot, y)\colon C_1 \to \mathbb{R} \text{ ist konkav für alle } y \in C_2, \\ f(x, \cdot)\colon C_2 \to \mathbb{R} \text{ ist konvex für alle } x \in C_1, \end{aligned}$$

dann gilt

$$\max_{x\in C_1} \min_{y\in C_2} f(x,y) = \min_{y\in C_2} \max_{x\in C_1} f(x,y).$$

Dieses Resultat ist mittlerweile in zahlreiche Richtungen verallgemeinert worden.

Satz 19.3 (Richtung des steilsten Abstiegs). *Es sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex und $x_0 \in \mathbb{R}^n$ kein globaler Minimierer von f. Dann ist die eindeutige Lösung von (19.2) gegeben durch*

$$d := -\frac{g}{\|g\|}, \quad \text{wobei } g := \operatorname{proj}_{\partial f(x_0)}(0) \text{ ist.} \tag{19.4}$$

Es gilt $f'(x;d) = -\|g\|$.

Beweis. Wir zeigen zunächst die Eindeutigkeit der Lösung von (19.2). Da x_0 kein globaler und damit auch kein lokaler Minimierer von f ist, gibt es ein $d \in \mathbb{R}^n$ mit $f'(x_0;d) < 0$. (**Quizfrage:** Genaue Begründung?) Die Eindeutigkeit der Richtung des steilsten Abstiegs folgt nun aus Lemma 19.1.

Da x_0 kein globaler Minimierer ist, gilt weiter $0 \notin \partial f(x_0)$. Da $\partial f(x_0)$ nichtleer ist (Satz 16.5) sowie abgeschlossen und konvex (Satz 16.6), existiert $g := \operatorname{proj}_{\partial f(x_0)}(0) \neq 0$ nach Satz 15.3 und ist charakterisiert durch

$$(g-0)^\mathsf{T}(s-g) \geq 0 \quad \text{für alle } s \in \partial f(x_0).$$

Mit der Definition $d := -g/\|g\|$ folgt also

$$s^\mathsf{T} d \leq d^\mathsf{T} g = -\|g\| \quad \text{für alle } s \in \partial f(x_0).$$

Nach Satz 16.15 folgt also (**Quizfrage:** Warum ist (16.19) anwendbar?)

$$f'(x_0;d) = \max\{s^\mathsf{T} d \mid s \in \partial f(x_0)\} \leq -\|g\|.$$

Wir können also den Optimalwert der Aufgabe (19.2) schreiben als

$$\min_{\|d\|\leq 1} f'(x_0;d) = \min_{\|d\|\leq 1} \max_{s\in\partial f(x_0)} s^\mathsf{T} d \leq -\|g\|. \tag{19.5}$$

Wir betrachten jetzt die Aufgabe nach Vertauschung von min und max, also

$$\max_{s\in\partial f(x_0)} \min_{\|d\|\leq 1} s^\mathsf{T} d.$$

Zunächst ist noch unklar, ob wir hier überhaupt max bzw. min schreiben dürfen oder sup bzw. inf verwenden müssen. Die innere Aufgabe hat, für gegebenes $s \in \partial f(x_0)$, aber offenbar $d = -s/\|s\|$ als eindeutigen Minimierer mit Optimalwert $-\|s\|$. (**Quizfrage:** Warum ist $s \neq 0$?) Wir können die Aufgabe also auch schreiben als

$$\text{Maximiere} \quad -\|s\| \text{ über } s \in \partial f(x_0). \tag{19.6}$$

Da $\partial f(x_0)$ nichtleer und kompakt ist und $-\|s\|$ stetig, wird das Maximum angenommen. Die Aufgabe ist weiter äquivalent zu

$$\text{Minimiere} \quad \|s-0\| \text{ über } s \in \partial f(x_0),$$

deren (eindeutige) Lösung wir bereits kennen: $s = \operatorname{proj}_{\partial f(x_0)}(0) = g$. Der Optimalwert von (19.6) ist also $-\|g\|$.

Wir fassen zusammen und erhalten unter Zuhilfenahme des Minimax-Lemmas 19.2:

$$-\|g\| = \max_{s\in\partial f(x_0)} \min_{\|d\|\leq 1} s^\mathsf{T} d \leq \min_{\|d\|\leq 1} \max_{s\in\partial f(x_0)} s^\mathsf{T} d \leq -\|g\|.$$

Es gilt also überall Gleichheit, und aus (19.5) folgt wie behauptet $f'(x;d) = -\|g\|$. □

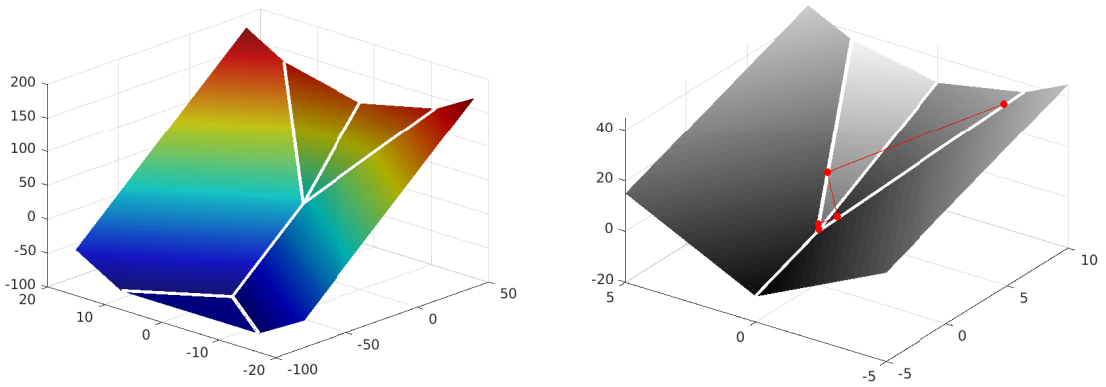


Abbildung 19.1: Darstellung der Funktion aus Beispiel 19.4 mit Linien der Nichtdifferenzierbarkeit (links) und Ausschnitt mit den ersten Iterierten des Verfahrens des steilsten Abstiegs mit exakter Liniensuche, ausgehend von $x^{(0)} = (9, -3)^\mathsf{T}$ (rechts).

Es folgt nun das eingangs erwähnte Beispiel, das zeigt, dass man unter Verwendung der Richtung des steilsten Abstiegs (19.4) selbst mit exakter Liniensuche ein Abstiegsverfahren erhält, das gegen „uninteressante“ Punkte konvergieren kann.

Beispiel 19.4 (aus Bonnans u. a., 2003, Example 9.1). *Die konvexe Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ sei gegeben durch $f(x) := \max\{f_0(x), f_{-1}(x), f_{-2}(x), f_1(x), f_2(x)\}$ mit*

$$f_0(x) := -100, \quad f_{\pm 1}(x) := 3\,x_1 \pm 2\,x_2, \quad f_{\pm 2}(x) := 2\,x_1 \pm 5\,x_2.$$

Der Optimalwert von $f^* = -100$ wird auf der konvexen Menge der globalen Minimierer $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \le -50,\ |x_2| \le 0.4\,x_1 + 20\}$ angenommen, vgl. Abbildung 19.1.

Man kann zeigen, dass das Verfahren des steilsten Abstiegs mit exakter Liniensuche, gestartet bei $x^{(0)} = (9, -3)^\mathsf{T}$, die Iterationsfolge

$$x^{(k)} = \left(3^{2-k}, (-1)^k 3^{1-k}\right)^\mathsf{T}$$

erzeugt. Das Subdifferential ist jeweils

$$\partial f(x^{(k)}) = \operatorname{conv}\left\{\begin{pmatrix} 3 \\ (-1)^{k+1}2 \end{pmatrix}, \begin{pmatrix} 2 \\ (-1)^{k+1}5 \end{pmatrix}\right\}.$$

Wir erhalten also $g^{(k)} = \operatorname{proj}_{\partial f(x^{(k)})}(0) = \left(3, (-1)^{k+1}2\right)^\mathsf{T}$ mit $\|g^{(k)}\| = \sqrt{13}$ und Suchrichtungen $d^{(k)} = -g^{(k)}/\|g^{(k)}\|$.

Die Funktionswerte $f(x^{(k)}) = 11 \cdot 3^{1-k}$ sind streng monoton fallend. Die Folge $(x^{(k)})$ konvergiert (q-linear) gegen den Punkt $x^* = (0, 0)^\mathsf{T}$ mit $f(x^*) = 0$, der ein nicht-optimaler, „nichtglatter“ Punkt der Zielfunktion f ist.

Die Schwierigkeiten lassen sich auf folgende Beobachtungen zurückführen:

(1) Das Subdifferential $\partial f : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ ist als mengenwertige Funktion zwar **außerhalbstetig** (englisch: *outer semicontinuous*), d. h.:

Für alle Folgen $x^{(k)} \to x$ und alle Folgen $s^{(k)} \in \partial f(x^{(k)})$ mit $s^{(k)} \to s$ gilt $s \in \partial f(x)$,

(**Quizfrage:** Beweis?) jedoch *nicht* **innerhalbstetig** (englisch: *inner semicontinuous*). Die Innerhalbstetigkeit würde bedeuten:

Für alle Folgen $x^{(k)} \to x$ und alle $s \in \partial f(x)$ gibt es eine Folge $s^{(k)} \in \partial f(x^{(k)})$ mit $s^{(k)} \to s$.

(**Quizfrage:** Beispiel, das die fehlende Innerhalbstetigkeit von ∂f zeigt?)

(2) Die fehlende Innerhalbstetigkeit von ∂f bedeutet, dass bereits kleine Änderungen in x große Änderungen in $\partial f(x)$ hervorrufen können. Da die Richtung des steilsten Abstiegs an einer Iterierten $x^{(k)}$ gemäß (19.4) aus $\partial f(x^{(k)})$ berechnet wird, ergibt sich kein stabiles Verfahren.

Weitere praktische Nachteile von Verfahren des steilsten Abstiegs für (19.1) sind:

(3) Die Bestimmung von $g^{(k)} = \operatorname{proj}_{\partial f(x^{(k)})}(0)$ und damit die Bestimmung der Suchrichtung $d^{(k)} = -g^{(k)}/\|g^{(k)}\|$ erfordert i. W. die Bestimmung des *gesamten* Subdifferentials.

Ende der Woche 13

§ 19.2 Das Bundle-Teilproblem

Wir entwickeln im Folgenden einen einfachen Vertreter der Klasse der **Bundle-Verfahren**, einer Familie leistungsfähiger Verfahren für *allgemeine* konvexe Optimierungsaufgaben (19.1), die die oben genannten Nachteile nicht aufweisen. Bundle-Verfahren basieren auf der Idee, Subgradienten $s^{(j)} \in \partial f(x^{(j)})$ einer Reihe von Punkten $x^{(j)}$, $j = 0, 1, \ldots, k$ zu sammeln (daher der Name **Bündel**) und daraus ein stückweise lineares, konvexes Modell der Zielfunktion f zu erstellen:

$$f^{\mathrm{CP}}(x) := \max\left\{ f(x^{(j)}) + (s^{(j)})^{\mathsf{T}}(x - x^{(j)}) \,\middle|\, j = 0, 1, \ldots, k \right\} \tag{19.7}$$

Dieses sogenannte **Schnittebenenmodell** (englisch: *cutting plane model*) hat folgende Eigenschaften:

Lemma 19.5 (Eigenschaften des Schnittebenenmodells). *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex, $x^{(j)} \in \mathbb{R}^n$ und $s^{(j)} \in \partial f(x^{(j)})$ für $j = 0, 1, \ldots, k$. Dann ist $f^{\mathrm{CP}} : \mathbb{R}^n \to \mathbb{R}$ konvex, und es gilt $f^{\mathrm{CP}}(x) \le f(x)$ für alle $x \in \mathbb{R}^n$ sowie $f^{\mathrm{CP}}(x^{(j)}) = f(x^{(j)})$ für $j = 0, 1, \ldots, k$.*

Beweis. f^{CP} ist als Maximum konvexer (linearer) Funktionen konvex nach Satz 13.16. Nach der Subgradientenungleichung (16.1) gilt

$$f(x) \ge f(x^{(j)}) + (s^{(j)})^{\mathsf{T}}(x - x^{(j)}) \quad \text{für alle } x \in \mathbb{R}^n \text{ und alle } j = 0, 1, \ldots, k,$$

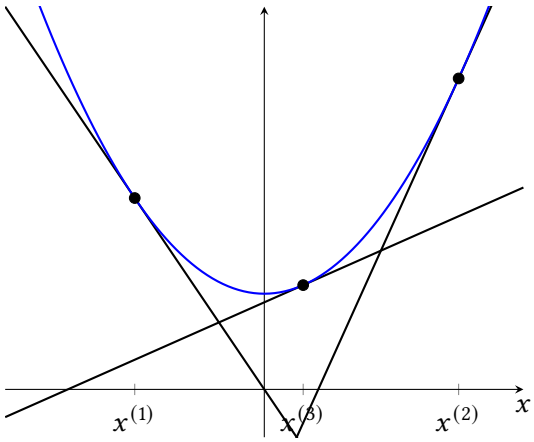


Abbildung 19.2: Illustration des Schnittebenenmodells (19.7).

also auch

$$f(x) \geq \max\left\{f(x^{(j)}) + (s^{(j)})^\mathsf{T}(x - x^{(j)}) \,\middle|\, j = 0, 1, \ldots, k\right\} = f^{\mathrm{CP}}(x) \quad \text{für alle } x \in \mathbb{R}^n.$$

Speziell für $x = x^{(j)}$ folgt

$$\begin{aligned} f(x^{(j)}) &\geq \max\left\{f(x^{(j)}) + (s^{(j)})^\mathsf{T}(x^{(j)} - x^{(j)}) \,\middle|\, j = 0, 1, \ldots, k\right\} = f^{\mathrm{CP}}(x^{(j)}) \\ &\geq f(x^{(j)}) + (s^{(j)})^\mathsf{T}(x^{(j)} - x^{(j)}) \\ &= f(x^{(j)}), \end{aligned}$$

also die Gleichheit. □

Das Schnittebenenmodell (19.7) und Varianten davon werden gleich als Zielfunktion eines Ersatzproblems für (19.1) verwendet. Zur Vereinfachung der Notation wählen wir einen Referenzpunkt $\overline{x}$ und ersetzen die Variable x durch die Richtungsvariable $d := x - \overline{x}$. Zur Abkürzung der Notation führen wir die **Linearisierungsfehler**

$$\overline{\alpha}^{(j)} := f(\overline{x}) - f(x^{(j)}) - (s^{(j)})^\mathsf{T}(\overline{x} - x^{(j)}) \geq 0, \quad j = 0, 1, \ldots, k \tag{19.8}$$

ein, die an der Stelle $\overline{x}$ die Differenz zwischen dem tatsächlichen Funktionswert und dem auf dem Subgradienten $s^{(j)}$ basierenden linearen Modell messen.

Beachte: Die Eigenschaft $\overline{\alpha}^{(j)} \geq 0$ folgt sofort aus Lemma 19.5.

Um das Modell (19.7) auf die Richtungsvariable d umzuschreiben, nutzen wir

$$\begin{aligned} f(x^{(j)}) + (s^{(j)})^\mathsf{T}(x - x^{(j)}) &= f(x^{(j)}) + (s^{(j)})^\mathsf{T}(x - \overline{x} + \overline{x} - x^{(j)}) \\ &= f(x^{(j)}) + (s^{(j)})^\mathsf{T}(d + \overline{x} - x^{(j)}) \\ &= (s^{(j)})^\mathsf{T} d - \overline{\alpha}^{(j)} + f(\overline{x}). \end{aligned}$$

Da bei der Minimierung bzgl. d die von j unabhängige Konstante $f(\overline{x})$ keine Rolle spielt, setzen wir als **Schnittebenenrichtungsmodell** jetzt die Funktion

$$m^{\mathrm{CP}}(d) := \max\left\{(s^{(j)})^\mathsf{T} d - \overline{\alpha}^{(j)} \,\middle|\, j = 0, 1, \ldots, k\right\} \tag{19.9}$$

an. Um die Minimierung dieser stückweise linearen, konvexen Funktion (19.9) durchzuführen, ist es günstig, zur sogenannten **Epigraph-Reformulierung** (19.10) überzugehen:

Lemma 19.6 (Epigraph-Reformulierung[11]). *Der Vektor $d \in \mathbb{R}^n$ ist ein (globaler) Minimierer von (19.9) genau dann, wenn $(d, \xi) \in \mathbb{R}^n \times \mathbb{R}$ ein (globaler) Minimierer der Aufgabe*

$$\begin{aligned} \textit{Minimiere} \quad & \xi \quad \textit{über } (d, \xi) \in \mathbb{R}^n \times \mathbb{R} \\ \textit{unter} \quad & (s^{(j)})^\mathsf{T} d - \overline{\alpha}^{(j)} \leq \xi, \quad j = 0, 1, \dots k \end{aligned} \tag{19.10}$$

ist.

Beweis. □

Wir haben also die Nichtglattheit von (19.9) gegen Nebenbedingungen in (19.10) „eingetauscht“. Die Aufgabe (19.10) ist nun ein LP. Wir untersuchen jetzt das dazu duale LP.[12] Wir können leicht nachrechnen, dass dieses durch die (Minimierungs-)Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & \sum_{j=0}^{k} \overline{\alpha}^{(j)} \lambda_j \quad \text{über } \lambda \in \mathbb{R}^{k+1} \\ \text{unter} \quad & \lambda \geq 0 \text{ und } \sum_{j=0}^{k} \lambda_j = 1 \\ \text{sowie} \quad & \sum_{j=0}^{k} \lambda_j s^{(j)} = 0 \end{aligned} \tag{19.11}$$

gegeben ist.

Das LP (19.10) kann unbeschränkt sein, was genau dann der Fall ist, wenn (19.11) unzulässig ist (Satz 8.7). Letzteres ist genau dann der Fall, wenn sich der Nullvektor nicht aus den Subgradienten $s^{(j)}$ konvexkombinieren lässt, also insbesondere dann, wenn nur wenige Subgradienten verwendet werden.

Um diese Schwierigkeit zu umgehen, wollen wir das primale Problem (19.10) durch Hinzufügen eines Terms der Bauart $\|d\|^2$ in der Zielfunktion regularisieren. Die Aufgabe ist dann kein LP mehr, sondern

[11] Wie sich aus dem Beweis ergibt, ist die Epigraph-Reformulierung immer möglich, wenn die Zielfunktion als das punktweise Maximum endlich vieler konvexer Funktionen definiert ist.

[12] Mit den Kenntnissen aus Kapitel 2 können wir leicht herleiten (siehe auch Übungsaufgabe 1 auf Übungsblatt 6), dass das zu

$$\begin{aligned} \text{Minimiere} \quad & c^\mathsf{T} x \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & A x \leq b \end{aligned}$$

duale LP durch

$$\begin{aligned} \text{Maximiere} \quad & -b^\mathsf{T} \lambda \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & A^\mathsf{T} \lambda = -c \\ \text{und} \quad & \lambda \geq 0 \end{aligned}$$

gegeben ist. (**Quizfrage:** Begründung?) Die notwendigen und hinreichenden Optimalitätsbedingungen bestehen neben der primalen und der dualen Zulässigkeit aus der Komplementaritätsbedingung $\lambda^\mathsf{T}(A x - b) = 0$.

ein konvexes QP (vgl. Definition 1.2). Auch für solche QPs gibt es eine Dualitätstheorie, die wir hier aber nicht im Detail ausführen. Stattdessen stellen wir ohne Beweis oder Herleitung die primalen und dualen Aufgaben für (19.10) und zwei Varianten davon in Tabelle 19.1 zusammen. Dabei verwenden wir für die in den dualen Aufgaben stets vorkommenden Bedingungen $\lambda \geq 0$ und $\sum_{j=0}^{k} \lambda_j = 1$ die Abkürzung $\lambda \in \Delta$ (Einheitssimplex, vgl. Beispiel 13.2). Außerdem führen wir zur Vermeidung von Summensymbolen die Matrix

$$S := \begin{bmatrix} | & & | \\ s^{(0)} & \dots & s^{(k)} \\ | & & | \end{bmatrix} \in \mathbb{R}^{n \times (k+1)}$$

sowie den Vektor

$$\overline{\alpha} := (\overline{\alpha}^{(0)}, \dots, \overline{\alpha}^{(k)})^\mathsf{T} \in \mathbb{R}^{k+1}$$

ein. Die Parameter τ und ε sind positive Zahlen.

Beachte: Auch die dualen Aufgaben sind hier immer in Minimierungsform angegeben.

Minimiere	ξ	über $(d, \xi) \in \mathbb{R}^n \times \mathbb{R}$	Minimiere	$\overline{\alpha}^\mathsf{T}\lambda$	über $\lambda \in \mathbb{R}^{k+1}$	
unter	$S^\mathsf{T}d - \overline{\alpha} \leq \xi \mathbf{1}$		unter	$\lambda \in \Delta$		
		(19.12)	und	$S\lambda = 0$		(19.13)

Komplementarität: $(S^\mathsf{T}d - \overline{\alpha} - \xi \mathbf{1})^\mathsf{T}\lambda = 0$

Minimiere	$\xi + \frac{\tau}{2}\|d\|^2$	über $(d, \xi) \in \mathbb{R}^n \times \mathbb{R}$	Minimiere	$\overline{\alpha}^\mathsf{T}\lambda + \frac{1}{2\tau}\|S\lambda\|^2$	über $\lambda \in \mathbb{R}^{k+1}$	
unter	$S^\mathsf{T}d - \overline{\alpha} \leq \xi \mathbf{1}$	(19.14)	unter	$\lambda \in \Delta$		(19.15)

Komplementarität: $(S^\mathsf{T}d - \overline{\alpha} - \xi \mathbf{1})^\mathsf{T}\lambda = 0$

Minimiere	$\xi + \frac{\tau}{2}\|d\|^2 + \varepsilon\eta$	über $(d, \xi, \eta) \in \mathbb{R}^n \times \mathbb{R}^2$	Minimiere	$\frac{1}{2\tau}\|S\lambda\|^2$	über $\lambda \in \mathbb{R}^{k+1}$	
unter	$S^\mathsf{T}d - \eta\overline{\alpha} \leq \xi \mathbf{1}$		unter	$\lambda \in \Delta$		
und	$\eta \geq 0$	(19.16)	und	$\overline{\alpha}^\mathsf{T}\lambda \leq \varepsilon$		(19.17)

Komplementarität: $(S^\mathsf{T}d - \overline{\alpha} - \xi \mathbf{1})^\mathsf{T}\lambda = 0$, $\eta^\mathsf{T}(\overline{\alpha}^\mathsf{T}\lambda - \varepsilon) = 0$

Tabelle 19.1: Zusammenstellung primaler und dualer Varianten der Epigraph-Reformulierung (19.10) des Schnittebenenproblems.

Bevor wir die Interpretationen und den Nutzen der verschiedenen Varianten angeben, benötigen wir weitere Informationen. Zunächst geben wir (ohne Beweis) ein bemerkenswertes Resultat für QPs, analog zum Existenzsatz für LPs 6.9 an.

Satz 19.7 (Frank-Wolfe lemma[13], Existenzsatz für QPs).
Es seien $Q \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$, $\gamma \in \mathbb{R}$ sowie $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ mit $m \in \mathbb{N}_0$. Wir betrachten das QP

$$\begin{aligned} \text{Minimiere} \quad & f(x) := \frac{1}{2} x^\mathsf{T} Q\, x + c^\mathsf{T} x + \gamma \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & A\, x \le b \end{aligned} \tag{19.18}$$

mit zulässiger Menge F. Ist der Optimalwert

$$f^* = \inf\{f(x) \mid x \in F\}$$

endlich, also die Aufgabe (19.18) weder unzulässig ($f^* = +\infty$) noch unbeschränkt ($f^* = -\infty$), so besitzt (19.18) mindestens einen globalen Minimierer.

Beachte: Der Satz gilt natürlich auch für QPs, die lineare Gleichungsnebenbedingungen enthalten, da man diese ja immer in der Form zweier Ungleichungen schreiben kann.

Mit Satz 19.7 können wir zeigen, dass die Aufgaben (19.14) und (19.15) jeweils mindestens einen globalen Minimierer besitzen (**Quizfrage:** Details?), während das für die LPs (19.12) und (19.13) ja nicht notwendigerweise der Fall war. Auch für (19.16) und (19.17) können wir die Existenz von Lösungen unter einer gewissen Bedingung zeigen. (**Quizfrage:** Welche Bedingung ist das?) Da die Zielfunktionen und zulässigen Mengen jeweils konvex sind, gibt es jeweils keine lokalen Minimierer, die nicht bereits globale Minimierer sind.

Um die Aufgaben aus Tabelle 19.1 besser einordnen zu können, benötigen wir einige weitere Begriffe.

Definition 19.8 (ε-Subdifferential, vgl. Definition 16.1). *Es sei $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ eine konvexe Funktion und $\varepsilon \ge 0$.*

(*i*) *Ein Vektor $s \in \mathbb{R}^n$ heißt ein (Euklidischer) **ε-Subgradient** von f im Punkt $x_0 \in \mathbb{R}^n$, wenn die **ε-Subgradientenungleichung** gilt:*

$$f(x) \ge f(x_0) + s^\mathsf{T}(x - x_0) - \varepsilon \quad \text{für alle } x \in \mathbb{R}^n. \tag{19.19}$$

(*ii*) *Die Menge $\partial_\varepsilon f(x_0)$* aller *$\varepsilon$-Subgradienten im Punkt x_0 heißt das **ε-Subdifferential** von f in x_0.*

(*iii*) *f heißt **ε-subdifferenzierbar** (kurz: **ε-subdiffbar**) im Punkt $x_0 \in \mathbb{R}^n$, wenn $\partial_\varepsilon f(x_0) \ne \emptyset$ ist.*

Das ε-Subdifferential hat u. a. folgende Eigenschaften:

Satz 19.9 (Eigenschaften des ε-Subdifferentials). *Es sei $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ eine konvexe Funktion und $x_0 \in \mathbb{R}^n$. Dann gilt*

[13] Dieses Resultat wurde ursprünglich in Frank, Wolfe, 1956 für konvexe QPs bewiesen. Ein direkter Beweis, der auch den nicht-konvexen Fall einschließt, findet sich in Blum, Oettli, 1972.

(*i*) $\partial f_0(x_0) = \partial f(x_0)$.

(*ii*) $\partial f_\varepsilon(x_0) \subseteq \partial f_{\varepsilon'}(x_0)$ *für alle* $0 \le \varepsilon \le \varepsilon'$ *und insbesondere* $\partial f(x_0) \subseteq \partial f_\varepsilon(x_0)$.

(*iii*) *Für alle* $\varepsilon \ge 0$ *ist* $\partial_\varepsilon f(x_0)$ *abgeschlossen und konvex und im Falle von* $x_0 \in \operatorname{int}\operatorname{dom} f$ *sogar kompakt.*

(*iv*) *Für* $\varepsilon > 0$ *ist* $\partial f_\varepsilon \colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ *außerhalbstetig und innerhalbstetig.*

Beweis. □

Mit dem ε-Subdifferential an Stelle des Subdifferentials könnte man im Prinzip ein Abstiegsverfahren mit Liniensuche aufbauen, wobei die Suchrichtung an einer Stelle $\overline{x}$ an Stelle von (19.4) durch

$$d := -\frac{g}{\|g\|} \quad \text{mit } g := \operatorname{proj}_{\partial_\varepsilon f(\overline{x})}(0) \tag{19.20}$$

bestimmt wird.[14] Jedoch muss man für (19.20) das gesamte ε-Subdifferential $\partial_\varepsilon f(\overline{x})$ berechnen können, was für viele Zielfunktionen unrealistisch ist.

Praktischer wäre es, eine implizit gegebene konvexe Teilmenge von $\partial_\varepsilon f(\overline{x})$ zu verwenden, auf die man einfach projizieren kann. Wir zeigen jetzt, dass die Nebenbedingungen in (19.17) genau eine solche Teilmenge

$$G_\varepsilon(\overline{x}) := \{S\lambda \mid \lambda \in \Delta \text{ und } \overline{\alpha}^\mathsf{T}\lambda \le \varepsilon\} \subseteq \partial_\varepsilon f(\overline{x}) \tag{19.21}$$

beschreiben. Aufgrund von $s^{(j)} \in \partial f(x^{(k)})$ gilt nämlich

$$\begin{aligned} f(x) &\ge f(x^{(j)}) + (s^{(j)})^\mathsf{T}(x - x^{(j)}) \\ &= f(x^{(j)}) \pm f(\overline{x}) + (s^{(j)})^\mathsf{T}(x - \overline{x} + \overline{x} - x^{(j)}) \\ &= f(\overline{x}) - \overline{\alpha}^{(j)} + (s^{(j)})^\mathsf{T}(x - \overline{x}) \end{aligned}$$

für alle $x \in \mathbb{R}^n$, d. h., $s^{(j)}$ gehört zu $\partial_{\overline{\alpha}^{(j)}} f(\overline{x})$. Unter Berücksichtigung von $\lambda \in \Delta$ ergibt die Summation dieser mit λ_j gewichteten Ungleichungen:

$$f(x) \ge f(\overline{x}) - \overline{\alpha}^\mathsf{T}\lambda + (S\lambda)^\mathsf{T}(x - \overline{x}).$$

Die Nebenbedingung $\overline{\alpha}^\mathsf{T}\lambda \le \varepsilon$ sichert also gerade

$$S\,\lambda \in \partial_\varepsilon f(\overline{x}). \tag{19.22}$$

Wir können nun unsere Interpretation der einzelnen Aufgaben (19.12)–(19.17) in einer Bemerkung festhalten:

[14] Die so festgelegte Richtung d minimiert dann gerade die **ε-Richtungsableitung**

$$f'_\varepsilon(x; d) := \lim_{t \searrow 0} \frac{f(x + t\,d) - f(x) + \varepsilon}{t},$$

vgl. (19.2).

Bemerkung 19.10 (zu den Aufgaben (19.12)–(19.15)).

(*i*) *Die Aufgabe* (19.14) *entspricht einer Minimierung des Schnittebenenrichtungsmodells* (19.9)*, wobei jedoch zur Zielfunktion der oft als* ***Proximalterm*** *(englisch:* proximal term*)* $\frac{\tau}{2}\|d\|^2$ *bezeichnete Term hinzugefügt wurde. Dieser bestraft Richtungen* d *mit großer Norm. Im ursprünglichen Schnittebenenmodell* (19.7) *mit der Variable* x *entspricht dies dem Hinzufügen des Terms* $\frac{\tau}{2}\|x-\overline{x}\|^2$*.*

(*ii*) *Die eindeutige (****Quizfrage:*** *Warum eindeutig?) Lösung* (d,ξ) *von* (19.14) *kann man aus einer Lösung* λ *des dualen QPs* (19.15) *erhalten, indem man*

$$d := -\frac{1}{\tau}S\lambda \quad \text{und} \quad \xi := -\frac{1}{\tau}\|S\lambda\|^2 - \overline{\alpha}^\mathsf{T}\lambda = -\tau\|d\|^2 - \overline{\alpha}^\mathsf{T}\lambda \tag{19.23}$$

setzt.

Beachte: *Während* λ *möglicherweise nicht eindeutig ist, ist es* $S\lambda$ *doch. (****Quizfrage:*** *Warum?)*

(*iii*) *Das duale QP* (19.17) *ist (und zwar für beliebiges* $\tau > 0$*) gerade die Aufgabe der orthogonalen Projektion der Null auf die kompakte Menge*

$$G_\varepsilon(\overline{x}) \subseteq \partial_\varepsilon f(\overline{x}).$$

Dabei ist $G_\varepsilon(\overline{x})$ *genau dann nichtleer, wenn* $\varepsilon \geq \|\overline{\alpha}\|_\infty$ *gilt.*

(*iv*) *Die dualen Aufgaben* (19.15) *und* (19.17) *sind eng verwandt. Während in* (19.17) *die Nebenbedingung* $\overline{\alpha}^\mathsf{T}\lambda \leq \varepsilon$ *explizit gefordert wird, werden in* (19.15) *große Werte von* $\overline{\alpha}^\mathsf{T}\lambda$ *bestraft, und zwar umso mehr, je größer der Parameter* τ *ist.*

Man kann zeigen, dass eine Lösung von (19.15) *auch eien Lösung von* (19.17) *ist, wenn man* $\varepsilon := \overline{\alpha}^\mathsf{T}\lambda$ *wählt. Umgekehrt ist eine Lösung von* (19.17) *auch eine Lösung von* (19.15) *für geeignetes* $\tau > 0$*.*

§ 19.3 Ein Bundle-Verfahren

Wir werden uns mit einem Bundle-Verfahren beschäftigen, das auf (19.15) basiert und damit gleichermaßen auf der Proximalpunkt-Regularisierung des Schnittebenenrichtungsmodells (19.9). Das Sammeln von Subgradienten $s^{(j)}$ an vorangegangen Iterierten dient also gleichzeitig dem Zweck, das Schnittebenenrichtungsmodell anzureichern, wie auch der besseren Ausschöpfung des ε-Subdifferentials $\partial_\varepsilon f(\overline{x})$ durch $G_\varepsilon(\overline{x})$ (für implizit festgelegtes ε). Gemeinsames Ziel ist es dabei, eine ausreichend gute Abstiegsrichtung zu erhalten.

Das Verfahren unterscheidet zwei Sorten von Iterierten, $x^{(k)}$ und $y^{(k)}$. Die sogenannten **wesentlichen Iterierten** (englisch: *serious iterates*) $x^{(k)}$ oder auch **Stabilitätszentren** (englisch: *stability centers*) dienen als Referenzpunkte und treten an die Position der bisher mit $\overline{x}$ bezeichneten Stelle. Ausgehend von der aktuellen wesentlichen Iterierten $x^{(k)}$ wird ein neuer Kandidat oder **Versuchspunkt** (englisch: *trial iterate*) $y^{(k+1)} := x^{(k)} + d$ bestimmt, basierend auf der Lösung von (19.14) bzw. gleichwertig von (19.15). Falls $y^{(k+1)}$ genügend Abstieg liefert, so wird dieser Punkt die nächste wesentliche Iterierte, also $x^{(k+1)} := y^{(k+1)}$ gesetzt. Das nennt man einen **wesentlichen Schritt** (englisch: *serious step*) des Verfahrens.

Andernfalls bleibt die wesentliche Iterierte unverändert, also $x^{(k+1)} := x^{(k)}$, aber die Subgradienteninformation an der Stelle $y^{(k+1)}$ fließt in das aktuelle Modell ein. Diesen Fall nennt man einen **Nullschritt** (englisch: *null step*).

Algorithmus 19.11 (Ein Bundle-Verfahren, vgl. Geiger, Kanzow, 2002, Algorithmus 6.72)**.**
Eingabe: *Startschätzung* $x^{(0)} \in \mathbb{R}^n$
Eingabe: *Parameter* $m \in (0,1)$
Ausgabe: *ein globaler Minimierer von* (19.1)

Setze $k := 0$
Setze $y^{(0)} := x^{(0)}$
Bestimme ein $s^{(0)} \in \partial f(y^{(0)})$
Setze $J^{(0)} := \{0\}$
Setze $S^{(0)} := \left[s^{(j)}\right]_{j \in J^{(0)}}$
Setze $\alpha^{(0)} := 0$
repeat
    Bestimme eine Lösung $\lambda^{(k)}$ *der Aufgabe*

$$\begin{aligned} \text{Minimiere} \quad & (\alpha^{(k)})^\mathsf{T}\lambda + \frac{1}{2}\|S^{(k)}\lambda\|^2 \quad \text{über } \lambda \in \mathbb{R}^{|J^{(k)}|} \\ \text{unter} \quad & \lambda \in \Delta \end{aligned} \tag{19.24}$$

    Setze $g^{(k)} := S^{(k)}\lambda^{(k)}$
    Setze $d^{(k)} := -g^{(k)}$
    Setze $\varepsilon^{(k)} := (\alpha^{(k)})^\mathsf{T}\lambda^{(k)}$
    Setze $\xi^{(k)} := -\|g^{(k)}\|^2 - \varepsilon^{(k)}$
    if $\xi^{(k)} < 0$ ***then***
        Setze $y^{(k+1)} := x^{(k)} + d^{(k)}$
        if $f(y^{(k+1)}) \le f(x^{(k)}) + m\,\xi^{(k)}$ ***then***
            Setze $x^{(k+1)} := y^{(k+1)}$ // *wesentlicher Schritt*
        else
            Setze $x^{(k+1)} := x^{(k)}$ // *Nullschritt*
        end if
        Bestimme ein $s^{(k+1)} \in \partial f(y^{(k+1)})$
        Setze $\overline{J}^{(k)} := \left\{j \in J^{(k)} \,\middle|\, \lambda_j^{(k)} > 0\right\}$
        Setze $J^{(k+1)} := \overline{J}^{(k)} \cup \{k+1\}$
        Setze $S^{(k)} := \left[s^{(j)}\right]_{j \in J^{(k)}}$
        Setze $\alpha_j^{(k+1)} := f(x^{(k+1)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k+1)} - y^{(j)})$ *für* $j \in J^{(k+1)}$
        Setze $k := k+1$
    end if
until $\xi^{(k)} = 0$

Bemerkung 19.12 (zu Algorithmus 19.11)**.**

(*i*) *Wie oben motiviert baut das Bundle-Verfahren implizit innere Approximationen* $G_{\varepsilon^{(k)}}(x^{(k)})$ *wie in* (19.21) *aus dem Bündel der aktuell verwendeten Subgradienten* $\{s^{(j)} \mid j \in J^{(k)}\}$ *auf.*

(*ii*) *Das Teilproblem* (19.24) *entspricht der oben besprochenen Aufgabe* (19.15)*, d. h., der Projektion der Null auf die Menge* $G_{\varepsilon^{(k)}}(x^{(k)})$ *mit implizit gegebenem* ε *und* $\tau = 1$.

(*iii*) *In Zeile 22 wird der neue Subgradient ins Bündel aufgenommen, und die nicht verwendeten Subgradienten werden ein für alle Mal aus dem Bündel entfernt.*

(*iv*) *Für die Lösung des konvexen QPs* (19.24) *über dem Einheitssimplex gibt es maßgeschneiderte Lösungsverfahren. Die Dimension dieses QPs entspricht der Anzahl der Subgradienten im aktuellen Bündel.*

Zur Durchführung des Verfahrens aus Algorithmus 19.11 werden folgende problemspezifische Routinen benötigt:

(1) Routine zur Auswertung der Zielfunktion $f(x)$.

(2) Routine, die einen beliebigen Subgradienten $s \in \partial f(x)$ bestimmt.

Wir analysieren jetzt noch die Konvergenz des Verfahrens. Dies erfordert einige Hilfsresultate, bis wir mit Satz 19.20 schließlich das Hauptresultat erhalten. Es bezeichnen jeweils $\cdot^{(k)}$ die durch den Algorithmus 19.11 erzeugten Folgen.

Wir beginnen mit einem einfachen Resultat über die Zugehörigkeit zu gewissen ε-Subdifferentialen.

Lemma 19.13 (erzeugte ε-Subgradienten, vgl. Geiger, Kanzow, 2002, Lemma 6.73)**.**

(*i*) $\alpha_j^{(k)} \geq 0$ *und* $s^{(j)} \in \partial_{\alpha_j^{(k)}} f(x^{(k)})$ *für alle* $j \in J^{(k)}$ *und* $k \in \mathbb{N}_0$.

(*ii*) $\varepsilon^{(k)} \geq 0$ *und* $g^{(k)} \in \partial_{\varepsilon^{(k)}} f(x^{(k)})$ *für alle* $k \in \mathbb{N}_0$.

Beweis. Wir erinnern zunächst an die Definition des Linearisierungsfehlers aus Zeile 24:

$$\alpha_j^{(k)} := f(x^{(k)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k)} - y^{(j)}), \tag{19.25}$$

vgl. (19.8). Die Eigenschaft $\alpha_j^{(k)} \geq 0$ wurde in (19.8) gezeigt, wir wählen einfach den Referenzpunkt $\overline{x} = x^{(k)}$ und beachten, dass die Subgradienten $s^{(j)}$ zu den Punkten $y^{(j)}$ gehören.

Für $k = 0$ gilt

$$s^{(0)} \in \partial f(x^{(0)}) = \partial_0 f(x^{(0)}) = \partial_{\alpha_0^{(0)}} f(x^{(0)}),$$

d. h., die Aussage (*i*) gilt für $k = 0$, da $J^{(0)} = \{0\}$ ist.

Für $k \geq 0$ folgt aus der Definition (19.25)

$$\alpha_j^{(k+1)} := f(x^{(k+1)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k+1)} - y^{(j)})$$

und aus $s^{(j)} \in \partial f(y^{(j)})$ die Ungleichung

$$\begin{aligned} f(x) &\geq f(y^{(j)}) + (s^{(j)})^\mathsf{T}(x - y^{(j)}) \\ &= f(x^{(k+1)}) + (s^{(j)})^\mathsf{T}(x - x^{(k+1)}) - \left[f(x^{(k+1)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k+1)} - y^{(j)})\right] \\ &= f(x^{(k+1)}) + (s^{(j)})^\mathsf{T}(x - x^{(k+1)}) - \alpha_j^{(k+1)} \end{aligned}$$

für alle $x \in \mathbb{R}^n$. Das heißt aber $s^{(j)} \in \partial_{\alpha_j^{(k+1)}} f(x^{(k+1)})$ für $j \in J^{(j)}$. Damit ist Aussage (i) gezeigt.

Zu Aussage (ii): Wegen der Nebenbedingungen $\lambda^{(k)} \geq 0$, der gerade gezeigten Aussage $\alpha^{(k)} \geq 0$ sowie der Definition $\varepsilon^{(k)} := (\alpha^{(k)})^\mathsf{T}\lambda^{(k)}$ folgt $\varepsilon^{(k)} \geq 0$. Weiterhin folgt wegen $s^{(j)} \in \partial_{\alpha_j^{(k)}} f(x^{(k)})$ für $j \in J^{(k)}$:

$$f(x) \geq f(x^{(k)}) + (s^{(j)})^\mathsf{T}(x - x^{(k)}) - \alpha_j^{(k)} \quad \text{für alle } j \in J^{(k)}$$

und alle $x \in \mathbb{R}^n$ und $k \in \mathbb{N}_0$. Unter Verwendung der Definitionen von $g^{(k)}$ und $\varepsilon^{(k)}$ ergibt sich weiterhin

$$\begin{aligned} f(x) &= \sum_{j \in J^{(k)}} \lambda_j^{(k)} f(x) \\ &\geq \sum_{j \in J^{(k)}} \lambda_j^{(k)} \left[f(x^{(k)}) + (s^{(j)})^\mathsf{T}(x - x^{(k)}) - \alpha_j^{(k)}\right] \\ &= f(x^{(k)}) + \sum_{j \in J^{(k)}} \left[(\lambda_j^{(k)} s^{(j)})^\mathsf{T}(x - x^{(k)})\right] - \sum_{j \in J^{(k)}} \lambda_j^{(k)} \alpha_j^{(k)} \\ &= f(x^{(k)}) + (g^{(k)})^\mathsf{T}(x - x^{(k)}) - \varepsilon^{(k)} \end{aligned}$$

für alle $x \in \mathbb{R}^n$ und $k \in \mathbb{N}_0$. Das heißt aber $g^{(k)} \in \partial_{\varepsilon^{(k)}} f(x^{(k)})$. □

Die Abbruchbedingung $\xi^{(k)} = 0$ in Algorithmus 19.11 wird durch folgendes Resultat motiviert:

Lemma 19.14 (Interpretation der Abbruchbedingung, vgl. Geiger, Kanzow, 2002, Lemma 6.74)**.**

(i) $\xi^{(k)} \leq 0$ *für alle* $k \in \mathbb{N}_0$.

(ii) *Ist* $\xi^{(k)} = 0$, *so ist* $x^{(k)}$ *ein Minimierer von* (19.1).

Beweis. Die Aussage (i) folgt sofort aus der Definition $\xi^{(k)} := -\|g^{(k)}\|^2 - \varepsilon^{(k)}$ im Verfahren und $\varepsilon^{(k)} \geq 0$, siehe Lemma 19.13. Im Fall von $\xi^{(k)}$ sind $g^{(k)} = 0$ und $\varepsilon^{(k)} = 0$. Aufgrund von Aussage (ii) in Lemma 19.13 gilt also $0 \in \partial_0 f(x^{(k)}) = \partial f(x^{(k)}$, d. h., $x^{(k)}$ ist ein globaler Minimierer von (19.1). □

Lemma 19.15 (Eigenschaften der Iterierten, vgl. Geiger, Kanzow, 2002, Lemma 6.75)**.** *Es gelte* $f(x^{(k)}) \geq \underline{f}$ *für alle* $k \in \mathbb{N}_0$. *Dann gelten:*

(i)

$$f(x^{(k)}) - f(x^{(k+1)}) \to 0 \quad \text{für } k \to \infty.$$

(*ii*)
$$\sum_{k=0}^{\infty} t^{(k)} \left(\|g^{(k)}\|^2 + \varepsilon^{(k)} \right) \leq \left(f(x^{(0)}) - \underline{f} \right) / m.$$

Dabei ist $t^{(k)} := 0$ *für Nullschritte und* $t^{(k)} := 1$ *für wesentliche Schritte.*

(*iii*) *Falls es unendlich viele wesentliche Schritt gibt und wir die entsprechende Teilfolge von Indizes mit* $\left(k^{(\ell)}\right)$ *bezeichnen, dann gilt* $g^{(k^{(\ell)})} \to 0$ *und* $\varepsilon^{(k^{(\ell)})} \to 0$ *für* $\ell \to \infty$.

Beweis. Aussage (*i*): Per Konstruktion ist die Folge $(f(x^{(k)}))$ monoton fallend. (**Quizfrage:** Details?) Da sie nach Voraussetzung nach unten beschränkt ist, konvergiert sie. Damit konvergiert die Folge der Differenzen $f(x^{(k)}) - f(x^{(k+1)})$ gegen Null.

Aussage (*ii*): Aus Zeile 15 im Algorithmus 19.11, also der Entscheidung, ob ein wesentlicher oder ein Nullschritt durchgeführt wird, ergibt sich unter Berücksichtigung der Definition von $t^{(k)}$

$$f(x^{(k+1)}) \leq f(x^{(k)}) + m\, t^{(k)}\, \xi^{(k)}$$

für alle $k \in \mathbb{N}_0$, also

$$f(x^{(k)}) - f(x^{(k+1)}) \geq -m\, t^{(k)}\, \xi^{(k)}.$$

Durch Aufsummieren erhalten wir

$$f(x^{(0)}) - \underline{f} \geq f(x^{(0)}) - f(x^{(k)}) \geq -m \sum_{j=0}^{k-1} t^{(j)}\, \xi^{(j)}$$

und im Grenzübergang

$$f(x^{(0)}) - \underline{f} \geq -m \sum_{j=0}^{\infty} t^{(j)}\, \xi^{(j)} = m \sum_{j=0}^{\infty} t^{(j)} \left(\|g^{(j)}\|^2 + \varepsilon^{(j)} \right).$$

Die Division durch $m \in (0,1)$ ergibt die Behauptung.

Aussage (*iii*): Die wesentlichen Schritte sind genau die mit $t^{(k)} = 1$. Ist dies für unendlich viele Indizes, die die Teilfolge $\left(k^{(\ell)}\right)$ bilden, der Fall, dann folgt aus der gerade gezeigten Summierbarkeit von

$$\sum_{k=0}^{\infty} t^{(k)} \left(\|g^{(k)}\|^2 + \varepsilon^{(k)} \right) = \sum_{\ell=0}^{\infty} \underbrace{t^{(k^{(\ell)})}}_{=1} \left(\|g^{(k^{(\ell)})}\|^2 + \varepsilon^{(k^{(\ell)})} \right),$$

dass notwendigerweise $g^{(k^{(\ell)})} \to 0$ und $\varepsilon^{(k^{(\ell)})} \to 0$ gelten. □

Lemma 19.16 (unendlich viele wesentliche Schritte, vgl. Geiger, Kanzow, 2002, Lemma 6.76)**.** *Es gebe unendlich viele wesentliche Schritte in der Folge* $\left(x^{(k)}\right)$. *Dann ist jeder Häufungspunkt ein Minimierer von* (19.1).

Beweis. Es sei x^* ein Häufungspunkt der Folge $\left(x^{(k)}\right)$. Da $\left(f(x^{(k)})\right)$ monoton fallend ist und auf einer Teilfolge gegen $f(x^*)$ konvergiert, gilt $f(x^{(k)}) \geq f(x^*) =: f^*$ für alle $k \in \mathbb{N}_0$ und $f(x^{(k)}) \to f(x^*)$.

Nach Lemma 19.13 *(ii)* gilt weiter

$$g^{(k)} \in \partial_{\varepsilon^{(k)}} f(x^{(k)}). \tag{19.26}$$

Da x^* ein Häufungspunkt der Folge $\left(x^{(k)}\right)$ ist und sich $x^{(k)}$ in einem Nullschritt nicht ändert, ist x^* ebenfalls ein Häufungspunkt der Teilfolge $\left(x^{k^{(\ell)}}\right)$ der wesentlichen Schritte, also der Grenzwert einer Teilfolge von $\left(x^{k^{(\ell)}}\right)$. Um Dreifach-Indizierung zu vermeiden, bezeichnen wir diese einfach weiter mit $\left(x^{k^{(\ell)}}\right)$. Es gilt also

$$x^{(k^{(\ell)})} \to x^* \quad \text{für } \ell \to \infty. \tag{19.27}$$

Wegen $f(x^{(k)}) \geq f^*$ für alle $k \in \mathbb{N}_0$ gilt nach Lemma 19.15 *(iii)*

$$g^{(k^{(\ell)})} \to 0 \quad \text{und} \quad \varepsilon^{(k^{(\ell)})} \to 0 \quad \text{für } \ell \to \infty. \tag{19.28}$$

Die ε-Subgradientenungleichung für (19.26) ergibt

$$f(x) \geq f(x^{(k^{(\ell)})}) + \left(g^{(k^{(\ell)})}\right)^\mathsf{T}(x - x^{(k^{(\ell)})}) - \varepsilon^{(k^{(\ell)})}$$

für alle $x \in \mathbb{R}^n$ und $\ell \in \mathbb{N}$. Der Grenzübergang $\ell \to \infty$ zusammen mit der Stetigkeit von f (Satz 16.18), (19.27) und (19.28) zeigt nun

$$f(x) \geq f(x^*) + 0 - 0$$

für alle $x \in \mathbb{R}^n$, d. h., x^* ist ein globalerer Minimerer von (19.1). □

Lemma 19.17 (nur endliche viele wesentliche Schritte, vgl. Geiger, Kanzow, 2002, Lemma 6.77)**.** *Die in Algorithmus 19.11 erzeugte Folge beinhalte nur* endlich *viele wesentliche Schritte, also gilt $x^{(k)} = x^{(k^*)}$ für alle $k \geq k^*$. Dann ist $x^* := x^{(k^*)}$ ein Minimierer von (19.1).*

Beweis. Nach Voraussetzung gilt $x^{(k+1)} = x^{(k)}$ für alle $k \geq k_0$. Wir haben also

$$\begin{aligned} \alpha_j^{(k+1)} &= f(x^{(k+1)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k+1)} - y^{(j)}) \quad \text{für } j \in J^{(k+1)}, \\ \alpha_j^{(k)} &= f(x^{(k)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k)} - y^{(j)}) \qquad \text{für } j \in J^{(k)}. \end{aligned}$$

Wegen $\overline{J}^{(k)} \subseteq J^{(k)}$ und $J^{(k+1)} = \overline{J}^{(k)} \cup \{k+1\}$ ist die Schnittmenge beider Indexmengen gerade $\overline{J}^{(k)}$. Es gilt also

$$\alpha_j^{(k+1)} = \alpha_j^{(k)} \quad \text{für } j \in \overline{J}^{(k)}.$$

Es folgt

$$\varepsilon^{(k)} = \sum_{j \in J^{(k)}} \alpha_j^{(k)} \lambda_j^{(k)} = \sum_{j \in \overline{J}^{(k)}} \alpha_j^{(k)} \lambda_j^{(k)} = \sum_{j \in \overline{J}^{(k)}} \alpha_j^{(k+1)} \lambda_j^{(k)} =: \sigma^{(k)} \tag{19.29}$$

für alle $k \geq k_0$.

Es sei nun $\mu \in [0,1]$ beliebig. Wir definieren den Vektor $\overline{\lambda}$ mit Komponenten $\overline{\lambda}_j$ für $j \in J^{(k)} = \overline{J}^{(k-1)} \cup \{k\}$ (disjunkte Vereinigung) gemäß

$$\overline{\lambda}_j := \begin{cases} \mu, & \text{falls } j = k, \\ (1-\mu)\,\lambda_j^{(k-1)}, & \text{falls } j \in \overline{J}^{(k-1)}. \end{cases}$$

Dabei sind die $\lambda_j^{(k-1)}$ für $j \in \overline{J}^{(k-1)}$ gerade die echt positiven Komponenten der Lösung $\lambda^{(k-1)}$ von (19.24) in der Iteration $k-1$. Dann ist $\overline{\lambda}$ zulässig für (19.24) in der Iteration k (**Quizfrage:** Begründung?)

Wir bezeichnen zur Abkürzung die Zielfunktion von (19.24) in Iteration k mit

$$q^{(k)}(\lambda) := (\alpha^{(k)})^\mathsf{T}\lambda + \frac{1}{2}\|S^{(k)}\lambda\|^2 = \sum_{j\in J^{(k)}} \alpha_j^{(k)}\lambda_j + \frac{1}{2}\Big\| \sum_{j\in J^{(k)}} \lambda_j\, s^{(j)}\Big\|^2.$$

Da $\lambda^{(k)}$ ein globaler Minimierer von (19.24) ist, gilt

$$q^{(k)}(\lambda^{(k)}) \le q^{(k)}(\overline{\lambda}).$$

Wir werten nun die Terme in der rechten Seite aus:

$$\begin{aligned}
\sum_{j\in J^{(k)}} \overline{\lambda}_j\, s^{(j)} &= \overline{\lambda}_k\, s^{(k)} + \sum_{j\in\overline{J}^{(k-1)}} \overline{\lambda}_j\, s^{(j)}, && \text{da } J^{(k)} = \overline{J}^{(k-1)}\cup\{k\} \\
&= \mu\, s^{(k)} + \sum_{j\in\overline{J}^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)}\, s^{(j)} && \text{nach Definition von } \overline{\lambda} \\
&= \mu\, s^{(k)} + \sum_{j\in J^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)}\, s^{(j)}, && \text{da } \lambda_j^{(k-1)} = 0 \text{ für } j \in J^{(k-1)} \setminus \overline{J}^{(k-1)} \\
&= \mu\, s^{(k)} + (1-\mu)\, g^{(k-1)} && \text{nach Definition von } g^{(k-1)}
\end{aligned}$$

und

$$\begin{aligned}
\sum_{j\in J^{(k)}} \alpha_j^{(k)}\overline{\lambda}_j &= \alpha_k^{(k)}\,\overline{\lambda}_k + \sum_{j\in\overline{J}^{(k-1)}} \alpha_j^{(k)}\,\overline{\lambda}_j, && \text{da } J^{(k)} = \overline{J}^{(k-1)}\cup\{k\} \\
&= \mu\,\alpha_k^{(k)} + \sum_{j\in\overline{J}^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)}\,\alpha_j^{(k)} && \text{nach Definition von } \overline{\lambda} \\
&= \mu\,\alpha_k^{(k)} + \sum_{j\in J^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)}\,\alpha_j^{(k)}, && \text{da } \lambda_j^{(k-1)} = 0 \text{ für } j \in J^{(k-1)} \setminus \overline{J}^{(k-1)} \\
&= \mu\,\alpha_k^{(k)} + (1-\mu)\,\sigma^{(k-1)} && \text{nach Definition (19.29) von } \sigma^{(k-1)}.
\end{aligned}$$

Damit ergibt sich

$$q^{(k)}(\lambda^{(k)}) \le q^{(k)}(\overline{\lambda}) = \mu\,\alpha_k^{(k)} + (1-\mu)\,\sigma^{(k-1)} + \frac{1}{2}\big\|\mu\, s^{(k)} + (1-\mu)\, g^{(k-1)}\big\|^2 =: h(\mu).$$

Da dies für alle $\mu \in [0,1]$ gilt, haben wir sogar

$$q^{(k)}(\lambda^{(k)}) \le \min\{h(\mu) \mid \mu \in [0,1]\}.$$

Beachte: Es ist $h(\mu) \geq 0$ für alle $\mu \in [0,1]$, da $\alpha_k^{(k)} = 0$ und $\sigma^{(k-1)} = \varepsilon^{(k-1)} \geq 0$ nach (19.29) und Lemma 19.13 gilt.

Es sei $\mu^{(k)}$ die eindeutige Lösung dieser Aufgabe (**Quizfrage:** Warum ist diese eindeutig?) mit Optimalwert $\gamma^{(k)} := h(\mu^{(k)}) \geq 0$. Dann haben wir also

$$q^{(k)}(\lambda^{(k)}) \leq \gamma^{(k)} = h(\mu^{(k)}) \leq h(\mu) \quad \text{für alle } \mu \in [0,1].$$

Zusammen mit (19.29) erhalten wir für alle $k > k_0$:

$$\begin{aligned}
\gamma^{(k)} &\leq h(0) \\
&= \sigma^{(k-1)} + \frac{1}{2}\left\|g^{(k-1)}\right\|^2 \\
&= \sum_{j \in J^{(k-1)}} \alpha_j^{(k-1)} \lambda_j^{(k-1)} + \frac{1}{2}\left\| \sum_{j \in J^{(k-1)}} \lambda_j^{(k-1)} s^{(j)} \right\|^2 && \text{nach (19.29) und Definition von } g^{(k-1)} \\
&= q^{(k-1)}(\lambda^{(k-1)}) \\
&\leq \gamma^{(k-1)}. && (19.30)
\end{aligned}$$

Es gilt also

$$0 \leq \gamma^{(k)} \leq \gamma^{(k-1)} \leq \gamma^{(k_0)} \quad \text{für alle } k > k_0$$

und somit

$$\frac{1}{2}\|g^{(k)}\|^2 \leq \gamma^{(k_0)} \quad \text{und} \quad \sigma^{(k)} \leq \gamma^{(k_0)} \quad \text{für alle } k \geq k_0. \tag{19.31}$$

Mit

$$\frac{1}{2}\|g^{(k-1)}\|^2 \leq \frac{1}{2}\|g^{(k-1)}\|^2 + \sigma^{(k-1)} \leq \gamma^{(k-1)}, \tag{19.32}$$

siehe (19.30), erhalten wir

$$\begin{aligned}
h(\mu) &= \mu\,\alpha_k^{(k)} + (1-\mu)\,\sigma^{(k-1)} + \frac{1}{2}\left\|\mu\, s^{(k)} + (1-\mu)\, g^{(k-1)}\right\|^2 && \text{nach Definition von } h \\
&= \mu\,\alpha_k^{(k)} + (1-\mu)\,\sigma^{(k-1)} \\
&\quad + \frac{\mu^2}{2}\left\|s^{(k)} - g^{(k-1)}\right\|^2 + \frac{2\mu}{2}\left(s^{(k)} - g^{(k-1)}\right)^{\mathsf{T}} g^{(k-1)} + \frac{1}{2}\left\|g^{(k-1)}\right\|^2 \\
&\leq \mu\left(\alpha_k^{(k)} - \sigma^{(k-1)}\right) + \gamma^{(k-1)} \\
&\quad + \frac{\mu^2}{2}\left\|s^{(k)} - g^{(k-1)}\right\|^2 + \mu\left(s^{(k)}\right)^{\mathsf{T}} g^{(k-1)} - \mu\left\|g^{(k-1)}\right\|^2 && \text{nach (19.32)}
\end{aligned}$$

für alle $k > k_0$.

Aus $x^{(k)} = x^{(k_0)}$ für alle $k \geq k_0$, der Beziehung

$$\begin{aligned}
\alpha_k^{(k)} &= f(x^{(k)}) - f(y^{(k)}) - (s^{(k)})^{\mathsf{T}}(x^{(k)} - y^{(k)}) \\
&= f(x^{(k)}) - f(y^{(k)}) - (s^{(k)})^{\mathsf{T}}(x^{(k)} - x^{(k-1)} - d^{(k)}) \\
&= f(x^{(k-1)}) - f(y^{(k)}) + (s^{(k)})^{\mathsf{T}} d^{(k)}
\end{aligned}$$

für alle $k > k_0$ sowie $f(y^{(k)}) > f(x^{(k-1)}) + m\,\xi^{(k-1)}$ für alle $k > k_0$ (der Bedingung für einen Nullschritt) folgt

$$\begin{aligned} -\alpha_k^{(k)} + (s^{(k)})^\mathsf{T} d^{(k-1)} &= f(y^{(k)}) - f(x^{(k-1)}) \\ &> m\,\xi^{(k-1)} = -m\left(\|g^{(k-1)}\|^2 + \sigma^{(k-1)}\right) \quad \text{nach Definition von } \xi^{(k-1)} \text{ und (19.29)} \end{aligned}$$

für alle $k > k_0$. Mit $d^{(k-1)} = -g^{(k-1)}$ können wir die Ungleichung umformen zu

$$(s^{(k)})^\mathsf{T} g^{(k-1)} < m\left(\|g^{(k-1)}\|^2 + \sigma^{(k-1)}\right) - \alpha_k^{(k)}.$$

Wir erhalten somit für alle $k > k_0$

$$\begin{aligned} h(\mu) &\le \mu\left(\alpha_k^{(k)} - \sigma^{(k-1)}\right) + \gamma^{(k-1)} + \frac{\mu^2}{2}\left\|s^{(k)} - g^{(k-1)}\right\|^2 + \mu\left(s^{(k)}\right)^\mathsf{T} g^{(k-1)} - \mu\left\|g^{(k-1)}\right\|^2 \\ &\le \mu\left(\alpha_k^{(k)} - \sigma^{(k-1)}\right) + \gamma^{(k-1)} + \frac{\mu^2}{2}\left\|s^{(k)} - g^{(k-1)}\right\|^2 + \mu\, m\left(\|g^{(k-1)}\|^2 + \sigma^{(k-1)}\right) - \mu\,\alpha_k^{(k)} - \mu\left\|g^{(k-1)}\right\|^2 \\ &= \gamma^{(k-1)} + \frac{\mu^2}{2}\left\|s^{(k)} - g^{(k-1)}\right\|^2 - \mu\,(1-m)\,\sigma^{(k-1)} - \mu\,(1-m)\left\|g^{(k-1)}\right\|^2. \end{aligned} \tag{19.33}$$

Nach (19.31) sind die Folgen $(\sigma^{(k)})$ und $(g^{(k)})$ beschränkt und daher auch $(d^{(k)})$ und $(y^{(k)})$. Wegen $s^{(k)} \in \partial f(y^{(k)})$ und Satz 16.23 ist auch $(s^{(k)})$ beschränkt. Wir halten also fest, dass es eine Konstante c gibt, für die wir o. B. d. A. als $c \ge 1/2$ annehmen können, sodass

$$\|g^{(k)}\| \le c, \quad \|s^{(k)}\| \le c \quad \text{und} \quad \sigma^{(k)} \le c$$

für alle $k \ge k_0$ gilt. Das zeigt

$$\|s^{(k)} - g^{(k-1)}\|^2 \le \left(\|s^{(k)}\| + \|g^{(k-1)}\|\right)^2 \le 4\,c^2,$$

und Einsetzen in (19.33) ergibt

$$\begin{aligned} h(\mu) &\le \gamma^{(k-1)} + \frac{\mu^2}{2}\left\|s^{(k)} - g^{(k-1)}\right\|^2 - \mu\,(1-m)\,\sigma^{(k-1)} - \mu\,(1-m)\left\|g^{(k-1)}\right\|^2 \\ &\le \gamma^{(k-1)} + 2\,c^2\mu^2 - \mu\,(1-m)\,\sigma^{(k-1)} - \mu\,(1-m)\left\|g^{(k-1)}\right\|^2 \\ &= 2\,c^2\mu^2 - (1-m)\left[\sigma^{(k-1)} + \left\|g^{(k-1)}\right\|^2\right]\mu + \gamma^{(k-1)} =: \theta(\mu) \end{aligned}$$

für alle $k > k_0$. Die Minimalwerte von θ über $\mathbb{R}$ und über $[0,1]$ stimmen überein und sind gleich (Minimalstelle ausrechnen) $\theta^* = \gamma^{(k-1)} - (1-m)^2\left[\sigma^{(k-1)} + \left\|g^{(k-1)}\right\|^2\right]/(8\,c^2)$.

Wir fassen zusammen:

$$\begin{aligned} \gamma^{(k)} &= h(\mu^{(k)}) && \text{nach Definition von } \gamma^{(k)} \\ &= \min\{h(\mu) \mid \mu \in [0,1]\} \\ &\le \min\{\theta(\mu) \mid \mu \in [0,1]\} \\ &= \theta^* \\ &= \gamma^{(k-1)} - \frac{(1-m)^2}{(8\,c^2)}\left[\sigma^{(k-1)} + \left\|g^{(k-1)}\right\|^2\right]. \end{aligned}$$

Das Aufsummieren dieser Ungleichung für $j = k_0 + 1, \dots, k + 1$ liefert

$$\frac{(1-m)^2}{8\,c^2} \sum_{j=k_0}^{k} \left(\|g^{(k-1)}\|^2 + \sigma^{(k-1)} \right) \leq \gamma^{k_0} - \gamma^{(k+1)}.$$

Mit $\gamma^{(k+1)}$ folgt daraus

$$\sum_{j=k_0}^{\infty} \left(\sigma^{(k-1)} + \|g^{(k-1)}\|^2 \right) < \infty.$$

Daraus folgt nun schließlich $g^{(k)} \to 0$ und $\varepsilon^{(k)} = \sigma^{(k)} \to 0$ für $k \to \infty$. Eine solche Situation hatten wir schon im Beweis von Lemma 19.16. Wie dort folgt, dass x^* ein Minimierer von (19.1) ist. □

Aus Lemma 19.16 und Lemma 19.17 erhalten wir sofort die folgende vorläufige Konvergenzaussage:

Satz 19.18 (Häufungspunkte sind Minimierer, vgl. Geiger, Kanzow, 2002, Satz 6.78)**.** *Jeder Häufungspunkt einer von Algorithmus 19.11 erzeugten Folge $\left(x^{(k)}\right)$ ist ein globaler Minimierer von (19.1).*

Um das Ergebnis noch zu verbessern, wollen wir zeigen, dass solche Häufungspunkte bereits unter einer schwachen Voraussetzung existieren.

Lemma 19.19. *Die Aufgabe (19.1) besitze mindestens einen globalen Minimierer. Ist x^* einer der Minimierer und $\left(x^{(k)}\right)$ eine von Algorithmus 19.11 erzeugte Folge, dann gelten:*

(i)

$$\|x^{(k)} - x^*\|^2 \leq \|x^{(m)} - x^*\|^2 + \sum_{j=m}^{k-1} \left(\|x^{(j+1)} - x^{(j)}\|^2 + 2\,t^{(j)}\,\varepsilon^{(j)} \right) \quad \text{für alle } m \in \mathbb{N}_0 \text{ und alle } k \geq m.$$

(ii)

$$\sum_{j=0}^{\infty} \left(\|x^{(j+1)} - x^{(j)}\|^2 + 2\,t^{(j)}\,\varepsilon^{(j)} \right) \text{ ist endlich.}$$

(iii) Die Folge $\left(x^{(k)}\right)$ ist beschränkt.

Beweis. Aussage *(i)*: Wir nutzen $g^{(k)} \in \partial_{\varepsilon^{(k)}} f(x^{(k)})$ (Lemma 19.13) und die Voraussetzung $f(x^{(k)}) \geq f(x^*)$ und erhalten

$$0 \geq f(x^*) - f(x^{(k)}) \geq (g^{(k)})^{\mathsf{T}}(x^* - x^{(k)}) - \varepsilon^{(k)}$$

und somit

$$(g^{(k)})^{\mathsf{T}}(x^* - x^{(k)}) \leq \varepsilon^{(k)}.$$

Mit $x^{(k+1)} - x^{(k)} = t^{(k)} d^{(k)} = -t^{(k)} g^{(k)}$ und $t^{(k)} \geq 0$ erhalten wir

$$-(x^* - x^{(k)})^{\mathsf{T}}(x^{(k+1)} - x^{(k)}) \leq t^{(k)} \varepsilon^{(k)} \quad \text{für alle } k \in \mathbb{N}.$$

Das impliziert

$$\begin{aligned}\|x^* - x^{(k+1)}\|^2 &= \|x^* - x^{(k)} + x^{(k)} - x^{(k+1)}\|^2 \\ &= \|x^* - x^{(k)}\|^2 - 2\,(x^* - x^{(k)})^\mathsf{T}(x^{(k+1)} - x^{(k)}) + \|x^{(k+1)} - x^{(k)}\|^2 \\ &\leq \|x^* - x^{(k)}\|^2 + \|x^{(k+1)} - x^{(k)}\|^2 + 2\,t^{(k)}\varepsilon^{(k)}.\end{aligned}$$

Durch Summation erhalten wir die Aussage.

Aussage (*ii*): Die Aussage

$$\|x^{(k+1)} - x^{(k)}\|^2 = (t^{(k)})^2\|d^{(k)}\|^2 = (t^{(k)})^2\|g^{(k)}\|^2 \leq 2\,(t^{(k)})^2\|g^{(k)}\|^2$$

und $t^{(k)} \in [0,1]$ ergibt zusammen mit Lemma 19.15

$$\begin{aligned}\sum_{j=0}^{\infty}\left(\|x^{(j+1)} - x^{(j)}\|^2 + 2\,t^{(j)}\,\varepsilon^{(j)}\right) &\leq 2\sum_{j=0}^{\infty}\left(2\,(t^{(j)})^2\|g^{(j)}\|^2 + t^{(j)}\,\varepsilon^{(j)}\right) \\ &\leq 2\sum_{j=0}^{\infty}\left(t^{(j)}\|g^{(j)}\|^2 + t^{(j)}\,\varepsilon^{(j)}\right) \\ &< \infty\end{aligned}$$

Aussage (*iii*): Die Behauptung folgt aus Aussagen (*i*) und (*ii*). □

Es folgt nun unser Hauptergebnis zur Konvergenz des Bundle-Verfahrens aus Algorithmus 19.11.

Satz 19.20 (Konvergenzsatz für Algorithmus 19.11)**.** *Die Aufgabe* (19.1) *besitze mindestens einen globalen Minimierer. Dann konvergiert jede von Algorithmus 19.11 erzeugte Folge* $\left(x^{(k)}\right)$ *gegen einen globaler Minimierer von* (19.1).

Beweis. Nach Lemma 19.19 ist die Folge $\left(x^{(k)}\right)$ beschränkt. Es existiert also mindestens ein Häufungspunkt x^*. Nach Satz 19.18 ist dieser ein globaler Minimierer von f. Es bleibt zu zeigen, dass die gesamte Folge $\left(x^{(k)}\right)$ gegen x^* konvergiert.

Es sei $\varepsilon > 0$. Da $\left(x^{(k)}\right)$ auf einer Teilfolge gegen x^* konvergiert und die Reihe aus Lemma 19.19 (*ii*) konvergiert, gibt es ein $m \in \mathbb{N}$ mit

$$\|x^{(m)} - x^*\| \leq \frac{\varepsilon}{2} \quad \text{und} \quad \sum_{j=m}^{\infty}\left(\|x^{(j+1)} - x^{(j)}\|^2 + 2\,t^{(j)}\,\varepsilon^{(j)}\right) \leq \frac{\varepsilon}{2}.$$

Damit ergibt sich aus Lemma 19.19 (*i*)

$$\|x^{(k)} - x^*\|^2 \leq \|x^{(m)} - x^*\|^2 + \frac{\varepsilon}{2} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

für alle $k \geq m$. Da $\varepsilon > 0$ beliebig war, folgt die Behauptung. □

Ende der Woche 14

Index

Literatur

Armijo, L. (1966). „Minimization of functions having Lipschitz continuous first partial derivatives“. *Pacific Journal of Mathematics* 16.1, S. 1–3. DOI: 10.2140/pjm.1966.16.1.

Blum, E.; W. Oettli (1972). „Direct proof of the existence theorem for quadratic programming“. *Operations Research* 20, S. 165–167. DOI: 10.1287/opre.20.1.165.

Bonnans, F.; C. Gilbert; C. Lemaréchal; C. Sagastizábal (2003). *Numerical Optimization*. 1. Aufl. Berlin: Springer. DOI: 10.1007/978-3-662-05078-1.

Forsgren, A. (2008). *An elementary proof of optimality conditions for linear programming*. TRITA–MAT 2008–OS6. Department of Mathematics, Royal Institute of Technology (KTH) Stockholm.

Frank, M.; P. Wolfe (1956). „An algorithm for quadratic programming“. *Naval Research Logistics Quarterly* 3, S. 95–110. DOI: 10.1002/nav.3800030109.

Gass, S. I.; A. A. Assad (2005). *An Annotated Timeline of Operations Research: An Informal History*. Bd. 75. International Series in Operations Research & Management Science. Boston, MA: Kluwer Academic Publishers.

Geiger, C.; C. Kanzow (1999). *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. New York: Springer. DOI: 10.1007/978-3-642-58582-1.

Geiger, C.; C. Kanzow (2002). *Theorie und Numerik restringierter Optimierungsaufgaben*. New York: Springer. DOI: 10.1007/978-3-642-56004-0.

Gerdts, M.; F. Lempio (2011). *Mathematische Optimierungsverfahren des Operations Research*. de Gruyter. DOI: 10.1515/9783110249989.

Gill, P. E.; W. Murray; M. H. Wright (1981). *Practical Optimization*. London: Academic Press.

Hamacher, H.; B. Klamroth (2006). *Lineare Optimierung und Netzwerkoptimierung*. 2. Aufl. Vieweg. DOI: 10.1007/978-3-8348-9031-3.

Heuser, H. (2002). *Lehrbuch der Analysis. Teil 2*. 12. Aufl. Stuttgart: B.G.Teubner. DOI: 10.1007/978-3-322-96826-5.

Heuser, H. (2003). *Lehrbuch der Analysis. Teil 1*. 15. Aufl. Vieweg+Teubner Verlag. DOI: 10.1007/978-3-322-96828-9.

Jarre, F.; J. Stoer (2004). *Optimierung*. Springer. DOI: 10.1007/978-3-642-18785-8.

Lemke, C. E. (1954). „The dual method of solving the linear programming problem“. *Naval Research Logistics Quarterly* 1, S. 36–47. DOI: 10.1002/nav.3800010107.

Nocedal, J.; S. J. Wright (2006). *Numerical Optimization*. 2. Aufl. New York: Springer. DOI: 10.1007/978-0-387-40065-5.

Phelps, R. R. (1993). *Convex Functions, Monotone Operators and Differentiability*. 2. Aufl. Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-46077-0.

Rockafellar, R. T. (1970). *Convex Analysis*. Bd. 28. Princeton Mathematical Series. Princeton, New Jersey: Princeton University Press. URL: https://www.jstor.org/stable/j.ctt14bs1ff.

Schrijver, A. (2003). *Combinatorial Optimization. Polyhedra and Efficiency. Volume A*. Bd. 24. Algorithms and Combinatorics. Paths, flows, matchings, Chapters 1–38. Springer-Verlag, Berlin, S. xxxviii+647.

Vanderbei, R. J. (2008). *Linear Programming: Foundations and Extensions*. Ooperations Research, Management Science. New York, NY: Springer. DOI: 10.1007/978-0-387-74388-2.

Von Neumann, J. (1928). „Zur Theorie der Gesellschaftsspiele“. *Mathematische Annalen* 100.1, S. 295–320. DOI: 10.1007/bf01448847.

Werner, J. (2007). *Vorlesung über Optimierung*. Lecture Notes, Department of Mathematics, University of Hamburg, Germany. URL: http://num.math.uni-goettingen.de/werner/.