

# VORLESUNGSSKRIPT GRUNDLAGEN DER OPTIMIERUNG

WINTERSEMESTER 2021

Roland Herzog\*

2022-12-11

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

Material für 13 Wochen.

# Inhaltsverzeichnis

o	Einführung	5
§ 1	Grundbegriffe und Klassifikation von Optimierungsaufgaben	5
§ 2	Notation und Wiederholung von Diffbarkeitsbegriffen	10
1	Unrestringierte Optimierung	14
§ 3	Optimalitätsbedingungen der unrestringierten Optimierung	14
§ 4	Das Gradientenverfahren	16
§ 4.1	Vorstellung des Verfahrens	18
§ 4.2	Das Gradientenverfahren in einem alternativen Skalarprodukt	23
§ 4.3	Konvergenz bei quadratischer Zielfunktion und exakter Liniensuche	25
§ 5	Das Newton-Verfahren	32
§ 5.1	Einige Hilfsresultate	33
§ 5.2	Das lokale Newton-Verfahren für die Nullstellenbestimmung $F(x) = 0$	36
§ 5.3	Das lokale Newton-Verfahren in der Optimierung	38
§ 5.4	Ein globalisiertes Newton-Verfahren in der Optimierung	39
2	Lineare Optimierung	42
§ 6	Einführung	42
§ 6.1	Existenz von Lösungen	47
§ 6.2	Die Bedeutung der Ecken	51
§ 7	Simplex-Algorithmus	56
§ 7.1	Der Simplex-Schritt	56
§ 7.2	Der Simplex-Algorithmus	62
§ 8	Optimalitätsbedingungen der linearen Optimierung (Dualität)	66
§ 9	Duales Simplex-Verfahren	74
§ 10	Sensitivitätsanalyse	79
§ 11	Lineare Optimierungsaufgaben auf Graphen	85
§ 12	Ganzzahlige Lösungen	91
3	Konvexe Optimierung	97

---

§ 13	Einführung	97
§ 13.1	Konvexe Mengen	97
§ 13.2	Konvexe Funktionen	101
§ 14	Konvexe Optimierungsaufgaben	112
§ 15	Trennungssätze für konvexe Mengen	114
§ 15.1	Die Aufgabe der orthogonalen Projektion	114
§ 15.2	Affine Unterräume	116
§ 15.3	Topologische Eigenschaften konvexer Mengen	121
§ 15.4	Trennungssätze	129
§ 16	Das Subdifferential und die Richtungsableitung konvexer Funktionen	136
§ 16.1	Das Subdifferential	136

# Kapitel 0 Einführung

## § 1 GRUNDBEGRIFFE UND KLASSIFIKATION VON OPTIMIERUNGSAUFGABEN

Die mathematische Optimierung beschäftigt sich mit Aufgaben der Form

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

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

- die Funktionen  $f, g_i, h_j: \mathbb{R}^n \rightarrow \mathbb{R}$  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

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

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

(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:

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

- (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 1.1:** Welche Eigenschaften haben die in [Abbildung 1.1](#) markierten Punkte?

**Quizfrage 1.2:** Was ist der Unterschied zwischen einem lokalen und einem globalen Minimierer?

**Quizfrage 1.3:** Ist jeder globale Minimierer auch ein lokaler Minimierer? Ist jeder lokale Minimierer auch ein globaler Minimierer?

**Quizfrage 1.4:** Gibt es Optimierungsaufgaben, die einen lokalen Minimierer besitzen, aber keinen globalen?

**Quizfrage 1.5:** Wie definiert man die Begriffe (strikt) globaler Maximierer und (strikt) lokaler Maximierer?

**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?

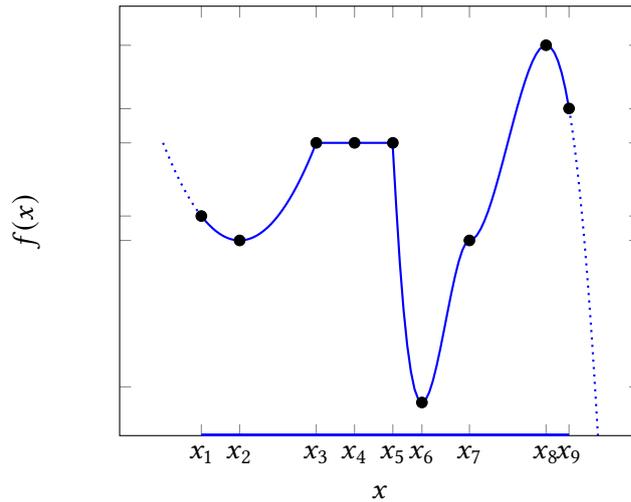


Abbildung 1.1: Illustration der Begriffe aus Definition 1.1 anhand einer Zielfunktion  $f: \mathbb{R} \rightarrow \mathbb{R}$ . Die zulässige Menge ist das auf der  $x$ -Achse markierte Intervall.

- (2) Wie erkennt man sie? ( $\leadsto$  Optimalitätsbedingungen)
- (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**).<sup>1</sup>
- (ii) *Ungleichungsbeschränkungen der besonders einfachen Art*

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

mit  $\ell_i \in \mathbb{R} \cup \{-\infty\}$  und  $u_i \in \mathbb{R} \cup \{\infty\}$  heißen **Box-Beschränkungen** (englisch: **box constraints**, **bound constraints**) mit **oberer Schranke**  $u$  und **unterer Schranke**  $\ell$ .

- (iii) Sind  $f, g$  und  $h$  (affin-)lineare Funktionen von  $x$ , so sprechen wir von **linearer Optimierung**.<sup>2</sup> Eine lineare Optimierungsaufgabe heißt auch **lineares Programm** (englisch: **linear program**, **LP**),

<sup>1</sup>Wir behandeln unrestringierte Aufgaben in Kapitel 1.

<sup>2</sup>Diese werden in Kapitel 2 behandelt.

also z. B.

Minimiere  $c^T x$  sodass  $Ax = b$  und  $x \geq 0$ .

- (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.<sup>3</sup>
- (v) Ist  $f$  ein quadratisches Polynom und sind  $g$  und  $h$  (affin-)linear, so sprechen wir von **quadratischer Optimierung**. Eine quadratische 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, besitzt also einen globalen Minimierer. Ein Kriterium für die Lösbarkeit liefert der Satz von Weierstraß aus der Analysis: „Stetige reellwertige Funktionen nehmen auf kompakten Mengen ihr Minimum (und ihr Maximum) an.“ Damit folgt sofort: Wenn die Zielfunktion  $f: F \rightarrow \mathbb{R}$  stetig und die zulässige Menge  $F$  kompakt ist, dann besitzt die Aufgabe

Minimiere  $f(x)$  über  $x \in F$

mindestens einen globalen Minimierer. Wir wollen diese Voraussetzungen hier in zwei Richtungen abschwächen:

- (1) Es ist bereits ausreichend, dass es ein Niveau  $m \in \mathbb{R}$  gibt, sodass die zugehörige **Sublevelmenge** (englisch: *sublevel set*) von  $f$

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

nichtleer und kompakt ist.

- (2) An Stelle der Stetigkeit von  $f$  wird nur die Unterhalbstetigkeit benötigt:

**Definition 1.4** (Unterhalbstetigkeit). Es sei  $F \subseteq \mathbb{R}^n$  eine nichtleere Menge. Eine Funktion  $f: F \rightarrow \mathbb{R}$  heißt **unterhalbstetig** (auch: **halbstetig von unten**, englisch: **lower semicontinuous**) auf  $F$ , wenn gilt:

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

<sup>3</sup>Diese Aufgaben werden in [Kapitel 3](#) besprochen.

**Lemma 1.5** (Äquivalenz zur Unterhalbstetigkeit).

Es sei  $F \subseteq \mathbb{R}^n$  eine nichtleere Menge. Für eine Funktion  $f: F \rightarrow \mathbb{R}$  sind äquivalent:

- (i)  $f$  ist unterhalbstetig auf  $F$ .
- (ii) Alle Sublevelmengen  $\{x \in F \mid f(x) \leq m\}$ ,  $m \in \mathbb{R}$ , sind abgeschlossen in  $F$ .

*Beweis.* Wir nehmen zunächst **Aussage (i)** an. Es sei  $L := \{x \in F \mid f(x) \leq m\}$  eine Sublevelmenge von  $f$ . Wenn  $L$  leer ist, ist nichts zu zeigen. Andernfalls betrachten wir eine Folge  $(x^{(k)}) \subseteq L$ , die in  $F$  konvergiert, also  $x^{(k)} \rightarrow x^* \in F$ . Dann gilt nach Definition der Unterhalbstetigkeit  $f(x^*) \leq \liminf_{k \rightarrow \infty} f(x^{(k)})$  und weiter  $f(x^{(k)}) \leq m$  wegen  $x^{(k)} \in L$ . Daraus folgt auch  $f(x^*) \leq m$ , d. h.,  $x^* \in L$ . Das zeigt, dass  $L$  in  $F$  abgeschlossen ist.

Umgekehrt gelte **Aussage (ii)**. Wir betrachten eine Folge  $(x^{(k)}) \subseteq F$ ,  $x^{(k)} \rightarrow x^* \in F$ . Wir nehmen das Gegenteil von **Aussage (i)** an, also  $C := \liminf_{k \rightarrow \infty} f(x^{(k)}) < f(x^*)$ . Nach Definition des Limes inferior gibt es eine Teilfolge mit den Indizes  $k^{(\ell)}$ , sodass  $f(x^{(k^{(\ell)})}) \rightarrow C < f(x^*)$  konvergiert. Aufgrund der Annahme gibt es also ein  $\varepsilon > 0$  und ein  $\ell_0 \in \mathbb{N}$ , sodass für alle  $\ell > \ell_0$  die Beziehung  $f(x^{(k^{(\ell)})}) + \varepsilon \leq f(x^*)$  erfüllt ist. Mit anderen Worten: Alle „späten“ Folgenglieder von  $x^{(k^{(\ell)})}$  gehören zur Sublevelmenge mit dem Niveau  $f(x^*) - \varepsilon$ . Nach Voraussetzung ist diese Menge abgeschlossen, also gehört auch der Grenzwert  $x^*$  zu dieser Menge. Das bedeutet aber, dass  $f(x^*) \leq f(x^*) - \varepsilon$  ist – ein Widerspruch.  $\square$

Wir formulieren nun einen allgemeinen Existenzsatz für globale Minimierer.

**Satz 1.6** (Existenz eines globalen Minimierers).

Die zulässige Menge  $F \subseteq \mathbb{R}^n$  sei nichtleer. Weiter sei  $f: F \rightarrow \mathbb{R}$  unterhalbstetig, und für irgendein  $m \in \mathbb{R}$  sei die Sublevelmenge

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

nichtleer und kompakt.<sup>4</sup> Dann besitzt die Aufgabe

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

mimdestens einen globalen Minimierer.

*Beweis.* Wir zeigen zuerst, dass  $f$  auf  $F$  nach unten beschränkt sein muss. Andernfalls gibt es eine Folge  $(x^{(k)}) \subseteq F$  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  $(x^{(k^{(\ell)})}) \subseteq L$  mit der Eigenschaft  $x^{(k^{(\ell)})} \rightarrow x^* \in L$  für  $\ell \rightarrow \infty$ . Aufgrund der Unterhalbstetigkeit von  $f$  folgt  $f(x^*) \leq \liminf_{\ell \rightarrow \infty} f(x^{(k^{(\ell)})}) = -\infty$ ; Widerspruch.

<sup>4</sup>**Beachte:** Bei der Kompaktheit von Mengen kommt es nicht auf die Teilraumtopologie an, in der wir diese betrachten. Bei der Kompaktheit von  $L = \{x \in F \mid f(x) \leq m\}$  kommt es also nicht darauf an, ob wir sie als kompakte Teilmenge von  $F$  oder von  $\mathbb{R}^n$  ansehen. Die Charakterisierung „kompakt  $\Leftrightarrow$  abgeschlossen und beschränkt“ gilt jedoch in  $\mathbb{R}^n$  und nicht in beliebigen Teilmengen.

Es sei nun  $f^* := \inf_{x \in F} f(x) \in \mathbb{R}$  der endliche Optimalwert und  $m \in \mathbb{R}$  in Niveau wie in der Voraussetzung. Dann gibt es eine Folge  $(x^{(k)}) \subseteq F$  mit der Eigenschaft<sup>5</sup>  $f(x^{(k)}) \searrow f^*$ . Wir unterscheiden zwei Fälle: Falls  $m = f^*$  ist, dann besteht die Sublevelmenge  $L$  nur aus globalen Minimierern und ist nach Annahme nichtleer; fertig. Andernfalls ist  $m > f^*$ , und wegen  $f(x^{(k)}) \searrow f^*$  gilt: Für hinreichend große  $k \in \mathbb{N}$  gehört die Folge zur Sublevelmenge  $L$ , und aufgrund der Kompaktheit existiert eine konvergente Teilfolge  $x^{(k^{(\ell)})} \rightarrow x^*$ , deren Grenzwert  $x^*$  in  $L$  liegt und insbesondere zulässig ist. Wegen der Unterhalbstetigkeit von  $f$  gilt  $\lim_{\ell \rightarrow \infty} f(x^{(k^{(\ell)})}) \geq f(x^*)$ , aber auch  $\lim_{\ell \rightarrow \infty} f(x^{(k^{(\ell)})}) = f^*$ . Dies zeigt, dass  $x^*$  ein globaler Minimierer ist.  $\square$

## § 2 NOTATION UND WIEDERHOLUNG VON DIFFERENZIERBARKEITSBEGRIFFEN

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

- Die natürlichen Zahlen sind  $\mathbb{N} = \{1, 2, \dots\}$ . 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  $\text{Id}$  bezeichnet. Wir unterscheiden den Vektorraum der Spaltenvektoren  $\mathbb{R}^n$  vom Vektorraum der Zeilenvektoren  $\mathbb{R}_n$ .
- Unendliche skalarwertige oder vektorwertige Folgen  $\mathbb{N} \rightarrow \mathbb{R}^n$  bezeichnen wir mit  $(x^{(k)})$  und nicht mit  $x_k$  etc., um einen Konflikt mit der Bezeichnung der Komponenten eines Vektors  $x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$  zu vermeiden. Endlich viele Vektoren werden dennoch auch gelegentlich mit  $x_1, x_2$  etc. bezeichnet, wenn wir deren Komponenten nicht benötigen.
- Die durch die streng monoton wachsende Folge  $\mathbb{N} \ni \ell \mapsto k^{(\ell)} \in \mathbb{N}$  gebildete **Teilfolge** einer Folge  $(x^{(k)})$  wird mit  $(x^{(k^{(\ell)})})$  bezeichnet.
- Für Vektoren  $x, y \in \mathbb{R}^n$  bezeichnet  $x^\top y$  das Euklidische Skalarprodukt (Innenprodukt) und  $\|x\|$  die euklidische Norm:

$$\|x\| = \sqrt{x^\top x}.$$

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  $x^\top M y$  und eine Norm  $\|x\|_M = \sqrt{x^\top M x}$  auf  $\mathbb{R}^n$ . Es gilt  $\|x\| = \|x\|_{\text{Id}}$ .
- Für  $\varepsilon > 0$  und  $x^* \in \mathbb{R}^n$  ist

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

<sup>5</sup>Für eine reelle Zahlenfolge  $(y^{(k)})$  bedeutet  $y^{(k)} \searrow y$ , dass  $y^{(k)} > y$  gilt und  $y^{(k)} \rightarrow y$ . Die Monotonie der Folge wird nicht verlangt.

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

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

- Das **Innere** einer Menge  $M \subseteq \mathbb{R}^n$  bezeichnen wir mit  $\text{int } M$  und den **Abschluss** mit  $\overline{M}$ .
- Für eine Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  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  $e_i = (0, \dots, 0, 1, 0, \dots, 0)^\top$  einer der Vektoren der Standardbasis von  $\mathbb{R}^n$ . Mit anderen Worten:

$$\frac{\partial}{\partial x_i} f(x) = \lim_{t \rightarrow 0} \frac{f(x + t e_i) - f(x)}{t}.$$

- 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  $d \in \mathbb{R}^n$ , kurz:

$$\frac{\partial}{\partial d} f(x) = \lim_{t \rightarrow 0} \frac{f(x + t d) - f(x)}{t}.$$

- 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  $d \in \mathbb{R}^n$ , kurz:

$$f'(x; d) = \lim_{t \searrow 0} \frac{f(x + t d) - f(x)}{t}.$$

- Eine Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  heißt **differenzierbar** (kurz: **diffbar**) an der Stelle  $x \in \mathbb{R}^n$ , falls ein Vektor  $v \in \mathbb{R}_n$  (Zeilenvektor) existiert, sodass gilt:

$$\frac{f(x + d) - f(x) - v d}{\|d\|} \rightarrow 0 \quad \text{für } d \rightarrow 0.$$

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  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  gilt

$$f'(x) = \left( \frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n} \right) \in \mathbb{R}_n.$$

Den transponierten Vektor (Spaltenvektor)

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = f'(x)^\top \in \mathbb{R}^n$$

bezeichnen wir als den **Gradienten** (bzgl. des Euklidischen Skalarprodukts) von  $f$  an der Stelle  $x$ .

- Für diffbare Funktionen  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  gilt:

$$f'(x; d) = \frac{\partial}{\partial d} f(x) = f'(x) d = \nabla f(x)^\top d.$$

- Eine Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  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.
- Eine vektorwertige Funktion  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  heißt an der Stelle  $x$  **diffbar**, wenn alle Komponentenfunktionen  $F_1, \dots, F_m$  dort diffbar sind. In diesem Fall ist die Ableitung  $F'(x)$  durch die **Jacobimatrix** von  $F$  an der Stelle  $x$ , also durch

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

gegeben.

- $F$  heißt **stetig partiell diffbar**, wenn alle Einträge der Jacobimatrix als Funktionen von der Stelle  $x$  stetig sind.  $C^1$ -Funktionen sind überall diffbar.
- Eine Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  heißt **zweimal differenzierbar** (kurz: **zweimal diffbar**) an der Stelle  $x \in \mathbb{R}^n$ , falls  $f$  in einer Umgebung von  $x$  diffbar ist und die Ableitung  $x \mapsto f'(x) \in \mathbb{R}^n$  an der Stelle  $x$  diffbar ist. In diesem Fall ist die zweite Ableitung  $f''(x)$  durch die **Hessematrix** von  $f$  an der Stelle  $x$ , also durch die Matrix der zweiten partiellen Ableitungen

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

gegeben. Diese ist dann symmetrisch (Satz von Schwarz)<sup>6</sup>.

- Eine Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  heißt **zweimal stetig partiell differenzierbar** oder kurz:  $C^2(\mathbb{R}^n, \mathbb{R})$ , wenn alle Einträge der Hessematrix als Funktionen von der Stelle  $x$  stetig sind.  $C^2$ -Funktionen sind überall zweimal diffbar.

Schließlich benötigen wir auch den Satz von Taylor, den wir in zwei Versionen angeben:

**Satz 2.1** (Taylor, siehe Cartan, 1967, Theorem 5.6.3). *Es sei  $G \subseteq \mathbb{R}^n$  offen,  $k \in \mathbb{N}_0$  und  $f: G \rightarrow \mathbb{R}$   $k$ -mal diffbar sowie  $(k+1)$ -mal diffbar im Punkt  $x_0 \in G$ . Dann gilt: Für alle  $\varepsilon > 0$  existiert  $\delta > 0$ , sodass gilt:*

$$\text{im Fall } k = 0: \quad |f(x_0 + d) - f(x_0) - f'(x_0) d| \leq \varepsilon \|d\|,$$

$$\text{im Fall } k = 1: \quad |f(x_0 + d) - f(x_0) - f'(x_0) d - \frac{1}{2} d^\top f''(x_0) d| \leq \varepsilon \|d\|^2.$$

<sup>6</sup>siehe z. B. Cartan, 1967, Proposition 5.2.2

für alle  $\|d\| < \delta$ .

**Satz 2.2** (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  $f: G \rightarrow \mathbb{R}$   $(k+1)$ -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:

im Fall  $k = 0$ :  $f(x_0 + d) = f(x_0) + f'(x_0 + \xi d) d$  (**Mittelwertsatz**),

im Fall  $k = 1$ :  $f(x_0 + d) = f(x_0) + f'(x_0) d + \frac{1}{2} d^T f''(x_0 + \xi d) d$ .

# 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

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

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.

## § 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 an der Stelle  $x^*$  diffbar. Dann ist die Ableitung  $f'(x^*) = 0$ .

*Beweis.* Es sei  $d \in \mathbb{R}^n$  beliebig. Wir betrachten die Kurve  $\gamma: (-\delta, \delta) \rightarrow \mathbb{R}^n$ ,  $\gamma(t) := x^* + t d$ . Für hinreichend kleines  $\delta > 0$  verläuft diese Kurve innerhalb der Umgebung der lokalen Optimalität von  $x^*$ . Daraus folgt, dass  $f \circ \gamma$  bei  $t = 0$  einen lokalen Minimierer besitzt. **Quizfrage 3.1:** Ist das klar?

Aufgrund dieser lokalen Optimalität gilt für den Differenzenquotienten

$$\frac{f(\gamma(t)) - f(\gamma(0))}{t} = \frac{f(x^* + t d) - f(x^*)}{t} \begin{cases} \geq 0 & \text{für } t > 0, \\ \leq 0 & \text{für } t < 0. \end{cases}$$

Andererseits konvergiert aber der Differenzenquotient für  $t \rightarrow 0$  gegen den Grenzwert  $f'(x^*) d$ . Es muss daher notwendigerweise  $f'(x^*) d = 0$  gelten. Da  $d \in \mathbb{R}^n$  beliebig war, bedeutet das  $f'(x^*) = 0$ .  $\square$

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

**Quizfrage 3.2:** Wie kann man sich die Eigenschaft „ $f'(x) = 0$ “ beispielsweise für eine Funktion  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  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 an der Stelle  $x^*$  zweimal diffbar. Dann ist die Hessematrix  $f''(x^*)$  positiv semidefinit.<sup>1</sup>

*Beweis.* Es sei  $d \in \mathbb{R}^n$  beliebig. Wie in Satz 3.1 definieren wir  $\gamma(t) := x^* + t d$  und betrachten wieder  $\varphi := f \circ \gamma$ , das bei  $t = 0$  einen lokalen Minimierer besitzt. Da  $\varphi$  an dieser Stelle zweimal diffbar ist, folgt aus Satz 2.1: Für alle  $\varepsilon > 0$  existiert  $\delta > 0$ , sodass

$$|\varphi(t) - \varphi(0) - \varphi'(0)t - \frac{1}{2}\varphi''(0)t^2| \leq \varepsilon t^2$$

für alle  $|t| < \delta$  ist. Aufgrund von Satz 3.1 ist  $\varphi'(0) = 0$ , und aus der lokalen Optimalität folgt  $\varphi(0) \leq \varphi(t)$  für alle  $|t|$  hinreichend klein. Wir erhalten also

$$-\frac{1}{2}\varphi''(0)t^2 \leq \varphi(t) - \varphi(0) - \frac{1}{2}\varphi''(0)t^2 \leq \varepsilon t^2$$

für alle  $|t|$  hinreichend klein, folglich

$$\frac{1}{2}\varphi''(0) \geq -\varepsilon.$$

Da  $\varepsilon > 0$  beliebig war, folgt  $\varphi''(0) = d^T f''(x^*) d \geq 0$ . Da wiederum  $d \in \mathbb{R}^n$  beliebig war, ist  $f''(x^*)$  positiv semi-definit.  $\square$

**Quizfrage 3.3:** Wie kann man sich die Eigenschaft „ $f''(x)$  ist positiv semidefinit“ beispielsweise für eine Funktion  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  vorstellen?

**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  $f$  zweimal diffbar an der Stelle  $x^*$ , und es gelte

- (i)  $f'(x^*) = 0$  und
- (ii)  $f''(x^*)$  ist positiv definit mit kleinstem Eigenwert  $\mu > 0$ .

Dann gilt: Zu jedem  $\beta \in (0, \mu)$  gibt es eine Umgebung  $U_\beta(x^*)$  von  $x^*$  mit der Eigenschaft

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

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

*Beweis.* Wir nutzen dieses Mal Satz 2.1 direkt für die Funktion  $f$ . Für jedes  $\varepsilon > 0$  existiert  $\delta > 0$ , sodass gilt:

$$|f(x^* + d) - f(x^*) - f'(x^*)d - \frac{1}{2}d^T f''(x^*)d| \leq \varepsilon \|d\|^2$$

<sup>1</sup>Aufgrund der Symmetrie von  $f''(x^*)$  ist dies äquivalent dazu, dass alle Eigenwerte von  $f''(x^*)$  nicht-negativ sind.

für alle  $\|d\| < \delta$ . Nach Voraussetzung ist  $f'(x^*) = 0$ . Es folgt also

$$-\varepsilon \|d\|^2 \leq f(x^* + d) - f(x^*) - \frac{1}{2} d^T f''(x^*) d$$

für alle  $\|d\| < \delta$ . Das bedeutet aber

$$f(x^* + d) \geq f(x^*) + \frac{1}{2} d^T f''(x^*) d - \varepsilon \|d\|^2$$

für alle  $\|d\| < \delta$ .

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:

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

Nun können wir die Behauptung zeigen: Zu  $\beta \in (0, \mu)$  wähle  $\varepsilon := (\mu - \beta)/2 > 0$  und ein dazugehöriges  $\delta > 0$ . Dann gilt also

$$\begin{aligned} f(x^* + d) &\geq f(x^*) + \frac{1}{2} d^T f''(x^*) d - \varepsilon \|d\|^2 \\ &\geq f(x^*) + \frac{\mu}{2} \|d\|^2 - \varepsilon \|d\|^2 \\ &= f(x^*) + \frac{\beta}{2} \|d\|^2 \end{aligned}$$

für alle  $\|d\| < \delta$ . □

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, Definition 13.8).

**Quizfrage 3.4:** Wie kann man sich die Eigenschaft (3.1) beispielsweise für eine Funktion  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  vorstellen?

**Quizfrage 3.5:** Welche Eigenschaft der Funktion  $f$  beschreibt der kleinste Eigenwert  $\mu$  von  $f''(x^*)$ ?

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. Als **Abstiegsverfahren** bezeichnet man ein iteratives Verfahren, das entlang von Abstiegsrichtungen voranschreitet und dabei (in der Regel) eine monoton nicht-wachsende Folge von Zielfunktionswerten erzeugt. Es entsteht also eine Folge von Iterierten  $x^{(k)} \subseteq \mathbb{R}^n$ , und 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) Aufdatieren durch  $x^{(k+1)} := x^{(k)} + t^{(k)} d^{(k)}$ .
- (4) Erhöhen des Iterationszählers  $k \rightsquigarrow k + 1$ .

In diesem § 4 sei  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  mindestens einmal stetig partiell diffbar, kurz:  $C^1$ . Es gilt also für die (beidseitige) Richtungsableitung

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

Außerdem dürfen wir den **Satz von Taylor 2.2** in Form des Mittelwertsatzes (also für  $k = 0$ ) verwenden.

**Definition 4.1** (Abstiegsrichtung).

Ein Vektor  $d \in \mathbb{R}^n$  heißt **Abstiegsrichtung** (englisch: *descent direction*) 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** (englisch: *direction of steepest descent*) von  $f$  im Punkt  $x$ . Er ist immer eine Abstiegsrichtung, außer in einem stationären Punkt. Wir können (4.1) auch schreiben als  $\nabla f(x)^T d < 0$ . Anschaulich bedeutet dies, dass der Winkel zwischen der Richtung  $d$  und dem negativen Gradienten  $-\nabla f(x)$  kleiner als  $90^\circ$  ist, siehe **Abbildung 4.1**.

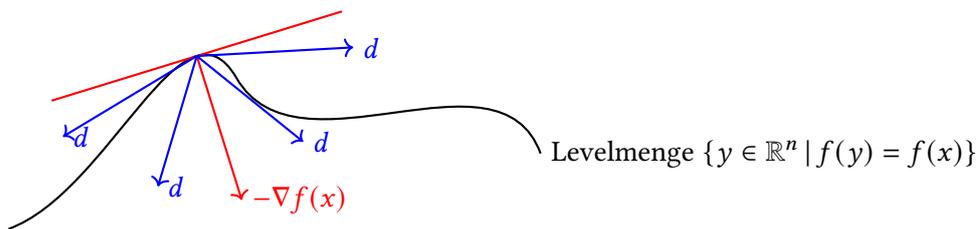


Abbildung 4.1: Verschiedene Abstiegsrichtungen  $d$  für  $f$  im Punkt  $x$ .

**Quizfrage 4.1:** Mit welchem Begriff könnte man die Menge aller Abstiegsrichtungen einer Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  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 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** (englisch: *line search*), 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  $t^{(k)} = (\frac{1}{2})^{k+2}$  gewählt. Dann ist die Folge der Iterierten gegeben durch

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

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 4.2:** 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** (englisch: *exact line search*)

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

ist außer in Sonderfällen für besonders einfache Zielfunktionen  $f$  nicht praktikabel.

**Quizfrage 4.3:** Welche Schwierigkeiten könnten sich beim Versuch, die Schrittweite nach (4.2) zu wählen, 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**<sup>2</sup> erfüllt ist:

$$f(x + t d) \leq f(x) + \sigma t f'(x) d. \quad (4.3)$$

Dabei ist  $\sigma \in (0, 1)$  der **Armijo-Parameter**. **Quizfrage 4.4:** Welche anschauliche Bedeutung hat der Parameter  $\sigma$ ?

<sup>2</sup>Armijo, 1966

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

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

zur **Suchrichtung**  $d$  ein. Man nennt  $\varphi$  auch den **Schnitt** durch die Funktion  $f$  am Punkt  $x$  in Richtung  $d$ . Die Funktion  $\varphi$  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  $\varphi'(0) = f'(x) d = -\|\nabla f(x)\|^2$ .

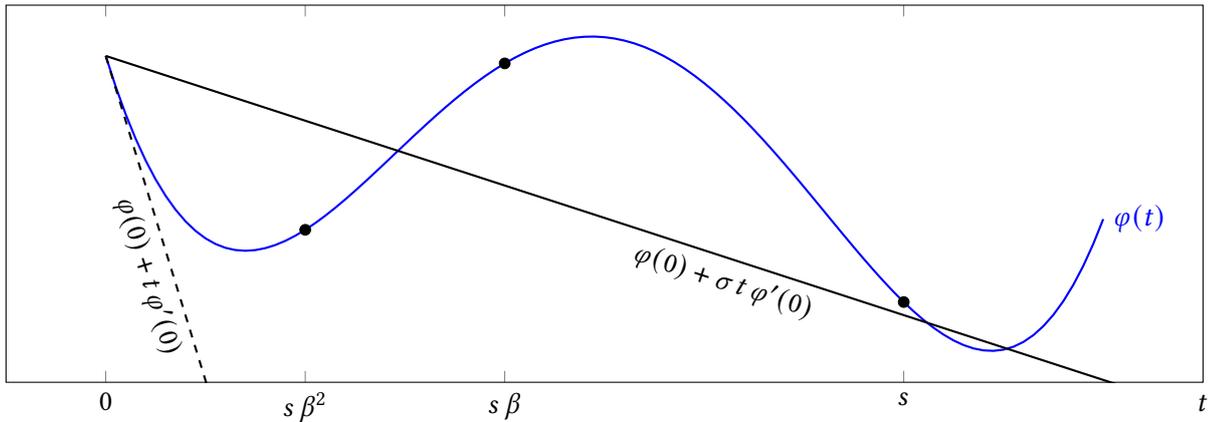


Abbildung 4.2: Darstellung der Armijo-Bedingung (4.5) und einigen Test-Schrittweiten beim Backtracking. Der Armijo-Parameter ist hier als  $\sigma = 0.1$  und der Backtracking-Parameter als  $\beta = 0.5$  gewählt.

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 ein Exponent  $\ell_0 \in \mathbb{N}_0$  existiert, sodass Schrittweiten der Form  $t = s\beta^\ell$  für  $\ell \geq \ell_0$  immer in  $[0, T]$  liegen.

*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  $k \rightarrow \infty$  folgt

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

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

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 mit Armijo-Schrittweitensuche werden die folgenden problemspezifischen 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 4.5:** Angenommen, für die Funktion  $f$  liegt (neben der Routine für die Auswertung der Funktionswerte) eine Routine vor, die zu einer gegebenen 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  $x^{(k)} \rightarrow x$  und  $d^{(k)} \rightarrow d$  sowie  $t^{(k)} \searrow 0$ . Dann gilt

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

*Beweis.* Wegen des [Mittelwertsatzes 2.2](#) existiert zu jedem  $k \in \mathbb{N}$  ein  $\xi^{(k)} \in (0, 1)$  mit

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

□

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  $x^{(k^{(\ell)})} \rightarrow x^*$ , und wegen der Stetigkeit von  $f$  gilt  $f(x^{(k^{(\ell)})}) \rightarrow f(x^*)$ . Da  $f(x^{(k)})$  aber monoton fällt, konvergiert die gesamte Folge  $f(x^{(k)}) \rightarrow f(x^*)$ . Somit gilt also auch  $f(x^{(k+1)}) - f(x^{(k)}) \rightarrow 0$ .

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

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

also

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

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

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

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

Die Grenzübergänge  $x^{(k^{(\ell)})} \rightarrow x^*$ ,  $d^{(k^{(\ell)})} = -\nabla f(x^{(k^{(\ell)})}) \rightarrow -\nabla f(x^*)$  und  $t^{(k^{(\ell)})} \rightarrow 0$  für  $\ell \rightarrow \infty$  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<sup>3</sup> 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 4.6:** Welche Bedeutung haben die Bedingungen (i) bis (iii)?

**Quizfrage 4.7:** 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:

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

Speziell beim Gradientenverfahren ergibt sich dann also

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

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

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

für die Startschrittweite.

Auch unter Verwendung dieser Startschrittweiten kann man Satz 4.6 beweisen.

<sup>3</sup>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  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  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 Euklidischen Skalarprodukt  $(x, y) := x^T y$  ausgestattet ist.

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

$$(x, y)_M := x^T M y$$

mit einer symmetrischen, positiv definiten (s. p. d.) Matrix  $M$  wählt. Dementsprechend ändert sich auch die Norm zur Längen- und Abstandsmessung im Raum  $\mathbb{R}^n$  der Optimierungsvariablen zu

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

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

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

Die Normierung auf die Länge 1 ist willkürlich gewählt. Alternativ könnten wir auch die äquivalente Aufgabe

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

betrachten. **Quizfrage 4.8:** Warum ist diese Aufgabe gleichwertig?

Aufgabe (4.9) ist eine restringierte Optimierungsaufgabe, die wir jedoch ohne Kenntnisse der Theorie lösen können: Wir schreiben die Zielfunktion als  $M$ -Skalarprodukt um:<sup>4</sup>

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

wobei die Symmetrie  $M = M^T$  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:

<sup>4</sup>Das heißt, wir bestimmen hier den Riesz-Repräsentanten von  $f'(x)$  bzgl. des  $M$ -Skalarprodukts.

**Lemma 4.9** (Richtung des steilsten Abstiegs im  $M$ -Skalarprodukt).

Falls  $f'(x) \neq 0$  gilt, ist die dann eindeutige Lösung  $d^*$  von (4.9) gegeben durch

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

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

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

$$M d^* = -\nabla f(x). \quad (4.11)$$

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

$$f'(x) d^* = -\nabla f(x)^T M^{-1} \nabla f(x) = -\|\nabla f(x)\|_{M^{-1}}^2 = -\|\nabla_M f(x)\|_M^2 < 0, \quad (4.12)$$

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

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

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

**Quizfrage 4.9:** Warum ändert sich **Abbruchbedingung (i)** nicht?

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

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

Zur Unterscheidung vom Euklidischen Fall heißt das Verfahren dann auch das **vorkonditionierte Gradientenverfahren** (englisch: *preconditioned steepest descent method*). 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  $M = \text{Id}$  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 (siehe Kapitel 3) quadratisches Polynom:

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

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

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

denn dies ist die einzige Lösung der notwendigen Bedingungen (Satz 3.1), und die hinreichenden Bedingungen (Satz 3.3) sind dort erfüllt. Für ein beliebiges  $x \in \mathbb{R}^n$  bezeichnen wir die Größe

$$r := Q x + c = \nabla f(x)$$

auch als das zu  $x$  gehörige **Residuum** (englisch: *residual*).

**Quizfrage 4.10:** An welcher Stelle geht die Symmetrie der Matrix  $Q$  ein?

Natürlich wird man das Gradientenverfahren (Algorithmus 4.11) zur Lösung von (4.14) bzw. des linearen Gleichungssystems (4.15) 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, sondern nur eine Funktion, die Matrix-Vektor-Produkte  $Qx$  auswertet.

**Beachte:** Das Verfahren kommt bereits mit Matrix-Vektor-Produkten  $Qx$  aus. Diese werden bei der Berechnung des Gradienten  $\nabla f(x) = Qx + c$  in [Zeile 3](#) benötigt.

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

$$t_{\min} = \arg \min_{t \geq 0} f(x^{(k)} + t d^{(k)})$$

im  $k$ -ten Schritt berechnen:

$$t^{(k)} := t_{\min} = \frac{(d^{(k)})^\top M d^{(k)}}{(d^{(k)})^\top Q d^{(k)}}. \quad (4.16)$$

In diesem Abschnitt wählen wir statt der Armijo-Strategie in [Algorithmus 4.4](#) stets die exakte Schrittweite ([4.16](#)). Der Vollständigkeit halber geben wir das Verfahren für diesen Spezialfall nochmals an:

**Algorithmus 4.11** (Vorkonditioniertes Gradientenverfahren bei quadratischer Zielfunktion).

**Eingabe:** Startschätzung  $x^{(0)} \in \mathbb{R}^n$

**Eingabe:** s. p. d. Matrix  $M \in \mathbb{R}^{n \times n}$

- 1: Setze  $k := 0$
- 2: Setze  $r^{(0)} := Qx^{(0)} + c$
- 3: Setze  $d^{(0)} := -M^{-1}r^{(0)}$
- 4: **while** *Qbruchkriterium nicht erfüllt* **do**
- 5:     Setze  $q^{(k)} := Qd^{(k)}$
- 6:     Setze  $t^{(k)} := -\frac{(r^{(k)})^\top d^{(k)}}{(d^{(k)})^\top q^{(k)}}$
- 7:     Setze  $x^{(k+1)} := x^{(k)} + t^{(k)} d^{(k)}$
- 8:     Setze  $r^{(k+1)} := r^{(k)} + t^{(k)} q^{(k)}$
- 9:     Setze  $d^{(k+1)} := -M^{-1}r^{(k+1)}$
- 10:    Setze  $k := k + 1$
- 11: **end while**
- 12: **return**  $x^{(k)}$

Es stellt sich nun die Frage nach dem Konvergenzverhalten sowie nach der Rolle des Vorkonditionierers/Skalarprodukts  $M$ . Dazu geben wir zunächst ein Hilfsresultat an, das die Funktionswerte, den Fehler  $x - x^*$  und das Residuum in Beziehung setzt:

**Lemma 4.12.** *Es gilt*

$$f(x) - f(x^*) = \frac{1}{2} \|x - x^*\|_Q^2 = \frac{1}{2} \|r\|_{Q^{-1}}^2. \quad (4.17)$$

*Beweis.* Durch direkte Rechnung ergibt sich

$$\begin{aligned}
 f(x) - f(x^*) &= \frac{1}{2}x^\top Q x + c^\top x + \gamma - \frac{1}{2}(x^*)^\top Q x^* - c^\top x^* - \gamma \\
 &= \frac{1}{2}x^\top Q x - (x^*)^\top Q x - \frac{1}{2}(x^*)^\top Q x^* + (x^*)^\top Q x^* \quad \text{denn } c = -Q x^* \\
 &= \frac{1}{2}x^\top Q x - (x^*)^\top Q x + \frac{1}{2}(x^*)^\top Q x^* \\
 &= \frac{1}{2}\|x - x^*\|_Q^2 \\
 &= \frac{1}{2}(x - x^*)^\top r = \frac{1}{2}r^\top Q^{-1}r \quad \text{denn } r = Q(x - x^*) \\
 &= \frac{1}{2}\|r\|_{Q^{-1}}^2.
 \end{aligned}$$

□

Für die Iterierten von [Algorithmus 4.11](#) zur Minimierung von (4.14) gilt nun die folgende Rekursion:

$$\begin{aligned}
 f(x^{(k+1)}) - f(x^*) &= \frac{1}{2}\|r^{(k+1)}\|_{Q^{-1}}^2 = \frac{1}{2}\|r^{(k)} + t^{(k)}Q d^{(k)}\|_{Q^{-1}}^2 \quad \text{wegen (4.17)} \\
 &= \frac{1}{2}\|r^{(k)}\|_{Q^{-1}}^2 + t^{(k)}(r^{(k)})^\top d^{(k)} + \frac{1}{2}(t^{(k)})^2(d^{(k)})^\top Q d^{(k)} \\
 &= \frac{1}{2}\|r^{(k)}\|_{Q^{-1}}^2 - \frac{[(r^{(k)})^\top d^{(k)}]^2}{(d^{(k)})^\top Q d^{(k)}} + \frac{1}{2} \frac{[(r^{(k)})^\top d^{(k)}]^2}{(d^{(k)})^\top Q d^{(k)}} \quad \text{wegen } t^{(k)} = -\frac{(r^{(k)})^\top d^{(k)}}{(d^{(k)})^\top Q d^{(k)}} \\
 &= \frac{1}{2}\|r^{(k)}\|_{Q^{-1}}^2 - \frac{1}{2} \frac{[(r^{(k)})^\top d^{(k)}]^2}{(d^{(k)})^\top Q d^{(k)}} \\
 &= \left(1 - \frac{[(r^{(k)})^\top d^{(k)}]^2}{[(d^{(k)})^\top Q d^{(k)}][(r^{(k)})^\top Q^{-1}r^{(k)}]}\right) (f(x^{(k)}) - f(x^*)) \quad \text{wegen (4.17)}.
 \end{aligned}$$

Hier setzen wir nun den speziellen Zusammenhang  $r^{(k)} = -M d^{(k)}$  für die Iterierten aus [Algorithmus 4.11](#) ein:

$$= \left(1 - \frac{[(d^{(k)})^\top M d^{(k)}]^2}{[(d^{(k)})^\top Q d^{(k)}][(d^{(k)})^\top M Q^{-1}M d^{(k)}]}\right) (f(x^{(k)}) - f(x^*)) \quad (4.18)$$

Für die weitere Abschätzung des Bruches benutzen wir die **Kantorovich-Ungleichung**, die wir zunächst für den Fall  $M = \text{Id}$  angeben:

**Lemma 4.13** (Kantorovich-Ungleichung). *Es sei  $Q \in \mathbb{R}^{n \times n}$  s. p. d. und  $\alpha := \lambda_{\min}(Q)$  sowie  $\beta := \lambda_{\max}(Q)$ . Dann gilt*

$$1 \leq \frac{(x^\top Q x)(x^\top Q^{-1}x)}{\|x\|^4} \leq \frac{(\alpha + \beta)^2}{4\alpha\beta} \leq \frac{\beta}{\alpha} \quad (4.19)$$

für alle  $x \in \mathbb{R}^n, x \neq 0$ .

**Beachte:** Für den Rayleigh-Quotienten von  $Q$  gilt

$$\frac{x^T Q x}{\|x\|^2} \leq \lambda_{\max}(Q) = \beta \quad \text{und analog} \quad \frac{x^T Q^{-1} x}{\|x\|^2} \leq \lambda_{\max}(Q^{-1}) = 1/\alpha.$$

Die zugehörigen Eigenvektoren erfüllen die Ungleichungen jeweils mit Gleichheit. Die offensichtliche Abschätzung

$$\frac{(x^T Q x) (x^T Q^{-1} x)}{\|x\|^4} \leq \frac{\beta}{\alpha}$$

ist jedoch nicht scharf, da derselbe Vektor  $x$  i. A. nicht gleichzeitig Eigenvektor zum größten und zum kleinsten Eigenwert sein kann. Die Kantorovich-Ungleichung (4.19) verbessert diese Abschätzung.

*Beweis.* Aus der Cauchy-Schwarz-Ungleichung folgt<sup>5</sup>

$$\|x\|^2 = x^T x = x^T Q^{-1/2} Q^{1/2} x \leq \|Q^{-1/2} x\| \|Q^{1/2} x\|.$$

Durch Quadrieren erhalten wir

$$\|x\|^4 \leq \|Q^{-1/2} x\|^2 \|Q^{1/2} x\|^2 = (x^T Q x) (x^T Q^{-1} x)$$

und damit die untere Schranke in (4.19).

Es seien nun<sup>6</sup>  $\lambda_1, \dots, \lambda_n > 0$  die Eigenwerte von  $Q$  und  $v_1, \dots, v_n$  ein Satz zugehöriger orthonormaler Eigenvektoren. Es sei  $x \in \mathbb{R}^n$ ,  $x \neq 0$  beliebig. Wir stellen  $x$  als  $x = \sum_{i=1}^n \gamma_i v_i$  dar. O. B. d. A. sei  $\|x\|^2 = \sum_{i=1}^n \gamma_i^2 = 1$ . Einsetzen in die linke Seite von (4.19) ergibt:

$$\frac{(x^T Q x) (x^T Q^{-1} x)}{\|x\|^4} = \underbrace{\left[ \sum_{i=1}^n \lambda_i \gamma_i^2 \right]}_{=\mathbb{E}(T)} \underbrace{\left[ \sum_{i=1}^n \frac{1}{\lambda_i} \gamma_i^2 \right]}_{=\mathbb{E}(1/T)}.$$

Es ist jetzt aus Gründen der Übersichtlichkeit hilfreich, diese Faktoren als Erwartungswerte einer „Zufallsvariablen“  $T$  bzw.  $1/T$  zu interpretieren, wobei  $T$  die Werte  $\lambda_i \in [\alpha, \beta]$  mit „Wahrscheinlichkeit“  $\gamma_i^2$  annimmt. Für  $0 < \alpha \leq T \leq \beta$  gilt

$$0 \leq (\beta - T)(T - \alpha) = (\beta + \alpha - T)T - \alpha\beta,$$

also auch

$$\frac{1}{T} \leq \frac{\alpha + \beta - T}{\alpha\beta}$$

und daher (Erwartungswert nehmen)

$$\begin{aligned} \mathbb{E}(T) \mathbb{E}(1/T) &\leq \mathbb{E}(T) \frac{\alpha + \beta - \mathbb{E}(T)}{\alpha\beta} \\ &= \frac{(\alpha + \beta)^2}{4\alpha\beta} - \frac{1}{\alpha\beta} \left[ \mathbb{E}(T) - \frac{1}{2}(\alpha + \beta) \right]^2 \\ &\leq \frac{(\alpha + \beta)^2}{4\alpha\beta}. \end{aligned}$$

<sup>5</sup>Hierbei ist  $Q^{1/2}$  die Matrixwurzel der s. p. d. Matrix  $Q$ , also diejenige eindeutig bestimmte s. p. d. Matrix, deren Quadrat wieder  $Q$  ist. Weiter ist  $Q^{-1/2}$  die Inverse von  $Q^{1/2}$ .  $Q^{-1/2}$  ist gleichzeitig die Matrixwurzel der s. p. d. Matrix  $Q^{-1}$ .

<sup>6</sup>Wir folgen ab hier dem Beweis von Anderson, 1971, wie er in der Masterarbeit Alpargu, 1996, Abschnitt 1.2.2 wiedergegeben ist.

Damit ist die erste (wesentliche) obere Schranke in (4.19) bewiesen. Die noch fehlende Ungleichung folgt elementar aus  $0 < \alpha \leq \beta$ .  $\square$

Um die Kantorovich-Ungleichung zur Abschätzung von (4.18) verwenden zu können, benötigen wir noch eine Verallgemeinerung von der Euklidischen Norm  $\|x\|$  auf die  $M$ -Norm  $\|x\|_M$ . Im Folgenden seien  $\lambda_{\min}(Q; M) > 0$  und  $\lambda_{\max}(Q; M) > 0$  der kleinste und größte Eigenwert des **verallgemeinerten Eigenwertproblems** (englisch: *generalized eigenvalue problem*)

$$Qx = \lambda Mx \quad \text{oder äquivalent} \quad M^{-1}Qx = \lambda x$$

mit den s. p. d. Matrizen  $Q$  und  $M$ .

**Folgerung 4.14** (verallgemeinerte Kantorovich-Ungleichung). *Es seien  $Q \in \mathbb{R}^{n \times n}$  und  $M \in \mathbb{R}^{n \times n}$  beide s. p. d. und  $\alpha := \lambda_{\min}(Q; M)$  sowie  $\beta := \lambda_{\max}(Q; M)$ . Dann gilt*

$$1 \leq \frac{(x^T Q x) (x^T M Q^{-1} M x)}{\|x\|_M^4} \leq \frac{(\alpha + \beta)^2}{4 \alpha \beta} \leq \frac{\beta}{\alpha} \quad (4.20)$$

für alle  $x \in \mathbb{R}^n$ ,  $x \neq 0$ .

*Beweis.* Für die Abschätzung nach unten verwenden wir

$$\|x\|_M^2 = x^T M x = x^T Q^{1/2} Q^{-1/2} M x \leq (x^T Q x)^{1/2} (x^T M Q^{-1} M x)^{1/2}$$

und daher  $\|x\|_M^4 \leq (x^T Q x) (x^T M Q^{-1} M x)$ .

Wir benutzen nun die Cholesky-Zerlegung<sup>7</sup>  $M = LL^T$  und setzen  $y := L^T x$ , also  $x = L^{-T} y$  ein:

$$\frac{(x^T Q x) (x^T M Q^{-1} M x)}{(x^T M x)^2} = \frac{(y^T L^{-1} Q L^{-T} y) (y^T L^T Q^{-1} L y)}{(y^T y)^2}.$$

Dies entspricht der Form in (4.19) mit der s. p. d. Matrix  $\tilde{Q} := L^{-1} Q L^{-T}$ . Deren Eigenpaare  $(\lambda, v)$  erfüllen

$$\tilde{Q} v = L^{-1} Q L^{-T} v = \lambda v, \quad v \neq 0,$$

also auch

$$Q L^{-T} v = \lambda L v.$$

Ersetzen wir noch  $v$  durch  $L^T w$ , so erhalten wir

$$Q w = \lambda M w. \quad (4.21)$$

Damit ist gezeigt, dass  $(\lambda, v)$  genau dann ein Eigenpaar von  $\tilde{Q} = L^{-1} Q L^{-T}$  ist, wenn  $(\lambda, w = L^{-T} v)$  ein Eigenpaar des verallgemeinerten Eigenwertproblems (4.21) ist. Insbesondere sind die Eigenwerte gleich. Es seien nun wie angenommen  $0 < \alpha \leq \beta$  die extremalen Eigenwerte von (4.21), dann sind dies auch die extremalen Eigenwerte von  $\tilde{Q}$ , und die Behauptung folgt aus der gewöhnlichen Kantorovich-Ungleichung (4.19).  $\square$

<sup>7</sup>Stattdessen könnten wir auch mit der Matrix-Wurzel  $M^{1/2}$  arbeiten.

Mit Hilfe der **verallgemeinerten (spektralen) Konditionszahl** (englisch: *generalized (spectral) condition number*) von  $Q$  bzgl.  $M$ ,

$$\kappa := \text{cond}_2(Q; M) = \frac{\lambda_{\max}(Q; M)}{\lambda_{\min}(Q; M)} \quad (4.22)$$

können wir die Abschätzung (4.20) auch in der äquivalenten Form

$$1 \leq \frac{(x^\top Q x) (x^\top M Q^{-1} M x)}{\|x\|_M^4} \leq \frac{(\kappa + 1)^2}{4 \kappa} \leq \kappa \quad (4.23)$$

schreiben.

Mit Hilfe der verallgemeinerten Kantorovich-Ungleichung (4.20) folgt nun für die in (4.18) abzuschätzende Klammer:

$$1 - \frac{[(d^{(k)})^\top M d^{(k)}]^2}{[(d^{(k)})^\top Q d^{(k)}] [(d^{(k)})^\top M Q^{-1} M d^{(k)}]} \leq 1 - \frac{4 \alpha \beta}{(\alpha + \beta)^2} = \frac{(\beta - \alpha)^2}{(\beta + \alpha)^2} = \left( \frac{\kappa - 1}{\kappa + 1} \right)^2.$$

Damit haben wir das klassische Konvergenzresultat des (vorkonditionierten) Gradientenverfahrens bewiesen:

**Satz 4.15** (Globaler Konvergenzsatz für quadratische Zielfunktionen).

Es seien  $Q$  und  $M$  s. p. d. Matrizen und  $\kappa$  die verallgemeinerte Konditionszahl von  $Q$  bzgl.  $M$ , siehe (4.22). 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

$$f(x^{(k+1)}) - f(x^*) \leq \left( \frac{\kappa - 1}{\kappa + 1} \right)^2 (f(x^{(k)}) - f(x^*)) \quad (4.24)$$

und deswegen auch

$$\|x^{(k+1)} - x^*\|_Q \leq \left( \frac{\kappa - 1}{\kappa + 1} \right) \|x^{(k)} - x^*\|_Q, \quad (4.25a)$$

$$\|x^{(k)} - x^*\|_Q \leq \left( \frac{\kappa - 1}{\kappa + 1} \right)^k \|x^{(0)} - x^*\|_Q. \quad (4.25b)$$

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

**Bemerkung 4.16** (Zum Konvergenzverhalten des Gradientenverfahrens).

- (i) Für große Konditionszahlen  $\kappa$  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.

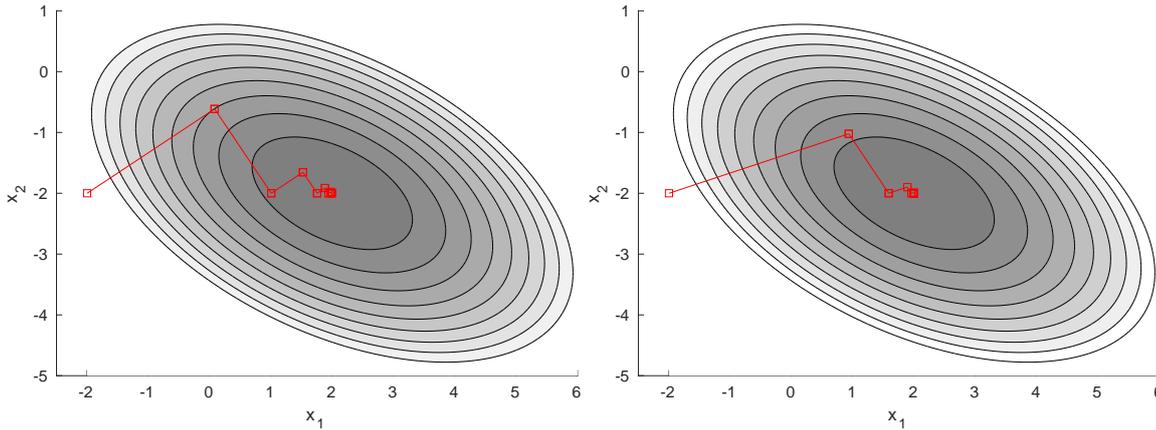


Abbildung 4.3: Illustration des Gradientenverfahrens (Algorithmus 4.10) mit Startpunkt  $x^{(0)} = (-2, -2)^T$  und exakter Schrittweite (4.16) für die Minimierung von (4.14) mit  $Q = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}$  und  $c = \begin{pmatrix} -2 \\ -8 \end{pmatrix}$ . Die exakte Lösung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

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

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  $\kappa$  möglichst klein, lineare Gleichungssysteme mit  $M$  als Koeffizientenmatrix aber noch leicht zu lösen sind. Manchmal ist bereits die Wahl

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

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 (4.25b) die Konvergenzabschätzung

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

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

## § 5 DAS NEWTON-VERFAHREN

Wir untersuchen in diesem Abschnitt das Newton-Verfahren zur Lösung der (nichtlinearen) Gleichung  $F(x) = 0$ . Dabei wird  $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$  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.

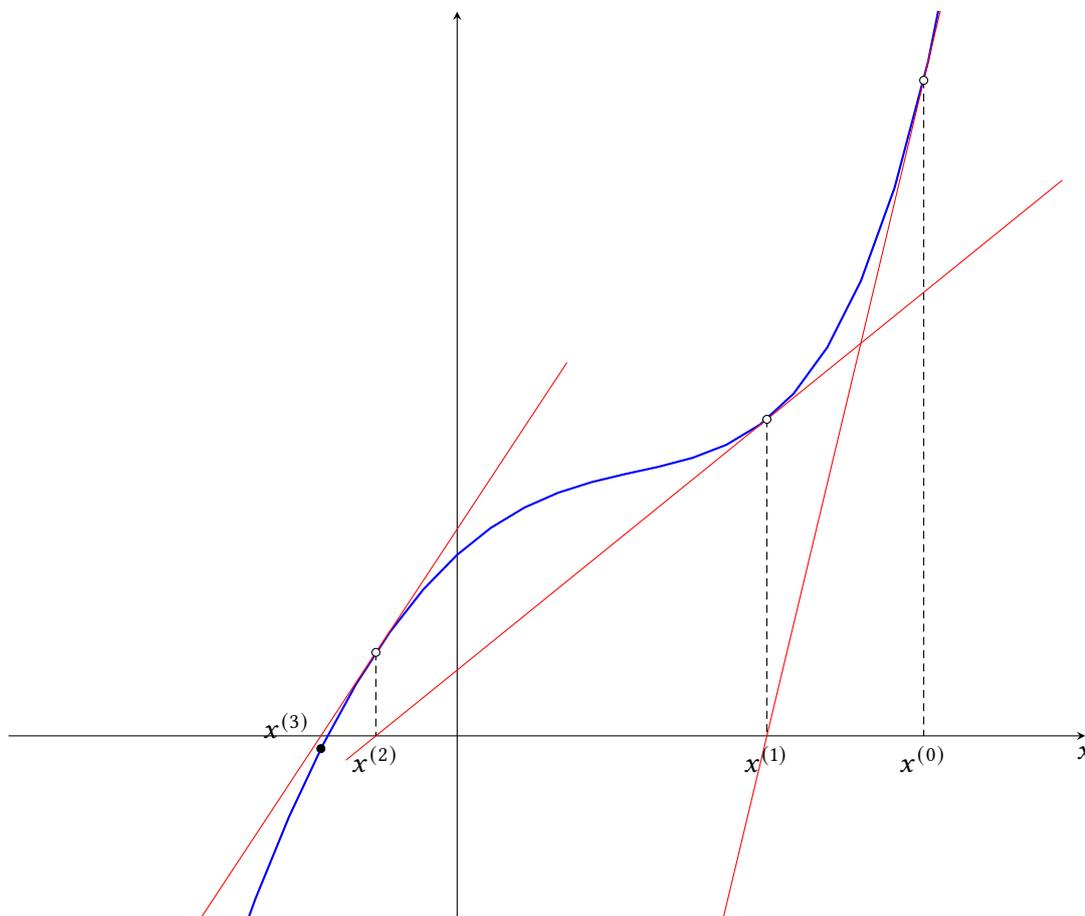


Abbildung 5.1: Illustration des Newton-Verfahrens zur Suche einer Nullstelle der Funktion  $F(x) = \exp(0.9x) - x^2$ .

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

**Jacobimatrix:**

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

**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

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

mit dem größten Singulärwert  $\sigma_{\max}$  von  $A$  und dem größten Eigenwert  $\lambda_{\max}$  von  $A^T A$ . 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  $\text{Id} - M$  regulär (invertierbar), und es gilt

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

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

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

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

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

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

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

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

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

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

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

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 offene Kugel  $B_\delta(x^*)$  und eine Konstante  $c > 0$ , sodass  $F'(x)$  für alle  $x \in B_\delta(x^*)$  regulär ist, und es gilt:*

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

*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

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

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

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

□

**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 5.1:** 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

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

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  $x^{(k)} \rightarrow x^*$ . Die Folge  $x^{(k)}$  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 (4.25a) zeigt beispielsweise die Q-lineare Konvergenz des Gradientenverfahrens bei quadratischer Zielfunktion.

**Quizfrage 5.2:** 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? (Siehe auch [Hausaufgabe 3.1](#).)

## § 5.2 DAS LOKALE NEWTON-VERFAHREN FÜR DIE NULLSTELLENBESTIMMUNG $F(x) = 0$

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

**Satz 5.8** (Lokaler Konvergenzsatz für das lokale 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 offene Kugel  $B_\delta(x^*)$  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

$$\|F'(x)^{-1}\| \leq c = 2 \|F'(x^*)^{-1}\|. \quad (5.1)$$

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

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

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

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

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

**Aussage (ii):** Wir stellen zunächst eine Gleichung für den Fehler auf:<sup>8</sup>

$$\begin{aligned} x^{(k+1)} - x^* &= x^{(k)} - x^* - F'(x^{(k)})^{-1}(F(x^{(k)}) - F(x^*)) \\ &= F'(x^{(k)})^{-1} [F'(x^{(k)})(x^{(k)} - x^*) - (F(x^{(k)}) - F(x^*))] \\ &= F'(x^{(k)})^{-1} \left[ F'(x^{(k)})(x^{(k)} - x^*) - \int_0^1 F'(x^{(k)} + t(x^* - x^{(k)}))(x^{(k)} - x^*) dt \right] \\ &= F'(x^{(k)})^{-1} \left[ \int_0^1 F'(x^{(k)}) - F'(x^{(k)} + t(x^* - x^{(k)})) dt \right] (x^{(k)} - x^*). \end{aligned}$$

Daraus erhalten wir folgende wichtige Abschätzung:

$$\|x^{(k+1)} - x^*\| \leq \|F'(x^{(k)})^{-1}\| \int_0^1 \overbrace{\|F'(x^{(k)}) - F'(x^{(k)} + t(x^* - x^{(k)}))\|}^{=:D^{(k)}(t)} dt \|x^{(k)} - x^*\|. \quad (5.3)$$

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

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

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

$$\|x^{(k+1)} - x^*\| \leq c \int_0^1 \|D^{(k)}(t)\| dt \|x^{(k)} - x^*\|,$$

also die Q-superlineare Konvergenz.

**Quizfrage 5.3:** 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:

$$\int_0^1 \|F'(x^{(k)}) - F'(x^{(k)} + t(x^* - x^{(k)}))\| dt \leq \int_0^1 L t \|x^* - x^{(k)}\| dt = \frac{L}{2} \|x^{(k)} - x^*\|.$$

Aus (5.3) erhalten wir nun:

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

□

**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.

<sup>8</sup>**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

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

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

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

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

$$m^{(k)}(x) := f(x^{(k)}) + \nabla f(x^{(k)})^\top (x - x^{(k)}) + \frac{1}{2} (x - x^{(k)})^\top f''(x^{(k)}) (x - x^{(k)}). \quad (5.6)$$

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):

$$f'(x^{(k)}) d^{(k)} = \nabla f(x^{(k)})^\top d^{(k)} = -\nabla f(x^{(k)})^\top \underbrace{f''(x^{(k)})^{-1}}_{\text{positiv definit}} \nabla f(x^{(k)}) < 0, \quad \text{falls } \nabla f(x^{(k)}) \neq 0.$$

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

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

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 5.4:** 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 **es eindeutig möglich ist<sup>RH</sup>**, 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

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

so setze  $\tilde{d}^{(k)} := -\nabla_M f(x^{(k)})$

*// Fallback auf Gradientenrichtung*

5:     Bestimme eine Schrittweite  $t^{(k)}$  mit der Armijo-Backtracking-Strategie zur Startschrittweite  $s = 1$ , sodass (4.3) erfüllt ist, also:

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

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 die folgenden problemspezifischen 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  $x^{(k)} \rightarrow x^*$ .

*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

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

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	$B^{(k)} = \text{Id}$	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 folgt 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:

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

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 6.1:** In der Regel setzt man den konstanten Term  $\gamma$  in der Zielfunktion gleich null. Warum stellt das keine Einschränkung in der Aufgabenstellung dar?

**Quizfrage 6.2:** 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 14.2). 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  $f(x) := c^\top x + \gamma$  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  $\hat{x} \in F$  mit  $f(\hat{x}) < f(x^*)$ . Wir betrachten Punkte  $x_\alpha$  entlang der Verbindungsstrecke zwischen  $x^*$  und  $\hat{x}$ , also

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

Alle diese Punkte  $x_\alpha$  sind zulässig, denn:

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

Für die Werte der Zielfunktion gilt

$$\begin{aligned} f(x_\alpha) &= c^\top x_\alpha + \gamma \\ &= c^\top(\alpha \hat{x} + (1 - \alpha) x^*) + \gamma \\ &= \alpha (c^\top \hat{x} + \gamma) + (1 - \alpha) (c^\top x^* + \gamma) \\ &= \alpha f(\hat{x}) + (1 - \alpha) f(x^*). \end{aligned}$$

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

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

Nun liegt aber für hinreichend kleines  $\alpha > 0$  der Punkt  $x_\alpha$  in der Umgebung  $U(x^*)$  und damit in  $U(x^*) \cap F$ . Dies steht im Widerspruch zur lokalen Optimalität von  $x^*$ . Also kann ein  $\hat{x}$  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

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

Wir erhalten folgende lineare Optimierungsaufgabe:

$$\left. \begin{array}{l}
 \text{Maximiere} \quad 9x_1 + 8x_2 \quad \text{über } x \in \mathbb{R}^2 \\
 \text{sodass} \quad x_1 + x_2 \leq 6 \quad (\text{Marzipanbedingung}) \\
 \quad \quad 2x_1 + x_2 \leq 11 \quad (\text{Nougatbedingung}) \\
 \quad \quad x_1 + 2x_2 \leq 9 \quad (\text{Schokoladenbedingung}) \\
 \text{und} \quad x_1 \geq 0 \\
 \quad \quad x_2 \geq 0.
 \end{array} \right\} \quad (6.2)$$

**Quizfrage 6.3:** Welche Bedeutung hat die Nichtnegativitätsbedingung  $x \geq 0$  in dieser Aufgabe? Was würde etwa eine negative Produktionsmenge  $x_1 < 0$  praktisch bedeuten?

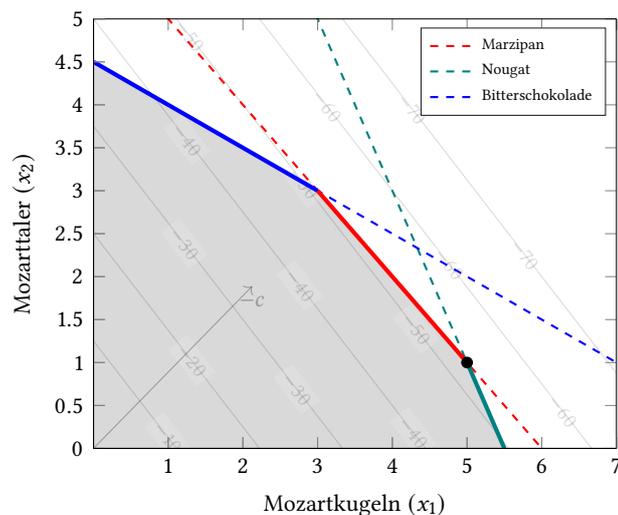


Abbildung 6.1: Zulässige Menge (Fünfeck), Niveaulinien der Zielfunktion und globaler Maximierer beim Mozartproblem (Beispiel 6.2).

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

$$\left. \begin{array}{l}
 \text{Maximiere} \quad c^T x \quad \text{über } x \in \mathbb{R}^n \\
 \text{sodass} \quad Ax \leq b \\
 \text{und} \quad x \geq 0.
 \end{array} \right\} \quad (6.3)$$

Dabei ist  $c \in \mathbb{R}^n$  der **Kostenvektor** (englisch: *cost vector*),  $A \in \mathbb{R}^{m \times n}$  die **Bedarfs-** oder **Aufwandsmatrix** (englisch: *demand matrix*) und  $b \in \mathbb{R}^m$  der **Ressourcenvektor** (englisch: *resource vector*), wobei  $m \in \mathbb{N}$  ist. Ein LP der Gestalt (6.3) heißt in **kanonischer Form**. Beim Mozartproblem (6.2) ist z. B.

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

Wie können wir ein LP der allgemeinen Form (6.1) in kanonischer Form schreiben? Enthält ein solches LP

- (i) eine Gleichungsbeschränkung  $a^T x = \beta$ , so können wir diese in Form zweier Ungleichungen,  $a^T x \leq \beta$  und  $-a^T x \leq -\beta$ , schreiben.
- (ii) für eine Variable  $x_i$  keine Beschränkung der Form  $x_i \geq 0$  (**freie Variable**, englisch: *free 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 6.4:** Was bedeutet diese Äquivalenz genau? (Siehe auch [Hausaufgabe 4.3](#).)

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

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

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

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

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

$$\begin{aligned} H^-(a, \beta) &:= \{x \in \mathbb{R}^n \mid a^T x \leq \beta\} && \text{negativer Halbraum,} \\ H^+(a, \beta) &:= \{x \in \mathbb{R}^n \mid a^T x \geq \beta\} && \text{positiver Halbraum.} \end{aligned} \quad (6.5)$$

- (iii) Der Durchschnitt endlich vieler abgeschlossener 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).

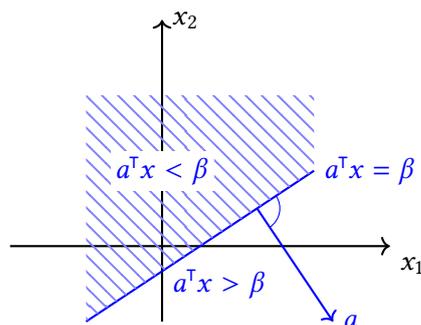


Abbildung 6.2: Darstellung einer Hyperebene mit Normalenvektor  $a$  und der beiden Halbräume.

Ein LP der Gestalt

$$\left. \begin{array}{l} \text{Minimiere } c^T x \text{ über } x \in \mathbb{R}^n \\ \text{sodass } Ax = b \\ \text{und } x \geq 0 \end{array} \right\} \quad (6.6)$$

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

Wie können wir ein LP der allgemeinen Form (6.1) in Normalform schreiben? Enthält ein solches LP

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

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

- (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:

$$\left. \begin{array}{l} \text{Minimiere } -9x_1 - 8x_2 \\ \text{sodass } x_1 + x_2 + s_1 = 6 \\ \quad \quad 2x_1 + x_2 + s_2 = 11 \\ \quad \quad x_1 + 2x_2 + s_3 = 9 \\ \text{und } x_1, x_2 \geq 0 \\ \quad \quad s_1, s_2, s_3 \geq 0. \end{array} \right\} \quad (6.7)$$

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

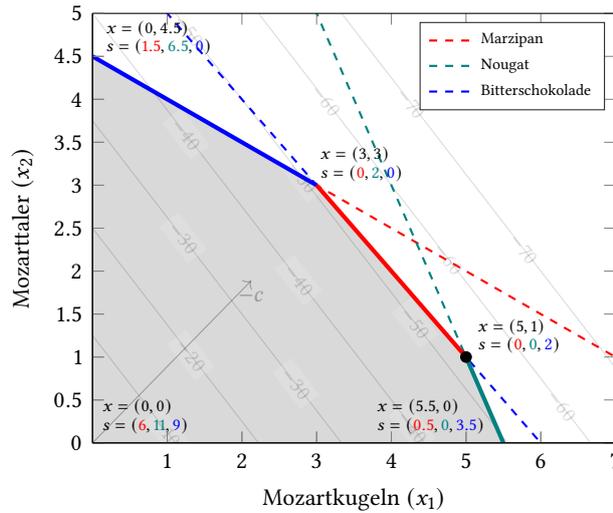


Abbildung 6.3: Darstellung der Ecken der zulässigen Menge beim Mozartproblem in Normalform (6.7) mitsamt den jeweiligen Werten der Schlupfvariablen.

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

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

heißt dann ein **in Normalform beschriebenes Polyeder**, kurz: **Polyeder in Normalform**. Für die Dimension der Matrix  $A \in \mathbb{R}^{m \times n}$  nehmen wir dabei an:<sup>1</sup>

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

### § 6.1 EXISTENZ VON LÖSUNGEN

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

**Vorüberlegung:** Die zulässige Menge  $P \subseteq \mathbb{R}^n$  von (6.6) ist immer abgeschlossen. (**Quizfrage 6.5:** Warum eigentlich?) Falls sie auch nichtleer und beschränkt (also kompakt) ist, dann besitzt die stetige Zielfunktion  $c^T x$  über  $P$  nach dem Satz von Weierstraß bzw. **Satz 1.6** einen Minimierer. Allerdings ist die zulässige Menge  $P$  im Allgemeinen nicht beschränkt, siehe **Abbildung 6.4**. Mit **Satz 1.6** können wir dann nicht argumentieren, da die Sublevelmengen  $L := \{x \in P \mid c^T x \leq m\}$  möglicherweise nicht beschränkt (also nicht kompakt) sind.

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

<sup>1</sup>Für  $m = 0$  ist die Aufgabe entweder unbeschränkt (wenn mindestens 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  $Ax = b$  ist unlösbar, d. h. (6.6) ist unzulässig.

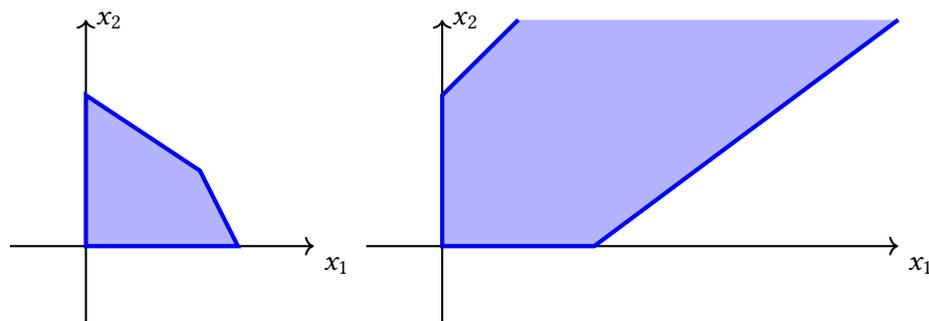


Abbildung 6.4: Kompaktes (abgeschlossenes und beschränktes) Polyeder (links) und unbeschränktes Polyeder (rechts).

**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  $c^\top d \geq 0$  für alle Richtungen in der Menge

$$\{d \in \mathbb{R}^n \mid Ad = 0, d \geq 0\}. \quad (6.9)$$

Die Menge in (6.9) wird **Rezessionskegel** (englisch: *recession cone*) des Polyeders (6.8) genannt.<sup>2</sup>

**Quizfrage 6.6:** Welche Bedeutung hat der Rezessionskegel?

*Beweis.* Wir führen einen Widerspruchsbeweis. Angenommen, es gebe eine Richtung  $d$  aus der Menge (6.9) mit der Eigenschaft  $c^\top d < 0$ . (**Quizfrage 6.7:** Was bedeutet  $c^\top d < 0$ ?) Dann ist  $x + td$  für alle  $t \geq 0$  zulässig, und es gilt

$$c^\top(x + td) = c^\top x + t c^\top d \rightarrow -\infty \text{ für } t \rightarrow \infty.$$

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

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

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 nicht-negativen Linearkombinationen einer gegebenen Menge von Vektoren abgeschlossen ist.

**Lemma 6.10** (Abgeschlossenheit der Menge nicht-negativer Linearkombinationen<sup>3</sup>).

Es sei  $B \in \mathbb{R}^{m \times n}$  eine Matrix (ohne Einschränkungen an die Dimensionen  $n, m \in \mathbb{N}$ ). Die Menge

$$K := \{Bd \mid d \in \mathbb{R}^n, d \geq 0\} \quad (6.10)$$

der nichtnegativen Linearkombinationen der Spalten von  $B$  ist abgeschlossen.

<sup>2</sup>Eine Menge  $K \subseteq \mathbb{R}^n$  heißt ein **Kegel** (englisch: *cone*), wenn  $\beta x \in K$  gilt für  $x \in K$  und alle  $\beta > 0$ .

<sup>3</sup>Beweis aus Werner, 2007, Lemma 1.5

**Quizfrage 6.8:** Wie kann man sich die Menge in (6.10) grafisch vorstellen?

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

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

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  $b_1, \dots, b_n$  erzeugt. Wir müssen zeigen, dass  $K$  abgeschlossen ist.

Wir machen eine Fallunterscheidung:

Fall 1: Falls diese  $n$  Vektoren linear **un**abhängig sind, dann hat  $B \in \mathbb{R}^{m \times n}$  vollen (Spalten-)Rang, also  $\text{Rang}(B) = n$ . Also ist  $B^T B \in \mathbb{R}^{n \times n}$  positiv definit, insbesondere invertierbar. Wir betrachten eine Folge  $d^{(k)} \subseteq \mathbb{R}^n$  nichtnegativer Vektoren, sodass  $Bd^{(k)} \subseteq K$  gilt und  $Bd^{(k)} =: y^{(k)} \rightarrow y$  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:

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

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

Fall 2: Falls die Vektoren  $b_1, \dots, b_n$  linear **ab**hängig sind, dann existieren Zahlen  $\gamma_1, \dots, \gamma_n$  mit der Eigenschaft

$$\sum_{i=1}^n \gamma_i b_i = 0, \quad (6.11)$$

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

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

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

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

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:

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

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

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

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

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

$$K = \{Bd \mid d \in \mathbb{R}^n, d \geq 0\} = \text{cone}\{b_1, \dots, b_n\}. \quad (6.13)$$

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. **Quizfrage 6.9:** Warum können wir o. B. d. A. von einem LP in Normalform ausgehen?

Es sei  $f^* = \inf\{c^T x \mid x \in P\}$  der endliche Optimalwert von (6.6). Es existiert also eine sogenannte Minimalfolge  $x^{(k)} \subseteq P$  mit der Eigenschaft  $c^T x^{(k)} \searrow f^*$ .

Wir betrachten die Folge

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

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

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

Diese Menge ist, abgesehen von der Verschiebung um den konstanten Vektor  $\begin{pmatrix} 0 \\ b \end{pmatrix}$ , von der Bauart (6.10) mit der Matrix  $B = \begin{pmatrix} c^T \\ A \end{pmatrix} \in \mathbb{R}^{(1+m) \times n}$ . Nach Lemma 6.10 ist  $\widetilde{K}$  abgeschlossen. Daraus folgt, dass der Grenzwert  $\begin{pmatrix} f^* \\ 0 \end{pmatrix}$  der Folge  $(y^{(k)})$  in  $\widetilde{K}$  liegt. Das heißt, es existiert ein  $x^* \geq 0$  mit der Eigenschaft  $c^T x^* = f^*$  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** (englisch: **extremal point**) oder **Ecke** (englisch: **vertex**) 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  $P$ .

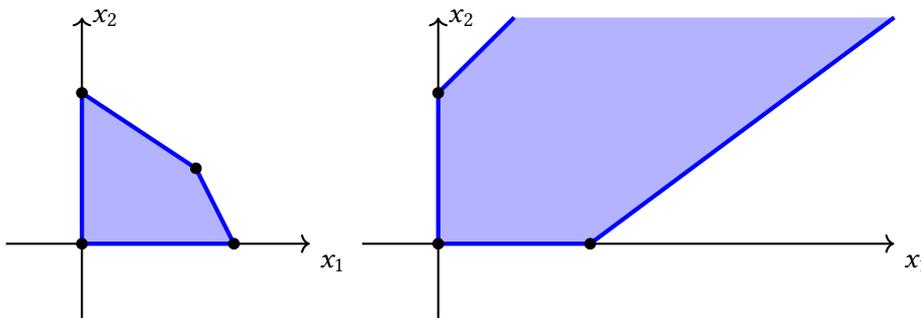


Abbildung 6.5: Polyeder (nicht in Normalform) und ihre Ecken: Viereck im  $\mathbb{R}^2$  und unbeschränktes Polyeder mit drei Ecken im  $\mathbb{R}^2$ .

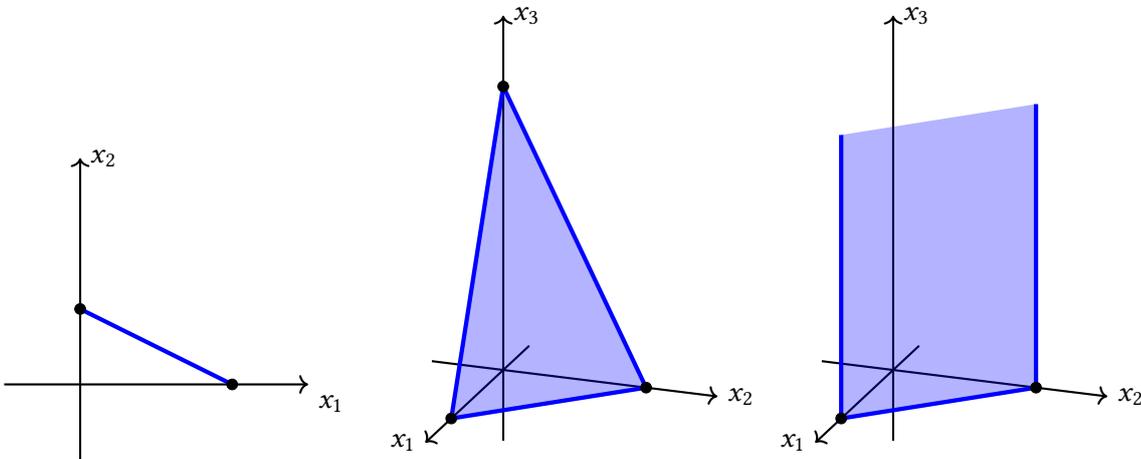


Abbildung 6.6: Polyeder in Normalform und ihre Ecken: Strecke im  $\mathbb{R}^2$  ( $n = 2, m = 1$ ) und Flächen im  $\mathbb{R}^3$  ( $n = 3, m = 1$ ).

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  $I(x) = \{1 \leq i \leq n \mid x_i > 0\}$  die Menge der **inaktiven Indizes** (bzgl. der Ungleichungen  $x \geq 0$ ). Dann sind äquivalent:

- (i)  $x$  ist eine Ecke von  $P$
- (ii) Die Menge der Spalten  $(a_i)_{i \in I(x)}$  von  $A$  ist linear **unabhängig**.

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

*Beweis.* **Aussage (i)  $\Rightarrow$  Aussage (ii):** Es sei  $x$  eine Ecke von  $P$ . Wir nehmen an, dass die Spalten  $(a_i)_{i \in I(x)}$  linear **abhängig** sind. Damit muss natürlich notwendigerweise  $I(x) \neq \emptyset$  sein. Wegen der linearen Abhängigkeit gibt es Koeffizienten  $\gamma_i$ , sodass gilt:

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

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

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

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

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

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  $y \neq z$ .

**Aussage (ii)  $\Rightarrow$  Aussage (i):** Umgekehrt seien nun die Spaltenvektoren  $(a_i)_{i \in I(x)}$  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  $j \notin I(x)$  gilt  $x_j = y_j = z_j = 0$  wegen  $y, z \geq 0$ . Also ist

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

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

**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  $N \subseteq \{1, 2, \dots, n\}$  mit  $|N| = n - m$  und setze  $x_i = 0$  für  $i \in N$ .
- Die restlichen Indizes bilden die Menge  $B = \{1, 2, \dots, n\} \setminus N$  mit  $|B| = m$ .

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

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

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  $B \subseteq \{1, \dots, n\}$  mit  $|B| = m$  eine (geordnete) Indexmenge (ein  $m$ -Tupel) von Spaltenindizes und  $N = \{1, \dots, n\} \setminus B$ .

- Ist die mit den Spaltenindizes  $B$  gebildete Untermatrix  $A_B$  regulär, so heißt die Indexmenge  $B$  eine **Basis** (englisch: **basis**) und  $A_B$  die zugehörige **Basismatrix** (englisch: **basis matrix**).  $N$  heißt dann **Nichtbasis** (englisch: **nonbasis**) und  $A_N$  die zugehörige **Nichtbasismatrix** (englisch: **nonbasis matrix**).
- Es sei  $A_B$  eine Basismatrix. Ein Punkt  $x \in \mathbb{R}^n$  heißt ein **Basisvektor** (englisch: **basic vector**) von  $P$  zur Basis  $B$ , wenn  $A_B x_B = b$  und  $x_N = 0$  gilt.
- In der Literatur wird ein Basisvektor auch häufig als **Basislösung** (englisch: **basic solution**) bezeichnet. Der Begriff „-lösung“ weist darauf hin, dass der Vektor das lineare Gleichungssystem  $Ax = b$  löst. Dieser Praxis folgen wir hier nicht, um Verwechslungen mit Optimallösungen zu vermeiden.
- Ein Basisvektor heißt **zulässig** (englisch: **feasible basic vector**), wenn  $x_B \geq 0$  gilt.
- Ist  $x$  ein Basisvektor zur Basis  $B$ , dann heißen die Komponenten von  $x_B$  **abhängige Variable** (englisch: **dependent variable**) und die Komponenten von  $x_N$  **unabhängige Variable** (englisch: **independent variable**).

**Beachte:** Damit überhaupt eine Basis existiert, muss notwendig  $A$  vollen Rang haben, also  $\text{Rang}(A) = m$  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).

Es seien

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

gegeben. Der Vektor  $x = (2, 0, 0, 2, 6, 0, 3)^T$  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  $x_B = (2, 2, 6, 3)^T \geq 0$ ,  $x_N = (0, 0, 0)^T$  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  $\text{Rang}(A) = m$ . 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.

*Beweis.* Der Beweis ist Inhalt von ??.

□

**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  $\text{Rang}(A) = m$ . 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

$$\text{Minimiere } c^T x \quad \text{sodass } x \in P$$

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  $\text{Rang}(A) = m$  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](#)  $\delta = \min \left\{ -\frac{x_i^*}{\gamma_i} \mid \gamma_i < 0, i \in \mathcal{I}(x^*) \right\} > 0$  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.<sup>4</sup> Zu jeder Basis gehört nur genau ein Basisvektor (der auch unzulässig sein kann).

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

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

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

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

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  $\text{Rang} \left( \begin{bmatrix} A \\ c^T \end{bmatrix} \right) = m+1$ .<sup>5</sup> 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.

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

<sup>4</sup>Hierbei ignorieren wir die Anordnung der Basiselemente, da sie keinen Einfluss auf den zugehörigen Basisvektor hat.

<sup>5</sup>Zu den Gleichungen  $Ax = b$  ist eine neue Zeile dazukommen, die wesentlich ist.

Also gilt  $c^T y = c^T z = f^*$ , 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.  $\square$

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 mit besserem (kleinerem) Funktionswert, bis es keinen besseren Nachbarn 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  $\text{Rang}(A) = m$ .

### § 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

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

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

Durch Einsetzen von (7.1) erhalten wir folgende Darstellungen der Werte der Zielfunktion in Abhän-

gigkeit von  $t \geq 0$ :

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

und

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

Die Größe  $\tilde{c}_N$  bezeichnet man als den Vektor der **reduzierten Kosten** (englisch: *reduced cost vector*). 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.

**Beachte:** Wir können die reduzierten Kosten als Kostenvektor einer reduzierten Aufgabe verstehen, bei der die Basis-Variablen mit Hilfe von  $x_B := A_B^{-1}(b - A_N x_N)$  eliminiert worden sind. **Quizfrage 7.1:** Wie sieht diese reduzierte Aufgabe genau aus?

**Lemma 7.1** (Erkennen einer Lösung).

Es sei  $x$  ein zulässiger Basisvektor zur Basis  $B$ . Wenn für die reduzierten Kosten

$$\tilde{c}_N := c_N - A_N^\top A_B^{-1} c_B \geq 0 \tag{7.4}$$

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  $c^\top x$  und  $c^\top z$  mit einer Rechnung ähnlich wie in (7.3):

$$\begin{aligned}
 c^\top z &= c_B^\top z_B + c_N^\top z_N \\
 &= c_B^\top A_B^{-1}(b - A_N z_N) + c_N^\top z_N \\
 &= c^\top x + (c_N - A_N^\top A_B^{-1} c_B)^\top z_N \\
 &= c^\top x + \tilde{c}_N^\top z_N.
 \end{aligned}$$

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

**Quizfrage 7.2:** 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  $\tilde{c}_N$  noch nicht in allen Einträgen  $\geq 0$  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  $\tilde{c}_r < 0$  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):

$$c^\top x(t) = c^\top x + t \begin{pmatrix} c_B \\ c_N \end{pmatrix}^\top \begin{pmatrix} \Delta x_B \\ e_r \end{pmatrix} = c^\top x + t \underbrace{\tilde{c}_r}_{<0} \rightarrow -\infty \quad \text{für } t \rightarrow \infty.$$

□

**Beachte:** Das ist genau die Situation, die in Lemma 6.8 beschrieben wird: Die Richtung  $d = \begin{pmatrix} \Delta x_B \\ e_r \end{pmatrix}$  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:

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

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:

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

Es ist also  $\ell$  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  $\ell$  in die neue Nichtbasis auf.

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

- (i) die Berechnung der reduzierten Kosten  $\tilde{c}_N$  nach (7.4) (lineares Gleichungssystem mit  $A_B^T$  lösen),
- (ii) die Auswahl eines Index'  $r \in N$  mit  $\tilde{c}_r < 0$ ,
- (iii) die Bestimmung des Vektors  $\Delta x_B$  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  $B^+ := (B \cup \{r\}) \setminus \{\ell\}$  und Nichtbasis  $N^+ := (N \cup \{\ell\}) \setminus \{r\}$ .

**Satz 7.3** (Simplex-Schritt).

Es sei  $x$  ein zulässiger Basisvektor von  $P$  zur Basis  $B$ , und es sei  $N = \{1, \dots, n\} \setminus B$ . Es gelte  $\tilde{c}_r < 0$  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

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

- (ii) Für die Zielfunktionswerte gilt

$$c^T x^+ = c^T x + \hat{t} \tilde{c}_r \leq c^T x.$$

*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:

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

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  $c^\top x^+ = c^\top x$ ).

Es gelten die Voraussetzungen von Satz 7.3.

- (i) Der Fall  $c^\top x^+ = c^\top x$  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** (englisch: **degenerate**).<sup>6</sup>

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

$$c^\top x^+ = c^\top x + \hat{t} \tilde{c}_r < c^\top x.$$

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  $x = (2, 0, 0, 2, 6, 0, 3)^\top$  zur Basis  $B = \{1, 4, 5, 7\}$ . Der Zielfunktionswert ist  $c^\top x = (-2, -3, -4, 0, 0, 0, 0) x = -4$ .

<sup>6</sup>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.

(i) Die reduzierten Kosten sind

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

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

(iii) Wir berechnen

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

und führen den Quotiententest durch:

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

**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  $B^+ = (B \cup \{r\}) \setminus \{\ell\} = \{1, 3, 5, 7\}$  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  $c^T x^+ = -12$ . Wie erwartet hat sich der Funktionswert also um  $\hat{t} \tilde{c}_r = 2(-4) = -8$  verändert.

Würden wir in *Schritt (ii)* stattdessen den Index  $r = 2$  wählen, so erhielten wir  $\Delta x_B = (0, -1, -3, 0)^T$  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  $x^+ = (2, 2, 0, 0, 0, 0, 3)^T$  mit neuem Funktionswert  $c^T x^+ = -10$ .

## § 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  $\text{Rang}(A) = m$  an. Der leichten 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

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

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

4:  $x$  ist eine Lösung von (6.6), **STOP**

5: **else**

6: Wähle einen Index  $r \in N$  mit  $\tilde{c}_r < 0$

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

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

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

14: Setze  $N^+ := (N \cup \{\ell\}) \setminus \{r\}$

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 7.3:** 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  $c^T x^+ < c^T x$  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.  $\square$

Der **Simplex-Algorithmus 7.6** lässt noch Freiheiten

- bei der Wahl der Austausch-Indizes  $r$  in **Zeile 6**
- und evtl. bei der Wahl von  $\ell$  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

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

und

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

gilt.

**Satz 7.8** (Regel von Bland).

*Wählt man in **Zeile 6** den Index  $r$  und in **Zeile 11** den Index  $\ell$  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.  $\square$

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

**Quizfrage 7.5:** 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 7.6:** 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 16) im Simplex-Verfahren (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

$$\left. \begin{array}{l} \text{Maximiere} \quad c^T x \\ \text{sodass} \quad Ax \leq b \\ \text{und} \quad x \geq 0 \end{array} \right\}$$

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  $(x, s) \in \mathbb{R}^n \times \mathbb{R}^m$ :

$$\begin{array}{l} \text{Minimiere} \quad -c^T x \\ \text{sodass} \quad Ax + s = b \\ \text{und} \quad x \geq 0, \quad s \geq 0. \end{array}$$

Falls  $b \geq 0$  ist, dann ist  $\begin{pmatrix} 0 \\ b \end{pmatrix}$  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.)  $b \geq 0$ .<sup>7</sup> Dann gelten für das Hilfsproblem

$$\left. \begin{array}{l} \text{Minimiere } \mathbf{1}^\top z \quad \text{über } (x, z) \in \mathbb{R}^n \times \mathbb{R}^m \\ \text{sodass } Ax + z = b \\ \text{und } x \geq 0, \quad z \geq 0 \end{array} \right\} \quad (7.7)$$

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

- (i) Der Vektor  $\begin{pmatrix} x \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix}$  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  $\begin{pmatrix} x^* \\ z^* \end{pmatrix}$  ein optimaler Basisvektor für (7.7). Ist  $z^* \neq 0$ , so besitzt das LP (6.6) keinen zulässigen Punkt. Ist dagegen  $z^* = 0$  und gilt  $\text{Rang}(A) = m$ , so ist  $x^*$  ein (zulässiger) Basisvektor für (6.6) zu einer geeigneten Basis.

Die Voraussetzung  $b \geq 0$  ist keine Einschränkung, ggf. multiplizieren wir betreffende Zeilen von  $Ax = b$  mit  $-1$ .

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

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

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

Wir betrachten nun den Fall  $z^* = 0$ . Es sei  $B^*$  mit  $|B^*| = m$  eine zu  $\begin{pmatrix} x^* \\ z^* \end{pmatrix}$  gehörige Basis. Es ist also  $[A, \text{Id}]_{B^*}$  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  $\text{Rang}(A) = m$  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  $[A, \text{Id}]$  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  $\text{Rang}(A) = m$  nach endlich vielen Schritten

<sup>7</sup>Über den Rang von  $A$  muss hier nichts vorausgesetzt werden. Der Rang von  $[A, \text{Id}]$  ist immer gleich  $m$ .

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  $\begin{pmatrix} x^* \\ z^* \end{pmatrix}$  entartet, so enthält die Basis  $B^*$  möglicherweise noch Indizes in  $\{n+1, \dots, n+m\}$ , 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  $[A, \text{Id}]$  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  $\{1, \dots, n\}$  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 7.7:** Wieviele Iterationen benötigt das Simplex-Verfahren in Phase I mindestens, um einen zulässigen Basisvektor der Aufgabe (6.6) zu einer Basis in  $\{1, \dots, n\}$  zu finden?

**Bemerkung 7.12** (Zur Komplexität des Simplex-Verfahrens).

Es gibt ein konstruiertes Beispiel von Klee, Minty, 1972<sup>8</sup>, 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. Im Mittel verhält sich das Simplex-Verfahren jedoch deutlich besser als in diesem schlechtesten Fall.

Ende der Woche 5

## § 8 OPTIMALITÄTSBEDINGUNGEN DER LINEAREN OPTIMIERUNG (DUALITÄT)

**Literatur:** Geiger, Kanzow, 2002, Kapitel 3.1.2

Wir betrachten weiterhin ein LP in Normalform, also

$$\left. \begin{array}{l} \text{Minimiere } c^\top x \text{ über } x \in \mathbb{R}^n \\ \text{sodass } Ax = b \\ \text{und } x \geq 0 \end{array} \right\} \quad (8.1)$$

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.

<sup>8</sup>siehe z. B. Hamacher, Klamroth, 2006, S.81

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

$$\left. \begin{array}{l} \text{Maximiere } b^T \lambda \quad \text{über } \lambda \in \mathbb{R}^m \\ \text{sodass } A^T \lambda \leq c. \end{array} \right\} \quad (8.2)$$

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

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

**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** (englisch: *dual LP*). In diesem Zusammenhang heißt (8.1) das **primale LP** (englisch: *primal LP*). Man spricht auch von **primal-dualen Paaren** (englisch: *primal-dual pair*).

**Quizfrage 8.1:** Was ist die duale Aufgabe der dualen Aufgabe (8.3)? (Siehe auch [Hausaufgabe 6.1](#))

**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  $d^*$ :

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

**Quizfrage 8.2:** Welchen Wert hat  $d^*$ , wenn die duale Aufgabe unbeschränkt bzw. unzulässig ist?

**Satz 8.2 (Schwache Dualität)** (englisch: *weak duality*).

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

$$b^T \lambda \leq c^T x.$$

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

*Beweis.* Aus der Zulässigkeit ergibt sich

$$b^T \lambda = (Ax)^T \lambda = x^T (A^T \lambda) = x^T (c - \mu) = c^T x - x^T \mu \leq c^T x, \quad (8.4)$$

denn wegen  $x \geq 0$  und  $\mu \geq 0$  gilt  $x^T \mu \geq 0$ . □

**Bemerkung 8.3** (Veranschaulichung des dualen LPs).

Jedes zulässige Paar  $(\lambda, \mu)$  des dualen LPs (8.3) liefert mit  $b^\top \lambda$  eine untere Schranke für den optimalen Zielfunktionswert des primalen LPs, also

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

Wegen  $Ax = b$  im primalen LP gilt auch  $\lambda^\top Ax = \lambda^\top b$  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.<sup>9</sup>

**Folgerung 8.4** (Erkennen primal-dualer Lösungen).

Es sei  $x^* \in \mathbb{R}^n$  zulässig für das primale LP (8.1), und es sei  $(\lambda^*, \mu^*) \in \mathbb{R}^m \times \mathbb{R}^n$  zulässig für das duale LP (8.3). Falls

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

gilt, dann ist  $x^*$  eine Lösung des primalen LP, und  $(\lambda^*, \mu^*)$  ist eine Lösung des dualen LP.

*Beweis.* Es seien  $x$  und  $(\lambda, \mu)$  irgendwelche zulässigen Punkte für das primale bzw. das duale LP. Aus der schwachen Dualität (Satz 8.2) folgt

$$\underbrace{b^\top \lambda \leq c^\top x^* = b^\top \lambda^* \leq c^\top x}_{\text{primale Optimalität}}, \quad \underbrace{b^\top \lambda^* \leq c^\top x^* = b^\top \lambda^* \leq c^\top x}_{\text{duale Optimalität}}$$

d. h.,  $x^*$  ist eine Lösung von (8.1), und  $(\lambda^*, \mu^*)$  ist Lösung von (8.3). □

Der folgende Satz zeigt, dass das System

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

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

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

$$x^\top \mu = \sum_{i=1}^n x_i \mu_i = 0.$$

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

<sup>9</sup>Eine analoge Aussage gilt natürlich auch für das primale LP. Die Sprechweise von primal-dualen Paaren ist daher gerechtfertigt.

- (i) Ist  $x^*$  eine Lösung für das primale LP (8.1), dann existieren  $(\lambda^*, \mu^*)$ , sodass  $(x^*, \lambda^*, \mu^*)$  das System (8.6) erfüllt.
- (ii) Ist  $(\lambda^*, \mu^*)$  eine Lösung für das duale LP (8.3), dann existiert  $x^*$ , sodass  $(x^*, \lambda^*, \mu^*)$  das System (8.6) erfüllt.
- (iii) Erfüllt  $(x^*, \lambda^*, \mu^*)$  das System (8.6), dann ist  $x^*$  eine Lösung von (8.1), und  $(\lambda^*, \mu^*)$  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  $B^T \xi = c$  besitzt eine Lösung  $\xi \geq 0$ .
- (ii) Es gilt  $c^T d \geq 0$  für alle Elemente der Menge  $\{d \in \mathbb{R}^n \mid B d \geq 0\}$ .

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

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

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

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

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, der Punkt  $c$  aber nicht. Man nennt dann  $H(d, 0)$  eine **trennende Hyperebene** (englisch: *separating hyperplane*).

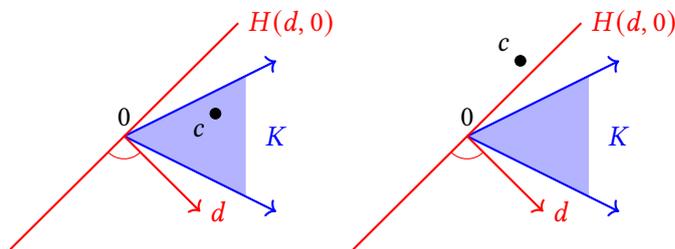


Abbildung 8.1: Illustration der beiden Fälle (links: Aussagen (i) und (ii) sind beide erfüllt und rechts: beide nicht erfüllt) im Farkas-Lemma 8.6.

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

$$c^T d = (B^T \xi)^T d = \xi^T (Bd) \geq 0.$$

Um **Aussage (ii)  $\Rightarrow$  Aussage (i)** zu zeigen, verwenden wir Kontraposition ( $\neg$  **Aussage (i)**  $\Rightarrow$   $\neg$  **Aussage (ii)**). 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 1.6 besitzt (8.7) daher einen globalen Minimierer  $w$ . Der Punkt  $w$  ist gleichzeitig 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 8.3:** Warum können Punkte außerhalb von  $\overline{B_R(c)}$  nicht globaler Minimierer von (8.8) sein?)

**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 8.4:** Warum?). Wir erhalten

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

Daraus folgt

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

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

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

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

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

Außerdem erhalten wir

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

Insgesamt folgt

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

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.  $\square$

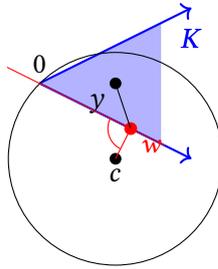


Abbildung 8.2: Illustration der Konstruktion des Normalenvektors  $d = w - c$  der trennenden Hyperebene (rot) im Beweis des Farkas-Lemmas 8.6.

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  $(\lambda^*, \mu^*)$  zulässig sind für (8.1) und (8.3). Wegen (8.4) gilt

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

Folgerung 8.4 zeigt nun, dass  $x^*$  und  $(\lambda^*, \mu^*)$  bereits Lösungen von (8.1) bzw. (8.3) sind.

Aussage (i): Um die notwendigen Bedingungen zu zeigen, benötigen wir das Farkas-Lemma 8.6.<sup>10</sup> Es sei also  $x$  eine Lösung des primalen LP (8.1). Insbesondere ist  $f^*$  endlich, und aus Lemma 6.8 folgt, dass  $c^T d \geq 0$  für alle Richtungen im Rezessionskegel

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

gilt. Setzen wir

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

so ist  $B d \geq 0$  äquivalent zu  $A d = 0$  und  $d \geq 0$ . Es ist also gerade die Aussage (ii) des Farkas-Lemma 8.6 erfüllt. Daraus folgt, dass ein Vektor  $\xi =: (\lambda^+, \lambda^-, \mu) \geq 0$  existiert mit  $B^T \xi = c$ . Setzen wir noch  $\lambda := \lambda^+ - \lambda^-$ , dann folgt  $A^T \lambda + \mu = c$  und  $\mu \geq 0$ . 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  $(\lambda^*, \mu^*)$ , sodass  $b^T \lambda^* = d^*$  gilt. Wir müssen noch zeigen, dass  $d^* = f^*$

<sup>10</sup>Ein direkter Beweis ohne Rückgriff auf das Farkas-Lemma oder den Simplex-Algorithmus findet sich bei Forsgren, 2008.

gilt und nicht etwa  $d^* < f^*$ . Dann folgt aus (8.12) die noch fehlende Komplementaritätsbedingung  $(x^*)^\top \mu^* = 0$ , 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  $\neg$  Aussage (i)  $\Rightarrow \neg$  Aussage (ii). Es sei dazu  $\varepsilon > 0$  beliebig. Wir wissen, dass das System

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

für  $\begin{pmatrix} x \\ \alpha \end{pmatrix} \geq 0$  nicht lösbar ist, denn das würde bedeuten:  $Ax = b$ ,  $x \geq 0$  und  $c^\top x \leq c^\top x + \alpha = f^* - \varepsilon$ ; 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  $\begin{pmatrix} b \\ f^* - \varepsilon \end{pmatrix}^\top d < 0$ . Wir partitionieren  $d =: \begin{pmatrix} -\lambda \\ \alpha \end{pmatrix}$  und erhalten

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

also

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

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

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

Es muss also  $\alpha > 0$  sein, und wir können durch Skalierung  $\alpha = 1$  in (8.13) erreichen.<sup>11</sup> Damit gilt also nun

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

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

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

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  $(\lambda, \mu)$  implizit vorkommen. Bei der Berechnung der reduzierten Kosten treten die Größen

$$\begin{aligned} \lambda &:= A_B^{-\top} c_B, \\ \tilde{c}_N &:= \mu_N := c_N - A_N^\top \lambda, \\ \mu_B &:= 0 \end{aligned} \quad (8.14)$$

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

<sup>11</sup>Wir ersetzen dazu  $\alpha$  durch  $\alpha/\alpha = 1$  und  $\lambda$  durch  $\lambda/\alpha$ .

Genauer heißt **Algorithmus 7.6** auch **primaler Simplex-Algorithmus** (englisch: *primal simplex algorithm*). Es gibt auch ein **duales Simplex-Verfahren**, 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:

- $f^* \in \mathbb{R}$
- (Satz 6.9)  $\Leftrightarrow$  das primale Problem (8.1) besitzt eine Lösung
- (Satz 8.5)  $\Leftrightarrow$  die Optimalitätsbedingungen (8.6) besitzen eine Lösung
- (Satz 8.5)  $\Leftrightarrow$  das duale Problem (8.3) besitzt eine Lösung
- (Satz 6.9)  $\Leftrightarrow d^* \in \mathbb{R}$ .

Für „ $\Leftarrow$ “ 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 8.2 schwachen 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 8.5:** Kennen Sie ein Beispiel? (Siehe auch **Hausaufgabe 6.3.**) )  $\square$

**Bemerkung 8.8** (Starke Dualität).

Zu der Erkenntnis  $d^* = f^*$  im Fall (I) sagt man auch: „Es tritt **keine Dualitätslücke** auf“ (zwischen den Optimalwerten, englisch: **no duality gap**) oder „Es herrscht **starke Dualität**“ (englisch: **strong duality**).

	Eigenschaft	primales Simplex-Verfahren	duales Simplex-Verfahren
primale Zulässigkeit	$x_B \geq 0$	✓	erst in der Lösung
	$x_N = 0$	✓	✓
	$Ax = b$	✓	✓
duale Zulässigkeit	$\mu_B = 0$	✓	✓
	$\mu_N \geq 0$	erst in der Lösung	✓
	$A^T \lambda + \mu = c$	✓	✓
Komplementarität	$x^T \mu = 0$	✓	✓

Tabelle 9.1: Unterschiede zwischen primalem und dualem Simplex-Verfahren.

## § 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** (englisch: *dual simplex method*). 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

$$\begin{aligned}
 \lambda &:= A_B^{-T} c_B, \\
 x_B &:= A_B^{-1} b, & \mu_B &:= 0, \\
 x_N &:= 0, & \mu_N &:= c_N - A_N^T \lambda
 \end{aligned} \tag{9.1}$$

sowohl die primalen wie auch die dualen Variablen eindeutig fest.

Eine Basis  $B$  heißt **primal zulässig** (englisch: *primal feasible*), 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** (englisch: *dual feasible*), wenn das durch (9.1) beschriebene Paar von Vektoren  $(\lambda, \mu)$  dual zulässig ist, also die Bedingung  $\mu_N \geq 0$  erfüllt. Im Unterschied zum primalen Simplex-Verfahren werden wir mit primal unzulässigen Basisvektoren arbeiten. Dafür sind die Größen  $(\lambda, \mu)$  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  $(\lambda, \mu)$  die dazugehörigen dualen Variablen

gemäß (9.1). Zur Motivation des *pricing*-Schritts untersuchen wir, was passiert, wenn wir einem der Indizes in  $\mu_B = 0$  erlauben, sich von der Null zu lösen. Wir machen also den Ansatz  $\mu_B(t) := t e_\ell$  mit einem Standard-Basisvektor  $e_\ell \in \mathbb{R}^m$ ,  $t \geq 0$ . In Abhängigkeit von  $t$  ergibt sich der Wert von  $\lambda$  nun aus

$$A_B^\top \lambda(t) + \mu_B(t) = c_B,$$

also

$$\lambda(t) = A_B^{-\top}(c_B - t e_\ell) = \lambda + t \underbrace{(-A_B^{-\top} e_\ell)}_{=:\Delta\lambda}.$$

Welchen Index  $\ell$  wählen wir? Dazu betrachten wir die Werte der dualen Zielfunktion:

$$b^\top \lambda(t) = b^\top \lambda - t b^\top A_B^{-\top} e_\ell = b^\top \lambda - t e_\ell^\top x_B = b^\top \lambda - t x_\ell.$$

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 9.1:** Begründung?)

Nach diesem **pricing**-Schritt berechnen wir  $\Delta\lambda := -A_B^{-\top} e_\ell$ . Die Aufdatierung von  $\mu_N$  erhalten wir aus

$$\mu_N(t) = c_N - A_N^\top \lambda(t) = c_N - A_N^\top (\lambda + t \Delta\lambda) = \mu_N + t \underbrace{(-A_N^\top \Delta\lambda)}_{=:\Delta\mu_N}.$$

Die Wahl der Schrittweite ergibt sich aus der Bedingung der dualen Zulässigkeit, also  $\mu_N(t) \geq 0$ . Wir erhalten ähnlich zum primalen **Quotiententest** (englisch: *ratio test*)

$$\hat{t} := \min \left\{ -\frac{\mu_i}{\Delta\mu_i} \mid i \in N, \Delta\mu_i < 0 \right\} = -\frac{\mu_r}{\Delta\mu_r}.$$

Falls  $\Delta\mu_N \geq 0$  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

$$\lambda^+ := \lambda + \hat{t} \Delta\lambda \quad \text{und} \quad \mu_i^+ := \begin{cases} \mu_i + \hat{t} \Delta\mu_i & \text{für } i \in N, i \neq r, \\ \hat{t} & \text{für } i = \ell, \\ 0 & \text{sonst} \end{cases}$$

und die Basis/Nichtbasis auf:

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

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

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  $(\lambda, \mu)$  umzubenennen.

**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 (dualen Variablen)

$$\begin{aligned}\lambda &:= A_B^{-T} c_B \\ \mu_N &:= c_N - A_N^T \lambda \\ \mu_B &:= 0\end{aligned}$$

- 3: **if**  $\mu_N \geq 0$  **then**
- 4:  $x$  ist eine Lösung von (8.1), und  $(\lambda, \mu)$  ist eine Lösung von (8.3), **STOP**
- 5: **else**
- 6: Wähle einen Index  $r \in N$  mit  $\mu_r < 0$
- 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

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

- 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  $B^+ := (B \cup \{r\}) \setminus \{\ell\}$
- 14: Setze  $N^+ := \{1, \dots, n\} \setminus B^+$
- 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  $(\lambda, \mu)$  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)

$$\begin{aligned}x_B &:= A_B^{-1} b \\ x_N &:= 0\end{aligned}$$

- 3: **if**  $x_B \geq 0$  **then**
- 4:  $(\lambda, \mu)$  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

$$\begin{aligned}\Delta \lambda &:= -A_B^{-T} e_\ell \\ \Delta \mu_N &:= -A_N^T \Delta \lambda\end{aligned}$$

- 8: **if**  $\Delta \mu_N \geq 0$  **then**
- 9: Aufgabe (8.3) ist unbeschränkt, **STOP**
- 10: **else**
- 11: Bestimme  $\hat{t} \geq 0$  und  $r \in N$  gemäß

$$\hat{t} := \min \left\{ -\frac{\mu_i}{\Delta \mu_i} \mid i \in N, \Delta \mu_i < 0 \right\} = -\frac{\mu_r}{\Delta \mu_r}$$

- 12: Setze  $\lambda^+ := \lambda + \hat{t} \Delta \lambda$  und

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

- 13: Setze  $B^+ := (B \cup \{r\}) \setminus \{\ell\}$
- 14: Setze  $N^+ := \{1, \dots, n\} \setminus B^+$
- 15: Setze  $\lambda := \lambda^+$  und  $\mu := \mu^+$
- 16: Setze  $B := B^+$  und  $N := N^+$
- 17: Setze  $k := k + 1$

- 18: **end if**
- 19: **end if**
- 20: Gehe zu Zeile 2

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  $(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^n$  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

$$\begin{aligned} \text{Minimize} \quad & \begin{pmatrix} c \\ \bar{c} \end{pmatrix}^\top \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \quad \text{über} \quad \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \in \mathbb{R}^{n+1} \\ \text{sodass} \quad & [A \quad \bar{a}] \begin{pmatrix} x \\ \bar{x} \end{pmatrix} = b \\ \text{und} \quad & \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \geq 0 \end{aligned} \tag{9.2}$$

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

$$\begin{bmatrix} A^\top \\ \bar{a}^\top \end{bmatrix} \lambda + \begin{pmatrix} \mu \\ \bar{\mu} \end{pmatrix} = \begin{pmatrix} c \\ \bar{c} \end{pmatrix}, \quad \begin{pmatrix} \mu \\ \bar{\mu} \end{pmatrix} \geq 0. \tag{9.3}$$

Wir können die neue duale Schlupfvariable  $\bar{\mu}$  mit  $\bar{c} - \bar{a}^\top \lambda$  initialisieren, aber sie wird i. A. nicht  $\bar{\mu} \geq 0$  erfüllen. Die Komplementaritätsbedingung  $x^\top \mu + \bar{x}^\top \bar{\mu} = 0$  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.

## HINZUFÜGEN EINER NEBENBEDINGUNG

Wir betrachten jetzt eine andere Veränderung der primalen Aufgabe (8.1) und fügen ihr eine neue Ungleichungsnebenbedingung  $\bar{a}^\top x \leq \bar{b}$  bzw.  $\bar{a}^\top x + \bar{x} = \bar{b}$  mit zugehöriger Schlupfvariable  $\bar{x}$  hinzu:

$$\begin{aligned} & \text{Minimiere} && \begin{pmatrix} c \\ 0 \end{pmatrix}^\top \begin{pmatrix} x \\ \bar{x} \end{pmatrix} && \text{über} && \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \\ & \text{sodass} && \begin{bmatrix} A & 0 \\ \bar{a}^\top & 1 \end{bmatrix} x = \begin{pmatrix} b \\ \bar{b} \end{pmatrix} && && (9.4) \\ & \text{und} && \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \geq 0. && && \end{aligned}$$

Die Bedingungen der dualen Zulässigkeit für die neue Aufgabe lauten

$$\begin{bmatrix} A^\top & \bar{a} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \lambda \\ \bar{\lambda} \end{pmatrix} + \begin{pmatrix} \mu \\ \bar{\mu} \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \mu \\ \bar{\mu} \end{pmatrix} \geq 0. \quad (9.5)$$

Die bisherige Lösung  $x$  wird i. A. nicht länger primal zulässig sein, da  $\bar{a}^\top x \leq \bar{b}$  verletzt ist. Wir können aber die bisherige dual optimale Lösung durch  $\bar{\lambda} = \bar{\mu} = 0$  erweitern und sind weiterhin dual zulässig. Genauer erweitern wir die bisherige Basis zu  $B \cup \{n+1\}$ . Die neue Basismatrix ist daher

$$\begin{bmatrix} A_B & 0 \\ \bar{a}_B^\top & 1 \end{bmatrix}.$$

Diese ist weiterhin regulär (**Quizfrage 9.2:** 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*

$$\left. \begin{aligned} & \text{Minimiere} && c^\top x && \text{über} && x \in \mathbb{Z}^n \\ & \text{sodass} && Ax = b \\ & \text{und} && x \geq 0. \end{aligned} \right\} \quad (9.6)$$

*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 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

$$\left. \begin{array}{l} \text{Minimiere } c^\top x \text{ über } x \in \mathbb{R}^n \\ \text{sodass } Ax = b \\ \text{und } x \geq 0 \\ \text{sowie } x_i \geq \lceil x_i^* \rceil \end{array} \right\} \left\{ \begin{array}{l} \text{Minimiere } c^\top x \text{ über } x \in \mathbb{R}^n \\ \text{sodass } Ax = b \\ \text{und } x \geq 0 \\ \text{sowie } x_i \leq \lfloor x_i^* \rfloor \end{array} \right. \quad (9.7)$$

gelöst. Dabei sind  $\lceil \cdot \rceil$  und  $\lfloor \cdot \rfloor$  die **obere** bzw. **untere Gaußklammer** (englisch: ceiling and floor functions), d. h.,

$$\begin{aligned} \lceil z \rceil &:= \min\{y \in \mathbb{Z} \mid y \geq z\} && \text{(kleinste ganze Zahl oberhalb von } z), \\ \lfloor z \rfloor &:= \max\{y \in \mathbb{Z} \mid y \leq z\} && \text{(kleinste ganze Zahl unterhalb von } z). \end{aligned}$$

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 10.1:** 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 10.2:** Was sind Beispiele von Veränderungen in der Aufgabenstellungen, die *nicht* durch Änderungen in  $b$  und/oder  $c$  dargestellt werden können? (Siehe auch [Hausaufgaben 7.3](#) und [7.4](#).)

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  $(\lambda^*, \mu^*)$  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  $t \in \mathbb{R}$ :

$$\left. \begin{array}{l} \text{Minimiere } (c + t \Delta c)^T x \\ \text{sodass } Ax = b \\ \text{und } x \geq 0 \end{array} \right\} \left\{ \begin{array}{l} \text{Maximiere } b^T \lambda \\ \text{sodass } A^T \lambda + \mu = c + t \Delta c \\ \text{und } \mu \geq 0. \end{array} \right. \quad (10.1)$$

Welche Aussagekraft besitzen die Lösungen  $x^*$  und  $(\lambda^*, \mu^*)$  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):

$$\lambda(t) = \lambda^* + t \Delta \lambda \quad \text{mit } \Delta \lambda := A_B^{-T} \Delta c_B \quad (10.2a)$$

$$\mu_N(t) = \mu_N^* + t \Delta \mu_N \quad \text{mit } \Delta \mu_N := \Delta c_N - A_N^T \Delta \lambda \quad (10.2b)$$

$$\mu_B(t) \equiv \mu_B^* = 0. \quad (10.2c)$$

Wann sind die auf diese Art und Weise erhaltenen Vektoren  $x^*$  und  $(\lambda(t), \mu(t))$  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:

$$x_B^* \underbrace{\mu_B(t)}_{=0} + \underbrace{x_N^*}_{=0} \mu_N(t) = 0.$$

Bzgl. der dualen Zulässigkeit ist die erste Bedingung

$$\begin{aligned} A^T \lambda(t) + \mu(t) &= \begin{pmatrix} A_B^T \lambda^* + t A_B^T A_B^{-T} \Delta c_B + \mu_B(t) \\ A_N^T \lambda^* + t A_N^T A_B^{-T} \Delta c_B + \mu_N(t) \end{pmatrix} = \begin{pmatrix} c_B + t \Delta c_B + 0 \\ A_N^T \lambda^* + t A_N^T A_B^{-T} \Delta c_B + \mu_N^* + t \Delta c_N - t A_N^T A_B^{-T} \Delta c_B \end{pmatrix} \\ &= \begin{pmatrix} c_B + t \Delta c_B \\ c_N + t \Delta c_N \end{pmatrix} = c + t \Delta c \end{aligned}$$

nach Konstruktion von  $\lambda(t)$  und  $\mu(t)$  erfüllt. Die Vorzeichenbedingung  $\mu_N(t) \geq 0$  jedoch gilt nicht automatisch, sondern genau dann, wenn der Störungsparameter  $t$  der Bedingung<sup>12</sup>

$$\sup_{\substack{i \in N \\ \Delta \mu_i > 0}} \underbrace{\left\{ -\frac{\mu_i^*}{\Delta \mu_i} \right\}}_{\leq 0} \leq t \leq \inf_{\substack{i \in N \\ \Delta \mu_i < 0}} \underbrace{\left\{ -\frac{\mu_i^*}{\Delta \mu_i} \right\}}_{\geq 0}. \quad (10.3)$$

<sup>12</sup>Wir schreiben hier sup und inf statt max und min, da die betreffenden Indexmengen durchaus leer sein können. Beispielsweise ist, sofern  $\mu_N \geq 0$  gilt, die Indexmenge für die obere Schranke leer, sodass die betreffende Ungleichung in (10.3) zu „ $t \leq \inf \emptyset = \infty$ “ wird, was als  $t < \infty$  zu interpretieren ist.

genügt.

**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:

$$f^*(t) = (c + t \Delta c)^T x^* = f^* + t \Delta c^T x^*. \quad (10.4)$$

Wir fassen unsere Erkenntnisse zusammen:

**Satz 10.1** (Sensitivitätssatz bei LP bei Änderungen im Kostenvektor).

Es seien  $x^*$  und  $(\lambda^*, \mu^*)$  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)<sub>primal</sub> weiterhin ein optimaler Basisvektor, und  $(\lambda(t), \mu(t))$  aus (10.2) ist ein optimaler Basisvektor für (10.1)<sub>dual</sub>. Der gemeinsame Optimalwert beider Aufgaben ist  $c^T x^* + t (\Delta c)^T x^*$ .
- (ii) Ist die rechte Grenze des Intervalls (10.3) echt positiv, dann ist die Optimalwertfunktion

$$c \mapsto \Phi(c) := \text{gemeinsamer Optimalwert von (8.1) und (8.3)}$$

an der Stelle  $c$  in Richtung  $\Delta c$  (einseitig) richtungsdiffbar, und die Richtungsableitung ist gegeben durch

$$\Phi'(c; \Delta c) = (\Delta c)^T x^*.$$

- (iii) Ist  $\mu^*$  nicht entartet, gilt also  $\mu_N^* > 0$ , dann ist die Optimalwertfunktion in einer offenen Kugel  $B_r(c)$  von  $c$  linear mit

$$\Phi(c + \Delta c) = (c + \Delta c)^T x^* \quad \text{für } \Delta c \in B_r(0).$$

Damit ist  $\Phi$  überall in dieser Kugel differenzierbar, und es gilt

$$\Phi'(c + \Delta c) \equiv (x^*)^T \quad \text{für } \Delta c \in B_r(0).$$

**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  $\Phi$  betrachten wir den Differenzenquotienten für solche  $t$ :

$$\frac{\Phi(c + t \Delta c) - \Phi(c)}{t} = \frac{c^T x^* + t (\Delta c)^T x^* - c^T x^*}{t} = (\Delta c)^T x^*,$$

also ist das auch der Wert im Grenzwert  $t \searrow 0$ , der Richtungsableitung  $\Phi'(c; \Delta c)$ .

**Aussage (iii):** Wenn  $\mu^*$  nicht entartet ist, dann enthält das zulässige Intervall (10.3) für jede beliebige Richtung  $\Delta c$  immer ein offenes Intervall um die 0. Wir müssen aber zeigen, dass die Länge dieses Intervalls über alle Richtungen  $\Delta c$  konstanter Norm gleichmäßig von 0 weg beschränkt bleibt. Wir zeigen dazu, dass die Funktion, die die obere Intervallgrenze angibt,

$$\Delta c \mapsto \inf_{\substack{i \in N \\ \Delta \mu_i < 0}} \underbrace{\left\{ -\frac{\mu_i^*}{\Delta \mu_i} \right\}}_{\geq 0}, \quad (10.5)$$

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  $\Delta \mu_N$  linear (und damit stetig) von  $\Delta c$  ab:

$$\Delta \mu_N = \Delta c_N - A_N^T A_B^{-T} \Delta c_B.$$

Da die Sphäre kompakt ist, existiert für jede Komponente  $i \in N$  von  $\Delta \mu_N$  ein endliches

$$\beta_i := \max\{\Delta \mu_i = [\Delta c_N - A_N^T A_B^{-T} \Delta c_B]_i \mid \|\Delta c\| = 1\}.$$

Es gilt  $\beta_i > 0$  für alle  $i \in N$ . (**Quizfrage 10.3:** Warum?) Wir setzen nun  $\beta := \max\{\beta_i \mid i \in N\} > 0$  und  $\alpha := \min\{\mu_i^* \mid i \in N\} > 0$ .

Für beliebiges  $\Delta c$  aus der Einheitssphäre und das zugehörige  $\Delta \mu$  gilt: Falls  $\Delta \mu_N \geq 0$  ist, dann erhalten wir

$$\inf_{\substack{i \in N \\ \Delta \mu_i < 0}} \left\{ -\frac{\mu_i^*}{\Delta \mu_i} \right\} = \inf \emptyset = \infty.$$

Andernfalls gilt

$$\inf_{\substack{i \in N \\ \Delta \mu_i < 0}} \left\{ -\frac{\mu_i^*}{\Delta \mu_i} \right\} = \min_{\substack{i \in N \\ \Delta \mu_i < 0}} \left\{ -\frac{\mu_i^*}{\Delta \mu_i} \right\} \geq \frac{\min_{i \in N} \mu_i^*}{\max_{\substack{i \in N \\ \Delta \mu_i < 0}} \{-\Delta \mu_i\}} \geq \frac{\alpha}{\max_{\substack{i \in N \\ \Delta \mu_i < 0}} \{-\Delta \mu_i\}} \geq \frac{\alpha}{\beta} > 0.$$

Die letzte Ungleichung gilt, da wir jeden der im Nenner vorkommenden Werte mit  $0 < -\Delta \mu_i \leq \beta_i \leq \beta$  abschätzen können und daher auch  $\max\{-\Delta \mu_i \mid i \in N, \Delta \mu_i < 0\} \leq \beta$  gilt. Zusammenfassend bekommen wir also die gewünschte Aussage

$$\inf_{\|\Delta c\|=1} \inf_{\substack{i \in N \\ \Delta \mu_i < 0}} \left\{ -\frac{\mu_i^*}{\Delta \mu_i} \right\} \geq \frac{\alpha}{\beta} =: r > 0.$$

Daraus folgt, dass die Vereinigung der Menge aller zulässigen Störungen  $t \Delta c$  die offene Kugel  $B_r(c)$  enthält:

$$\bigcup_{\|\Delta c\|=1} \{t \Delta c \mid t \in I(\Delta c)\} \supseteq \bigcup_{\|\Delta c\|=1} \{t \Delta c \mid t \in (-r, r)\} = B_r(c).$$

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  $\Phi(c + \Delta c) = \Phi(c) + \Delta c^T x^* = (c + \Delta c)^T x^*$  hat. Die Differenzierbarkeit von  $\Phi$  in  $B_r(c)$  mit Ableitung  $(x^*)^T$  ist eine unmittelbare Konsequenz.  $\square$

## 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

$$\left. \begin{array}{l} \text{Minimiere } c^\top x \\ \text{sodass } Ax = b + t \Delta b \\ \text{und } x \geq 0 \end{array} \right\} \left\{ \begin{array}{l} \text{Maximiere } (b + t \Delta b)^\top \lambda \\ \text{sodass } A^\top \lambda + \mu = c \\ \text{und } \mu \geq 0. \end{array} \right. \quad (10.6)$$

Dieses Mal ist  $(\lambda^*, \mu^*)$  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:

$$x_B(t) = x_B + t \Delta x_B \quad \text{mit } \Delta x_B := A_B^{-1} \Delta b \quad (10.7a)$$

$$x_N(t) \equiv x_N^* = 0. \quad (10.7b)$$

Wann sind die auf diese Art und Weise erhaltenen Vektoren  $x(t)$  und  $(\lambda^*, \mu^*)$  optimal für (10.6)? Wir überprüfen dazu die Optimalitätsbedingungen (8.6). Die duale Zulässigkeit

$$A^\top \lambda^* + \mu^* = c, \quad \mu^* \geq 0$$

ist erfüllt, ebenso die Komplementaritätsbedingung:

$$x_B(t) \underbrace{\mu_B^*}_{=0} + x_N(t) \underbrace{\mu_N^*}_{=0} = 0.$$

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

$$\sup_{\substack{i \in B \\ \Delta x_i > 0}} \underbrace{\left\{ -\frac{x_i^*}{\Delta x_i} \right\}}_{\leq 0} \leq t \leq \inf_{\substack{i \in B \\ \Delta x_i < 0}} \underbrace{\left\{ -\frac{x_i^*}{\Delta x_i} \right\}}_{\geq 0} \quad (10.8)$$

genügt.

Für diese  $t$  ist also tatsächlich  $(\lambda^*, \mu^*)$  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:

$$d^*(t) = (b + t \Delta b)^\top \lambda^* = d^* + t \Delta b^\top \lambda^*. \quad (10.9)$$

Die Erkenntnisse fassen wir wie folgt zusammen:

**Satz 10.2** (Sensitivitätssatz bei LP bei Änderungen in der rechten Seite).

Es seien  $x^*$  und  $(\lambda^*, \mu^*)$  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  $(\lambda^*, \mu^*)$  für (10.6)<sub>dual</sub> weiterhin ein optimaler Basisvektor, und  $x(t)$  aus (10.7) ist ein optimaler Basisvektor für (10.6)<sub>primal</sub>. Der gemeinsame Optimalwert beider Aufgaben ist  $b^T \lambda^* + t (\Delta b)^T \lambda^*$ .
- (ii) Ist die rechte Grenze des Intervalls (10.8) echt positiv, dann ist die Optimalwertfunktion

$$b \mapsto \Psi(b) := \text{gemeinsamer Optimalwert von (8.1) und (8.3)}$$

an der Stelle  $b$  in Richtung  $\Delta b$  (einseitig) richtungsdiffbar, und die Richtungsableitung ist gegeben durch

$$\Psi'(b; \Delta b) = (\Delta b)^T \lambda^*.$$

- (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

$$\Psi(b + \Delta b) = (b + \Delta b)^T \lambda^* \quad \text{für } \Delta b \in B_r(0).$$

Damit ist  $\Psi$  überall in dieser Kugel differenzierbar, und es gilt

$$\Psi'(b + \Delta b) \equiv (\lambda^*)^T \quad \text{für } \Delta b \in B_r(0).$$

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

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 1 \end{pmatrix}, \quad c = \begin{pmatrix} -9 \\ -8 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad b = \begin{pmatrix} 6 \\ 11 \\ 9 \\ 0 \end{pmatrix} \begin{array}{l} \text{Marzipan} \\ \text{Nougat} \\ \text{Schokolade} \end{array}$$

gegeben. Die eindeutige primal optimale Lösung ist  $x^* = (5, 1, 0, 0, 2)^T$  zur Basis  $B = \{1, 2, 5\}$ . Auch die duale Lösung  $(\lambda^*, \mu^*)$  ist eindeutig, und zwar  $\lambda^* = (-7, -1, 0)^T$  und  $\mu^* = (0, 0, 7, 1, 0)^T$ . Beide Basisvektoren sind nicht entartet, denn  $x^*$  hat Nulleinträge nur in der Nichtbasis  $N = \{3, 4\}$ , und  $\mu^*$  hat Nulleinträge nur in der Basis  $B$ . Wir erhalten also folgende Darstellung des Optimalwerts als Funktion des Ressourcenvektors  $b$

$$\Psi(b + \Delta b) = (b + \Delta b)^T \lambda^* = -53 + (\Delta b)^T \begin{pmatrix} -7 \\ -1 \\ 0 \end{pmatrix}$$

für  $\Delta b$  mit hinreichend kleiner Norm. Das bedeutet beispielsweise, dass wir pro Einheit an Marzipan, das wir zusätzlich zur Verfügung haben ( $\Delta b = (1, 0, 0)^T$ ), 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** (englisch: shadow price). 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:

$$\begin{aligned} \sup_{\substack{i \in B \\ \Delta x_i > 0}} \underbrace{\left\{ -\frac{x_i^*}{\Delta x_i} \right\}}_{\leq 0} &\leq t \leq \inf_{\substack{i \in B \\ \Delta x_i < 0}} \underbrace{\left\{ -\frac{x_i^*}{\Delta x_i} \right\}}_{\geq 0} \\ \Leftrightarrow \max_{i=2} \left\{ -\frac{1}{2} \right\} &\leq t \leq \min_{\substack{i=1 \\ i=5}} \left\{ -\frac{5}{-1}, -\frac{2}{-3} \right\} \\ \Leftrightarrow -\frac{1}{2} &\leq t \leq \frac{2}{3}. \end{aligned}$$

Bis zu 2/3 Einheiten Marzipan können also zugekauft werden, ohne dass sich die Lösungsstruktur ändert.

**Quizfrage 10.4:** Was passiert an der Stelle  $t = 2/3$ ? Wie sieht die Rechnung aus, wenn man stattdessen  $\Delta b = (0, 1, 0)^T$  (Nougat) verwendet? Und bei  $\Delta b = (0, 0, 1)^T$  (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** (englisch: transportation network). 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** (englisch: supplies) der Produktionsstätten sowie die **Bedarfe** (englisch: demands) 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

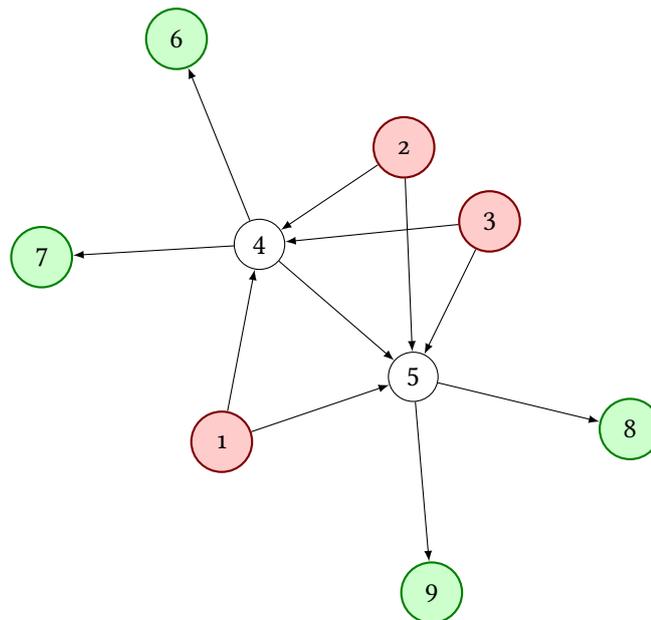


Abbildung 11.1: Transportnetzwerk eines Unternehmens (siehe [Beispiel 11.1](#)) mit Produktionsstätten 1–3 (rot), Zwischenlagern 4–5 und Verkaufsstätten 6–9 (grün).

*Belegung der Kanten mit den darüber zu transportierenden Warenmengen, sodass (unter Beachtung aller Restriktionen) die Gesamttransportkosten minimiert werden.*

**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 11.1:** 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

$$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9\} \quad (11.1a)$$

$$E = \{(1, 4), (1, 5), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5), (4, 6), (4, 7), (5, 8), (5, 9)\}. \quad (11.1b)$$

Diese Beschreibung ist allerdings für die Formulierung von Optimierungsaufgaben ungeeignet.

**Definition 11.3** (Inzidenzmatrix). *Es sei  $(V, E)$  ein einfacher Digraph. Die Knotenmenge sei  $V = \{v_1, v_2, \dots, v_m\}$  und die Kantenmenge  $E = \{e_1, e_2, \dots, e_n\}$ . 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

$$A = \begin{bmatrix} -1 & -1 & \cdot \\ \cdot & \cdot & -1 & -1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & -1 & -1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & 1 & \cdot & 1 & \cdot & -1 & -1 & -1 & \cdot & \cdot \\ \cdot & 1 & \cdot & 1 & \cdot & 1 & 1 & \cdot & \cdot & -1 & -1 \\ \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot \\ \cdot & 1 & \cdot \\ \cdot & 1 \end{bmatrix} \in \mathbb{R}^{9 \times 11}. \tag{11.2}$$

Der besseren Lesbarkeit wegen wurden die Nullen durch „ $\cdot$ “ ersetzt.

**Quizfrage 11.2:** 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  $\mathbf{1}^T A = 0$ .

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  $Ax$  gibt also offenbar den Vektor der **Knotenbilanzen** (englisch: *nodal balances*) 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:

$$\sum_{j=1}^n c_j x_j = c^\top x.$$

Dabei sind  $c_j$  die gegebenen Transportkosten pro Wareneinheit über die Kante  $j$ . Weiter sind die geforderten Bilanzen  $b_i$  aller Knoten  $i = 1, \dots, m$  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

$$Ax = b$$

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:

$$0 \leq x_i \leq u_i \quad \text{für alle } i = 1, \dots, n.$$

**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**  $Ax = b$  erfüllt, heißt ein **Fluss** oder **Flussvektor** auf diesem Netzwerk. Ein Fluss heißt **zulässig**, wenn zusätzlich die **Kapazitätsbeschränkungen**  $0 \leq x \leq u$  erfüllt sind.
- (iii) Eine lineare Optimierungsaufgabe der Form

$$\begin{aligned} &\text{Minimiere} && c^\top x && \text{über } x \in \mathbb{R}^n \\ &\text{unter} && Ax = b \\ &\text{sowie} && 0 \leq x \leq u \end{aligned} \tag{11.3}$$

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) dargestellten Digraphen aus [Beispiel 11.1](#) und die Beispieldaten

$$\begin{aligned} b &= (-100, -200, -300, 0, 0, 150, 150, 150, 150)^\top, \\ c &= (0.8, 2.0, 2.5, 1.0, 1.2, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0)^\top, \\ u &= (\infty, \infty, \infty, \infty, \infty, \infty, \infty, \infty, \infty, \infty, \infty)^\top, \end{aligned}$$

erhalten wir den Fluss

$$x^* = (100, 0, 0, 200, 200, 100, 0, 150, 150, 150, 150)^\top \quad (11.4)$$

als optimale Lösung der Aufgabe (11.3) des kostenminimalen Transports. Die Lösung ist in [Abbildung 11.2](#) dargestellt. Die zugehörigen Transportkosten betragen  $c^\top x^* = 1320$ . 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  $\text{Rang}(A) = 8$  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  $\mathbf{1}^\top b = 0$  gelten, denn

$$Ax = b \quad \text{impliziert} \quad \underbrace{\mathbf{1}^\top Ax}_{=0} = \mathbf{1}^\top b. \quad (11.5)$$

Die Bedarfe und Angebote in einem Transportnetzwerk müssen sich also ausgleichen. Sollte in einem Transportnetzwerk  $\mathbf{1}^\top b < 0$  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 11.3:** 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  $\mathbf{1}^\top b > 0$  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 11.4:** 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?

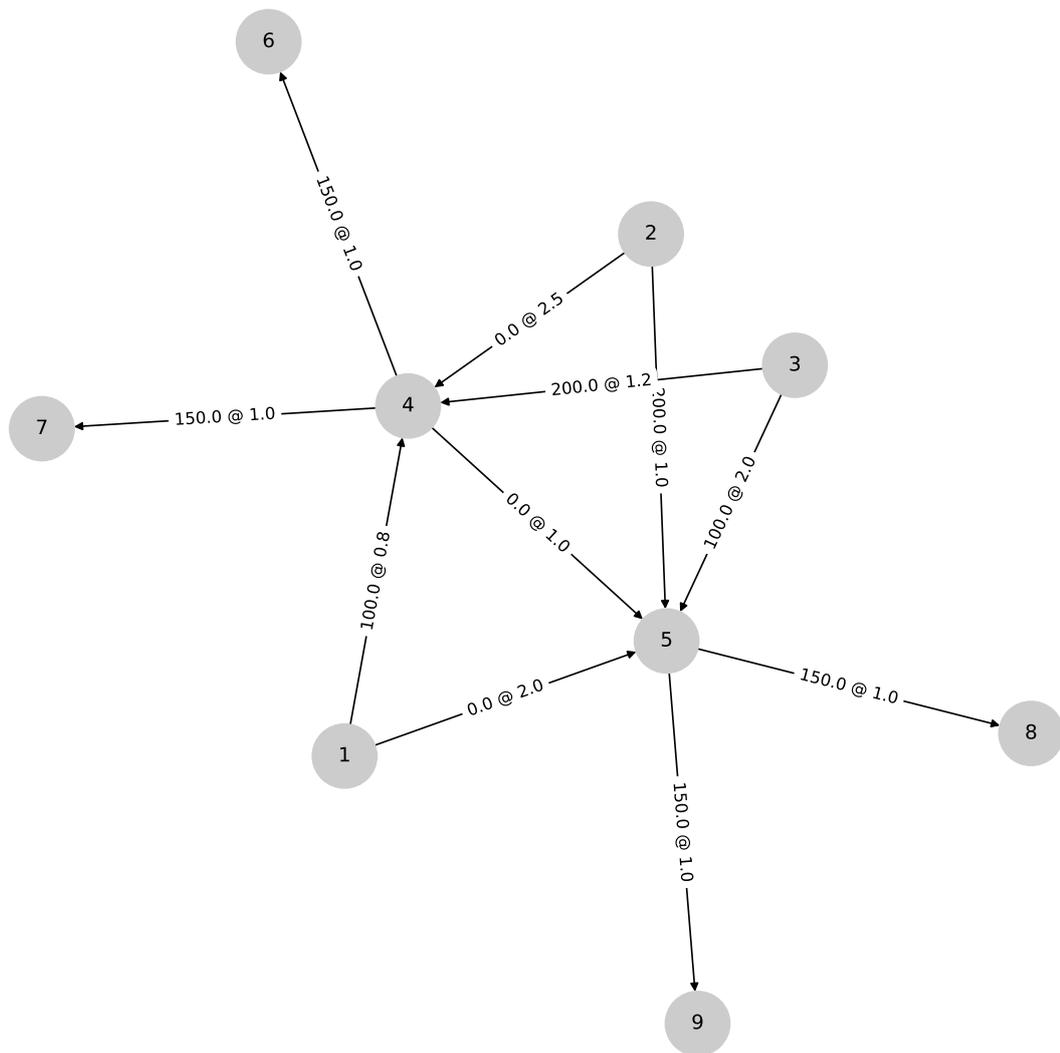


Abbildung 11.2: Eine optimale Lösung von [Beispiel 11.5](#).

**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 [Gerdt's, 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

$$\{x \in \mathbb{R}^n \mid Ax = b, 0 \leq x \leq u\} \quad (12.1)$$

```

# 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](#).

**Definition 12.1** (Unimodularität und totale Unimodularität).

- (i) Eine Matrix  $A \in \mathbb{Z}^{m \times n}$  heißt **unimodular** (englisch: **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** (englisch: **totally 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.

**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 Dimensi-

on  $r = \min\{m, n\}$ . 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:

$$((\widehat{A})^{-1})_{ij} = \frac{1}{\det(\widehat{A})} \det \begin{bmatrix} \widehat{a}_1 & \cdots & \widehat{a}_{i-1} & e_j & \widehat{a}_{i+1} & \cdots & \widehat{a}_m \end{bmatrix}.$$

Die Matrix im Zähler hat nur ganzzahlige Einträge, also ist auch ihre Determinante ganzzahlig. (**Quizfrage 12.1:** Warum eigentlich?) Damit ist auch  $((\widehat{A})^{-1})_{ij}$  ganzzahlig. Im Fall  $r = n \leq m$  argumentiert man ähnlich. (**Quizfrage 12.2:** 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  $\det((\widehat{A})^{-1})$  beide ganzzahlig. Wegen

$$\det(\widehat{A}) \det((\widehat{A})^{-1}) = \det(\widehat{A} (\widehat{A})^{-1}) = \det(\text{Id}) = 1$$

bleiben nur die Möglichkeiten  $\det(\widehat{A}) = \det((\widehat{A})^{-1}) = 1$  oder  $\det(\widehat{A}) = \det((\widehat{A})^{-1}) = -1$ . Andererseits erfüllt jede Untermatrix  $\widehat{A}$  der Dimension  $r$ , die nicht regulär ist,  $\det(\widehat{A}) = 0$ . Also ist  $A$  unimodular.  $\square$

**Satz 12.3** (Bedeutung unimodularer Matrizen).

Es sei  $A \in \mathbb{Z}^{m \times n}$  mit  $\text{Rang}(A) = m$ . 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

$$P_{\text{NF}} := \{x \in \mathbb{R}^n \mid Ax = b, x \geq 0\}$$

nur ganzzahlige Ecken.

- (iii) Für jedes Paar von Vektoren  $b \in \mathbb{Z}^m$ ,  $u \in \mathbb{Z}^n$  besitzt das Polyeder in Normalform mit zusätzlicher oberer Schranke

$$P_{\text{NFB}} := \{x \in \mathbb{R}^n \mid Ax = b, x \geq 0, x \leq u\}$$

nur ganzzahlige Ecken.

*Beweis.* Der Beweis ist Inhalt von [Hausaufgabe 8.2](#).  $\square$

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)  $A^\top$  ist total unimodular.
- (iv)  $[A, -A]$  ist total unimodular.
- (v)  $[A, \text{Id}_m]$  ist total unimodular.
- (vi)  $\begin{bmatrix} A \\ \text{Id}_n \end{bmatrix}$  ist total unimodular.

*Beweis.* Der Beweis ist Inhalt von [Hausaufgabe 8.1](#). □

**Satz 12.5** (Bedeutung total unimodularer Matrizen).

Es sei  $A \in \mathbb{Z}^{m \times n}$ . 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

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

nur ganzzahlige Ecken.

- (iii) Für jedes Paar von Vektoren  $b \in \mathbb{Z}^m, u \in \mathbb{Z}^n$  besitzt das Polyeder in kanonischer Form mit zusätzlicher oberer Schranke

$$P_{\text{KFB}} := \{x \in \mathbb{R}^n \mid Ax \leq b, x \geq 0, x \leq u\}$$

nur ganzzahlige Ecken.

*Beweis.* Der Beweis ist Inhalt von [Hausaufgabe 8.2](#). □

Man kann zeigen, dass Inzidenzmatrizen  $A$  für einfache Digraphen total unimodular sind:

**Satz 12.6** (aus [Schrijver, 2003](#), Theorem 13.9).

Inzidenzmatrizen einfacher Digraphen sind total unimodular.

*Beweis.* Es sei  $\hat{A}$  eine quadratische Untermatrix der Dimension  $r$  der Inzidenzmatrix  $A$ . Wir zeigen die Behauptung  $\det(\hat{A}) \in \{0, \pm 1\}$  durch Induktion über  $r$ . Der Fall  $r = 1$  ist klar, da  $A$  und damit  $\hat{A}$  nur Einträge in  $\{0, \pm 1\}$  besitzt. Wir zeigen nun den Schluss von  $r$  auf  $r + 1$  und unterscheiden dabei drei Fälle.

Im ersten Fall besitzt  $\widehat{A}$  eine Nullspalte, dann ist  $\det(\widehat{A}) = 0$ . Im zweiten Fall besitzt  $\widehat{A}$  mindestens eine Spalte mit genau einem Eintrag ungleich Null. Dann hat  $\widehat{A}$  (ggf. nach einigen Permutationen von Spalten und Zeilen) die Gestalt

$$\widehat{A} = \begin{bmatrix} \pm 1 & a^T \\ 0 & A' \end{bmatrix}.$$

Dabei ist  $A'$  eine Untermatrix von  $A$  der Dimension  $r - 1$ . Aufgrund des Determinantenentwicklungssatzes und der Induktionsvoraussetzung folgt  $\det(\widehat{A}) \in \{0, \pm 1\}$ . Die Permutationen ändern daran nichts, da sie nur das Vorzeichen ändern können.

Im dritten Fall besitzt jede Spalte von  $\widehat{A}$  genau zwei von Null verschiedene Einträge, d. h., einen Eintrag  $+1$  und einen Eintrag  $-1$ . Das bedeutet, die Summe jedes Spaltenvektors von  $\widehat{A}$  ist Null, damit ist  $\det(\widehat{A}) = 0$ . □

Aus Satz 12.5 und Satz 12.6 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 (falls das LP überhaupt lösbar ist) automatisch eine ganzzahlige Lösung für Aufgaben des kostenminimalen Flusses, wie wir es z. B. in Beispiel 11.5 beobachtet hatten, siehe (11.4).

**Beachte:** Auf den Kostenvektor  $c$  kommt es dabei nicht an!

**Quizfrage 12.3:** 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 12.7.** 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

$$\begin{aligned} &\text{Minimiere} && (c^{(1)})^T x^{(1)} + (c^{(2)})^T x^{(2)} && \text{über } (x^{(1)}, x^{(2)}) \in \mathbb{R}^n \times \mathbb{R}^n \\ &\text{unter} && \begin{cases} A_0 x^{(1)} & = b^{(1)} & \text{(Knotenbilanzen Transportgut 1)} \\ & A_0 x^{(2)} = b^{(2)} & \text{(Knotenbilanzen Transportgut 2)} \\ & x^{(1)} + x^{(2)} \leq u & \text{(Kapazitätsbeschränkung)} \end{cases} && (12.2) \\ &\text{sowie} && x^{(1)} \geq 0, \quad x^{(2)} \geq 0. \end{aligned}$$

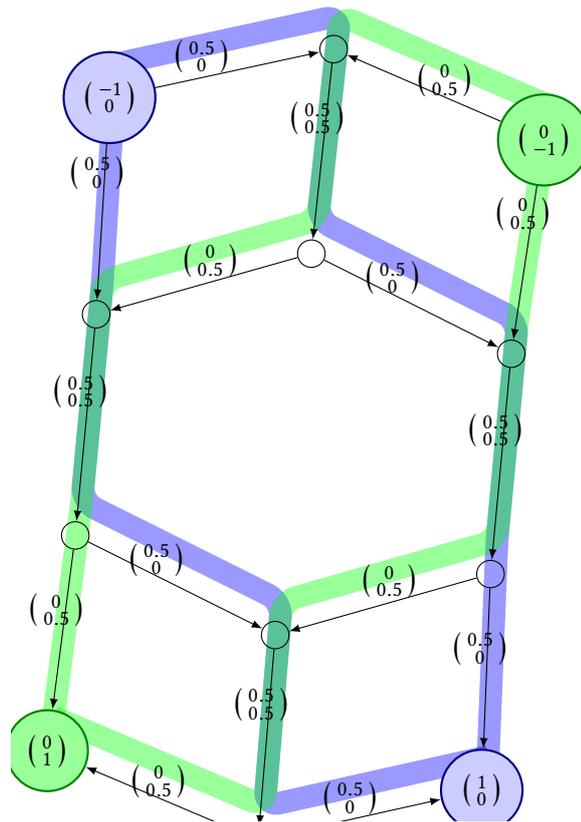


Abbildung 12.1: Darstellung eines Transportnetzwerks für ein Zweigüterflussproblem. Die Menge eins des ersten Gutes (blau) soll von der Quelle oben links zur Senke unten rechts transportiert werden. Dieselbe Menge des zweiten Gutes (grün) soll von der Quelle oben rechts zur Senke unten links transportiert werden. Die Kantenkapazitäten sind alle gleich eins. Der einzige zulässige Fluss ist an den jeweiligen Kanten eingezeichnet. Er ist nicht ganzzahlig.

Obwohl  $A_0$  total unimodular ist, trifft das auf die Matrix

$$\begin{bmatrix} A_0 & 0 \\ 0 & A_0 \\ \text{Id} & \text{Id} \end{bmatrix}$$

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**.

# 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.

Bemerkenswert dabei ist, welche starken topologischen bzw. analytischen Eigenschaften aus der Konvexität von Mengen bzw. Funktionen folgen, siehe z. B. Satz 15.19 und ??.

### § 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*)

$$\Delta_n = \left\{ x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i \leq 1, x_i \geq 0, i = 1, \dots, n \right\}.$$

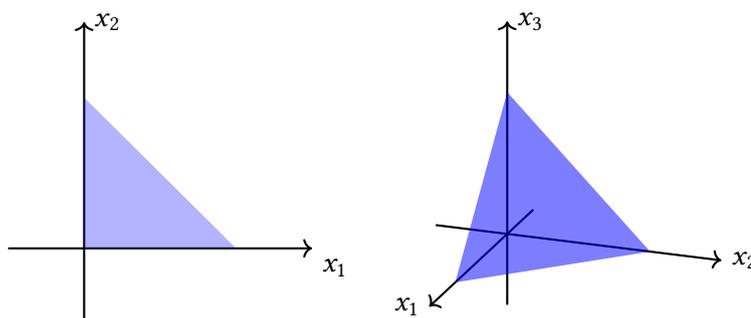


Abbildung 13.2: Einheitssimplizes im  $\mathbb{R}^2$  und  $\mathbb{R}^3$ .

**Quizfrage 13.1:** 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 der Durchschnitt  $\bigcap_{j \in J} C_j$  konvex.
- (ii) Es seien  $C_i \subseteq \mathbb{R}^{n_i}$  konvex,  $i = 1, \dots, k$ . Dann ist das kartesische Produkt  $C_1 \times \dots \times C_k$  konvex in  $\mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_k}$ .
- (iii) Es sei  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  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 ist Gegenstand von [Hausaufgabe 9.1](#). □

**Quizfrage 13.2:** Ist die Vereinigung konvexer Mengen wieder konvex?

**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  $x_1, \dots, x_m \in M$  ist.

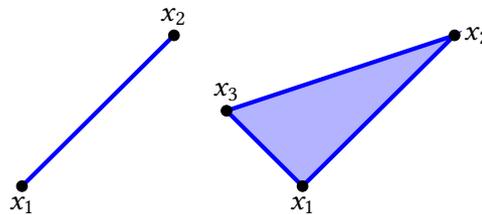


Abbildung 13.3: Konvexkombinationen von zwei und drei Punkten in  $\mathbb{R}^2$ . Bei zwei Punkten kann man auch  $\alpha x_1 + (1 - \alpha) x_2$  mit  $\alpha \in [0, 1]$  schreiben.

**Beachte:** An dieser Stelle wird die Bemerkung nach [Definition 6.12](#) klar, dass die Ecken eines Polyeders  $P$  genau diejenigen Punkte von  $P$  sind, die sich nicht als echte Konvexkombination zweier verschiedener Elemente von  $P$  schreiben lassen.

**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.* „ $\Rightarrow$ “: Es sei  $M$  konvex. Wir haben zu zeigen: Für  $m \in \mathbb{N}$  und  $x_1, \dots, x_m \in M$  sowie  $\alpha_1, \dots, \alpha_m \geq 0$  mit  $\sum_{i=1}^m \alpha_i = 1$  gilt:  $x = \sum_{i=1}^m \alpha_i x_i \in M$ . Wir verwenden dazu Induktion nach der Anzahl  $m$  der beteiligten Vektoren: Für  $m = 1, 2$  ist die Behauptung erfüllt. (**Quizfrage 13.3:** Warum?) 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. gelte  $\alpha_{m+1} < 1$ . (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 die folgende Konvexkombination von  $y$  und  $x_{m+1}$ :  $x = (1 - \alpha_{m+1}) y + \alpha_{m+1} x_{m+1}$ .

„ $\Leftarrow$ “: Es seien  $x_1, x_2 \in M$ . Nach Voraussetzung enthält  $M$  alle Konvexkombinationen  $\alpha x_1 + (1 - \alpha) x_2$  mit  $\alpha \in [0, 1]$ , d. h.,  $M$  ist nach Definition 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

$$\text{conv}(M) = \bigcap \{C \subseteq \mathbb{R}^n \mid C \text{ ist konvex und } M \subseteq C\}, \quad (13.1)$$

heißt die **konvexe Hülle** von  $M$ .  $\text{conv}(M)$  ist also die kleinste konvexe Menge, die  $M$  enthält.

**Beachte:** Es gilt  $M \subseteq \underbrace{\text{conv}(M)}_{\text{„Hülle“}},$  daher der Name **konvexe Hülle**.

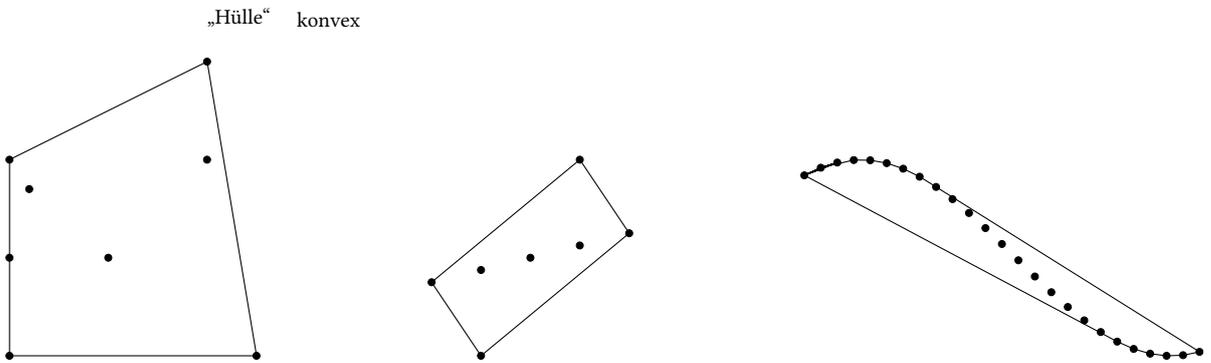


Abbildung 13.4: Konvexe Hüllen einiger Punktmenge in  $\mathbb{R}^2$ .

**Lemma 13.7** (Charakterisierung der konvexen Hülle).

Es sei  $M \subseteq \mathbb{R}^n$ . Dann gilt:

$$\text{conv}(M) = \text{Menge aller Konvexkombinationen von } M.$$

*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.

„ $\text{conv}(M) \subseteq C$ “: Wir zeigen:  $C$  ist konvex. Damit kommt diese Menge im Durchschnitt (13.1) vor, also gilt  $\text{conv}(M) \subseteq C$ .

Es seien  $x, y \in C$ , also gibt es Zahlen  $m, \ell \in \mathbb{N}$  und  $\beta_1, \dots, \beta_m \geq 0$  sowie  $\gamma_1, \dots, \gamma_\ell \geq 0$  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  $x_1, \dots, x_m \in M$  und  $y_1, \dots, y_\ell \in M$ . 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  $\geq 0$  und ergeben in der Summe 1. Damit ist  $\alpha x + (1 - \alpha) y \in C$ , also  $C$  konvex.

„ $\text{conv}(M) \supseteq C$ “: Es sei  $x \in C$ , also eine Konvexkombination von  $M$ . Wegen  $M \subseteq \text{conv}(M)$  ist  $x$  auch eine Konvexkombination von  $\text{conv}(M)$ .  $\text{conv}(M)$  ist konvex, stimmt also nach Lemma 13.5 mit der Menge seiner Konvexkombinationen überein. Also ist  $x \in \text{conv}(M)$ .  $\square$

**Beachte:**  $M$  ist konvex  $\Leftrightarrow M = \text{conv}(M)$ .

## § 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  $f: C \rightarrow \mathbb{R}$  heißt

(i) **konvex** (englisch: **convex**) auf  $C$ , falls

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) \quad (13.2)$$

für alle  $x, y \in C$  und  $\alpha \in [0, 1]$  gilt.

(ii) **strikt konvex** (englisch: **strictly convex**) auf  $C$ , falls

$$f(\alpha x + (1 - \alpha)y) < \alpha f(x) + (1 - \alpha)f(y) \quad (13.3)$$

für alle  $x, y \in C$  mit  $x \neq y$  und  $\alpha \in (0, 1)$  gilt.

(iii)  **$\mu$ -stark konvex** (englisch:  **$\mu$ -strongly convex**) oder **stark konvex** mit Parameter  $\mu > 0$  auf  $C$ , falls

$$f(\alpha x + (1 - \alpha)y) + \frac{\mu}{2} \alpha (1 - \alpha) \|x - y\|^2 \leq \alpha f(x) + (1 - \alpha)f(y) \quad (13.4)$$

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 Definitionsmenge*.

**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 13.4:** Was hat die starke Konvexität von  $f$  mit der Konvexität von  $f(\cdot) - \frac{\mu}{2} \|\cdot\|^2$  zu tun?

**Beachte:** Für Funktionen  $f: C \rightarrow \mathbb{R}$  wie in [Definition 13.8](#) gilt:

$$f \text{ stark konvex} \quad \Rightarrow \quad f \text{ strikt konvex} \quad \Rightarrow \quad f \text{ konvex.}$$

**Beispiel 13.9** (Beispiele konvexer Funktionen).

(i) Die *affin-lineare Funktion*  $f(x) = a^T x + \beta$  ist gleichzeitig konvex und konkav auf  $\mathbb{R}^n$ .

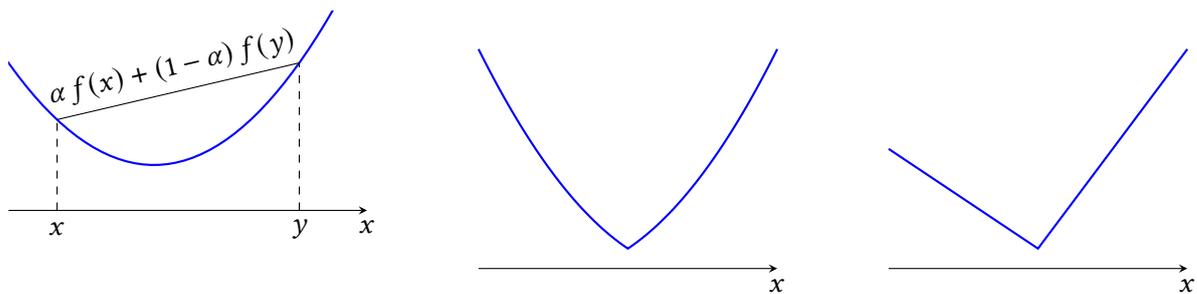


Abbildung 13.5: Beispiele strikt konvexer Funktionen (links und Mitte) und konvexe, aber nicht strikt konvexe Funktion (rechts) auf Intervallen in  $\mathbb{R}$ .

(ii) Die quadratische Funktion  $f(x) = \frac{1}{2}x^T Q x + c^T x + y$  mit symmetrischer Matrix  $Q \in \mathbb{R}^{n \times n}$  ist

- konvex  $\Leftrightarrow Q$  ist positiv semidefinit,
- $\mu$ -stark konvex  $\Leftrightarrow Q$  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 13.5:** Gibt es außer affin-linearen Funktionen noch weitere Funktionen auf  $\mathbb{R}^n$ , die gleichzeitig konvex und konkav sind?

**Quizfrage 13.6:** 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  $\mathbb{R} \cup \{\pm\infty\}$  liegen. Man spricht dann von **erweitert reellwertigen Funktionen**. In  $\mathbb{R} \cup \{\pm\infty\}$  gelten 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  $a \in \mathbb{R}$  sowie für  $a = -\infty$ .
- (iii)  $a \infty = \infty a = \infty$  für alle  $a > 0$  sowie für  $a = \infty$ .

- (iv)  $a \infty = \infty a = -\infty$  für alle  $a < 0$  sowie für  $a = -\infty$ .
- (v)  $a(-\infty) = (-\infty)a = -\infty$  für alle  $a > 0$  sowie für  $a = \infty$ .
- (vi)  $a(-\infty) = (-\infty)a = \infty$  für alle  $a < 0$  sowie für  $a = -\infty$ .
- (vii)  $0 \infty = \infty 0 = 0(-\infty) = (-\infty) 0 = 0$ .
- (viii)  $-\infty < a < \infty$  für alle  $a \in \mathbb{R}$ .
- (ix)  $-\infty \leq a \leq \infty$  für alle  $a \in \mathbb{R} \cup \{\pm\infty\}$ .

Die Kombinationen  $\infty - \infty$  und  $-\infty + \infty$  sind undefiniert und müssen vermieden werden.

Die Kommutativität und Assoziativität der Addition und der Multiplikation sowie das Distributivgesetz gelten auch in  $\mathbb{R} \cup \{\pm\infty\}$  weiter, sofern in den betreffenden Ausdrücken jeweils alle Terme definiert sind. Beispielsweise gilt  $(2 + 1)\infty = 2 \infty + 1 \infty = \infty + \infty = \infty$ . Beim Ausdruck  $(-3 + 1)\infty = -2\infty = -\infty$  jedoch darf man das Distributivgesetz nicht anwenden, da  $-\infty + \infty$  nicht erklärt ist.

Der Mehrwert erweitert reellwertiger Funktionen liegt in folgenden Überlegungen begründet:

- (1) Wir können jede reellwertige Funktion  $f: M \rightarrow \mathbb{R}$ , die auf einer Teilmenge  $M \subseteq \mathbb{R}^n$  definiert ist, auf ganz  $\mathbb{R}^n$  fortsetzen, in dem wir  $\bar{f}: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  definieren als

$$\bar{f}(x) := \begin{cases} f(x) & \text{falls } x \in M, \\ \infty & \text{falls } x \notin M. \end{cases} \quad (13.5)$$

Daher haben wir es nur noch mit Funktionen zu tun, die auf ganz  $\mathbb{R}^n$  definiert sind.

- (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  $\infty$  setzen. Dies gelingt einfach durch Addition einer Indikatorfunktion.

**Definition 13.10** (Indikatorfunktion).

Es sei  $M \subseteq \mathbb{R}^n$  irgendeine Menge. Die Funktion  $I_M: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ , definiert durch

$$I_M(x) = \begin{cases} 0, & \text{falls } x \in M, \\ \infty, & \text{falls } x \notin M, \end{cases} \quad (13.6)$$

heißt die **Indikatorfunktion** (englisch: indicator function) von  $M$ .

Der Mehrwert von Funktionen, die auch den Wert  $-\infty$  annehmen, wird bei den Richtungsableitungen in ?? deutlich werden. In manchen Resultaten, etwa in [Satz 13.16](#), wird dieser Fall auch ausgeschlossen.

**Definition 13.11** (Eigentlicher Definitionsbereich, eigentliche Funktion).

Es sei  $f: M \rightarrow \mathbb{R} \cup \{\pm\infty\}$  eine Funktion. Dann heißt

$$\text{dom } f := \{x \in M \mid f(x) < \infty\} \quad (13.7)$$

der **eigentliche Definitionsbereich** oder (englisch: (effective) domain) von  $f$ . Die Funktion  $f$  heißt **eigentlich** (englisch: proper function), wenn  $f$  nirgendwo den Wert  $-\infty$  annimmt und  $f$  nicht identisch  $\infty$  (also  $\text{dom } f \neq \emptyset$ ) ist.

Wir erweitern jetzt die [Definition 13.8](#) auf erweitert reellwertige Funktionen. Änderungen gegenüber [Definition 13.8](#) sind farblich **hervorgehoben**. Insbesondere sind die Begriffe **strikte** und **starke Konvexität** nicht für beliebige erweitert reellwertige Funktionen sinnvoll. Die eigentlichen Funktionen sind aber eine geeignete Funktionenklasse dafür.

**Definition 13.12** (Erweitert reellwertige konvexe Funktion).

Es sei  $C \subseteq \mathbb{R}^n$  konvex.

(i) Eine Funktion  $f: C \rightarrow \mathbb{R} \cup \{\pm\infty\}$  heißt **konvex** (englisch: convex) auf  $\mathbb{R}^n$ , falls

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) \quad (13.8)$$

gilt für alle  $x, y \in C$  und  $\alpha \in [0, 1]$ , für die die rechte Seite definiert ist.

(ii) Eine **eigentliche** Funktion  $f: C \rightarrow \mathbb{R} \cup \{\infty\}$  heißt **strikt konvex** auf  $\mathbb{R}^n$ , falls

$$f(\alpha x + (1 - \alpha)y) < \alpha f(x) + (1 - \alpha)f(y) \quad (13.9)$$

gilt für alle  $x, y \in \text{dom } f \subseteq C$  mit  $x \neq y$  und  $\alpha \in (0, 1)$ .

(iii) Eine **eigentliche** Funktion  $f: C \rightarrow \mathbb{R} \cup \{\infty\}$  heißt  **$\mu$ -stark konvex** oder **stark konvex** mit Parameter  $\mu > 0$  auf  $C$ , falls

$$f(\alpha x + (1 - \alpha)y) + \frac{\mu}{2} \alpha(1 - \alpha) \|x - y\|^2 \leq \alpha f(x) + (1 - \alpha)f(y) \quad (13.10)$$

gilt für alle  $x, y \in C$  und  $\alpha \in [0, 1]$ .

**Beachte:** Die rechte Seite in (13.8) ist genau dann nicht für alle  $\alpha \in [0, 1]$  definiert, wenn  $f(x) = \infty$  und  $f(y) = -\infty$  gilt oder umgekehrt. In (13.9) und (13.10) ist das ausgeschlossen, da  $f$  eigentlich ist.

**Quizfrage 13.7:** Gilt für erweitert reellwertige, eigentliche Funktionen noch immer die Aussage „ $f$  stark konvex  $\Rightarrow f$  strikt konvex  $\Rightarrow f$  konvex“?

**Satz 13.13** (Konvexität der Erweiterung konvexer Funktionen).

(i) Es sei  $C \subseteq \mathbb{R}^n$  konvex und  $f: C \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex. Dann ist die Fortsetzung  $\bar{f}: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  auf  $\mathbb{R}^n$  konvex.

(ii) Es sei  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex. Dann ist  $\text{dom } f \subseteq \mathbb{R}^n$  konvex, und die Einschränkung  $f|_{\text{dom } f}: \text{dom } f \rightarrow \mathbb{R} \cup \{-\infty\}$  ist auf  $\text{dom } f$  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  $\bar{f}$  unterscheiden.

*Beweis.* **Aussage (i):** Es sei  $C \subseteq \mathbb{R}^n$  konvex und  $f: C \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex. Wir müssen zeigen:

$$\bar{f}(\alpha x + (1 - \alpha) y) \leq \alpha \bar{f}(x) + (1 - \alpha) \bar{f}(y) \quad (13.11)$$

für alle  $x, y \in \mathbb{R}^n$  und alle  $\alpha \in [0, 1]$ , für die die rechte Seite definiert ist.

Falls  $x, y \in C$  liegen, dann ist (13.11) nach Definition erfüllt, da  $\bar{f}$  auf der gesamten Verbindungsstrecke zwischen  $x$  und  $y$  mit  $f$  übereinstimmt.

Falls  $x \notin C$  liegt, also  $\bar{f}(x) = \infty$ , und  $y \in C$  gilt, dann ist die rechte Seite in (13.11) entweder gleich  $\infty$  für alle  $\alpha \in [0, 1]$  oder (im Fall  $\bar{f}(y) = f(y) = -\infty$ ) nur für  $\alpha \in \{0, 1\}$  definiert. In jedem Fall ist (13.11) erfüllt.

Dieselbe Argumentation ist natürlich gültig, wenn  $y \notin C$  und  $x \in C$  liegt.

Im Fall  $x, y \notin C$  gilt  $\bar{f}(x) = \bar{f}(y) = \infty$ , daher ist (13.11) ebenfalls für alle  $\alpha \in [0, 1]$  erfüllt.

**Aussage (ii):** Es sei  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex, es gilt also

$$f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y) \quad (13.12)$$

für alle  $x, y \in \mathbb{R}^n$  und alle  $\alpha \in [0, 1]$ , für die die rechte Seite definiert ist.

Wir zeigen zunächst, dass  $\text{dom } f \subseteq \mathbb{R}^n$  konvex ist. Es seien dazu  $x, y \in \text{dom } f$ . Dann ist die rechte Seite in (13.12) für alle  $\alpha \in [0, 1]$  definiert. (**Quizfrage 13.8:** Warum?) Aus (13.12) folgt

$$f(\alpha x + (1 - \alpha) y) \leq \underbrace{\alpha f(x)}_{< \infty} + \underbrace{(1 - \alpha) f(y)}_{< \infty} < \infty$$

für alle  $\alpha \in [0, 1]$ , sodass die linke Seite nicht  $\infty$  ist. Also gehört  $\alpha x + (1 - \alpha) y$  zu  $\text{dom } f$ . Dieselbe Ungleichung zeigt auch bereits, dass  $f|_{\text{dom } f}: \text{dom } f \rightarrow \mathbb{R}$  wie behauptet konvex ist.  $\square$

**Folgerung 13.14** (Konvexität der Indikatorfunktion). *Die Indikatorfunktion  $I_M: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  einer Menge  $M \subseteq \mathbb{R}^n$  ist genau dann konvex, wenn die Menge  $M$  konvex ist.*

Es gilt folgende wichtige Charakterisierung konvexer Funktionen:

**Satz 13.15** (Epigraph-Charakterisierung konvexer Funktionen).

*Eine Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  ist genau dann konvex, wenn ihr **Epigraph***

$$\text{epi } f := \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \mid y \geq f(x) \right\} \quad (13.13)$$

*eine konvexe Menge ist.*

*Beweis.* Es sei zunächst  $f$  konvex, und es seien  $\begin{pmatrix} x \\ \gamma \end{pmatrix}$  und  $\begin{pmatrix} y \\ \delta \end{pmatrix}$  Punkte in  $\text{epi } f$ . Insbesondere ist also  $f(x) < \infty$  und  $f(y) < \infty$ . Die Ungleichung (13.8) gilt also für alle  $\alpha \in [0, 1]$ , d. h.,

$$f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y) \leq \alpha \gamma + (1 - \alpha) \delta.$$

Das bedeutet aber

$$\begin{pmatrix} \alpha x + (1 - \alpha) y \\ \alpha \gamma + (1 - \alpha) \delta \end{pmatrix} = \alpha \begin{pmatrix} x \\ \gamma \end{pmatrix} + (1 - \alpha) \begin{pmatrix} y \\ \delta \end{pmatrix} \in \text{epi } f.$$

Also ist  $\text{epi } f$  konvex.

Umgekehrt sei nun  $\text{epi } f$  konvex und  $x, y \in \mathbb{R}^n$ . Wir müssen (13.8) für diejenigen  $\alpha \in [0, 1]$  nachweisen, für die rechte Seite definiert ist.

Wir unterscheiden einige Fälle. Wenn  $f(x)$  und  $f(y)$  beide endlich sind, dann gehören  $\begin{pmatrix} x \\ f(x) \end{pmatrix}$  und  $\begin{pmatrix} y \\ f(y) \end{pmatrix}$  zu  $\text{epi } f$ . Die Konvexität von  $\text{epi } f$  zeigt, dass auch  $\alpha \begin{pmatrix} x \\ f(x) \end{pmatrix} + (1 - \alpha) \begin{pmatrix} y \\ f(y) \end{pmatrix}$  zu  $\text{epi } f$  gehört, also gilt (13.8), was zu zeigen war.

Im Fall  $f(x) = \infty$  und  $f(y) > -\infty$  ist die Ungleichung (13.8) trivialerweise für alle  $\alpha \in [0, 1]$  erfüllt. Dasselbe gilt im umgekehrten Fall  $f(x) > -\infty$  und  $f(y) = \infty$ .

Es verbleiben drei Fälle. In einem davon ist  $f(x) = -\infty$  und  $f(y)$  endlich. Nach Voraussetzung gehört  $\begin{pmatrix} x \\ \gamma \end{pmatrix}$  zu  $\text{epi } f$  für jedes  $\gamma \in \mathbb{R}$ . Weiter sei  $\begin{pmatrix} y \\ \delta \end{pmatrix}$  ebenfalls in  $\text{epi } f$ . Die Konvexität von  $\text{epi } f$  impliziert, dass auch  $\alpha \begin{pmatrix} x \\ \gamma \end{pmatrix} + (1 - \alpha) \begin{pmatrix} y \\ \delta \end{pmatrix}$  zu  $\text{epi } f$  gehört, also gilt

$$f(\alpha x + (1 - \alpha) y) \leq \alpha \gamma + (1 - \alpha) \delta$$

für alle  $\alpha \in [0, 1]$ . Da  $\gamma \in \mathbb{R}$  beliebig ist, folgt daraus

$$f(\alpha x + (1 - \alpha) y) = -\infty$$

für alle  $\alpha \in (0, 1]$ . Deshalb ist die Ungleichung (13.8) für alle diese  $\alpha$  und auch für  $\alpha = 0$  erfüllt.

Eine ähnliche Argumentation zeigt die Behauptung im Fall  $f(x)$  endlich und  $f(y) = -\infty$  sowie auch im Fall  $f(x) = f(y) = -\infty$ . □

**Satz 13.16** (Operationen auf konvexen Funktionen).

(i) Sind  $f_i: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  konvex auf  $\mathbb{R}^n$  und  $\beta_i \geq 0$  für  $i = 1, \dots, m$ , dann ist die durch

$$f(x) := \sum_{i=1}^m \beta_i f_i(x)$$

definierte Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  konvex auf  $\mathbb{R}^n$ .

**Beachte:** Wir lassen hier nicht zu, dass eine der Funktionen  $f_i$  irgendwo den Wert  $-\infty$  annimmt. Wegen der Summenbildung wäre sonst möglicherweise  $f(x)$  nicht definiert.

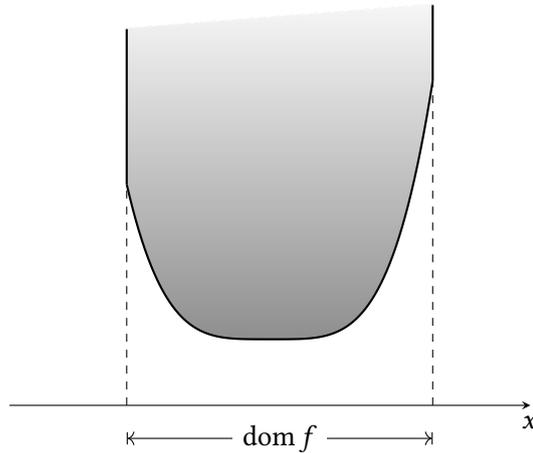


Abbildung 13.6: Epigraph einer Funktion  $f: \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$ .

(ii) Sind die Funktionen  $f_i: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex auf  $\mathbb{R}^n$  für alle  $i$  aus irgendeiner Indexmenge  $I$ , dann ist die durch das punktweise Supremum

$$f(x) := \sup\{f_i(x) \mid i \in I\}$$

definierte Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex auf  $\mathbb{R}^n$ .

(iii) Ist  $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$  affin-linear und  $f: \mathbb{R}^m \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex, so ist  $(f \circ g): \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex auf  $\mathbb{R}^n$ .

(iv) Ist  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  konvex und ist  $f: \mathbb{R} \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex und monoton wachsend, so ist  $(f \circ g): \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex.

**Beachte:** Aus Aussage (iv) folgt insbesondere, dass die Funktion  $g^2$  konvex auf  $\mathbb{R}^n$  ist, wenn  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  konvex ist und  $g(x) \geq 0$  gilt für alle  $x \in \mathbb{R}^n$ . (Quizfrage 13.9: Genaue Begründung?)

*Beweis.* Der Beweis ist Inhalt von Hausaufgabe 9.4. □

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 die 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  $f: C \rightarrow \mathbb{R}$  diffbar. Dann gelten:

(a) Es sind äquivalent:

(i)  $f$  ist konvex auf  $C$ .

(ii) Für alle  $x, y \in C$  gilt:

$$f(x) - f(y) \geq f'(y)(x - y). \quad (13.14)$$

(iii) Für alle  $x, y \in C$  gilt:

$$(f'(x) - f'(y))(x - y) \geq 0. \quad (13.15)$$

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:

$$f(x) - f(y) > f'(y)(x - y). \quad (13.16)$$

(iii) Für alle  $x, y \in C$  mit  $x \neq y$  gilt:

$$(f'(x) - f'(y))(x - y) > 0. \quad (13.17)$$

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:

$$f(x) - f(y) \geq f'(y)(x - y) + \frac{\mu}{2} \|x - y\|^2. \quad (13.18)$$

(iii) Es existiert  $\mu > 0$ , sodass für alle  $x, y \in C$  gilt:

$$(f'(x) - f'(y))(x - y) \geq \mu \|x - y\|^2. \quad (13.19)$$

Man sagt zu (13.19), die Ableitung  $f'$  sei auf  $C$  ein **stark monotoner Operator**.

**Beachte:** Nach [Aussage \(a\)](#) ist eine diffbare Funktion  $f: C \rightarrow \mathbb{R}$  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.

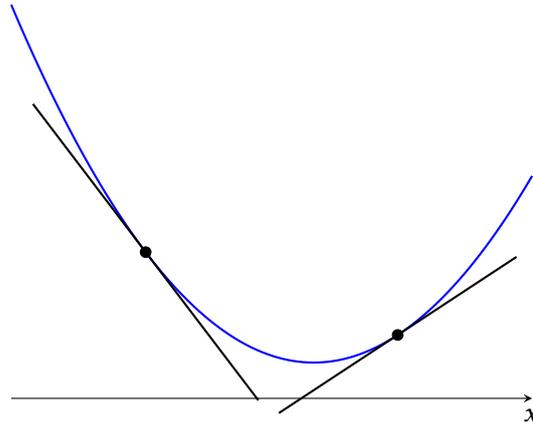


Abbildung 13.7: Charakterisierung der Konvexität diffbarer Funktionen über ihre Tangenten.

*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

$$\begin{aligned} f(y + \alpha(x - y)) &= f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) - \frac{\mu}{2} \alpha(1 - \alpha) \|x - y\|^2 \\ \Rightarrow \frac{f(y + \alpha(x - y)) - f(y)}{\alpha} &\leq f(x) - f(y) - \frac{\mu}{2} (1 - \alpha) \|x - y\|^2 \\ \Rightarrow f'(y)(x - y) &\leq f(x) - f(y) - \frac{\mu}{2} \|x - y\|^2 \quad (\text{Grenzübergang } \alpha \searrow 0), \end{aligned}$$

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

$$\begin{aligned} f(x) - f(z) &\geq f'(z)(x - z) + \frac{\mu}{2} \|x - z\|^2, \\ f(y) - f(z) &\geq f'(z)(y - z) + \frac{\mu}{2} \|y - z\|^2. \end{aligned}$$

Wir multiplizieren die erste Ungleichung mit  $\alpha$ , 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

$$\begin{aligned} f(x) - f(y) &\geq f'(y)(x - y) + \frac{\mu}{2} \|x - y\|^2, \\ f(y) - f(x) &\geq f'(x)(y - x) + \frac{\mu}{2} \|x - y\|^2, \end{aligned}$$

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  $t \mapsto d(t) := f'(y + t(x - y))^T(x - y)$  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

$$\begin{aligned} (d(t) - d(s))(t - s) &= [f'(y + t(x - y))^T(x - y) - f'(y + s(x - y))^T(x - y)](t - s) \\ &= [f'(y + t(x - y)) - f'(y + s(x - y))]^T(t - s)(x - y) \\ &\geq \mu \|(t - s)(x - y)\|^2 \quad \text{wegen (13.19)} \\ &= \mu \|x - y\|^2 |t - s|^2. \end{aligned}$$

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

$$D(1) - D(0) = \int_0^1 d(t) dt.$$

Daher gilt weiter

$$\begin{aligned} D(1) - D(0) - d(0) &= \int_0^1 [d(t) - d(0)] dt \\ &= \int_0^1 \frac{1}{t} [d(t) - d(0)] (t - 0) dt \\ &\geq \mu \|x - y\|^2 \int_0^1 t dt \\ &= \frac{\mu}{2} \|x - y\|^2. \end{aligned}$$

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  $f: C \rightarrow \mathbb{R}$  zweimal diffbar. 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

$$\begin{aligned} d^\top f''(x) d &= \lim_{t \rightarrow 0} \frac{f'(x + td) - f'(x)}{t} d \\ &= \lim_{t \rightarrow 0} \frac{f'(x + td) - f'(x)}{t^2} (td) \\ &\geq \lim_{t \rightarrow 0} \frac{1}{t^2} \mu \|td\|^2 \quad \text{wegen (13.19)} \\ &= \mu \|d\|^2. \end{aligned}$$

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. Die Funktion  $D(t) = f'(y + t(x - y))(x - y)$  ist diffbar und monoton wachsend in  $t \in [0, 1]$ , da ihre Ableitung  $d(t) = (x - y)^\top f''(y + t(x - y))(x - y)$  auf  $[0, 1]$  nach Voraussetzung nichtnegativ ist.

Aus dem Hauptsatz der Differential- und Integralrechnung, angewendet auf die Funktion  $D$  und ihre Ableitung  $d$ , folgt

$$D(1) - D(0) = \int_0^1 d(t) dt,$$

also

$$\begin{aligned} [f'(x) - f'(y)](x - y) &= \int_0^1 (x - y)^\top f''(y + t(x - y))(x - y) dt \\ &\geq \mu \int_0^1 \|x - y\|^2 dt \\ &= \mu \|x - y\|^2. \end{aligned}$$

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

$$[f'(x) - f'(y)](x - y) = \int_0^1 (x - y)^T f''(y + t(x - y))(x - y) dt > 0.$$

Das heißt,  $f'$  ist strikt monoton, und aus [Satz 13.17 \(b\)](#) folgt die strikte Konvexität von  $f$  auf  $C$ .  $\square$

Ende der Woche 9

## § 14 KONVEXE OPTIMIERUNGSAUFGABEN

Wir betrachten die **konvexe Optimierungsaufgabe**

$$\text{Minimiere } f(x) \text{ über } x \in \mathbb{R}^n \tag{14.1}$$

mit konvexer Zielfunktion  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$ . Diese Problemklasse enthält insbesondere Aufgaben der Form

$$\text{Minimiere } g(x) \text{ über } x \in C, \tag{14.2}$$

wobei die zulässige Menge  $C \subseteq \mathbb{R}^n$  konvex und  $g: C \rightarrow \mathbb{R}$  eine konvexe reellwertige Funktion ist. Setzen wir dann  $f := g + \delta_C$  mit der Indikatorfunktion  $\delta_C$  von  $C$ , setzen also  $g$  durch den Wert  $\infty$  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  $x \in C = \text{dom } f$  enthält. Beispiele für (14.1) sind sämtliche linearen Optimierungsaufgaben aus [Kapitel 2](#), nicht notwendigerweise in Normalform. Weitere Beispiele folgen.

Die Grundbegriffe aus [Definition 1.1](#) gelten i. W. auch für Aufgaben mit erweitert reellwertigen Zielfunktionen weiter. Wir wiederholen sie hier jedoch der Klarheit wegen.

**Definition 14.1** (Grundbegriffe für die konvexe Optimierung).

(i) Der Wert

$$f^* := \inf \{f(x) \mid x \in \mathbb{R}^n\}$$

heißt der **Optimalwert** der Aufgabe (14.1).

(ii) Ein Punkt  $x^* \in \text{dom } f$  heißt ein **globaler Minimierer**, **globale Minimalstelle** oder **globale optimale Lösung**, wenn gilt:

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

Äquivalent dazu ist:  $f(x^*) = f^*$ . In diesem Fall heißt die Zahl  $f^*$  dann auch das **globale Minimum** oder der **globale Minimalwert** von (14.1).

(iii) Ein Punkt  $x^* \in \text{dom } f$  heißt ein **lokaler Minimierer**, **lokale Minimalstelle** oder **lokal optimale Lösung**, wenn es eine Umgebung  $U(x^*)$  gibt, sodass gilt:

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

In diesem Fall heißt  $f(x^*)$  dann auch ein **lokales Minimum** oder ein **lokaler Minimalwert** von (14.1).

**Beachte:** Dem Begriff des lokalen und globalen Minimierers haben wir die Bedingung  $x^* \in \text{dom } f$  hinzugefügt. Der Funktionswert dort darf also nicht  $\infty$  sein. Das erfolgt vor dem Hintergrund, dass wir bei der Formulierung von (14.2) in der Form (14.1) den Funktionswert  $f(x) = \infty$  ja gerade als Kennzeichen dafür verwenden, dass der zugehörige Punkt  $x$  unzulässig war.

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  $\text{dom } f = \emptyset$  gilt. Diesen Fall werden wir aber im Folgenden oft ausschließen, ebenso wie den Fall, dass  $f$  den Wert  $-\infty$  annimmt. Mit anderen Worten: Wir werden oft annehmen, dass  $f$  eine eigentliche Funktion ist.

Die fundamentale Bedeutung der Konvexität in der Optimierung erläutert der folgende Satz.

**Satz 14.2** (Hauptsatz der konvexen Optimierung).

- (i) Jeder lokale Minimierer von (14.1) ist bereits ein globaler Minimierer.
- (ii) Die Lösungsmenge von (14.1) ist konvex (evtl. leer).
- (iii) Ist  $f$  eigentlich und strikt konvex auf  $\mathbb{R}^n$ , so besitzt (14.1) höchstens eine Lösung.

**Beachte:** Wir brauchen also in der konvexen Optimierung nicht zwischen lokalen und globalen Minimierern zu unterscheiden!

*Beweis. Aussage (i):* Es sei  $x^*$  lokaler Minimierer, d. h., es gilt  $f(x^*) < \infty$ , und es existiert eine Umgebung  $U(x^*)$  mit  $f(x^*) \leq f(x)$  für alle  $x \in U(x^*)$ , vgl. Definition 14.1. Im Fall  $f(x^*) = -\infty$  ist  $x^*$  zweifelsohne auch ein globaler Minimierer. Wir nehmen also nun an, dass  $f(x^*)$  endlich ist. Angenommen, es gäbe ein  $\hat{x} \in \mathbb{R}^n$  mit  $f(\hat{x}) < f(x^*)$ . Für  $\alpha \in (0, 1]$  gilt dann

$$f(\alpha \hat{x} + (1 - \alpha) x^*) \leq \alpha f(\hat{x}) + (1 - \alpha) f(x^*) < f(x^*),$$

wobei alle Terme wohldefiniert sind. Für  $\alpha$  hinreichend klein liegt aber  $\alpha \hat{x} + (1 - \alpha) x^* \in U(x^*)$ , im Widerspruch zur lokalen Optimalität von  $x^*$ . Also kann ein solches  $\hat{x}$  nicht existieren, d. h., es ist

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

*Aussage (ii):* Es seien  $x^*$  und  $x^{**}$  Lösungen von (14.1), also  $f(x^*) = f(x^{**}) = f^*$  (Optimalwert). Für  $\alpha \in [0, 1]$  gilt

$$f(\alpha x^* + (1 - \alpha) x^{**}) \leq \alpha f(x^*) + (1 - \alpha) f(x^{**}) = f^*.$$

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), insbesondere gilt  $x^*, x^{**} \in \text{dom } f$ . Da  $f$  eigentlich ist, haben wir außerdem  $f(x^*) = f(x^{**}) = f^* \in \mathbb{R}$  (also nicht  $-\infty$ ). Die Konvexkombination  $(x^* + x^{**})/2$  liegt ebenfalls in  $\text{dom } f$ , daher folgt aus der strikten Konvexität (13.9)

$$f\left(\frac{x^* + x^{**}}{2}\right) < \frac{1}{2}f(x^*) + \frac{1}{2}f(x^{**}) = f^*,$$

im Widerspruch zur Optimalität von  $x^*$  und  $x^{**}$ . □

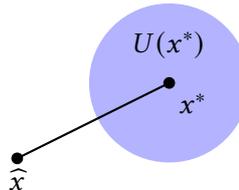


Abbildung 14.1: Illustration zum Beweis von Satz 14.2 (i) (lokale Minimierer sind globale Minimierer).

## § 15 TRENNUNGSSÄTZE FÜR KONVEXE MENGEN

Die Trennung zweier konvexer Mengen mittels einer Hyperebene sind ein zentrales Hilfsmittel in der konvexen Analysis und damit auch in der konvexen Optimierung. Nach einigen Vorbereitungen zu den topologischen Eigenschaften konvexer Mengen folgen der **Trennungssatz 15.26**, der **eigentliche Trennungssatz 15.30** und der **strikte Trennungssatz 15.32** in § 15.4.

### § 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** (Projektionsaufgabe). *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:  $\text{proj}_C(p)$ . Als Optimierungsaufgabe können wir dies in der Form*

$$\text{Minimiere } f(x) := \|x - p\| + \delta_C(x) \quad \text{über } x \in \mathbb{R}^n \quad (15.1)$$

oder auch als

$$\text{Minimiere } g(x) := \frac{1}{2}\|x - p\|^2 + \delta_C(x) \quad \text{über } x \in \mathbb{R}^n \quad (15.2)$$

schreiben. Beide Zielfunktionen in (15.1) und (15.2) sind eigentlich (**Quizfrage 15.1: Warum?**).

**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:  $\text{proj}_C(p)$ .

*Beweis.* Es sei  $p \in \mathbb{R}^n$ . Wir betrachten zunächst (15.2). Um die Existenz einer Lösung zu zeigen, führen wir 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 15.2:** 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 1.6 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 14.2 (iii).

Mit Hilfe von Hausaufgabe 1.3 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.  $\square$

**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  $x^* = \text{proj}_C(p)$  genau dann, wenn  $x^* \in C$  ist und gilt:

$$(x^* - p)^\top (x - x^*) \geq 0 \quad \text{für alle } x \in C. \tag{15.4}$$

**Beachte:** (15.4) ist eine **Variationsungleichung**. Sie besagt, dass der Winkel zwischen  $x^* - p$  und  $x - x^*$   $90^\circ$  nicht übersteigen darf. Anders ausgedrückt:  $C$  ist enthalten im Halbraum  $H^+(a, \beta) = \{x \in \mathbb{R}^n \mid a^\top x \geq \beta\}$  mit Normalenvektor  $a = x^* - p$  und  $\beta = a^\top x^*$ , vgl. Abbildung 15.1.

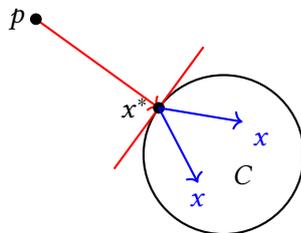


Abbildung 15.1: Der Winkel zwischen  $x^* - p$  und  $x - x^*$  darf  $90^\circ$  nicht übersteigen.

*Beweis.* Wir definieren wie in (15.1) und (15.2) die Funktionen

$$f(x) := \|x - y\| + \delta_C(x) \quad \text{und} \quad g(x) := \frac{1}{2}\|x - y\|^2 + \delta_C(x).$$

„ $\Rightarrow$ “: Es sei  $x^* = \text{proj}_C(p)$ , 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

$$\begin{aligned} \frac{1}{2}\|x^* - p\|^2 = g(x^*) &\leq g(x^* + \alpha(x - x^*)) = \frac{1}{2}\|(x^* - p) + \alpha(x - x^*)\|^2 \\ \Rightarrow 0 &\leq \alpha(x^* - p)^\top(x - x^*) + \frac{\alpha^2}{2}\|x - x^*\|^2. \end{aligned}$$

Division durch  $\alpha$  und Grenzübergang  $\alpha \searrow 0$  liefern die Behauptung (15.4).

„ $\Leftarrow$ “: Es gelte  $x^* \in C$  und (15.4). Daraus folgt

$$\begin{aligned} 0 &\geq (p - x^*)^\top(x - x^*) \quad \text{für alle } x \in C \\ &= (p - x^*)^\top(x - p + p - x^*) \\ &= (p - x^*)^\top(x - p) + \|p - x^*\|^2 \\ &\geq -\|p - x^*\|\|x - p\| + \|p - x^*\|^2 \quad (\text{Cauchy-Schwarz}). \end{aligned}$$

Daraus folgt weiter  $f(x) = \|x - p\| \geq \|p - x^*\| = f(x^*)$  für alle  $x \in C$ , d. h.,  $x^* = \text{proj}_C(p)$ .  $\square$

## § 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.* „ $\Leftarrow$ “: 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).

„ $\Rightarrow$ “: 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

$$\begin{aligned}
 u_1 + u_2 &= x_1 - x_0 + x_2 - x_0 \\
 &= \overbrace{\frac{1}{2}(2x_1 - x_0)}_{\in A} + \left(1 - \frac{1}{2}\right) \overbrace{(2x_2 - x_0)}_{\in A} - x_0 \in A - x_0.
 \end{aligned}
 \tag{15.6}$$

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

$$x - x_0 + \tilde{x}_0 = x - x_0 + \underbrace{\tilde{x}_0 - x_0}_{\in \{x - x_0 \mid x \in A\}} + x_0 \in A,$$

ist der Unterraum  $U$  unabhängig vom gewählten Aufpunkt  $x_0$ . □

**Beachte:** Ein affiner Unterraum ist nach (15.5) also ein „verschobener“ Unterraum.

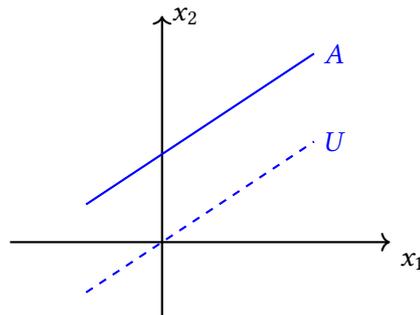


Abbildung 15.2: Ein 1-dimensionaler affiner Unterraum  $A$  von  $\mathbb{R}^2$ . Der zugehörige Richtungsraum  $U$  ist gestrichelt gezeichnet.

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  $\dim A := \dim U = -1$ .

**Quizfrage 15.3:** Welche Gestalt hat  $A$  im Fall  $\dim A = 0$ ?

**Definition 15.6** (Affine Unabhängigkeit, Affinkombination).

- (i) Eine Menge von Vektoren  $\{x_0, x_1, \dots, x_k\}$  in  $\mathbb{R}^n$ ,  $k \in \mathbb{N}_0$ , heißt **affin unabhängig** (englisch: **affine independent**), wenn die Vektoren  $x_1 - x_0, \dots, x_k - x_0$  linear unabhängig sind.<sup>1</sup>
- (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  $x_0, \dots, x_m \in \mathbb{R}^n$ ,  $m \in \mathbb{N}_0$ , 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  $x_0, \dots, x_m \in M$  ist.

**Lemma 15.7** (Dimension eines affinen Unterraums).

Es sei  $A \subset \mathbb{R}^n$  ein affiner Unterraum.

- (i)  $A$  besitzt genau dann eine affine Basis  $\{x_0, x_1, \dots, x_k\}$  mit  $k + 1$  Elementen, wenn  $\dim A = k$  ist.
- (ii) Ist  $\{x_0, x_1, \dots, x_k\}$  eine affine Basis von  $A$ , dann lässt sich jedes Element von  $A$  auf eindeutige Art und Weise aus  $\{x_0, x_1, \dots, x_k\}$  affinkombinieren. Genauer hat jedes  $x \in A$  die Darstellung

$$x = \sum_{i=0}^k \alpha_i x_i \quad (15.7)$$

mit Koeffizienten  $\alpha = (\alpha_0, \dots, \alpha_k)^\top$ , die sich aus der eindeutigen Lösung des linearen Gleichungssystems

$$\underbrace{\begin{bmatrix} 1 & \cdots & 1 \\ | & & | \\ x_0 & \cdots & x_k \\ | & & | \end{bmatrix}}_{=:B} \underbrace{\begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_k \end{pmatrix}}_{=:b} = \underbrace{\begin{pmatrix} 1 \\ | \\ x \\ | \end{pmatrix}}_{=:b} \quad (15.8)$$

ergeben. Die Matrix  $B \in \mathbb{R}^{(n+1) \times (k+1)}$  hat Rang  $k + 1$ . Daher ist  $B^\top B$  regulär, und (15.8) kann äquivalent als

$$B^\top B \alpha = B^\top b \quad (15.9)$$

geschrieben werden.

**Beachte:** Im Fall von  $A = \emptyset$  ist  $k = -1$  und die affine Basis die leere Menge.

<sup>1</sup>Ein einzelner Vektor  $x_0 \in \mathbb{R}^n$  ist also immer affin unabhängig.

*Beweis.* Der Beweis ist Inhalt von [Hausaufgabe 10.3](#). □

**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 letzterer 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

$$\text{aff}(M) = \bigcap \{A \subseteq \mathbb{R}^n \mid A \text{ ist affiner Unterraum von } \mathbb{R}^n \text{ und } M \subseteq A\}, \quad (15.10)$$

heißt die **affine Hülle** von  $M$ .  $\text{aff}(M)$  ist also der kleinste affine Unterraum, der  $M$  enthält.

**Beachte:** Es gilt  $M \subseteq \underbrace{\text{aff}(M)}_{\text{„Hülle“}} \subseteq \underbrace{\text{aff}(M)}_{\text{affin}}$ , daher der Name **affine Hülle**.

**Definition 15.10** (Dimension einer Menge). Es sei  $M \subseteq \mathbb{R}^n$ . Die (**affine**) **Dimension** von  $M$  ist  $\dim M := \dim \text{aff}(M)$ .

**Quizfrage 15.4:** 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)  $\text{aff}(M)$  ist gleich der Menge aller Affinkombinationen von  $M$ .
- (iii) Es gilt  $M \subseteq \text{conv}(M) \subseteq \text{aff}(M)$ .

**Beachte:**  $A$  ist affiner Unterraum  $\Leftrightarrow A = \text{aff}(A)$ .

*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  $m \geq 2$ : 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  $\alpha_{m+1} \neq 1$ . (Es können nur dann alle Koeffizienten  $\alpha_i = 1$  sein, wenn  $m = 0$  ist.) 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.

„ $\text{aff}(M) \subseteq A$ “: Wir zeigen:  $A$  ist ein affiner Unterraum. Damit kommt diese Menge im Durchschnitt (15.10) vor, also gilt  $\text{aff}(M) \subseteq A$ .

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.

„ $\text{aff}(M) \supseteq A$ “: Es sei  $x \in A$ , also eine Affinkombination von  $M$ . Wegen  $M \subseteq \text{aff}(M)$  ist  $x$  auch eine Affinkombination von  $\text{aff}(M)$ .  $\text{aff}(M)$  ist ein affiner Unterraum, stimmt also nach **Aussage (i)** mit der Menge seiner Affinkombinationen überein. Also ist  $x \in \text{aff}(M)$ .

**Aussage (iii):** Nach **Lemma 13.7** gilt  $M \subseteq \text{conv}(M)$ , und  $\text{conv}(M)$  sind gerade die Konvexkombinationen von  $M$ . Da jede Konvexkombination auch eine Affinkombination ist, gilt weiter  $\text{conv}(M) \subseteq \text{aff}(M)$ .  $\square$

**Quizfrage 15.5:** Welche Beziehung besteht zwischen  $\text{aff}(M_1 + M_2)$  und  $\text{aff}(M_1) + \text{aff}(M_2)$  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  $\text{aff}(M)$  bilden.*

*Beweis.* Die affine Hülle  $\text{aff}(M)$  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.  $\square$

### § 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  $x \in \text{conv}(M)$  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  $x \in \text{conv}(M)$  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 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.

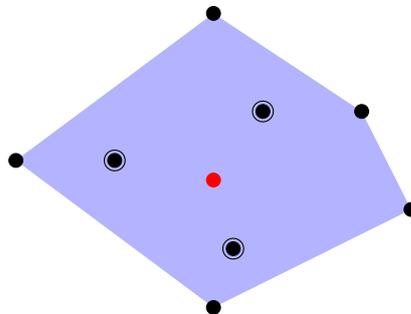


Abbildung 15.3: Illustration des [Satzes von Carathéodory 15.13](#) für eine Menge der Dimension  $k = 2$ . Die schwarzen Punkte bilden die Menge  $M$ . Der rote Punkt  $x \in \text{conv}(M)$  ist bereits eine Konvexkombination der drei hervorgehobenen Punkte.

*Beweis.* Der Beweis ist Inhalt von [Hausaufgabe 10.4](#). □

Eine konvexe Menge  $C \subseteq \mathbb{R}^n$ , deren Dimension  $\dim C < n$  ist, besitzt keine inneren Punkte in  $\mathbb{R}^n$ . ([Quizfrage 15.6](#): Ist das anschaulich klar?) Das ändert sich, wenn man zur Relativtopologie in  $\text{aff}(C)$  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  $B_\varepsilon(x) \cap \text{aff}(C) \subseteq C$  liegt. Die Menge aller relativ inneren Punkte von  $C$  heißt das **relative Innere** (englisch: **relative interior**) und wird mit  $\text{rel int}(C)$  bezeichnet.
- (ii) Ein Punkt  $x \in \mathbb{R}^n$  heißt ein **relativer Randpunkt** von  $C$ , wenn  $x \in \overline{C} \setminus \text{rel int}(C)$  liegt. Die Menge aller relativen Randpunkte von  $C$  heißt der **relative Rand** (englisch: **relative boundary**) und wird mit  $\text{rel } \partial(C)$  bezeichnet.

**Quizfrage 15.7:** Was ist  $\text{rel int}(C)$ , wenn die konvexe Menge  $C \subseteq \mathbb{R}^n$  die volle Dimension  $n$  hat?

**Quizfrage 15.8:** Was ist  $\text{rel int}(A)$  für einen affinen Unterraum  $A \subseteq \mathbb{R}^n$ ?

**Quizfrage 15.9:** Wenn  $C_1 \subseteq C_2$  ist, gilt dann auch immer  $\text{rel int}(C_1) \subseteq \text{rel int}(C_2)$ ?

**Quizfrage 15.10:** Warum benötigen wir keinen Begriff des relativen Abschlusses?

**Satz 15.15** (vgl. [Jarre, Stoer, 2004](#), Satz 7.2.5).

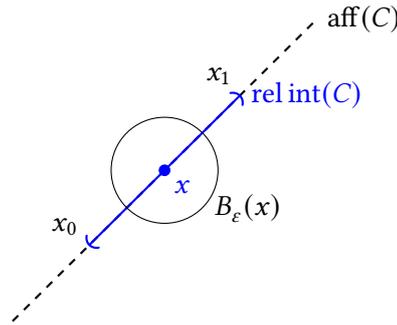


Abbildung 15.4: Illustration einer 1-dimensionalen konvexen Menge  $C = \text{conv}(\{x_0, x_1\}) = \{\alpha x_0 + (1 - \alpha) x_1 \mid \alpha \in [0, 1]\}$  und ihres relativen Inneren  $\text{rel int}(C) = \{\alpha x_0 + (1 - \alpha) x_1 \mid \alpha \in (0, 1)\}$  in  $\mathbb{R}^2$ . Insbesondere der Punkt  $x$  ist ein relativ innerer Punkt.

Jede nichtleere konvexe Menge  $C \subseteq \mathbb{R}^n$  besitzt ein nichtleeres relatives Inneres. Es gilt:  $\dim \text{rel int}(C) = \dim C$ .

*Beweis.* Es sei  $k := \dim C$ . Da  $C$  nichtleer ist, gilt,  $k \geq 0$ . Es gibt also nach Satz 15.12 affin unabhängige Punkte  $x_0, \dots, x_k \in C$ . Nach Lemma 15.7 lässt sich jeder Punkt  $x \in \text{aff}(C)$  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:

$$B^T B \alpha = B^T b.$$

Wir zeigen jetzt, dass der Mittelpunkt der Punkte  $\{x_0, \dots, x_k\}$

$$\bar{x} := \frac{1}{k+1} \sum_{i=0}^k x_i$$

ein relativ innerer Punkt von  $C$  ist. Dazu konstruieren wir eine abgeschlossene Kugel

$$\overline{B_\epsilon^{\|\cdot\|_\infty}(\bar{x})} = \{x \in \mathbb{R}^n \mid \|x - \bar{x}\|_\infty \leq \epsilon\}$$

mit der Eigenschaft  $\overline{B_\epsilon^{\|\cdot\|_\infty}(\bar{x})} \cap \text{aff}(C) \subseteq \text{conv}(\{x_0, \dots, x_k\}) \subseteq \text{conv}(C) = C$ . (Quizfrage 15.11: Warum gilt die letzte Inklusion?)

Wir setzen dazu  $\epsilon := 1/((k+1)\|(B^T B)^{-1} B^T\|_\infty)$ . Dabei ist  $\|\cdot\|_\infty$  die durch die  $\infty$ -Norm im Definitionsbereich und Bildbereich induzierte Matrixnorm, also<sup>2</sup>

$$\|A\|_\infty = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{\|x\|_\infty=1} \|Ax\|_\infty. \tag{15.11}$$

<sup>2</sup>Diese ist auch als Zeilensummennorm bekannt, da für  $A \in \mathbb{R}^{m \times n}$  die folgende Beziehung gilt:

$$\|A\|_\infty = \max \left\{ \sum_{j=1}^n |a_{ij}| \mid i = 1, \dots, m \right\}.$$

$\|A\|_\infty$  ist also das Maximum der betragsweisen Zeilensummen.

Es sei nun  $x \in \overline{B_\varepsilon^{\|\cdot\|_\infty}(\bar{x})} \cap \text{aff}(C)$ . Wir bezeichnen mit  $\alpha$  die Koeffizienten in der Affinkombination  $x = \sum_{i=0}^k \alpha_i x_i$ . Weiterhin sind  $\bar{\alpha}$  die Koeffizienten von  $\bar{x}$ , also  $\bar{\alpha} = \frac{1}{k+1} \mathbf{1}$ . Um zu zeigen, dass tatsächlich  $x \in \text{conv}(\{x_0, \dots, x_k\})$  liegt, müssen wir für die Koeffizienten zeigen:  $\alpha \geq 0$ . Es gilt

$$\begin{aligned} \|\alpha - \bar{\alpha}\|_\infty &= \|(B^\top B)^{-1} B^\top \left[ \begin{pmatrix} 1 \\ x \end{pmatrix} - \begin{pmatrix} 1 \\ \bar{x} \end{pmatrix} \right]\|_\infty \leq \|(B^\top B)^{-1} B^\top\|_\infty \left\| \begin{pmatrix} 1 \\ x \end{pmatrix} - \begin{pmatrix} 1 \\ \bar{x} \end{pmatrix} \right\|_\infty \\ &= \|(B^\top B)^{-1} B^\top\|_\infty \|x - \bar{x}\|_\infty \leq \|(B^\top B)^{-1} B^\top\|_\infty \varepsilon = \frac{1}{k+1}. \end{aligned}$$

Daraus folgt wie gewünscht  $\alpha \geq 0$  mit Hilfe der Dreiecksungleichung.

Wir zeigen jetzt noch, dass die Dimension von  $\text{rel int}(C)$  ebenfalls gleich  $k$  ist. Dazu geben wir  $k+1$  affin unabhängige Punkte in  $\text{rel int}(C)$  an. Damit ist  $\dim \text{rel int}(C) \geq k$ , und da außerdem  $\text{rel int}(C) \subseteq C$  die Beziehung  $\dim \text{rel int}(C) \leq \dim C = k$  impliziert, ist dann die Behauptung gezeigt. Zur Konstruktion der Punkte machen wir den Ansatz

$$\bar{x}_i := \beta \bar{x} + (1 - \beta) x_i$$

und wählen  $\beta \in (0, 1)$  so klein, dass  $\|\bar{x}_i - \bar{x}\|_\infty \leq \varepsilon$  bleibt für alle  $i = 0, \dots, k$ . Damit liegen alle  $\bar{x}_i \in \overline{B_\varepsilon^{\|\cdot\|_\infty}(\bar{x})}$  und außerdem in  $\text{aff}(C)$  (**Quizfrage 15.12**: Begründung?).

Um die affine Unabhängigkeit der Punkte  $\{\bar{x}_0, \dots, \bar{x}_k\}$  zu zeigen, machen wir den Ansatz

$$\begin{aligned} 0 &= \sum_{i=1}^k \gamma_i (\bar{x}_i - \bar{x}_0) = \sum_{i=1}^k \gamma_i (\beta \bar{x} + (1 - \beta) x_i - \beta \bar{x} - (1 - \beta) x_0) \\ &= (1 - \beta) \sum_{i=1}^k \gamma_i (x_i - x_0). \end{aligned}$$

Da die Punkte  $\{x_0, \dots, x_k\}$  aber affin unabhängig sind, folgt  $\gamma_i = 0$  für alle  $i = 1, \dots, k$ . Also sind auch die Punkte  $\{\bar{x}_0, \dots, \bar{x}_k\}$  affin unabhängig.  $\square$

**Lemma 15.16 (Accessibility lemma, vgl. Jarre, Stoer, 2004, Lemma 7.2.6).**

Es seien  $C \subseteq \mathbb{R}^n$  konvex,  $x_1 \in \bar{C}$  und  $x_2 \in \text{rel int}(C)$ . Dann gilt  $\alpha x_1 + (1 - \alpha) x_2 \in \text{rel int}(C)$  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  $x_1 \in \bar{C}$  gilt  $x_1 \in C + B_\varepsilon(0)$  für alle  $\varepsilon > 0$ . Wegen  $x_2 \in \text{rel int}(C)$  gibt es ein  $r > 0$ , sodass  $B_r(x_2) \cap \text{aff}(C) \subseteq C$  liegt.

Es sei nun  $\alpha \in [0, 1)$  beliebig und  $x := \alpha x_1 + (1 - \alpha) x_2$ . Wir zeigen: Mit  $r(\alpha) := \frac{1-\alpha}{1+\alpha} r$  gilt  $B_{r(\alpha)}(x) \cap \text{aff}(C) \subseteq C$ , d. h.,  $x \in \text{rel int}(C)$ . Dazu halten wir zunächst fest:

$$\begin{aligned} B_{r(\alpha)}(x) &= B_{r(\alpha)}(\alpha x_1 + (1 - \alpha) x_2) \\ &= \alpha x_1 + (1 - \alpha) x_2 + B_{r(\alpha)}(0) \\ &\subseteq \alpha (C + B_{r(\alpha)}(0)) + (1 - \alpha) x_2 + B_{r(\alpha)}(0) \\ &= \alpha C + (1 - \alpha) x_2 + (1 + \alpha) B_{r(\alpha)}(0) \\ &= \alpha C + (1 - \alpha) [x_2 + B_r(0)] \\ &= \alpha C + (1 - \alpha) B_r(x_2), \end{aligned}$$

siehe auch [Abbildung 15.5](#). Durch den Schnitt mit  $\text{aff}(C)$  folgt

$$B_{r(\alpha)}(x) \cap \text{aff}(C) \subseteq [\alpha C + (1 - \alpha) B_r(x_2)] \cap \text{aff}(C).$$

Wegen  $\text{aff}(C) = \alpha \text{aff}(C) + (1 - \alpha) \text{aff}(C)$  (**Quizfrage 15.13**: Begründung?) gilt weiter

$$B_{r(\alpha)}(x) \cap \text{aff}(C) \subseteq \alpha \underbrace{[C \cap \text{aff}(C)]}_{\subseteq C} + (1 - \alpha) \underbrace{[B_r(x_2) \cap \text{aff}(C)]}_{\subseteq C} \subseteq C.$$

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 einer konvexen Menge).

Es sei  $C \subseteq \mathbb{R}^n$  konvex und nichtleer. Für einen Punkt  $x \in \mathbb{R}^n$  sind folgende Aussagen äquivalent:

- (i)  $x \in \text{rel int}(C)$ .
- (ii) Zu jedem  $y \in \text{aff}(C)$  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 15.14**: Was bedeutet die Bedingung aus [Aussage \(iii\)](#) anschaulich?

*Beweis.* [Aussage \(i\)](#)  $\Rightarrow$  [Aussage \(ii\)](#): Es sei  $x \in \text{rel int}(C)$  und  $y \in \text{aff}(C)$ . Wir können  $y \neq x$  annehmen, sonst ist die Aussage klar. Wegen  $x \in \text{rel int}(C)$  gibt es ein  $\tilde{\varepsilon} > 0$ , sodass  $B_{\tilde{\varepsilon}}(x) \cap \text{aff}(C) \subseteq C$ . Für  $\varepsilon := \tilde{\varepsilon}/\|y - x\|$  sind dann

$$\underbrace{x + \varepsilon(y - x)}_{=\varepsilon y + (1-\varepsilon)x \in \text{aff}(C)} \in B_{\tilde{\varepsilon}}(x) \cap \text{aff}(C) \subseteq C \quad \text{und} \quad \underbrace{x - \varepsilon(y - x)}_{=-\varepsilon y + (1+\varepsilon)x \in \text{aff}(C)} \in B_{\tilde{\varepsilon}}(x) \cap \text{aff}(C) \subseteq C.$$

[Aussage \(ii\)](#)  $\Rightarrow$  [Aussage \(iii\)](#): Klar, da  $C \subseteq \text{aff}(C)$  ist.

[Aussage \(iii\)](#)  $\Rightarrow$  [Aussage \(i\)](#): Nach [Satz 15.15](#) können wir ein  $y \in \text{rel int}(C)$  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  $y \in \text{rel int}(C)$ . Nach dem [Accessibility lemma 15.16](#) gehört also  $x$  zu  $\text{rel int}(C)$ . □

**Lemma 15.18** (Relatives Inneres des Epigraphen, vgl. Rockafellar, 1970, Lemma 7.3).  
Es sei  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex. Dann gilt

$$\text{rel int epi } f = \left\{ \begin{pmatrix} x \\ \gamma \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \mid x \in \text{rel int dom } f, \gamma > f(x) \right\}. \quad (15.12)$$

Beweis. □

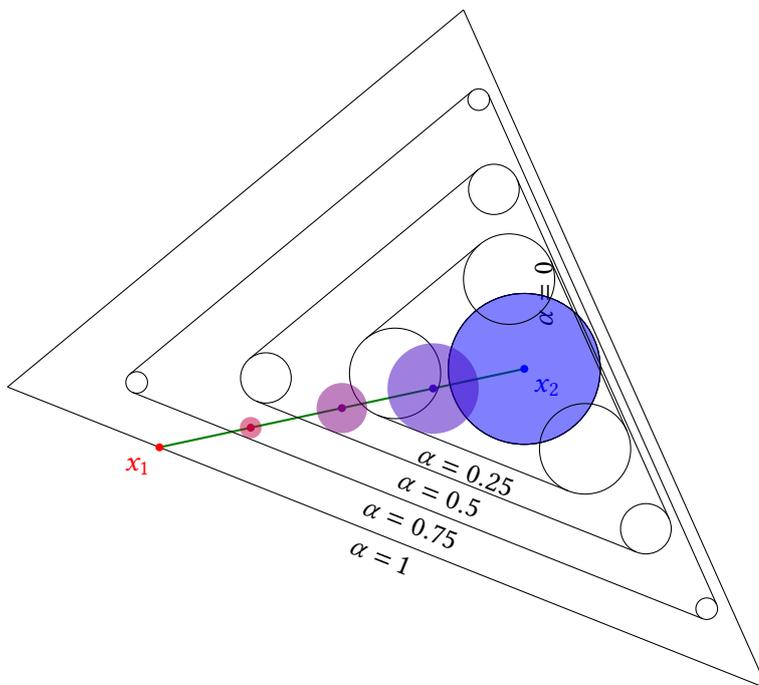


Abbildung 15.5: Illustration der Inklusion  $B_{r(\alpha)}(x) \subseteq \alpha C + (1 - \alpha) B_r(x_2)$  aus dem Beweis des **Accessibility lemmas 15.16** für  $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ . Die farbig hervorgehobenen Kugeln sind die Mengen  $B_{r(\alpha)}(x)$ . Die Mengen  $\alpha C + (1 - \alpha) B_r(x_2)$  sind die Dreiecke mit den abgerundeten Ecken. Die Extremfälle sind  $\alpha = 1$  (die Menge  $C$  selbst, also das äußere Dreieck) und  $\alpha = 0$  (die blaue Kugel, kein Dreieck mehr zu erkennen).

**Satz 15.19** (Rockafellar, 1970, Theorem 6.2).

- (i) Ist  $C \subseteq \mathbb{R}^n$  konvex, dann sind das relative Innere  $\text{rel int}(C)$  und der Abschluss  $\bar{C}$  konvex.
- (ii) Die Mengen  $C$ ,  $\text{rel int}(C)$  und  $\bar{C}$  haben alle dieselbe affine Hülle, also auch dieselbe Dimension.

*Beweis. Aussage (i):* Es sei  $C \subseteq \mathbb{R}^n$  konvex. Wir zeigen zuerst, dass  $\text{rel int}(C)$  konvex ist. Es seien dazu  $x_1, x_2 \in \text{rel int}(C)$  und  $\alpha \in [0, 1]$ . Dann gehört nach **Lemma 15.16** auch die Konvexkombination  $\alpha x_1 + (1 - \alpha) x_2$  zu  $\text{rel int}(C)$ , also ist  $\text{rel int}(C)$  konvex.

Der Abschluss von  $C$  erfüllt die Beziehung

$$\bar{C} = \bigcap \{C + \overline{B_\varepsilon(0)} \mid \varepsilon > 0\}.$$

Die Mengen, über die der Durchschnitt gebildet wird, sind nach [Beispiel 13.2](#) und [Satz 13.3](#) konvex, also ist auch  $\bar{C}$  konvex.

**Aussage (ii):** Wir zeigen jetzt noch  $\text{aff rel int}(C) = \text{aff } C = \text{aff } \bar{C}$ . Aufgrund von  $\text{rel int}(C) \subseteq C \subseteq \bar{C}$  folgt

$$\text{aff rel int}(C) \subseteq \text{aff } C \subseteq \text{aff } \bar{C}.$$

Wegen  $\bar{C} \subseteq \text{aff}(C)$  gilt auch  $\text{aff } \bar{C} \subseteq \text{aff } \text{aff}(C) = \text{aff}(C)$ . Also bleibt noch  $\text{aff rel int}(C) = \text{aff } C$  zu zeigen. Wir wissen aber bereits aus [Satz 15.15](#), dass  $\dim \text{rel int}(C) = \dim C$  gilt, also muss  $\text{aff rel int}(C) = \text{aff } C$  sein.  $\square$

**Satz 15.20** ([Rockafellar, 1970](#), Theorem 6.3).

Es sei  $C \subseteq \mathbb{R}^n$  konvex. Dann gelten:

- (i)  $\overline{\text{rel int}(C)} = \bar{C}$ , d. h.,  $\text{rel int}(C)$  and  $C$  haben denselben Abschluss.
- (ii)  $\text{rel int}(C) = \text{rel int}(\bar{C})$ , d. h.,  $C$  und  $\bar{C}$  haben dasselbe relative Innere.
- (iii)  $\text{rel } \partial(C) = \text{rel } \partial(\bar{C})$ , d. h.,  $C$  und  $\bar{C}$  haben denselben relativen Rand.

*Beweis.* Wir können  $C \neq \emptyset$  annehmen, ansonsten sind alle Mengen leer.

**Aussage (i):** Die Inklusion  $\overline{\text{rel int}(C)} \subseteq \bar{C}$  folgt unmittelbar aus  $\text{rel int}(C) \subseteq C$ . Für die umgekehrte Aussage sei nun  $x_1 \in \bar{C}$  und  $x_2 \in \text{rel int}(C)$ . Dann ist nach [Lemma 15.16](#)  $\alpha x_1 + (1 - \alpha) x_2 \in \text{rel int}(C)$ . Der Grenzübergang  $\alpha \nearrow 1$  zeigt  $x_1 \in \text{rel int}(C)$ .

**Aussage (ii):** Aus  $C \subseteq \bar{C}$  und der Tatsache, dass beide Mengen dieselbe affine Hülle haben ([Satz 15.19](#)) folgt  $\text{rel int}(C) \subseteq \text{rel int}(\bar{C})$ . Für die umgekehrte Aussage sei nun  $x \in \text{rel int}(\bar{C})$ , also existiert  $\varepsilon > 0$  mit  $B_\varepsilon(x) \cap \text{aff}(C) \subseteq \bar{C}$ . Es sei außerdem  $y \in \text{rel int}(C)$  ([Satz 15.15](#)). Ist  $y = x$ , so sind wir fertig. Es sei also jetzt  $y \neq x$ . Ziel ist die Konstruktion eines Punktes  $z \in \bar{C}$ , sodass  $x$  als echte Konvexkombination von  $y \in \text{rel int}(C)$  und  $z \in \bar{C}$  geschrieben werden kann. Denn dann folgt aus [Lemma 15.16](#), dass  $x \in \text{rel int}(C)$  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  $z \in B_\varepsilon(x)$ . Wegen  $B_\varepsilon(x) \cap \text{aff}(C) \subseteq \bar{C}$  (**Quizfrage 15.15**: Begründung?) folgt auch  $z \in \bar{C}$ . Wir können nun  $x$  schreiben als echte Konvexkombination

$$x = \frac{1}{1 + \delta} z + \left(1 - \frac{1}{1 + \delta}\right) y$$

mit  $z \in \bar{C}$  und  $y \in \text{rel int}(C)$ . Nach [Lemma 15.16](#) gilt  $x \in \text{rel int}(C)$ , was zu zeigen war.

**Aussage (iii):** Nach Definition des relativen Randes gilt

$$\text{rel } \partial(C) = \bar{C} \setminus \text{rel int}(C)$$

und weiter

$$\text{rel int}(\overline{C}) = \overline{C} \setminus \text{rel int}(\overline{C}) = \overline{C} \setminus \text{rel int}(C),$$

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)  $\text{rel int}(C_1) = \text{rel int}(C_2)$ .

*Beweis.* [Aussage \(i\)](#)  $\Rightarrow$  [Aussage \(ii\)](#): Aus  $\overline{C_1} = \overline{C_2}$  folgt  $\text{rel int}(\overline{C_1}) = \text{rel int}(\overline{C_2})$ . [Satz 15.20 \(ii\)](#) impliziert  $\text{rel int}(C_1) = \text{rel int}(C_2)$ .

[Aussage \(ii\)](#)  $\Rightarrow$  [Aussage \(i\)](#): Aus  $\text{rel int}(C_1) = \text{rel int}(C_2)$  folgt  $\overline{\text{rel int}(C_1)} = \overline{\text{rel int}(C_2)}$ . [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:  $\text{core } M$ .

Es ist leicht zu sehen, dass  $\text{int } M \subseteq \text{core } M$  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  $\text{int } C = \text{core } C$ .

*Beweis.* Es ist nur  $\text{core } C \subseteq \text{int } C$  zu bestätigen. Im Fall  $\text{core } C = \emptyset$  ist nichts zu zeigen. Es sei also  $x_0 \in \text{core } C$ . Dann ist  $\dim C = n$  und damit  $\text{aff } C = \mathbb{R}^n$ . (**Quizfrage 15.16:** Begründung?) Nach Definition des algebraischen Inneren ist die Bedingung aus [Lemma 15.17 \(ii\)](#) erfüllt. Damit gehört  $x_0$  zu  $\text{rel int } C = \text{int } C$ . □

§ 15.4 TRENNUNGSSÄTZE

**Literatur:** Geiger, Kanzow, 2002, Kapitel 2.1.4

Wir kommen nun zu den wichtigsten Resultaten in diesem Abschnitt, dem **Trennungssatz 15.26**, dem **eigentlichen Trennungssatz 15.30** und dem **strikten Trennungssatz 15.32**. Das Ziel ist jeweils die Trennung zweier konvexer Mengen durch eine Hyperebene, sodass jede der Mengen in einem anderen Halbraum liegt und – je nach Voraussetzung – gewisse Zusatzeigenschaften gelten.

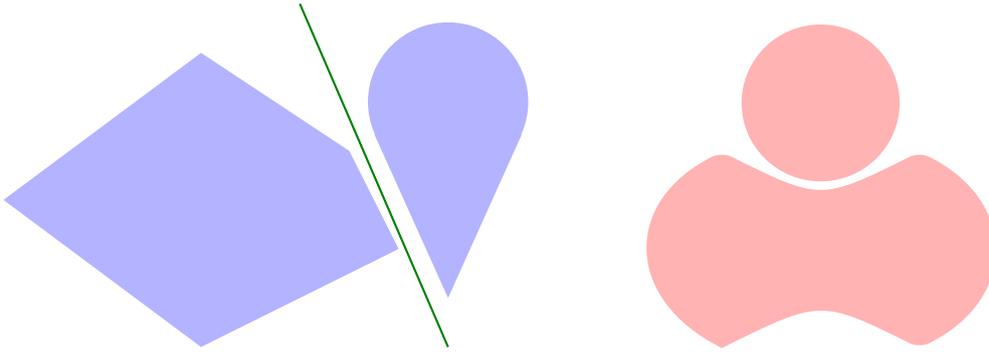


Abbildung 15.6: Zwei disjunkte konvexe Mengen (links) sind durch eine Hyperebene trennbar. Ist eine der Mengen nichtkonvex, stimmt diese Aussage nicht mehr (rechts).

**Definition 15.24** (Trennende Hyperebene).

Es seien  $A, B \subseteq \mathbb{R}^n$  zwei nichtleere Mengen und  $H(a, \beta) = \{x \in \mathbb{R}^n \mid a^\top x = \beta\}$  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

$$a^\top x \leq \beta \leq a^\top y \quad \text{für alle } x \in A \text{ und alle } y \in B. \tag{15.13}$$

(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:

$$a^\top x \leq \beta \leq a^\top y \quad \text{für alle } x \in A \text{ und alle } y \in B \tag{15.14a}$$

und

$$a^\top \bar{x} < a^\top \bar{y} \quad \text{für ein } \bar{x} \in A \text{ und ein } \bar{y} \in B. \tag{15.14b}$$

(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  $\text{int} H^-(a, \beta)$  und die andere im offenen Halbraum  $\text{int} H^+(a, \beta)$  enthalten ist, wenn also gilt

$$a^\top x < \beta < a^\top y \quad \text{für alle } x \in A \text{ und alle } y \in B. \tag{15.15}$$

**Beachte:** Der „Offset“  $\beta$  ist nicht die entscheidende Größe in der [Definition 15.24](#). Gilt beispielsweise  $a^\top x \leq a^\top y$  für alle  $x \in A$  und alle  $y \in B$ , dann lässt sich  $\beta$  in [\(15.13\)](#) immer nachträglich passend definieren.

**Quizfrage 15.17:** Wie kann man in den einzelnen Fällen der [Definition 15.24](#) das  $\beta$  jeweils passend definieren?

Wir beginnen mit einer Aussage zur Trennung eines Punktes und einer konvexen Menge. Aus diesem Spezialfall leiten wir dann anschließend den [Trennungssatz 15.26](#) her.

**Lemma 15.25** (Trennung von Punkt und konvexer Menge).

Es sei  $C \subseteq \mathbb{R}^n$  konvex und nichtleer und  $\bar{x} \in \mathbb{R}^n$  ein Punkt, sodass  $\bar{x} \notin \text{int}(C)$ . Dann existiert eine Hyperebene  $H(a, \beta)$ , die  $\bar{x}$  von  $C$  trennt, sodass also gilt:

$$a^\top x \geq \beta \geq a^\top \bar{x} \quad \text{für alle } x \in C. \quad (15.16)$$

*Beweis.* Wir wählen eine Folge  $x^{(k)}$  mit der Eigenschaft  $x^{(k)} \notin \bar{C}$  und  $x^{(k)} \rightarrow \bar{x}$  gilt. (**Quizfrage 15.18:** Wieso ist das möglich?) Die Menge  $\bar{C}$  ist nichtleer, abgeschlossen und konvex. Nach [Lemma 15.2](#) existiert die orthogonale Projektion  $\hat{x}^{(k)} := \text{proj}_{\bar{C}}(x^{(k)})$ , und nach [Satz 15.3](#) ist diese charakterisiert durch

$$\begin{aligned} (\hat{x}^{(k)} - x^{(k)})^\top x &\geq (\hat{x}^{(k)} - x^{(k)})^\top \hat{x}^{(k)} \quad \text{für alle } k \in \mathbb{N} \text{ und alle } x \in \bar{C} \\ &= (\hat{x}^{(k)} - x^{(k)})^\top (\hat{x}^{(k)} - x^{(k)} + x^{(k)}) \\ &= \|\hat{x}^{(k)} - x^{(k)}\|^2 + (\hat{x}^{(k)} - x^{(k)})^\top x^{(k)} \\ &\geq (\hat{x}^{(k)} - x^{(k)})^\top x^{(k)}. \end{aligned} \quad (15.17)$$

Wir setzen

$$a^{(k)} := \frac{\hat{x}^{(k)} - x^{(k)}}{\|\hat{x}^{(k)} - x^{(k)}\|}.$$

**Beachte:** Der Nenner ist  $\neq 0$ , da  $\hat{x}^{(k)} \in \bar{C}$  und  $x^{(k)} \notin \bar{C}$ .

Damit erhalten wir aus [\(15.17\)](#)

$$a^{(k)\top} x \geq a^{(k)\top} x^{(k)} \quad \text{für alle } k \in \mathbb{N} \text{ und alle } x \in \bar{C},$$

d. h., mit dem Normalenvektor  $a^{(k)}$  können wir jeweils den Punkt  $x^{(k)}$  von  $\bar{C}$  trennen. Wegen  $\|a^{(k)}\| = 1$  und der Kompaktheit der Einheitskugel in  $\mathbb{R}^n$  existiert eine konvergente Teilfolge  $a^{(k(\ell))} \rightarrow a$  mit  $\|a\| = 1$ , und der Grenzübergang  $\ell \rightarrow \infty$  zeigt die Behauptung

$$a^\top x \geq a^\top \bar{x} \quad \text{für alle } x \in \bar{C}. \quad \square$$

**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, sodass also gilt:

$$a^\top x_1 \leq \beta \leq a^\top x_2 \quad \text{für alle } x_1 \in C_1 \text{ und alle } x_2 \in C_2. \quad (15.18)$$

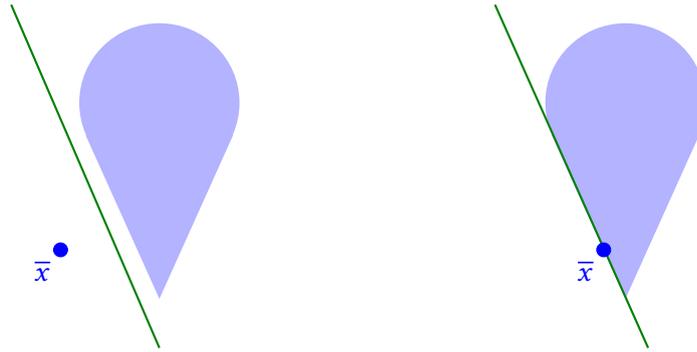


Abbildung 15.7: Illustration von Lemma 15.25 in zwei Fällen.

*Beweis.* Wir betrachten

$$C := C_2 - C_1 = \{x_2 - x_1 \mid x_1 \in C_1, x_2 \in C_2\},$$

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  $0 \notin \text{int}(C)$ . Aus Lemma 15.25 bekommen wir die Existenz einer Hyperebene  $H(a, \beta)$ , sodass  $0 \leq a^T x$  gilt für alle  $x \in C$ . Das heißt aber

$$a^T x_1 \leq a^T x_2 \quad \text{für alle } x_1 \in C_1 \text{ und alle } x_2 \in C_2.$$

□

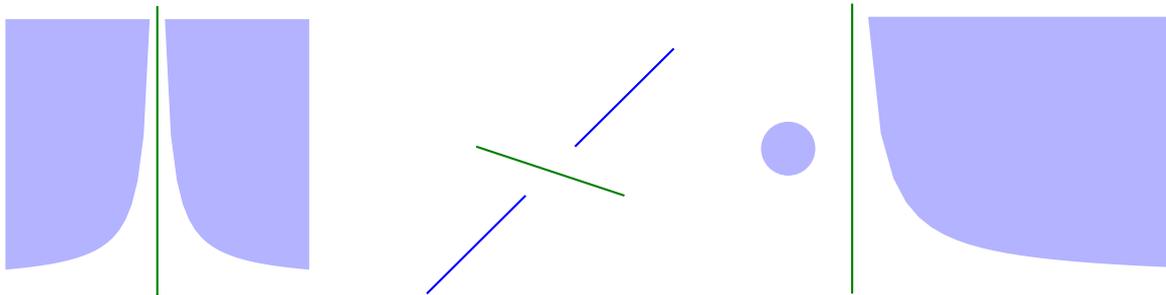


Abbildung 15.8: Illustration der Aussagen von Satz 15.26 (Trennung disjunkter konvexer Mengen), Satz 15.30 (eigentliche Trennung) und Satz 15.32 (strikte Trennung).

**Quizfrage 15.19:** Gilt auch die folgende Umkehrung des Satzes: „Wenn zwei nichtleere, konvexe Mengen  $C_1$  und  $C_2$  durch eine Hyperebene getrennt werden können, dann gilt notwendigerweise  $C_1 \cap C_2 = \emptyset$ “?

Für den folgenden, sogenannten **eigentlichen Trennungssatz 15.30** benötigen wir einige vorbereitende Aussagen.

**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 auch  $AC = \{Ax \mid x \in C\} \subseteq \mathbb{R}^m$  konvex, und es gilt:

$$(i) \quad \overline{AC} \subseteq \overline{AC}.$$

$$(ii) \quad A \operatorname{rel int}(C) = \operatorname{rel int}(AC).$$

*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  $\bar{x} \in \overline{C}$ , dann existiert eine Folge  $x^{(k)} \subseteq C$  mit der Eigenschaft  $x^{(k)} \rightarrow \bar{x}$ . Das impliziert  $Ax^{(k)} \in AC$  und  $Ax^{(k)} \rightarrow A\bar{x}$ , also gilt  $A\bar{x} \in \overline{AC}$ .

**Aussage (ii):** Es gilt

$$\begin{aligned} A \operatorname{rel int}(C) \subseteq AC \subseteq \overline{AC} &= \overline{A \operatorname{rel int}(C)} && \text{nach Satz 15.20} \\ &\subseteq \overline{A \operatorname{rel int}(C)} && \text{nach Aussage (i).} \end{aligned}$$

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 Inklusion 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.  $\square$

**Beachte:** Die [Aussage \(i\)](#) verwendet nur die Stetigkeit linearer Abbildungen und gilt auch für nicht-konvexe Mengen.

**Folgerung 15.28** (Relatives Inneres der Minkowski-Summe, vgl. [Jarre, Stoer, 2004](#), Satz 7.2.8).  
Es seien  $C_1, C_2 \subseteq \mathbb{R}^n$  konvex. Dann gilt

$$\operatorname{rel int}(C_1) + \operatorname{rel int}(C_2) = \operatorname{rel int}(C_1 + C_2). \quad (15.19)$$

*Beweis.* Setze  $C := C_1 \times C_2 \subseteq \mathbb{R}^n \times \mathbb{R}^n \cong \mathbb{R}^{2n}$  und  $A := \begin{bmatrix} \operatorname{Id} & \operatorname{Id} \end{bmatrix}$ , sodass also  $A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 + x_2$  gilt. Dann ist  $C$  nach [Satz 13.3](#) konvex, und aus [Lemma 15.27](#) folgt

$$\operatorname{rel int}(C_1) + \operatorname{rel int}(C_2) = A \operatorname{rel int}(C) = \operatorname{rel int}(AC) = \operatorname{rel int}(C_1 + C_2). \quad \square$$

Vor dem Beweis des [eigentlichen Trennungssatzes 15.30](#) betrachten wir wieder zunächst den Spezialfall, dass eine der zu trennenden Mengen nur aus einem Punkt besteht.

**Lemma 15.29** (Eigentliche Trennung von Punkt und konvexer Menge).

Es sei  $C \subseteq \mathbb{R}^n$  konvex und nichtleer. Falls  $0 \notin \text{rel int}(C)$  liegt, dann lassen sich  $0$  und  $C$  durch eine Hyperebene  $H(a, \beta)$  eigentlich trennen.

*Beweis.* Im Fall  $0 \notin \bar{C}$ , wählen wir  $a := \text{proj}_{\bar{C}}(0)$ . Dann ist  $a \neq 0$ . Der [Projektionssatz 15.3](#) impliziert

$$(a - 0)^\top(x - a) \geq 0 \quad \text{für alle } x \in C,$$

also gilt  $a^\top x \geq \|a\|^2 > 0 = a^\top 0$  für alle  $x \in C$ . Dies zeigt die eigentliche (und sogar die starke) Trennung von  $0$  und  $C$  in diesem Fall.

Andernfalls gilt  $0 \in \bar{C} \setminus \text{rel int}(C) = \text{rel } \partial(C)$ . Wegen  $0 \in \bar{C} \subseteq \text{aff}(C)$  ist  $U := \text{aff}(C)$  ein Unterraum von  $\mathbb{R}^n$ . Da  $0$  ein relativer Randpunkt von  $C$  ist, schneidet jede Kugel  $B_\varepsilon(0)$  sowohl  $U$  als auch  $\bar{C}$  als auch das Komplement  $\mathbb{R}^n \setminus \bar{C}$ . Durch Setzen von  $\varepsilon := 1/k$  können wir also eine Folge  $x^{(k)} \subseteq U$  konstruieren mit der Eigenschaft  $x^{(k)} \notin \bar{C}$  und  $x^{(k)} \rightarrow 0$ . Wir setzen

$$\hat{x}^{(k)} := \text{proj}_{\bar{C}}(x^{(k)}).$$

Diese Folge erfüllt  $\hat{x}^{(k)} \neq x^{(k)}$  und  $\hat{x}^{(k)} \rightarrow 0$ , und der [Projektionssatz 15.3](#) impliziert

$$(\hat{x}^{(k)} - x^{(k)})^\top(x - \hat{x}^{(k)}) \geq 0 \quad \text{für alle } x \in C. \quad (15.20)$$

Wir setzen

$$a^{(k)} := \frac{\hat{x}^{(k)} - x^{(k)}}{\|\hat{x}^{(k)} - x^{(k)}\|}$$

und schreiben (15.20) als

$$(a^{(k)})^\top x \geq (a^{(k)})^\top \hat{x}^{(k)} \quad \text{für alle } x \in C. \quad (15.21)$$

Wegen  $\|a^{(k)}\| = 1$  gibt es eine konvergente Teilfolge  $a^{(k(\ell))} \rightarrow a$  mit  $\|a\| = 1$  und insbesondere  $a \neq 0$ . Weiter gehört wegen  $x^{(k)} \in U$  und  $\hat{x}^{(k)} \in \bar{C} \subseteq U$  auch  $a^{(k)}$  zu  $U$ , und damit gilt auch  $a \in U$ . Der Grenzübergang auf der Teilfolge in (15.21) zeigt schließlich

$$a^\top x \geq a^\top 0 = 0 \quad \text{für alle } x \in C$$

Es bleibt zu zeigen, dass es ein  $\bar{x} \in C$  gibt, sodass  $a^\top \bar{x} > 0$  gilt. Nehmen wir an, dies sei nicht der Fall, d. h., es gelte  $a^\top x = 0$  für alle  $x \in C$ . Weil  $a \in U = \text{aff}(C)$  liegt, gibt es Punkte  $x_i \in C$  und Koeffizienten  $\alpha_i$ , 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  $\bar{x} \in C$  geben, sodass  $a^\top \bar{x} > 0$  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  $\text{rel int}(C_1) \cap \text{rel int}(C_2) = \emptyset$  gilt.

*Beweis.* Wir nehmen zunächst an, dass  $C_1$  und  $C_2$  durch  $H(a, \beta)$  eigentlich getrennt werden. Es gilt also  $a^\top x \leq \beta \leq a^\top y$  für alle  $x \in C_1$  und alle  $y \in C_2$ , und es gibt  $\bar{x} \in C_1$  und  $\bar{y} \in C_2$ , für die  $a^\top \bar{x} < a^\top \bar{y}$  gilt. Wir zeigen nun:

$$a^\top x < a^\top y \quad \text{für alle } x \in \text{rel int}(C_1) \text{ und alle } y \in \text{rel int}(C_2), \quad (15.22)$$

woraus dann  $\text{rel int}(C_1) \cap \text{rel int}(C_2) = \emptyset$  folgt. Nehmen wir an, dass (15.22) unwahr ist, dann gibt es ein  $x \in \text{rel int}(C_1)$  und ein  $y \in \text{rel int}(C_2)$  mit  $a^\top x = a^\top y$ . Für hinreichend kleines  $\varepsilon > 0$  sind  $\hat{x} := x - \varepsilon(\bar{x} - x) \in C_1$  und  $\hat{y} := y - \varepsilon(\bar{y} - y) \in C_2$ . Dann ist aber

$$a^\top(\hat{x} - \hat{y}) = a^\top(x - \varepsilon(\bar{x} - x) - y + \varepsilon(\bar{y} - y)) = \varepsilon a^\top(\bar{y} - \bar{x}) > 0$$

im Widerspruch zu  $a^\top x \leq a^\top y$  für alle  $x \in C_1$  und alle  $y \in C_2$ .

Umgekehrt nehmen wir nun an, dass  $\text{rel int}(C_1) \cap \text{rel int}(C_2) = \emptyset$  gilt. Insbesondere ist dann

$$0 \notin \text{rel int}(C_1) - \text{rel int}(C_2) = \text{rel int}(C_1 - C_2),$$

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  $a^\top x \leq 0$  gilt für alle  $x \in C$ , und es existiert ein  $\bar{x} \in C$  mit  $a^\top \bar{x} < 0$ . Das bedeutet aber  $a^\top x_1 \leq a^\top x_2$  für alle  $x_1 \in C_1$  und alle  $x_2 \in C_2$ , und für gewisse  $\bar{x}_1 \in C_1$  und  $\bar{x}_2 \in C_2$  gilt  $a^\top \bar{x}_1 < a^\top \bar{x}_2$ .  $\square$

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  $x^{(k(\ell))} \rightarrow x$  und  $y^{(k(\ell))} \rightarrow y$ , sodass auch  $z^{(k(\ell))} = x^{(k(\ell))} + y^{(k(\ell))} \rightarrow x + y$  für  $\ell \rightarrow \infty$ .  $M_1$  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.  $\square$

**Satz 15.32** (Strikter Trennungssatz, vgl. Geiger, Kanzow, 2002, Satz 2.24).

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

$$a^\top x_1 < \beta < a^\top x_2 \quad \text{für alle } x_1 \in C_1 \text{ und alle } x_2 \in C_2. \quad (15.23)$$

*Beweis.* Wir betrachten die Optimierungsaufgabe

$$\text{Minimiere } \|x_1 - x_2\|, \quad (x_1, x_2) \in C_1 \times C_2. \quad (15.24)$$

Eine Lösung  $(x_1^*, x_2^*)$  dieser Aufgabe, sofern existent, realisiert den Abstand zwischen  $C_1$  und  $C_2$ , also gilt  $\|x_1^* - x_2^*\| = \inf\{\|x_1 - x_2\| \mid x_1 \in C_1, x_2 \in C_2\}$ .

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 die konvexe, abgeschlossene Menge  $C$ . 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  $(a, \beta)$  der Hyperebene und setzen dazu

$$a := \frac{x_2^* - x_1^*}{2} \neq 0, \quad \hat{x} := \frac{x_1^* + x_2^*}{2}, \quad \beta := a^\top \hat{x}.$$

Wir zeigen nun, dass

$$x_1^* = \text{proj}_{C_1}(\hat{x}), \quad x_2^* = \text{proj}_{C_2}(\hat{x}). \quad (15.25)$$

gilt. Dazu seien  $x_1 \in C_1$  und  $x_2 \in C_2$  beliebig, dann gilt

$$\begin{aligned} \|x_1^* - \hat{x}\| + \|\hat{x} - x_2^*\| &= \|x_1^* - \hat{x} + (\hat{x} - x_2^*)\|, \quad \text{denn } x_1^* - \hat{x} = \hat{x} - x_2^* \\ &= \|x_1^* - x_2^*\| \\ &\leq \|x_1 - x_2\|, \quad \text{denn } (x_1^*, x_2^*) \text{ ist optimal für (15.24)} \\ &\leq \|x_1 - \hat{x}\| + \|\hat{x} - x_2\| \quad \text{wegen der Dreiecksungleichung.} \end{aligned}$$

Setzen wir speziell  $x_2 = x_2^*$  ein, so folgt

$$\|x_1^* - \hat{x}\| \leq \|x_1 - \hat{x}\| \quad \text{für alle } x_1 \in C_1.$$

Setzen wir dagegen  $x_1 = x_1^*$ , so folgt

$$\|x_2^* - \hat{x}\| \leq \|x_2 - \hat{x}\| \quad \text{für alle } x_2 \in C_2.$$

Dies bestätigt (15.25). Aus dem Projektionssatz 15.3 folgt daher

$$-a^\top (x_1 - x_1^*) = (x_1^* - \hat{x})^\top (x_1 - x_1^*) \geq 0 \quad \text{für alle } x_1 \in C_1.$$

Dies impliziert

$$a^\top x_1 \leq a^\top x_1^* = a^\top \hat{x} + a^\top (x_1^* - \hat{x}) = \beta - \|a\|^2 < \beta \quad \text{für alle } x_1 \in C_1.$$

Analog zeigt man

$$a^\top x_2 \geq \beta + \|a\|^2 > \beta \quad \text{für alle } x_2 \in C_2. \quad \square$$

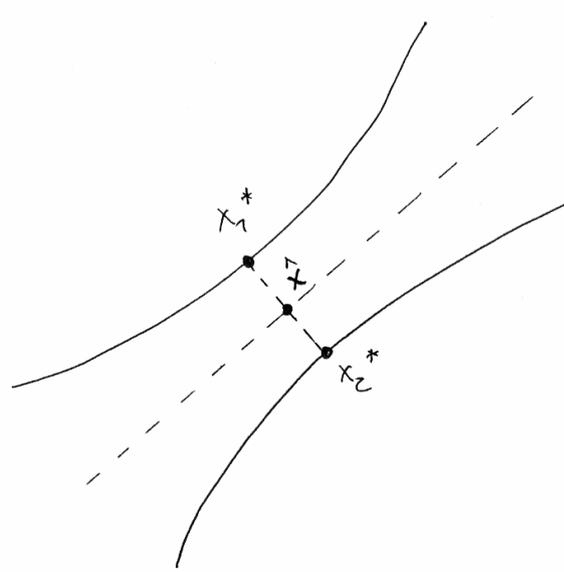


Abbildung 15.9: Illustration der Lage der Punkte  $x_1^*$ ,  $x_2^*$  und  $\hat{x}$  im Beweis von Satz 15.32.

**Quizfrage 15.20:** Wie kann man die „Lücke“  $\inf\{a^T x_2 \mid x_2 \in C_2\} - \sup\{a^T x_1 \mid x_1 \in C_1\}$  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)

$$C_1 = \{B^T \xi \mid \xi \in \mathbb{R}^m, \xi \geq 0\}$$

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  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  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:

$$f(x) \geq f(x_0) + s^\top(x - x_0) \quad \text{für alle } x \in \mathbb{R}^n. \quad (16.1)$$

Wenn  $f(x_0) \in \mathbb{R}$  liegt, dann sagt man: Die rechte Seite in (16.1) ist eine **lineare Minorante** (mit Euklidischem Gradienten  $s$ ), die  $f$  in  $x_0$  **stützt**, kurz: eine **lineare Stützfunktion**.<sup>3</sup>

- (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. Das Subdifferential an der Stelle  $x_0$  besteht gerade aus allen Gradienten linearer Funktionen, die  $f$  im Punkt  $x_0$  stützen.

**Beachte:** Wenn  $f$  eigentlich ist, also nicht  $f \equiv \infty$  gilt, dann ist  $\partial f(x_0) = \emptyset$  in allen Punkten  $x_0 \notin \text{dom } f$ .

**Quizfrage 16.1:** Warum?

Eine einfache, aber wichtige Anwendung des Subdifferentials ist folgende.

**Satz 16.2** (0 im Subdifferential).

Es sei  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  eine konvexe Funktion. Dann gilt:

- (i) Wenn  $x_0$  ein globaler Minimierer von  $f$  ist, dann ist  $0 \in \partial f(x_0)$ .
- (ii) Wenn  $f \not\equiv \infty$  ist und  $0 \in \partial f(x_0)$  ist, dann ist  $x_0$  ein globaler Minimierer von  $f$ .

**Beweis.** **Aussage (i):** Es sei  $x_0$  ein globaler Minimierer von  $f$ , insbesondere gilt also  $f(x_0) < \infty$ . Aufgrund der globalen Optimalität von  $x_0$  folgt

$$f(x) \geq f(x_0) + 0^\top(x - x_0),$$

was  $0 \in \partial f(x_0)$  bestätigt.

**Aussage (ii):** Aus  $0 \in \partial f(x_0)$  folgt sofort mit (16.1), dass  $f(x) \geq f(x_0)$  für alle  $x \in \mathbb{R}^n$  gilt. Außerdem kann nicht  $f(x_0) = \infty$  gelten, da  $f$  nach Voraussetzung nicht identisch  $\infty$  ist. (**Quizfrage 16.2:** Genaue Begründung?) Damit ist  $x_0$  ein globaler Minimierer von  $f$ , siehe **Definition 14.1**.  $\square$

**Satz 16.3** (Elementare Eigenschaften des Subdifferentials). Es sei  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  eine konvexe Funktion und  $x_0 \in \mathbb{R}^n$ . Dann ist  $\partial f(x_0)$  abgeschlossen und konvex (möglicherweise leer).

<sup>3</sup>Eine andere Anschauung:  $\begin{pmatrix} s \\ -1 \end{pmatrix}$  ist Normalenvektor einer Hyperebene durch den Punkt  $\begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix}$ , die den (Epi-)Graphen in diesem Punkt stützt.

Beweis. Der Beweis ist Inhalt von [Hausaufgabe 11.5](#). □

**Bemerkung 16.4** (Zum Begriff „Subgradient“). Die Bezeichnung **Subgradient** ist eigentlich irreführend, weil das Subdifferential den Begriff der 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 16.3:** Wie würde die Definition eines Subgradienten von  $f$  im Punkt  $x_0$  bzgl. des  $M$ -Skalarprodukts aussehen?

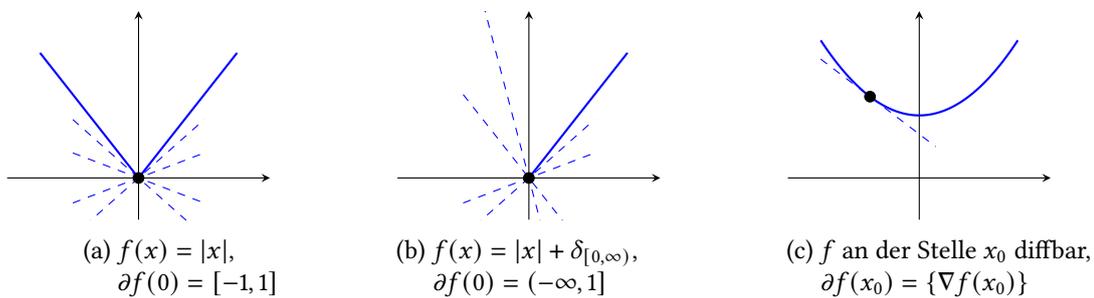


Abbildung 16.1: Das Subdifferential  $\partial f(x_0)$  besteht aus den Steigungen aller linearen Minoranten an  $f$  im Punkt  $x_0$ .

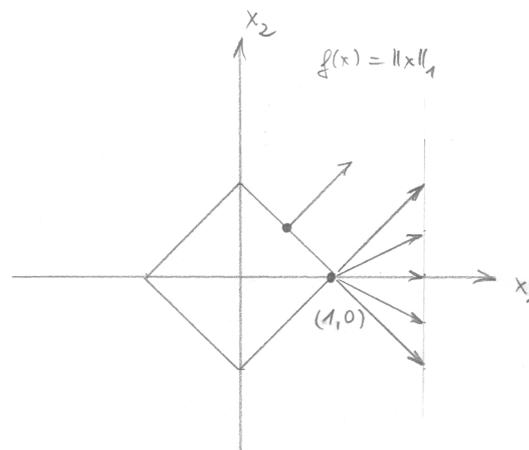


Abbildung 16.2: Das Subdifferential der 1-Norm  $\|x\|_1 = |x_1| + |x_2|$  ([Beispiel 16.5](#)) in zwei Punkten derselben Levelmenge.

**Beispiel 16.5** (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  $x \neq 0$

$$\partial f(x) = \left\{ s \in \mathbb{R}^n \mid \begin{array}{l} \|s\|_1 = 1, \quad s_i \geq 0 \text{ für diejenigen } i \text{ mit } x_i = \|x\|_\infty, \\ s_i \leq 0 \text{ für diejenigen } i \text{ mit } -x_i = \|x\|_\infty, \\ \text{und } s_i = 0 \text{ für diejenigen } i \text{ mit } |x_i| < \|x\|_\infty \end{array} \right\}$$

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  $\text{dom } f$  gehören, leer sein kann.

**Beispiel 16.6** (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  $\text{dom } f$  sind, in denen das Subdifferential leer sein kann. In der Tat gilt folgender Satz:

**Satz 16.7** (Wann ist das Subdifferential leer bzw. nichtleer? vgl. Rockafellar, 1970, Theorem 23.4).

Es sei  $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  konvex.

- (i) Für alle  $x_0 \notin \text{dom } f$  ist  $\partial f(x_0) = \emptyset$ .
- (ii) Für alle  $x_0 \in \text{rel int}(\text{dom } f)$  ist  $\partial f(x_0) \neq \emptyset$ .

*Beweis.* **Aussage (i):** Es sei  $x_0 \notin \text{dom } f$ , also  $f(x_0) = \infty$ . Die Subgradientenungleichung (16.1), also

$$f(x) \geq \underbrace{f(x_0)}_{=\infty} + s^\top(x - x_0)$$

kann mit jedem  $x \in \text{dom } f \neq \emptyset$  für kein  $s \in \mathbb{R}^n$  erfüllt sein.

**Aussage (ii):** Es sei  $x_0 \in \text{rel int}(\text{dom } f) \subseteq \text{dom } f$ . Falls  $f(x_0) = -\infty$  gilt, so ist offensichtlich  $\partial f(x_0) = \mathbb{R}^n$  und die Aussage bewiesen. Wir gehen also nun von  $f(x_0) \in \mathbb{R}$  aus und betrachten den Epigraphen

$$\text{epi } f = \left\{ \begin{pmatrix} x \\ \gamma \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \geq f(x) \right\}.$$

Nach Satz 13.15 ist  $\text{epi } f \subseteq \mathbb{R}^n \times \mathbb{R}$  konvex. Wir wenden den **eigentlichen Trennungssatz 15.30** an, um die Mengen  $C_1 = \left\{ \begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix} \right\}$  und  $C_2 = \text{epi } f$  eigentlich zu trennen. Beachte, dass  $\text{rel int}(C_1) \cap \text{rel int}(C_2) = \emptyset$  ist, da der Punkt  $\begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix}$  zu  $\text{epi } f$ , aber nicht zu  $\text{rel int}(\text{epi } f)$  gehört (**Quizfrage 16.4:** 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  $\begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix}$  und  $\text{epi } f$  eigentlich trennt, demnach gilt

$$\begin{pmatrix} s \\ \sigma \end{pmatrix}^\top \begin{pmatrix} x \\ \gamma \end{pmatrix} \leq \begin{pmatrix} s \\ \sigma \end{pmatrix}^\top \begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix} \quad \text{für alle } \begin{pmatrix} x \\ \gamma \end{pmatrix} \in \text{epi } f, \quad (16.2)$$

und es existiert ein Punkt  $\begin{pmatrix} \bar{x} \\ \bar{\gamma} \end{pmatrix} \in \text{epi } f$  mit der Eigenschaft

$$\begin{pmatrix} s \\ \sigma \end{pmatrix}^\top \begin{pmatrix} \bar{x} \\ \bar{\gamma} \end{pmatrix} < \begin{pmatrix} s \\ \sigma \end{pmatrix}^\top \begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix}. \quad (16.3)$$

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  $\gamma = 1$ :

$$s^\top x_0 + \sigma \leq s^\top x_0 + 0.$$

- (ii) Ist  $f(x_0) > 0$ , so folgt aus (16.2) mit der Wahl  $x = x_0$  und  $\gamma = 2f(x_0)$ :

$$s^\top x_0 + 2\sigma f(x_0) \leq s^\top x_0 + \sigma f(x_0).$$

- (iii) Ist  $f(x_0) < 0$ , so folgt aus (16.2) mit der Wahl  $x = x_0$  und  $\gamma = 0$ :

$$s^\top x_0 + 0 \leq s^\top x_0 + \sigma f(x_0).$$

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)  $s^\top(\bar{x} - x_0) < 0$ . Da  $x_0 \in \text{rel int}(\text{dom } f)$  war, existiert nach Lemma 15.17 ein  $\varepsilon > 0$ , sodass auch  $\tilde{x} := x_0 - \varepsilon(\bar{x} - x_0)$  noch zu  $\text{dom } f$  gehört. Durch Einsetzen von  $\begin{pmatrix} \tilde{x} \\ f(\tilde{x}) \end{pmatrix} \in \text{epi } f$  in (16.2) erhalten wir den Widerspruch

$$s^\top x_0 - \varepsilon \underbrace{s^\top(\bar{x} - x_0)}_{<0} \leq s^\top x_0.$$

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

$$s^\top x - f(x) \leq s^\top x_0 - f(x_0) \quad \text{für alle } x \in \text{dom } f.$$

Da dieselbe Ungleichung trivialerweise auch für  $x \notin \text{dom } f$  gilt, erhalten wir schließlich

$$f(x) \geq f(x_0) + s^\top(x - x_0) \quad \text{für alle } x \in \mathbb{R}^n,$$

d. h.,  $s \in \partial f(x_0)$ . □

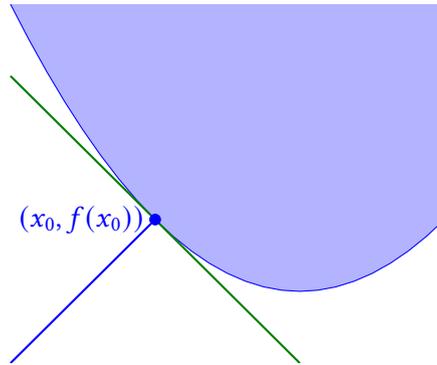


Abbildung 16.3: Konstruktion eines Subgradienten aus dem Beweis von Satz 16.7 mit Hilfe des Normalenvektors einer Hyperebene, die den Punkt  $(x_0, f(x_0))$  eigentlich von  $\text{epi } f$  trennt.

Ein wichtiger Satz über das Subdifferential ist die folgende Summenregel, deren Beweis wiederum den eigentlichen Trennungssatz 15.30 verwendet.

**Satz 16.8** (Summenregel für das Subdifferential, vgl. Rockafellar, 1970, Theorem 23.8).

Es seien  $f_1, f_2: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  zwei konvexe Funktionen.

(i) Es gilt

$$\partial f_1(x_0) + \partial f_2(x_0) \subseteq \partial(f_1 + f_2)(x_0) \quad \text{für alle } x_0 \in \mathbb{R}^n. \tag{16.4}$$

(ii) Falls

$$(\text{rel int dom } f_1) \cap (\text{rel int dom } f_2) \neq \emptyset \quad (16.5)$$

erfüllt ist, dann gilt sogar

$$\partial f_1(x_0) + \partial f_2(x_0) = \partial(f_1 + f_2)(x_0) \quad \text{für alle } x_0 \in \mathbb{R}^n. \quad (16.6)$$

*Beweis.* **Aussage (i):** 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 (16.1) liefert

$$\begin{aligned} f_1(x) &\geq f_1(x_0) + s_1^\top(x - x_0), \\ f_2(x) &\geq f_2(x_0) + s_2^\top(x - x_0) \end{aligned}$$

für alle  $x \in \mathbb{R}^n$ . Die Addition der Ungleichungen ergibt (**Quizfrage 16.5:** Warum stimmt das auch dann, wenn hier nicht-endliche Funktionswerte vorkommen?)

$$(f_1 + f_2)(x) \geq (f_1 + f_2)(x_0) + (s_1 + s_2)^\top(x - x_0),$$

für alle  $x \in \mathbb{R}^n$ , d. h., wir haben  $s = s_1 + s_2 \in \partial(f_1 + f_2)(x_0)$ .

**Aussage (ii):** Wir nehmen jetzt an, dass  $(\text{rel int dom } f_1) \cap (\text{rel int dom } f_2) \neq \emptyset$  gilt 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) auch  $\partial f_1(x_0) = \partial f_2(x_0) = \emptyset$  und die Behauptung gezeigt.

Es sei also  $s \in \partial(f_1 + f_2)(x_0)$ . Aus **Satz 16.7 (i)** folgt  $x_0 \in \text{dom}(f_1 + f_2)$ , also sind  $f_1(x_0)$  und  $f_2(x_0)$  beide endlich. 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

$$\begin{aligned} g_1(x) &:= f_1(x + x_0) - f_1(x_0) - s^\top x \\ g_2(x) &:= f_2(x + x_0) - f_2(x_0). \end{aligned}$$

Diese Funktionen sind zunächst wohldefiniert, denn der Fall  $\infty - \infty$  tritt nicht auf, da  $f_1(x_0)$  und  $f_2(x_0)$  beide endlich sind. Die Funktionen  $g_1$  und  $g_2$  erfüllen außerdem noch immer die Voraussetzung  $(\text{rel int dom } g_1) \cap (\text{rel int dom } g_2) \neq \emptyset$ , und die Voraussetzung  $s \in \partial(f_1 + f_2)(x_0)$  wird zu  $0 \in \partial(g_1 + g_2)(\tilde{x}_0)$  mit  $\tilde{x}_0 := 0$ . (**Quizfrage 16.6:** Details?) Wir müssen jetzt also zeigen: Es gibt ein  $\tilde{s} \in \partial g_1(0)$ , sodass  $-\tilde{s} \in \partial g_2(0)$  ist. Dies erfolgt unter Anwendung des **eigentlichen Trennungssatzes 15.30**.

Wir betrachten dazu die konvexen Mengen<sup>4</sup>

$$\begin{aligned} C_1 &:= \left\{ \begin{pmatrix} x \\ \gamma \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \geq g_1(x) \right\} = \text{epi } g_1, \\ C_2 &:= \left\{ \begin{pmatrix} x \\ \gamma \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \leq -g_2(x) \right\} = \text{hypo } -g_2. \end{aligned}$$

<sup>4</sup>Der **Hypograph** einer Funktion  $f := \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  ist die Menge

$$\text{hypo } f := \left\{ \begin{pmatrix} x \\ \gamma \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \leq f(x) \right\}.$$

Der Hypograph von  $f$  ist konvex genau dann, wenn  $f$  konkav ist.

**(Quizfrage 16.7:** Warum sind beide dieser Mengen nichtleer?) **Lemma 15.18** zeigt

$$\begin{aligned} \text{rel int } C_1 &= \left\{ \begin{pmatrix} x \\ \gamma \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \mid x \in \text{rel int dom } g_1, \gamma > g_1(x) \right\}, \\ \text{rel int } C_2 &= \left\{ \begin{pmatrix} x \\ \gamma \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \mid x \in \text{rel int dom } g_2, \gamma < -g_2(x) \right\}. \end{aligned}$$

Wegen  $0 \in \partial(g_1 + g_2)(\tilde{x}_0)$  ist  $\tilde{x}_0 = 0$  nach **Satz 16.2 (ii)** ein globaler Minimierer für  $g_1 + g_2$  mit Funktionswert  $(g_1 + g_2)(0) = 0$ . Daraus folgt, dass  $\text{rel int } C_1 \cap \text{rel int } C_2 = \emptyset$  ist. **(Quizfrage 16.8:** Details?)

Nach dem **eigentlichen Trennungssatz 15.30** lassen sich nun  $C_1$  und  $C_2$  durch eine Hyperebene  $H(a, \beta)$  eigentlich trennen. Wir schreiben den Normalenvektor als  $a = \begin{pmatrix} \tilde{s} \\ \sigma \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}$ . Es gilt also

$$\begin{pmatrix} \tilde{s} \\ \sigma \end{pmatrix}^\top \begin{pmatrix} x_1 \\ \gamma_1 \end{pmatrix} \leq \begin{pmatrix} \tilde{s} \\ \sigma \end{pmatrix}^\top \begin{pmatrix} x_2 \\ \gamma_2 \end{pmatrix} \quad \text{für alle } \begin{pmatrix} x_1 \\ \gamma_1 \end{pmatrix} \in C_1 \text{ und } \begin{pmatrix} x_2 \\ \gamma_2 \end{pmatrix} \in C_2, \tag{16.7}$$

und es existieren  $\begin{pmatrix} \bar{x}_1 \\ \bar{\gamma}_1 \end{pmatrix} \in C_1$  und  $\begin{pmatrix} \bar{x}_2 \\ \bar{\gamma}_2 \end{pmatrix} \in C_2$ , für die die Ungleichung strikt ist. Die Komponente  $\sigma$  kann nicht gleich Null sein, denn sonst hätten wir  $\text{dom } g_1$  und  $\text{dom } g_2$  durch eine Hyperebene mit Normalenvektor  $\tilde{s}$  eigentlich getrennt, was aufgrund der Voraussetzung  $(\text{rel int dom } g_1) \cap (\text{rel int dom } g_2) \neq \emptyset$  und **Satz 15.30** aber nicht möglich ist.

Wegen  $(g_1 + g_2)(0) = 0$  liegt  $\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in C_1 \cap C_2$ . 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  $\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in C_1 \cap C_2$ , dass die eigentlich trennende Hyperebene  $H(a, \beta)$  den Offset  $\beta = 0$  hat. Es folgt nun aus (16.7)

$$\begin{pmatrix} \tilde{s} \\ -1 \end{pmatrix}^\top \begin{pmatrix} x_1 \\ \gamma_1 \end{pmatrix} \leq 0 \leq \begin{pmatrix} \tilde{s} \\ -1 \end{pmatrix}^\top \begin{pmatrix} x_2 \\ \gamma_2 \end{pmatrix} \quad \text{für alle } \begin{pmatrix} x_1 \\ \gamma_1 \end{pmatrix} \in C_1 \text{ und } \begin{pmatrix} x_2 \\ \gamma_2 \end{pmatrix} \in C_2,$$

also

$$\begin{aligned} \tilde{s}^\top x &\leq \gamma \quad \text{für alle } \begin{pmatrix} x \\ \gamma \end{pmatrix} \in C_1, \\ \tilde{s}^\top x &\geq \gamma \quad \text{für alle } \begin{pmatrix} x \\ \gamma \end{pmatrix} \in C_2. \end{aligned}$$

Wegen  $g_1(0) = g_2(0) = 0$  und der Definition von  $C_1$  und  $C_2$  heißt das aber

$$\begin{aligned} g_1(x) &\geq g_1(0) + \tilde{s}^\top(x - 0) \quad \text{für alle } x \in \mathbb{R}^n, \\ g_2(x) &\geq g_2(0) + (-\tilde{s})^\top(x - 0) \quad \text{für alle } x \in \mathbb{R}^n. \end{aligned}$$

Also folgt schließlich  $\tilde{s} \in \partial g_1(\tilde{x}_0)$  und  $-\tilde{s} \in \partial g_2(\tilde{x}_0)$ , was zu zeigen war. □

Wie das folgende Beispiel zeigt, ist die Gleichheit der Subdifferenziale (16.6) ohne eine Regularitätsbedingung wie  $(\text{rel int dom } f_1) \cap (\text{rel int dom } f_2) \neq \emptyset$  im Allgemeinen falsch.

**Beispiel 16.9 (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

$$\begin{aligned} C_1 &:= \{(x, \gamma) \mid \gamma \geq x^2\} = \text{epi}(x \mapsto x^2) \\ C_2 &:= \{(x_1, x_2) \mid x_2 = 0\}. \end{aligned}$$

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  $\text{dom } f_1 \cap \text{dom } f_2 = C_1 \cap C_2 = \{(0, 0)\}$ , 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 16.10** (Kettenregel für das Subdifferential, vgl. [Rockafellar, 1970](#), Theorem 23.9<sup>5</sup>).

Es sei  $f: \mathbb{R}^m \rightarrow \mathbb{R} \cup \{\infty\}$  konvex und  $A \in \mathbb{R}^{m \times n}$ . Weiter sei  $g: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  definiert durch  $g(x) := f(Ax)$ .

(i) Es gilt:

$$A^\top \partial f(Ax_0) \subseteq \partial g(x_0) \quad \text{für alle } x_0 \in \mathbb{R}^n. \quad (16.8)$$

(ii) Falls

$$(\text{Bild } A) \cap (\text{rel int dom } f) \neq \emptyset \quad (16.9)$$

erfüllt ist, dann gilt sogar

$$A^\top \partial f(Ax_0) = \partial g(x_0) \quad \text{für alle } x_0 \in \mathbb{R}^n. \quad (16.10)$$

Weitere Eigenschaften des Subdifferentials folgen in ??.

Ende der Woche 11

<sup>5</sup>In [Rockafellar, 1970](#), Theorem 23.9 wird  $f$  als eigentlich vorausgesetzt. Der Fall  $f \equiv \infty$ , den unsere Formulierung des Satzes mit abdeckt, ist aber klar, da dann auch  $g(x) \equiv \infty$  ist und (16.8) dann  $\mathbb{R}^n \subseteq \mathbb{R}^n$  lautet.

# Index

- $\mu$ -stark konvexe Funktion, 101, 104
- abgeschlossene  $\varepsilon$ -Kugel, 11
- abgeschlossene  $\varepsilon$ -Umgebung, 11
- abhängige Variable, 53
- Ableitung, 11
- Abschluss einer Menge, 11
- Abstiegsrichtung, 17
- Abstiegsverfahren, 17
- Accessibility lemma, 124
- affin unabhängig, 118
- affine Basis, 118
- affine Dimension einer Menge, 120
- affine Hülle, 120
- affiner Unterraum, 116
- Affinkombination, 118
- aktive Ungleichung, 5
- algebraisch innerer Punkt, 128
- algebraisches Inneres, 128
- Anfangsknoten, 86
- Angebotsknoten, 88
- Armijo-Bedingung, 18
- Armijo-Parameter, 18
- Aufwandsmatrix, 44
  
- Backtracking-Parameter, 19
- Backtracking-Strategie, 19
- Basis, 53
- Basislösung, 53
- Basismatrix, 53
- Basisvektor, 53
  - benachbart, 56
- Bedarfsknoten, 88
- Bedarfsmatrix, 44
- beidseitige Richtungsableitung, 11
- Box-Beschränkungen, 7
  
- $C^1$ -Funktion, 12
- $C^2$ -Funktion, 12
- CG-Verfahren, 31
  
- differenzierbare Funktion, 11
- Digraph, 86
- Dimension einer Menge, 120
- Dimension eines affinen Unterraums, 117
- diskrete Optimierung, 5
- dual zulässige Basis, 74
- duale Schlupfvariablen, 67
- duales LP, 67
- duales Simplex-Verfahren, 73, 74
- Dualitätslücke, 73
- Durchflussknoten, 88
  
- echte Konvexkombination, 99
- Ecke, 51
- eigentlich trennende Hyperebene, 129
- eigentliche Funktion, 104
- eigentlicher Definitionsbereich, 104
- einfacher Digraph, 86
- einseitige Richtungsableitung, 11
- Endknoten, 86
- entarteter Basisvektor, 60
- Epigraph, 105
- Erhaltungsbedingung, 88
- erweitert reellwertige Funktion, 102
- exakte Liniensuche, 18
- Extremalpunkt, 51
  
- Farkas-Lemma, 69, 136
- Fluss, 88
- Flusserhaltungsgleichungen, 88
- Flussnetzwerk, 88
- Flussvektor, 88
- freie Optimierungsaufgabe, 7
- freie Variable, 45
  
- ganzzahlige lineare Optimierungsaufgabe, 78
- ganzzahliges lineares Programm, 78
- Gaußklammer
  - obere, 79
  - untere, 79

- gerichtete Kante, 86  
gerichteter Graph, 86  
gleichungsbeschränkte Optimierungsaufgabe, 7  
Gleichungsnebenbedingung, 5  
global optimale Lösung, 6, 112  
globale Minimalstelle, 6, 112  
globaler Minimalwert, 6, 112  
globaler Minimierer, 6, 112  
globales Minimum, 6, 112  
Globalisiertes Newton-Verfahren, 39  
Gradient, 11  
Gradientenverfahren, 18  
Grundmenge, 5
- Halbraum, 45  
Hessematrix, 12  
Hyperebene, 45  
Hypograph, 142
- inaktive Indizes, 52  
inaktive Ungleichung, 5  
Indikatorfunktion, 103  
Inneres einer Menge, 11  
Inzidenzmatrix, 87
- Jacobimatrix, 12
- kanonische Form, 44  
Kantenkapazität, 88  
Kantenkostenvektor, 88  
Kantorovich-Ungleichung, 27  
Kapazitätsbeschränkungen, 88  
Kegel, 48  
Knoten, 86  
Knoten-Kanten-Inzidenzmatrix, 87  
Knotenbilanz, 88  
Knotenbilanzen, 88  
konische Hülle, 50  
konkave Funktion, 101  
kontinuierliche Optimierung, 5  
konvexe Funktion, 101, 104  
konvexe Hülle, 100  
konvexe Menge, 97  
konvexe Optimierungsaufgabe, 8, 112  
Konvexkombination, 99  
kostenminimaler Fluss, 88  
kostenminimaler Transport, 88
- Kostenvektor, 42, 44
- Laplacematrix, 87  
lineare Minorante, 137  
lineare Optimierungsaufgabe, 7  
lineare Stützfunktion, 137  
lineares Modell, 32  
lineares Programm, 7, 42  
Liniensuche, 18  
Liniensuchfunktion, 19  
Liniensuchverfahren, 41  
lokal optimale Lösung, 6, 113  
lokale Minimalstelle, 6, 113  
lokaler Minimalwert, 6, 113  
lokaler Minimierer, 6, 113  
lokales Minimum, 6, 113  
Lokales Newton-Verfahren, 33  
LP, *siehe* lineares Programm  
lösbare Optimierungsaufgabe, 6
- Matrixnorm, 33  
Mehrgüterflussprobleme, 95  
Mehrgütertransportprobleme, 95  
MILP, *siehe* ganzzahliges lineares Programm  
Minkowski-Summe, 98  
Mittelwertsatz, 13  
monotoner Operator, 108
- negativer Halbraum, 45  
Netzwerk-Simplex-Verfahren, 90  
Newton-Richtung, 33, 38  
Nichtbasis, 53  
Nichtbasismatrix, 53  
nichtlineare Optimierungsaufgabe, 8  
nichtlineares Programm, 8  
NLP, *siehe* nichtlineares Programm  
Normalenvektor, 45  
Normalform, 46
- obere Schranke, 7  
offene  $\varepsilon$ -Kugel, 11  
offene  $\varepsilon$ -Umgebung, 11  
Optimalwert, 5, 112  
Optimierungsvariable, 5  
orthogonale Projektion, 114, 115
- partielle Ableitung, 11  
Phase-I-Problem, 64

- Phase-II-Problem, 66
- Polyeder, 45
- Polyeder in Normalform, 47
- positiver Halbraum, 45
- pricing im dualen Simplex-Verfahren, 75
- pricing im primalen Simplex-Verfahren, 58
- primal zulässige Basis, 74
- primal-duales Paar, 67
- primales LP, 67
- primales Simplex-Verfahren, 73
- Q-lineare Konvergenz, 35
- Q-quadratische Konvergenz, 35
- Q-superlineare Konvergenz, 35
- QP, *siehe* quadratisches Programm
- quadratische Optimierungsaufgabe, 8
- quadratisches Ersatzmodell, 38
- quadratisches Programm, 8
- quadratisches Wachstum, 16
- Quelle in einem Transportnetzwerk, 88
- Quotiententest im dualen Simplex-Verfahren, 75
- Quotiententest im primalen Simplex-Verfahren, 58
- reduzierte Kosten im primalen Simplex-Verfahren, 57
- relativ innerer Punkt, 122
- relativer Rand, 122
- relativer Randpunkt, 122
- relatives Inneres, 122
- Residuum, 25, 32
- Ressourcenvektor, 44
- Rezessionskegel, 48
- Richtung des steilsten Abstiegs, 17
- Richtung des steilsten Abstiegs im  $M$ -Skalarprodukt, 24
- Richtungsraum, 116
- Schattenpreis, 85
- Schleife, 86
- Schlupfvariable, 46
- Schnitt durch eine Funktion, 19
- schwache Dualität, 67
- Senke in einem Transportnetzwerk, 88
- Simplex-Schritt, 58
- Spektralnrm, 33
- stark konkave Funktion, 101
- stark konvexe Funktion, 101, 104
- stark monotoner Operator, 108
- starke Dualität, 73
- stationärer Punkt, 14
- strikt konkave Funktion, 101
- strikt konvexe Funktion, 101, 104
- strikt monotoner Operator, 108
- strikt trennende Hyperebene, 129
- strikter globaler Minimierer, 6
- strikter lokaler Minimierer, 6
- Subableitung, 138
- Subdifferential, 137
- subdifferenzierbare Funktion, 137
- Subgradient, 137
- Subgradientenungleichung, 137
- Sublevelmenge, 8
- Suchrichtung, 19
- Teilfolge, 10
- total unimodulare Matrix, 92
- Transportnetzwerk, 88
- trennende Hyperebene, 69, 129
- Trust-Region-Verfahren, 41
- Überschussvariable, 46
- Umladeknoten, 88
- unabhängige Variable, 53
- unbeschränkte Optimierungsaufgabe, 5
- ungleichungsbeschränkte Optimierungsaufgabe, 7
- Ungleichungsnebenbedingung, 5
- unimodulare Matrix, 92
- unlösbare Optimierungsaufgabe, 6
- unrestringierte Optimierungsaufgabe, 7
- untere Schranke, 7
- unterhalbstetige Funktion, 8
- Untermatrix, 92
- unzulässige Optimierungsaufgabe, 5
- Variationsungleichung, 115
- verallgemeinerte Konditionszahl, 30
- verallgemeinertes Eigenwertproblem, 29
- vereinfachtes Newton-Verfahren, 37
- Verfahren der konjugierten Gradienten, 31
- Verfahren des steilsten Abstiegs, 18
- verletzte Ungleichung, 5
- von unten halbsetige Funktion, 8
- vorkonditioniertes Gradientenverfahren, 24

Vorkonditionierung, [24](#)

Zeilensummennorm, [123](#)

Zielfunktion, [5](#)

zulässige Menge, [5](#)

zulässiger Basisvektor, [53](#)

zulässiger Fluss, [88](#)

zulässiger Punkt, [5](#)

zweimal differenzierbare Funktion, [12](#)

# Literatur

- Alpargu, G. (1996). „The Kantorovich Inequality, with Some Extensions and with Some Statistical Applications“. Magisterarb. Department of Mathematics und Statistics, McGill University, Montreal, Canada.
- Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley & Sons, Inc., New York-London-Sydney. DOI: [10.1002/9781118186428](https://doi.org/10.1002/9781118186428).
- 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](https://doi.org/10.2140/pjm.1966.16.1).
- Cartan, H. (1967). *Calcul Différentiel*. Paris: Hermann, S. 178.
- 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.
- 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](https://doi.org/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](https://doi.org/10.1007/978-3-642-56004-0).
- Gerdts, M.; F. Lempio (2011). *Mathematische Optimierungsverfahren des Operations Research*. de Gruyter. DOI: [10.1515/9783110249989](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1007/978-3-322-96828-9).
- Jarre, F.; J. Stoer (2004). *Optimierung*. Springer. DOI: [10.1007/978-3-642-18785-8](https://doi.org/10.1007/978-3-642-18785-8).
- Klee, V.; G. J. Minty (1972). „How good is the simplex algorithm?“ *Inequalities III: Proceedings of the Third Symposium on Inequalities held at the University of California, Los Angeles, September 1–9, 1969*. Hrsg. von O. Shisha. Academic Press, New York, S. 159–175.
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/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.
- Vanderbei, R. J. (2008). *Linear Programming: Foundations and Extensions*. Operations Research, Management Science. New York, NY: Springer. DOI: [10.1007/978-0-387-74388-2](https://doi.org/10.1007/978-0-387-74388-2).
- Werner, J. (2007). *Vorlesung über Optimierung*. Lecture Notes, Department of Mathematics, University of Hamburg, Germany. URL: <http://num.math.uni-goettingen.de/werner/>.