Vorlesungsskript
Grundlagen der Optimierung

Wintersemester 2024

Roland Herzog*

2024-12-18

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

Material für 27 Vorlesungen.

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

Expertenwissen: Thema

Ein solcher Block kennzeichnet weiterführende Informationen. Diese sind nicht prüfungsrelevant.

Inhaltsverzeichnis

Kapitel 0 Einführung

§ 1 Grundbegriffe und Klassifikation von Optimierungsaufgaben

Die mathematische Optimierung beschäftigt sich mit Aufgaben der Form

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

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

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

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

Definition 1.1 (Grundbegriffe)**.**

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

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

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

(*ii*) Die Ungleichung $g_i(x) \leqslant 0$ heißt an der Stelle x **aktiv** (englisch: *active*), wenn $g_i(x) = 0$ gilt. Sie heißt **inaktiv** (englisch: *inactive*), wenn $g_i(x) < 0$ ist. Sie heißt **verletzt** (englisch: *violated*), wenn $g_i(x) > 0$ ist.

(*iii*) Der Wert

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

heißt der **Infimalwert** (englisch: *infimal 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** (englisch: *global minimizer, globally optimal solution*), wenn gilt:

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

Äquivalent dazu ist: $f(x^*) = f^*$. In diesem Fall heißt die Zahl f^* dann auch der **Optimalwert**, das **globale Minimum**, **globale Minimalwert** (englisch: *optimal value, global minimum*) von (1.1).

(*vi*) Ein globaler Minimierer x^* heißt **strikt** (englisch: *strict global minimizer, strict globally optimal solution*), 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** (englisch: *local minimizer, locally optimal solution*), wenn es eine Umgebung $U(x^*)$ gibt, sodass gilt:

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

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** (englisch: *strict local minimizer, strict locally optimal solution*), 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** (englisch: *solvable*), wenn sie mindestens einen globalen Minimierer besitzt, also einen zulässigen Punkt, an dem der Infimalwert angenommen wird. Ansonsten heißt die Aufgabe **unlösbar**. △

Die Abbildung 1.1 illustriert die Begriffe aus Definition 1.1 anhand einer Zielfunktion $f\colon \mathbb{R} \to \mathbb{R}$.
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 Minimierer?

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

Quizfrage 1.6: Was gilt an Punkten, die gleichzeitig lokaler Minimierer und lokaler Maximierer sind?

Beachte: Es kann drei Gründe geben, aufgrund derer eine Optimierungsaufgabe (1.1) unlösbar ist:

(1) Die Aufgabe ist unzulässig, also $f^* = \infty$.

(2) Die Aufgabe ist unbeschränkt, also $f^* = -\infty$.

(3) Der Infimalwert f^* ist endlich, wird aber über der zulässigen Menge nicht als Funktionswert angenommen.

Beispiel 1.2 (Unlösbare Optimierungsaufgaben)**.**

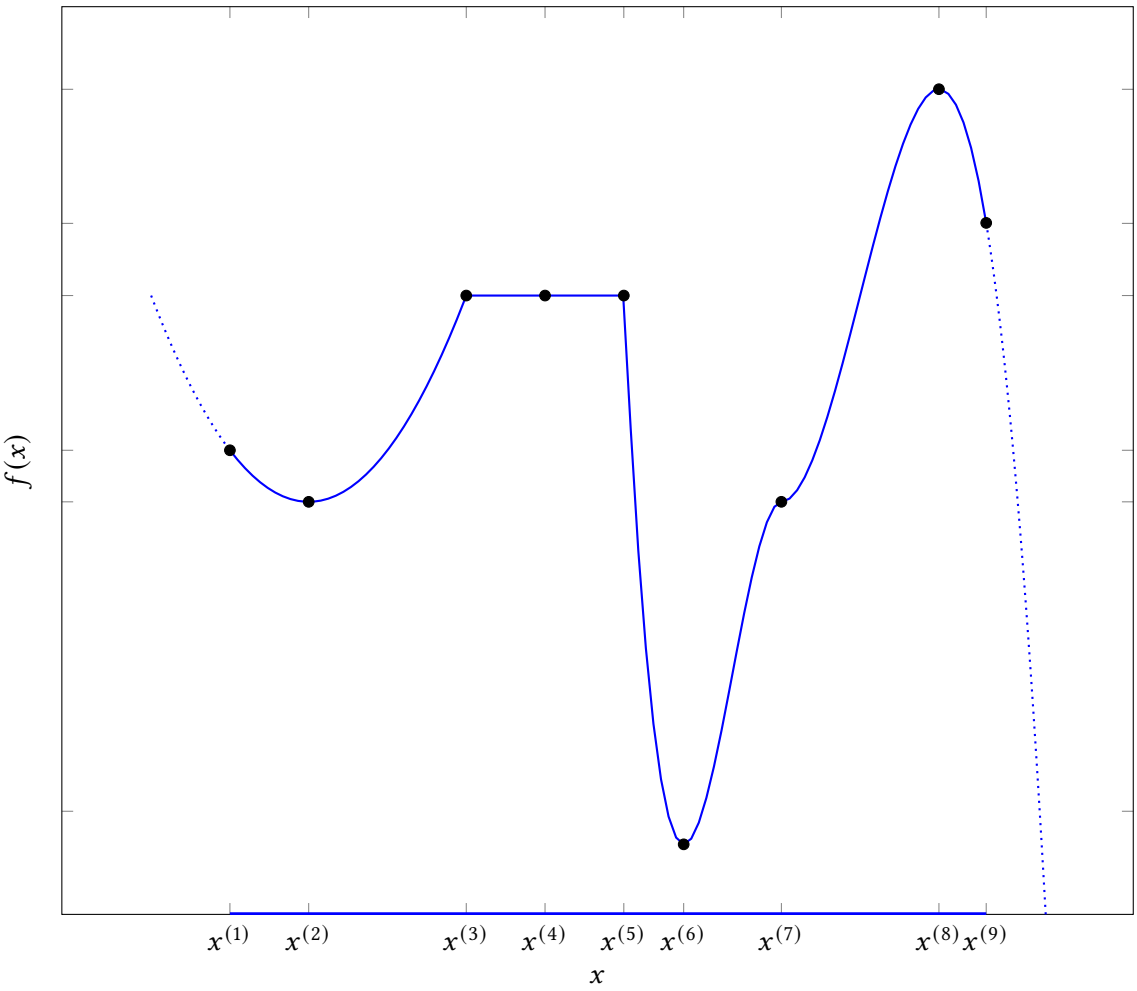


Abbildung 1.1.: Illustration der Begriffe aus Definition 1.1 anhand einer Zielfunktion $f\colon \mathbb{R} \to \mathbb{R}$. Die zulässige Menge ist das auf der x-Achse markierte Intervall.

(i) Die Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & x \quad \text{über } x \in \mathbb{R} \\ \text{sodass} \quad & x \geqslant 2 \\ \text{und} \quad & x \leqslant 1 \end{aligned}$$

ist unzulässig, also $F = \emptyset$ und $f^* = \infty$.

(ii) Die Aufgabe

$$\text{Minimiere} \quad x \quad \text{über } x \in \mathbb{R}$$

ist unbeschränkt, also $f^* = -\infty$.

(iii) Bei der Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & 1/x \quad \text{über } x \in \mathbb{R} \\ \text{sodass} \quad & x \geqslant 1 \end{aligned}$$

ist der Infimalwert $f^* = 0$, also endlich, wird aber über der zulässigen Menge $F = [1, \infty)$ nicht von der Zielfunktion als Funktionswert angenommen. Mit anderen Worten, der Infimalwert

$\inf\{f(x) \mid x \in F\}$ ist kein Minimum. △

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 optimale Lösungen?

(2) Wie erkennt man sie? ($\rightsquigarrow$ 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 *Nonlinear Optimization* behandelt. Später schließen sich Veranstaltungen beispielsweise zu *Infinite-Dimensional Optimization*, unter anderem mit Aufgaben der Optimierung mit partiellen Differentialgleichungen als Nebenbedingungen, an. Optimierungsaufgaben in nichtlinearen Räumen behandelt die Lehrveranstaltung *Optimization on Manifolds*.

Um die Modellierung von Aufgaben der Form (1.1) einzuüben, geben wir noch drei Beispiele an.

Beispiel 1.3 (Angebotsauswertung)**.**

Ein Unternehmen will eine bestimmte Menge M eines Gutes einkaufen und holt dazu Angebote von n Lieferfirmen ein, von denen keine die gewünschte Gesamtmenge alleine liefern kann. Anbieter i liefert maximal die Menge m_i, wobei die Funktion $f_i(x_i)$ den Gesamtpreis in Abhängigkeit der Bestellmenge x_i angibt. Die Funktionen f_i werden i. d. R. monoton wachsend sein und evtl. nichtlinear (konkav).

Die optimalen Einkaufsmengen x_i und die minimalen Beschaffungskosten (Optimalwert) ergeben sich durch Lösung der folgenden Optimierungsaufgabe:

$$\begin{aligned} \text{Minimiere} \quad & f(x) := \sum_{i=1}^{n} f_i(x_i) \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} \quad & \sum_{i=1}^{n} x_i = M \\ \text{und} \quad & 0 \leqslant x_i \leqslant m_i, \quad i = 1, \dots, n. \end{aligned}$$

Quizfrage 1.7: Könnten wir hier auch $\sum_{i=1}^{n} x_i \geqslant M$ zulassen? △

Beispiel 1.4 (Transportproblem)**.**

Eine Firma besitzt zwei Fabriken F_1, F_2 und zwölf Verkaufsstellen $V_1, \dots, V_{12}$. Jede Fabrik F_i produziert pro Woche die Menge p_i eines bestimmten Produktes. Jede Verkaufsstelle V_j hat eine bekannte wöchentliche Nachfrage n_j an diesem Produkt, die gedeckt werden muss. Die Kosten, eine Einheit von Fabrik F_i zur Verkaufsstelle V_j zu transportieren, seien c_{ij}, und die zu transportierende Menge sei x_{ij}.

Wie muss die Produktion auf die Verkaufsstellen verteilt werden, um die Transportkosten zu minimieren und die Nachfrage zu decken?

$$\begin{aligned} \text{1Minimiere} \quad & f(x) \coloneqq \sum_{i=1}^{2} \sum_{j=1}^{12} c_{ij}\, x_{ij} \quad \text{über } x \in \mathbb{R}^{2\times 12} \\ \text{sodass} \quad & \sum_{j=1}^{12} x_{ij} \leqslant p_i, \quad i = 1, 2 \quad \text{(Produktionsrestriktionen)} \\ \text{und} \quad & \sum_{i=1}^{2} x_{ij} \geqslant n_j, \quad j = 1, \ldots, 12 \quad \text{(Bedarfsdeckung)} \\ \text{sowie} \quad & x_{ij} \geqslant 0, \quad i = 1, 2, \quad j = 1, \ldots, 12. \end{aligned}$$ △

Beispiel 1.5 (Norm-Projektionsaufgabe)**.**
Zu einer abgeschlossenen konvexen Menge $C \subseteq \mathbb{R}^n$ und einem Punkt $p \in \mathbb{R}^n$ suchen wir denjenigen Punkt $x \in C$, der p am nächsten liegt. Diese Aufgabe können wir wie folgt als Optimierungsaufgabe schreiben:

$$\begin{aligned} \text{Minimiere} \quad & f(x) \coloneqq \|x - p\|_* \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} \quad & x \in C. \end{aligned}$$

Die Norm $\|\cdot\|_*$ gibt dabei an, in welchem Sinne „am nächsten“ gemeint ist. △

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

Definition 1.6 (Klassifikation von Optimierungsaufgaben)**.**

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

(*ii*) Ungleichungsbeschränkungen der besonders einfachen Art

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

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

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

$$\text{Minimiere} \quad c^{\mathsf{T}} x \quad \text{sodass} \quad A\,x = b \quad \text{und} \quad x \geqslant 0.$$

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

[1] Wir behandeln unrestringierte Aufgaben in Kapitel 1.
[2] Lineare Optimierungsaufgaben werden in Kapitel 2 behandelt.
[3] Diese Aufgaben werden in Kapitel 3 besprochen.

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

(vi) Im allgemeinen Fall spricht man von **nichtlinearer Optimierung** (englisch: *nonlinear optimization*) und von einem **nichtlinearen Programm** (englisch: *nonlinear program*, **NLP**). Nichtlineare Optimierungsaufgaben und zugehörige Lösungsalgorithmen werden in der Lehrveranstaltung *Nonlinear Optimization* behandelt. △

Expertenwissen: Die Ursprünge der linearen Optimierung

Die Grundsteine der linearen Optimierung wurden in den 1940er Jahren von Leonid Kantorovich (1912–1986) und später von der 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.

Ende der Vorlesung 1

Wie wir bereits in Beispiel 1.2 gesehen haben, ist nicht jede Optimierungsaufgabe 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 (in beliebigen topologischen Räumen) ihr Minimum (und ihr Maximum) an.“ Damit folgt sofort: Wenn die Zielfunktion $f\colon F \to \mathbb{R}$ stetig und die zulässige Menge $F \subseteq \mathbb{R}^n$ kompakt ist, dann besitzt die Aufgabe

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

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

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

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

nichtleer und kompakt ist.

(2) An Stelle der Stetigkeit von f wird nur die Unterhalbstetigkeit benötigt.

Definition 1.7 (Unterhalbstetigkeit)**.**
Es sei $F \subseteq \mathbb{R}^n$ eine nichtleere Menge. Eine Funktion $f\colon F \to \mathbb{R}$ heißt **unterhalbstetig** (auch: **halbstetig von unten**, englisch: *lower semicontinuous*) auf F, wenn gilt:

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

Lemma 1.8 (Äquivalente Charakterisierung der Unterhalbstetigkeit)**.**
Es sei $F \subseteq \mathbb{R}^n$ eine nichtleere Menge. Für eine Funktion $f\colon F \to \mathbb{R}$ sind äquivalent:

(i) f ist unterhalbstetig auf F.

(*ii*) Alle Sublevelmengen $\{x \in F \mid f(x) \leqslant m\}$, $m \in \mathbb{R}$, sind abgeschlossen in F.

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

Umgekehrt gelte Aussage (*ii*). Wir betrachten eine Folge $\left(x^{(k)}\right) \subseteq F$, $x^{(k)} \to x^* \in F$. Wir nehmen das Gegenteil von Aussage (*i*) an, also $C \coloneqq \liminf_{k\to\infty} f(x^{(k)}) < f(x^*)$. Nach Definition des Limes inferior gibt es eine Teilfolge mit den Indizes $k^{(\ell)}$, sodass $f(x^{(k^{(\ell)})}) \to C < f(x^*)$ 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 $f(x^{(k^{(\ell)})}) + \varepsilon \leqslant f(x^*)$ 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 $f(x^*) \leqslant f(x^*) - \varepsilon$ ist — ein Widerspruch. □

Expertenwissen: Eine weitere äquivalente Charakterisierung der Unterhalbstetigkeit

Die Unterhalbstetigkeit einer Funktion $f\colon F \to \mathbb{R}$ ist weiterhin äquivalent dazu, dass ihr **Epigraph**

$$\operatorname{epi} f \coloneqq \left\{ \binom{x}{\gamma} \in \mathbb{R}^n \times \mathbb{R} \,\middle|\, x F,\ \gamma \geqslant f(x) \right\}$$

abgeschlossen ist.

Wir formulieren nun einen allgemeinen Existenzsatz für globale Minimierer.

Satz 1.9 (Existenz eines globalen Minimierers)**.**
Die zulässige Menge $F \subseteq \mathbb{R}^n$ sei nichtleer. Weiter sei $f\colon F \to \mathbb{R}$ unterhalbstetig, und für irgendein $m \in \mathbb{R}$ sei die Sublevelmenge

$$L \coloneqq \{x \in F \mid f(x) \leqslant m\}$$

nichtleer und kompakt.[4] Dann besitzt die Aufgabe

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

mimdestens einen globalen Minimierer.

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

[4] **Beachte:** Bei der Kompaktheit von Mengen kommt es nicht auf die Teilraumtopologie an, in der wir diese betrachten. Bei der Kompaktheit von $L = \{x \in F \mid f(x) \leqslant m\}$ kommt es also nicht darauf an, ob wir sie als kompakte Teilmenge von F oder von $\mathbb{R}^n$ ansehen. Die Charakterisierung „kompakt ⇔ 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 Infimalwert und $m \in \mathbb{R}$ ein Niveau wie in der Voraussetzung. Dann gibt es eine Folge $\big(x^{(k)}\big) \subseteq F$ mit der Eigenschaft[5] $f(x^{(k)}) \searrow f^*$. Wir unterscheiden zwei Fälle:

Fall 1: Falls $m = f^*$ ist, dann besteht die Sublevelmenge L ausschließlich aus globalen Minimierern und ist nach Annahme nichtleer; fertig.

Fall 2: 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 $x^{(k^{(\ell)})} \to x^*$, deren Grenzwert x^* in L liegt und insbesondere zulässig ist. Wegen der Unterhalbstetigkeit von f gilt $\lim_{\ell\to\infty} f(x^{(k^{(\ell)})}) \geqslant f(x^*)$, aber auch $\lim_{\ell\to\infty} f(x^{(k^{(\ell)})}) = f^*$. Dies zeigt, dass x^* ein globaler Minimierer ist. □

Das folgende Beispiel zeigt, dass die Voraussetzungen von Satz 1.9 alle wesentlich sind.

Beispiel 1.10 (Die Voraussetzungen von Satz 1.9 sind wesentlich)**.**

(i) Die Unterhalbstetigkeit von f ist wesentlich:

Ist die Funktion f nicht unterhalbstetig, so besitzt sie also mindestens eine Sublevelmenge, die in F nicht abgeschlossen ist: Betrachte zum Beispiel die Funktion

$$f(x) = \begin{cases} x^2 & \text{für } x \leqslant 0 \\ x^2 - 1 & \text{für } x > 0. \end{cases}$$

Die Minimierung von f über $F = [-1, 1]$ besitzt keine Lösung. Die Sublevelmenge zu $m = -0.5$ ist $(0, 1/\sqrt{2}]$, und diese ist in F nicht abgeschlossen.

(ii) Die Abgeschlossenheit der nichtleeren Sublevelmengen von f ist wesentlich:

Im Beispiel

$$\text{Minimiere} \quad f(x) := 1/x \quad \text{über } x \in F = [1, 2)$$

sind alle nichtleeren Sublevelmengen $L := \{x \in F \,|\, f(x) \leqslant m\}$ in $\mathbb{R}$ nicht abgeschlossen. **Beachte:** In $\mathbb{R}_{>0}$ sind die Sublevelmengen $f^{-1}\big((-\infty, m]\big)$ von f abgeschlossen. Auch in $\mathbb{R}$ sind die Sublevelmengen abgeschlossen, aber durch den Schnitt mit der nicht-abgeschlossenen Menge F geht diese Eigenschaft verloren!

(iii) Die Beschränktheit der nichtleeren Sublevelmengen von f ist wesentlich:

Im Beispiel

$$\text{Minimiere} \quad f(x) := 1/x \quad \text{über } x \in F = [1, \infty)$$

sind alle nichtleeren Sublevelmengen $L := \{x \in F \,|\, f(x) \leqslant m\}$ in $\mathbb{R}$ unbeschränkt. △

§ 2 Notation und Wiederholung von Differenzierbarkeitsbegriffen

Wir verwenden folgende Notation:

- Die natürlichen Zahlen sind $\mathbb{N} = \{1, 2, \ldots\}$. Wir schreiben $\mathbb{N}_0$ für $\mathbb{N} \cup \{0\}$.

[5] Für eine reelle Zahlenfolge $\big(y^{(k)}\big)$ bedeutet $y^{(k)} \searrow y$, dass $y^{(k)} > y$ gilt und $y^{(k)} \to y$. Die Monotonie der Folge wird nicht verlangt.

- Wir bezeichnen offene Intervalle mit (a, b) und abgeschlossene Intervalle mit $[a, b]$.
- Matrizen werden üblicherweise mit lateinischen Großbuchstaben bezeichnet, Vektoren mit lateinischen Kleinbuchstaben und Skalare mit griechischen oder lateinischen Kleinbuchstaben. Die Einheitsmatrix wird mit Id bezeichnet. Wir unterscheiden den Vektorraum der Spaltenvektoren $\mathbb{R}^n$ vom Vektorraum der Zeilenvektoren $\mathbb{R}_n$.
- Unendliche skalarwertige oder vektorwertige Folgen $\mathbb{N} \to \mathbb{R}^n$ bezeichnen wir mit $\left(x^{(k)}\right)$ und nicht mit x_k etc., um einen Konflikt mit der Bezeichnung der Komponenten eines Vektors $x = (x_1, \dots, x_n)^\mathsf{T} \in \mathbb{R}^n$ zu vermeiden. Auch endlich viele Vektoren werden mit $x^{(1)}, x^{(2)}$ etc. bezeichnet.

  Beachte: Diese Konvention ist noch nicht überall konsequent umgesetzt.
- Die durch die streng monoton wachsende Folge $\mathbb{N} \ni \ell \mapsto k^{(\ell)} \in \mathbb{N}$ gebildete **Teilfolge** (englisch: *subsequence*) einer Folge $\left(x^{(k)}\right)$ wird mit $\left(x^{(k^{(\ell)})}\right)$ bezeichnet.
- Für Vektoren $x, y \in \mathbb{R}^n$ bezeichnet $x^\mathsf{T} y$ das Euklidische Innenprodukt und $\|x\|$ die Euklidische Norm:
$$\|x\| = \sqrt{x^\mathsf{T} x}.$$
  Wir schreiben also *nicht* $\langle x\,,\, y\rangle$ oder $x \cdot y$ für das Euklidische Innenprodukt.
- Ist $M \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix, so erzeugt sie ein Innenprodukt $x^\mathsf{T} M y$ und eine Norm $\|x\|_M = \sqrt{x^\mathsf{T} M x}$ auf $\mathbb{R}^n$. Es gilt $\|x\| = \|x\|_{\mathrm{Id}}$.
- Für $\varepsilon > 0$ und $x^* \in \mathbb{R}^n$ ist
$$B_\varepsilon(x^*) := \{x \in \mathbb{R}^n \mid \|x - x^*\| < \varepsilon\}$$
  die **offene ε-Umgebung** (englisch: *open ε-neighborhood*) von x^* oder auch die **offene ε-Kugel** (englisch: *open ε-ball*) um x^*. Die **abgeschlossene ε-Umgebung** (englisch: *closed ε-neighborhood*) von x^* oder auch die **abgeschlossene ε-Kugel** (englisch: *closed ε-ball*) notieren wir als
$$\overline{B_\varepsilon(x^*)} := \{x \in \mathbb{R}^n \mid \|x - x^*\| \leqslant \varepsilon\}.$$
- Das **Innere** (englisch: *interior*) einer Menge $M \subseteq \mathbb{R}^n$ bezeichnen wir mit $\operatorname{int} M$ und den **Abschluss** (englisch: *closure*) mit $\overline{M}$.
- Für eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ und gegebenes $x \in \mathbb{R}^n$ heißt die Ableitung der partiellen Funktion $t \mapsto f(x + t\, e^{(i)})$ an der Stelle $t = 0$ die i-te **partielle Ableitung** von f an der Stelle x, kurz: $\frac{\partial}{\partial x_i} f(x)$. Dabei ist $e^{(i)} = (0, \dots, 0, 1, 0, \dots, 0)^\mathsf{T}$ einer der Vektoren der Standardbasis von $\mathbb{R}^n$. Mit anderen Worten:
$$\frac{\partial}{\partial x_i} f(x) = \lim_{t\to 0} \frac{f(x + t\, e^{(i)}) - f(x)}{t}.$$
- Allgemeiner heißt die Ableitung der Funktion $t \mapsto f(x + t\, d)$ an der Stelle $t = 0$ die **(beidseitige) Richtungsableitung** (englisch: *(two-sided) directional derivative*) von f an der Stelle x in Richtung $d \in \mathbb{R}^n$, kurz:
$$\frac{\partial}{\partial d} f(x) = \lim_{t\to 0} \frac{f(x + t\, d) - f(x)}{t}.$$
- Die rechtsseitige Ableitung der Funktion $t \mapsto f(x + t\, d)$ an der Stelle $t = 0$ heißt die **(einseitige) Richtungsableitung** (englisch: *(one-sided) directional derivative*) 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 $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt **differenzierbar** (kurz: **diffbar**, englisch: *differentiable*) an der Stelle $x \in \mathbb{R}^n$, falls ein Vektor $v \in \mathbb{R}_n$ (Zeilenvektor) existiert, sodass gilt:

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

  Der Vektor v heißt in dem Fall die **(totale) Ableitung** von f (englisch: *(total) derivative* an der Stelle x und wird mit $f'(x)$ bezeichnet.

- Für diffbare Funktionen $f\colon \mathbb{R}^n \to \mathbb{R}$ gilt

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

  Den transponierten Vektor (Spaltenvektor)

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

  bezeichnen wir als den **Gradienten** (englisch: *gradient*) bzgl. des Euklidischen Innenprodukts von f an der Stelle x.

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

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

- Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt **stetig partiell diffbar** (englisch: *continuously partially differentiable*) 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 $F\colon \mathbb{R}^n \to \mathbb{R}^m$ heißt an der Stelle x **diffbar**, wenn alle Komponentenfunktionen $F_1, \ldots, F_m$ dort diffbar sind. In diesem Fall ist die Ableitung $F'(x)$ durch die **Jacobimatrix** (englisch: *Jacobian*) von F an der Stelle x, also durch

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

  gegeben.

- F heißt **stetig partiell diffbar** (englisch: *continuously partially differentiable*), wenn alle Einträge der Jacobimatrix in einer Umgebung von x existieren und an der Stelle x stetig sind. C^1-Funktionen sind überall diffbar.

- Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt **zweimal differenzierbar** (kurz: **zweimal diffbar**, englisch: *twice differentiable*) 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** (englisch: *Hessian*) von f an der Stelle x, also durch die Matrix der zweiten partiellen Ableitungen

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

gegeben. Diese ist dann symmetrisch (Satz von Schwarz).[6]

- Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt **zweimal stetig partiell differenzierbar** (englisch: *twice continuously partially differentiable*) oder kurz: $C^2(\mathbb{R}^n, \mathbb{R})$, wenn alle Einträge der Hessematrix in einer Umgebung von x existieren und an der Stelle x stetig sind. C^2-Funktionen sind überall zweimal diffbar.

Schließlich benötigen wir auch den **Satz von Taylor** (englisch: *Taylor's theorem*), 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 $f\colon G \to \mathbb{R}$ k-mal diffbar sowie $(k+1)$-mal diffbar im Punkt $\overline{x} \in G$. Dann gilt: Für alle $\varepsilon > 0$ existiert $\delta > 0$, sodass gilt:

$$\text{im Fall } k = 0: \quad \left|f(\overline{x}+d) - f(\overline{x}) - f'(\overline{x})\, d\right| \leqslant \varepsilon \|d\|,$$
$$\text{im Fall } k = 1: \quad \left|f(\overline{x}+d) - f(\overline{x}) - f'(\overline{x})\, d - \frac{1}{2} d^{\mathsf{T}} f''(\overline{x}) d\right| \leqslant \varepsilon \|d\|^2.$$

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 $f\colon G \to \mathbb{R}$ $(k+1)$-mal stetig partiell diffbar, kurz: $C^{k+1}(G, \mathbb{R})$. Falls $\overline{x}$ und $\overline{x} + d$ und die gesamte Verbindungsstrecke in G liegen, dann existiert $\xi \in (0, 1)$, sodass gilt:

$$\text{im Fall } k = 0: \quad f(\overline{x}+d) = f(\overline{x}) + f'(\overline{x}+\xi\, d)\, d \quad (\textbf{Mittelwertsatz}, \text{englisch: } \textit{mean value theorem}),$$
$$\text{im Fall } k = 1: \quad f(\overline{x}+d) = f(\overline{x}) + f'(\overline{x})\, d + \frac{1}{2} d^{\mathsf{T}} f''(\overline{x}+\xi\, d)\, d.$$

[6] siehe z. B. Cartan, 1967, Proposition 5.2.2

Kapitel 1 Unrestringierte Optimierung

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

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

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

§ 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 $\gamma\colon (-\delta, \delta) \to \mathbb{R}^n$, $\gamma(t) := x^* + t\,d$. 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

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

Andererseits konvergiert aber der Differenzenquotient für $t \to 0$ 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 $f'(x^*) = 0$. □

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

Quizfrage 3.2: Wie kann man sich die Eigenschaft „$f'(x) = 0$“ beispielsweise für eine Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ vorstellen?

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

Satz 3.2 (Notwendige Bedingung 2. Ordnung)**.**
Es sei x^* ein lokaler Minimierer, und f sei an der Stelle x^* zweimal diffbar. Dann ist die Hessematrix $f''(x^*)$ positiv semidefinit.[1]

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

$$\left|\varphi(t) - \varphi(0) - \varphi'(0)\,t - \frac{1}{2}\varphi''(0)\,t^2\right| \leqslant \varepsilon\,t^2$$

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

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

für alle $|t|$ hinreichend klein, folglich

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

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

Quizfrage 3.3: Wie kann man sich die Eigenschaft „$f''(x)$ ist positiv semidefinit“ beispielsweise für eine Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ 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, siehe Hausaufgabe 2.2.

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

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

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

Beweis. Wir nutzen dieses Mal Satz 2.1 direkt für die Funktion f (nicht entlang einer Kurve). Für jedes $\varepsilon > 0$ existiert $\delta > 0$, sodass gilt:

$$\left|f(x^* + d) - f(x^*) - f'(x^*)\,d - \frac{1}{2}d^\mathsf{T} f''(x^*)d\right| \leqslant \varepsilon\,\|d\|^2$$

[1] 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

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

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

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

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:

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

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

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

für alle $\|d\| < \delta$. □

Zu der Eigenschaft (3.1) sagt man auch, die Funktion f habe mindestens **quadratisches Wachstum** (englisch: *quadratic growth*) in der Nähe von x^* bzw. f verhalte sich lokal **stark konvex** (englisch: *strongly convex*), siehe Definition 13.9.

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

Quizfrage 3.5: Welche Eigenschaft der Funktion f beschreibt der kleinste Eigenwert μ 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.

Beispiel 3.4 (Möglichkeiten und Grenzen von Bedingungen 1. und 2. Ordnung)**.**
Nachfolgend stehen Beispiele für folgende Situationen in $\mathbb{R}$ bzw. in $\mathbb{R}^2$:

(*i*) Es liegt eine lokale Minimalstelle vor, und die hinreichende Bedingung 2. Ordnung ist erfüllt:

$$\begin{aligned} f(x) &= x^2 && \text{bei } x = 0, \\ f(x) &= x_1^2 + x_2^2 && \text{bei } x = (0,0)^\mathsf{T}. \end{aligned}$$

(*ii*) Es liegt eine lokale Minimalstelle vor, und die notwendigen Bedingungen 1. und 2. Ordnung sind erfüllt, aber nicht die hinreichende Bedingung 2. Ordnung:

$$f(x) = x^4 \qquad \text{bei } x = 0,$$
$$f(x) = x_1^2 + x_2^4 \qquad \text{bei } x = (0,0)^\mathsf{T}.$$

(*iii*) Es liegt keine lokale Minimalstelle vor, die notwendige Bedingung 1. Ordnung ist erfüllt, aber die notwendige Bedingung 2. Ordnung ist nicht erfüllt:

$$f(x) = -x^2 \qquad \text{bei } x = 0,$$
$$f(x) = -x_1^2 + x_2^2 \qquad \text{bei } x = (0,0)^\mathsf{T}.$$

(*iv*) Es liegt keine lokale Minimalstelle vor, aber die notwendigen Bedingungen 1. Ordnung und 2. Ordnung sind beide erfüllt.

$$f(x) = x^3 \qquad \text{bei } x = 0,$$
$$f(x) = x_1^2 - x_2^4 \qquad \text{bei } x = (0,0)^\mathsf{T}.$$ △

Quizfrage 3.6: Kann ein Punkt x^*, der die notwendigen Bedingungen 1. und 2. Ordnung erfüllt, ein lokaler Maximierer sein?

Ende der Vorlesung 2

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** (englisch: *descent method*) 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** (englisch: *iterates*) $x^{(k)} \subseteq \mathbb{R}^n$. In jeder Iteration werden folgende Schritte ausgeführt:

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

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

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

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

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

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

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 eine **Abstiegsrichtung** (englisch: *descent direction*) für f im Punkt $x \in \mathbb{R}^n$, wenn

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

gilt. △

Differenzierbare Funktionen besitzen in jedem Punkt eine Abstiegsrichtung, außer in stationären Punkten. (**Quizfrage 4.1:** Klar?) Der negative Gradient $-\nabla f(x)$ ist dabei eine **Richtung des steilsten Abstiegs** (englisch: *direction of steepest descent*) von f im Punkt x. Er ist immer eine Abstiegsrichtung, außer wenn x ein stationärer Punkt ist. Wir können (4.1) auch als Innenprodukt schreiben:

$$\nabla f(x)^\mathsf{T} d < 0.$$

Diese Aussage bedeutet, dass der Winkel zwischen der Richtung d und dem negativen Gradienten $-\nabla f(x)$ kleiner als 90° ist, siehe Abbildung 4.1.

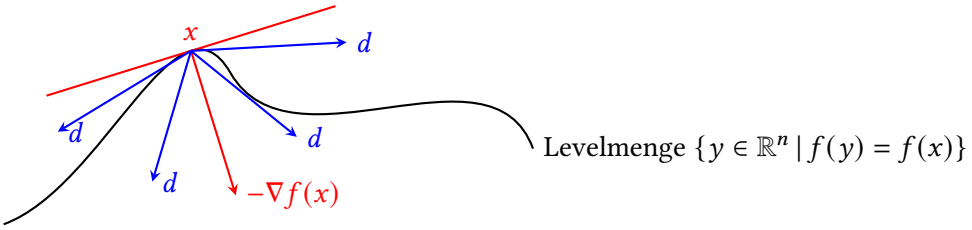


Abbildung 4.1.: Verschiedene Abstiegsrichtungen d für f im Punkt x.

Quizfrage 4.2: Mit welchem Begriff könnte man die Menge aller Abstiegsrichtungen einer Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ in einem Punkt x geometrisch beschreiben?

Zur Übung können Sie den negativen Gradienten und die Menge aller Abstiegsrichtungen in den Konturplot in Abbildung 4.2 an ausgewählten Punkte einzeichnen.

§ 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:

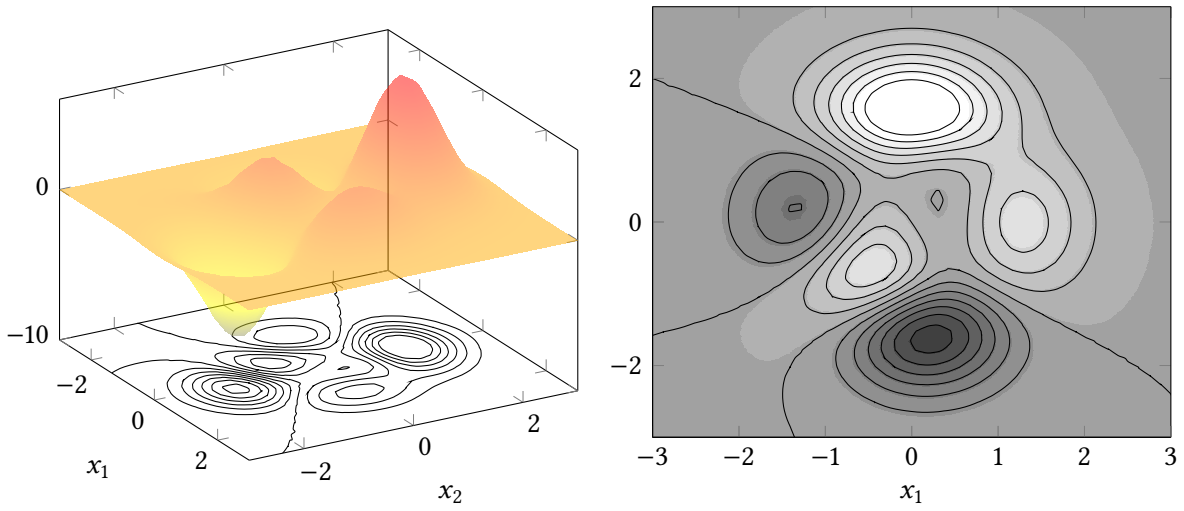


Abbildung 4.2.: Konturplot der peaks-Funktion aus MATLAB, also $f(x) = 3\,(1-x_1)^2\,\exp\!\left(-x_1^2-(x_2+1)^2\right) - 10\left(\frac{x_1}{5}-x_1^3-x_2^5\right)\exp\!\left(-x_1^2-x_2^2\right) - \frac{1}{3}\exp\!\left(-(x_1+1)^2-x_2^2\right)$, auf der Menge $[-3,3]\times[-3,3] \subseteq \mathbb{R}^2$.

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

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

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

Quizfrage 4.3: 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*)

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

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

Quizfrage 4.4: Für welche Funktionen könnte die exakte Liniensuche praktisch durchführbar sein?

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**[2] (englisch: *Armijo condition*) erfüllt ist:

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

[2] vorgeschlagen in Armijo, 1966

Dabei ist $\sigma \in (0,1)$ der **Armijo-Parameter** (englisch: *Armijo parameter*). (**Quizfrage 4.5:** Welche anschauliche Bedeutung hat der Parameter σ?)

Zur Veranschaulichung der Bedingung (4.3) führen wir die **Liniensuchfunktion** (englisch: *line search function*)

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

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

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

Also lautet die Armijo-Bedingung (4.3) alternativ

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

Diese Bedingung wird in Abbildung 4.3 illustriert. **Beachte:** Beim Gradientenverfahren gilt $\varphi'(0) = f'(x)\,d = -\|\nabla f(x)\|^2$.

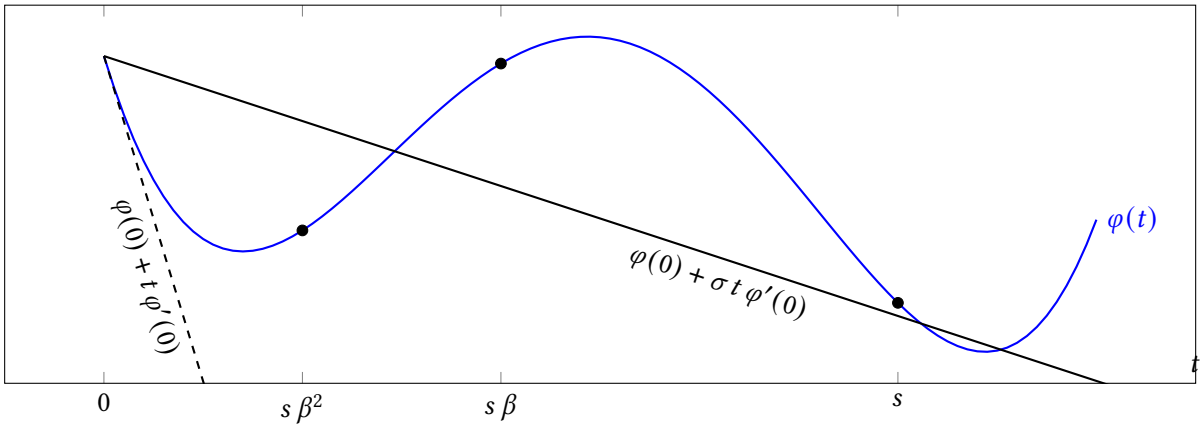


Abbildung 4.3.: Darstellung der Armijo-Bedingung (4.5) und einige Test-Schrittweiten beim Backtracking. Der Armijo-Parameter ist hier als $\sigma = 0.1$ und der Backtracking-Parameter als $\beta = 0.5$ gewählt.

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**. Ein Algorithmus für das Armijo-Backtracking-Verfahren ist

in Algorithmus 4.3 angegeben.

Algorithmus 4.3 (Armijo-Backtracking)**.**
Eingabe: Liniensuchfunktion $\varphi\colon \mathbb{R} \to \mathbb{R}$
Eingabe: Funktionswert $\varphi(0)$ und $\varphi'(0)$
Eingabe: Armijo-Parameter $\sigma \in (0,1)$, Backtracking-Parameter $\beta \in (0,1)$, Startschrittweite $s > 0$
Ausgabe: Schrittweite $0 < t \leqslant s$, die die Armijo-Bedingung $\varphi(t) \leqslant \varphi(0) + \sigma\, t\, \varphi'(0)$ erfüllt
Ausgabe: zugehöriger Funktionswert $\varphi(t)$

Setze $\ell := 0$
Setze $t^{(0)} := s$
Setze fertig := false
while nicht fertig **do**
    Bestimme den Funktionswert $\varphi(t^{(\ell)})$
    if $\varphi(t^{(\ell)}) \leqslant \varphi(0) + \sigma\, t^{(\ell)}\, \varphi'(0)$ **then**
        Setze fertig := true // $t^{(\ell)}$ erfüllt die Armijo-Bedingung (4.5)
    else
        Setze $t^{(\ell+1)} := \beta\, t^{(\ell)}$
        Setze $\ell := \ell + 1$
    end if
    return $t^{(\ell)}$ und $\varphi(t^{(\ell)})$
end while

Satz 4.4 (Wohldefiniertheit der Armijo-Backtracking-Strategie)**.**
Es sei $\sigma \in (0,1)$ und 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 für jede Startschrittweite $s > 0$ ein Exponent $\ell_0 \in \mathbb{N}_0$ existiert, sodass Schrittweiten der Form $t = s\,\beta^{\ell}$ für $\ell \geqslant \ell_0$ immer in $[0,T]$ liegen. Spätestens beim ℓ_0-ten Backtracking-Schritt ist also die Armijo-Bedingung erfüllt. Es sind auch andere Strategien als die fortlaufende Multiplikation mit einem festen Faktor $\beta \in (0,1)$ möglich, solange nur sichergestellt ist, dass man nach endlich vielen Iterationen in jedem noch so kleinen Intervall $[0,T]$ landet.

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

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

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

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

Im Grenzübergang $k \to \infty$ folgt

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

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

Wir geben nun das Gradientenverfahren mit Armijo-Liniensuche an:

Algorithmus 4.5 (Gradientenverfahren mit Armijo-Schrittweitensuche).
Eingabe: Zielfunktion $f\colon \mathbb{R}^n \to \mathbb{R}$
Eingabe: Startschätzung $x^{(0)} \in \mathbb{R}^n$
Eingabe: Armijo-Parameter $\sigma \in (0,1)$, Backtracking-Parameter $\beta \in (0,1)$, Startschrittweite $s > 0$
Ausgabe: $x \in \mathbb{R}^n$ mit $f(x) \leqslant f(x^{(0)})$

1: Setze $k := 0$
2: **while** Abbruchkriterium nicht erfüllt **do**
3: Setze $f^{(k)} := f(x^{(k)})$
4: Setze $d^{(k)} := -\nabla f(x^{(k)})$
5: Bilde die Liniensuchfunktion $\varphi := t \mapsto f(x^{(k)} + t\, d^{(k)})$
6: Bestimme eine Schrittweite $t > 0$, die die Armijo-Bedingung (4.5) erfüllt. Nutze dafür die Armijo-Backtracking-Strategie (Algorithmus 4.3) mit den Daten $\varphi(0) = f(x^{(k)})$ und $\varphi'(0) = f'(x^{(k)})\, d^{(k)} = -\|\nabla f(x)\|^2 = -\|d^{(k)}\|^2$, den gegebenen Parametern σ und β sowie der Startschrittweite $s > 0$
7: Setze $x^{(k+1)} := x^{(k)} + t^{(k)}\, d^{(k)}$
8: Setze $k := k + 1$
9: **end while**
10: **return** $x^{(k)}$

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 4.6: Angenommen, für die Funktion f liegt (neben der Routine für die Auswertung der Funktionswerte) eine Routine vor, die zu einer gegegenen Stelle x und einer gegebenen Richtung d die Richtungsableitung $f'(x)\, d$ bestimmt. Wieso reicht das für die Durchführung von Algorithmus 4.5 aus? Wie bestimmt man insbesondere den negativen Gradienten in Zeile 4?

Für den Beweis eines Konvergenzsatzes für das Gradientenverfahrens benötigen wir folgendes Resultat, das die vorausgesetzte C^1-Eigenschaft der Zielfunktion nutzt:

Lemma 4.6 (Konvergenz des Differenzenquotienten bei variabler Stelle und Richtung).
Es seien $x, d \in \mathbb{R}^n$, $x^{(k)}, d^{(k)} \subseteq \mathbb{R}^n$ mit $x^{(k)} \to x$ und $d^{(k)} \to d$ sowie $t^{(k)} \searrow 0$. Dann gilt

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

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

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

□

Wir analysieren jetzt Algorithmus 4.5 ohne Abbruchbedingung, sodass eine unendliche Folge $x^{(k)}$ entsteht. Insbesondere nehmen wir an, dass kein Punkt $x^{(k)}$ stationär ist.

Satz 4.7 (Ein globaler Konvergenzsatz für das Gradientenverfahren)**.**
Jeder Häufungspunkt x^* einer durch Algorithmus 4.5 erzeugten Folge $x^{(k)}$ ist ein stationärer Punkt von f, erfüllt also $f'(x^*) = 0$.

Beweis. Es sei $x^* \in \mathbb{R}^n$ ein Häufungspunkt von $x^{(k)}$. Es gibt also eine Teilfolge $x^{(k^{(\ell)})}$ mit $x^{(k^{(\ell)})} \to x^*$, und wegen der Stetigkeit von f gilt $f(x^{(k^{(\ell)})}) \to f(x^*)$. Da $f(x^{(k)})$ aber monoton fällt, konvergiert die gesamte Folge $f(x^{(k)}) \to f(x^*)$, siehe auch Hausaufgabe 2.2. Somit gilt also auch $f(x^{(k+1)}) - f(x^{(k)}) \to 0$.

Angenommen, es sei $f'(x^*) \neq 0$. Aus Zeilen 4, 6 und 7 des Algorithmus 4.5 folgt

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

also

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

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

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

$$\begin{aligned} & f(x^{(k^{(\ell)})} + \beta^{-1} t^{(k^{(\ell)})}\, d^{(k^{(\ell)})}) > f(x^{(k^{(\ell)})}) + \sigma\, \beta^{-1} t^{(k^{(\ell)})}\, f'(x^{(k^{(\ell)})})\, d^{(k^{(\ell)})} \\ \Rightarrow \quad & \frac{f(x^{(k^{(\ell)})} + \beta^{-1} t^{(k^{(\ell)})}\, d^{(k^{(\ell)})}) - f(x^{(k^{(\ell)})})}{\beta^{-1} t^{(k^{(\ell)})}} > \sigma\, f'(x^{(k^{(\ell)})})\, d^{(k^{(\ell)})} = -\sigma\, \|\nabla f(x^{(k^{(\ell)})})\|^2. \end{aligned}$$

Die Grenzübergänge $x^{(k^{(\ell)})} \to x^*$, $d^{(k^{(\ell)})} = -\nabla f(x^{(k^{(\ell)})}) \to -\nabla f(x^*)$ und $t^{(k^{(\ell)})} \to 0$ für $\ell \to \infty$ ergeben mit Lemma 4.6:

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

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

Bemerkung 4.8 (Zur praktischen Implementierung des Gradientenverfahrens)**.**
Typische Abbruchkriterien beim Gradientenverfahren[3] sind:

(*i*) $f(x^{(k-1)}) - f(x^{(k)}) \leqslant \mathrm{ATOL}_f + \mathrm{RTOL}_f\, |f(x^{(k-1)})|$,

(*ii*) $\|x^{(k-1)} - x^{(k)}\| \leqslant \mathrm{ATOL}_x + \mathrm{RTOL}_x\, \|x^{(k-1)}\|$.

Gefordert werden beide Bedingungen gleichzeitig. Dabei wird oft $\mathrm{RTOL}_f = \mathrm{RTOL}_x^2$ gewählt. Als „Notbremsen“ dienen zusätzlich die Abfragen

(*iii*) $\|\nabla f(x^{(k)})\| \leqslant \mathrm{ATOL}_{\nabla f(x)} + \mathrm{RTOL}_{\nabla f(x)} \|\nabla f(x^{(0)})\|$,

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

(*iv*) $k \leqslant k_{\max}$

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

Quizfrage 4.7: Welche Bedeutung haben die Bedingungen (*i*) bis (*iii*)?

Quizfrage 4.8: Wie setzt man ATOL und RTOL, wenn man in Bedingungen (*i*) bis (*iii*) entweder ausschließlich eine absolute oder ausschließlich eine relative Abbruchbedingung verwenden möchte?

Bemerkung 4.9 (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:

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

Speziell beim Gradientenverfahren ergibt sich dann also

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

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

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

für die Startschrittweite.

Auch unter Verwendung dieser Startschrittweiten kann man Satz 4.7 beweisen. △

Expertenwissen: Das Gradientenverfahren als diskretisierter Gradientenfluss

Man kann das Gradientenverfahren als Diskretisierungsverfahren für den sogenannten (Euklidischen) **Gradientenfluss** (englisch: *gradient flow*) der Zielfunktion f verstehen, das ist die Differentialgleichung

$$\dot{x}(r) = -\nabla f(x(r))$$

mit der Anfangsbedingung $x(0) = x^{(0)}$.

Die Diskretisierung mit dem expliziten Euler-Verfahren zur Schrittweite $t^{(k)}$ ergibt die Iterationsvorschrift

$$\frac{x^{(k+1)} - x^{(k)}}{t^{(k)}} = -\nabla f(x^{(k)}),$$

also

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

Dabei approximiert $x^{(k)}$ die Lösung x an der Stelle (zur Zeit) $r^{(k)}$, und $t^{(k)} = r^{(k+1)} - r^{(k)}$ bezeichnet die gewählte Schrittweite.

Ende der Vorlesung 3

§ 4.2 Das Gradientenverfahren in einem alternativen Innenprodukt

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

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

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

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

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

Per Definition maximiert die/eine **Richtung des steilsten Anstiegs** die Richtungsableitung $f'(x)\, d$ über alle Vektoren $d \in \mathbb{R}^n$ konstanter Länge. Alternativ können wir die Richtungsableitung auch mit der Länge von d normieren. Es ergibt sich die Optimierungsaufgabe

$$\text{Maximiere} \quad \frac{f'(x)\, d}{\|d\|_M} \quad \text{über } d \in \mathbb{R}^n \setminus \{0\}. \tag{4.9}$$

Wenn x ein stationärer Punkt ist, dann ist die Zielfunktion in (4.9) konstant Null. Diesen uninteressanten Fall schließen wir jetzt aus.

Wir können die Aufgabe (4.9) explizit lösen. Dazu schreiben wir den Zähler der Zielfunktion als M-Innenprodukt um:

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

wobei die Symmetrie $M = M^\mathsf{T}$ benutzt wurde, wodurch auch M^{-1} symmetrisch ist. Die Cauchy-Schwarz-Ungleichung besagt, dass dieser Ausdruck genau dann maximal wird, wenn d parallel zu $M^{-1}\nabla f(x)$ liegt, also wenn d irgendein positives Vielfaches von $M^{-1}\nabla f(x)$ ist. Er wird dagegen minimal, wenn d antiparallel zu $M^{-1}\nabla f(x)$ liegt, also wenn d irgendein negatives Vielfaches von $M^{-1}\nabla f(x)$ ist. Wir fassen zusammen:

Lemma 4.10 (Richtung des steilsten Abstiegs im M-Innenprodukt)**.**
Falls $f'(x) \neq 0$ gilt, dann ist eine Lösung d^* von (4.9) gegeben durch den ***M*-Gradienten** (englisch: *M-gradient*) von f an der Stelle x:

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

(Genau die positiven Vielfachen von d^* sind ebenfalls Lösungen.)

Daher nennt man $d^* = -\nabla_M f(x)$ eine **Richtung des steilsten Abstiegs bzgl. des M-Innenprodukts** (englisch: *steepest descent direction w.r.t. the M-inner product*) oder den **negativen M-Gradienten**. Wir berechnen diese durch Lösung des linearen Gleichungssystems

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

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

Expertenwissen: Der Gradient ist der Riesz-Repräsentant der Ableitung

Die Ableitung einer Funktion $f\colon X \to \mathbb{R}$ auf einem (reellen) Vektorraum X (hier $X = \mathbb{R}^n$) an einer Stelle $x \in X$ ist per Definition eine stetige lineare Abbildung $f'(x) \in \mathrm{Hom}(X, \mathbb{R})$, also ein Element des Dualraumes $f'(x) \in X^*$. Um aus dem dualen Objekt *Ableitung* in X^* eine primale *Richtung* im Raum X zu gewinnen, benötigen wir ein Innenprodukt auf dem Raum X.

In dem für uns relevanten Fall $X = \mathbb{R}^n$ wird dieses Innenprodukt durch die s. p. d. Matrix M repräsentiert. X wird damit (wegen der endlichen Dimension ohne weitere Annahmen) zu einem Hilbert-Raum. Mit dem Innenprodukt auf X kommt auch das zugehörige Innenprodukt M^{-1} auf X^* sowie der Riesz-Isomorphismus $R\colon X \to X^*$, mit dem primale und duale Objekte ineinander umgerechnet werden können. (Mehr dazu in der Vorlesung *Funktionalanalysis*.) Im Fall von $X = \mathbb{R}^n$ ist der Riesz-Isomorphismus gerade gegeben durch die Multiplikation eines Vektors $x \in \mathbb{R}^n$ mit M. Das Ergebnis $M\,x$ repräsentiert dann einen Vektor in X^*. Die inverse Riesz-Abbildung ist dementsprechend durch Multiplikation mit M^{-1} gegeben.

Daraus erkennen wir, dass der M-Gradient $\nabla_M f(x) \in X$ nichts anderes ist als der Riesz-Repräsentant der Ableitung $f'(x) \in X^*$. Auch der Euklidische Gradient $\nabla f(x)$ ist der Riesz-Repräsentant der Ableitung $f'(x)$, nur eben im Spezialfall, dass das Euklidische Innenprodukt ($M = \mathrm{Id}$) genutzt wird.

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):

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

falls nicht x bereits ein stationärer Punkt ist.

Expertenwissen: Einfluss des M-Innenprodukts

Angenommen, $f'(x)$ ist bekannt. Welche verschiedenen Abstiegsrichtungen kann man durch Variation des Innenprodukts M daraus erzeugen? Welche möglichen Richtungen d

$$R := \{d \in \mathbb{R}^n \mid d = -M^{-1}\nabla f(x),\ M \text{ ist s. p. d.}\}$$

ergeben sich also aus (4.11), wenn man M über alle s. p. d. Matrizen laufen lässt?

Die Antwort ist: R ist ein offener Halbraum, der $\nabla f(x)$ als Euklidischen Normalenvektor hat, also

$$H := \{d \in \mathbb{R}^n \mid \nabla f(x)^{\mathsf{T}} d < 0\}.$$

Zum Beweis sei zunächst M eine s. p. d. Matrix und $d = -M^{-1}\nabla f(x)$. Dann ist $\nabla f(x)^\mathsf{T} d = -\|\nabla f(x)\|^2_{M^{-1}} < 0$, also liegt d in besagtem Halbraum: $R \subseteq H$.

Umgekehrt sei $d \in \mathbb{R}^n$ ein Vektor mit der Eigenschaft $\nabla f(x)^\mathsf{T} d < 0$. Dann können wir M so wählen, dass d auf $-\nabla f(x)$ abgebildet wird. Das ist nicht ganz offensichtlich. Mögliche Konstruktionen von M ergeben sich aus der BFGS-Formel oder aus der DFP-Formel (die wir hier verwenden). Diese sogenannten **Quasi-Newton-Formeln** haben ursprünglich die Funktion der Approximation der Hessematrix in der algorithmischen Optimierung, siehe dazu die Vorlesung *Nonlinear Optimization*.

Wir schreiben zur Abkürzung $g := \nabla f(x)$ und $\gamma := -\frac{1}{g^\mathsf{T} d} > 0$ und wählen

$$M := \left(\mathrm{Id} + \gamma\, g\, d^\mathsf{T}\right)\left(\mathrm{Id} + \gamma\, d\, g^\mathsf{T}\right) + \gamma\, g\, g^\mathsf{T}.$$

Die Matrix M ist s. p. d., denn: Die Symmetrie von M ist offensichtlich. Weiter gilt für Vektoren $v \in \mathbb{R}^n$:

$$\begin{aligned} v^\mathsf{T}\left(\mathrm{Id} + \gamma\, g\, d^\mathsf{T}\right)\left(\mathrm{Id} + \gamma\, d\, g^\mathsf{T}\right) v + \gamma\, v^\mathsf{T} g\, g^\mathsf{T} v \\ = \left(v^\mathsf{T} + \gamma\, v^\mathsf{T} g\, d^\mathsf{T}\right)\left(v + \gamma\, d\, g^\mathsf{T} v\right) + \gamma\, v^\mathsf{T} g\, g^\mathsf{T} v \\ \geqslant 0. \end{aligned}$$

Beide Summanden sind $\geqslant 0$. Wir prüfen nun auf positive Definitheit. Dazu sei nun $v \neq 0$. Damit Gleichheit in der obigen Ungleichung gilt, muss notwendigerweise für den ersten Summanden

$$v + \gamma\,(g^\mathsf{T} v)\, d = 0$$

gelten, also ist v ein (nicht-triviales) Vielfaches von d. Daraus folgt aber, dass der zweite Summand

$$\gamma\, v^\mathsf{T} g\, g^\mathsf{T} v$$

wegen $g^\mathsf{T} d < 0$ ungleich null ist. Wir erhalten also die positive Definitheit:

$$v^\mathsf{T} M\, v > 0 \quad \text{für alle } v \neq 0.$$

Weiter gilt tatsächlich

$$\begin{aligned} M\, d &= \left(\mathrm{Id} + \gamma\, g\, d^\mathsf{T}\right)\left(\mathrm{Id} + \gamma\, d\, g^\mathsf{T}\right) d + \gamma\, g\, g^\mathsf{T} d \\ &= \left(\mathrm{Id} - \frac{1}{g^\mathsf{T} d}\, g\, d^\mathsf{T}\right)\left(\mathrm{Id} - \frac{1}{g^\mathsf{T} d}\, d\, g^\mathsf{T}\right) d - \frac{1}{g^\mathsf{T} d}\, g\, g^\mathsf{T} d \\ &= \left(\mathrm{Id} - \frac{1}{g^\mathsf{T} d}\, g\, d^\mathsf{T}\right)(d - d) - g, \end{aligned}$$

also $M\, d = -g = -\nabla f(x)$, und folglich $H \subseteq R$.

Algorithmisch ergeben sich durch Verwendung des M-Innenprodukts an Stelle des Euklidischen Innenprodukts folgende Änderungen: In Algorithmus 4.5 lautet Zeile 4 nun $d^{(k)} := -\nabla_M f(x^{(k)})$. Dieser Schritt wird durch Lösung des linearen Gleichungssystems

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

ausgeführt. Die übrigen Schritte, insbesondere die Formulierung der Armijo-Bedingung

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

bleiben unverändert. Der globale Konvergenz-Satz 4.7 gilt weiter. (**Quizfrage 4.9:** Können Sie die Änderungen im Beweis nachvollziehen?) Als Abbruchbedingung *(ii)* in Bemerkung 4.8 dient nun $\|x^{(k)} - x^{(k-1)}\|_M \leqslant \mathrm{ATOL}_x + \mathrm{RTOL}_x\, \|x^{(k-1)}\|_M$ und als Bedingung *(iii)* $\|\nabla_M f(x^{(k)})\|_M \leqslant \mathrm{ATOL}_{\nabla f(x)} + \mathrm{RTOL}_{\nabla_M f(x)} \|\nabla f(x^{(0)})\|_M$.

Quizfrage 4.10: Warum ändert sich Abbruchbedingung *(i)* nicht?

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

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

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 4.11 (Vorkonditioniertes Gradientenverfahren mit Armijo-Schrittweitensuche)**.**
Eingabe: Zielfunktion $f\colon \mathbb{R}^n \to \mathbb{R}$
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 $M \in \mathbb{R}^{n\times n}$
Ausgabe: $x \in \mathbb{R}^n$ mit $f(x) \leqslant f(x^{(0)})$

1: Setze $k := 0$
2: **while** Abbruchkriterium nicht erfüllt **do**
3: Setze $f^{(k)} := f(x^{(k)})$
4: Bestimme $d^{(k)}$ durch Lösung des linearen Gleichungssystems $M\, d^{(k)} = -\nabla f(x^{(k)})$
5: Bilde die Liniensuchfunktion $\varphi := t \mapsto f(x^{(k)} + t\, d^{(k)})$
6: Bestimme eine Schrittweite $t > 0$, die die Armijo-Bedingung (4.5) erfüllt. Nutze dafür die Armijo-Backtracking-Strategie (Algorithmus 4.3) mit den Daten $\varphi(0) = f(x^{(k)})$ und $\varphi'(0) = f'(x^{(k)})\, d^{(k)} = -\|\nabla_M f(x)\|_M^2 = -\|d^{(k)}\|_M$, den gegebenen Parametern σ und β sowie der Startschrittweite $s > 0$
7: Setze $x^{(k+1)} := x^{(k)} + t^{(k)}\, d^{(k)}$
8: Setze $k := k + 1$
9: **end while**
10: **return** $x^{(k)}$

Beachte: Das Verfahren verallgemeinert das unvorkonditionierte Gradientenverfahren (Algorithmus 4.5), das sich im Spezialfall $M = \mathrm{Id}$ aus Algorithmus 4.11 ergibt.

§ 4.3 Konvergenz bei quadratischer Zielfunktion und exakter Liniensuche

Literatur: Geiger, Kanzow, 1999, Kapitel 8.2

Aussagen zur Konvergenz von Gradientenverfahren für allgemeine nichtlineare Zielfunktionen folgen in der Vorlesung *Nonlinear Optimization*. Man kann unter relativ allgemeinen Voraussetzungen zeigen, dass jeder Häufungspunkt der Folge der Iterierten $\left(x^{(k)}\right)$ ein stationärer Punkt ist. Damit konvergieren Teilfolgen zumindest gegen interessante Punkte, wenn auch nicht notwendig gegen Minimierer. Diese Eigenschaft gilt bereits als eine globale Konvergenzaussage und unterstreicht die vergleichsweise hohe Robustheit des Gradientenverfahrens.

Wir betrachten hier nur den einfachsten sinnvollen Fall unrestringierter Optimierungsaufgaben. Bei diesen ist die Zielfunktion ein stark konvexes (siehe Kapitel 3) quadratisches Polynom:

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

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

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

denn dies ist die einzige Lösung der notwendigen Bedingungen (Satz 3.1), und die hinreichenden Bedingungen (Satz 3.3) sind dort erfüllt. 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 4.11: An welcher Stelle geht die Symmetrie der Matrix Q ein?

Wir untersuchen die Konvergenzgeschwindigkeit des (vorkonditionierten) Gradientenverfahrens (Algorithmus 4.12) bei Anwendung auf (4.14). Natürlich wird man das Gradientenverfahren zur Minimierung von (4.14) bzw. äquivalent zur Lösung 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 bzw. der Berechnung der Cholesky-Zerlegung 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 $Q\, x$ auswertet.

Beachte: Das Verfahren kommt bereits mit Matrix-Vektor-Produkten $Q\, x$ aus. Nur diese werden bei der Berechnung des Gradienten $\nabla f(x) = Q\, x + c$ in Zeile 4 benötigt. Auch für die Bestimmung des Funktionswertes $f(x) = \frac{1}{2} x^\mathsf{T} Q\, x + c^\mathsf{T} x + \gamma$ (Zeile 3) sind Matrix-Vektor-Produkte $Q\, x$ ausreichend.

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

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

im k-ten Schritt berechnen:

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

Das wird in Hausaufgabe 2.2 nachgerechnet.

In diesem Abschnitt wählen wir statt der Armijo-Strategie in Algorithmus 4.5 stets die exakte Schrittweite (4.16). Der Vollständigkeit halber geben wir das vorkonditionierte Gradientenverfahren für diesen Spezialfall nochmals an. Da wir der Übersichtlichkeit halber auf die Berechnung der Funktionswerte verzichten, ist der Offset $\gamma \in \mathbb{R}$ in der Zielfunktion (4.14) irrelevant.

Algorithmus 4.12 (Vorkonditioniertes Gradientenverfahren bei quadratischer Zielfunktion (4.14))**.**
Eingabe: Daten der Zielfunktion $Q \in \mathbb{R}^{n\times n}$ s. p. d. und $c \in \mathbb{R}^n$
Eingabe: Startschätzung $x^{(0)} \in \mathbb{R}^n$
Eingabe: s. p. d. Matrix $M \in \mathbb{R}^{n\times n}$
Ausgabe: $x \in \mathbb{R}^n$ mit $f(x) \leqslant f(x^{(0)})$

Setze $k := 0$
Setze $r^{(0)} := Q\,x^{(0)} + c$ // $r^{(0)} = \nabla f(x^{(0)})$
Setze $d^{(0)} := -M^{-1}r^{(0)}$ // $d^{(0)} = -\nabla_M f(x^{(0)})$
while Abbruchkriterium nicht erfüllt **do**
    Setze $q^{(k)} := Q\,d^{(k)}$
    Setze $t^{(k)} := -\frac{(r^{(k)})^\top d^{(k)}}{(d^{(k)})^\top q^{(k)}}$ // exakte Schrittweite
    Setze $x^{(k+1)} := x^{(k)} + t^{(k)}\,d^{(k)}$
    Setze $r^{(k+1)} := r^{(k)} + t^{(k)}\,q^{(k)}$
    Setze $d^{(k+1)} := -M^{-1}r^{(k+1)}$
    Setze $k := k + 1$
end while
return $x^{(k)}$

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

Lemma 4.13 (Zusammenhang zwischen Funktionswerten, Fehler und Residuum)**.**

Es sei f wie in (4.14) mit einer s. p. d. Matrix $Q \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$ und $\gamma \in \mathbb{R}$. Weiter sei $x^* = -Q^{-1}c$ der eindeutige globale Minimierer von f. Dann gilt für alle $x \in \mathbb{R}^n$ und daas zugehörige Residuum $r = Q\,x + c$

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

Beweis. Durch direkte Rechnung ergibt sich

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

□

Für die Iterierten von Algorithmus 4.12 zur Minimierung von (4.14) gilt nun die folgende Rekursion:

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

Hier setzen wir nun den speziellen Zusammenhang $r^{(k)} = -M\,d^{(k)}$ für die Iterierten aus Algorithmus 4.12 ein:

$$= \left(1 - \frac{\left[(d^{(k)})^\mathsf{T}M\,d^{(k)}\right]^2}{\left[(d^{(k)})^\mathsf{T}Q\,d^{(k)}\right]\left[(d^{(k)})^\mathsf{T}M\,Q^{-1}M\,d^{(k)}\right]}\right)\left(f(x^{(k)}) - f(x^*)\right). \tag{4.18}$$

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

Lemma 4.14 (Kantorovich-Ungleichung im Euklidischen Innenprodukt)**.**
Es sei $Q \in \mathbb{R}^{n\times n}$ s. p. d. und $\alpha := \lambda_{\min}(Q)$ sowie $\beta := \lambda_{\max}(Q)$. Dann gilt

$$1 \leqslant \frac{(x^\mathsf{T}Q\,x)\,(x^\mathsf{T}Q^{-1}x)}{\|x\|^4} \leqslant \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} \leqslant \frac{\beta}{\alpha} \tag{4.19}$$

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

Vor dem Beweis wollen wir die Ungleichung (4.19) interpretieren. Für den Rayleigh-Quotienten von Q gilt

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

Die erste Ungleichung ist genau für die Eigenvektoren von Q zum größten Eigenwert $\lambda_{\max}(Q)$ mit Gleichheit erfüllt. Die zweite Ungleichung ist genau für die Eigenvektoren von Q^{-1} zum größten Eigenwert $\lambda_{\max}(Q^{-1})$ mit Gleichheit erfüllt. Das sind aber genau die Eigenvektoren von Q zum *kleinsten* Eigenwert $\lambda_{\min}(Q)$.

Die offensichtliche Abschätzung

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

ist jedoch nicht scharf, da derselbe Vektor x i. A. nicht gleichzeitig Eigenvektor zum größten und zum kleinsten Eigenwert sein kann. (**Quizfrage 4.12:** Außer in welchem Fall?) Die Kantorovich-Ungleichung (4.19) verbessert diese Abschätzung.

Beweis von Lemma 4.14. Aus der Cauchy-Schwarz-Ungleichung folgt[4]

$$\|x\|^2 = x^\mathsf{T} x = x^\mathsf{T} Q^{-1/2}\, Q^{1/2} x \leqslant \|Q^{-1/2} x\|\, \|Q^{1/2} x\|.$$

Durch Quadrieren erhalten wir

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

und damit die untere Schranke in (4.19).

Es seien nun[5] $\lambda^{(1)}, \dots, \lambda^{(n)} > 0$ 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:

$$\frac{(x^\mathsf{T} Q\, x)\,(x^\mathsf{T} Q^{-1} x)}{\|x\|^4} = \underbrace{\Big[\sum_{i=1}^n \lambda^{(i)}\, (\gamma^{(i)})^2\Big]}_{=\mathbb{E}(T)}\, \underbrace{\Big[\sum_{i=1}^n \frac{1}{\lambda^{(i)}}\, (\gamma^{(i)})^2\Big]}_{=\mathbb{E}(1/T)}.$$

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 $0 < \alpha \leqslant T \leqslant \beta$ gilt

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

also auch

$$\frac{1}{T} \leqslant \frac{\alpha + \beta - T}{\alpha\, \beta}$$

[4] 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}.

[5] Wir folgen ab hier dem Beweis von Anderson, 1971, wie er in der Masterarbeit Alpargu, 1996, Abschnitt 1.2.2 wiedergegeben ist. Siehe auch Horn, Johnson, 1990, Theorem 7.4.41 für einen anderen Beweis mit Hilfe der Wielandt-Ungleichung.

und daher (Erwartungswert nehmen)

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

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

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*)

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

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

Folgerung 4.15 (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

$$1 \leqslant \frac{(x^{\mathsf{T}}Q\,x)\,(x^{\mathsf{T}}M\,Q^{-1}M\,x)}{\|x\|_M^4} \leqslant \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} \leqslant \frac{\beta}{\alpha} \tag{4.20}$$

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

Beweis. Für die Abschätzung nach unten verwenden wir

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

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

Wir benutzen nun die Cholesky-Zerlegung[6] $M = LL^{\mathsf{T}}$ und setzen $y := L^{\mathsf{T}}x$, also $x = L^{-\mathsf{T}}y$ ein:

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

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

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

also auch

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

[6] Stattdessen könnten wir auch mit der Matrix-Wurzel $M^{1/2}$ arbeiten.

Ersetzen wir noch v durch $L^\mathsf{T} w$, so erhalten wir

$$Q\,w = \lambda\,M\,w. \tag{4.21}$$

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

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

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

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

$$1 \leqslant \frac{(x^\mathsf{T} Q\, x)\,(x^\mathsf{T} M\, Q^{-1} M\, x)}{\|x\|_M^4} \leqslant \frac{(\kappa+1)^2}{4\,\kappa} \leqslant \kappa \tag{4.23}$$

schreiben.

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

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

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

Satz 4.16 (Globaler Konvergenzsatz für quadratische Zielfunktionen)**.**
Es seien Q und M s. p. d. Matrizen und κ die verallgemeinerte Konditionszahl von Q bzgl. M, siehe (4.22). Das Gradientenverfahren im M-Innenprodukt mit exakter Schrittweite zur Minimierung der Zielfunktion (4.14) (Algorithmus 4.12) konvergiert für jeden Startvektor $x^{(0)} \in \mathbb{R}^n$ gegen den eindeutigen globalen Minimierer $x^* = -Q^{-1}c$, und es gelten die Abschätzungen

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

und deswegen auch

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

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

Expertenwissen: Zur Bedeutung der Konditionszahl

In der Numerik lernt man, dass die Kondition einer zu lösenden Aufgabe ein Maß dafür darstellt, wie die Lösung von (Störungen in) den Eingangsdaten abhängt. Sie ist eine Eigenschaft der zu lösenden Aufgabe und hat nichts mit dem verwendeten Lösungsverfahren zu tun. Hier lernen wir, dass wir durch die Wahl von M die Konditionszahl beeinflussen können. Wir verändern sozusagen die Metrik, bzgl. der wir die Kondition messen.

Wir betrachten die Aufgabe $Q\,x = -c$ und die gestörte Aufgabe $Q\,(x + \Delta x) = -c + \Delta c$. Dann gilt für die Störung der Lösung $\Delta x = -Q^{-1}\Delta c$.

Betrachtung in absoluten Termen: Es gilt

$$\lambda_{\min}(Q^{-1})\|c\| \leqslant \|x\| \quad \leqslant \lambda_{\max}(Q^{-1})\|c\| = \frac{1}{\lambda_{\min}(Q)}\|c\|$$

und analog

$$\lambda_{\min}(Q^{-1})\|\Delta c\| \leqslant \|\Delta x\| \leqslant \lambda_{\max}(Q^{-1})\|\Delta c\| = \frac{1}{\lambda_{\min}(Q)}\|\Delta c\|,$$

also in relativen Termen:

$$\frac{\|\Delta x\|}{\|x\|} \leqslant \frac{\lambda_{\max}(Q^{-1})\|\Delta c\|}{\lambda_{\min}(Q^{-1})\|c\|} = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}\,\frac{\|\Delta c\|}{\|c\|} = \kappa\,\frac{\|\Delta c\|}{\|c\|}$$

mit der Euklidischen Konditionszahl $\kappa = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$.

Verwenden wir für x und Δx die M-Norm und für das duale Objekt c die M^{-1}-Norm, so ergibt sich analog zur obigen Rechnung allgemeiner:

$$\frac{\|\Delta x\|_M}{\|x\|_M} \leqslant \frac{\lambda_{\max}(Q;M)}{\lambda_{\min}(Q;M)}\,\frac{\|\Delta c\|_{M^{-1}}}{\|c\|_{M^{-1}}} = \kappa\,\frac{\|\Delta c\|_{M^{-1}}}{\|c\|_{M^{-1}}}$$

mit der verallgemeinerten Konditionszahl $\kappa = \frac{\lambda_{\max}(Q;M)}{\lambda_{\min}(Q;M)}$.

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

Bemerkung 4.17 (Zum Konvergenzverhaltens des Gradientenverfahrens für quadratische Zielfunktionen)**.**

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

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

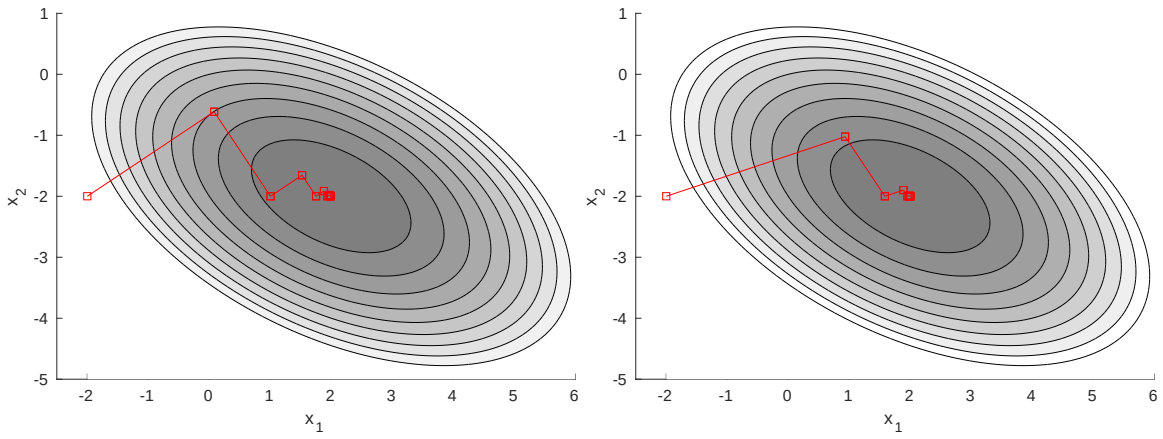


Abbildung 4.4.: Illustration des Gradientenverfahrens (Algorithmus 4.11) mit Startpunkt $x^{(0)} = (-2,-2)^\mathsf{T}$ und exakter Schrittweite (4.16) für die Minimierung von (4.14) mit $Q = \left(\begin{smallmatrix} 3 & 2 \\ 2 & 6 \end{smallmatrix}\right)$ und $c = \left(\begin{smallmatrix} -2 \\ 8 \end{smallmatrix}\right)$. Die exakte Lösung ist $x^* = (2,-2)^\mathsf{T}$. Verlauf bei Verwendung des Innenprodukts $M = \mathrm{Id}$ (links) und $M = \mathrm{diag}(Q)$ (rechts).

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

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

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

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

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

konvergenzbeschleunigend.

(*v*) An Stelle der exakten Schrittweiten kann man auch andere Schrittweitenstrategien verwenden, die das Konvergenzverhalten verbessern können. Mehr dazu in der Vorlesung *Nonlinear Optimization*.

(*vi*) Satz 4.16 liefert eine Abschätzung, die für alle Startschätzungen $x^{(0)}$ anwendbar ist. In der Praxis kann die Konvergenz besser sein. Insbesondere kann man zeigen, dass man Konvergenz in einer Iteration erhält, wenn der Anfangsfehler $e^{(0)} = x^{(0)} - x^*$ ein Eigenvektor zu irgendeinem Eigenwert des verallgemeinerten Eigenwertproblems zum Paar $(Q; M)$ ist, wenn also $Q\,e^{(0)} = \lambda\,M\,e^{(0)}$ gilt für irgendein λ (das notwendigerweise > 0 ist). △

Expertenwissen: Das Verfahren der konjugierten Gradienten

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**, englisch: *conjugate gradient method, CG method*), siehe Vorlesung *Nonlinear Optimization* oder *Numerische Lineare Algebra*. Beim CG-Verfahren erhält man mit

i. W. demselben Aufwand pro Iteration an Stelle von (4.25b) die Konvergenzabschätzung

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

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

Ende der Vorlesung 4

Ende der Woche 2

§ 5 Das Newton-Verfahren

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

Idee: Es sei $x^{(0)}$ die Schätzung einer Nullstelle von F. Wir legen im Punkt $x^{(0)}$ die Tangente (ein **lineares Modell**, englisch: *linear model*) 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.

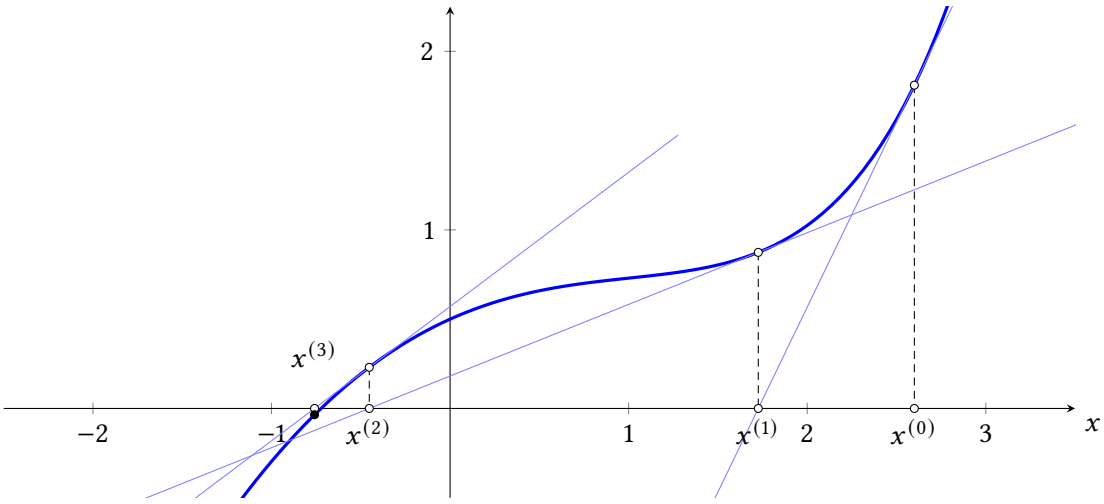


Abbildung 5.1.: Illustration des Newton-Verfahrens zur Suche einer Nullstelle der Funktion $F(x) = 0.5\exp(0.9\,x) - x^2$, ausgehend von der Startschätzung $x^{(0)} = 2.6$.

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

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

Das (lokale) Newton-Verfahren wird in Algorithmus 5.1 skizziert.

Algorithmus 5.1 (**Lokales Newton-Verfahren**[7])**.**
Eingabe: Funktion $F\colon \mathbb{R}^n \to \mathbb{R}^n$
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**[8] $d^{(k)}$
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

Wir führen die Konvergenzanalysis für das Newton-Verfahren in dieser Vorlesung nur in der Euklidischen Norm durch. Ähnlich wie beim Gradientenverfahren könnten wir analog auch hier ein benutzer-definiertes M-Innenprodukt im Raum $\mathbb{R}^n$ verwenden. Das wird in der Vorlesung *Nonlinear Optimization* nachgeholt.

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** oder **Operatornorm** (englisch: *matrix norm, operator norm*)

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|. \tag{5.1}$$

△

$\|A\|$ wird auch als **Spektralnorm** (englisch: *spectral norm*) 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 $A^T A$ (der $\geqslant 0$ ist; **Quizfrage 5.1:** Warum?). Weiter gilt $\|Ax\| \leqslant \|A\|\|x\|$ und $\|AB\| \leqslant \|A\|\|B\|$ für alle Matrizen A, B und Vektoren x passender Größe.

[7] englisch: *local Newton method*
[8] englisch: *Newton direction*

Lemma 5.3 (Banach-Lemma)**.**

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

$$\|(\mathrm{Id} - C)^{-1}\| \leqslant \frac{1}{1 - \|C\|}.$$

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

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

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

Beweis von Lemma 5.3. Aussage (*i*): Für $x \in \mathbb{R}^n$ gilt

$$\|(\mathrm{Id} - C)\,x\| = \|x - Cx\| \geqslant \|x\| - \|Cx\| \geqslant \underbrace{(1 - \|C\|)}_{>0}\,\|x\|.$$

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

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

$$\|y\| \geqslant (1 - \|C\|)\,\|x\|, \quad \text{also } \frac{\|x\|}{\|y\|} \leqslant \frac{1}{1 - \|C\|}$$

und damit

$$\|(\mathrm{Id} - C)^{-1}\| = \max_{y\neq 0} \frac{\|(\mathrm{Id} - C)^{-1}y\|}{\|y\|} = \max_{\substack{y\neq 0 \\ x=(\mathrm{Id}-C)^{-1}y}} \frac{\|x\|}{\|y\|} \leqslant \frac{1}{1 - \|C\|}.$$

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

$$(\mathrm{Id} - C)^{-1} = (BA)^{-1} = A^{-1}B^{-1},$$

also

$$A\,(\mathrm{Id} - C)^{-1} = B^{-1}.$$

Es folgt

$$\begin{aligned} \|B^{-1}\| &\leqslant \|A\|\|(\mathrm{Id} - C)^{-1}\| \\ &\leqslant \frac{\|A\|}{1 - \|C\|} && \text{wegen Aussage (i)} \\ &= \frac{\|A\|}{1 - \|\mathrm{Id} - BA\|}. \end{aligned}$$

Die andere Ungleichung folgt analog. □

Lemma 5.4 (Die Menge invertierbarer Matrizen ist offen).
Es sei $F\colon \mathbb{R}^n \to \mathbb{R}^n$ 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:

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

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

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

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

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

Nach dem Banach-Lemma 5.3 *(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

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

Expertenwissen: 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 (und die Norm der Inversen ist in einer Umgebung gleichmäßig beschränkt).

Die folgende Abschätzung sieht ähnlich aus wie eine Abschätzung des Linearisierungsfehler. Dieser würde sich ergeben, wenn x^* an Stelle von x stehen würde.

Lemma 5.5 (Abschätzung ähnlich dem Linearisierungsfehler).
Es sei $F\colon \mathbb{R}^n \to \mathbb{R}^n$ eine C^1-Funktion und $x^* \in \mathbb{R}^n$. Für alle $\varepsilon > 0$ existiert $\delta > 0$ mit

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

für alle $\|x - x^*\| \leqslant \delta$, also $x \in B_\Delta(x^*)$.

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

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

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

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

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

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

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

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

Definition 5.6 (Q-Konvergenzraten[9])**.**
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

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

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

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

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

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

Die Abschätzung (4.25a) zeigt beispielsweise die Q-lineare Konvergenz des Gradientenverfahrens bei quadratischer Zielfunktion, wobei an Stelle der Euklidischen Norm die durch die Matrix Q induzierte Norm verwendet wird.

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.)

Ende der Vorlesung 5

§ 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 5.7 (Lokaler Konvergenzsatz für das lokale Newton-Verfahren)**.**
Es sei $F\colon \mathbb{R}^n \to \mathbb{R}^n$ 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 jede Startschätzung $x^{(0)} \in B_\delta(x^*)$ gilt:

[9] Das „Q“ steht für „Quotient“. Das kommt daher, dass manche Autoren etwa die Bedingung für die Q-lineare Konvergenz in der Form $\frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|} \leqslant c$ schreiben, also den Quotienten der Fehlernorm in aufeinanderfolgenden Iterationen betrachten. Natürlich muss dann vorausgesetzt werden, dass der Fehler nicht Null wird. Unsere Formulierung vermeidet dieses Problem.

(*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

$$\|F'(x)^{-1}\| \leqslant c = 2\,\|F(x^*)^{-1}\|. \tag{5.2}$$

Erinnerung: Als Norm für Matrizen verwenden wir die durch die Euklidische Vektornorm induzierte Matrixnorm, siehe (5.1).

Nach Lemma 5.5 existiert zu $\varepsilon = 1/(2c)$ ein $\delta_2 > 0$ mit

$$\|F(x) - F(x^*) - F'(x)(x - x^*)\| \leqslant \frac{1}{2c}\|x - x^*\| \tag{5.3}$$

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

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

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

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

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

Beachte: Unter dem Integral stehen Matrizen.

Mit Hilfe der obigen Gleichung erhalten wir folgende wichtige Abschätzung:

$$\|x^{(k+1)} - x^*\| \leqslant \big\|F'(x^{(k)})^{-1}\big\| \int_0^1 \overbrace{\big\|F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\big\|}^{=:D^{(k)}(t)}\,\mathrm{d}t\,\|x^{(k)} - x^*\|. \tag{5.4}$$

Wegen $x^{(k)} \to x^*$ gilt $x^{(k)} + t\,(x^* - x^{(k)}) \to x^*$ gleichmäßig auf $t \in [0,1]$. (**Quizfrage 5.3:** Klar?) Außerdem ist F' stetig. Zu jedem $\varepsilon > 0$ existiert also ein Index $k_0 \in \mathbb{N}$ mit

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

Das bedeutet aber: $\int_0^1 \|D^{(k)}(t)\|\,\mathrm{d}t \to 0$. Jetzt liefern (5.2) und (5.4):

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

Da $\varepsilon > 0$ beliebig war, bedeutet dies die Q-superlineare Konvergenz der Iterierten.

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:

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

Aus (5.4) erhalten wir nun:

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

Bemerkung 5.8 (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** (englisch: *simplified Newton method*), 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. △

Expertenwissen: Das vereinfachte Newton-Verfahren

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

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

(*ii*) Die Konvergenzrate ist Q-linear.

Beweis. Nach Lemma 5.4 existieren $\delta_1 > 0$ und $c := 2\,\|F(x^*)^{-1}\| > 0$, sodass $F'(x)$ für alle $x \in B_{\delta_1}(x^*)$ regulär ist mit

$$\|F'(x)^{-1}\| \leqslant c,$$

wie in (5.2). Andererseits ist F' stetig in x^*, sodass $\delta_2 > 0$ existiert mit

$$\|F'(x) - F'(y)\| \leqslant \|F'(x) - F'(x^*)\| + \|F'(x^*) - F'(y)\| \leqslant \frac{1}{2c}$$

für alle $x, y \in B_{\delta_2}(x^*)$. Wir betrachten ähnlich wie im Beweis von Satz 5.7 die Gleichung für den Fehler:

$$\begin{aligned} x^{(k+1)} - x^* &= x^{(k)} - x^* - F'(x^{(0)})^{-1}\big(F(x^{(k)}) - F(x^*)\big) \\ &= F'(x^{(0)})^{-1}\big[F'(x^{(0)})(x^{(k)} - x^*) - \big(F(x^{(k)}) - F(x^*)\big)\big]. \end{aligned}$$

Unter der Voraussetzung, dass $x^{(0)}$ und $x^{(k)}$ in $B_\delta(x^*)$ mit $\delta = \min\{\delta_1, \delta_2\}$ liegen, können wir abschätzen:

$$\begin{aligned} \|x^{(k+1)} - x^*\| &\leqslant \big\|F'(x^{(0)})^{-1}\big\| \int_0^1 \big\|F'(x^{(0)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\big\|\,\mathrm{d}t\, \|x^{(k)} - x^*\| \\ &\leqslant c\,\frac{1}{2c}\,\|x^{(k)} - x^*\|. \end{aligned}$$

Mit einem einfachen Induktionsargument wie im Beweis von Satz 5.7 können wir zeigen, dass $x^{(0)} \in B_\delta(x^*)$ impliziert, dass alle $x^{(k)}$ in dieser Menge liegen. Damit ist das Verfahren wohldefiniert und, wie die Abschätzung oben zeigt, Q-linear konvergent. □

§ 5.3 Das lokale Newton-Verfahren in der Optimierung

Literatur: Geiger, Kanzow, 1999, Kapitel 9

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

$$\text{Minimiere } f(x) \text{ über } x \in \mathbb{R}^n. \tag{5.5}$$

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

(1) Die notwendige Optimalitätsbedingung 1. Ordnung für (5.5) lautet

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

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

$$x^{(k+1)} = x^{(k)} - f''(x^{(k)})^{-1}\nabla f(x^{(k)}). \tag{5.6}$$

(2) Im aktuellen Iterationspunkt $x^{(k)}$ ersetzen wir (5.5) durch die Minimierung des **quadratischen Ersatzmodells** (Taylorpolynoms)

$$m^{(k)}(x) := f(x^{(k)}) + \nabla f(x^{(k)})^\top (x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^\top f''(x^{(k)})(x - x^{(k)}). \tag{5.7}$$

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

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

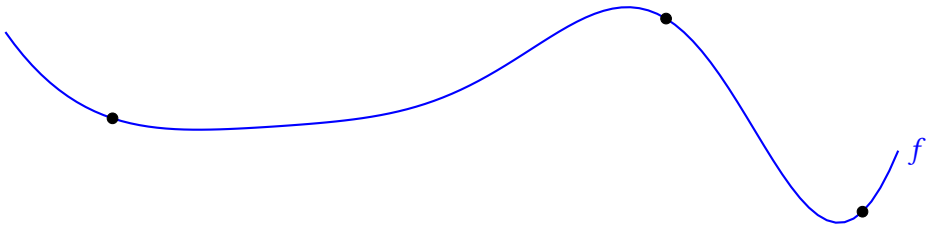


Abbildung 5.2.: Zeichnen Sie an den markierten Punkten das Taylorpolynom 2. Ordnung ein und bestimmen Sie grafisch den Newton-Schritt.

charakterisiert, vgl. (4.15). In jedem Fall erhalten wir dadurch einen stationären Punkt des quadratischen Models.

Wir wählen die Lösung dieses linearen Gleichungssystem als nächste Iterierte $x^{(k+1)}$ und erhalten wiederum die Iterationsvorschrift (5.6)

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

Zur Übung können Sie das Taylorpolynom 2. Ordnung an den markierten Punkten in Abbildung 5.2 einzeichnen und den Newton-Schritt grafisch bestimmen. (**Quizfrage 5.4:** Ergibt sich in jedem Fall ein Abstieg im Funktionswert?)

Expertenwissen: Vom Gradienten- zum Newton-Verfahren

Es gibt noch eine weitere Interpretation des Gradienten- bzw. des Newton-Verfahrens in der Optimierung. Wir betrachten dazu den M-Gradientenfluss zur Zielfunktion $f\colon \mathbb{R}^n \to \mathbb{R}$, also

$$\dot{x}(r) = -\nabla_M f(x(r)).$$

Ein Schritt des *impliziten* Euler-Verfahrens mit der Schrittweite $t^{(k)}$ führt auf die Gleichung

$$\frac{x^{(k+1)} - x^{(k)}}{t^{(k)}} = -\nabla_M f(x^{(k+1)}) \quad \Leftrightarrow \quad M\,(x^{(k+1)} - x^{(k)}) + t^{(k)}\nabla f(x^{(k+1)}) = 0$$

für die Unbekannte $x^{(k+1)}$. Wir führen nun für diese *nichtlineare* Gleichung einen einzigen Newton-Schritt mit der Startschätzung $x^{(k)}$ durch. Mit anderen Worten, wir ersetzen die nichtlineare Gleichung durch ihr Taylormodell erster Ordnung und erhalten

$$M\,(x^{(k)} - x^{(k)}) + M\,(x^{(k+1)} - x^{(k)}) + t^{(k)}\nabla f(x^{(k)}) + t^{(k)} f''(x^{(k)})(x^{(k+1)} - x^{(k)}) = 0,$$

also

$$\left[M + t^{(k)} f''(x^{(k)})\right](x^{(k+1)} - x^{(k)}) = -t^{(k)}\nabla f(x^{(k)}).$$

Für große Schrittweiten $t^{(k)} \to \infty$ konvergiert die Lösung gegen die Lösung der Newton-Gleichung (5.6). Für kleine Schrittweiten $t^{(k)} \searrow 0$ dagegen konvergiert die Lösung gegen die eines Gradientenschrittes mit Schrittweite $t^{(k)}$. Wir können also über die Schrittweite kontinuierlich zwischen Gradienten- und Newton-Verfahren hin- und herschalten.

Bemerkung 5.9 (Zum lokalen Newton-Verfahren).

(*i*) Satz 5.7 liefert die lokal Q-superlineare bzw. Q-quadratische Konvergenz von Algorithmus 5.1 mit $F(x) = \nabla f(x)$ gegen einen stationären Punkt x^* von f. Dieser kann auch ein lokaler Maximierer oder ein Sattelpunkt von f sein, da wir $f''(x^*)$ nur als regulär und nichts über die Definitheit voraussetzen.

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

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

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

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

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

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

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

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: Gilt für die Startschätzungen der Zusammenhang $x^{(0)} = Ay^{(0)} + b$, dann gilt auch $x^{(k)} = Ay^{(k)} + b$ für alle $k \in \mathbb{N}$.

Quizfrage 5.5: Stimmt diese Eigenschaft auch für das Gradientenverfahren? △

Quizfrage 5.6: Was unterscheidet ein vereinfachtes Newton-Verfahren vom Gradientenverfahren in der Optimierung?

§ 5.4 Abschliessende Bemerkungen zu Verfahren der unrestringierten Optimierung

Für den Einsatz in der Praxis ist das lokale Newton-Verfahren nicht geeignet, weil man hinreichend nah an einem (unbekannten) stationären Punkt starten muss, um überhaupt Konvergenz zu bekommen.

Idee: Kombiniere die robusten globalen Konvergenzeigenschaften des Gradientenverfahrens (Algorithmus 4.11) mit der schnellen lokalen Konvergenz des Newton-Verfahrens (Algorithmus 5.1).

Dazu wird in jedem Iterationsschritt versucht, das lineare Gleichungssystem

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

für die Newton-Richtung $d^{(k)}$ zu lösen. Ist dieses System nicht oder nicht eindeutig lösbar, so wird ersatzweise auf die (vorkonditionierte) Gradientenrichtung zurückgefallen. Andernfalls wird noch die Güte der Newton-Richtung geprüft, indem der Wert der Richtungsableitung $f'(x^{(k)})\, d^{(k)}$ mit dem für die Gradientenrichtung verglichen wird. Schneidet die Newton-Richtung dabei zu schlecht ab, wird ebenfalls die Gradientenrichtung verwendet.

Für die Details verweisen wir auf Geiger, Kanzow, 1999, Kapitel 9.2 oder auf die Vorlesung *Nonlinear Optimization*.

Im praktischen Einsatz kommt 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.

Alle in Kapitel 1 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 *Nonlinear Optimization* und Geiger, Kanzow, 1999, Abschnitt 14.

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

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

gewinnen, d. h. aus dem linearen Gleichungssystem

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

Folgende Tabelle fasst typische Eigenschaften dieser Verfahren zusammen:

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

Mehr insbesondere zu Quasi-Newton-Verfahren folgt in der Vorlesung *Nonlinear Optimization*.

Ende der Vorlesung 6

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, und diese waren auch die Motivation für ihre Entwicklung.[1] 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 der Optimierungsvariablen x, und die Grundmenge ist $\Omega = \mathbb{R}^n$.

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

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

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

Quizfrage 6.1: In der Regel setzt man den konstanten Term γ 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 $A_{\text{ineq}} x \geqslant b_{\text{ineq}}$ 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

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

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

[1] Leonid Kantorovich, Nobelpreis für Wirtschaftswissenschaften 1975

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

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

Alle diese Punkte x_α sind zulässig, denn:

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

Für die Werte der Zielfunktion gilt

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

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

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

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

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

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:

$$\left.\begin{aligned} \text{Maximiere} \quad & 9\,x_1 + 8\,x_2 \quad \text{über } x \in \mathbb{R}^2 \\ & x_1 + x_2 \leqslant 6 \quad \text{(Marzipanbedingung)} \\ \text{sodass} \quad & 2\,x_1 + x_2 \leqslant 11 \quad \text{(Nougatbedingung)} \\ & x_1 + 2\,x_2 \leqslant 9 \quad \text{(Schokoladenbedingung)} \\ \text{und} \quad & x_1 \geqslant 0 \\ & x_2 \geqslant 0. \end{aligned}\right\} \tag{6.2}$$

△

Abbildung 6.1 stellt die Aufgabe grafisch dar.

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

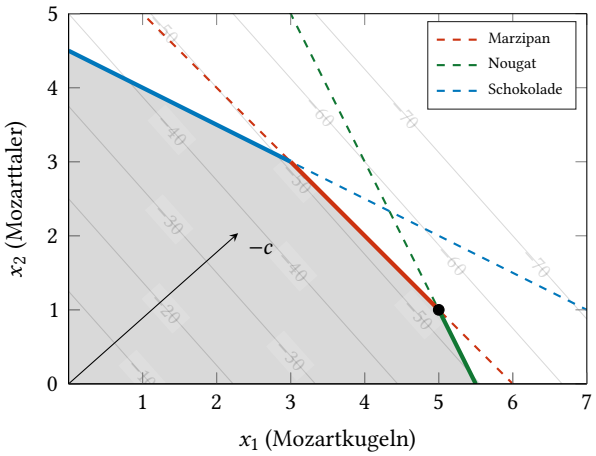


Abbildung 6.1.: Zulässige Menge (Fünfeck), Niveaulinien der Zielfunktion und globaler Maximierer beim Mozartproblem (Beispiel 6.2).

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

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

Dabei sind $m, n \in \mathbb{N}$ und

- $c \in \mathbb{R}^n$ der **Kostenvektor** (englisch: *cost vector*),
- $A \in \mathbb{R}^{m \times n}$ die **Bedarfs-** oder **Aufwandsmatrix** (englisch: *demand matrix*)
- $b \in \mathbb{R}^m$ der **Ressourcenvektor** (englisch: *resource vector*).

Ein LP der Gestalt (6.3) heißt in **kanonischer Form** (englisch: *canonical 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 $a^\mathsf{T}x = \beta$, so können wir diese in Form zweier Ungleichungen, $a^\mathsf{T}x \leqslant \beta$ und $-a^\mathsf{T}x \leqslant -\beta$, formulieren.

(*ii*) für eine Variable x_i keine Beschränkung der Form $x_i \geqslant 0$ (**freie Variable**, englisch: *free variable*), so ersetzen wir $x_i := x_i^+ - x_i^-$ und fordern $x_i^+ \geqslant 0$ und $x_i^- \geqslant 0$.

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 4.2.)

Wir wollen nun die Geometrie der zulässigen Menge einer linearen Optimierungsaufgabe näher beschreiben.

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

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

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

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

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

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

(*iii*) Der Durchschnitt endlich vieler abgeschlossener Halbräume wird als **(konvexes) Polyeder** (altgriechisch: *πολύ*: viele, altgriechisch: *ἕδρον*: Fläche, englisch: *polyhedron*) bezeichnet. △

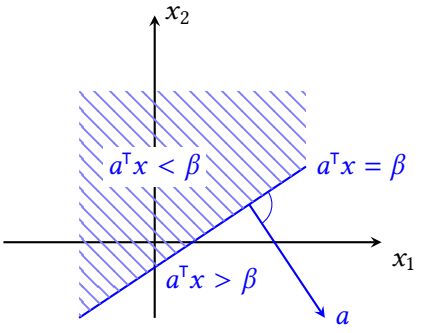


Abbildung 6.2.: Darstellung einer Hyperebene mit Normalenvektor a und der beiden Halbräume.

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

$$\{x \in \mathbb{R}^n \mid Ax \leqslant b,\ x \geqslant 0\}$$

der Aufgabe (6.3) ein Polyeder. Ein Beispiel in $\mathbb{R}^2$ war bereits in Abbildung 6.1 zu sehen.

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

Definition 6.5 (LP in Normalform)**.**

Ein LP der Gestalt

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

mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ heißt in **Normalform** (englisch: *normal form*) bzw.**Standardform** (englisch: *standard form*). △

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

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

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

(*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 $x_i^+ \geqslant 0$ und $x_i^- \geqslant 0$.

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

Mit obigen Umformungen kann man zeigen:

Lemma 6.6 (Transformierbarkeit in Normalform)**.**

Jedes LP ist äquivalent zu einem LP in Normalform.

Beispiel 6.7 (Mozartproblem in Normalform)**.**

Wir führen drei Schlupfvariablen s_1, s_2, s_3 ein:

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

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

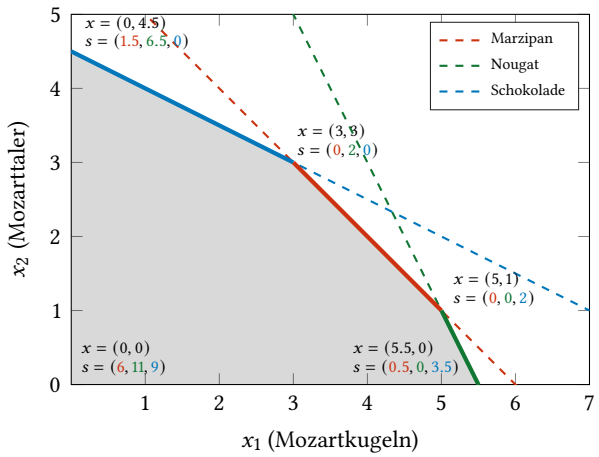


Abbildung 6.3.: Darstellung der Ecken der zulässigen Menge beim Mozartproblem in Normalform (6.7) mitsamt den jeweiligen Werten der Schlupfvariablen.

Ende der Vorlesung 7

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

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

heißt dann ein **in Normalform beschriebes Polyeder**, kurz: **Polyeder in Normalform** (englisch: *polyhedron in normal form*). Für die Dimension der Matrix $A \in \mathbb{R}^{m\times n}$ nehmen wir dabei an:[2]

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

Es gibt eine Klasse von Aufgaben, die zunächst nicht wie in lineares Optimierungsproblem aussehen, aber als ein solches geschrieben werden können. Es handelt sich dabei um Aufgaben mit ℓ_1- bzw. ℓ_∞-Normen in der Zielfunktion

Beispiel 6.8 (Aufgaben mit ℓ_1- und ℓ_∞-Normen)**.**

(*i*) Die Aufgabe

$$\text{Minimiere} \quad \|A\,x - b\|_1 \quad \text{über } x \in \mathbb{R}^n \tag{6.9}$$

kann wie folgt als lineare Optimierungsaufgabe umformuliert werden:

$$\left.\begin{aligned} \text{Minimiere} \quad & \mathbf{1}^\mathsf{T} t \quad \text{über } (x,t) \in \mathbb{R}^n \times \mathbb{R}^m \\ \text{unter} \quad & -t \leqslant A\,x - b \leqslant t \\ \text{und} \quad & t \geqslant 0 \quad \text{(redundant).} \end{aligned}\right\} \tag{6.10}$$

Anschließend kann (6.10) bei Bedarf in Normalform überführt werden. Die Aufgabe (6.10), also auch die Originalaufgabe (6.9), ist immer lösbar, da der Infimalwert $f^* \geqslant 0$ und endlich ist (Satz 6.12). (**Quizfrage 6.5:** Warum ist (6.10) immer zulässig?)

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

(*ii*) Die Aufgabe

$$\text{Minimiere} \quad \|Ax - b\|_\infty \quad \text{über } x \in \mathbb{R}^n \tag{6.11}$$

kann ebenfalls als lineare Optimierungsaufgabe umformuliert werden, und zwar in der Form

$$\left.\begin{aligned} \text{Minimiere} \quad & t \quad \text{über } (x,t) \in \mathbb{R}^n \times \mathbb{R} \\ \text{unter} \quad & -t\,\mathbf{1} \leqslant Ax - b \leqslant t\,\mathbf{1} \\ \text{und} \quad & t \geqslant 0 \quad \text{(redundant)}. \end{aligned}\right\} \tag{6.12}$$

Auch (6.12) kann bei Bedarf noch in Normalform überführt werden. Auch hier ist die Aufgabe (6.12) und damit auch die Originalaufgabe (6.11) immer lösbar, da der Infimalwert $f^* \geqslant 0$ und endlich ist (Satz 6.12). △

§ 6.1 Existenz von Lösungen

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

Vorüberlegung: Die zulässige Menge $P \subseteq \mathbb{R}^n$ von (6.6) ist immer abgeschlossen. (**Quizfrage 6.6:** Warum eigentlich?) *Falls* sie auch nichtleer und beschränkt (also kompakt) ist, dann besitzt die stetige Zielfunktion $c^\mathsf{T}x$ über P nach dem **Satz von Weierstraß** bzw. Satz 1.9 einen Minimierer. Allerdings ist die zulässige Menge P im Allgemeinen nicht beschränkt, siehe Abbildung 6.4. Mit der hinreichenden Bedingung aus Satz 1.9 können wir ebenfalls nicht argumentieren, da die Sublevelmengen $L := \{x \in P \mid c^\mathsf{T}x \leqslant m\}$ möglicherweise nicht beschränkt (also nicht kompakt) sind, wie das folgende Beispiel 6.9 zeigt.

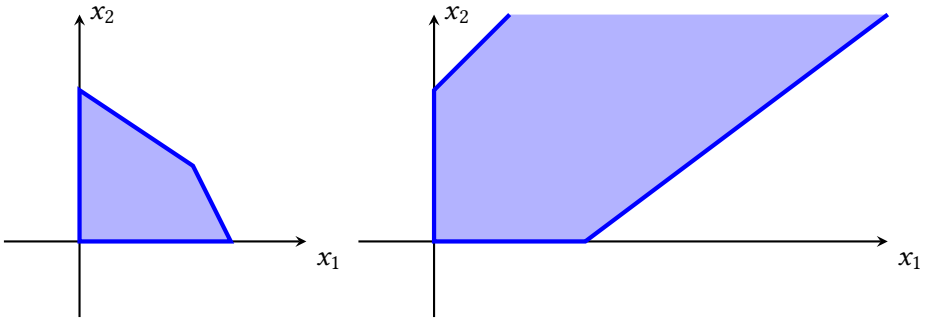


Abbildung 6.4.: Kompaktes (abgeschlossenes und beschränktes) Polyeder (links) und unbeschränktes Polyeder (rechts).

Beispiel 6.9 (Die Nicht-Kompaktheit aller nichtleeren Sublevelmengen sagt nichts über die Lösbarkeit eines LPs aus)**.**

(*i*) Bei der folgenden Aufgabe sind die Sublevelmengen $L = [0,m] \times \mathbb{R}_{\geqslant 0}$ genau für $m \geqslant 0$ nichtleer, aber nicht kompakt:

$$\begin{aligned} \text{Minimiere} \quad & \begin{pmatrix} 1 \\ 0 \end{pmatrix}^\mathsf{T} x \quad \text{über } x \in \mathbb{R}^2 \\ \text{unter} \quad & x \geqslant 0. \end{aligned}$$

Dennoch besitzt die Aufgabe eine nichtleere Lösungsmenge, und zwar $\{0\} \times \mathbb{R}_{\geqslant 0}$. Der Optimalwert ist $f^* = 0$.

(*ii*) Bei der folgenden Aufgabe sind die Sublevelmengen $L = \big[\max\{-m,\ 0\}, \infty\big) \times \mathbb{R}_{\geqslant 0}$ für alle $m \leqslant 0$ nichtleer, aber niemals kompakt:

$$\begin{aligned} \text{Minimiere} \quad & \begin{pmatrix} -1 \\ 0 \end{pmatrix}^{\mathsf{T}} x \quad \text{über } x \in \mathbb{R}^2 \\ \text{unter} \quad & x \geqslant 0. \end{aligned}$$

Diese Aufgabe ist unbeschränkt, also nicht lösbar. Der Infimalwert ist $f^* = -\infty$. △

Intuitiv können wir uns vorstellen, dass ein LP unbeschränkt ist, falls die Zielfunktion entlang eines Strahles $t \mapsto x + t\,d$ abfällt, der für alle $t \geqslant 0$ in der zulässigen Menge P bleibt. Wir betrachten dazu für einen beliebigen Punkt $x \in P$ die Menge der Richtungen möglicher Strahlen:

$$\begin{aligned} &\{d \in \mathbb{R}^n \mid x + t\,d \in P \text{ für alle } t \geqslant 0\} \\ &\quad = \{d \in \mathbb{R}^n \mid A\,(x + t\,d) = b \text{ und } (x + t\,d) \geqslant 0 \text{ für alle } t \geqslant 0\} \\ &\quad = \{d \in \mathbb{R}^n \mid A\,d = 0 \text{ und } d \geqslant 0\}. \end{aligned}$$

Beachte: Diese Menge hängt nicht davon ab, welchen Punkt $x \in P$ wir wählen!

Definition 6.10 (Kegel, Rezessionskegel)**.**

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

(*ii*) Es sei

$$P = \{x \in \mathbb{R}^n \mid A\,x = b,\ x \geqslant 0\}$$

ein Polyeder. Dann heißt die Menge

$$\operatorname{rec}(P) := \{d \in \mathbb{R}^n \mid A\,d = 0,\ d \geqslant 0\} \tag{6.13}$$

der **Rezessionskegel** (englisch: *recession cone*) von P. △

Wir formulieren unsere Überlegung zur Unbeschränktheit eines LPs als Resultat:

Lemma 6.11 (Endlichkeit des Infimalwertes und Abfall entlang von Strahlen)**.**
Wir betrachten ein LP in Normalform (6.6) mit nichtleerer zulässiger Menge $P = \{x \in \mathbb{R}^n \mid A\,x = b,\ x \geqslant 0\}$ wie in (6.8). Gilt $c^{\mathsf{T}}d < 0$ für eine Richtung $d \in \operatorname{rec}(P)$, dann ist $f^* = -\infty$.

(**Quizfrage 6.7:** Was bedeutet $c^{\mathsf{T}}d < 0$?)

Beweis. Angenommen, für $d \in \operatorname{rec}(P)$ gilt $c^{\mathsf{T}}d < 0$. Dann ist $x + t\,d$ für alle $t \geqslant 0$ zulässig, und es gilt

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

Daraus folgt $f^* = -\infty$. □

Tatsächlich gilt auch die Umkehrung der Aussage ($c^{\mathsf{T}}d \geqslant 0$ für alle $d \in \operatorname{rec}(P)$ impliziert die Endlichkeit von f^*), aber das werden wir erst später zeigen können.

Wir zeigen nun, dass LPs bereits dann lösbar sind, wenn der Infimalwert endlich ist:

Satz 6.12 (Existenzsatz für LPs)**.**
Wir betrachten ein LP in allgemeiner Form (6.1) mit zulässiger Menge F. Ist der Infimalwert

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

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

Zum Beweis von Satz 6.12 benötigen wir ein Hilfsresultat. Dieses sagt aus, dass die Menge der nicht-negativen Linearkombinationen einer gegebenen Menge von Vektoren abgeschlossen ist:

Lemma 6.13 (Abgeschlossenheit der Menge nicht-negativer Linearkombinationen[3])**.**
Es sei $B \in \mathbb{R}^{m\times n}$ eine Matrix (ohne Einschränkungen an die Dimensionen $n, m \in \mathbb{N}$). Die Menge

$$K := \left\{ B\,d \,\middle|\, d \in \mathbb{R}^n,\ d \geqslant 0 \right\} \tag{6.14}$$

der nichtnegativen Linearkombinationen der Spalten von B ist abgeschlossen.

Quizfrage 6.8: Wie kann man sich die Menge in (6.14) grafisch vorstellen?

Beweis. Wir betrachten zunächst einen einzelnen Vektor $y = B\,d \in K$, also mit $d \geqslant 0$.

Schritt 1: Wir schreiben diesen Vektor in einer alternativen Form, und zwar als $y = \alpha\, B\, d^*$. Dabei soll der Faktor $\alpha \geqslant 0$ sein sowie $\|d^*\| = 1$ und $d^* \geqslant 0$. Außerdem soll der Vektor d^* so gewählt sein, dass in der Menge $\{c \in \mathbb{R}^n_{\geqslant 0} \mid B\,c = B\,d^*\}$ kein Vektor c existiert, der eine geringe Anzahl an Nicht-Nulleinträgen aufweist als d^*.

Falls $y = 0$ ist, so wählen wir $\alpha = 0$ und $d^* = e^{(1)}$ aus der Standardbasis von $\mathbb{R}^n$. Falls $y \neq 0$ ist, so wählen wir aus der Menge $\{z \in \mathbb{R}^n_{\geqslant 0} \mid y = B\,z\}$ einen Vektor z^* mit einer minimalen Anzahl an Nicht-Nulleinträgen. Dann setzen wir $\alpha := \|z^*\|$ und $d^* := \frac{z^*}{\|z^*\|}$. Die Darstellung $y = \alpha\, B\, d^*$ hat dann die gewünschten Eigenschaften.

Es sei nun $y^{(k)}$ eine Folge in K. Aufgrund von Schritt 1 können wir von der Darstellung $y^{(k)} = \alpha^{(k)} B\, d^{(k)}$ ausgehen mit $\alpha^{(k)} \geqslant 0$ und $\|d^{(k)}\| = 1$, wobei alle $d^{(k)} \geqslant 0$ sind und die jeweils minimale Anzahl von Nicht-Nulleinträgen haben.

Um die Abgeschlossenheit zu zeigen, nehmen wir an, dass $y^{(k)} \to y$ konvergiert. Da die Folge $\left(d^{(k)}\right)$ in der Einheitssphäre liegt und diese kompakt ist, existiert eine konvergente Teilfolge $\left(d^{(k^{(\ell)})}\right)$ mit Grenzwert d. Für den Grenzwert gilt $\|d\| = 1$ und außerdem $d \geqslant 0$.

Schritt 2: Wir zeigen $B\,d \neq 0$.

Es sei $I \subseteq \{1, \ldots, n\} \neq \emptyset$ die Menge der Indizes mit $d_i > 0$. Wir wählen irgendeinen Index $k^* \in \left(k^{(\ell)}\right)$ aus, sodass $d_i^{(k^*)} > 0$ für alle $i \in I$ gilt. (**Quizfrage 6.9:** Warum ist das möglich?) Wir definieren $\mu := \min\left\{ \frac{d_i^{(k^*)}}{d_i} \,\middle|\, i \in I \right\} > 0$ und $z := d^{(k^*)} - \mu\, d$. Dann hat z mindestens

[3] Dieser Beweis folgt Kager, 2023. Für einen anderen Beweis per Induktion über die Anzahl n der Spalten siehe etwa Werner, 2007, Lemma 1.5.

einen Nulleintrag mehr als $d^{(k^*)}$. Aufgrund der Minimalitätseigenschaft von $d^{(k^*)}$ muss $B\,z \neq B\,d^{(k^*)}$ sein. Damit folgt

$$\mu\,B\,d = B\,d^{(k^*)} - B\,z \neq 0,$$

also $B\,d \neq 0$.

Schritt 3: Wir zeigen die Konvergenz von $\left(\alpha^{(k^{(\ell)})}\right)$.

Aus

$$y^{(k^{(\ell)})} = \alpha^{(k^{(\ell)})}\,B\,d^{(k^{(\ell)})} \tag{6.15}$$

folgt

$$\left\|y^{(k^{(\ell)})}\right\| = \alpha^{(k^{(\ell)})}\left\|B\,d^{(k^{(\ell)})}\right\|$$

und damit

$$\alpha^{(k^{(\ell)})} = \frac{\left\|y^{(k^{(\ell)})}\right\|}{\left\|B\,d^{(k^{(\ell)})}\right\|},$$

zumindest für alle hinreichend großen $\ell \in \mathbb{N}$, da wir ja in Schritt 2 $B\,d \neq 0$ gezeigt haben. Durch Grenzübergang folgt

$$\alpha \coloneqq \lim_{\ell\to\infty} \alpha^{(k^{(\ell)})} = \frac{\|y\|}{\|B\,d\|}.$$

Schritt 4: Wir zeigen schließlich $y \in K$.

Aus (6.15) erhalten wir durch Grenzübergang $\ell \to \infty$:

$$y = \alpha\,B\,d = B\,(\alpha\,d)$$

mit $\alpha \geqslant 0$, $d \geqslant 0$ und $\|d\| = 1$. Das heißt aber: $y \in K$. □

Wir können nun den Existenzsatz 6.12 beweisen.

Beweis von Satz 6.12. 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 6.10:** Warum können wir von einem LP in Normalform ausgehen?)

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

Wir betrachten die Folge

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

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

$$\widetilde{K} \coloneqq \left\{ \begin{pmatrix} c^\mathsf{T}z \\ A\,z \end{pmatrix} - \begin{pmatrix} 0 \\ b \end{pmatrix} \middle| \, z \geqslant 0 \right\} = \left\{ \begin{pmatrix} c^\mathsf{T} \\ A \end{pmatrix} z \middle| \, z \geqslant 0 \right\} - \begin{pmatrix} 0 \\ b \end{pmatrix} \subseteq \mathbb{R}^{m+1}.$$

Diese Menge ist, abgesehen von der Verschiebung um den konstanten Vektor $\left(\begin{smallmatrix} 0 \\ b \end{smallmatrix}\right)$, von der Bauart (6.14) mit der Matrix $B = \left(\begin{smallmatrix} c^\mathsf{T} \\ A \end{smallmatrix}\right) \in \mathbb{R}^{(1+m)\times n}$. Nach Lemma 6.13 ist $\widetilde{K}$ abgeschlossen. Daraus folgt, dass der Grenzwert $\left(\begin{smallmatrix} f^* \\ 0 \end{smallmatrix}\right)$ der Folge $\left(y^{(k)}\right)$ in $\widetilde{K}$ liegt. Das heißt, es existiert ein $x^* \geqslant 0$ mit der Eigenschaft $c^\mathsf{T}x^* = f^*$ und $A\,x^* - b = 0$. Damit ist x^* eine Lösung des LPs (6.6). □

Bemerkung 6.14 (zum Existenzsatz 6.12)**.**
Es ist eine durchaus bemerkenswerte Eigenschaft linearer Optimierungsaufgaben, dass sie bereits dann eine optimale Lösung besitzen, wenn der Infimalwert endlich ist. Wie wir aus Beispielen wie

$$\text{Minimiere} \quad 1/x \quad \text{über } x \in [1, \infty)$$

wissen, ist das für nichtlineare Aufgaben i. A. nicht der Fall. △

Ende der Vorlesung 8

Ende der Woche 4

§ 6.2 Die Bedeutung der Ecken

Definition 6.15 (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), wenn aus

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

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

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

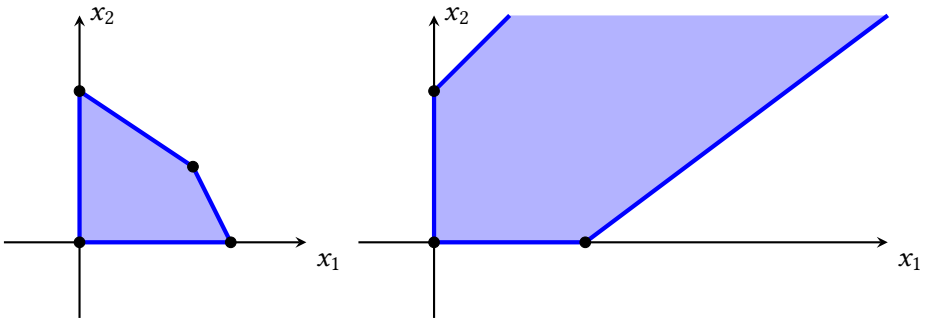


Abbildung 6.5.: Polyeder (nicht in Normalform) und ihre Ecken: Viereck im $\mathbb{R}^2$ und unbeschränktes Polyeder mit drei Ecken im $\mathbb{R}^2$.

Ecken eines Polyeders in Normalform können wie folgt charakterisiert werden:

Satz 6.16 (Charakterisierung der Ecken eines Polyeders in Normalform)**.**
Es sei P ein Polyeder in Normalform wie in (6.8) und $x \in P$ gegeben. Ferner sei $\mathcal{I}(x) = \{1 \leqslant i \leqslant n \,|\, x_i > 0\}$ die Menge der **inaktiven Indizes** (bzgl. der Ungleichungen $x \geqslant 0$). Dann sind äquivalent:

(*i*) x ist eine Ecke von P

(*ii*) Die Menge der Spalten $(a^{(i)})_{i \in \mathcal{I}(x)}$ von A ist linear unabhängig.

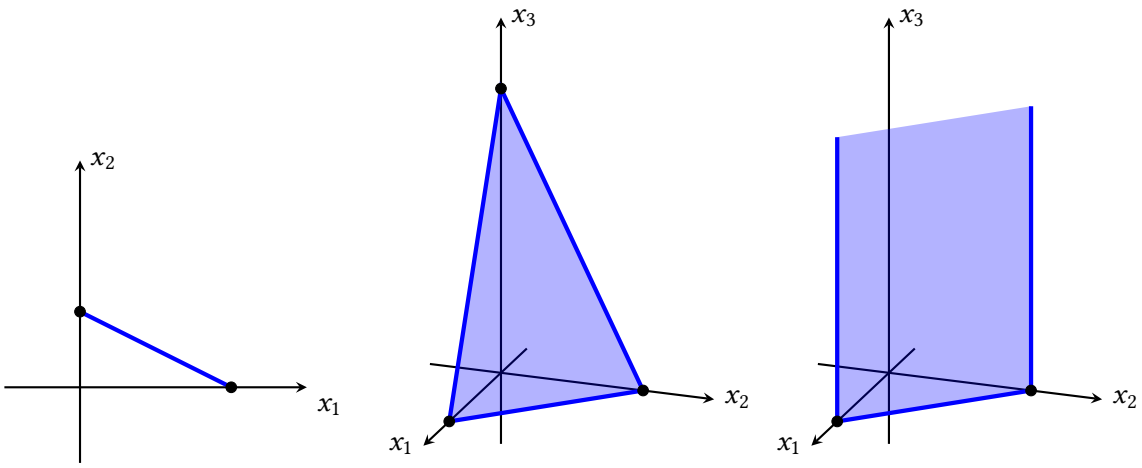


Abbildung 6.6.: Polyeder in Normalform und ihre Ecken: Strecke im $\mathbb{R}^2$ ($n = 2$, $m = 1$) und Flächen im $\mathbb{R}^3$ ($n = 3$, $m = 1$).

Beachte: Insbesondere ist $x = 0$, sofern zu P gehörig (also für $b = 0$), wegen $\mathcal{I}(x) = \emptyset$ immer eine Ecke.

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

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

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

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

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

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

also liegt $y \in P$. Ganz analog folgt auch $z \in P$. Dies ist aber ein Widerspruch zur Definition 6.15 einer Ecke, denn es gilt $x = \frac{y+z}{2}$ mit $y \neq z$.

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

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

und aus der linearen Unabhängigkeit der $(a^{(i)})_{i \in \mathcal{I}(x)}$ folgt $y_i = z_i$ auch für $i \in \mathcal{I}(x)$. Insgesamt gilt also $y = z$, d. h., nach Definition 6.15 ist x eine Ecke von P. □

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

Aus Satz 6.16 ergibt sich folgende Idee zur Generierung potentieller Ecken:

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

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

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

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

Definition 6.17 (Basisvektor, Basis)**.**
Es sei P wie in (6.8) ein Polyeder in Normalform. Weiter sei B mit $\#B = m$ ein m-Tupel (eine „Indexmenge mit Reihenfolge“) von m verschiedenen Spaltenindizes aus $\{1, \ldots, n\}$. Schließlich sei N ein zu B komplementäres $(n - m)$-Tupel, sodass also jeder Index aus $\{1, \ldots, n\}$ entweder in B oder in N genau einmal vorkommt.[4]

(i) 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*).

(ii) 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.

(iii) 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 $A\,x = b$ löst. Dieser Praxis folgen wir hier nicht, um Verwechslungen mit optimalen Lösungen zu vermeiden.

(iv) Ein Basisvektor heißt **zulässig** (englisch: *feasible basic vector*), wenn $x_B \geqslant 0$ gilt.

(v) Ist x ein Basisvektor zur Basis B, dann heißen die Komponenten von x_B **abhängige Variablen** (englisch: *dependent variables*) und die Komponenten von x_N **unabhängige Variablen** (englisch: *independent variables*). △

Beachte: Damit überhaupt eine Basis existiert, muss notwendig A vollen Rang haben, also $\operatorname{Rang}(A) = m$ gelten. Das heißt, die Zeilen von A müssen linear unabhängig sein. Dies kann zumindest theoretisch immer durch Streichen von Zeilen erreicht werden.

[4] Wir arbeiten im Kontext der Beschreibung von Ecken immer mit solchen komplementären Tupeln B und N. Wir verwenden für diese auch die von Mengen bekannte Operation $\setminus$, um das Entfernen eines Mitglieds (und Aufrücken der anderen) zu notieren sowie $\cup$, um ein Mitglied hinten anzufügen.

Beispiel 6.18 (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 $x = (2, 0, 0, 2, 6, 0, 3)^\mathsf{T}$ ist zulässiger Basisvektor zur Basis $B = (1, 4, 5, 7)$, denn: Die Untermatrix

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

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

Satz 6.19 (Zusammenhang zwischen Ecken und zulässigen Basisvektoren)**.**

Es sei P wie in (6.8) ein Polyeder in Normalform, und es gelte $\operatorname{Rang}(A) = m$. Dann sind äquivalent:

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

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

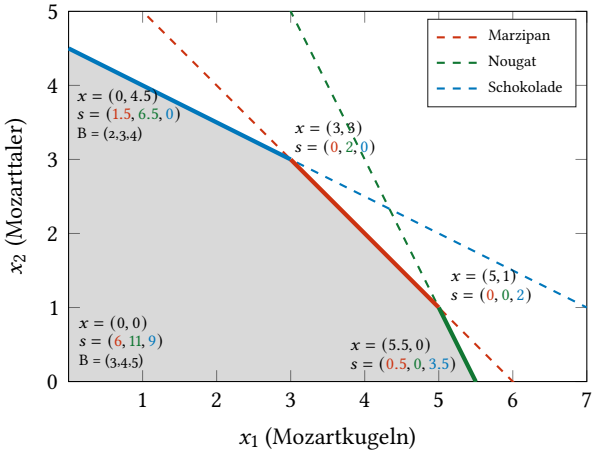
Beweis. Aussage (*i*) ⇒ Aussage (*ii*): Es sei x eine Ecke von P. Wie in Satz 6.16 sei $\mathcal{I}$ eine Familie mit Elementen aus $\{1, \dots, n\}$, die die inaktiven Indizes bzgl. der Ungleichungen $x \geqslant 0$ genau einmal enthält. Die Spalten der Teilmatrix $A_\mathcal{I}$ sind dann nach Satz 6.16 linear unabhängig. Damit muss notwendig $\#\mathcal{I} \leqslant m$ gelten. Im Fall $\#\mathcal{I} < m$ kann die Familie $\mathcal{I}$ zu einer Basis B ergänzt werden, sodass also die Untermatrix A_B regulär ist. Nach Konstruktion gilt $x_N = 0$, $A_B\, x_b = b$ und $x_B \geqslant 0$, da x Element von P ist. Damit ist bestätigt, dass x ein Basisvektor ist.

Aussage (*ii*) ⇒ Aussage (*i*): Umgekehrt sei nun x ein Basisvektor zur Basis B. Wir setzen wir oben $\mathcal{I}$ als eine Familie mit Elementen aus $\{1, \dots, n\}$, die die inaktiven Indizes bzgl. der Ungleichungen $x \geqslant 0$ genau einmal enthält. Da $\mathcal{I}$ eine Teilfamilie von B ist, folgt, dass die Spalten $A_\mathcal{I}$ linear unabhängig sind. Satz 6.16 zeigt nun, dass x eine Ecke ist. □

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

Beispiel 6.20 (Ecken beim Mozartproblem)**.**

Wir betrachten die zulässige Menge des Mozartproblems in Normalform, vgl. Abbildung 6.3:



Für einige Ecken ist bereits die Basis angegeben. (**Quizfrage 6.11:** Finden Sie die Basis der anderen Ecken. Ist die Basis jeweils eindeutig?)

Für die zum Mozartproblem in Normalform (6.7) gehörige Matrix

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 1 \end{bmatrix}$$

ergibt sich für *jede* Auswahl von 3 Spalten eine reguläre Untermatrix. Es gibt also $\binom{5}{3} = 10$ verschiedene Basen. **Quizfrage 6.12:** Wo sind die anderen 5 Basisvektoren?

Quizfrage 6.13: Wie könnte man die Aufgabe verändern, damit eine Ecke entsteht, die ein Basisvektor zu verschiedenen Basen ist (**entarteter Basisvektor**)? △

Satz 6.21 (Hauptsatz der linearen Optimierung)**.**
Es sei P wie in (6.8) ein Polyeder in Normalform, und es gelte $\operatorname{Rang}(A) = m$. Dann gilt:

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

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

(*iii*) Besitzt das Problem

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

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

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

Beweis. Aussage (*i*): Zunächst stellen wir fest, dass es mindestens eine Basis gibt, da $\operatorname{Rang}(A) = m$ gilt. Gehört der Nullvektor zu P, dann ist er ein zulässiger Basisvektor zu jeder Basis. Andernfalls wählen wir ein $x^* \in P$ mit der minimalen Anzahl positiver Komponenten. Die Indexmenge $\mathcal{I}(x^*) = \{1 \leqslant i \leqslant n \mid x_i^* > 0\}$ ist nicht leer. Wir zeigen, dass die Spaltenvektoren $(a^{(i)})$, $i \in \mathcal{I}(x^*)$ linear unabhängig sind. Nach Satz 6.16 ist dann x^* eine Ecke, und nach Satz 6.19 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 6.13 $\delta = \min \left\{ -\frac{x_i^*}{\gamma_i} \,\middle|\, \gamma_i < 0,\ i \in \mathcal{I}(x^*) \right\} > 0$ wählen. Daraus folgt, dass

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

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 6.16), 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.[5] Zu jeder Basis gehört nur genau ein Basisvektor (der auch unzulässig sein kann).

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

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

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

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

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

Ist dagegen $\widehat{P} \subsetneq P$, so gilt Rang $\left(\left[\begin{smallmatrix} A \\ c^\mathsf{T} \end{smallmatrix}\right]\right) = m+1$.[6] Nach Aussage *(i)* besitzt $\widehat{P}$ mindestens einen zulässigen Basisvektor x^*, der nach Satz 6.19 eine Ecke von $\widehat{P}$ ist. Es bleibt noch zu zeigen, dass x^* auch Ecke von P ist. Es seien also $y, z \in P$ und $\alpha \in (0,1)$, sodass $x^* = \alpha\, y + (1-\alpha)\, z$ gilt.

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

Also gilt $c^\mathsf{T} y = c^\mathsf{T} z = f^*$, d. h., $y, z \in \widehat{P}$. Da x^* eine Ecke von $\widehat{P}$ ist, muss $y = z$ gelten. Damit ist x^* eine Ecke von P und nach Satz 6.19 auch zulässiger Basisvektor von P, und wegen $x^* \in \widehat{P}$ ist x^* eine Lösung des LPs. □

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

Das folgende Beispiel zeigt, dass die Voraussetzung der Normalform für die Aussagen des Hauptsatzes 6.21 wesentlich ist.

Beispiel 6.22 (die Voraussetzung der Normalform ist wesentlich für den Hauptsatz 6.21)**.**
Betrachte das LP

$$\begin{aligned} \text{Minimiere} \quad & x_1 \quad \text{über } x \in \mathbb{R}^2 \\ \text{unter} \quad & x_1 \geqslant 0. \end{aligned}$$

Diese Aufgabe hat genau die Punkte $\{0\}\times\mathbb{R}$ als Lösungen, aber keiner davon ist eine Ecke der zulässigen Menge. Die zulässige Menge besitzt keine Ecken. △

Bemerkung 6.23. Hat man ein Polyeder in Normalform gegeben, dann kann man zu jeder seiner Ecken einen Kostenvektor c finden, sodass diese Ecke die einzige Lösung des LPs ist. △

Ende der Vorlesung 9

§ 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 Basismengen in genau einem Index unterscheiden.

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

§ 7.1 Der Simplex-Schritt

Literatur: Geiger, Kanzow, 2002, Kapitel 3.2

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

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

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

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

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

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

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

und

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

Die Größe $\widetilde{c}_N$ bezeichnet man als den Vektor der **reduzierten Kosten** (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.

Expertenwissen: Das reduzierte LP

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. Wie in (7.3) berechnet, können wir die Zielfunktion für jedes $x \in \mathbb{R}^n$ ausdrücken als

$$c^\mathsf{T} x = \underbrace{\left(c_N - A_N^\mathsf{T} A_B^{-\mathsf{T}} c_B\right)^\mathsf{T}}_{\widetilde{c}_N} x_N + \underbrace{c_B^\mathsf{T} A_B^{-1} b}_{\text{const}}.$$

Durch Elimination von x_B können wir ein zum Ausgangsproblem äquivalentes, reduziertes Problem formulieren:

$$\begin{aligned} \text{Minimiere} \quad & \widetilde{c}_N^\mathsf{T} x_N \quad \text{über } x_N \in \mathbb{R}^{n-m} \\ \text{unter} \quad & A_B^{-1} A_N x_N \leqslant A_B^{-1} b \quad (\text{und damit } x_B = A_B^{-1}(b - A_N x_N) \geqslant 0) \\ \text{und} \quad & x_N \geqslant 0. \end{aligned}$$

Der Kostenvektor $\widetilde{c}_N$ dieser reduzierten Aufgabe ist gerade der Vektor, der als reduzierte Kosten bezeichnet wurde.

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

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

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

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

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

Da $z_N \geqslant 0$ ist, gilt $c^\mathsf{T}z \geqslant c^\mathsf{T}x$. Da z ein beliebiger zulässiger Punkt war, ist x ein Minimierer der Aufgabe (6.6). □

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

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

Es ergibt sich die Frage, wie groß t werden darf, sodass $x_B(t)$ noch zulässig, also $x_B(t) \geqslant 0$ 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 $\Delta x_B \geqslant 0$, so ist das LP (6.6) unbeschränkt, also nicht lösbar.

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

Es gilt nach (7.1) und (7.2):

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

Beachte: Das ist genau die Situation, die in Lemma 6.11 beschrieben wird: Die Richtung $d = \left(\begin{smallmatrix} \Delta x_B \\ e_r \end{smallmatrix}\right)$ liegt im Rezessionskegel $\operatorname{rec}(P)$ der zulässigen Menge P 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

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

gilt oder äquivalent dazu:

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

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

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

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

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

(*i*) die Berechnung der reduzierten Kosten $\widetilde{c}_N$ nach (7.4), d. h., Lösung eines linearen Gleichungssystems mit A_B^T,

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

(*iii*) die Bestimmung des Vektors Δx_B nach (7.1), d. h., Lösung eines linearen Gleichungssystem mit A_B

(*iv*) die Bestimmung der Schrittlänge $\hat{t}$ nach (7.5)

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

Satz 7.3 (Simplex-Schritt)**.**
Es sei x ein zulässiger Basisvektor von P zur Basis B, und es sei N die zugehörige Nichtbasis. Es gelte $\widetilde{c}_r < 0$ für mindestens ein $r \in N$, und es sei $\Delta x_B := -A_B^{-1} a_r$. Es gelte weiter $\Delta x_i < 0$ für mindestens ein $i \in B$. Wird dann $\hat{t} \geqslant 0$ nach (7.5) bestimmt und wird das Minimum für den Index $\ell \in B$ angenommen, so gelten für den Vektor x^+ mit

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

die folgenden Aussagen:

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

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

(*ii*) Für die Zielfunktionswerte gilt

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

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

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

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

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

(d) A_{B^+} ist regulär.

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

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

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

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

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

Die Aussage *(ii)* folgt aus (7.3). □

Bemerkung 7.4 (Der Fall $c^\mathsf{T} x^+ = c^\mathsf{T} x$)**.**

Es gelten die Voraussetzungen von Satz 7.3.

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

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

gelten. Ein Basisvektor x, für den (7.6) zutrifft, heißt **entartet** (englisch: *degenerate*).[7]

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

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

Der Zielfunktionswert nimmt dann also strikt ab. △

Beispiel 7.5 (Nochmal Beispiel 6.18)**.**

Wir führen einen Simplex-Schritt für Beispiel 6.18 durch, ausgehend vom (zulässigen) Basisvektor $x = (2, 0, 0, 2, 6, 0, 3)^\mathsf{T}$ zur Basis $B = (1, 4, 5, 7)$. Der Zielfunktionswert ist $c^\mathsf{T} x = (-2, -3, -4, 0, 0, 0, 0)\, x = -4$.

[7] 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

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

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

(*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:

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

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

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

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

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

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

§ 7.2 Der Simplex-Algorithmus

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

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

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

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

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

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

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

12: Setze

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

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

Quizfrage 7.2: 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?

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

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

Beweis. Nach Bemerkung 7.4 Punkt (*ii*) gilt $c^\mathsf{T} x^+ < c^\mathsf{T} x$ für alle Iterierten. Daher kann kein Basisvektor mehrfach im Verfahren auftreten. Da es nach Satz 6.21 nur endlich viele zulässige Basisvektoren gibt, muss das Verfahren in Zeile 4 oder in Zeile 9 abbrechen. □

Der Simplex-Algorithmus 7.6 lässt noch Freiheiten

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

vgl. Beispiel 7.5. Durch eine geeignete Zusatzregel 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

$$\begin{aligned} & x^{(k)} = x^{(k+1)} = \ldots = x^{(k+p)} \\ \text{und} \quad & B^{(k)} \rightsquigarrow B^{(k+1)} \rightsquigarrow \ldots \rightsquigarrow B^{(k+p)} = B^{(k)} \end{aligned}$$

gilt. Man kann zeigen, dass Zyklen mindestens die Länge $p = 3$ haben.

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

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

Bemerkung 7.9 (Alternativer Beweis von Satz 6.12)**.**
Der Simplex-Algorithmus in Verbindung mit der Regel von Bland bietet eine konstruktive Möglichkeit, den Existenzsatz 6.12 zu beweisen. △

Quizfrage 7.3: 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 7.4: Was könnte der Grund sein, warum man im Simplex-Verfahren mit benachbarten Ecken arbeitet? Könnte man nicht auch größere Änderungen in den Basis-Indizes zulassen?

Quizfrage 7.5: Ist das Simplex-Verfahren ein Abstiegsverfahren? Wenn ja, können Sie die Schritte (1) bis (3) eines allgemeinen Abstiegsverfahrens (siehe Anfang von § 4 auf Seite 19) im Simplex-Verfahren (Algorithmus 7.6) wiederfinden?

Finden der ersten Ecke

Beachte: Für den Start des Simplex-Algorithmus 7.6 muss ein zulässiger Basisvektor des in Normalform beschriebenen Polyeders P bekannt sein.

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

$$\left.\begin{aligned} \text{Maximiere} \quad & c^\mathsf{T} x \\ \text{sodass} \quad & A x \leqslant b \\ \text{und} \quad & x \geqslant 0 \end{aligned}\right\}$$

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

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

Falls $b \geqslant 0$ ist, dann ist $\binom{0}{b}$ 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.) $b \geqslant 0$.[8] Dann gelten für das Hilfsproblem

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

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

- (i) Der Vektor $\binom{x}{z} = \binom{0}{b}$ 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 $\binom{x^*}{z^*}$ 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 $\operatorname{Rang}(A) = m$, so ist x^* ein (zulässiger) Basisvektor für (6.6) zu einer geeigneten Basis.

Beachte: Die Voraussetzung $b \geqslant 0$ ist keine Einschränkung, ggf. multiplizieren wir betreffende Zeilen von $A x = b$ mit -1.

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

Aussage (ii): Wegen $z \geqslant 0$ ist die Zielfunktion $\mathbf{1}^\mathsf{T} z = \sum_{i=1}^m z_i$ über der zulässigen Menge selbst $\geqslant 0$, d. h., (7.7) ist nicht unbeschränkt. Aus Satz 6.12 folgt die Existenz einer Lösung.

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

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

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

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

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

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

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

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

Quizfrage 7.6: Wieviele Iterationen benötigt das Simplex-Verfahren in Phase I mindestens, um einen zulässigen Basisvektor der Aufgabe (6.6) zu einer Basis in $\{1, \ldots, n\}$ zu finden?

Expertenwissen: Zur Komplexität des Simplex-Verfahrens

Es gibt ein konstruiertes Beispiel von Klee, Minty, 1972 (siehe z. B. Hamacher, Klamroth, 2006, S.81), 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 Anzahl der Iterationen nicht polynomial in n. Dies ist eine Motivation für sogenannte Innere-Punkte-Verfahren. Im Mittel verhält sich das Simplex-Verfahren jedoch deutlich besser als in diesem schlechtesten Fall. Die probabilistische Laufzeitanalyse für das Simplex-Verfahren geht auf https://de.wikipedia.org/wiki/Karl_Heinz_Borgwardt zurück.

Ende der Vorlesung 10

Ende der Woche 5

§ 8 Optimalitätsbedingungen der linearen Optimierung (Dualität)

Literatur: Geiger, Kanzow, 2002, Kapitel 3.1.2

Bei der Untersuchung einiger Optimierungsprobleme spielt das Konzept der **Dualität** eine wichtige Rolle. Jedem Optimierungsproblem der Form

$$\left.\begin{array}{rll} \text{Minimiere} & f(x) & \text{über } x \in \mathbb{R}^n \\ \text{sodass} & g_i(x) \leqslant 0 & \text{für } i \in \{1,\dots,p\} \\ \text{und} & h_j(x) = 0 & \text{für } j \in \{1,\dots,m\}. \end{array}\right\}, \tag{8.1}$$

welches in diesem Kontext als das **primale Problem** bezeichnet wird, wird hierbei ein dazugehöriges duales Problem gegenübergestellt, aus dem sich (unter passenden Zusatzbedingungen) häufig wichtige Informationen zu dem primalen Problem generieren lassen. Für lineare Optimierungsprobleme lassen sich hierbei besonders starke Resultate *ohne* weitere Zusatzannahmen zeigen.

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

Für die Herleitung des dualen Problems beginnen wir mit der Beobachtung, dass das Problem (8.1) äquivalent ist zu dem Problem

$$\inf_{x\in\mathbb{R}^n} \sup_{\substack{\lambda\in\mathbb{R}^m \\ \mu\in\mathbb{R}^p_{\geqslant 0}}} \quad f(x) + \lambda^\mathsf{T} h(x) + \mu^\mathsf{T} g(x), \tag{8.2}$$

in dem die Größen $\lambda \in \mathbb{R}^m$ und $\mu \in \mathbb{R}^p$ mit $\mu \geqslant 0$ die Rolle von Strafparametern übernehmen – mit betragsmäßig wachsenden λ und μ können in der modifizierten Zielfunktion für unzulässige Punkte beliebig große Werte erzeugt werden. [9] Als **duales Problem** bezeichnet man dasjenige Optimierungsproblem, das sich ergibt, wenn man in (8.3) die Reihenfolge der Infimierung und Supremierung vertauscht, also das Problem

$$\sup_{\substack{\lambda\in\mathbb{R}^m \\ \mu\in\mathbb{R}^p_{\geqslant 0}}} \inf_{x\in\mathbb{R}^n} \quad f(x) + \lambda^\mathsf{T} h(x) + \mu^\mathsf{T} g(x). \tag{8.3}$$

Um den Zusammenhang des primalen und des dualen Problems weiter zu untersuchen benötigt es dann die konkrete Problemstruktur, in unserem Anwendungsfall also die (affin) linearen Funktionen eines LPs in Normalform der Form

$$\left.\begin{array}{rll} \text{Minimiere} & c^\mathsf{T}x & \text{über } x \in \mathbb{R}^n \\ \text{sodass} & Ax = b & \\ \text{und} & x \geqslant 0 & \end{array}\right\} \tag{8.4}$$

mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$, also $f(x) = c^\mathsf{T}x$, $h(x) = b - Ax$ und $g(x) = -x$. Hier ergeben sich direkt das primale- und duale Problem

$$\inf_{x\in\mathbb{R}^n} \sup_{\substack{\lambda\in\mathbb{R}^m \\ \mu\in\mathbb{R}^p_{\geqslant 0}}} \quad c^\mathsf{T}x + \lambda^\mathsf{T}(b - Ax) + \mu^\mathsf{T}(-x),$$

$$\sup_{\substack{\lambda\in\mathbb{R}^m \\ \mu\in\mathbb{R}^p_{\geqslant 0}}} \inf_{x\in\mathbb{R}^n} \quad b^\mathsf{T}\lambda + (-A^\mathsf{T}\lambda - \mu + c)^\mathsf{T}x.$$

[9] Die modifizierte Zielfunktion $f(x) + \lambda^\mathsf{T} h(x) + \mu^\mathsf{T} g(x)$ wird auch als **Lagrangefunktion** bezeichnet.

Im Infimierungsprozess können hier für diejenigen λ, μ, für die $-A^\mathsf{T} - \mu + c \neq 0$ ist, wieder durch betragsmäßig wachsende $x \in \mathbb{R}^n$ beliebig kleine Werte erzeugt werden. Das duale Problem lässt sich also äquivalent als das duale Problem

$$\left.\begin{array}{rl} \text{Maximiere} & b^\mathsf{T}\lambda \quad \text{über } (\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^n \\ \text{sodass} & A^\mathsf{T}\lambda + \mu = c \\ \text{und} & \mu \geqslant 0. \end{array}\right\} \tag{8.5}$$

formulieren.

Beachte: Das LP (8.5) liegt nicht in Normalform vor, da die Bedingung $\lambda \geqslant 0$ fehlt (und die Zielfunktion maximiert wird). Die Komponenten von $\mu \in \mathbb{R}^n$ werden als die **dualen Schlupfvariablen** (englisch: *dual slack variables*) bezeichnet. Sie sind die Slackvariablen der äquivalenten Formulierung

$$\left.\begin{array}{rl} \text{Maximiere} & b^\mathsf{T}\lambda \quad \text{über } \lambda \in \mathbb{R}^m \\ \text{sodass} & A^\mathsf{T}\lambda \leqslant c. \end{array}\right\} \tag{8.6}$$

Definition 8.1 (Duales LP)**.**

Das LP (8.5) bzw. (8.6) heißt das zu (8.4) gehörige **duale LP** (englisch: *dual LP*). In diesem Zusammenhang heißt (8.4) das **primale LP** (englisch: *primal LP*). Man spricht auch von **primal-dualen Paaren** (englisch: *primal-dual pair*). △

Beispiel 8.2 (Duales Mozartproblem)**.**

Die zum Mozartproblem in Normalform

$$\begin{aligned} \text{Minimiere} \quad & \begin{pmatrix} -9, & -8, & 0, & 0, & 0 \end{pmatrix} x \\ \text{sodass} \quad & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 1 \end{bmatrix} x = \begin{pmatrix} 6 \\ 11 \\ 9 \end{pmatrix} \\ \text{und} \quad & x \geqslant 0 \end{aligned}$$

zugehörige duale Aufgabe lautet

$$\begin{aligned} \text{Maximiere} \quad & \begin{pmatrix} 6, & 11, & 9 \end{pmatrix} \lambda + \begin{pmatrix} 0, & 0, & 0 \end{pmatrix} \mu \\ \text{sodass} \quad & \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \lambda + \mu = \begin{pmatrix} -9 \\ -8 \\ 0 \\ 0 \\ 0 \end{pmatrix} \\ \text{und} \quad & \mu \geqslant 0. \end{aligned}$$

△

Quizfrage 8.1: Was ist die duale Aufgabe der dualen Aufgabe (8.5)? (Siehe auch Hausaufgabe 6.1.)

Ziel: Verständnis des Zusammenhangs von (8.4) und (8.5)

Wir bezeichnen wie bisher auch den Infimalwert von (8.4) mit f^* und den **Supremalwert** (englisch: *supremal value*) von (8.6) bzw. von (8.5) mit d^*:

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

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

Wir beweisen einen ersten Zusammenhang zwischen primalen und dualen LPs.

Satz 8.3 (**Schwache Dualität** (englisch: *weak duality*))**.**
Es sei $x \in \mathbb{R}^n$ zulässig für das primale LP (8.4), und es sei $(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^n$ zulässig für das duale LP (8.5). Dann gilt für die Funktionswerte

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

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

Beweis. Aus der Zulässigkeit ergibt sich

$$b^\mathsf{T}\lambda = (A\,x)^\mathsf{T}\lambda = x^\mathsf{T}(A^\mathsf{T}\lambda) = x^\mathsf{T}(c - \mu) = c^\mathsf{T} x - x^\mathsf{T}\mu \leqslant c^\mathsf{T} x, \tag{8.7}$$

denn wegen $x \geqslant 0$ und $\mu \geqslant 0$ gilt $x^\mathsf{T}\mu \geqslant 0$. □

Folgerung 8.4 (Erkennen primal-dualer Lösungen durch gleiche Zielfunktionswerte)**.**
Es sei $x^* \in \mathbb{R}^n$ zulässig für das primale LP (8.4), und es sei $(\lambda^*, \mu^*) \in \mathbb{R}^m \times \mathbb{R}^n$ zulässig für das duale LP (8.5). Falls

$$c^\mathsf{T} x^* = b^\mathsf{T}\lambda^* \tag{8.8}$$

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

Beweis. Es seien x und (λ, μ) irgendwelche zulässigen Punkte für das primale bzw. das duale LP. Aus der schwachen Dualität (Satz 8.3) folgt

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

d. h., x^* ist eine Lösung von (8.4), und (λ^*, μ^*) ist Lösung von (8.5). □

Der folgende Satz zeigt, dass das System

$$\left.\begin{aligned} A^\mathsf{T}\lambda + \mu = c, \quad \mu \geqslant 0 \qquad & \text{duale Zulässigkeit} \\ A\,x = b, \quad x \geqslant 0 \qquad & \text{primale Zulässigkeit} \\ x_i\,\mu_i = 0, \quad i = 1, \ldots, n \qquad & \text{Komplementarität} \end{aligned}\right\} \tag{8.9}$$

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)**.**

(*i*) Ist x^* eine Lösung für das primale LP (8.4), dann existieren (λ^*, μ^*), sodass (x^*, λ^*, μ^*) das System (8.9) erfüllt.

(*ii*) Ist (λ^*, μ^*) eine Lösung für das duale LP (8.5), dann existiert x^*, sodass (x^*, λ^*, μ^*) das System (8.9) erfüllt.

(*iii*) Erfüllt (x^*, λ^*, μ^*) das System (8.9), dann ist x^* eine Lösung von (8.4), und (λ^*, μ^*) ist eine Lösung von (8.5).

In jedem Fall sind der Infimalwert des primalen LPs und der Supremalwert des dualen LPs gleich: $f^* = d^*$.

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

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

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

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

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

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

liegt. (Nach Lemma 6.13 ist K ein abgeschlossener Kegel.) Um Aussage (*ii*) zu veranschaulichen, machen wir folgende Überlegung:

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

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

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

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

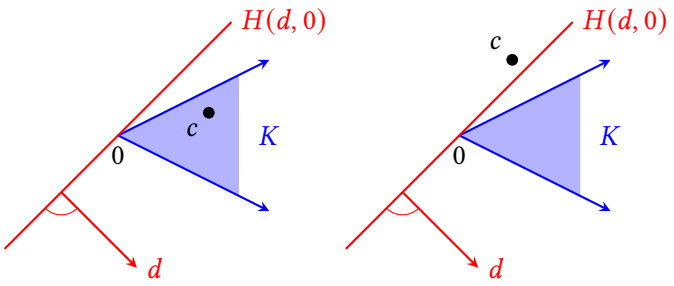


Abbildung 8.1.: Illustration der beiden Fälle (links: Aussagen (i) und (ii) sind beide erfüllt und rechts: beide nicht erfüllt) im Farkas-Lemma 8.6.

Um Aussage $(ii) \Rightarrow$ Aussage (i) zu zeigen, verwenden wir Kontraposition, also $\neg$ 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

$$\begin{aligned} \text{Minimiere} \quad & \|x - c\| \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} \quad & x \in K \cap \overline{B_R(c)}. \end{aligned} \tag{8.10}$$

Da K nach Lemma 6.13 abgeschlossen und $\overline{B_R(c)}$ kompakt ist, ist auch $K \cap \overline{B_R(c)}$ kompakt. Nach dem Satz von Weierstraß bzw. Satz 1.9 besitzt (8.10) daher einen globalen Minimierer w. Der Punkt w ist gleichzeitig ein globaler Minimierer der relaxierten Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & \|x - c\| \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} \quad & x \in K, \end{aligned} \tag{8.11}$$

weil Punkte außerhalb von $\overline{B_R(c)}$ als globale Minimierer von (8.11) nicht in Betracht kommen.
(**Quizfrage 8.3:** Warum können Punkte außerhalb von $\overline{B_R(c)}$ nicht globaler Minimierer von (8.11) sein?)

Behauptung: Der Vektor $d = w - c$ dient als Normalenvektor einer Hyperebene, die den Kegel 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

$$\begin{aligned} \|w - c\|^2 &\leqslant \|\alpha\, y + (1-\alpha)\, w - c\|^2 \quad \text{denn } w \text{ ist optimal für (8.11)} \\ &= \|\alpha\,(y - w) + (w - c)\|^2 \\ &= \alpha^2 \|y - w\|^2 + 2\,\alpha (y - w)^\mathsf{T} (w - c) + \|w - c\|^2. \end{aligned}$$

Daraus folgt

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

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

$$(y - w)^\mathsf{T} d \geqslant 0 \text{ für alle } y \in K. \tag{8.12}$$

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

$$w^\mathsf{T} d = 0. \tag{8.13}$$

Außerdem erhalten wir

$$c^\mathsf{T} d = (c - w)^\mathsf{T} d + w^\mathsf{T} d = -\|\underbrace{w - c}_{=d\neq 0}\|^2 + \underbrace{w^\mathsf{T} d}_{=0} < 0. \tag{8.14}$$

Insgesamt folgt

$$y^\mathsf{T} d \overset{(8.12)}{\geqslant} w^\mathsf{T} d = 0 \overset{(8.14)}{>} c^\top d \quad \text{for all } y \in K.$$

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

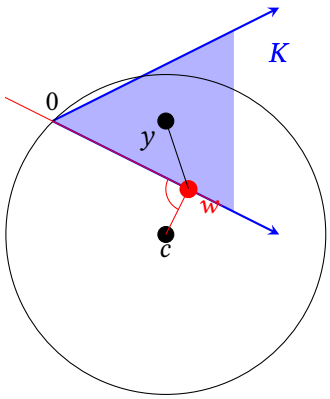


Abbildung 8.2.: Illustration der Konstruktion des Normalenvektors $d = w - c$ der trennenden Hyperebene (die rote Gerade) im Beweis des Farkas-Lemmas 8.6.

Ende der Vorlesung 11

Wir können nun Satz 8.5 beweisen.

Beweis von Satz 8.5. Wir zeigen zunächst die hinreichenden Bedingungen.
Aussage *(iii)*: Aus (8.9) folgt insbesondere, dass x^* und (λ^*, μ^*) zulässig sind für (8.4) und (8.5). Wegen (8.7) gilt

$$b^\mathsf{T}\lambda^* = c^\mathsf{T} x^* - \underbrace{(x^*)^\mathsf{T}\mu^*}_{=0} = c^\mathsf{T} x^*. \tag{8.15}$$

Folgerung 8.4 zeigt nun, dass x^* und (λ^*, μ^*) bereits Lösungen von (8.4) bzw. (8.5) sind.

Aussage *(i)*: Um die notwendigen Bedingungen zu zeigen, benötigen wir das Farkas-Lemma 8.6.[10] Es sei also x eine Lösung des primalen LP (8.4). Insbesondere ist f^* endlich, und aus Lemma 6.11 folgt, dass für alle Richtungen im Rezessionskegel

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

[10] Ein direkter Beweis ohne Rückgriff auf das Farkas-Lemma oder den Simplex-Algorithmus findet sich bei Forsgren, 2008. In seinem Beewis wird allerdings vorausgesetzt, dass A vollen Zeilenrang besitzt, was für die Aussage von Satz 8.5 nicht notwendig ist.

$c^\mathsf{T} d \geqslant 0$ gilt. Setzen wir

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

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

Wegen der schwachen Dualität (Satz 8.3) $d^* \leqslant f^*$ ist der duale Supremalwert d^* endlich. Aus dem Existenzsatz 6.12 (angewendet auf das duale LP) folgt, dass die duale Aufgabe (8.5) lösbar ist. Es existiert also ein für die duale Aufgabe zulässiges Paar (λ^*, μ^*), sodass $b^\mathsf{T}\lambda^* = d^*$ gilt.

Wir müssen noch zeigen, dass $d^* = f^*$ gilt und nicht etwa $d^* < f^*$. Dann folgt aus (8.15) die noch fehlende Komplementaritätsbedingung $(x^*)^\mathsf{T}\mu^* = 0$, die den Beweis von (8.9) vervollständigt.

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

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

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

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

also

$$A^\mathsf{T}\lambda \leqslant \alpha\, c, \quad \alpha \geqslant 0 \quad \text{und} \quad b^\mathsf{T}\lambda > \alpha\,(f^* - \varepsilon). \tag{8.16}$$

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

$$0 \geqslant \underbrace{x^\mathsf{T}}_{\geqslant 0}\,\underbrace{(A^\mathsf{T}\lambda)}_{\leqslant 0} = \lambda^\mathsf{T}(A\,x) = b^\mathsf{T}\lambda > 0.$$

Es muss also $\alpha > 0$ sein, und wir können durch Skalierung $\alpha = 1$ in (8.16) erreichen.[11] Damit gilt also nun

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

Damit ist λ dual zulässig, und aufgrund der Optimalität von x^* und des schwachen Dualitätssatzes 8.3 gilt $f^* = c^\mathsf{T}x^* \geqslant b^\mathsf{T}\lambda > f^* - \varepsilon$. Da $\varepsilon > 0$ beliebig war, muss

$$d^* := \sup\{b^\mathsf{T}\lambda \mid A^\mathsf{T}\lambda \leqslant c\} = \sup\{b^\mathsf{T}\lambda \mid A^\mathsf{T}\lambda + \mu = c,\ \mu \geqslant 0\} = f^*$$

gelten.

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

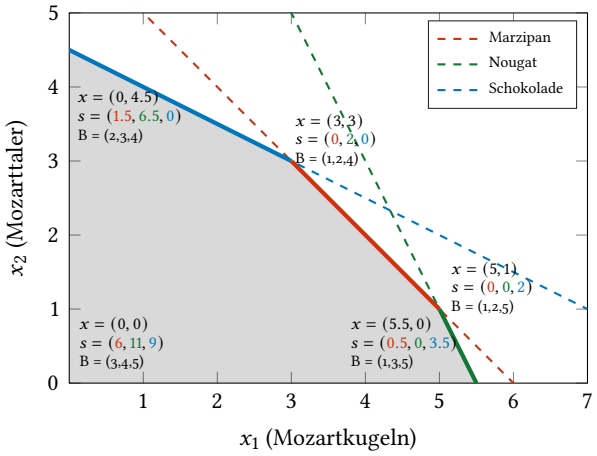
[11] Wir ersetzen dazu α durch $\alpha/\alpha = 1$ und λ durch λ/α.

Der Satz 8.5 sagt im Prinzip aus, dass wir das primale LP nicht lösen können, ohne auch das duale LP gleichzeitig zu lösen. Insbesondere können wir einen primal zulässigen Punkt nur dadurch als optimal bestätigen, dass wir einen zugehörigen dual zulässigen Punkt finden, sodass diese gemeinsam das Optimalitätssystem (8.9) erfüllen. Es ist daher nicht verwunderlich, dass im Simplex-Algorithmus, den wir in § 7 besprochen haben, auch die dualen Optimierungsvariablen (λ, μ) implizit vorkommen, und zwar als Nebenprodukte bei der Berechnung der reduzierten Kosten $\widetilde{c}_N := c_N - A_N^\mathsf{T} A_B^{-\mathsf{T}} c_B$:

$$\begin{aligned} \lambda &:= A_B^{-\mathsf{T}} c_B, \\ \widetilde{c}_N := \mu_N &:= c_N - A_N^\mathsf{T} \lambda, \\ \mu_B &:= 0. \end{aligned} \tag{8.17}$$

In jedem Simplex-Schritt sind alle Bedingungen im Optimalitätssystem (8.9) erfüllt mit Ausnahme von $\mu_N \geqslant 0$. Die Iterierten sind also primal zulässig und *dual unzulässig*, bis eine optimale Ecke gefunden wurde.

Beispiel 8.7 (Duale Variablen beim Mozartproblem (6.7))**.**



Das Mozartproblem in Normalform hat die Daten

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

Wir betrachten zunächst die Ecke $x = (0, 0, 6, 11, 9)^\mathsf{T}$, also die Basis $B = (3, 4, 5)$ und Nichtbasis $N = (1, 2)$. Die Basismatrix lautet

$$A_B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Laut (9.1) haben wir $A_B^\mathsf{T} \lambda = c_B$, also

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \lambda = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

und damit $\lambda = (0,0,0)^\mathsf{T}$. Weiter ist $\mu_B = (0,0,0)^\mathsf{T}$ und

$$\mu_N = c_N - A_N^\mathsf{T}\lambda = \begin{pmatrix}-9\\-8\end{pmatrix} - \begin{bmatrix}1&2&1\\1&1&2\end{bmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix} = \begin{pmatrix}-9\\-8\end{pmatrix}.$$

Im Optimalitätssystem (8.9) sind alle Bedingungen erfüllt mit Ausnahme der dualen Zulässigkeit für μ_N, d. h., die reduzierten Kosten sind nicht $\geqslant 0$.

Für die Ecke unten rechts haben wir $x = (5.5, 0, 0.5, 0, 3.5)^\mathsf{T}$ und die Basis $B = (1,3,5)$ und Nichtbasis $N = (2,4)$. Die Basismatrix lautet

$$A_B = \begin{bmatrix}1&1&0\\2&0&0\\1&0&1\end{bmatrix}.$$

Laut (9.1) gilt wir $A_B^\mathsf{T}\lambda = c_B$, also

$$\begin{bmatrix}1&2&1\\1&0&0\\0&0&1\end{bmatrix}\lambda = \begin{pmatrix}-9\\0\\0\end{pmatrix}$$

und damit $\lambda = (0,-4.5,0)^\mathsf{T}$. Weiter ist $\mu_B = (0,0,0)^\mathsf{T}$ und

$$\mu_N = c_N - A_N^\mathsf{T}\lambda = \begin{pmatrix}-8\\-0\end{pmatrix} - \begin{bmatrix}1&1&2\\0&1&0\end{bmatrix}\begin{pmatrix}0\\-4.5\\0\end{pmatrix} = \begin{pmatrix}-3.5\\4.5\end{pmatrix}.$$

Im Optimalitätssystem (8.9) sind wieder alle Bedingungen erfüllt mit Ausnahme der dualen Zulässigkeit für μ_N, d. h., die reduzierten Kosten sind nicht $\geqslant 0$. △

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 (8.9) erfüllt sind mit Ausnahme von $x_B \geqslant 0$. Die Iterierten des dualen Simplex-Verfahrens sind also dual zulässig und *primal unzulässig*. Wir besprechen das duale Simplex-Verfahren in § 9.

Satz 8.8 (Mögliche primal-duale Situationen)**.**
Für jedes primal-duale Paar von LP können folgende Situationen auftreten:

		duales LP (8.6) bzw. (8.5)		
		lösbar $d^* \in \mathbb{R}$	unbeschränkt $d^* = \infty$	unzulässig $d^* = -\infty$
primales LP (8.4)	lösbar $f^* \in \mathbb{R}$	(I) $d^* = f^*$	—	—
	unbeschränkt $f^* = -\infty$	—	—	(III)
	unzulässig $f^* = \infty$	—	(III)	(II)

Beweis. Zu Zeile 1 und Spalte 1:

$$
\begin{aligned}
& f^* \in \mathbb{R} \\
\overset{\text{(Satz 6.12)}}{\Leftrightarrow} \quad & \text{das primale Problem (8.4) besitzt eine Lösung} \\
\overset{\text{(Satz 8.5)}}{\Leftrightarrow} \quad & \text{die Optimalitätsbedingungen (8.9) besitzen eine Lösung} \\
\overset{\text{(Satz 8.5)}}{\Leftrightarrow} \quad & \text{das duale Problem (8.5) besitzt eine Lösung} \\
\overset{\text{(Satz 6.12)}}{\Leftrightarrow} \quad & d^* \in \mathbb{R}.
\end{aligned}
$$

Für „⇐“ in der letzten Aussage: Bringe (8.5) in Normalform und benutze Satz 6.12. Aus dem schwachen Dualitätssatz 8.3 und (8.15) 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 schwachen Dualitätssatz 8.3 $d^* \leqslant f^* = -\infty$ gelten, Widerspruch, also muss $D = \emptyset$ und $d^* = -\infty$ sein. Analoges gilt für die 2. Spalte (Fall (III)).

Fall (II) kann auftreten. (**Quizfrage 8.5:** Kennen Sie ein Beispiel? (Siehe auch Hausaufgabe 6.3.)) □

Bemerkung 8.9 (Starke Dualität)**.**
Zu der Erkenntnis $d^* = f^*$ im Fall (I) sagt man auch: „Es tritt **keine Dualitätslücke** auf“ (zwischen dem Infimalwert des primalen und dem Supremalwert der dualen Aufgabe, englisch: *no duality gap*) oder „Es herrscht **starke Dualität**“ (englisch: *strong duality*). Wir können also den Satz 8.8 auch wie folgt formulieren: In der linearen Optimierung herrscht starke Dualität genau dann, wenn sowohl das primale als auch das duale LP einen zulässigen Punkt besitzen. △

Ende der Vorlesung 12

Ende der Woche 6

§ 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 6.17, also als eine Auswahl von m Indizes aus $\{1, \dots, n\}$, sodass die Untermatrix A_B regulär ist.

Beachte: Eine Basis B legt gemäß

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

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

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

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 $x_B \geqslant 0$ erfüllt. Eine Basis B heißt **dual zulässig** (englisch: *dual feasible*), wenn das durch (9.1) beschriebene Paar von Vektoren (λ, μ) dual zulässig ist, also die Bedingung $\mu_N \geqslant 0$ erfüllt. Im Unterschied zum primalen Simplex-Verfahren werden wir mit primal unzulässigen Basisvektoren arbeiten. Dafür sind die Größen (λ, μ) stets dual zulässig, siehe Tabelle 9.1.

Wir leiten jetzt einen Schritt des dualen Simplex-Verfahrens analog zu § 7.1 her. Es sei dazu als Ausgangspunkt eine dual zulässige Basis B gegeben und (λ, μ) 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 $e_\ell \in \mathbb{R}^m$, $t \geqslant 0$. In Abhängigkeit von t ergibt sich der Wert von λ nun aus

$$A_B^\mathsf{T}\lambda(t) + \mu_B(t) = c_B,$$

also

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

Welchen Index ℓ wählen wir? Dazu betrachten wir die Werte der dualen Zielfunktion:

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

Hier übernimmt also $x_B := A_B^{-1}b$ die Rolle der reduzierten Kosten. Da wir die duale Zielfunktion maximieren wollen, wählen wir $\ell \in B$ so, dass $x_\ell < 0$ ist. Falls bereits $x_B \geqslant 0$ gilt, so haben wir eine primal und dual optimale Lösung gefunden. (**Quizfrage 9.1:** Begründung?)

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

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

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

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

Falls $\Delta\mu_N \geqslant 0$ ist, so ist die duale Aufgabe unbeschränkt und damit auch die primale Aufgabe nicht lösbar. (Die primale Aufgabe ist dann notwendigerweise unzulässig, siehe Satz 8.8).

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:

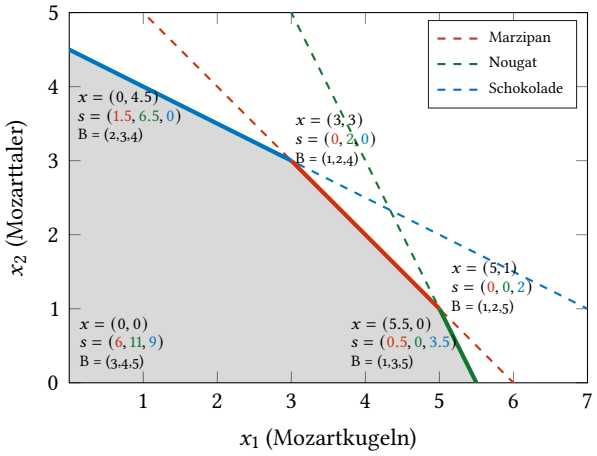
$$B^+ := \big(B \cup \{r\}\big) \setminus \{\ell\}$$
$$N^+ := \big(N \cup \{\ell\}\big) \setminus \{r\}.$$

Der Vollständigkeit halber geben wir das duale Simplex-Verfahren nochmal komplett an und stellen es dem primalen Verfahren gegenüber (Algorithmen 9.1 und 9.2). Nachdem wir in (8.17) gesehen haben, in welcher Beziehung die reduzierten Kosten zu den dualen Variablen stehen, nutzen wir die Gelegenheit, die dualen Variablen im primalen Verfahren nochmal mit den üblichen Bezeichnungen (λ, μ) umzubenennen.

Eine erste dual zulässige Ecke (sofern existent) kann mit Hilfe eines dualen Phase-I-Problems gefunden werden.

Beispiel 9.3 (Duales Simplex-Verfahren für das Mozartproblem (6.7))**.**

Wir berechnen und illustrieren einen Schritt des dualen Simplex-Verfahrens, angewendet auf das Mozartproblem (6.7), ausgehend von einem primal unzulässigen Basisvektor.



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 (8.4) (und ein optimaler Basisvektor (8.5)) oder die Aussage, dass (8.4) unbeschränkt ist

1: Setze $k := 0$

2: Berechne die primalen reduzierten Kosten (duale Variablen)

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

3: **if** $\mu_N \geqslant 0$ **then**

4: x ist eine Lösung von (8.4), und (λ, μ) ist eine Lösung von (8.5), **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** $\Delta x_B \geqslant 0$ **then**

9: Aufgabe (8.4) ist unbeschränkt, **STOP**

10: **else**

11: Bestimme $\hat{t} \geqslant 0$ und $\ell \in B$ gemäß

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

12: Setze

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

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

14: Setze $N^+ := \{1, \ldots, n\} \setminus B^+$

15: Setze $x := x^+$

16: Setze $B := B^+$ und $N := N^+$

17: Setze $k := k + 1$

18: **end if**

19: **end if**

20: Gehe zu Zeile 2

Algorithmus 9.2 (Dualer Simplex-Algorithmus (Lemke, 1954))**.**

Eingabe: Aufgabenbeschreibung durch A, b, c

Eingabe: dual zulässiger Basisvektor (λ, μ) von P mit zugehöriger Basis B und Nichtbasis N

Ausgabe: ein optimaler Basisvektor von (8.5) (und ein optimaler Basisvektor von (8.4)) oder die Aussage, dass (8.5) 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** $x_B \geqslant 0$ **then**

4: (λ, μ) ist eine Lösung von (8.5), und x ist eine Lösung von (8.4), **STOP**

5: **else**

6: Wähle einen Index $\ell \in B$ mit $x_\ell < 0$

7: Berechne

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

8: **if** $\Delta\mu_N \geqslant 0$ **then**

9: Aufgabe (8.5) ist unbeschränkt, **STOP**

10: **else**

11: Bestimme $\hat{t} \geqslant 0$ und $r \in N$ gemäß

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

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

14: Setze $N^+ := \{1, \ldots, n\} \setminus B^+$

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

Wir verwenden als Beispiel den Schnittpunkt der Beschränkungen „Nougat“ und „Schokolade“. Das ist der primal unzulässige Basisvektor, der durch $B = (1, 2, 3)$ und $N = (4, 5)$ charakterisiert ist. Die dualen Variablen dort erfüllen

$$\begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 2 \\ 1 & 0 & 0 \end{bmatrix} \lambda = \begin{pmatrix} -9 \\ -8 \\ 0 \end{pmatrix},$$

also ist $\lambda = \frac{1}{3}(0, -10, -7)^\mathsf{T}$, und weiter

$$\mu_N = c_N - A_N^\mathsf{T}\lambda = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \frac{1}{3} \begin{pmatrix} 0 \\ -10 \\ -7 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 10 \\ 7 \end{pmatrix},$$

sodass wir insgesamt $\mu = \frac{1}{3}(0, 0, 0, 10, 7)^\mathsf{T}$ erhalten. Wie erwartet ist die gewählte Basis also dual zulässig aber primal unzulässig, denn zu ihr gehört der primale Basisvektor x, der

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 0 \\ 1 & 2 & 0 \end{bmatrix} x_B = \begin{pmatrix} 6 \\ 11 \\ 9 \end{pmatrix}$$

erfüllt, also $x = \frac{1}{3}(13, 7, -2, 0, 0)^\mathsf{T}$.

Für den Index $\ell \in B$ mit $x_\ell < 0$ haben wir nur eine Wahl, und zwar $\ell = 3$. Wir berechnen $\Delta\lambda = -A_B^{-\mathsf{T}} e_\ell$, lösen also das lineare Gleichungssystem

$$\begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 2 \\ 1 & 0 & 0 \end{bmatrix} \Delta\lambda = -\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

was $\Delta\lambda = \frac{1}{3}(-3, 1, 1)^\mathsf{T}$ ergibt. Für die Richtung der Aufdatierung von μ erhalten wir

$$\Delta\mu_N = -A_N^\mathsf{T}\Delta\lambda = -\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \frac{1}{3} \begin{pmatrix} -3 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

Daraus ergibt sich die Wahl

$$\hat{t} = \min\left\{ -\frac{\mu_4}{\Delta\mu_4}, -\frac{\mu_5}{\Delta\mu_5} \right\} = \min\left\{ -\frac{10/3}{-1/3}, -\frac{7/3}{-1/3} \right\} = 7$$

mit $r = 5$.

Dadurch erhalten wir für die neuen Werte von λ und μ Folgendes:

$$\lambda^+ = \lambda + \hat{t}\,\Delta\lambda = \frac{1}{3} \begin{pmatrix} 0 \\ -10 \\ -7 \end{pmatrix} + 7\,\frac{1}{3} \begin{pmatrix} -3 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -7 \\ -1 \\ 0 \end{pmatrix}$$

und

$$\mu^+ = \begin{pmatrix} 0 \\ 0 \\ \hat{t} \\ \mu_4 + 7\,\Delta\mu_4 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 7 \\ 10/3 + 7\,(-1/3) \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 7 \\ 1 \\ 0 \end{pmatrix} \begin{array}{l} \text{bleibt in } B \\ \text{bleibt in } B \\ \text{verlässt } B\text{, kommt neu in } N \\ \text{bleibt in } N \\ \text{verlässt } N\text{, kommt neu in } B \end{array}$$

Die neuen Indizes sind $B^+ = \{1, 2, 5\}$ und $N^+ = \{3, 4\}$. Die Bestimmung der dualen reduzierten Kosten zu Beginn der nächsten Iteration ergibt (mit Umbennenung B statt B^+)

$$\begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} x_B = \begin{pmatrix} 6 \\ 11 \\ 9 \end{pmatrix},$$

also $x_B = (5, 1, 2)^\mathsf{T}$. Damit ist die neue Ecke bereits optimal, da $x_B \geqslant 0$ gilt. △

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 (8.4) 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

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

erweitern wollen. Wir können die neue Variable mit $\bar{x} = 0$ initialisieren und erhalten dadurch einen weiterhin primal zulässigen Basisvektor zur bisherigen Basis B. Die neue Nichtbasis ist $N \cup \{n + 1\}$. Die Bedingungen der dualen Zulässigkeit für das neue Problem lauten

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

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

Wir können an Stelle einer einzelnen Variablen mit denselben Argumenten auch mehrere Variablen gleichzeitig hinzufügen.

Hinzufügen einer Nebenbedingung

Wir betrachten jetzt eine andere Veränderung der primalen Aufgabe (8.4) und fügen ihr eine neue Ungleichungsnebenbedingung $\bar{a}^\mathsf{T} x \leqslant \bar{b}$ bzw. $\bar{a}^\mathsf{T} x + \bar{x} = \bar{b}$ mit zugehöriger Schlupfvariable $\bar{x}$ hinzu:

$$\begin{aligned} \text{Minimiere} \quad & \begin{pmatrix} c \\ 0 \end{pmatrix}^\mathsf{T} \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \quad \text{über } \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \\ \text{sodass} \quad & \begin{bmatrix} A & 0 \\ \bar{a}^\mathsf{T} & 1 \end{bmatrix} \begin{pmatrix} x \\ \bar{x} \end{pmatrix} = \begin{pmatrix} b \\ \bar{b} \end{pmatrix} \\ \text{und} \quad & \begin{pmatrix} x \\ \bar{x} \end{pmatrix} \geqslant 0. \end{aligned} \tag{9.4}$$

Die Bedingungen der dualen Zulässigkeit für die neue Aufgabe lauten

$$\begin{bmatrix} A^\mathsf{T} & \bar{a} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \lambda \\ \bar{\lambda} \end{pmatrix} + \begin{pmatrix} \mu \\ \bar{\mu} \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \mu \\ \bar{\mu} \end{pmatrix} \geqslant 0. \tag{9.5}$$

Die bisherige Lösung x wird i. A. nicht länger primal zulässig sein, da $\bar{a}^\mathsf{T} x \leqslant \bar{b}$ 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

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

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.

Auch hier können an Stelle einer einzigen Nebenbedingung auch wieder mehrere gleichzeitig hinzugefügt werden.

Bemerkung 9.4 (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** (auch: **(gemischt-)ganzzahliger lineare Programme**, 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

$$\left.\begin{aligned} \text{Minimiere} \quad & c^\mathsf{T} x \quad \text{über } x \in \mathbb{Z}^n \\ \text{sodass} \quad & A x = b \\ \text{und} \quad & x \geqslant 0. \end{aligned}\right\} \tag{9.6}$$

Solche Aufgaben fallen in den Bereich der ganzzahligen Optimierung. In einem gängigen Lösungsansatz, den man **branch and bound** nennt, wird zunächst ein relaxiertes LP gelöst, bei dem die

Ganzzahligkeitsbedingungen vernachlässigt werden, also (8.4). Dessen Lösung bezeichnen wir jetzt mit x^*. Dann wird eine Variable x_i^* ausgewählt, die die Ganzzahligskeitsbedingung verletzt, und es werden die zwei LPs

$$\left.\begin{array}{rl} \text{Minimiere} & c^\mathsf{T}x \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} & A\,x = b \\ \text{und} & x \geqslant 0 \\ \text{sowie} & x_i \geqslant \lceil x_i^* \rceil \end{array}\right\} \quad \left\{\begin{array}{rl} \text{Minimiere} & c^\mathsf{T}x \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} & A\,x = b \\ \text{und} & x \geqslant 0 \\ \text{sowie} & x_i \leqslant \lfloor x_i^* \rfloor \end{array}\right. \tag{9.7}$$

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

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

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 Vorlesung 13

§ 10 Sensitivitätsanalyse

In diesem Abschnitt gehen wir der Frage nach, wie empfindlich (sensitiv) der Infimalwert (also der Zielfunktionswert an einer optimalen Lösung) eines LPs in Normalform (8.4) gegenüber Änderungen im Kostenvektor c und in der rechten Seite b sind.

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 Umstände 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 (λ^*, μ^*) Lösungen der primalen Aufgabe (8.4) bzw. der dualen Aufgabe (8.5) zu einer Basis B, also optimale Ecken, wie sie mit dem primalen oder dem dualen Simplex-Verfahren berechnet werden.

Änderungen im Kostenvektor

Wir bezeichnen mit $\Delta c \in \mathbb{R}^n$ eine Änderungsrichtung im Kostenvektor c der primalen Aufgabe und betrachten folgende Familie primal-dualer Aufgaben mit Parameter $t \in \mathbb{R}$:

$$\left.\begin{array}{rl} \text{Minimiere} & (c + t\,\Delta c)^\mathsf{T} x \\ \text{sodass} & A\,x = b \\ \text{und} & x \geqslant 0 \end{array}\right\} \quad \left\{\begin{array}{rl} \text{Maximiere} & b^\mathsf{T}\lambda \\ \text{sodass} & A^\mathsf{T}\lambda + \mu = c + t\,\Delta c \\ \text{und} & \mu \geqslant 0. \end{array}\right. \tag{10.1}$$

Welche Aussagekraft besitzen die Lösungen x^* und (λ^*, μ^*) des „ungestörten“ Aufgabenpaares ($t = 0$) noch für (10.1)?

Da (10.1) dieselbe primal zulässige Menge besitzt wie die ungestörte Aufgabe (8.4), ist x^* weiterhin primal zulässig. Die dual zulässige Menge hat sich jedoch gegenüber (8.5) geändert. Wir können aber den Versuch unternehmen, die duale Lösung aufzudatieren. Dazu gehen wir wie in § 9 bei der Herleitung des dualen Simplex-Verfahrens vor. Durch die Basis B sind die dualen Variablen wie folgt festgelegt, vgl. (9.1):

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

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

$$\mu_B(t) \equiv \mu_B^* = 0. \tag{10.2c}$$

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 (8.9). Die primale Zulässigkeit

$$A\,x^* = b, \quad x^* \geqslant 0$$

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

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

nach Konstruktion von $\lambda(t)$ und $\mu(t)$ erfüllt. Die Vorzeichenbedingung $\mu_N(t) \geqslant 0$ jedoch gilt nicht automatisch, sondern genau dann, wenn der Störungsparameter t der Bedingung[12]

$$\sup_{\substack{i \in N \\ \Delta\mu_i > 0}} \underbrace{\left\{-\frac{\mu_i^*}{\Delta\mu_i}\right\}}_{\leqslant 0} \leqslant t \leqslant \inf_{\substack{i \in N \\ \Delta\mu_i < 0}} \underbrace{\left\{-\frac{\mu_i^*}{\Delta\mu_i}\right\}}_{\geqslant 0}. \tag{10.3}$$

[12] Wir schreiben hier sup und inf statt max und min, da die betreffenden Indexmengen durchaus leer sein können. Beispielsweise ist, sofern $\mu_N \geqslant 0$ gilt, die Indexmenge für die obere Schranke leer, sodass die betreffende Ungleichung in (10.3) zu „$t \leqslant \inf \emptyset = \infty$“ 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 Infimalwert lässt sich daher bequem aus der primalen Aufgabe ablesen:

$$f^*(t) = (c + t\,\Delta c)^\mathsf{T} x^* = f^* + t\,\Delta c^\mathsf{T} x^*. \tag{10.4}$$

Wir fassen unsere Erkenntnisse zusammen:

Satz 10.1 (Sensitivitätssatz bei LP bei Änderungen im Kostenvektor)**.**
Es seien x^* und (λ^*, μ^*) Lösungen der ungestörten primalen Aufgabe $(10.1)_{\text{primal}}$ bzw. der ungestörten dualen Aufgabe $(10.1)_{\text{dual}}$ zu einer Basis B. Dann gilt:

(*i*) Für beliebiges $\Delta c \in \mathbb{R}^n$ und zugehörige t gemäß (10.3) ist x^* für $(10.1)_{\text{primal}}$ weiterhin ein optimaler Basisvektor, und $(\lambda(t), \mu(t))$ aus (10.2) ist ein optimaler Basisvektor für $(10.1)_{\text{dual}}$. Der gemeinsame Optimalwert beider Aufgaben ist $c^\mathsf{T} x^* + t\,(\Delta c)^\mathsf{T} x^*$.

(*ii*) Ist die rechte Grenze des Intervalls (10.3) echt positiv, dann ist die Optimalwertfunktion

$$c \mapsto \Phi(c) := \text{gemeinsamer Optimalwert von (8.4) und (8.5)}$$

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

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

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

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

Damit ist Φ überall in dieser Kugel differenzierbar, und es gilt

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

Beweis. Aussage (*i*): Diese Aussage haben wir durch Bestätigung der Optimalitätsbedingungen (8.9) bereits bewiesen.

Aussage (*ii*): Unter der genannten Voraussetzung ist $\Phi(c + t\,\Delta c)$ für hinreichend kleine $t > 0$ durch (10.4) gegeben. Für die Richtungsdiffbarkeit von Φ betrachten wir den Differenzenquotienten für solche t:

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

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

Aussage (*iii*): Wenn μ^* nicht entartet ist, dann enthält das zulässige Intervall (10.3) für jede beliebige Richtung Δc immer ein offenes Intervall um die 0. Wir müssen aber zeigen, dass die Länge dieses

Intervalls über alle Richtungen Δc konstanter Norm gleichmäßig von 0 weg beschränkt bleibt. Wir zeigen dazu, dass die Funktion, die die obere Intervallgrenze angibt,

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

auf der Einheitssphäre $\{\Delta c \in \mathbb{R}^n \mid \|\Delta c\| = 1\}$ gleichmäßig von 0 weg beschränkt ist. Das ist ausreichend, weil sich die untere Intervallgrenze durch den Übergang $\Delta c \rightsquigarrow -\Delta c$ ergibt.

Wegen (10.2) hängt $\Delta\mu_N$ linear (und damit stetig) von Δc ab:

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

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

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

Es gilt $\beta_i > 0$ für alle $i \in N$. (**Quizfrage 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 Δc aus der Einheitssphäre und das zugehörige $\Delta\mu$ gilt: Falls $\Delta\mu_N \geqslant 0$ 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

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

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

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

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

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

Weiter folgt in Verbindung mit (10.4), dass für alle Kostenvektoren $c + \Delta c$ mit $\Delta c \in B_r(0)$ die Optimalwertfunktion die Gestalt $\Phi(c + \Delta c) = \Phi(c) + \Delta c^\mathsf{T} x^* = (c + \Delta c)^\mathsf{T} x^*$ hat. Die Differenzierbarkeit von Φ in $B_r(c)$ mit Ableitung $(x^*)^\mathsf{T}$ ist eine unmittelbare Konsequenz. □

Änderungen in der rechten Seite

Wir betrachten jetzt Änderungen in der rechten Seite b und bezeichnen mit $\Delta b \in \mathbb{R}^m$ eine entsprechende Änderungsrichtung. Das primal-duale Paar von Aufgaben hat nun die Gestalt

$$\left.\begin{array}{rl} \text{Minimiere} & c^\mathsf{T} x \\ \text{sodass} & A\,x = b + t\,\Delta b \\ \text{und} & x \geqslant 0 \end{array}\right\} \quad \left\{\begin{array}{rl} \text{Maximiere} & (b + t\,\Delta b)^\mathsf{T}\lambda \\ \text{sodass} & A^\mathsf{T}\lambda + \mu = c \\ \text{und} & \mu \geqslant 0. \end{array}\right. \tag{10.6}$$

Dieses Mal ist (λ^*, μ^*) weiterhin dual zulässig, die primal zulässige Menge hat sich jedoch geändert. Wir unternehmen daher jetzt den Versuch, die primale Lösung aufzudatieren. Das Vorgehen ähnelt dem bei der Herleitung des primalen Simplex-Verfahrens in § 7. Durch die Basis B ist die primale Variable wie folgt festgelegt:

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

$$x_N(t) \equiv x_N^* = 0. \tag{10.7b}$$

Wann sind die auf diese Art und Weise erhaltenen Vektoren $x(t)$ und (λ^*, μ^*) optimal für (10.6)? Wir überprüfen dazu die Optimalitätsbedingungen (8.9). Die duale Zulässigkeit

$$A^\mathsf{T}\lambda^* + \mu^* = c, \quad \mu^* \geqslant 0$$

ist erfüllt, ebenso die Komplementaritätsbedingung:

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

Bzgl. der primalen Zulässigkeit ist die erste Bedingung

$$A\,x(t) = A_B\,x_B(t) + A_N\,x_N(t) = A_B\,x_B^* + t\,A_B\,\Delta x_B + A_N x_N = A_B\,x_B^* + t\,A_B A_B^{-1}\Delta b + 0 = b + t\,\Delta b$$

nach Konstruktion von $x(t)$ erfüllt. Die Vorzeichenbedingung $x_B(t) \geqslant 0$ jedoch gilt nicht automatisch, sondern genau dann, wenn der Störungsparameter t der Bedingung

$$\sup_{\substack{i\in B\\ \Delta x_i>0}} \underbrace{\left\{-\frac{x_i^*}{\Delta x_i}\right\}}_{\leqslant 0} \leqslant t \leqslant \inf_{\substack{i\in B\\ \Delta x_i<0}} \underbrace{\left\{-\frac{x_i^*}{\Delta x_i}\right\}}_{\geqslant 0} \tag{10.8}$$

genügt.

Für diese t ist also tatsächlich (λ^*, μ^*) auch für die gestörten Probleme (10.6) weiterhin eine optimale Ecke. Der zugehörige Optimalwert lässt sich daher dieses Mal bequem aus der dualen Aufgabe ablesen:

$$d^*(t) = (b + t\,\Delta b)^\mathsf{T}\lambda^* = d^* + t\,\Delta b^\mathsf{T}\lambda^*. \tag{10.9}$$

Die Erkenntnisse fassen wir wie folgt zusammen:

Satz 10.2 (Sensitivitätssatz bei LP bei Änderungen in der rechten Seite).
Es seien x^* und (λ^*, μ^*) Lösungen der ungestörten primalen Aufgabe $(10.6)_{\text{primal}}$ bzw. der ungestörten dualen Aufgabe $(10.6)_{\text{dual}}$ 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 $(10.6)_{\text{dual}}$ weiterhin ein optimaler Basisvektor, und $x(t)$ aus (10.7) ist ein optimaler Basisvektor für $(10.6)_{\text{primal}}$. Der gemeinsame Optimalwert beider Aufgaben ist $b^\mathsf{T}\lambda^* + t\,(\Delta b)^\mathsf{T}\lambda^*$.

(*ii*) Ist die rechte Grenze des Intervalls (10.8) echt positiv, dann ist die Optimalwertfunktion

$$b \mapsto \Psi(b) := \text{gemeinsamer Optimalwert von (8.4) und (8.5)}$$

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

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

(*iii*) Ist x^* nicht entartet, gilt also $x_B^* > 0$, dann ist die Optimalwertfunktion in einer offenen Kugel $B_r(b)$ von b linear mit

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

Damit ist Ψ überall in dieser Kugel differenzierbar, und es gilt

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

Der Beweis erfolgt analog zum Beweis von Satz 10.1.

Expertenwissen: Auch μ^* hat eine Interpretation als Sensitivität

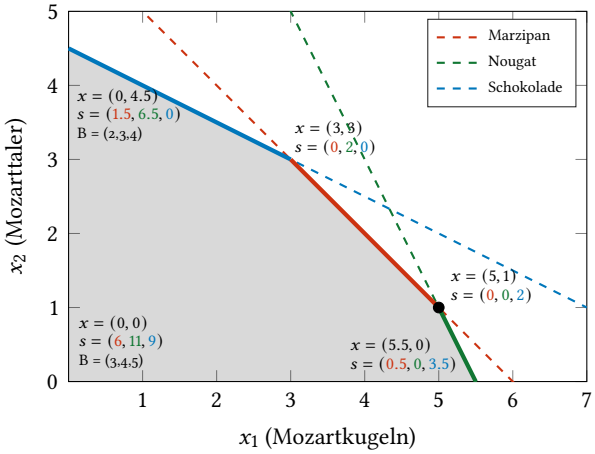
Satz 10.1 besagt, dass die primale Lösung x^* unter geeigneten Voraussetzungen die Sensitivität (Ableitung) der Optimalwertfunktion bei Störungen Δc im Kostenvektor ist. Analog besagt Satz 10.2, dass die duale Lösung λ^* unter geeigneten Voraussetzungen die Sensitivität der Optimalwertfunktion bei Störungen Δb in der rechten Seite der Gleichungsnebenbedingung ist.

Auch die zweite duale Variable μ^* besitzt eine solche Interpretation, nämlich als Sensitivität der Optimalwertfunktion bei Störungen der rechten Seite der Ungleichungsnebenbedingung $x \geqslant 0$. (**Quizfrage 10.4:** Können Sie ein entsprechendes Resultat formulieren und beweisen?) Insbesondere zeigt ein Wert von $\mu_i^* = 0$ in der dualen Lösung an, dass die zugehörige primale Ungleichung $x_i \geqslant 0$ irrelevant ist.

Beispiel 10.3 (Sensitivitäten beim Mozartproblem bei Änderungen in den Ressourcen)**.**
Das Mozartproblem in Normalform (Beispiel 6.7) ist durch die Daten

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

gegeben. Die eindeutige primal optimale Lösung ist $x^* = (5, 1, 0, 0, 2)^\mathsf{T}$ zur Basis $B = (1, 2, 5)$. Auch die duale Lösung (λ^*, μ^*) ist eindeutig, und zwar $\lambda^* = (-7, -1, 0)^\mathsf{T}$ und $\mu^* = (0, 0, 7, 1, 0)^\mathsf{T}$.



An diesem Bild sehen wir schon, dass wir den Ressourcenvektor in jede beliebige Richtung Δb um $t\,\Delta b$ innerhalb eines Intervalls für t um die 0 herum stören können. (**Quizfrage 10.5:** Wie sieht man das?) Wir sind also hier im differenzierbaren Fall von Satz 10.2. Das liegt daran, dass der optimale Basisvektor $x^* = (5,1,0,0,2)^\mathsf{T}$ nicht entartet ist. Wir erhalten also aus Satz 10.2 *(iii)* folgende Darstellung des Infimalwerts als Funktion des Ressourcenvektors b:

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

für Δb mit hinreichend kleiner Norm.

Wir betrachten im Folgenden nur Störungen jeder Ressource einzeln und berechnen auch die Änderungsrichtungen der B-Komponenten ($B = (1,2,5)$) der optimalen Lösungen in den drei Fällen:

Marzipan	Nougat	Schokolade
$\Delta b = (1,0,0)^\mathsf{T}$	$\Delta b = (0,1,0)^\mathsf{T}$	$\Delta b = (0,0,1)^\mathsf{T}$
$\Delta x_B = A_B^{-1}\Delta b = (-1,2,-3)^\mathsf{T}$	$\Delta x_B = A_B^{-1}\Delta b = (1,-1,1)^\mathsf{T}$	$\Delta x_B = A_B^{-1}\Delta b = (0,0,1)^\mathsf{T}$.

Diese Richtungen sollten Sie im obigen Bild einzeichnen und interpretieren! Als zulässige Intervalle für die Größe der Störung ergeben sich aus (10.8)

$-\frac{1}{2} \leqslant t \leqslant \frac{2}{3}$	$-2 \leqslant t \leqslant 1$	$-2 \leqslant t < \infty$
Bei $t = -\frac{1}{2}$ geht x_2 auf Null.	Bei $t = -2$ geht x_5 auf Null.	Bei $t = -2$ geht x_5 auf Null.
Bei $t = \frac{2}{3}$ geht x_5 auf Null.	Bei $t = 1$ geht x_2 auf Null.	

Bei einer Änderung der Marzipanressourcen lautet die Rechnung für das zulässige Intervall (10.8)

beispielsweise

$$\sup_{\substack{i\in B\\ \Delta x_i>0}} \underbrace{\left\{-\frac{x_i^*}{\Delta x_i}\right\}}_{\leqslant 0} \leqslant t \leqslant \inf_{\substack{i\in B\\ \Delta x_i<0}} \underbrace{\left\{-\frac{x_i^*}{\Delta x_i}\right\}}_{\geqslant 0}.$$

$$\Leftrightarrow \quad \max\Big\{\underbrace{-\frac{1}{2}}_{i=2}\Big\} \leqslant t \leqslant \min\Big\{\underbrace{-\frac{5}{-1}}_{i=1}, \underbrace{-\frac{2}{-3}}_{i=5}\Big\}$$

$$\Leftrightarrow \quad -\frac{1}{2} \leqslant t \leqslant \frac{2}{3}.$$

Wir wollen den Wert der dualen Variable $\lambda^* = (-7,-1,0)^\mathsf{T}$ interpretieren. Er bedeutet, dass wir pro Einheit an Marzipan $\Delta b = (1,0,0)^\mathsf{T}$, das wir zusätzlich zur Verfügung haben, wegen $(\lambda^*)^\mathsf{T}\Delta b = -7$ 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.

Der Zukauf von Marzipan ist aber nur sinnvoll, solange der Faktor t im Ausdruck $t\,\Delta b$ im zulässigen Intervall bleibt. Bei Erreichen der Intervallgrenzen ändert sich die Lösungsstruktur.

Unser Schattenpreis für Nougat beläuft sich auf eine Geldeinheit pro Einheit Nougat, und der Schattenpreis für Schokolade ist sogar gleich Null. (**Quizfrage 10.6:** Warum?) △

Ende der Vorlesung 14

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

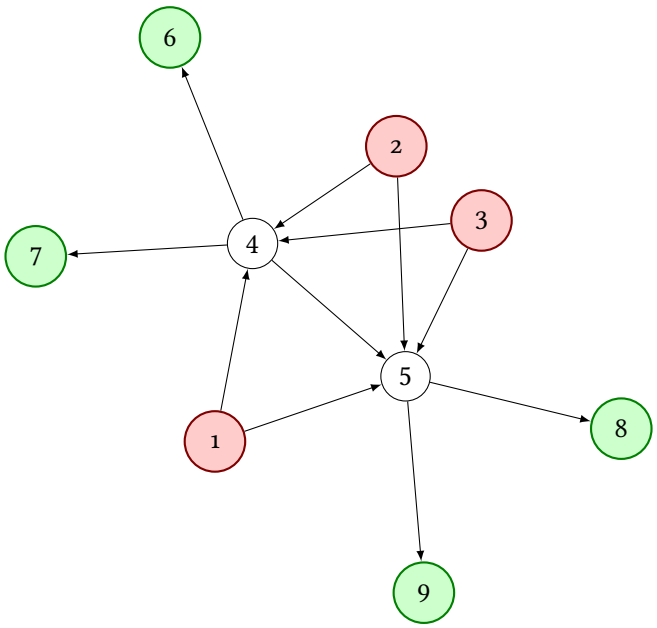


Abbildung 11.1.: Transportnetzwerk eines Unternehmens (siehe Beispiel 11.1) mit Produktionsstätten 1–3 (rot), Zwischenlagern 4–5 und Verkaufsstätten 6–9 (grün).

üblicherweise **Kapazitätsbeschränkungen** (englisch: *capacity constraints*) auf jedem Transportweg geben. Gesucht ist nun eine optimale Belegung der Kanten mit den darüber zu transportierenden Warenmengen, sodass (unter Beachtung aller Restriktionen) die Gesamttransportkosten minimiert werden. △

Definition 11.2 (Gerichtete 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:** Wieviele Kanten kann ein Digraph höchstens besitzen?)

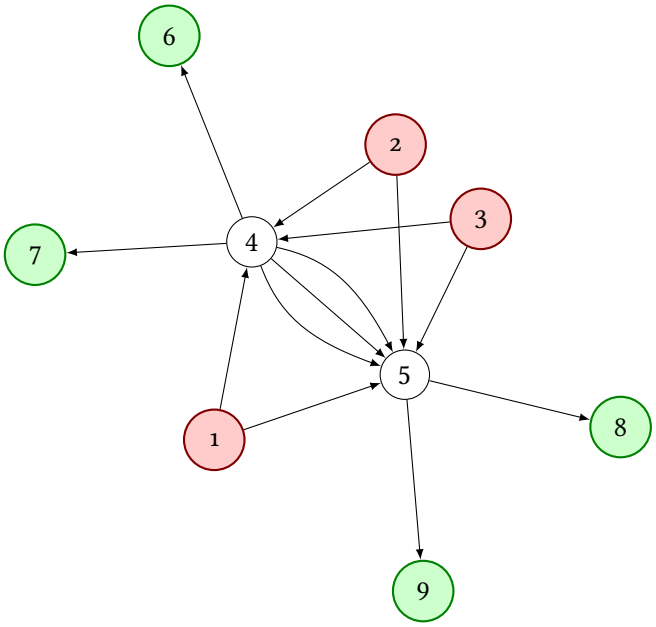
Der Graph aus Abbildung 11.1 wird beispielsweise beschrieben durch

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

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

Bemerkung 11.3 (Gerichtete Multigraphen)**.**
Es ist auch möglich, mehrfache Kanten zwischen Knoten zuzulassen. Diese entsprechen beispielsweise verschiedenen Transportwegen. Das führt zum Begriff des **gerichteten Multigraphen** (englisch: *directed multigraph, multidigraph*).

Andererseits können wir Multigraphen aber auch in äquivalente Graphen ohne Mehrfachkanten umschreiben. (**Quizfrage 11.2:** Wie könnte das beispielsweise bei folgendem Multigraphen funktionieren?)



△

Die Beschreibung eines Graphen durch Auflistung der Knoten und Kanten ist für die Formulierung von Optimierungsaufgaben ungeeignet. Stattdessen arbeiten wir mit Inzidenzmatrizen.

Definition 11.4 (Inzidenzmatrix)**.**

Es sei (V, E) ein einfacher Digraph. Die Knotenmenge sei $V = \{v^{(1)}, v^{(2)}, \ldots, v^{(m)}\}$ und die Kantenmenge $E = \{e^{(1)}, e^{(2)}, \ldots, e^{(n)}\}$. Die zu diesem Digraphen gehörende **Knoten-Kanten-Inzidenzmatrix** (englisch: *node-edge incidence matrix*) $A = (a_{ij})$ hat die Dimension $m \times n$ und ist wie folgt definiert:

$$a_{ij} = \begin{cases} -1, & \text{falls die Kante } e^{(j)} \text{ im Knoten } v^{(i)} \text{ startet,} \\ 1, & \text{falls die Kante } e^{(j)} \text{ im Knoten } v^{(i)} \text{ endet,} \\ 0, & \text{sonst.} \end{cases}$$

Wir sprechen auch kurz von der **Inzidenzmatrix** des Digraphen (V, E). △

Nummerieren wir die Kanten wie sie in (11.1) aufgezählt werden, so ist die Inzidenzmatrix des Digraphen

in Abbildung 11.1 gegeben durch

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

Der besseren Lesbarkeit wegen wurden die Nullen durch „·“ ersetzt. Es sollte klar sein, dass jeder einfache Digraph durch seine Inzidenzmatrix (bis auf Umordnung der Knoten und Kanten) eindeutig beschrieben wird.

Beachte: Da jede Kante (Spalte) genau einen Anfang (Eintrag −1) und ein Ende (Eintrag +1) hat, sind alle Spaltensummen gleich null, also $\mathbf{1}^\mathsf{T} A = 0$.

Expertenwissen: Laplacematrix eines Digraphen

Die Matrix AA^T wird die **Laplacematrix des Digraphen** (englisch: *Graph Laplacian*) genannt. Sie hat viele interessante Eigenschaften und Anwendungen in der Graphentheorie, die wir in dieser Vorlesung aber nicht weiter betrachten. (**Quizfrage 11.3:** Können Sie eine Vermutung anstellen, was die Einträge der Matrix AA^T aussagen?)

Wir überlegen uns jetzt am Beispiel von (11.2), was das Matrix-Vektor-Produkt $A\,x$ bedeutet. Der Vektor $x \in \mathbb{R}^{11}$ steht dabei für die Warenmengen, die über die Kanten fließen. Wir betrachten die Zeile 5 des Matrix-Vektor-Produkts, also

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & -1 & -1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{11} \end{pmatrix} = x_2 + x_4 + x_6 + x_7 - x_{10} - x_{11}.$$

Im Ergebnis spielen also nur die Warenströme der an den Knoten 5 (eines der Zwischenlager) angrenzenden Kanten (mit den Nummern 2, 4, 6, 7, 10, 11) eine Rolle. Dabei werden die über die Kanten 2, 4, 6 und 7 *eingehenden* Warenströme positiv gezählt und die über die Kanten 10 und 11 *ausgehenden* Warenströme negativ.

Das Matrix-Vektor-Produkt $A\,x$ gibt also offenbar den Vektor der **Knotenbilanzen** (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**, englisch: *minimum cost flow*) nun als lineare Optimierungsaufgabe formulieren. Die zu minimierenden Gesamtkosten aller Transportströme setzen sich als Summe der Kosten über die einzelnen Kanten zusammen:

$$\sum_{j=1}^{n} c_j\, x_j = c^\mathsf{T} x.$$

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

- **Bedarfsknoten** oder **Senken** (englisch: *demand nodes, sinks*) $b_i > 0$,
- **Angebotsknoten** oder **Quellen** (englisch: *supply nodes, sources*) $b_i < 0$,
- **Durchfluss-** oder **Umladeknoten** (englisch: *transshipment nodes*) $b_i = 0$.

Die Erfüllung aller Knotenbilanzen wird durch das lineare Gleichungssystem

$$A\,x = b$$

ausgedrückt. Dieses heißen auch **Flusserhaltungsgleichungen** (englisch: *flow conservation equations*) oder **Erhaltungsbedingungen** (englisch: *conservation constraints*). 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:

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

(**Quizfrage 11.4:** Wenn man auf einer Kante Rückflüsse zulassen will, kann man dann nicht einfach die Kapazitätsbeschränkung als $-u_i \leqslant x_i \leqslant u_i$ formulieren und sich eine der beiden Kanten sparen?)

Definition 11.5 (Flussnetzwerk, kostenminimaler Fluss)**.**

(*i*) Ein einfacher gerichteter Digraph mit Inzidenzmatrix $A \in \mathbb{R}^{m\times n}$, **Kantenkapazitäten** (englisch: *edge capacities*) $u \in \mathbb{R}^n$ und **Knotenbilanzen** (englisch: *node balances*) $b \in \mathbb{R}^m$ wird als **Transportnetzwerk** (englisch: *transport network*) oder **Flussnetzwerk** (englisch: *flow network*) bezeichnet.

(*ii*) Eine Kantenbelegungsvektor $x \in \mathbb{R}^n$, der die **Erhaltungsbedingung** $A\,x = b$ erfüllt, heißt ein **Fluss** oder **Flussvektor** (englisch: *flow, flow vector*) auf diesem Netzwerk. Ein Fluss heißt **zulässig** (englisch: *admissible flow*), wenn zusätzlich die **Kapazitätsbeschränkungen** (englisch: *capacity constraints*) $0 \leqslant x \leqslant u$ erfüllt sind.

(*iii*) Eine lineare Optimierungsaufgabe der Form

$$\begin{aligned} \text{Minimiere} \quad & c^\mathsf{T} x \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & A\,x = b \\ \text{sowie} \quad & 0 \leqslant x \leqslant u \end{aligned} \tag{11.3}$$

mit gegebenem **Kantenkostenvektor** (englisch: *edge cost vector*) $c \in \mathbb{R}^n$ 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 11.6 (Kostenminimaler Fluss)**.**
Für den durch die Inzidenzmatrix (11.2) dargerstellten Digraphen aus Beispiel 11.1 und die Beispieldaten

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

erhalten wir den Fluss

$$x^* = (100,\ 0,\ 0,\ 200,\ 200,\ 100,\ 0,\ 150,\ 150,\ 150,\ 150)^\mathsf{T} \tag{11.4}$$

als (eindeutige) optimale Lösung der Aufgabe (11.3) des kostenminimalen Transports. Diese ist in Abbildung 11.2 dargestellt. Die zugehörigen minimalen Transportkosten betragen $c^\mathsf{T} x^* = 1320$. Die Lösung wurde unter Verwendung des Simplex-Verfahrens in `linprog` aus dem Modul `scipy.optimize` bestimmt, siehe Abbildung 11.3 für den Python-Code. Abbildung 11.4 zeigt eine alternative Lösung mit `cvxpy`. △

Beachte: Die Matrix A der Nebenbedingung $Ax = b$ hat hier $m = 9$ Zeilen, effektiv besitzt die Nebenbedingung jedoch nur $m = 8$ Gleichungen, da $\mathrm{Rang}(A) = 8$ beträgt. Da A außerdem $n = 11$ Spalten besitzt, hat jede Nichtbasis im Simplex-Verfahren die Mächtigkeit $\#N = 11 - 8 = 3$. 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.

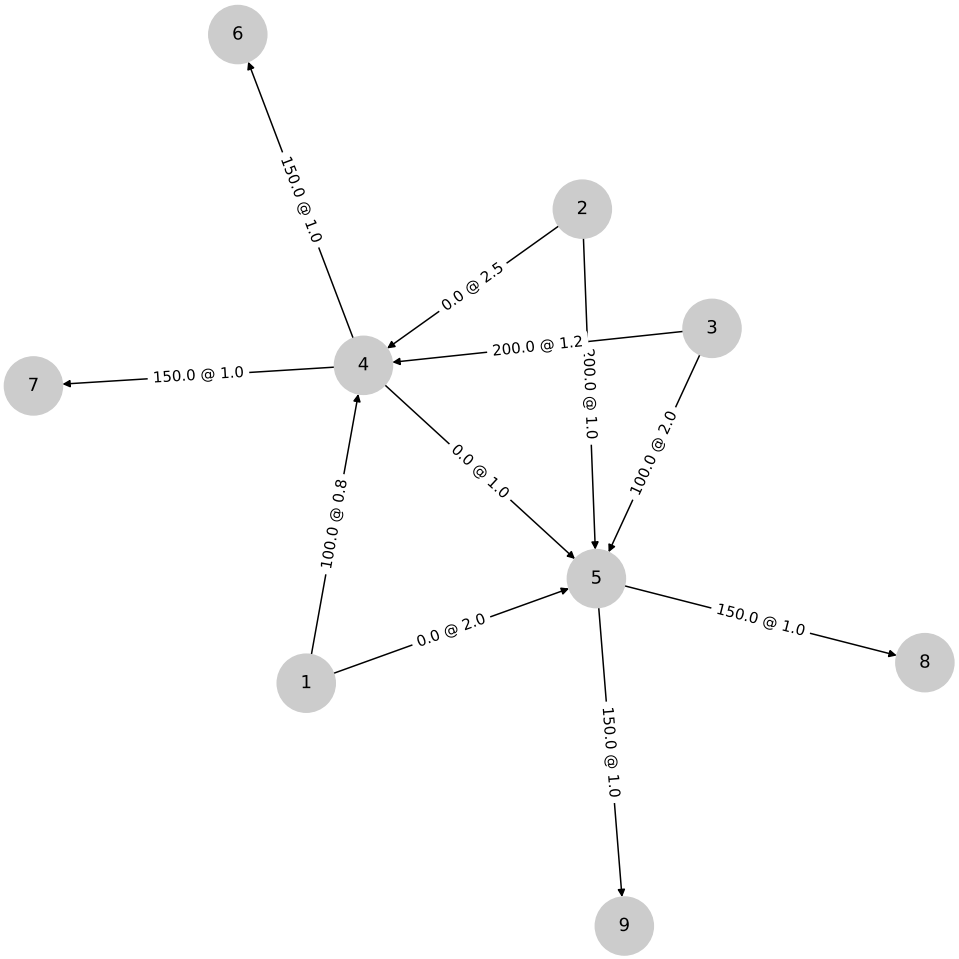


Abbildung 11.2.: Eine optimale Lösung von Beispiel 11.6. Gezeigt wird der jeweilige Fluss über die Kante und der zugehörige Eintrag im Kostenvektor (Kantenkosten).

```
# This code solves a minimal cost network flow problem
# using scipy.optimize.linprog.

# 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.: Bestimmung der Lösung von Beispiel 11.6 mit Hilfe von `linprog` aus dem Modul `scipy.optimize`.

```
# This code solves a minimal cost network flow problem
# using cvxpy.

# Resolve the dependencies.
import cvxpy as cp
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 = np.array([float(cost) for cost in 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])

# Define the problem in cvxpy.
n = A.shape[1]
x = cp.Variable(n)
objective = c.T @ x
constraints = [A @ x == b, x >= 0]
problem = cp.Problem(cp.Minimize(objective), constraints)

# Solve the problem
problem.solve()
```

Abbildung 11.4.: Bestimmung der Lösung von Beispiel 11.6 mit Hilfe des Moduls cvxpy.

Damit eine Aufgabe der Form (11.3) überhaupt zulässige Punkte (Flüsse) besitzt, muss notwendig $\mathbf{1}^\mathsf{T} b = 0$ gelten, denn

$$A\,x = b \quad \text{impliziert} \quad \underbrace{\mathbf{1}^\mathsf{T} A}_{=0}\, x = \mathbf{1}^\mathsf{T} b. \tag{11.5}$$

Die Bedarfe und Angebote in einem Transportnetzwerk müssen sich also ausgleichen. Sollte in einem Transportnetzwerk $\mathbf{1}^\mathsf{T} b < 0$ gelten, dann liegt ein **Überangebot** (oversupply) 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** (englisch: *artificial sink*) wird mit dann z. B. mit allen Angebotsknoten durch neue Kanten verbunden, und es werden Kosten für diese Kanten gesetzt.

Quizfrage 11.5: Was bedeutet es in Beispiel 11.1, wenn die künstliche Senke mit den drei Produktionsstätten verbunden wird? Und was bedeutet es, wenn sie mit den vier Verkaufsstellen verbunden wird? Wofür könnten dabei z. B. Kosten anfallen? Was bedeutet es, mehrere künstliche Senken in den Digraphen aufzunehmen?

Im Fall $\mathbf{1}^\mathsf{T} b > 0$ liegt dagegen ein **Mangel** (englisch: *undersupply*) an dem zu transportierenden Gut vor. Durch Schaffung eines **zusätzlichen Angebotsknotens** (englisch: *additional source*), den wir mit geeigneten Knoten im Netzwerk verbinden, können wir den Mangel kompensieren.

Quizfrage 11.6: 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?

Expertenwissen: Das Netzwerksimplex-Verfahren

Das Simplex-Verfahren ist nicht die effizienteste Lösungsmöglichkeit für Aufgaben kostenminimaler Flüsse auf Transportnetzwerken. Es gibt dafür eine spezielle Variante, das **Netzwerk-Simplex-Verfahren**, siehe etwa Gerdts, Lempio, 2011, Abschnitt 4.2 oder Vanderbei, 2008, Kapitel 14. Diese Variante 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.

Ende der Vorlesung 15

§ 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, wann man die Ganzzahligkeit der Lösung einer linearen Optimierungsaufgabe garantieren kann, ohne sie explizit zu fordern. Da das Simplex-Verfahren auf den Ecken der zulässigen Menge

$$\{x \in \mathbb{R}^n \mid A\,x = b,\ 0 \leqslant x \leqslant u\} \tag{12.1}$$

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 $1 \leqslant r \leqslant \min\{m, n\}$ die Eigenschaft $\det(\widehat{A}) \in \{0, \pm 1\}$ besitzt. △

Eine Matrix $\widehat{A}$ heißt dabei eine **Untermatrix** (englisch: *submatrix*) 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 Dimension $r = \min\{m, n\}$. Es gilt also $\det(\widehat{A}) = 1$ oder $\det(\widehat{A}) = -1$. Wir betrachten den Fall $r = m \leqslant n$. Die Einträge von $(\widehat{A})^{-1}$ können mit Hilfe der Cramerschen Regel wie folgt dargestellt werden:

$$\big((\widehat{A})^{-1}\big)_{ij} = \frac{1}{\det(\widehat{A})} \det\left[\widehat{a}^{(1)} \quad \cdots \quad \widehat{a}^{(i-1)} \quad e^{(j)} \quad \widehat{a}^{(i+1)} \quad \cdots \quad \widehat{a}^{(m)}\right].$$

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

$$\det(\widehat{A})\ \det\big((\widehat{A})^{-1}\big) = \det\big(\widehat{A}\,(\widehat{A})^{-1}\big) = \det(\mathrm{Id}) = 1$$

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

Satz 12.3 (Bedeutung unimodularer Matrizen)**.**
Es sei $A \in \mathbb{Z}^{m\times n}$ mit $\mathrm{Rang}(A) = m$. Dann sind die folgenden Aussagen äquivalent:

(*i*) Die Matrix A ist unimodular.

(*ii*) Für jeden Vektor $b \in \mathbb{Z}^m$ besitzt das Polyeder in Normalform

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

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

$$P_{\text{NFB}} := \{x \in \mathbb{R}^n \mid A\,x = b,\ x \geqslant 0,\ x \leqslant u\}$$

nur ganzzahlige Ecken.

Beweis. Der Beweis ist Inhalt von Hausaufgabe 8.2. □

Wir wenden uns nun der Bedeutung der totalen Unimodularität zu.

Lemma 12.4 (Totale Unimodularität verwandter Matrizen)**.**
Es sei $A \in \mathbb{Z}^{m\times n}$. Dann sind die folgenden Aussagen äquivalent:

(*i*) A ist total unimodular.

(*ii*) $-A$ ist total unimodular.

(*iii*) A^T ist total unimodular.

(*iv*) $\begin{bmatrix} A, & -A \end{bmatrix}$ ist total unimodular.

(*v*) $\begin{bmatrix} A, & \mathrm{Id}_m \end{bmatrix}$ ist total unimodular.

(*vi*) $\begin{bmatrix} A \\ \mathrm{Id}_n \end{bmatrix}$ 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

$$P_{\text{KF}} = \{x \in \mathbb{R}^n \mid A\,x \leqslant b,\ x \geqslant 0\}$$

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

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

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 (Inzidenzmatrizen einfacher Digraphen sind total unimodular; siehe Schrijver, 2003, Theorem 13.9)**.**
Inzidenzmatrizen einfacher Digraphen sind total unimodular.

Beweis. Es sei $\widehat{A}$ eine quadratische Untermatrix der Dimension r der Inzidenzmatrix A. Wir zeigen die Behauptung $\det(\widehat{A}) \in \{0, \pm 1\}$ durch Induktion über r. Der Fall $r = 1$ ist klar, da A und damit $\widehat{A}$ 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

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

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 11.6 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

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

Obwohl A_0 total unimodular ist, trifft das auf die Matrix

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

nicht mehr zu!

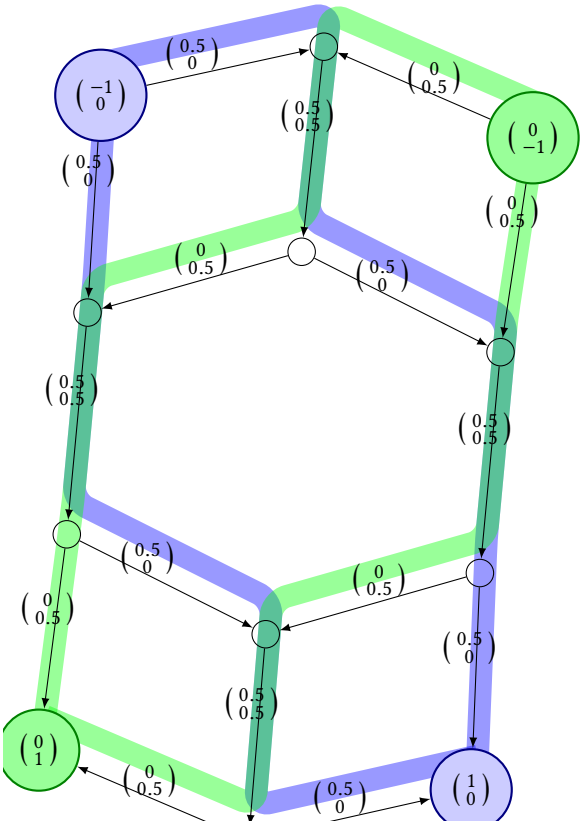


Abbildung 12.1.: Darstellung eines Transportnetzwerks für ein Zweigüterflussproblem. Die Menge eins des ersten Gutes (blau) soll von der Quelle oben links zur Senke unten rechts transportiert werden. Dieselbe Menge des zweiten Gutes (grün) soll von der Quelle oben rechts zur Senke unten links transportiert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

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 9.4.

Ende der Vorlesung 16

Ende der Woche 8

Kapitel 3 Konvexe Optimierung

§ 13 Einführung

Unser Ziel ist es auch in diesem Kapitel über konvexe Optimierungsaufgaben wieder, notwendige und hinreichende Optimalitätsbedingungen herzuleiten. 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 15.20 und Satz 16.22.

§ 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** (englisch: *convex*), 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 $\overline{B_\varepsilon(y)} = \{x \in \mathbb{R}^n \mid \|x - y\| \leqslant \varepsilon\}$,

(*iii*) Hyperebenen $H(a, \beta) = \{x \in \mathbb{R}^n \mid a^\mathsf{T} x = \beta\}$ mit $a \in \mathbb{R}^n$, $a \neq 0$ und $\beta \in \mathbb{R}$,

(*iv*) offene Halbräume $\{x \in \mathbb{R}^n \mid a^\mathsf{T} x < \beta\}$,

(*v*) abgeschlossene Halbräume $H^-(a, \beta) = \{x \in \mathbb{R}^n \mid a^\mathsf{T} x \leqslant \beta\}$ und $H^+(a, \beta) = \{x \in \mathbb{R}^n \mid a^\mathsf{T} x \geqslant \beta\}$,

(*vi*) das n-dimensionale **Einheitssimplex** (englisch: *unit simplex*) (siehe Abbildung 13.2)

$$\Delta_n := \left\{ x \in \mathbb{R}^n \,\middle|\, \sum_{i=1}^{n} x_i \leqslant 1,\ x_i \geqslant 0,\ i = 1, \dots, n \right\}.$$

(*vii*) das n-dimensionale **Standardsimplex** (englisch: *standard simplex, probability simplex*)

$$\left\{ x \in \mathbb{R}^{n+1} \,\middle|\, \sum_{i=1}^{n+1} x_i = 1,\ x_i \geqslant 0,\ i = 1, \dots, n+1 \right\}. \qquad \triangle$$

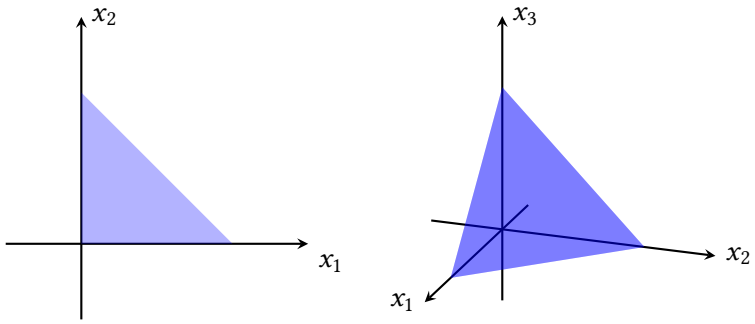


Abbildung 13.2.: Einheitssimplizes im $\mathbb{R}^2$ und $\mathbb{R}^3$.

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 $C_1 \times \cdots \times C_k$ konvex in $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$.

(*iii*) Es sei $f \colon \mathbb{R}^n \to \mathbb{R}^m$ eine (affin-)lineare Abbildung, also $f(x) = Ax + b$, und $C \subseteq \mathbb{R}^n$ sowie $D \subseteq \mathbb{R}^m$ 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 die **Skalierung** (englisch: *scaling*)

$$\beta C_1 = \{\beta x_1 \mid x_1 \in C_1\} \quad \text{für } \beta \in \mathbb{R}$$

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 ist die Verschiebung (**Translation**, englisch: *translation*) einer konvexen Menge, die sich ergibt, 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: Die Definition 13.1 können wir auch so lesen, dass eine Menge $C \subseteq \mathbb{R}^n$ genau dann konvex ist, 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?

Expertenwissen: Mittelpunktkonvexität von Mengen

Eine Menge $C \subseteq \mathbb{R}^n$ heißt **mittelpunkt-konvex** (englisch: *mid-point convex*), wenn mit $x, y \in C$ und $\alpha = \frac{1}{2}$ auch $\alpha\, x + (1-\alpha)\, y \in C$ ist, also der Mittelpunkt von x und y.

Jede konvexe Menge ist offensichtlich auch mittelpunkt-konvex. Gilt evtl. sogar die Äquivalenz?

Wir versuchen einen Beweis. Dazu sei C mittelpunkt-konvex und $x, y \in C$. Nach Voraussetzung ist $\alpha\, x + (1-\alpha)\, y \in C$ für $\alpha = \frac{1}{2}$. Indem wir x oder y durch den Mittelpunkt von x und y ersetzen, folgt, dass $\alpha\, x + (1-\alpha)\, y \in C$ liegt auch für $\alpha = \frac{1}{4}$ und $\alpha = \frac{3}{4}$. Per Induktion können wir zeigen, dass alle Koeffizienten $\alpha \in B := \left\{ \frac{r}{2^n} \,\middle|\, n \in \mathbb{N}_0,\ 0 \leqslant r \leqslant 2^n,\ r \in \mathbb{N}_0 \right\}$ möglich sind. (**Quizfrage 13.3:** Wie können wir Elemente der Menge B in Binärdarstellung angeben?)

Die Menge B liegt dicht in $[0,1]$. Wenn nun $\alpha \in [0,1]$ beliebig ist, dann gibt es eine Folge $\alpha^{(k)} \subseteq B$, sodass $\alpha^{(k)} \to \alpha$ konvergiert. Alle $\alpha^{(k)} x + (1-\alpha^{(k)})\, y$ liegen in C. Wenn nun C auch abgeschlossen wäre, dann würde auch der Grenzwert $\alpha\, x + (1-\alpha)\, y$ zu C gehören.

Wir haben damit gezeigt: Für *abgeschlossene* konvexe Mengen $C \subseteq \mathbb{R}^n$ ist die Mittelpunktkonvexität äquivalent zur Konvexität.

Definition 13.4 (Konvexkombination)**.**

(*i*) $x \in \mathbb{R}^n$ heißt eine **Konvexkombination** (englisch: *convex combination*) von $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$ mit $m \in \mathbb{N}$, falls $x = \sum_{i=1}^m \alpha^{(i)}\, x^{(i)}$ gilt mit Koeffizienten $\alpha^{(i)} \geqslant 0$ und $\sum_{i=1}^m \alpha^{(i)} = 1$. Eine solche Konvexkombination heißt **echt** (englisch: *proper*), 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 $x^{(1)}, \ldots, x^{(m)} \in M$ ist. △

Beachte: ~~Die Konvexkombination von $m = 0$ Vektoren ist per Definition immer der Nullvektor.~~ Bei $m = 2$ Vektoren können wir auch $\alpha\, x^{(1)} + (1-\alpha)\, x^{(2)}$ mit $\alpha \in [0,1]$ schreiben statt $\alpha^{(1)}\, x^{(1)} + \alpha^{(2)}\, x^{(2)}$ mit $\alpha^{(1)}, \alpha^{(2)} \geqslant 0$ und $\alpha^{(1)} + \alpha^{(2)} = 1$.

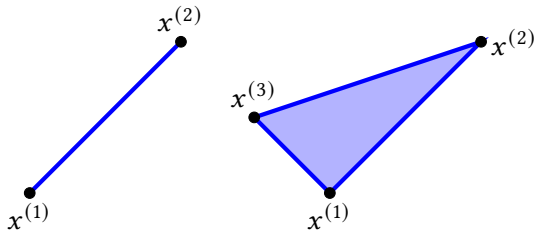


Abbildung 13.3.: Alle Konvexkombinationen von zwei und drei Punkten in $\mathbb{R}^2$.

Beachte: An dieser Stelle wird die Bemerkung nach Definition 6.15 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.

Per Definition 13.1 sind konvexe Mengen genau die Mengen, die alle Konvexkombinationen von je zwei Vektoren enthält.

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 $\alpha^{(1)}, \dots, \alpha^{(m)} \geqslant 0$ 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 13.4:** Warum?) Es sei bereits gezeigt, dass M alle Konvexkombinationen von höchstens m Elementen enthält.

Schluss auf $m+1$: Die Idee ist, eine Konvexkombination x von $m+1$ Punkten als Konvexkombination von zwei Punkten y und $x^{(m+1)}$ zu schreiben: Es seien $\alpha^{(i)} \geqslant 0$, $\sum_{i=1}^{m+1} \alpha^{(i)} = 1$ 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 $\beta^{(i)} \geqslant 0$ 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

$$\operatorname{conv}(M) = \bigcap \left\{ C \subseteq \mathbb{R}^n \,\middle|\, C \text{ ist konvex und } M \subseteq C \right\}, \tag{13.1}$$

heißt die **konvexe Hülle** (englisch: *convex hull*) von M. △

Beachte: Es gilt $M \subseteq \operatorname{conv}(M)$, und $\operatorname{conv}(M)$ ist als Schnitt konvexer Mengen wiederum eine konvexe Menge, daher der Name **konvexe Hülle**. $\operatorname{conv}(M)$ ist die kleinste konvexe Menge, die M enthält, genauer: $\operatorname{conv}(M)$ ist das eindeutige minimale Element der Teilmenge $\left\{ C \subseteq \mathbb{R}^n \,\middle|\, C \text{ ist konvex und } M \subseteq C \right\}$ im Sinne der Halbordnung der Mengeninklusion auf der Potenzmenge von $\mathbb{R}^n$.

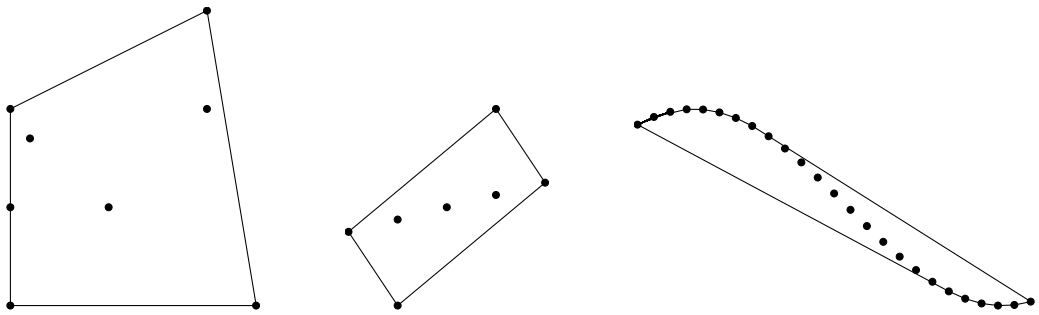


Abbildung 13.4.: Konvexe Hüllen einiger endlicher Punktmengen in $\mathbb{R}^2$.

Lemma 13.7 (Charakterisierung der konvexen Hülle)**.**
Es sei $M \subseteq \mathbb{R}^n$. Dann gilt:

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

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.

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

Es seien $x, y \in C$, also gibt es Zahlen $m, \ell \in \mathbb{N}$ und $\beta^{(1)}, \ldots, \beta^{(m)} \geqslant 0$ sowie $\gamma_1, \ldots, \gamma_\ell \geqslant 0$ mit $\sum_{i=1}^{m} \beta^{(i)} = 1$ und $\sum_{j=1}^{\ell} \gamma_j = 1$, sodass $x = \sum_{i=1}^{m} \beta^{(i)} x^{(i)}$ und $y = \sum_{j=1}^{\ell} \gamma_j y^{(j)}$ gelten mit irgendwelchen $x^{(1)}, \ldots, x^{(m)} \in M$ und $y^{(1)}, \ldots, y^{(\ell)} \in M$. Es sei $\alpha \in [0, 1]$. Dann gilt

$$\alpha x + (1 - \alpha)\, y = \alpha \sum_{i=1}^{m} \beta^{(i)} x^{(i)} + (1 - \alpha) \sum_{j=1}^{\ell} \gamma_j y^{(j)},$$

d. h., $\alpha x + (1 - \alpha)\, y$ ist Linearkombination der $\{x^{(i)}\}_{i=1}^{m} \cup \{y^{(j)}\}_{j=1}^{\ell}$. Die Koeffizienten sind $\geqslant 0$ und ergeben in der Summe 1. Damit ist $\alpha x + (1 - \alpha)\, y \in C$, also C konvex.

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

Folgerung 13.8 (Charakterisierungen der Konvexität)**.**
Für $M \subseteq \mathbb{R}^n$ sind äquivalent:

(*i*) M ist konvex.

(*ii*) M enthält alle Konvexkombinationen von M.

(*iii*) $M = \operatorname{conv}(M)$.

Beweis. Die Äquivalenz von Aussage (*i*) und Aussage (*ii*) gilt nach Lemma 13.5. Die Äquivalenz von Aussage (*ii*) und Aussage (*iii*) folgt aus Lemma 13.7. □

Ende der Vorlesung 17

§ 13.2 Konvexe Funktionen

Literatur: Geiger, Kanzow, 2002, Kapitel 2.1.2

Definition 13.9 (Konvexe Funktion)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex. Eine Funktion $f : C \to \mathbb{R}$ heißt

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

$$f(\alpha x + (1 - \alpha)\, y) \leqslant \alpha f(x) + (1 - \alpha) f(y) \tag{13.2}$$

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

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

$$f(\alpha x + (1 - \alpha)\, y) < \alpha f(x) + (1 - \alpha) f(y) \tag{13.3}$$

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

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

$$f(\alpha x + (1-\alpha)\, y) + \frac{\mu}{2}\,\alpha\,(1-\alpha)\|x-y\|^2 \leqslant \alpha f(x) + (1-\alpha) f(y) \tag{13.4}$$

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

(*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*.

Quizfrage 13.5: Was hat die μ-starke Konvexität von f mit der Konvexität von $f(\cdot) - \frac{\mu}{2}\|\cdot\|^2$ zu tun?

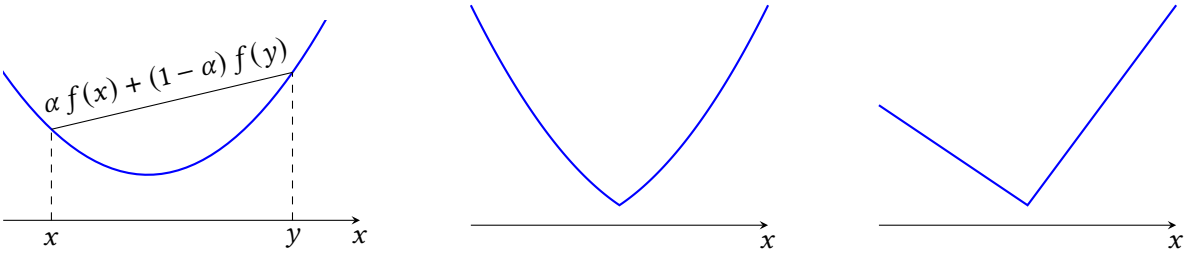


Abbildung 13.5.: Beispiele strikt konvexer Funktionen (links und Mitte) und konvexe, aber nicht strikt konvexe Funktion (rechts) auf Intervallen in $\mathbb{R}$.

Beachte: Für Funktionen $f\colon C \to \mathbb{R}$ wie in Definition 13.9 gilt:

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

Beispiel 13.10 (Beispiele konvexer Funktionen)**.**

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

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

- konvex $\quad\Leftrightarrow\quad$ Q ist positiv semidefinit,
- μ-stark konvex $\quad\Leftrightarrow\quad$ Q ist positiv definit mit kleinstem Eigenwert $\lambda_{\min}(Q) \geqslant \mu > 0$.

(*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. △

[1] In Geiger, Kanzow, 1999, Definition 3.2 wird diese Eigenschaft als **gleichmäßige Konvexität** (englisch: *uniform convexity*) bezeichnet. Das ist in der Literatur leider nicht einheitlich.

Quizfrage 13.6: Welche Konvexitätseigenschaften haben die 1-**Norm** $f(x) = \|x-z\|_1$ und die ∞-**Norm** $f(x) = \|x-z\|_\infty$ im $\mathbb{R}^n$?

Quizfrage 13.7: Gibt es außer affin-linearen Funktionen noch weitere Funktionen auf $\mathbb{R}^n$, die gleichzeitig konvex und konkav sind?

Expertenwissen: Ehegattensplitting und Konvexität

Was hat das sogenannte Ehegattensplitting mit Konvexität zu tun?

Betrachten wir ein Paar (Ehepaar oder eingetragene Lebenspartner), bei dem die Partner jährlich die zu versteuernden Einkommen y_1 und y_2 beziehen. Es bezeichne $T(y)$ die Steuerschuld einer einzeln veranlagten Person. T ist eine monoton wachsende und konvexe Funktion vom Einkommen.

In Deutschland kann das Paar wählen, wie die gemeinsame jährliche Steuerschuld $E(y_1, y_2)$ berechnet wird:

- (i) Einzelveranlagung (Individualbesteuerung) mit $E(y_1, y_2) = T(y_1) + T(y_2)$ und
- (ii) gemeinsame Veranlangug (Ehegattensplitting) mit $E(y_1, y_2) = 2\,T\big((y_1 + y_2)/2\big)$. Hierbei wird also das gemeinsame Einkommen rechnerisch auf die beiden Partner zu gleichen Teilen aufgeteilt. Dann werden diese aufgeteilten fiktiven Einkommen dem Steuertarif für einzeln veranlagte Personen unterworfen.

Aus der Konvexität der Steuerfunktion T folgt, dass das Ehegattensplitting zu einer geringeren Steuerlast führt als die Individualbesteuerung mit derselben Steuerfunktion, denn:

$$T\left(\frac{y_1 + y_2}{2}\right) \leqslant \frac{1}{2}T(y_1) + \frac{1}{2}T(y_2).$$

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** (englisch: *extended real-valued functions*). 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) $-\infty \leqslant a \leqslant \infty$ 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 $f\colon M \to \mathbb{R}$, die auf einer Teilmenge $M \subseteq \mathbb{R}^n$ definiert ist, auf ganz $\mathbb{R}^n$ fortsetzen, indem wir $\overline{f}\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ definieren als

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

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

(2) Wir können die zulässige Menge F einer Optimierungsaufgabe einfach dadurch in die Aufgabe einbauen, dass wir den Wert der Zielfunktion außerhalb der zulässigen Menge auf ∞ setzen. Dies gelingt einfach durch Addition der Indikatorfunktion I_F. Dadurch wird jede Optimierungsaufgabe formal zu einer unrestringierten Aufgabe.

Definition 13.11 (Indikatorfunktion)**.**
Es sei $M \subseteq \mathbb{R}^n$ irgendeine Menge. Die Funktion $I_M\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, definiert durch

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

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

Der Mehrwert von Funktionen, die auch den Wert $-\infty$ annehmen, wird bei den Richtungsableitungen in § 16.2 deutlich werden. In manchen Resultaten, etwa in Satz 13.18, wird dieser Fall auch ausgeschlossen.

Definition 13.12 (Eigentlicher Definitionsbereich, eigentliche Funktion)**.**
Es sei $f\colon M \to \mathbb{R} \cup \{\pm\infty\}$ eine Funktion.

(*i*) Die Menge

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

heißt der **eigentliche Definitionsbereich** (englisch: *(effective) domain*) von f.

(*ii*) Die Funktion f heißt **eigentlich** (englisch: *proper function*), wenn f nicht identisch ∞ (also $\operatorname{dom} f \neq \emptyset$) ist und nirgendwo den Wert $-\infty$ annimmt.

△

Wir erweitern jetzt die Definition 13.9 auf erweitert reellwertige Funktionen. Änderungen gegenüber Definition 13.9 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 13.13 (Erweitert reellwertige konvexe Funktion)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex.

(*i*) Eine Funktion $f\colon C \to \mathbb{R} \cup \{\pm\infty\}$ heißt **konvex** auf $\mathbb{R}^n$, falls

$$f(\alpha x + (1-\alpha)\, y) \leqslant \alpha f(x) + (1-\alpha) f(y) \tag{13.8}$$

gilt für alle $x, y \in C$ und $\alpha \in [0,1]$, für die die rechte Seite definiert ist.

(*ii*) Eine eigentliche Funktion $f\colon C \to \mathbb{R} \cup \{\infty\}$ heißt **strikt konvex** auf $\mathbb{R}^n$, falls

$$f(\alpha x + (1-\alpha)\, y) < \alpha f(x) + (1-\alpha) f(y) \tag{13.9}$$

gilt für alle $x, y \in \operatorname{dom} f \subseteq C$ mit $x \neq y$ und $\alpha \in (0,1)$.

(*iii*) Eine eigentliche Funktion $f\colon C \to \mathbb{R} \cup \{\infty\}$ heißt **μ-stark konvex** oder **stark konvex** mit Parameter $\mu > 0$ auf C, falls

$$f(\alpha x + (1-\alpha)\, y) + \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|^2 \leqslant \alpha f(x) + (1-\alpha) f(y) \tag{13.10}$$

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 13.8: Gilt für erweitert reellwertige, eigentliche Funktionen noch immer die Aussage „f stark konvex $\Rightarrow$ f strikt konvex $\Rightarrow$ f konvex“?

Satz 13.14 (Konvexität der Erweiterung konvexer Funktionen)**.**

(*i*) Es sei $C \subseteq \mathbb{R}^n$ konvex und $f\colon C \to \mathbb{R} \cup \{\pm\infty\}$ konvex. Dann ist die Fortsetzung $\overline{f}\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ auf $\mathbb{R}^n$ konvex.

(*ii*) Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex. Dann ist $\operatorname{dom} f \subseteq \mathbb{R}^n$ eine konvexe Menge, und die Einschränkung $f|_{\operatorname{dom} f}\colon \operatorname{dom} f \to \mathbb{R} \cup \{-\infty\}$ ist auf $\operatorname{dom} f$ 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 $\overline{f}$ unterscheiden.

Beweis. Aussage (*i*): Es sei $C \subseteq \mathbb{R}^n$ konvex und $f\colon C \to \mathbb{R} \cup \{\pm\infty\}$ konvex. Wir müssen zeigen:

$$\overline{f}(\alpha x + (1-\alpha)\, y) \leqslant \alpha \overline{f}(x) + (1-\alpha)\overline{f}(y) \tag{13.11}$$

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 $\overline{f}$ auf der gesamten Verbindungsstrecke zwischen x und y mit f übereinstimmt.

Falls $x \notin C$ liegt, also $\overline{f}(x) = \infty$, und $y \in C$ gilt, dann ist die rechte Seite in (13.11) entweder gleich ∞ für alle $\alpha \in [0,1]$ oder (im Fall $\overline{f}(y) = f(y) = -\infty$) 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 $\overline{f}(x) = \overline{f}(y) = \infty$, daher ist (13.11) ebenfalls für alle $\alpha \in [0,1]$ erfüllt.

Aussage (*ii*): Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex, es gilt also

$$f(\alpha\, x + (1-\alpha)\, y) \leqslant \alpha f(x) + (1-\alpha) f(y) \tag{13.12}$$

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 $\operatorname{dom} f \subseteq \mathbb{R}^n$ konvex ist. Es seien dazu $x, y \in \operatorname{dom} f$. Dann ist die rechte Seite in (13.12) für alle $\alpha \in [0,1]$ definiert. (**Quizfrage 13.9:** Warum?) Aus (13.12) folgt

$$f(\alpha\, x + (1-\alpha)\, y) \leqslant \alpha \underbrace{f(x)}_{<\infty} + (1-\alpha) \underbrace{f(y)}_{<\infty} < \infty$$

für alle $\alpha \in [0,1]$, sodass die linke Seite nicht ∞ ist. Also gehört $\alpha\, x + (1-\alpha)\, y$ zu $\operatorname{dom} f$. Dieselbe Ungleichung zeigt auch bereits, dass $f|_{\operatorname{dom} f}\colon \operatorname{dom} f \to \mathbb{R}$ wie behauptet konvex ist. □

Folgerung 13.15 (Konvexität der Indikatorfunktion)**.** Die Indikatorfunktion $I_M\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ einer Menge $M \subseteq \mathbb{R}^n$ ist genau dann konvex, wenn die Menge M konvex ist.

Mit Hilfe der folgenden Definition kann ein sehr nützlicher Zusammenhang zwischen konvexen Funktionen und konvexen Mengen hergestellt werden.

Definition 13.16 (Epigraph einer Funktion)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ eine Funktion. Die Menge

$$\operatorname{epi} f := \left\{ \binom{x}{\gamma} \in \mathbb{R}^n \times \mathbb{R} \,\middle|\, \gamma \geqslant f(x) \right\} \tag{13.13}$$

heißt der **Epigraph** (englisch: *epigraph*) von f. △

Abbildung 13.6 illustriert das Konzept des Epigraphen einer Funktion. **Beachte:** Aus dem Epigraphen einer Funktion können wir wie folgt die Funktionswerte rekonstruieren:

$$f(x) = \inf\{\gamma \in \mathbb{R} \mid (x,\gamma) \in \operatorname{epi} f\}. \tag{13.14}$$

Es gilt folgende wichtige Charakterisierung konvexer Funktionen:

Satz 13.17 (Epigraph-Charakterisierung konvexer Funktionen)**.**
Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ ist genau dann konvex, wenn $\operatorname{epi} f$ eine konvexe Menge ist.

Beweis. Es sei zunächst f konvex, und es seien $\binom{x}{\gamma}$ und $\binom{y}{\delta}$ Punkte in $\operatorname{epi} f$. Insbesondere ist also $f(x) < \infty$ und $f(y) < \infty$. Die Ungleichung (13.8) gilt also für alle $\alpha \in [0,1]$, d. h.,

$$f(\alpha\, x + (1-\alpha)\, y) \leqslant \alpha\, f(x) + (1-\alpha)\, f(y) \leqslant \alpha\, \gamma + (1-\alpha)\, \delta.$$

Das bedeutet aber

$$\begin{pmatrix} \alpha\, x + (1-\alpha)\, y \\ \alpha\, \gamma + (1-\alpha)\, \delta \end{pmatrix} = \alpha \begin{pmatrix} x \\ \gamma \end{pmatrix} + (1-\alpha) \begin{pmatrix} y \\ \delta \end{pmatrix} \in \operatorname{epi} f.$$

Also ist $\operatorname{epi} f$ konvex.

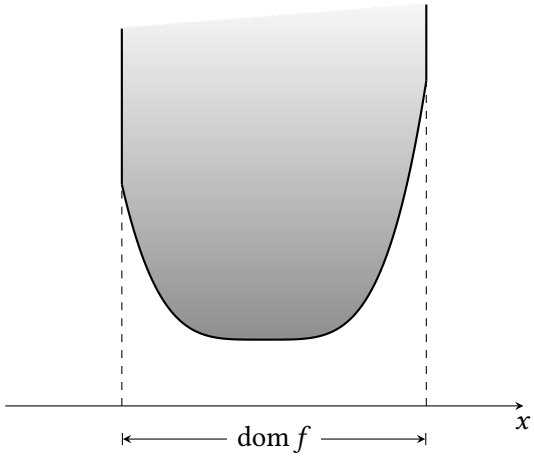


Abbildung 13.6.: Epigraph einer Funktion $f\colon \mathbb{R} \to \mathbb{R} \cup \{\infty\}$.

Umgekehrt sei nun 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 $\binom{x}{f(x)}$ und $\binom{y}{f(y)}$ zu epi f. Die Konvexität von epi f zeigt, dass auch $\alpha \binom{x}{f(x)} + (1-\alpha)\binom{y}{f(y)}$ zu 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 $\binom{x}{\gamma}$ zu epi f für jedes $\gamma \in \mathbb{R}$. Weiter sei $\binom{y}{\delta}$ ebenfalls in epi f. Die Konvexität von epi f impliziert, dass auch $\alpha \binom{x}{\gamma} + (1-\alpha)\binom{y}{\delta}$ zu epi f gehört, also gilt

$$f(\alpha x + (1-\alpha)\, y) \leqslant \alpha\gamma + (1-\alpha)\,\delta$$

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 α 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 13.18 (Operationen auf konvexen Funktionen)**.**

(*i*) Sind $f^{(i)}\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex auf $\mathbb{R}^n$ und $\beta^{(i)} \geqslant 0$ für $i = 1, \ldots, m$, dann ist die durch

$$f(x) := \sum_{i=1}^{m} \beta^{(i)} f^{(i)}(x)$$

definierte Funktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex auf $\mathbb{R}^n$.

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.

(*ii*) Sind die Funktionen $f^{(i)} : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ 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 $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex auf $\mathbb{R}^n$.

(*iii*) Ist $g : \mathbb{R}^n \to \mathbb{R}^m$ affin-linear und $f : \mathbb{R}^m \to \mathbb{R} \cup \{\pm\infty\}$ konvex, so ist $(f \circ g) : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex auf $\mathbb{R}^n$.

(*iv*) Ist $g : \mathbb{R}^n \to \mathbb{R}$ konvex und ist $f : \mathbb{R} \to \mathbb{R} \cup \{\pm\infty\}$ konvex und monoton wachsend, so ist $(f \circ g) : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex.

Beachte: Aus Aussage (*iv*) folgt insbesondere, dass die Funktion g^2 konvex auf $\mathbb{R}^n$ ist, wenn $g : \mathbb{R}^n \to \mathbb{R}$ konvex ist und $g(x) \geqslant 0$ gilt für alle $x \in \mathbb{R}^n$. (**Quizfrage 13.10:** Genaue Begründung?)

Beweis von Satz 13.18. Der Beweis ist Inhalt von Hausaufgabe 9.3. □

Expertenwissen: Mittelpunktkonvexität von Funktionen

Eine Funktion $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ heißt **mittelpunkt-konvex** (englisch: *mid-point convex*), wenn mit $x, y \in \mathbb{R}^n$ und $\alpha = \frac{1}{2}$ die Ungleichung $f(\alpha x + (1-\alpha)\, y) \leqslant \alpha f(x) + (1-\alpha) f(y)$ gilt.

Jede konvexe Funktion ist offensichtlich auch mittelpunkt-konvex. Gilt evtl. sogar die Äquivalenz?

Wir versuchen einen Beweis, erwarten aber wie bei der Mittelpunktkonvexität von Mengen, dass noch eine zusätzliche Bedingung an die Funktion f gestellt werden muss.

Dazu sei f mittelpunkt-konvex und $x, y \in \mathbb{R}^n$. Wieder kann man per Induktion zeigen, dass die Mittelpunktkonvexität impliziert, dass

$$f(\alpha x + (1-\alpha)\, y) \leqslant \alpha f(x) + (1-\alpha) f(y)$$

für alle $\alpha \in B := \left\{ \frac{r}{2^n} \,\middle|\, n \in \mathbb{N}_0,\ 0 \leqslant r \leqslant 2^n,\ r \in \mathbb{N}_0 \right\}$ gilt.

Die Menge B liegt dicht in $[0,1]$. Wenn nun $\alpha \in [0,1]$ beliebig ist, dann gibt es eine Folge $\alpha^{(k)} \subseteq B$, sodass $\alpha^{(k)} \to \alpha$ konvergiert. Es gilt wie gesagt

$$f(\alpha^{(k)} x + (1-\alpha^{(k)})\, y) \leqslant \alpha^{(k)} f(x) + (1-\alpha^{(k)}) f(y).$$

Die rechte Seite konvergiert gegen $\alpha f(x) + (1-\alpha) f(y)$. Wenn nun die Funktion f auch noch unterhalbstetig ist, dann gilt

$$f(\alpha x + (1-\alpha)\, y) \leqslant \lim_{k\to\infty} f(\alpha^{(k)} x + (1-\alpha^{(k)})\, y) \leqslant \lim_{k\to\infty} \alpha^{(k)} f(x) + (1-\alpha^{(k)}) f(y) = \alpha f(x) + (1-\alpha) f(y).$$

Wir haben damit gezeigt: Für *unterhalbstetige* konvexe Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ ist die Mittelpunktkonvexität äquivalent zur Konvexität.

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 13.19 und Satz 13.20 nur für reellwertige konvexe Funktionen.

Ende der Vorlesung 18

Ende der Woche 9

Satz 13.19 (Charakterisierung konvexer Funktionen mittels erster Ableitung)**.**
Es sei $C \subseteq \mathbb{R}^n$ offen und konvex sowie $f : C \to \mathbb{R}$ diffbar. Dann gelten:

(a) Es sind äquivalent:

(i) f ist konvex auf C.

(ii) Für alle $x, y \in C$ gilt:

$$f(x) - f(y) \geqslant f'(y)(x - y). \tag{13.15}$$

(iii) Für alle $x, y \in C$ gilt:

$$\big(f'(x) - f'(y)\big)(x - y) \geqslant 0. \tag{13.16}$$

Man sagt zu (13.16), die Ableitung f' sei auf C ein **monotoner Operator** (englisch: *monotone operator*).

(b) Es sind äquivalent:

(i) f ist strikt konvex auf C.

(ii) Für alle $x, y \in C$ mit $x \neq y$ gilt:

$$f(x) - f(y) > f'(y)(x - y). \tag{13.17}$$

(iii) Für alle $x, y \in C$ mit $x \neq y$ gilt:

$$\big(f'(x) - f'(y)\big)(x - y) > 0. \tag{13.18}$$

Man sagt zu (13.18), die Ableitung f' sei auf C ein **strikt monotoner Operator** (englisch: *strictly monotone 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:

$$f(x) - f(y) \geqslant f'(y)(x - y) + \frac{\mu}{2} \|x - y\|^2. \tag{13.19}$$

(iii) Es existiert $\mu > 0$, sodass für alle $x, y \in C$ gilt:

$$\big(f'(x) - f'(y)\big)(x - y) \geqslant \mu \|x - y\|^2. \tag{13.20}$$

Man sagt zu (13.20), die Ableitung f' sei auf C ein **stark monotoner Operator** (englisch: *strongly monotone operator*).

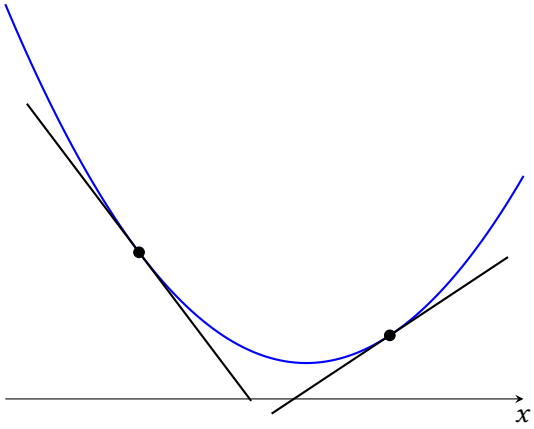


Abbildung 13.7.: Charakterisierung der Konvexität diffbarer Funktionen über ihre Tangenten.

Beachte: Nach Aussage (a) ist eine diffbare Funktion $f\colon C \to \mathbb{R}$ genau dann konvex, wenn der Graph oberhalb aller seiner Tangentialebenen

$$T(x;y) := f(y) + f'(y)(x-y) \quad \text{(Tangentialebene an } f \text{ im Punkt } y \in C)$$

verläuft, siehe Abbildung 13.7. Anders ausgedrückt: Eine diffbare Funktion ist genau dann konvex, wenn alle Taylormodelle erster Ordnung die Funktion unterschätzen.

Beweis. Wir zeigen nur die Aussage (c) über die Charakterisierung der starken Konvexität. Die Aussagen (a) und (b) lassen sich analog beweisen.

Aussage (i) $\Rightarrow$ Aussage (ii): Es sei f stark konvex auf C und $x, y \in C$, $\alpha \in (0,1)$. Dann gilt

$$\begin{aligned} & f(y+\alpha(x-y)) = f(\alpha x + (1-\alpha)\, y) \leqslant \alpha f(x) + (1-\alpha) f(y) - \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|^2 \\ \Rightarrow \quad & \frac{f(y+\alpha(x-y)) - f(y)}{\alpha} \leqslant f(x) - f(y) - \frac{\mu}{2}(1-\alpha)\,\|x-y\|^2 \\ \Rightarrow \quad & f'(y)(x-y) \leqslant f(x) - f(y) - \frac{\mu}{2}\|x-y\|^2 \quad \text{(Grenzübergang } \alpha \searrow 0\text{)}, \end{aligned}$$

d. h., es gilt (13.19).

Aussage (ii) $\Rightarrow$ Aussage (i): Es gelte (13.19), und es seien $x, y \in C$ und $\alpha \in (0,1)$. Setze $z := \alpha x + (1-\alpha)\, y$. Eine zweimalige Anwendung von (13.19) ergibt

$$\begin{aligned} f(x) - f(z) &\geqslant f'(z)(x-z) + \frac{\mu}{2}\|x-z\|^2, \\ f(y) - f(z) &\geqslant f'(z)(y-z) + \frac{\mu}{2}\|y-z\|^2. \end{aligned}$$

Wir multiplizieren die erste Ungleichung mit α, die zweite mit $(1-\alpha)$ und addieren:

$$\alpha f(x) + (1-\alpha) f(y) - f(z) \geqslant \underbrace{\alpha f'(z)x + (1-\alpha) f'(z)y - f'(z)z}_{=0 \text{ nach Definition von } z} + \underbrace{\frac{\mu}{2}\alpha\,\|x-z\|^2 + \frac{\mu}{2}(1-\alpha)\|y-z\|^2}_{=\frac{\mu}{2}\alpha(1-\alpha)\|x-y\|^2 \text{ (nachrechnen)}}.$$

Durch Einsetzen von $z = \alpha\, x + (1-\alpha)\, y$ folgt schließlich

$$\alpha f(x) + (1-\alpha) f(y) - f(\alpha\, x + (1-\alpha)\, y) \geqslant \frac{\mu}{2}\, \alpha\, (1-\alpha) \|x-y\|^2,$$

d. h., f ist stark konvex.

Aussage (ii) $\Rightarrow$ Aussage (iii): Es seien $x, y \in C$. Eine zweimalige Anwendung von (13.19) ergibt

$$\begin{aligned} f(x) - f(y) &\geqslant f'(y)(x-y) + \frac{\mu}{2} \|x-y\|^2, \\ f(y) - f(x) &\geqslant f'(x)(y-x) + \frac{\mu}{2} \|x-y\|^2, \end{aligned}$$

und aus der Addition der Ungleichungen folgt (13.20).

Aussage (iii) $\Rightarrow$ Aussage (ii): Es seien $x, y \in C$. Wir betrachten die Funktion $t \mapsto D(t) := f(y + t\,(x-y))$ und deren Ableitung $t \mapsto d(t) := f'(y + t\,(x-y))(x-y)$ auf $[0,1]$. Wir zeigen zunächst, dass d auf $[0,1]$ stark monoton ist. Es seien dazu $s, t \in [0,1]$ beliebig. Dann ist

$$\begin{aligned} (d(t) - d(s))(t-s) &= \big[f'(y + t\,(x-y))(x-y) - f'(y + s\,(x-y))(x-y)\big](t-s) \\ &= \big[f'(y + t\,(x-y)) - f'(y + s\,(x-y))\big](t-s)(x-y) \\ &\geqslant \mu \|(t-s)(x-y)\|^2 \quad \text{wegen (13.20)} \\ &= \mu \|x-y\|^2 |t-s|^2. \end{aligned}$$

Da monotone Funktionen Riemann-integrierbar sind (Heuser, 2003, Satz 83.3), ist der Hauptsatz der Differential- und Integralrechnung (Heuser, 2003, Satz 79.1) anwendbar, und es folgt

$$D(1) - D(0) = \int_0^1 d(t)\, \mathrm{d}t.$$

Daher gilt weiter

$$\begin{aligned} D(1) - D(0) - d(0) &= \int_0^1 [d(t) - d(0)]\, \mathrm{d}t \\ &= \int_0^1 \frac{1}{t}\, [d(t) - d(0)]\, (t-0)\, \mathrm{d}t \\ &\geqslant \mu \|x-y\|^2 \int_0^1 t\, \mathrm{d}t \\ &= \frac{\mu}{2} \|x-y\|^2. \end{aligned}$$

Das Einsetzen der Definitionen von D und d ergibt schließlich

$$f(x) - f(y) - f'(y)(x-y) \geqslant \frac{\mu}{2} \|x-y\|^2,$$

also (13.19). □

Satz 13.20 (Charakterisierung konvexer Funktionen mittels zweiter Ableitungen)**.**
Es sei $C \subseteq \mathbb{R}^n$ offen und konvex sowie $f\colon C \to \mathbb{R}$ 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 $\lambda_{\min}(f''(x)) \geqslant \mu > 0$ 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 13.19 (c) ist f' dann stark monoton auf C, erfüllt also (13.20). Für beliebiges $x \in C$ und $d \in \mathbb{R}^n$ ist daher

$$\begin{aligned} d^\mathsf{T} f''(x)\, d &= \lim_{t\to 0} \frac{f'(x+td) - f'(x)}{t}\, d \\ &= \lim_{t\to 0} \frac{f'(x+td) - f'(x)}{t^2}\, (td) \\ &\geqslant \lim_{t\to 0} \frac{1}{t^2}\, \mu \, \|td\|^2 \quad \text{wegen (13.20)} \\ &= \mu \, \|d\|^2. \end{aligned}$$

Daraus folgt, dass $\lambda_{\min}(f''(x)) \geqslant \mu$ ist.

(*ii*) $\Rightarrow$ (*i*): Es gelte nun umgekehrt $\lambda_{\min}(f''(x)) \geqslant \mu > 0$ 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 $d(t) = (x-y)^\mathsf{T} f''(y + t\,(x-y))\,(x-y)$ auf $[0,1]$ nach Voraussetzung nichtnegativ ist.

Aus dem Hauptsatz der Differential- und Integralrechnung, angewendet auf die Funktion D und ihre Ableitung d, folgt

$$D(1) - D(0) = \int_0^1 d(t)\, \mathrm{d}t,$$

also

$$\begin{aligned} [f'(x) - f'(y)]\,(x-y) &= \int_0^1 (x-y)^\mathsf{T} f''(y + t\,(x-y))\,(x-y)\, \mathrm{d}t \\ &\geqslant \mu \int_0^1 \|x-y\|^2\, \mathrm{d}t \\ &= \mu\, \|x-y\|^2. \end{aligned}$$

Das heißt, f' ist stark monoton, und nach Satz 13.19 (c) ist f stark konvex.

Der Beweis von Aussage (a) erfolgt genau auf die gleiche Weise mit $\mu = 0$. Zum Beweis von Aussage (b) nehmen wir an, dass $f''(x)$ für alle $x \in C$ positiv definit ist. Es seien $x, y \in C$, $x \neq y$, und wir setzen D und d wie oben. Dann ist $d(t) > 0$ für alle $t \in [0,1]$. Wieder mit dem Hauptsatz der Differential- und Integralrechnung folgt

$$[f'(x) - f'(y)]\,(x-y) = \int_0^1 (x-y)^\mathsf{T} f''(y + t\,(x-y))\,(x-y)\, \mathrm{d}t > 0.$$

Das heißt, f' ist strikt monoton, und aus Satz 13.19 (b) folgt die strikte Konvexität von f auf C. □

§ 14 Konvexe Optimierungsaufgaben

Wir betrachten die **konvexe Optimierungsaufgabe** (englisch: *convex optimization problem, convex programming problem*)

$$\text{Minimiere} \quad f(x) \quad \text{über } x \in \mathbb{R}^n \tag{14.1}$$

mit konvexer Zielfunktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$. Diese Problemklasse enthält insbesondere Aufgaben der Form

$$\text{Minimiere} \quad g(x) \quad \text{über } x \in C, \tag{14.2}$$

wobei die zulässige Menge $C \subseteq \mathbb{R}^n$ konvex und $g\colon C \to \mathbb{R}$ eine konvexe reellwertige Funktion ist. Setzen wir dann $f := g + I_C$ mit der Indikatorfunktion I_C von C, setzen also g durch den Wert ∞ außerhalb von C fort, so ergibt sich eine Aufgabe der Form (14.1). Formal handelt es sich bei (14.1) um eine unrestringierte Optimierungsaufgabe, die jedoch implizit die Nebenbedingung $x \in C = \operatorname{dom} f$ enthält. 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 die wichtigsten Begriffe hier mit einer Hervorhebung von Unterschieden.

Definition 14.1 (Grundbegriffe für die konvexe Optimierung)**.**

(*i*) Der Wert

$$f^* := \inf\{f(x) \mid x \in \mathbb{R}^n\}$$

heißt der **Infimalwert** der Aufgabe (14.1).

(*ii*) Ein Punkt $x^* \in \operatorname{dom} f$ heißt ein **globaler Minimierer**, **globale Minimalstelle** oder **global optimale Lösung**, wenn gilt:

$$f(x^*) \leqslant f(x) \text{ für alle } x \in \mathbb{R}^n.$$

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 $x^* \in \operatorname{dom} f$ heißt ein **lokaler Minimierer**, **lokale Minimalstelle** oder **lokal optimale Lösung**, wenn es eine Umgebung $U(x^*)$ gibt, sodass gilt:

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

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 $x^* \in \operatorname{dom} f$ hinzugefügt. Der Funktionswert dort darf also nicht ∞ 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 Infimalwert f^* der Aufgabe (14.1) ist wieder entweder $f^* = -\infty$, oder f^* ist endlich, oder es gilt $f^* = \infty$. Letzteres ist genau dann der Fall, wenn $f \equiv \infty$ ist, also $\operatorname{dom} f = \emptyset$ gilt. Diesen Fall werden wir aber im Folgenden oft ausschließen, ebenso wie den Fall, dass f den Wert $-\infty$ annimmt. Mit anderen Worten: Wir werden oft annehmen, dass f eine **eigentliche Funktion** ist (Definition 13.12).

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! Insbesondere gilt das für lineare Optimierungsaufgaben, die wir in der Form (14.2) schreiben können und daher auch in der Form (14.1), vgl. Satz 6.1.

Beweis. Aussage (*i*): Es sei x^* lokaler Minimierer, d. h., es gilt $f(x^*) < \infty$, und es existiert eine Umgebung $U(x^*)$ mit $f(x^*) \leqslant f(x)$ 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 $\widehat{x} \in \mathbb{R}^n$ mit $f(\widehat{x}) < f(x^*)$. Für $\alpha \in (0, 1]$ gilt dann

$$f(\alpha\,\widehat{x} + (1-\alpha)\,x^*) \leqslant \alpha f(\widehat{x}) + (1-\alpha) f(x^*) < f(x^*),$$

wobei alle Terme wohldefiniert sind. Für α hinreichend klein liegt aber $\alpha\,\widehat{x} + (1-\alpha)\,x^* \in U(x^*)$, im Widerspruch zur lokalen Optimalität von x^*. Also kann ein solches $\widehat{x}$ nicht existieren, d. h., es ist

$$f(x^*) \leqslant f(x) \quad \text{für alle } x \in \mathbb{R}^n.$$

Aussage (*ii*): Es seien x^* und x^{**} Lösungen von (14.1), also $f(x^*) = f(x^{**}) = f^*$ (Infimalwert). Für $\alpha \in [0, 1]$ gilt

$$f(\alpha\,x^* + (1-\alpha)\,x^{**}) \leqslant \alpha f(x^*) + (1-\alpha) f(x^{**}) = f^*.$$

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 $x^*, x^{**} \in \operatorname{dom} f$. 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 $\operatorname{dom} f$, daher folgt aus der strikten Konvexität (13.9)

$$f\Big(\frac{x^* + x^{**}}{2}\Big) < \frac{1}{2} f(x^*) + \frac{1}{2} f(x^{**}) = f^*,$$

im Widerspruch zur Optimalität von x^* und x^{**}. □

Ende der Vorlesung 19

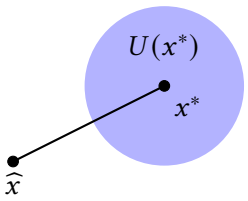


Abbildung 14.1.: Illustration zum Beweis von Satz 14.2 (*i*) (lokale Minimierer sind globale Minimierer).

§ 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 drei Aussagen dazu: der Trennungssatz 15.28, der eigentliche Trennungssatz 15.32 und der strikte Trennungssatz 15.35 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 (Orthogonale Projektionsaufgabe, vgl. Beispiel 1.5)**.**
Zu einer nichtleeren, abgeschlossenen, konvexen Menge $C \subseteq \mathbb{R}^n$ und einem gegebenen Punkt $p \in \mathbb{R}^n$ suchen wir einen Punkt $x \in C$, der p im Sinne der Euklidischen Norm am nächsten liegt. Als Optimierungsaufgabe können wir dies in der Form

$$\text{Minimiere} \quad f(x) := \|x - p\| + I_C(x) \quad \text{über } x \in \mathbb{R}^n \tag{15.1}$$

oder auch als (vgl. Hausaufgabe 1.2)

$$\text{Minimiere} \quad g(x) := \frac{1}{2}\|x - p\|^2 + I_C(x) \quad \text{über } x \in \mathbb{R}^n \tag{15.2}$$

schreiben. Beide Zielfunktionen in (15.1) und (15.2) sind eigentlich (**Quizfrage 15.1:** Warum?). △

Lemma 15.2 (orthogonale 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** (englisch: *orthogonal projection*) von p auf C **im Sinne des Euklidischen Innenprodukts**, 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 wie eine Hilfsaufgabe ein. Dazu wählen wir ein $w \in C$ beliebig und definieren B als die kompakte Kugel

$$B := \overline{B_r(p)} \quad \text{mit } r := \|p - w\|.$$

Die Hilfsaufgabe lautet

$$\begin{aligned} \text{Minimiere} \quad & \frac{1}{2}\|x-p\|^2 \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & x \in C \cap B. \end{aligned} \tag{15.3}$$

Ein Punkt $x^* \in C$ ist genau dann ein globaler Minimierer von (15.2), wenn er ein globaler Minimierer von (15.3) ist (**Quizfrage 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 1.9 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 1.2 und der strikten Monotonie der Wurzelfunktion auf $\mathbb{R}_{\geqslant 0}$ 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. □

Wir werden nun die eindeutige Lösung der Aufgabe (15.1), (15.2) der orthogonalen Projektion charakterisieren. Das ist ein zentrales Resultat, aus dem viele weitere folgen, insbesondere die Trennungssätze.

Satz 15.3 (orthogonaler Projektionssatz: notwendige und hinreichende Bedingungen für (15.2))**.**

Es sei $C \subseteq \mathbb{R}^n$ nichtleer, abgeschlossen und konvex und $p \in \mathbb{R}^n$. Es gilt $x^* = \operatorname{proj}_C(p)$ genau dann, wenn $x^* \in C$ ist und gilt:

$$(x^* - p)^\mathsf{T}(x - x^*) \geqslant 0 \quad \text{für alle } x \in C. \tag{15.4}$$

Beachte: (15.4) ist eine **Variationsungleichung** (englisch: *variational inequality*). Sie besagt, dass der Winkel zwischen $x^* - p$ und $x - x^*$ 90° nicht übersteigen darf. Anders ausgedrückt: C ist enthalten im Halbraum $H^+(a,\beta) = \{x \in \mathbb{R}^n \mid a^\mathsf{T}x \geqslant \beta\}$ mit Normalenvektor $a = x^* - p$ und $\beta = a^\mathsf{T}x^*$, vgl. Abbildung 15.1.

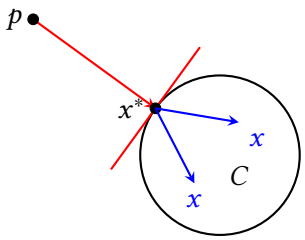


Abbildung 15.1.: Der Winkel zwischen $x^* - p$ und $x - x^*$ darf 90° nicht übersteigen.

Beweis von Satz 15.3. Wir definieren wie in (15.1) und (15.2) die Funktionen

$$f(x) := \|x-y\| + I_C(x) \quad \text{und} \quad g(x) := \frac{1}{2}\|x-y\|^2 + I_C(x).$$

„⇒“: Es sei $x^* = \operatorname{proj}_C(p)$, also insbesondere $x^* \in C$. Dann gilt $x^* + \alpha\,(x - x^*) \in C$ für alle $x \in C$ und $\alpha \in (0, 1)$. Aus der Optimalität von x^* folgt

$$\begin{aligned} & \frac{1}{2}\|x^* - p\|^2 = g(x^*) \leqslant g(x^* + \alpha\,(x - x^*)) = \frac{1}{2}\|(x^* - p) + \alpha\,(x - x^*)\|^2 \\ \Rightarrow \quad & 0 \leqslant \alpha\,(x^* - p)^\mathsf{T}(x - x^*) + \frac{\alpha^2}{2}\|x - x^*\|^2. \end{aligned}$$

Division durch α und Grenzübergang $\alpha \searrow 0$ liefern die Behauptung (15.4).

„⇐“: Es gelte $x^* \in C$ und (15.4). Daraus folgt

$$\begin{aligned} 0 &\geqslant (p - x^*)^\mathsf{T}(x - x^*) \quad \text{für alle } x \in C \\ &= (p - x^*)^\mathsf{T}(x - p + p - x^*) \\ &= (p - x^*)^\mathsf{T}(x - p) + \|p - x^*\|^2 \\ &\geqslant -\|p - x^*\|\|x - p\| + \|p - x^*\|^2 \quad \text{(Cauchy-Schwarz).} \end{aligned}$$

Daraus folgt weiter $f(x) = \|x - p\| \geqslant \|p - x^*\| = f(x^*)$ für alle $x \in C$, d. h., $x^* = \operatorname{proj}_C(p)$. □

§ 15.2 Affine Unterräume

Definition 15.4 (Affiner Unterraum)**.**
Eine nichtleere 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** (englisch: *direction space, direction*) von A. Der Vektor x_0 heißt ein **Stützvektor** (englisch: *support point*) von A. Zwei affine Unterräume heißen **parallel** (englisch: *parallel*), wenn sie denselben Richtungsraum besitzen.

Beweis von Lemma 15.5. „⇐“: Es sei A eine Menge von der Form (15.5). Weiter seien $x_1, x_2 \in A$, also $x_1 = u_1 + x_0$ und $x_2 = u_2 + x_0$ mit $u_1, u_2 \in U$. Schließlich sei $\alpha \in \mathbb{R}$. Dann ist auch

$$\alpha\,x_1 + (1 - \alpha)\,x_2 = \alpha\,(u_1 + x_0) + (1 - \alpha)\,(u_2 + x_0) = \underbrace{\alpha\,u_1 + (1 - \alpha)\,u_2}_{\in U} + x_0$$

von der Form (15.5).

„⇒“: Es sei A ein affiner Unterraum und $x_0 \in A$ beliebig, aber fest. Definiere $U := \{x - x_0 \mid x \in A\}$. Nach Konstruktion gilt dann $A = U + x_0$. Wir müssen zeigen: U ist ein Unterraum von $\mathbb{R}^n$. Es seien

dazu $u_1 = x_1 - x_0$ und $u_2 = x_2 - x_0$ Elemente von U mit irgendwelchen $x_1, x_2 \in A$. Dann sind $2\,x_1 + (1-2)\,x_0 = 2\,x_1 - x_0 \in A$ und $2\,x_2 + (1-2)\,x_0 = 2\,x_2 - x_0 \in A$ und daher

$$\begin{aligned} u_1 + u_2 &= x_1 - x_0 + x_2 - x_0 \\ &= \overbrace{\frac{1}{2}\underbrace{(2\,x_1 - x_0)}_{\in A} + \left(1 - \frac{1}{2}\right)\underbrace{(2\,x_2 - x_0)}_{\in A}}^{\in A} - x_0 \in A - x_0. \end{aligned} \tag{15.6}$$

Das zeigt $u_1 + u_2 \in U$, also $U + U \subseteq U$. Weiterhin sei $\beta \in \mathbb{R}$, dann gilt

$$\begin{aligned} \beta\,u_1 &= \beta\,(x_1 - x_0) \\ &= \underbrace{\beta\,x_1 + (1-\beta)\,x_0}_{\in A} - x_0. \end{aligned}$$

Das zeigt $\beta\,u_1 \in U$, also $\beta\,U \subseteq U$. Damit ist U ein Unterraum.

Der Beweis der letzten Implikation zeigt auch, dass (15.5) für jedes beliebige $x_0 \in A$ mit einem Unterraum der Form $\{x - x_0 \mid x \in A\}$ gilt. Es bleibt also lediglich zu zeigen, dass die Mengen $\{x - x_0 \mid x \in A\}$ und $\{x - \tilde{x}_0 \mid x \in A\}$ für beliebige $x_0, \tilde{x}_0 \in A$ übereinstimmen. Es seien dafür x, x_0, $\tilde{x}_0 \in A$. Wir schreiben $x - x_0 = x - x_0 + \tilde{x}_0 - \tilde{x}_0$, und weil

$$x - x_0 + \tilde{x}_0 = \underbrace{x - x_0 + \tilde{x}_0 - x_0}_{\in \{x - x_0 \mid x \in A\}} + x_0 \in A,$$

ist der Unterraum U unabhängig vom gewählten Aufpunkt x_0. □

Beachte: Ein affiner Unterraum entsteht nach (15.5) also aus einer Translation seines Richtungsraumes.

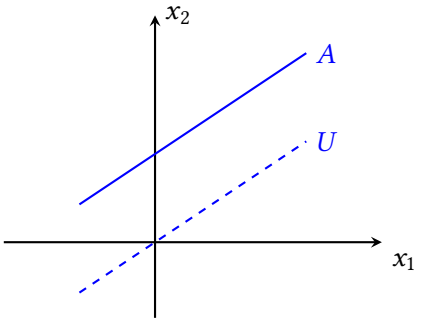


Abbildung 15.2.: Ein 1-dimensionaler affiner Unterraum A von $\mathbb{R}^2$. Der zugehörige Richtungsraum U ist gestrichelt gezeichnet.

Wir ordnen einem affinen Unterraum A die **Dimension** (englisch: *dimension*) seines Richtungsraumes zu:

$$\dim A := \dim U. \tag{15.7}$$

~~Aus technischen Gründen ist es günstig, den Fall zuzulassen, bei dem $A = \emptyset$ ist damit auch $U = \emptyset$. In diesem Fall setzen wir $\dim A := \dim U = -1$. Die einzige affine Basis ist dann die leere Menge.~~

Quizfrage 15.3: Welche Gestalt hat A im Fall $\dim A = 0$?

Definition 15.6 (Affine Unabhängigkeit, Affinkombination)**.**

(*i*) Eine (nichtleere) Menge von Vektoren $\{x^{(0)}, x^{(1)}, \ldots, x^{(k)}\}$ in $\mathbb{R}^n$, $k \in \mathbb{N}_0$, heißt **affin unabhängig** (englisch: *affine independent*), wenn die Menge der Vektoren $\{x^{(1)} - x^{(0)}, \ldots, x^{(k)} - x^{(0)}\}$ linear unabhängig ist.[2]

(*ii*) Eine maximale Menge affin unabhängiger Vektoren eines affinen Unterraumes A von $\mathbb{R}^n$ heißt eine **affine Basis** (englisch: *affine basis*) von A.

(*iii*) Ein Vektor $x \in \mathbb{R}^n$ heißt eine **Affinkombination** (englisch: *affine combination*) von $x^{(0)}, \ldots, x^{(k)} \in \mathbb{R}^n$, $k \in \mathbb{N}_0$, wenn gilt:

$$x = \sum_{i=0}^{k} \alpha^{(i)}\, x^{(i)}$$

mit Koeffizienten $\alpha^{(i)} \in \mathbb{R}$, die $\sum_{i=0}^{k} \alpha^{(i)} = 1$ erfüllen. Ist $M \subseteq \mathbb{R}^n$ irgendeine (nicht notwendig endliche) Menge, so heißt x eine Affinkombination von M, wenn x eine Affinkombination von endlich vielen Vektoren $x^{(0)}, \ldots, x^{(k)} \in M$ ist.[3] △

Quizfrage 15.4: Warum ist die Definition der affinen Unabhängigkeit invariant gegenüber einer Permutation der Vektoren $x^{(0)}, \ldots, x^{(k)}$?

Lemma 15.7 (Dimension eines affinen Unterraumes)**.**
Es sei $A \subseteq \mathbb{R}^n$ ein affiner Unterraum.

(*i*) A besitzt genau dann eine affine Basis $\{x^{(0)}, x^{(1)}, \ldots, x^{(k)}\}$ aus $k+1$ Elementen mit $k \in \mathbb{N}_0$, wenn $\dim A = k$ ist.

(*ii*) Ist $\{x^{(0)}, x^{(1)}, \ldots, x^{(k)}\}$ eine affine Basis von A, dann lässt sich jedes Element von A auf eindeutige Art und Weise aus $\{x^{(0)}, x^{(1)}, \ldots, x^{(k)}\}$ affinkombinieren. Genauer hat jedes $x \in A$ die Darstellung

$$x = \sum_{i=0}^{k} \alpha^{(i)}\, x^{(i)} \tag{15.8}$$

mit Koeffizienten $\alpha = (\alpha^{(0)}, \ldots, \alpha^{(k)})^\mathsf{T}$, die sich aus der eindeutigen Lösung des linearen Gleichungssystems

$$\underbrace{\begin{bmatrix} 1 & \cdots & 1 \\ | & & | \\ x^{(0)} & \cdots & x^{(k)} \\ | & & | \end{bmatrix}}_{=:B} \begin{pmatrix} \alpha^{(0)} \\ \vdots \\ \alpha^{(k)} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ | \\ x \\ | \end{pmatrix}}_{=:b} \tag{15.9}$$

ergeben. Die Matrix $B \in \mathbb{R}^{(n+1)\times(k+1)}$ hat Rang $k+1$. Daher ist $B^\mathsf{T}B$ regulär, und (15.9) kann äquivalent als

$$B^\mathsf{T}B\,\alpha = B^\mathsf{T}b \tag{15.10}$$

geschrieben werden.

[2] Ein einzelner Vektor $x^{(0)} \in \mathbb{R}^n$ ist also immer affin unabhängig, da die leere Menge von Vektoren linear unabhängig ist.

[3] Affinkombinationen sind also allgemeiner als Konvexkombinationen, weil die Vorzeichenbedingung $\alpha^{(i)} \geqslant 0$ für die Koeffizienten nicht gefordert wird.

Beweis. Der Beweis ist Inhalt von Hausaufgabe 10.3. □

Lemma 15.8 (Gleichheit von und Operationen auf affinen Unterräumen)**.**
Es seien $A = U + x_0$ 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*) Falls $A \cap B \neq \emptyset$ gilt, dann ist $A \cap B$ ist ein affiner Unterraum von $\mathbb{R}^n$.

(*iv*) Es sei $\{A_j\}_{j\in J}$ eine beliebige Familie affiner Unterräume von $\mathbb{R}^n$ mit nichtleerem Durchschnitt. Dann ist der Durchschnitt $\bigcap_{j\in J} A_j$ ein affiner Unterraum.

Beweis. Aussage (*i*): Wir zeigen zunächst (a) ⇒ (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) ⇒ (a). Es seien dazu $U = V$ und $0 \in A - B$. Aus letzerer Beziehung folgt, dass es $u \in U$ und $v \in V$ gibt, sodass $u + x_0 = v + y_0$ gilt. Es sei nun $x_1 = u_1 + x_0$ irgendein Punkt in A mit irgendeinem $u_1 \in U$. Dann ist $x_1 = u_1 - u + u + x_0 = u_1 - u + v + y_0$. Da $u_1 - u + v \in U = V$ gilt, liegt $x_1 \in B$, also $A \subseteq B$. Analog zeigt man $B \subseteq A$.

Die Äquivalenz (c) ⇔ (d) ist klar. Wir zeigen nun noch (b) ⇔ (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.

Aussage (*iv*): analog zu Aussage (*iii*) □

Definition 15.9 (Affine Hülle)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine nichtleere Menge. Der Durchschnitt aller affinen Unterräume von $\mathbb{R}^n$, die M enthalten, also

$$\operatorname{aff}(M) = \bigcap \big\{ A \subseteq \mathbb{R}^n \,\big|\, A \text{ ist affiner Unterraum von } \mathbb{R}^n \text{ und } M \subseteq A \big\}, \tag{15.11}$$

heißt die **affine Hülle** (englisch: *affine hull*) von M. △

Beachte: Es gilt $M \subseteq \operatorname{aff}(M)$, und $\operatorname{aff}(M)$ ist als nichtleerer Schnitt affiner Unterräume wiederum ein affiner Unterraum, daher der Name **affine Hülle**. $\operatorname{aff}(M)$ ist der kleinste affine Unterraum, der M enthält, genauer: $\operatorname{aff}(M)$ ist das eindeutige minimale Element der Teilmenge und $\{A \subseteq \mathbb{R}^n \mid A \text{ ist affiner Unterraum und } M \subseteq A\}$ im Sinne der Halbordnung der Mengeninklusion auf der Potenzmenge von $\mathbb{R}^n$.

Es gibt verschiedene Dimensionsbegriffe für allgemeine Teilmengen von $\mathbb{R}^n$. Einer davon verwendet die affine Hülle der Menge:

Definition 15.10 ((Affine) Dimension einer Menge)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine nichtleere Menge. Die **(affine) Dimension** (englisch: *(affine) dimension*) von M ist $\dim M := \dim \operatorname{aff}(M)$. △

Quizfrage 15.5: 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 nichtleere Menge.

(*i*) Die Menge $M \subseteq \mathbb{R}^n$ ist genau dann ein affiner Unterraum, wenn sie alle Affinkombinationen ihrer Elemente enthält.

(*ii*) $\operatorname{aff}(M)$ ist gleich der Menge aller Affinkombinationen von M.

(*iii*) Es gilt $M \subseteq \operatorname{conv}(M) \subseteq \operatorname{aff}(M)$.

Beachte: A ist affiner Unterraum $\Leftrightarrow A = \operatorname{aff}(A)$.

Beweis. Aussage (*i*): „⇒“: 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 $m \geqslant 2$: 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 $x = (1 - \alpha^{(m+1)})\, y + \alpha^{(m+1)} x^{(m+1)}$.

„⇐“: Es seien $x^{(1)}, x^{(2)} \in M$. Nach Voraussetzung enthält M alle Affinkombinationen $\alpha x^{(1)} + (1-\alpha) x^{(2)}$, d. h., M ist affiner Unterraum.

Aussage (*ii*): Es sei A die Menge aller Affinkombinationen von M. Natürlich gilt dann $M \subseteq A$. ~~Im Fall $M = \emptyset$ ist nichts zu zeigen, weil dann auch $A = \emptyset$ ist. Wir gehen also jetzt von $M \neq \emptyset$ aus.~~

„$\operatorname{aff}(M) \subseteq A$“: Wir zeigen: A ist ein affiner Unterraum. Damit kommt diese Menge im Durchschnitt (15.11) vor, also gilt $\operatorname{aff}(M) \subseteq A$.

Es seien $x, y \in A$, also gibt es Zahlen $m, \ell \in \mathbb{N}$ und $\beta^{(1)}, \dots, \beta^{(m)} \in \mathbb{R}$ sowie $\gamma_1, \dots, \gamma_\ell \in \mathbb{R}$ mit $\sum_{i=1}^m \beta^{(i)} = 1$ und $\sum_{j=1}^\ell \gamma^{(j)} = 1$, sodass $x = \sum_{i=1}^m \beta^{(i)} x^{(i)}$ und $y = \sum_{j=1}^\ell \gamma^{(j)} y^{(j)}$ gelten mit irgendwelchen $x^{(1)}, \dots, x^{(m)} \in M$ und $y^{(1)}, \dots, y^{(\ell)} \in M$. Es sei $\alpha \in \mathbb{R}$. Dann gilt

$$\alpha x + (1-\alpha) y = \alpha \sum_{i=1}^m \beta^{(i)} x^{(i)} + (1-\alpha) \sum_{j=1}^\ell \gamma^{(j)} y^{(j)},$$

d. h., $\alpha x + (1-\alpha) y$ ist Linearkombination der $\{x^{(i)}\}_{i=1}^m \cup \{y^{(j)}\}_{j=1}^\ell$. Die Koeffizienten ergeben in der Summe 1. Damit ist $\alpha x + (1-\alpha) y \in A$, also A ist affiner Unterraum.

„$\operatorname{aff}(M) \supseteq A$“: Es sei $x \in A$, also eine Affinkombination von M. Wegen $M \subseteq \operatorname{aff}(M)$ ist x auch eine Affinkombination von $\operatorname{aff}(M)$. $\operatorname{aff}(M)$ ist ein affiner Unterraum, stimmt also nach Aussage *(i)* mit der Menge seiner Affinkombinationen überein. Also ist $x \in \operatorname{aff}(M)$.

Aussage *(iii)*: Nach Lemma 13.7 gilt $M \subseteq \operatorname{conv}(M)$, und $\operatorname{conv}(M)$ sind gerade die Konvexkombinationen von M. Da jede Konvexkombination auch eine Affinkombination ist, gilt weiter $\operatorname{conv}(M) \subseteq \operatorname{aff}(M)$. □

Quizfrage 15.6: Welche Beziehung besteht zwischen $\operatorname{aff}(M_1 + M_2)$ und $\operatorname{aff}(M_1) + \operatorname{aff}(M_2)$ für beliebige Mengen $M_1, M_2 \subseteq \mathbb{R}^n$?

Satz 15.12 (Existenz einer affinen Basis)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge der Dimension $k \in \mathbb{N}_0$. Dann existieren $k+1$ affin unabhängige Punkte $x^{(0)}, \dots, x^{(k)} \in M$, die eine affine Basis von $\operatorname{aff}(M)$ bilden.

Beweis. Die affine Hülle $\operatorname{aff}(M)$ besteht aus den Affinkombinationen von M (Lemma 15.11). Da die Dimension einer Menge von Affinkombinationen gleich der maximalen Anzahl affin unabhängiger Punkte ist, folgt die Behauptung. □

Expertenwissen: Hüllenoperatoren

Wir haben bisher die folgenden **Hüllenoperationen** für Teilmengen von $\mathbb{R}^n$ verwendet:

abgeschlossene Hülle (**Abschluss**)	$\overline{M} = \{x \in \mathbb{R}^n \mid x \text{ ist Häufungspunkt von } M\}$
lineare Hülle (**Spann**)	$\operatorname{span}\{x^{(1)}, \dots, x^{(k)}\}$ $= \left\{x \in \mathbb{R}^n \,\middle|\, x \text{ ist Linearkombination von } x^{(1)}, \dots, x^{(k)}\right\}$
konvexe Kegelhülle (**konische Hülle**, **positive Hülle**)	$\operatorname{pos}\{b^{(1)}, \dots, b^{(k)}\}$ $= \{x \in \mathbb{R}^n \mid x \text{ ist nichtneg. Linearkombination von } b^{(1)}, \dots, b^{(k)}\}$ (Lemma 6.13)
konvexe Hülle	$\operatorname{conv}(M) = \bigcap \left\{C \subseteq \mathbb{R}^n \,\middle|\, C \text{ ist konvex und } M \subseteq C\right\}$ (Definition 13.6)
affine Hülle	$\operatorname{aff}(M)$ $= \bigcap \left\{A \subseteq \mathbb{R}^n \,\middle|\, A \text{ ist affiner Unterraum von } \mathbb{R}^n \text{ und } M \subseteq A\right\}$ (Definition 15.9)

Die lineare Hülle sowie die konvexe Kegelhülle können wir natürlich auch auf beliebige Mengen erweitern, wir haben dies nur bisher nicht benötigt.

Was ist den obigen Operationen gemeinsam? Die jeweilige Operation — nennen wir sie $H\colon \mathcal{P}(\mathbb{R}^n) \to \mathcal{P}(\mathbb{R}^n)$ — ist auf beliebigen Teilmengen von $\mathbb{R}^n$ definiert, und es gilt jeweils

$A \subseteq H(A)$	H ist **extensiv** (englisch: *extensive*),
$A \subseteq B \Rightarrow H(A) \subseteq H(B)$	H ist **monoton** bzgl. der Mengeninkl. (englisch: *monotone*),
$H(H(A)) = H(A)$	H ist **idempotent** (englisch: *idempotent*).

Eine solche Funktion nennt man einen **Hüllenoperator** (englisch: *hull operator*), siehe dazu etwa Westermann, 1976.

Man kann (leicht) zeigen (Cohn, 1981, p.42–43), dass Hüllenoperatoren in bijektiver Beziehung stehen mit **Abschlusssystemen**. Eine Menge $\mathcal{X} \subseteq \mathcal{P}(\mathbb{R}^n)$ heißt ein **Abschlusssystem** (englisch: *closure system*), wenn gilt:

$$\mathcal{Y} \subseteq \mathcal{X} \quad \Rightarrow \quad \bigcap \mathcal{Y} \in \mathcal{X}.$$

Der Zusammenhang wie folgt: Für jedes Abschlusssystem $\mathcal{X}$ definiert

$$H(A) := \bigcap \{B \in \mathcal{X} \mid A \subseteq B\}$$

einen Hüllenoperator.

Wir können die o. g. Hüllenoperationen daher einheitlich über solche Schnitte definieren, die die jeweils charakteristischen Eigenschaften erhalten:

$$\begin{aligned}
\overline{A} &= \bigcap \{B \subseteq \mathbb{R}^n \mid B \text{ ist abgeschlossen und } A \subseteq B\} \\
\operatorname{span} A &= \bigcap \{B \subseteq \mathbb{R}^n \mid B \text{ ist Unterraum und } A \subseteq B\} \\
\operatorname{pos} A &= \bigcap \{B \subseteq \mathbb{R}^n \mid B \text{ ist konvexer Kegel und } A \subseteq B\} \\
\operatorname{conv} A &= \bigcap \{B \subseteq \mathbb{R}^n \mid B \text{ ist konvex und } A \subseteq B\} \\
\operatorname{aff} A &= \bigcap \{B \subseteq \mathbb{R}^n \mid B \text{ ist affiner Unterraum und } A \subseteq B\}.
\end{aligned}$$

Ein weiteres Beispiel ist noch die konvexe abgeschlossene Hülle

$$\overline{\operatorname{conv}} A = \bigcap \{B \subseteq \mathbb{R}^n \mid B \text{ ist abgeschlossen und konvex und } A \subseteq B\},$$

wobei es hierbei auch reicht (Hausaufgabe 12.1), B nur über die abgeschlossenen Halbräume laufen zu lassen. Noch ein weiteres Beispiel für eine Hüllenoperation ist die von einer Teilmenge $\mathcal{A}$ der Potenzmenge $\mathcal{P}(A)$ von $A \subseteq \mathbb{R}^n$ erzeugte Σ-Algebra, also

$$\sigma(\mathcal{A}) = \bigcap \{\Sigma \subseteq \mathcal{P}(A) \mid \Sigma \text{ ist } \sigma\text{-Algebra und } \mathcal{A} \subseteq \Sigma\}.$$

wobei $\mathcal{A}$ hier eine Teilmenge der Potenzmenge $\mathcal{P}(A)$ einer Menge $A \subseteq \mathbb{R}^n$ ist.

Ende der Vorlesung 20

Ende der Woche 10

§ 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** (1911))**.**
Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge der Dimension k und $x \in \operatorname{conv}(M)$ eine Konvexkombination von Punkten $x^{(0)}, \dots, x^{(m)} \in M$ mit $m \geqslant 0$. Dann ist x bereits eine Konvexkombination von höchstens $k+1$ dieser Punkte.

Beachte: Dass jedes $x \in \operatorname{conv}(M)$ eine Konvexkombination von Punkten $x^{(0)}, \dots, x^{(m)} \in M$ mit $m \geqslant 0$ ist, ist durch Lemma 13.7 gesichert. Der Satz von Carathéodory 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.

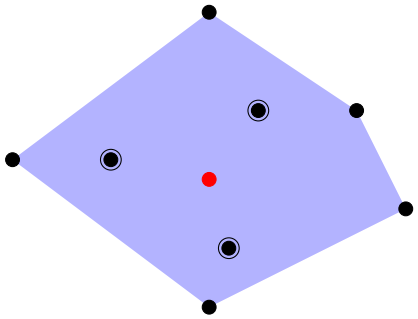


Abbildung 15.3.: Illustration des Satzes von Carathéodory 15.13 für eine Menge der Dimension $k = 2$. Die schwarzen Punkte bilden die Menge M. Der rote Punkt $x \in \operatorname{conv}(M)$ ist bereits eine Konvexkombination der drei hervorgehobenen Punkte.

Beweis. Der Beweis ist Inhalt von Hausaufgabe 11.1. □

Lemma 15.14 (Innere Punkte und Dimension einer Menge)**.**
Es sei $M \subseteq \mathbb{R}^n$. Ist $\operatorname{int} M \neq \emptyset$, dann gilt $\dim M = n$.

Beweis. Es sei $x \in \operatorname{int} M$ und $B_\varepsilon(x) \subseteq M$. Dann gilt $\operatorname{aff} B_\varepsilon(x) \subseteq \operatorname{aff} M$ und damit $n = \dim B_\varepsilon(x) \leqslant \dim M$. Andererseits gilt für jede Menge $M \subseteq \mathbb{R}^n$ natürlich $\dim(M) \leqslant n$. □

Lemma 15.14 zeigt, dass Mengen $M \subseteq \mathbb{R}^n$, die nicht die volle Dimension n besitzen, keine inneren Punkte in $\mathbb{R}^n$ haben. Wir würden aber, insbesondere für konvexe Mengen $C \subseteq \mathbb{R}^n$, gerne mit inneren Punkten arbeiten. Wir gehen dazu zur Relativtopologie in $\operatorname{aff}(C)$ über. Dies führt zu folgenden Begriffen:

Definition 15.15 (Relatives Inneres).
Es sei $C \subseteq \mathbb{R}^n$ eine konvexe Menge.

(*i*) Ein Punkt $x \in C$ heißt ein **relativ innerer Punkt** (englisch: *relative interior point*) von C, wenn es ein $\varepsilon > 0$ gibt, sodass $B_\varepsilon(x) \cap \operatorname{aff}(C) \subseteq C$ liegt. Die Menge aller relativ inneren Punkte von C heißt das **relative Innere** (englisch: *relative interior*) und wird mit $\operatorname{relint}(C)$ bezeichnet.

(*ii*) Ein Punkt $x \in \mathbb{R}^n$ heißt ein **relativer Randpunkt** (englisch: *relative boundary point*) von C, wenn $x \in \overline{C} \setminus \operatorname{relint}(C)$ liegt. Die Menge aller relativen Randpunkte von C heißt der **relative Rand** (englisch: *relative boundary*) (englisch: *relative boundary*) und wird mit $\operatorname{rel}\partial(C)$ bezeichnet. △

Siehe Abbildung 15.4 für eine Illustration.

Quizfrage 15.7: Was ist $\operatorname{relint}(C)$, wenn die konvexe Menge $C \subseteq \mathbb{R}^n$ die volle Dimension n hat?

Quizfrage 15.8: Was ist $\operatorname{relint}(A)$ für einen affinen Unterraum $A \subseteq \mathbb{R}^n$?

Quizfrage 15.9: Wenn $C_1 \subseteq C_2$ ist, gilt dann auch immer $\operatorname{relint}(C_1) \subseteq \operatorname{relint}(C_2)$?

Quizfrage 15.10: Warum benötigen wir keinen Begriff des relativen Abschlusses?

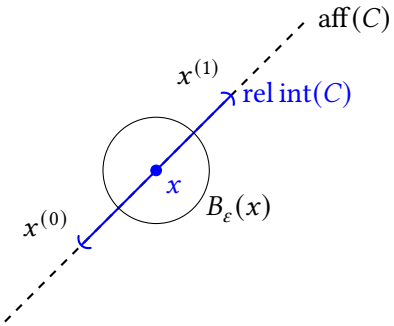


Abbildung 15.4.: Illustration einer 1-dimensionalen konvexen Menge $C = \operatorname{conv}(\{x^{(0)}, x^{(1)}\}) = \{\alpha\, x^{(0)} + (1-\alpha)\, x^{(1)} \mid \alpha \in [0,1]\}$ und ihres relativen Inneren $\operatorname{relint}(C) = \{\alpha\, x^{(0)} + (1-\alpha)\, x^{(1)} \mid \alpha \in (0,1)\}$ in $\mathbb{R}^2$. Insbesondere der Punkt x ist ein relativ innerer Punkt.

Satz 15.16 (vgl. Jarre, Stoer, 2004, Satz 7.2.5).
Jede nichtleere konvexe Menge $C \subseteq \mathbb{R}^n$ besitzt ein nichtleeres relatives Inneres. Es gilt: $\dim \operatorname{relint}(C) = \dim C$.

Beweis. Es sei $k := \dim C$. Da C nichtleer ist, gilt, $k \geqslant 0$. Es gibt also nach Satz 15.12 affin unabhängige Punkte $x^{(0)}, \ldots, x^{(k)} \in C$. Nach Lemma 15.7 lässt sich jeder Punkt $x \in \operatorname{aff}(C)$ eindeutig als Affinkombination $x = \sum_{i=0}^{k} \alpha^{(i)}\, x^{(i)}$ schreiben, wobei sich die Koeffizienten aus der eindeutigen Lösung des linearen Gleichungssystems (15.10) ergeben:

$$B^\mathsf{T} B\, \alpha = B^\mathsf{T} b.$$

Wir zeigen jetzt, dass der Mittelpunkt der Punkte $\{x^{(0)}, \ldots, x^{(k)}\}$

$$\overline{x} := \frac{1}{k+1} \sum_{i=0}^{k} x^{(i)}$$

ein relativ innerer Punkt von C ist. Dazu konstruieren wir eine abgeschlossene Kugel

$$\overline{B_\varepsilon^{\|\cdot\|_\infty}(\overline{x})} = \left\{ x \in \mathbb{R}^n \,\middle|\, \|x - \overline{x}\|_\infty \leqslant \varepsilon \right\}$$

mit der Eigenschaft $\overline{B_\varepsilon^{\|\cdot\|_\infty}(\overline{x})} \cap \operatorname{aff}(C) \subseteq \operatorname{conv}(\{x^{(0)}, \dots, x^{(k)}\}) \subseteq \operatorname{conv}(C) = C$. (**Quizfrage 15.11:** Warum gilt die letzte Inklusion?)

Wir setzen dazu $\varepsilon := 1/\big((k+1)\|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty\big)$. Dabei ist $\|\cdot\|_\infty$ die durch die ∞-Norm im Definitions- und Bildbereich induzierte Matrixnorm, also[4]

$$\|A\|_\infty = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{\|x\|_\infty = 1} \|Ax\|_\infty. \tag{15.12}$$

Es sei nun $x \in \overline{B_\varepsilon^{\|\cdot\|_\infty}(\overline{x})} \cap \operatorname{aff}(C)$. Wir bezeichnen mit α die Koeffizienten in der Affinkombination $x = \sum_{i=0}^k \alpha^{(i)} x^{(i)}$. Weiterhin sind $\overline{\alpha}$ die Koeffizienten von $\overline{x}$, also $\overline{\alpha} = \frac{1}{k+1}\mathbf{1}$. Um zu zeigen, dass tatsächlich $x \in \operatorname{conv}(\{x^{(0)}, \dots, x^{(k)}\})$ liegt, müssen wir für die Koeffizienten zeigen: $\alpha \geqslant 0$. Es gilt

$$\begin{aligned}
\|\alpha - \overline{\alpha}\|_\infty &= \left\| (B^\mathsf{T}B)^{-1}B^\mathsf{T} \left[\begin{pmatrix} \frac{1}{x} \end{pmatrix} - \begin{pmatrix} \frac{1}{\overline{x}} \end{pmatrix} \right] \right\|_\infty \leqslant \|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty \left\| \begin{pmatrix} \frac{1}{x} \end{pmatrix} - \begin{pmatrix} \frac{1}{\overline{x}} \end{pmatrix} \right\|_\infty \\
&= \|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty \, \|x - \overline{x}\|_\infty \leqslant \|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty \, \varepsilon = \frac{1}{k+1}.
\end{aligned}$$

Daraus folgt wie gewünscht $\alpha \geqslant 0$ mit Hilfe der Dreiecksungleichung.

Wir zeigen jetzt noch, dass die Dimension von $\operatorname{relint}(C)$ ebenfalls gleich k ist. Dazu geben wir $k+1$ affin unabhängige Punkte in $\operatorname{relint}(C)$ an. Damit ist $\dim \operatorname{relint}(C) \geqslant k$, und da außerdem $\operatorname{relint}(C) \subseteq C$ die Beziehung $\dim \operatorname{relint}(C) \leqslant \dim C = k$ impliziert, ist dann die Behauptung gezeigt. Zur Konstruktion der Punkte machen wir den Ansatz

$$\overline{x}_i := \beta\,\overline{x} + (1-\beta)\,x^{(i)}$$

und wählen $\beta \in (0,1)$ so klein, dass $\|\overline{x}_i - \overline{x}\|_\infty \leqslant \varepsilon$ bleibt für alle $i = 0, \dots, k$. Damit liegen alle $\overline{x}_i \in \overline{B_\varepsilon^{\|\cdot\|_\infty}(\overline{x})}$ und außerdem in $\operatorname{aff}(C)$ (**Quizfrage 15.12:** Begründung?).

Um die affine Unabhängigkeit der Punkte $\{\overline{x}_0, \dots, \overline{x}_k\}$ zu zeigen, machen wir den Ansatz

$$\begin{aligned}
0 = \sum_{i=1}^k \gamma^{(i)} \big(\overline{x}_i - \overline{x}_0\big) &= \sum_{i=1}^k \gamma^{(i)} \big(\beta\,\overline{x} + (1-\beta)\,x^{(i)} - \beta\,\overline{x} - (1-\beta)\,x^{(0)}\big) \\
&= (1-\beta) \sum_{i=1}^k \gamma^{(i)} (x^{(i)} - x^{(0)}).
\end{aligned}$$

Da die Punkte $\{x^{(0)}, \dots, x^{(k)}\}$ aber affin unabhängig sind, folgt $\gamma^{(i)} = 0$ für alle $i = 1, \dots, k$. Also sind auch die Punkte $\{\overline{x}_0, \dots, \overline{x}_k\}$ affin unabhängig. □

[4] Diese ist auch als **Zeilensummennorm** (englisch: *row-sum norm*) bekannt, da für $A \in \mathbb{R}^{m\times n}$ die folgende Beziehung gilt:

$$\|A\|_\infty = \max\left\{ \sum_{j=1}^n |a_{ij}| \,\middle|\, i = 1, \dots, m \right\}.$$

$\|A\|_\infty$ ist also das Maximum der betragsweisen Zeilensummen.

Folgerung 15.17 (Innere Punkte und Dimension einer konvexen Menge, vgl. Lemma 15.14)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex. Dann gilt $\operatorname{int} C \neq \emptyset \Leftrightarrow \dim C = n$.

Beweis. Die Behauptung folgt sofort aus Satz 15.16. □

Lemma 15.18 (**Accessibility lemma**, vgl. Jarre, Stoer, 2004, Lemma 7.2.6)**.**
Es seien $C \subseteq \mathbb{R}^n$ konvex, $x \in \overline{C}$ und $z \in \operatorname{relint}(C)$. Dann gilt $\alpha\, x + (1-\alpha)\, z \in \operatorname{relint}(C)$ für alle $\alpha \in [0,1)$, d. h., die gesamte Verbindungsstrecke (evtl. mit Ausnahme von x selbst) gehört zum relativen Inneren von C.

Beweis. Aufgrund der Voraussetzung $x \in \overline{C}$ gilt $x \in C + B_\varepsilon(0)$ für alle $\varepsilon > 0$. Wegen $z \in \operatorname{relint}(C)$ gibt es ein $r > 0$, sodass $B_r(z) \cap \operatorname{aff}(C) \subseteq C$ liegt.

Es sei nun $\alpha \in [0,1)$ beliebig und $y := \alpha\, x + (1-\alpha)\, z$. Wir zeigen: Mit $r(\alpha) := \frac{1-\alpha}{1+\alpha}\, r$ gilt $B_{r(\alpha)}(x) \cap \operatorname{aff}(C) \subseteq C$, d. h., $y \in \operatorname{relint}(C)$. Dazu halten wir zunächst fest:

$$\begin{aligned}
B_{r(\alpha)}(x) &= B_{r(\alpha)}\big(\alpha\, x + (1-\alpha)\, z\big) \\
&= \alpha\, x + (1-\alpha)\, z + B_{r(\alpha)}(0) \\
&\subseteq \alpha\, \big(C + B_{r(\alpha)}(0)\big) + (1-\alpha)\, z + B_{r(\alpha)}(0) \\
&= \alpha\, C + (1-\alpha)\, z + (1+\alpha)\, B_{r(\alpha)}(0) \\
&= \alpha\, C + (1-\alpha)\big[z + B_r(0)\big] \\
&= \alpha\, C + (1-\alpha)\, B_r(z),
\end{aligned}$$

siehe auch Abbildung 15.5. Durch den Schnitt mit $\operatorname{aff}(C)$ folgt

$$B_{r(\alpha)}(x) \cap \operatorname{aff}(C) \subseteq \big[\alpha\, C + (1-\alpha)\, B_r(z)\big] \cap \operatorname{aff}(C).$$

Wegen $\operatorname{aff}(C) = \alpha \operatorname{aff}(C) + (1-\alpha) \operatorname{aff}(C)$ (**Quizfrage 15.13:** Begründung?) gilt weiter

$$B_{r(\alpha)}(x) \cap \operatorname{aff}(C) \subseteq \alpha \underbrace{\big[C \cap \operatorname{aff}(C)\big]}_{\subseteq C} + (1-\alpha) \underbrace{\big[B_r(z) \cap \operatorname{aff}(C)\big]}_{\subseteq C} \subseteq C.$$

Die letzte Inklusion folgt aus der Konvexität von C. □

Wir geben noch eine nützliche Charakterisierung des relativen Inneren einer konvexen Menge an, die es uns erlaubt, mit einzelnen Richtungen zu argumentieren statt gleichzeitig mit allen Richtungen in einer Kugel:

Lemma 15.19 (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*) $x \in \operatorname{relint}(C)$.

(*ii*) Zu jedem $y \in \operatorname{aff}(C)$ existiert ein $\varepsilon > 0$, sodass $x + \varepsilon\,(y - x) \in C$ und $x - \varepsilon\,(y - x) \in C$ liegen.

(*iii*) Zu jedem $y \in C$ existiert ein $\varepsilon > 0$, sodass $x + \varepsilon\,(y - x) \in C$ und $x - \varepsilon\,(y - x) \in C$ liegen.

Quizfrage 15.14: Was bedeutet die Bedingung aus Aussage (*iii*) anschaulich?

Beweis. Aussage (i) $\Rightarrow$ Aussage (ii): Es sei $x \in \operatorname{relint}(C)$ und $y \in \operatorname{aff}(C)$. Wir können $y \neq x$ annehmen, sonst ist die Aussage klar. Wegen $x \in \operatorname{relint}(C)$ gibt es ein $\tilde{\varepsilon} > 0$, sodass $B_{\tilde{\varepsilon}}(x) \cap \operatorname{aff}(C) \subseteq C$. Für $\varepsilon := \tilde{\varepsilon}/\|y - x\|$ sind dann

$$\underbrace{x + \varepsilon\,(y - x)}_{=\varepsilon y + (1-\varepsilon)x \in \operatorname{aff}(C)} \in B_{\tilde{\varepsilon}}(x) \cap \operatorname{aff}(C) \subseteq C \quad \text{und} \quad \underbrace{x - \varepsilon\,(y - x)}_{=-\varepsilon y + (1+\varepsilon)x \in \operatorname{aff}(C)} \in B_{\tilde{\varepsilon}}(x) \cap \operatorname{aff}(C) \subseteq C.$$

Aussage (ii) $\Rightarrow$ Aussage (iii): Klar, da $C \subseteq \operatorname{aff}(C)$ ist.

Aussage (iii) $\Rightarrow$ Aussage (i): Nach Satz 15.16 können wir ein $y \in \operatorname{relint}(C)$ wählen, insbesondere gilt $y \in C$. Nach Voraussetzung existiert $\varepsilon > 0$, sodass $z := x - \varepsilon\,(y - x) \in C$ liegt. Das heißt aber auch

$$x = \frac{1}{1+\varepsilon}\, z + \frac{\varepsilon}{1+\varepsilon}\, y,$$

d. h., x ist eine echte Konvexkombination von $z \in C$ und $y \in \operatorname{relint}(C)$. Nach dem Accessibility lemma 15.18 gehört also x zu $\operatorname{relint}(C)$. □

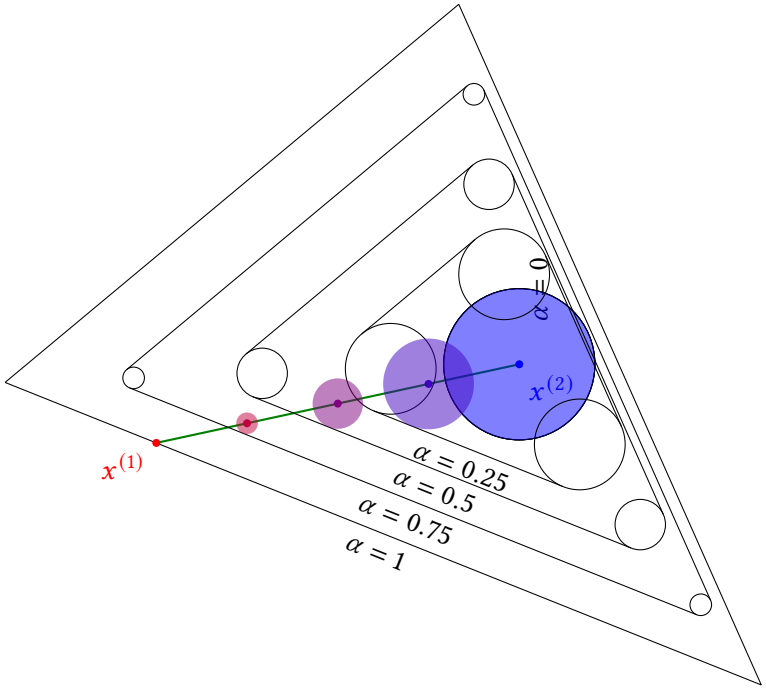


Abbildung 15.5.: Illustration der Inklusion $B_{r(\alpha)}(x) \subseteq \alpha C + (1-\alpha)\, B_r(x^{(2)})$ aus dem Beweis des Accessibility lemmas 15.18 für $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. Die farbig hervorgehobenen Kugeln sind die Mengen $B_{r(\alpha)}(x)$. Die Mengen $\alpha C + (1-\alpha)\, B_r(x^{(2)})$ 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 15.20 (Rockafellar, 1970, Theorem 6.2)**.**

(i) Ist $C \subseteq \mathbb{R}^n$ konvex, dann sind das relative Innere $\operatorname{relint}(C)$ und der Abschluss $\overline{C}$ konvex.

(ii) Die Mengen C, $\operatorname{relint}(C)$ und $\overline{C}$ haben alle dieselbe affine Hülle, also auch dieselbe Dimension.

Beweis. Aussage (*i*): Es sei $C \subseteq \mathbb{R}^n$ konvex. Wir zeigen zuerst, dass $\operatorname{relint}(C)$ konvex ist. Es seien dazu $x_1, x_2 \in \operatorname{relint}(C)$ und $\alpha \in [0,1]$. Dann gehört nach Lemma 15.18 auch die Konvexkombination $\alpha x_1 + (1-\alpha) x_2$ zu $\operatorname{relint}(C)$, also ist $\operatorname{relint}(C)$ konvex.

Der Abschluss von C erfüllt die Beziehung

$$\overline{C} = \bigcap \left\{ C + \overline{B_\varepsilon(0)} \,\middle|\, \varepsilon > 0 \right\}.$$

Die Mengen, über die der Durchschnitt gebildet wird, sind nach Beispiel 13.2 und Satz 13.3 konvex, also ist auch $\overline{C}$ konvex.

Aussage (*ii*): Wir zeigen jetzt noch $\operatorname{aff} \operatorname{relint}(C) = \operatorname{aff} C = \operatorname{aff} \overline{C}$. Aufgrund von $\operatorname{relint}(C) \subseteq C \subseteq \overline{C}$ folgt

$$\operatorname{aff} \operatorname{relint}(C) \subseteq \operatorname{aff} C \subseteq \operatorname{aff} \overline{C}.$$

Wegen $\overline{C} \subseteq \operatorname{aff}(C)$ gilt auch $\operatorname{aff} \overline{C} \subseteq \operatorname{aff} \operatorname{aff}(C) = \operatorname{aff}(C)$. Also bleibt noch $\operatorname{aff} \operatorname{relint}(C) = \operatorname{aff} C$ zu zeigen. Wir wissen aber bereits aus Satz 15.16, dass $\dim \operatorname{relint}(C) = \dim C$ gilt, also muss $\operatorname{aff} \operatorname{relint}(C) = \operatorname{aff} C$ sein. □

Satz 15.21 (Rockafellar, 1970, Theorem 6.3)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex. Dann gelten:

(*i*) $\overline{\operatorname{relint}(C)} = \overline{C}$, d. h., $\operatorname{relint}(C)$ and C haben denselben Abschluss.

(*ii*) $\operatorname{relint}(C) = \operatorname{relint}(\overline{C})$, d. h., C und $\overline{C}$ haben dasselbe relative Innere.

(*iii*) $\operatorname{rel} \partial(C) = \operatorname{rel} \partial(\overline{C})$, d. h., C und $\overline{C}$ haben denselben relativen Rand.

Beweis. Wir können $C \neq \emptyset$ annehmen, ansonsten sind alle Mengen leer.

Aussage (*i*): Die Inklusion $\overline{\operatorname{relint}(C)} \subseteq \overline{C}$ folgt unmittelbar aus $\operatorname{relint}(C) \subseteq C$. Für die umgekehrte Aussage sei nun $x_1 \in \overline{C}$ und $x_2 \in \operatorname{relint}(C)$. Dann ist nach Lemma 15.18 $\alpha x_1 + (1-\alpha) x_2 \in \operatorname{relint}(C)$. Der Grenzübergang $\alpha \nearrow 1$ zeigt $x_1 \in \overline{\operatorname{relint}(C)}$.

Aussage (*ii*): Aus $C \subseteq \overline{C}$ und der Tatsache, dass beide Mengen dieselbe affine Hülle haben (Satz 15.20) folgt $\operatorname{relint}(C) \subseteq \operatorname{relint}(\overline{C})$. Für die umgekehrte Aussage sei nun $x \in \operatorname{relint}(\overline{C})$, also existiert $\varepsilon > 0$ mit $B_\varepsilon(x) \cap \operatorname{aff}(C) \subseteq \overline{C}$. Es sei außerdem $y \in \operatorname{relint}(C)$ (Satz 15.16). Ist $y = x$, so sind wir fertig. Es sei also jetzt $y \neq x$. Ziel ist die Konstruktion eines Punktes $z \in \overline{C}$, sodass x als echte Konvexkombination von $y \in \operatorname{relint}(C)$ und $z \in \overline{C}$ geschrieben werden kann. Denn dann folgt aus Lemma 15.18, dass $x \in \operatorname{relint}(C)$ liegt.

Wir definieren

$$z := x + \delta (x - y) \quad \text{mit } \delta := \frac{\varepsilon}{\|x - y\|}.$$

Dann ist $\|z - x\| = \delta \|x - y\| = \varepsilon$, also gilt $z \in \overline{B_\varepsilon(x)}$. Wegen $\overline{B_\varepsilon(x)} \cap \operatorname{aff}(C) \subseteq \overline{C}$ (**Quizfrage 15.15:** Begründung?) folgt auch $z \in \overline{C}$. Wir können nun x schreiben als echte Konvexkombination

$$x = \frac{1}{1+\delta} z + \left(1 - \frac{1}{1+\delta}\right) y$$

mit $z \in \overline{C}$ und $y \in \operatorname{relint}(C)$. Nach Lemma 15.18 gilt $x \in \operatorname{relint}(C)$, was zu zeigen war.

Aussage (iii): Nach Definition des relativen Randes gilt

$$\operatorname{rel}\partial(C) = \overline{C} \setminus \operatorname{relint}(C)$$

und weiter

$$\operatorname{rel}\partial(\overline{C}) = \overline{C} \setminus \operatorname{relint}(\overline{C}) = \overline{C} \setminus \operatorname{relint}(C),$$

wobei die letzte Gleichheit aus Aussage (i) folgt. □

Folgerung 15.22 (Gleichheit der Abschlüsse und der relativen Inneren konvexer Mengen)**.**
Es seien $C_1, C_2 \subseteq \mathbb{R}^n$ konvex. Die folgenden Aussagen sind äquivalent:

(i) $\overline{C_1} = \overline{C_2}$.

(ii) $\operatorname{relint}(C_1) = \operatorname{relint}(C_2)$.

Beweis. Aussage (i) ⇒ Aussage (ii): Aus $\overline{C_1} = \overline{C_2}$ folgt $\operatorname{relint}(\overline{C_1}) = \operatorname{relint}(\overline{C_2})$. Satz 15.21 (ii) impliziert $\operatorname{relint}(C_1) = \operatorname{relint}(C_2)$.

Aussage (ii) ⇒ Aussage (i): Aus $\operatorname{relint}(C_1) = \operatorname{relint}(C_2)$ folgt $\overline{\operatorname{relint}(C_1)} = \overline{\operatorname{relint}(C_2)}$. Satz 15.21 (i) impliziert $\overline{C_1} = \overline{C_2}$. □

Lemma 15.23 (Relatives Inneres und Abschluss des Schnitts konvexer Mengen, vgl. Hiriart-Urruty, Lemaréchal, 2001, Proposition A.2.1.10)**.**
Es seien $C_1, C_2 \subseteq \mathbb{R}^n$ konvex und $\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2) \neq \emptyset$. Dann gilt:

(i) $\overline{C_1 \cap C_2} = \overline{C_1} \cap \overline{C_2}$.

(ii) $\operatorname{relint}(C_1 \cap C_2) = \operatorname{relint}(C_1) \cap \operatorname{relint}(C_2)$.

Beweis. Aussage (i): Es gilt

$$\begin{aligned}
\overline{C_1 \cap C_2} &\subseteq \overline{\overline{C_1} \cap \overline{C_2}} && \text{wegen } C_i \subseteq \overline{C_i} \text{ und der Monotonie des Abschlusses} \\
&= \overline{C_1} \cap \overline{C_2} && \text{denn der Schnitt zweier abgeschlossener Mengen ist abgeschlossen.}
\end{aligned}$$

Für die umgekehrte Inklusion sei $x \in \overline{C_1} \cap \overline{C_2}$. Wir wählen ein $z \in \operatorname{relint}(C_1) \cap \operatorname{relint}(C_2) \neq \emptyset$. Nach dem Accessibility lemma 15.18 gilt

$$\left\{\alpha x + (1-\alpha) z \,\middle|\, \alpha \in [0,1)\right\} \subseteq \operatorname{relint}(C_1) \cap \operatorname{relint}(C_2).$$

Nehmen wir auf beiden Seiten den Abschluss, so erhalten wir

$$\left\{\alpha x + (1-\alpha) z \,\middle|\, \alpha \in [0,1]\right\} \subseteq \overline{\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2)},$$

also insbesondere $x \in \overline{\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2)}$ und wegen der Monotonie des Abschlusses auch $x \in \overline{C_1 \cap C_2}$. Da x beliebig war, folgt

$$\overline{C_1} \cap \overline{C_2} \subseteq \overline{\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2)} \subseteq \overline{C_1 \cap C_2} \subseteq \overline{C_1} \cap \overline{C_2}, \tag{15.13}$$

also gilt überall die Gleichheit.

Aussage (*ii*): Wie in (15.13) gesehen, besitzen die konvexen Mengen $\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2)$ und $C_1 \cap C_2$ denselben Abschluss. Nach Folgerung 15.22 haben sie also auch dasselbe relative Innere, also

$$\operatorname{relint}(C_1 \cap C_2) = \operatorname{relint}(\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2)) \subseteq \operatorname{relint}(C_1) \cap \operatorname{relint}(C_2). \tag{15.14}$$

Die Inklusion folgt dabei aus der Eigenschaft $\operatorname{relint}(C) \subseteq C$ für alle Mengen $C \subseteq \mathbb{R}^n$.

Für die zu (15.14) umgekehrte Inklusion sei $x \in \operatorname{relint}(C_1) \cap \operatorname{relint}(C_2)$. Aufgrund von Lemma 15.19 können wir zu jedem $y^{(1)} \in C_1$ ein $\varepsilon^{(1)} > 0$ so wählen, dass $x - \varepsilon^{(1)}\,(y^{(1)} - x) \in C_1$ liegt. Analog können wir zu jedem $y^{(2)} \in C_2$ ein $\varepsilon^{(2)} > 0$ so wählen, dass $x - \varepsilon^{(2)}\,(y^{(2)} - x) \in C_2$ liegt. Insbesondere gilt das, wenn $y^{(1)} = y^{(2)} \in \operatorname{relint}(C_1 \cap C_2) \subseteq C_1 \cap C_2$ liegt.

Ein solches $y \in \operatorname{relint}(C_1 \cap C_2)$ wählen wir jetzt, und zwar so, dass $y \neq x$ ist. (Wenn das nicht möglich ist, dann ist $x \in \operatorname{relint}(C_1 \cap C_2)$ gezeigt.) Es gibt also ein $\varepsilon > 0$, sodass

$$z \coloneqq x - \varepsilon\,(y - x) = (1 + \varepsilon)\,x - \varepsilon\,y \in C_1 \cap C_2.$$

Das heißt andererseits, dass x eine echte Konvexkombination von y und z ist:

$$x = \frac{\varepsilon}{1+\varepsilon} \underbrace{y}_{\in \operatorname{relint}(C_1 \cap C_2)} + \frac{1}{1+\varepsilon} \underbrace{z}_{\in C_1 \cap C_2}.$$

Nach dem Accessibility lemma 15.18 gilt also $x \in \operatorname{relint}(C_1 \cap C_2)$, d. h., $\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2) \subseteq \operatorname{relint}(C_1 \cap C_2)$. □

Definition 15.24 (Algebraisches Inneres)**.**
Es sei $M \subseteq \mathbb{R}^n$. Ein Punkt $x_0 \in M$ heißt ein **algebraisch innerer Punkt** (englisch: *algebraically interior point*) von M, wenn gilt: Für jedes $d \in \mathbb{R}^n$ existiert ein $\varepsilon_d > 0$, sodass

$$\{x_0 + t\,d \mid 0 \leqslant t < \varepsilon_d\} \subseteq M$$

liegt. Die Menge aller algebraisch inneren Punkte von M heißt das **algebraische Innere** (englisch: *algebraic interior, core*) von M, kurz: $\operatorname{core} M$. △

Es ist leicht zu sehen, dass $\operatorname{int} M \subseteq \operatorname{core} M$ gilt. Die Umkehrung ist jedoch i. A. falsch. Es gilt aber:

Lemma 15.25 (Algebraisches Inneres konvexer Mengen)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex. Dann gilt $\operatorname{int} C = \operatorname{core} C$.

Beweis. Es ist nur $\operatorname{core} C \subseteq \operatorname{int} C$ zu bestätigen. Im Fall $\operatorname{core} C = \emptyset$ ist nichts zu zeigen. Es sei also $x \in \operatorname{core} C$. Dann ist $\dim C = n$ und damit $\operatorname{aff} C = \mathbb{R}^n$. (**Quizfrage 15.16:** Begründung?) Nach Definition des algebraischen Inneren ist die Bedingung aus Lemma 15.19 (*ii*) erfüllt. Damit gehört x zu $\operatorname{relint} C = \operatorname{int} C$. □

Quizfrage 15.17: Können Sie ein Beispiel einer Menge angeben und einen Punkt, der ein algebraisch innerer Punkt ist aber kein innerer Punkt?

Ende der Vorlesung 21

Ende der Woche 11

§ 15.4 Trennungssätze

Literatur: Geiger, Kanzow, 2002, Kapitel 2.1.4

Wir kommen nun zu den wichtigsten Resultaten in § 15, dem Trennungssatz 15.28, dem eigentlichen Trennungssatz 15.32 und dem strikten Trennungssatz 15.35. 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.

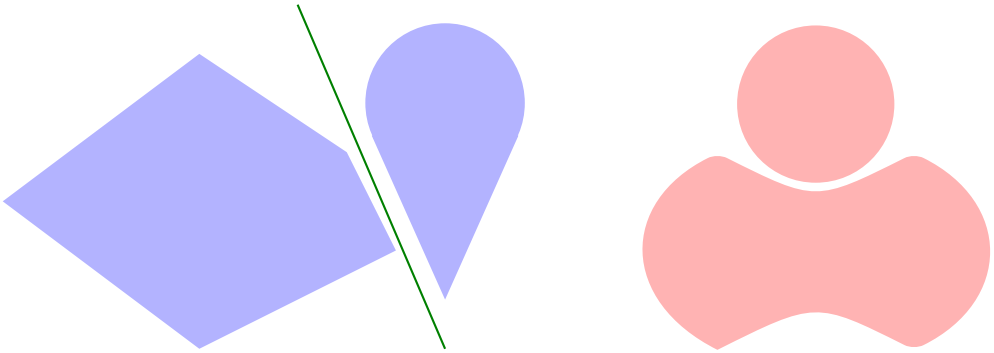


Abbildung 15.6.: Zwei disjunkte konvexe Mengen (links) sind durch eine Hyperebene trennbar. Ist eine der Mengen nichtkonvex, stimmt diese Aussage nicht mehr (rechts).

Definition 15.26 (Trennende Hyperebene)**.**
Es seien $A, B \subseteq \mathbb{R}^n$ zwei nichtleere Mengen und $H(a,\beta) = \{x \in \mathbb{R}^n \mid a^\mathsf{T} x = \beta\}$ eine Hyperebene.

(*i*) Wir sagen, $H(a,\beta)$ sei eine **trennende Hyperebene** (englisch: *separating hyperplane*) für die Mengen A und B, falls eine der Mengen in $H^-(a,\beta)$ und die andere in $H^+(a,\beta)$ enthalten ist, wenn also gilt

$$a^\mathsf{T} x \leqslant \beta \leqslant a^\mathsf{T} y \quad \text{für alle } x \in A \text{ und alle } y \in B. \tag{15.15}$$

(*ii*) Wir sagen, die Hyperebene $H(a,\beta)$ sei eine **eigentlich trennende Hyperebene** (englisch: *properly separating hyperplane*) für die Mengen A und B, falls $H(a,\beta)$ die Mengen A und B trennt, aber nicht beide Mengen A und B enthält, wenn also gilt:

$$a^\mathsf{T} x \leqslant \beta \leqslant a^\mathsf{T} y \quad \text{für alle } x \in A \text{ und alle } y \in B \tag{15.16a}$$

und

$$a^\mathsf{T} \overline{x} < a^\mathsf{T} \overline{y} \quad \text{für ein } \overline{x} \in A \text{ und ein } \overline{y} \in B. \tag{15.16b}$$

(*iii*) Wir sagen, die Hyperebene $H(a,\beta)$ sei eine **strikt trennende Hyperebene** (englisch: *strictly separating hyperplane*) für die Mengen A und B, falls genau eine der Mengen im offenen Halbraum $\operatorname{int} H^-(a,\beta)$ und die andere im offenen Halbraum $\operatorname{int} H^+(a,\beta)$ enthalten ist, wenn also gilt

$$a^\mathsf{T} x < \beta < a^\mathsf{T} y \quad \text{für alle } x \in A \text{ und alle } y \in B. \tag{15.17}$$

△

Beachte: Der „Offset“ β ist nicht die entscheidende Größe in der Definition 15.26. Gilt beispielsweise $a^\mathsf{T} x \leqslant a^\mathsf{T} y$ für alle $x \in A$ und alle $y \in B$, dann lässt sich β in (15.15) immer nachträglich passend

definieren. (**Quizfrage 15.18:** Wie kann man in den einzelnen Fällen der Definition 15.26 das β jeweils passend definieren, wenn man den Normalenvektor a bereits kennt?)

Wir beginnen mit einer Aussage zur Trennung eines Punktes und einer konvexen Menge. Aus diesem Speziallfall leiten wir dann anschließend den ersten Trennungssatz 15.28 her.

Lemma 15.27 (Trennung von Punkt und konvexer Menge)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex und nichtleer und $\overline{x} \in \mathbb{R}^n$ ein Punkt, sodass $\overline{x} \notin \operatorname{int}(C)$. Dann existiert eine Hyperebene $H(a, \beta)$, die $\overline{x}$ von C trennt, sodass also gilt:

$$a^\mathsf{T} x \geqslant \beta \geqslant a^\mathsf{T} \overline{x} \quad \text{für alle } x \in C. \tag{15.18}$$

Beweis. **Schritt 1:** Wir wählen eine Folge $x^{(k)}$ mit der Eigenschaft $x^{(k)} \notin \overline{C}$ und $x^{(k)} \to \overline{x}$.

Wir unterscheiden drei Fälle.

Fall 1: Falls $\overline{x} \notin \overline{C}$ liegt, wähle $x^{(k)} \equiv \overline{x}$.

Fall 2: Falls $\overline{x} \in \overline{C}$ liegt und C nicht die volle Dimension hat, so ist $C \subseteq x_0 + U$ mit einem Unterraum U einer Dimension $< n$. Wir können dann $x^{(k)} = \overline{x} + (1/k)\, a$ wählen mit einem $a \in U^\perp$.

Fall 3: Falls $\overline{x} \in \overline{C}$ liegt und C volle Dimension hat, dann liegt $\overline{x} \in \overline{C} \setminus \operatorname{int}(C) = \partial C = \partial(\overline{C})$ nach Satz 15.21(*iii*). Jede Umgebung eines Randpunktes von $\overline{C}$ schneidet die Menge $\overline{C}$ und ihr Komplement $\mathbb{R}^n \setminus \overline{C}$. Also können wir $x^{(k)} \in B_{1/k}(\overline{x}) \cap (\mathbb{R}^n \setminus \overline{C})$ auswählen, woraus $x^{(k)} \to \overline{x}$ folgt.

Schritt 2: Wir konstruieren eine Folge $a^{(k)}$ von Normalenvektoren, sodass die zugehörigen Hyperebene den Punkt $x^{(k)}$ von $\overline{C}$ trennt.

Die Menge $\overline{C}$ ist nichtleer, abgeschlossen und konvex. Nach Lemma 15.2 existiert die orthogonale Projektion $\widehat{x}^{(k)} := \operatorname{proj}_{\overline{C}}(x^{(k)})$, und nach Satz 15.3 ist diese charakterisiert durch

$$\begin{aligned}(\widehat{x}^{(k)} - x^{(k)})^\mathsf{T} x &\geqslant (\widehat{x}^{(k)} - x^{(k)})^\mathsf{T} \widehat{x}^{(k)} \quad \text{für alle } k \in \mathbb{N} \text{ und alle } x \in \overline{C} \\ &= (\widehat{x}^{(k)} - x^{(k)})^\mathsf{T} (\widehat{x}^{(k)} - x^{(k)} + x^{(k)}) \\ &= \|\widehat{x}^{(k)} - x^{(k)}\|^2 + (\widehat{x}^{(k)} - x^{(k)})^\mathsf{T} x^{(k)} \\ &\geqslant (\widehat{x}^{(k)} - x^{(k)})^\mathsf{T} x^{(k)}. \end{aligned} \tag{15.19}$$

Wir setzen

$$a^{(k)} := \frac{\widehat{x}^{(k)} - x^{(k)}}{\|\widehat{x}^{(k)} - x^{(k)}\|}.$$

Beachte: Der Nenner ist $\neq 0$, da $\widehat{x}^{(k)} \in \overline{C}$ und $x^{(k)} \notin \overline{C}$.

Damit erhalten wir aus (15.19)

$$a^{(k)\mathsf{T}} x \geqslant a^{(k)\mathsf{T}} x^{(k)} \quad \text{für alle } k \in \mathbb{N} \text{ und alle } x \in \overline{C},$$

d. h., mit dem Normalenvektor $a^{(k)}$ können wir jeweils den Punkt $x^{(k)}$ von $\overline{C}$ trennen.

Schritt 3: Wir konstruieren den Normalenvektor a der gesuchten Hyperebene durch derGrenzübergang:

Wegen $\|a^{(k)}\| = 1$ und der Kompaktheit der Einheitssphäre in $\mathbb{R}^n$ existiert eine konvergente Teilfolge $a^{(k^{(\ell)})} \to a$ mit $\|a\| = 1$, und der Grenzübergang $\ell \to \infty$ zeigt die Behauptung

$$a^\mathsf{T} x \geqslant a^\mathsf{T} \overline{x} \quad \text{für alle } x \in \overline{C}. \qquad \square$$

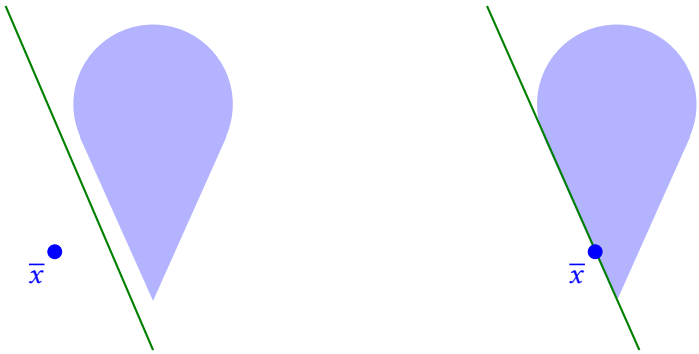


Abbildung 15.7.: Illustration von Lemma 15.27 in zwei Fällen.

Satz 15.28 (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:

$$a^\mathsf{T} x_1 \leqslant \beta \leqslant a^\mathsf{T} x_2 \quad \text{für alle } x_1 \in C_1 \text{ und alle } x_2 \in C_2. \tag{15.20}$$

Beweis. Wir betrachten

$$C := C_2 - C_1 = \{x_2 - x_1 \,|\, x_1 \in C_1,\ x_2 \in C_2\},$$

also die Minkowski-Summe von C_2 und $-C_1$. Nach Satz 13.3 ist C konvex, und wegen $C_1 \cap C_2 = \emptyset$ gilt $0 \notin C$, also erst recht $0 \notin \operatorname{int}(C)$. Aus Lemma 15.27 bekommen wir die Existenz einer Hyperebene $H(a, \beta)$, sodass $0 \leqslant a^\mathsf{T} x$ gilt für alle $x \in C$. Das heißt aber

$$a^\mathsf{T} x_1 \leqslant a^\mathsf{T} x_2 \quad \text{für alle } x_1 \in C_1 \text{ und alle } x_2 \in C_2. \qquad \square$$

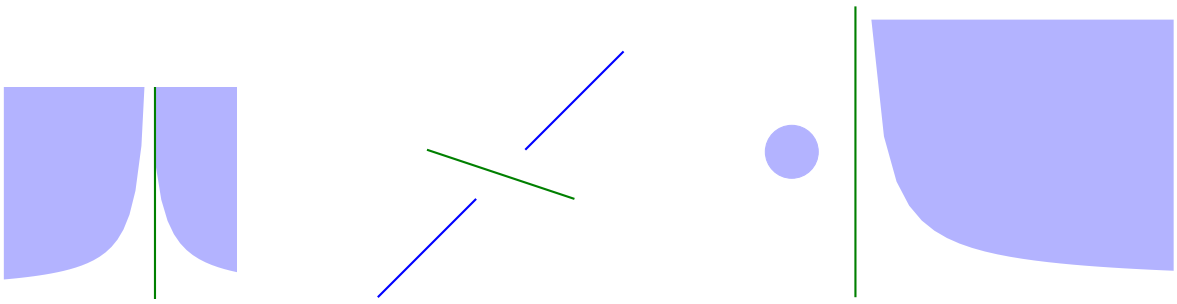


Abbildung 15.8.: Illustration der Aussagen von Satz 15.28 (Trennung disjunkter konvexer Mengen), Satz 15.32 (eigentliche Trennung) und Satz 15.35 (strikte Trennung).

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 15.32 benötigen wir einige vorbereitende Aussagen.

Lemma 15.29 (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*) $A\overline{C} \subseteq \overline{AC}$.

(*ii*) $A \operatorname{relint}(C) = \operatorname{relint}(AC)$.

Beweis. Im Fall $C = \emptyset$ ist nichts zu zeigen. Wir gehen also ab jetzt von $C \neq \emptyset$ aus. Die Konvexität von AC ist offensichtlich.

Aussage (*i*): Es sei $\overline{x} \in \overline{C}$, dann existiert eine Folge $x^{(k)} \subseteq C$ mit der Eigenschaft $x^{(k)} \to \overline{x}$. Das impliziert $Ax^{(k)} \in AC$ und $Ax^{(k)} \to A\overline{x}$, also gilt $A\overline{x} \in \overline{AC}$.

Aussage (*ii*): Es gilt

$$\begin{aligned} A \operatorname{relint}(C) \subseteq AC \subseteq A\overline{C} = A\overline{\operatorname{relint}(C)} \quad & \text{nach Satz 15.21} \\ \subseteq \overline{A \operatorname{relint}(C)} \quad & \text{nach Aussage (}i\text{).} \end{aligned}$$

Die Bildung des Abschlusses in allen Termen dieser Ungleichung zeigt $\overline{AC} = \overline{A \operatorname{relint}(C)}$. Aus Folgerung 15.22 folgt damit:

$$\operatorname{relint}(AC) = \operatorname{relint}(A \operatorname{relint}(C)) \subseteq A \operatorname{relint}(C).$$

Um die umgekehrte Inklusion zu zeigen, sei $z \in A \operatorname{relint}(C)$, d. h., $z = Az'$ für ein $z' \in \operatorname{relint}(C)$. Weiter sei x irgendein Element von AC, d. h., $x = Ax'$ für ein $x' \in C$. Aus Lemma 15.19 (*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 15.19 (*iii*) für die Menge AC erfüllt ist, also gehört z zu $\operatorname{relint}(AC)$, was $A \operatorname{relint}(C) \subseteq \operatorname{relint}(AC)$ zeigt. □

Beachte: Die Aussage (*i*) verwendet nur die Stetigkeit linearer Abbildungen und gilt auch für nichtkonvexe Mengen.

Folgerung 15.30 (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

$$\operatorname{relint}(C_1) + \operatorname{relint}(C_2) = \operatorname{relint}(C_1 + C_2). \tag{15.21}$$

Beweis. Setze $C := C_1 \times C_2 \subseteq \mathbb{R}^n \times \mathbb{R}^n \cong \mathbb{R}^{2n}$ und $A := \begin{bmatrix} \mathrm{Id} & \mathrm{Id} \end{bmatrix}$, sodass also $A\binom{x_1}{x_2} = x_1 + x_2$ gilt. Dann ist C nach Satz 13.3 konvex, und aus Lemma 15.29 folgt

$$\operatorname{relint}(C_1) + \operatorname{relint}(C_2) = A \operatorname{relint}(C) = \operatorname{relint}(A\,C) = \operatorname{relint}(C_1 + C_2). \qquad \square$$

Vor dem Beweis des eigentlichen Trennungssatzes 15.32 betrachten wir wieder zunächst den Spezialfall, dass eine der zu trennenden Mengen nur aus einem Punkt besteht.

Lemma 15.31 (Eigentliche Trennung von Punkt und konvexer Menge)**.**
Es sei $C \subseteq \mathbb{R}^n$ konvex und nichtleer. Falls $0 \notin \operatorname{relint}(C)$ liegt, dann lassen sich 0 und C durch eine Hyperebene $H(a, \beta)$ eigentlich trennen.

Beweis. Im Fall $0 \notin \overline{C}$, wählen wir $a := \operatorname{proj}_{\overline{C}}(0)$. Dann ist $a \neq 0$. Der Projektionssatz 15.3 impliziert

$$(a - 0)^\mathsf{T}(x - a) \geqslant 0 \quad \text{für alle } x \in C,$$

also gilt $a^\mathsf{T}x \geqslant \|a\|^2 > 0 = a^\mathsf{T}0$ für alle $x \in C$. Dies zeigt die eigentliche (und sogar die starke) Trennung von 0 und C in diesem Fall.

Andernfalls gilt $0 \in \overline{C} \setminus \operatorname{relint}(C) = \operatorname{rel}\partial(C)$. Wegen $0 \in \overline{C} \subseteq \operatorname{aff}(C)$ ist $U := \operatorname{aff}(C)$ ein Unterraum von $\mathbb{R}^n$. Da 0 ein relativer Randpunkt von C ist, schneidet jede Kugel $B_\varepsilon(0)$ sowohl U als auch $\overline{C}$ als auch das Komplement $\mathbb{R}^n \setminus \overline{C}$. Durch die Wahl $\varepsilon := 1/k$ können wir also eine Folge $x^{(k)} \subseteq U$ konstruieren mit der Eigenschaft $x^{(k)} \notin \overline{C}$ und $x^{(k)} \to 0$. Wir setzen

$$\widehat{x}^{(k)} := \operatorname{proj}_{\overline{C}}(x^{(k)}).$$

Diese Folge erfüllt $\widehat{x}^{(k)} \neq x^{(k)}$ und $\widehat{x}^{(k)} \to 0$, und der Projektionssatz 15.3 impliziert

$$(\widehat{x}^{(k)} - x^{(k)})^\mathsf{T}(x - \widehat{x}^{(k)}) \geqslant 0 \quad \text{für alle } x \in C. \tag{15.22}$$

Wir setzen

$$a^{(k)} := \frac{\widehat{x}^{(k)} - x^{(k)}}{\|\widehat{x}^{(k)} - x^{(k)}\|}$$

und schreiben (15.22) als

$$(a^{(k)})^\mathsf{T}x \geqslant (a^{(k)})^\mathsf{T}\widehat{x}^{(k)} \quad \text{für alle } x \in C. \tag{15.23}$$

Wegen $\|a^{(k)}\| = 1$ gibt es eine konvergente Teilfolge $a^{(k^{(\ell)})} \to a$ mit $\|a\| = 1$ und insbesondere $a \neq 0$. Weiter gehört wegen $x^{(k)} \in U$ und $\widehat{x}^{(k)} \in \overline{C} \subseteq U$ auch $a^{(k)}$ zu U, und damit gilt auch $a \in U$. Der Grenzübergang auf der Teilfolge in (15.23) zeigt schließlich

$$a^\mathsf{T}x \geqslant a^\mathsf{T}0 = 0 \quad \text{für alle } x \in C$$

Es bleibt zu zeigen, dass es ein $\overline{x} \in C$ gibt, sodass $a^\mathsf{T}\overline{x} > 0$ gilt. Nehmen wir an, dies sei nicht der Fall, d. h., es gelte $a^\mathsf{T}x = 0$ für alle $x \in C$. Weil $a \in U = \operatorname{aff}(C)$ liegt, gibt es Punkte $x^{(i)} \in C$ und Koeffzienten $\alpha^{(i)}$, 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 $\overline{x} \in C$ geben, sodass $a^\mathsf{T}\overline{x} > 0$ gilt. $\square$

Satz 15.32 (Eigentlicher Trennungssatz, vgl. Jarre, Stoer, 2004, Satz 7.2.8)**.**
Es seien $C_1, C_2 \subseteq \mathbb{R}^n$ konvex und nichtleer. Dann sind äquivalent:

(*i*) C_1 und C_2 lassen sich durch eine Hyperebene $H(a, \beta)$ eigentlich trennen.

(*ii*) $\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2) = \emptyset$.

Beweis. Wir nehmen zunächst an, dass C_1 und C_2 durch $H(a, \beta)$ eigentlich getrennt werden. Es gilt also $a^\mathsf{T} x \leqslant \beta \leqslant a^\mathsf{T} y$ für alle $x \in C_1$ und alle $y \in C_2$, und es gibt $\overline{x} \in C_1$ und $\overline{y} \in C_2$, für die $a^\mathsf{T}\overline{x} < a^\mathsf{T}\overline{y}$ gilt. Wir zeigen nun:

$$a^\mathsf{T} x < a^\mathsf{T} y \quad \text{für alle } x \in \operatorname{relint}(C_1) \text{ und alle } y \in \operatorname{relint}(C_2), \tag{15.24}$$

woraus dann $\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2) = \emptyset$ folgt. Nehmen wir an, dass (15.24) unwahr ist, dann gibt es ein $x \in \operatorname{relint}(C_1)$ und ein $y \in \operatorname{relint}(C_2)$ mit $a^\mathsf{T} x = a^\mathsf{T} y$. Für hinreichend kleines $\varepsilon > 0$ sind $\widehat{x} \coloneqq x - \varepsilon\,(\overline{x} - x) \in C_1$ und $\widehat{y} \coloneqq y - \varepsilon\,(\overline{y} - y) \in C_2$. Dann ist aber

$$a^\mathsf{T}(\widehat{x} - \widehat{y}) = a^\mathsf{T}\big(x - \varepsilon\,(\overline{x} - x) - y + \varepsilon\,(\overline{y} - y)\big) = \varepsilon\, a^\mathsf{T}(\overline{y} - \overline{x}) > 0$$

im Widerspruch zu $a^\mathsf{T} x \leqslant a^\mathsf{T} y$ für alle $x \in C_1$ und alle $y \in C_2$.

Umgekehrt nehmen wir nun an, dass $\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2) = \emptyset$ gilt. Insbesondere ist dann

$$0 \notin \operatorname{relint}(C_1) - \operatorname{relint}(C_2) = \operatorname{relint}(C_1 - C_2),$$

wobei die Gleichheit der Mengen wie in Folgerung 15.30 folgt. Nach Lemma 15.31 können also 0 und $C \coloneqq C_1 - C_2$ eigentlich getrennt werden. Das heißt, es gibt ein $a \in \mathbb{R}^n$, $a \neq 0$, sodass $a^\mathsf{T} x \leqslant 0$ gilt für alle $x \in C$, und es existiert ein $\overline{x} \in C$ mit $a^\mathsf{T}\overline{x} < 0$. Das bedeutet aber $a^\mathsf{T} x_1 \leqslant a^\mathsf{T} x_2$ für alle $x_1 \in C_1$ und alle $x_2 \in C_2$, und für gewisse $\overline{x}_1 \in C_1$ und $\overline{x}_2 \in C_2$ gilt $a^\mathsf{T}\overline{x}_1 < a^\mathsf{T}\overline{x}_2$. □

Wir bereiten nun den strikten Trennungssatz 15.35 vor.

Lemma 15.33 (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 $M \coloneqq M_1 + M_2$. Falls M_1 oder M_2 die leere Menge ist, dann ist $M = \emptyset$ und die Aussage klar. Es seien nun also $M_1, M_2 \neq \emptyset$ und damit $M \neq \emptyset$. Weiter sei $z^{(k)} \subseteq M$ eine konvergente Folge mit Grenzwert z. Es existieren also Folgen $x^{(k)} \subseteq M_1$ und $y^{(k)} \subseteq M_2$ mit $z^{(k)} = x^{(k)} + y^{(k)}$.

$$M_2 \text{ ist beschränkt} \quad \Rightarrow \quad y^{(k)} \text{ ist beschränkt} \quad \Rightarrow \quad x^{(k)} \text{ ist beschränkt.}$$

Es existieren also konvergente Teilfolgen $x^{(k^{(\ell)})} \to x$ und $y^{(k^{(\ell)})} \to y$, sodass auch $z^{(k^{(\ell)})} = x^{(k^{(\ell)})} + y^{(k^{(\ell)})} \to x + y$ für $\ell \to \infty$. M_1 und M_2 sind abgeschlossen, also liegt $x + y \in M_1 + M_2 = M$. Andererseits konvergiert $z^{(k^{(\ell)})}$ auch gegen z, also gilt $z = x + y \in M$, d. h., M ist abgeschlossen. □

Beispiel 15.34 (Die Minkowski-Summe zweier abgeschlossener Mengen ist nicht notwendig abgeschlossen)**.**
Das folgende Gegenbeispiel zeigt, dass die Minkowski-Summe zweier abgeschlossener Mengen nicht notwendig abgeschlossen ist, selbst wenn die Mengen konvex sind:

$$M_1 = \left\{ \binom{x}{y} \,\middle|\, xy \geqslant 1,\ x > 0 \right\} \quad \text{und} \quad M_2 = \mathbb{R} \times \{0\} \quad \text{mit} \quad M_1 + M_2 = \mathbb{R} \times (0, \infty). \qquad \triangle$$

Satz 15.35 (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

$$a^\mathsf{T} x_1 < \beta < a^\mathsf{T} x_2 \quad \text{für alle } x_1 \in C_1 \text{ und alle } x_2 \in C_2. \tag{15.25}$$

Beweis. Wir betrachten die Optimierungsaufgabe

$$\text{Minimiere} \quad \|x_1 - x_2\|, \quad (x_1, x_2) \in C_1 \times C_2. \tag{15.26}$$

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 15.33 abgeschlossen ist. (15.26) ist also gleichzeitig eine Aufgabe der orthogonalen Projektion (15.1). Dabei wird der Nullvektor $0 \in \mathbb{R}^n$ auf die konvexe, abgeschlossene Menge C projiziert. Es sei nun $(x_1^*, x_2^*) \in C_1 \times C_2$ die nach Lemma 15.2 eindeutige Lösung von (15.26). Wir konstruieren daraus nun die Daten (a, β) der Hyperebene und setzen dazu

$$a := \frac{x_2^* - x_1^*}{2} \neq 0, \quad \widehat{x} := \frac{x_1^* + x_2^*}{2}, \quad \beta := a^\mathsf{T} \widehat{x}.$$

Wir zeigen nun, dass

$$x_1^* = \operatorname{proj}_{C_1}(\widehat{x}), \quad x_2^* = \operatorname{proj}_{C_2}(\widehat{x}). \tag{15.27}$$

gilt. Dazu seien $x_1 \in C_1$ und $x_2 \in C_2$ beliebig, dann gilt

$$\begin{aligned}
\|x_1^* - \widehat{x}\| + \|\widehat{x} - x_2^*\| &= \|x_1^* - \widehat{x} + (\widehat{x} - x_2^*)\|, && \text{denn } x_1^* - \widehat{x} = \widehat{x} - x_2^* \\
&= \|x_1^* - x_2^*\| \\
&\leqslant \|x_1 - x_2\|, && \text{denn } (x_1^*, x_2^*) \text{ ist optimal für (15.26)} \\
&\leqslant \|x_1 - \widehat{x}\| + \|\widehat{x} - x_2\| && \text{wegen der Dreiecksungleichung.}
\end{aligned}$$

Setzen wir speziell $x_2 = x_2^*$ ein, so folgt

$$\|x_1^* - \widehat{x}\| \leqslant \|x_1 - \widehat{x}\| \quad \text{für alle } x_1 \in C_1.$$

Setzen wir dagegen $x_1 = x_1^*$, so folgt

$$\|x_2^* - \widehat{x}\| \leqslant \|x_2 - \widehat{x}\| \quad \text{für alle } x_2 \in C_2.$$

Dies bestätigt (15.27). Aus dem Projektionssatz 15.3 folgt daher

$$-a^\mathsf{T}(x_1 - x_1^*) = (x_1^* - \widehat{x})^\mathsf{T}(x_1 - x_1^*) \geqslant 0 \quad \text{für alle } x_1 \in C_1.$$

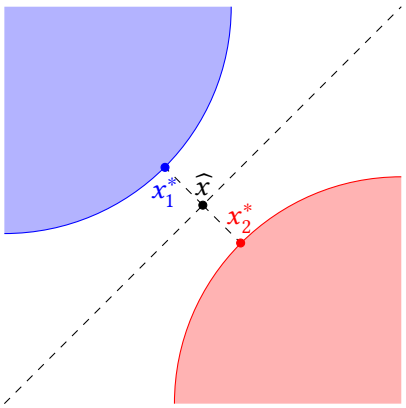


Abbildung 15.9.: Illustration der Lage der Punkte x_1^*, x_2^* und $\widehat{x}$ im Beweis des strikten Trennungssatzes 15.35.

Dies impliziert

$$a^\mathsf{T} x_1 \leqslant a^\mathsf{T} x_1^* = a^\mathsf{T}\widehat{x} + a^\mathsf{T}(x_1^* - \widehat{x}) = \beta - \|a\|^2 < \beta \quad \text{für alle } x_1 \in C_1.$$

Analog zeigt man

$$a^\mathsf{T} x_2 \geqslant a^\mathsf{T} x_2^* = a^\mathsf{T}\widehat{x} + a^\mathsf{T}(x_2^* - \widehat{x}) = \beta + \|a\|^2 > \beta \quad \text{für alle } x_2 \in C_2. \qquad \square$$

Quizfrage 15.20: Wie kann man die „Lücke“ $\inf\{a^\mathsf{T} x_2 \mid x_2 \in C_2\} - \sup\{a^\mathsf{T} x_1 \mid x_1 \in C_1\}$ interpretieren?

Bemerkung 15.36 (Der strikte Trennungssatz und das Farkas-Lemma)**.**
Das Farkas-Lemma 8.6 ist eine spezielle Version des strikten Trennungssatzes 15.35, und zwar für den Fall, dass die Menge C_1 im Trennungssatz der konvexe abgeschlossene *Kegel*

$$C_1 = \left\{ B^\mathsf{T}\xi \,\middle|\, \xi \in \mathbb{R}^m,\ \xi \geqslant 0 \right\}$$

und C_2 die kompakte *einpunktige* Menge $C_2 = \{c\}$ ist. $\triangle$

Expertenwissen: Beweis der Behauptung aus Bemerkung 15.36

Die Behauptung ist, das Farkas-Lemma 8.6 sei äquivalent zum strikten Trennungssatz 15.35.

Der Übersicht halber tragen wir die beiden untereinander äquivalenten Aussagen im Farkas-Lemma 8.6 hier nochmal zusammen:

- Aussage (i): Das System $B^\mathsf{T}\xi = c$ besitzt eine Lösung $\xi \geqslant 0$.
- Aussage (ii): Es gilt $c^\mathsf{T} d \geqslant 0$ für alle Elemente der Menge $\{d \in \mathbb{R}^n \mid B\,d \geqslant 0\}$.

Wie wir im Beweis von Lemma 8.6 gesehen haben, ist die Folgerung Aussage $(i) \Rightarrow$ Aussage (ii) elementar. Wir zeigen nun, dass wir mit Hilfe des strikten Trennungssatzes 15.35 die wesentliche Aussage $\neg$ Aussage $(i) \Rightarrow \neg$ Aussage (ii) beweisen kann.

Die Menge C_1 ist ein konvexer abgeschlossener Kegel (siehe Lemma 6.13), und C_2 ist konvex und kompakt. Aufgrund von $\neg$ Aussage (i) sind C_1 und C_2 disjunkt. Es gibt also eine Hyperebene

$H(d, \beta)$, sodass gilt:

$$d^\mathsf{T} x > \beta > d^\mathsf{T} c \quad \text{für alle } x \in C_1.$$

(Achtung, die Relationszeichen sind andersherum gewählt als in (15.25).) Aufgrund von $0 \in C_1$ ist $\beta < 0$, also $c^\mathsf{T} d < 0$. Weiter gilt $d^\mathsf{T} x \geqslant 0$ für alle $x \in C_1$, denn wäre $d^\mathsf{T} x < 0$ für ein $x \in C_1$, dann könnten wir x wegen der Kegeleigenschaft durch $\alpha\, x$ ersetzen, was mit $\alpha \to \infty$ zum Widerspruch führt.

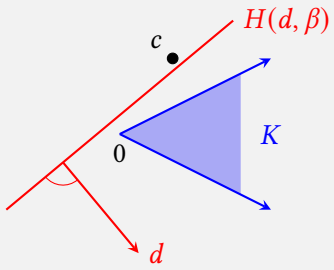
Wir haben also

$$d^\mathsf{T} x \geqslant 0 > \beta > c^\mathsf{T} d \quad \text{für alle } x \in C_1,$$

d. h.,

$$d^\mathsf{T} B^\mathsf{T} \xi \geqslant 0 > \beta > c^\mathsf{T} d \quad \text{für alle } \xi \geqslant 0.$$

Aus der ersten Ungleichung folgt aber $B\, d \geqslant 0$. Das heißt, wir haben tatsächlich ein $d \in \mathbb{R}^n$ konstruiert mit der Eigenschaft $B\, d \geqslant 0$, aber $c^\mathsf{T} d < 0$, das ist gerade ¬ Aussage (ii).



Ende der Vorlesung 22

§ 16 Das Subdifferential und die Richtungsableitung konvexer Funktionen

Literatur: Geiger, Kanzow, 2002, Kapitel 6.3, Rockafellar, 1970

Ziel: Verallgemeinerung der Ableitung für nicht-glatte konvexe Funktionen

§ 16.1 Das Subdifferential

Definition 16.1 (Subdifferential).
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ eine konvexe Funktion.

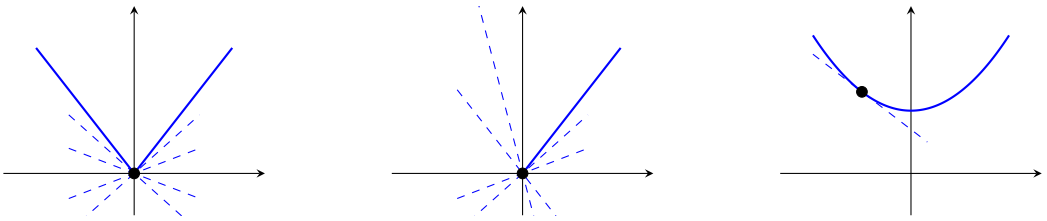
(i) Ein Vektor $s \in \mathbb{R}^n$ heißt ein (Euklidischer) **Subgradient** (englisch: *subgradient*) von f im Punkt $x_0 \in \mathbb{R}^n$, wenn die **Subgradientenungleichung** (englisch: *subgradient inequality*) gilt:

$$f(x) \geqslant f(x_0) + s^\mathsf{T}(x - x_0) \quad \text{für alle } x \in \mathbb{R}^n. \tag{16.1}$$

Wenn $f(x_0) \in \mathbb{R}$ liegt, dann sagt man: Die rechte Seite in (16.1) ist eine **affine Minorante** (englisch: *affine minorant*) mit Euklidischem Gradienten s, die f in x_0 **stützt**, kurz: eine **affine Stützfunktion** (englisch: *affine supporting function*), siehe Abbildung 16.1.

(*ii*) Die Menge $\partial f(x_0)$ *aller* Subgradienten im Punkt x_0 heißt das **Subdifferential** (englisch: *subdifferential*) von f in x_0.

(*iii*) f heißt **subdifferenzierbar** (kurz: **subdiffbar**, englisch: *subdifferentiable*) im Punkt $x_0 \in \mathbb{R}^n$, wenn $\partial f(x_0) \neq \emptyset$ ist. △



(a) $f(x) = |x|$, $\partial f(0) = [-1, 1]$

(b) $f(x) = |x| + I_{[0,\infty)}$, $\partial f(0) = (-\infty, 1]$

(c) f an der Stelle x_0 diffbar, $\partial f(x_0) = \{\nabla f(x_0)\}$

Abbildung 16.1.: Das Subdifferential $\partial f(x_0)$ besteht aus den *Steigungen* aller affinen Minoranten an f im Punkt x_0.

Bemerkung 16.2 (Zum Subdifferential)**.**

(*i*) (16.1) verallgemeinert die Ungleichung (13.15), die für konvexe *diffbare* Funktionen gilt. Das Subdifferential an der Stelle x_0 besteht gerade aus allen Gradienten affiner Funktionen, die f im Punkt x_0 stützen.

(*ii*) Eine andere Anschauung der Subgradientenungleichung (16.1) besagt, dass $\begin{pmatrix} s \\ -1 \end{pmatrix}$ der Normalenvektor einer Hyperebene durch den Punkt $\begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix}$ ist, die den (Epi-)Graphen in diesem Punkt stützt. △

Expertenwissen: Zum Begriff „Subgradient“

Die Bezeichnung **Subgradient** ist eigentlich irreführend, weil das Subdifferential den Begriff der Ableitung ersetzt (ein Element des Dualraumes von $\mathbb{R}^n$) und nicht den Gradienten (der ja nichts anderes als eine vom verwendeten Innenprodukt abhängige primale Darstellung der Ableitung ist). Konzeptionell wäre es daher besser, einen Vektor $s \in \mathbb{R}_n$ (Zeilenvektor) eine **Subableitung** (englisch: *subderivative*) von f im Punkt $x_0 \in \mathbb{R}^n$ zu nennen, wenn $f(x) \geqslant f(x_0) + s\,(x - x_0)$ 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 Innenprodukts). (**Quizfrage 16.1:** Wie würde die Definition eines Subgradienten von f im Punkt x_0 bzgl. des M-Innenprodukts aussehen?)

Eine einfache, aber wichtige Anwendung des Subdifferentials ist folgende.

Satz 16.3 (0 im Subdifferential)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ eine konvexe Funktion. Dann sind äquivalent:

(*i*) x_0 ist ein globaler Minimierer von f.

(*ii*) $f \not\equiv \infty$ und $0 \in \partial f(x_0)$.

Beweis. Aussage (*i*) $\Rightarrow$ Aussage (*ii*): Es sei x_0 ein globaler Minimierer von f, insbesondere gilt nach Definition 14.1 also $f(x_0) < \infty$. Aufgrund der globalen Optimalität von x_0 folgt

$$f(x) \geqslant f(x_0) + 0^\mathsf{T}(x - x_0),$$

was $0 \in \partial f(x_0)$ bestätigt.

Aussage (*ii*) $\Rightarrow$ Aussage (*i*): Aus $0 \in \partial f(x_0)$ folgt sofort mit (16.1), dass $f(x) \geqslant f(x_0)$ für alle $x \in \mathbb{R}^n$ gilt. Außerdem kann nicht $f(x_0) = \infty$ gelten, da f nach Voraussetzung nicht identisch ∞ ist. (**Quizfrage 16.2:** Genaue Begründung?) Damit ist x_0 ein globaler Minimierer von f, siehe Definition 14.1. □

Expertenwissen: Prüfung der Subdifferentialungleichung (16.1) nur in einer Umgebung

Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ eine konvexe Funktion und $x_0 \in \mathbb{R}^n$ mit $f(x_0) < \infty$. Dann gilt $s \in \partial f(x_0)$ genau dann, wenn

$$f(\overline{x}) \geqslant f(x_0) + s^\mathsf{T}(\overline{x} - x_0) \quad \text{für alle } \overline{x} \in B_\varepsilon(x_0) \tag{16.2}$$

gilt. Es ist also hinreichend, die Subgradientenungleichung (16.1) nur für alle x in einer Umgebung von x_0 zu prüfen, den Rest erledigt die Konvexität von f.

Offenbar impliziert (16.1) auch (16.2). Wir müssen nur zeigen, dass (16.2) auch (16.1) impliziert.

Im Fall $f(x_0) = -\infty$ ist (16.1) sofort klar. Es sei also von nun an $f(x_0)$ endlich und $x \in \mathbb{R}^n$ mit $\|x - x_0\| \geqslant \varepsilon$ beliebig, aber fest. Wir setzen $\overline{x} := \overline{\alpha}\, x + (1 - \overline{\alpha})\, x_0$ mit $\overline{\alpha} = \frac{\varepsilon/2}{\|x - x_0\|} \in (0, 1/2]$. Dann ist $\overline{x} \in B_\varepsilon(x_0)$.

Aus der Konvexität von f folgt (beachte, dass die rechte Seite für alle $\alpha \in [0,1]$ definiert ist)

$$f(\overline{x}) \leqslant \overline{\alpha}\, f(x) + (1 - \overline{\alpha})\, f(x_0).$$

Daraus folgt

$$\begin{aligned}
\overline{\alpha}\, f(x) &\geqslant f(\overline{x}) - (1 - \overline{\alpha})\, f(x_0) \\
&\geqslant f(x_0) - (1 - \overline{\alpha})\, f(x_0) + s^\mathsf{T}(\overline{x} - x_0) \quad \text{wegen (16.2)} \\
&= \overline{\alpha}\, f(x_0) + s^\mathsf{T}\big(\overline{\alpha}\, x + (1 - \overline{\alpha})\, x_0 - x_0\big) \quad \text{da } \overline{x} = \overline{\alpha}\, x + (1 - \overline{\alpha})\, x_0 \\
&= \overline{\alpha}\, f(x_0) + \overline{\alpha}\, s^\mathsf{T}(x - x_0)
\end{aligned}$$

und schließlich nach Division durch $\overline{\alpha}$:

$$f(x) \geqslant f(x_0) + s^\mathsf{T}(x - x_0).$$

Das heißt, (16.1) und (16.2) sind in der Tat äquivalent.

Der Beweis zeigt auch, dass in Wirklichkeit sogar folgende Bedingung reicht, um $s \in \partial f(x_0)$ zu zeigen: Für alle $d \in \mathbb{R}^n$ mit $\|d\| = 1$ existiert ein $\varepsilon_d > 0$, sodass

$$f(\overline{x}) \geqslant f(x_0) + s^\mathsf{T}(\overline{x} - x_0) \quad \text{für alle } \overline{x} \in \{x_0 + t\, d \mid 0 \leqslant t < \varepsilon_d\}$$

gilt.

Die Voraussetzung $f(x_0) < \infty$ ist aber wesentlich: Wenn nämlich $f(x_0) = \infty$ ist, dann zeigt (16.2), dass f in der ganzen Umgebung $B_\varepsilon(x_0)$ gleich ∞ ist. Außerhalb dieser Umgebung kann aber f wieder irgendwo endlich sein. Das heißt aber, die Subgradientenungleichung (16.1) ist an der Stelle x_0 für kein $s \in \mathbb{R}^n$ erfüllbar.

Ein konkretes Beispiel ist $f = I_C$ mit $C = [0, \infty)$ in $\mathbb{R}$, $x_0 = 1$. Dann ist (16.2) für $\varepsilon = 1/2$ mit allen $s \in \mathbb{R}^n$ erfüllt („$\infty \geqslant \infty$“). Jedoch ist $\partial f(x_0) = \emptyset$.

Mit Hilfe der gerade bewiesenen Äquivalenz von (16.1) und (16.2) sowie Satz 16.3 können wir nochmals bestätigen, dass lokale Minimierer konvexer Funktionen bereits globale Minimierer sind: Ist also x^* ein lokaler Minimierer, dann gilt nach Definition $f(x^*) < \infty$ und nach Voraussetzung $f(x^*) \leqslant f(\overline{x})$ für alle $\overline{x} \in B_\varepsilon(x^*)$. Also ist (16.2) für $s = 0$ erfüllt. Nach der obigen Erkenntnis ist also $0 \in \partial f(x^*)$. Das bedeutet nach Satz 16.3, dass x^* ein globaler Minimierer ist.

Satz 16.4 (Elementare Eigenschaften des Subdifferentials)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ eine konvexe Funktion und $x_0 \in \mathbb{R}^n$. Dann ist $\partial f(x_0)$ abgeschlossen und konvex (möglicherweise leer).

Beweis. Es sei $\left(s^{(n)}\right)$ eine Folge in $\partial f(x_0)$. Nach (16.1) gilt also

$$f(x) \geqslant f(x_0) + \left(s^{(n)}\right)^\mathsf{T}(x - x_0) \quad \text{für alle } x \in \mathbb{R}^n.$$

Konvergiert $s^{(n)} \to s \in \mathbb{R}^n$, dann folgt durch Grenzübergang

$$f(x) \geqslant \lim_{n\to\infty} \left[f(x_0) + \left(s^{(n)}\right)^\mathsf{T}(x - x_0)\right] = f(x_0) + s^\mathsf{T}(x - x_0) \quad \text{für alle } x \in \mathbb{R}^n.$$

Das heißt, $\partial f(x_0)$ ist abgeschlossen.

Es seien nun $r, s \in \partial f(x_0)$ und $\alpha \in [0, 1]$. Die gewichtete Addition der Subgradientenungleichungen

$$f(x) \geqslant f(x_0) + r^\mathsf{T}(x - x_0)$$
$$f(x) \geqslant f(x_0) + s^\mathsf{T}(x - x_0)$$

ergibt

$$f(x) \geqslant f(x_0) + \left(\alpha\, r + (1 - \alpha)\, s\right)^\mathsf{T}(x - x_0)$$

für alle $x \in \mathbb{R}^n$. Also gehört auch $\alpha\, r + (1 - \alpha)\, s$ zu $\partial f(x_0)$. Das heißt, $\partial f(x_0)$ ist konvex. □

Beispiel 16.5 (Beispiele zum Subdifferential)**.**
Wir betrachten als Beispiel für f verschiedene Normen auf $\mathbb{R}^n$. Der Nachweis der nachfolgenden Aussagen ist Inhalt von Hausaufgabe 12.3.

(*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 $i = 1, \ldots, n$; siehe auch Abbildung 16.2 für eine Illustration.

(*ii*) $f(x) = \|x\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$. Dann gilt

$$\partial f(x) = \begin{cases} \left\{\frac{x}{\|x\|_2}\right\}, & \text{falls } x \neq 0, \\ \{s \in \mathbb{R}^n \mid \|s\|_2 \leqslant 1\}, & \text{falls } x = 0. \end{cases}$$

(*iii*) $f(x) = \|x\|_\infty = \max_{i=1,\dots,n} |x_i|$. Dann gilt für $x \neq 0$

$$\partial f(x) = \left\{ s \in \mathbb{R}^n \;\middle|\; \begin{array}{r} \|s\|_1 = 1,\ s_i \geqslant 0 \text{ für diejenigen } i \text{ mit } \quad x_i = \|x\|_\infty, \\ s_i \leqslant 0 \text{ für diejenigen } i \text{ mit } -x_i = \|x\|_\infty, \\ \text{und } s_i = 0 \text{ für diejenigen } i \text{ mit } \ |x_i| < \|x\|_\infty \end{array} \right\}$$

sowie

$$\partial f(0) = \{s \in \mathbb{R}^n \mid \|s\|_1 \leqslant 1\}. \qquad \triangle$$

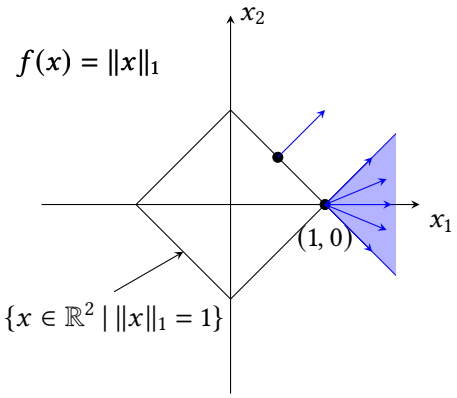


Abbildung 16.2.: Das Subdifferential der 1-Norm $\|x\|_1 = |x_1| + |x_2|$ (Beispiel 16.5) in zwei Punkten aus der Levelmenge $\{x \in \mathbb{R}^2 \mid \|x\|_1 = 1\}$.

Das folgende Beispiel zeigt, dass das Subdifferential eigentlicher konvexer Funktionen auch in Punkten, die zu $\operatorname{dom} f$ gehören, leer sein kann.

Beispiel 16.6 (Leeres Subdifferential)**.**

(*i*) 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 $\partial f(1) = \partial f(-1) = \emptyset$.

(*ii*) 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 $\partial f(0) = \emptyset$. $\triangle$

Das Beispiel 16.6 deutet schon darauf hin, dass das Subdifferential im relativen Inneren von dom f nicht leer ist, während es in Punkten des relativen Randes von dom f auch leer sein kann. Das werden wir auch gleich in Satz 16.8 beweisen, wofür wir folgendes Resultat benötigen:

Lemma 16.7 (Relatives Inneres des Epigraphen, vgl. Hiriart-Urruty, Lemaréchal, 2001, Proposition B.1.1.9)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex. Dann gilt

$$\operatorname{relint} \operatorname{epi} f = \left\{ \begin{pmatrix} x \\ \gamma \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R} \,\middle|\, x \in \operatorname{relint} \operatorname{dom} f,\ \gamma > f(x) \right\}. \tag{16.3}$$

Beweis. Es sei $A \in \mathbb{R}^{n\times(n+1)}$ die Matrix $A = \begin{bmatrix} \mathrm{Id}_n & 0 \end{bmatrix}$. Als lineare Abbildung bedeutet $x \mapsto A\,x$ das Weglassen der $(n+1)$-ten Komponente von x. Es gilt $\operatorname{dom} f = A\,(\operatorname{epi} f)$. (**Quizfrage 16.3:** Klar?) Aus Lemma 15.29 folgt daher

$$\operatorname{relint} \operatorname{dom} f = \operatorname{relint}(A\,(\operatorname{epi} f)) = A\ \operatorname{relint}(\operatorname{epi} f). \tag{16.4}$$

Das bedeutet, dass die x-Komponenten der Elemente $\binom{x}{\gamma} \in \operatorname{relint}(\operatorname{epi} f)$ genau $\operatorname{relint} \operatorname{dom} f$ abdecken.

Es sei nun $x \in \operatorname{relint} \operatorname{dom} f$. Nach (16.4) gibt es also ein Element in $\operatorname{relint}(\operatorname{epi} f)$ mit dieser x-Komponente. Mit anderen Worten:

$$\operatorname{relint}(C_1) \cap C_2 = \operatorname{relint}(\operatorname{epi} f) \ \cap\ \{x\} \times \mathbb{R} \neq \emptyset \tag{16.5}$$

mit den Abkürzungen

$$C_1 \coloneqq \operatorname{epi} f \quad \text{und} \quad C_2 \coloneqq \{x\} \times \mathbb{R}.$$

Wir haben:

$$\begin{aligned}
&\emptyset \neq \operatorname{relint}(C_1) \cap C_2 \\
&\quad = \operatorname{relint}(C_1) \cap \operatorname{relint}(C_2) && \text{denn } C_2 = \operatorname{relint}(C_2) \\
&\quad = \operatorname{relint}(C_1 \cap C_2) && \text{nach Lemma 15.23, denn } \operatorname{relint}(C_1) \cap \operatorname{relint}(C_2) \neq \emptyset \\
&\quad = \operatorname{relint}(\{x\} \times [f(x), \infty)) && \text{wegen der Definition von epi } f \\
&\quad = \{x\} \times (f(x), \infty).
\end{aligned}$$

Damit haben wir gezeigt:

$$\begin{pmatrix} x \\ \gamma \end{pmatrix} \in \operatorname{relint}(\operatorname{epi} f) \quad \Leftrightarrow \quad x \in \operatorname{relint} \operatorname{dom} f \quad \text{und} \quad \gamma > f(x). \qquad \square$$

Wir können nun das angekündigte Ergebnis beweisen:

Satz 16.8 (Wann ist das Subdifferential leer bzw. nichtleer? vgl. Rockafellar, 1970, Theorem 23.4)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex.

(*i*) Falls $f \not\equiv \infty$ ist, dann gilt: Für alle $x_0 \notin \operatorname{dom} f$ ist $\partial f(x_0) = \emptyset$.

(*ii*) Für alle $x_0 \in \operatorname{relint}(\operatorname{dom} f)$ ist $\partial f(x_0) \neq \emptyset$.

Beweis. Aussage (*i*): Es sei $x_0 \notin \operatorname{dom} f$, also $f(x_0) = \infty$. Nach Voraussetzung existiert ein $x \in \operatorname{dom} f$. Die Subgradientenungleichung (16.1), also

$$\underbrace{f(x)}_{<\infty} \geqslant \underbrace{f(x_0)}_{=\infty} + s^\mathsf{T}(x - x_0)$$

kann mit diesem x für kein $s \in \mathbb{R}^n$ erfüllt sein.

Aussage (*ii*): Es sei $x_0 \in \operatorname{relint}(\operatorname{dom} f) \subseteq \operatorname{dom} f$. 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

$$\operatorname{epi} f = \left\{ \binom{x}{\gamma} \in \mathbb{R}^n \times \mathbb{R} \,\middle|\, \gamma \geqslant f(x) \right\}.$$

Nach Satz 13.17 ist $\operatorname{epi} f \subseteq \mathbb{R}^n \times \mathbb{R}$ konvex.

Schritt 1: Wir wenden den eigentlichen Trennungssatz 15.32 an, um die Mengen $C_1 = \left\{ \binom{x_0}{f(x_0)} \right\}$ und $C_2 = \operatorname{epi} f$ eigentlich zu trennen.

Das ist möglich aufgrund von $\operatorname{relint}(C_1) \cap \operatorname{relint}(C_2) = \emptyset$, da der Punkt $\binom{x_0}{f(x_0)}$ zu $\operatorname{epi} f$, aber nicht zu $\operatorname{relint}(\operatorname{epi} f)$ gehört, siehe Lemma 16.7.

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 $\binom{x_0}{f(x_0)}$ und $\operatorname{epi} f$ eigentlich trennt. Demnach gilt

$$\binom{s}{\sigma}^\mathsf{T} \binom{x}{\gamma} \leqslant \binom{s}{\sigma}^\mathsf{T} \binom{x_0}{f(x_0)} \quad \text{für alle } \binom{x}{\gamma} \in \operatorname{epi} f, \tag{16.6}$$

und es existiert ein Punkt $\binom{\overline{x}}{\overline{\gamma}} \in \operatorname{epi} f$ mit der Eigenschaft

$$\binom{s}{\sigma}^\mathsf{T} \binom{\overline{x}}{\overline{\gamma}} < \binom{s}{\sigma}^\mathsf{T} \binom{x_0}{f(x_0)}. \tag{16.7}$$

Schritt 2: Wir zeigen $\sigma \leqslant 0$.

Dazu unterscheiden wir drei Fälle:

(*i*) Ist $f(x_0) = 0$, so folgt aus (16.6) mit der Wahl $x = x_0$ und $\gamma = 1$:

$$s^\mathsf{T} x_0 + \sigma \leqslant s^\mathsf{T} x_0 + 0.$$

(*ii*) Ist $f(x_0) > 0$, so folgt aus (16.6) mit der Wahl $x = x_0$ und $\gamma = 2\, f(x_0)$:

$$s^\mathsf{T} x_0 + 2\,\sigma\, f(x_0) \leqslant s^\mathsf{T} x_0 + \sigma\, f(x_0).$$

(*iii*) Ist $f(x_0) < 0$, so folgt aus (16.6) mit der Wahl $x = x_0$ und $\gamma = 0$:

$$s^\mathsf{T} x_0 + 0 \leqslant s^\mathsf{T} x_0 + \sigma\, f(x_0).$$

In allen drei Fällen folgt aus der jeweiligen Ungleichung $\sigma \leqslant 0$.

Schritt 3: Wir zeigen jetzt, dass sogar $\sigma < 0$ gilt, indem wir die Annahme $\sigma = 0$ zum Widerspruch führen. Aus $\sigma = 0$ folgt mit (16.7) $s^\mathsf{T}(\overline{x} - x_0) < 0$. Da $x_0 \in \operatorname{relint}(\operatorname{dom} f)$ war, existiert nach Lemma 15.19 ein $\varepsilon > 0$, sodass auch $\tilde{x} := x_0 - \varepsilon(\overline{x} - x_0)$ noch zu $\operatorname{dom} f$ gehört. Durch Einsetzen von $\begin{pmatrix} \tilde{x} \\ f(\tilde{x}) \end{pmatrix} \in \operatorname{epi} f$ in (16.6) erhalten wir den Widerspruch

$$s^\mathsf{T} x_0 - \varepsilon \underbrace{s^\mathsf{T}(\overline{x} - x_0)}_{<0} \leqslant s^\mathsf{T} x_0.$$

Schritt 4: Durch positive Skalierung des Normalenvektors $a = -(s, \sigma)$ können wir nun o. B. d. A. $\sigma = -1$ annehmen. Dann ergibt sich aus (16.6) mit der Wahl $\gamma = f(x)$ die Folgerung

$$s^\mathsf{T} x - f(x) \leqslant s^\mathsf{T} x_0 - f(x_0) \quad \text{für alle } x \in \operatorname{dom} f.$$

Da dieselbe Ungleichung trivialerweise auch für $x \notin \operatorname{dom} f$ gilt, erhalten wir schließlich

$$f(x) \geqslant f(x_0) + s^\mathsf{T}(x - x_0) \quad \text{für alle } x \in \mathbb{R}^n,$$

d. h., $s \in \partial f(x_0)$. □

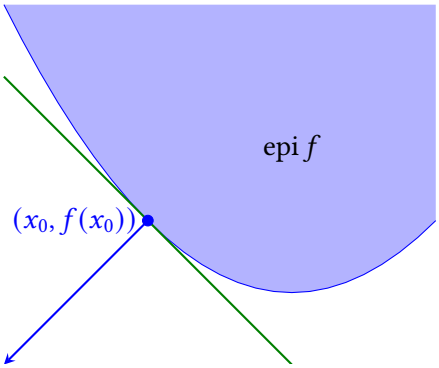


Abbildung 16.3.: Konstruktion eines Subgradienten aus dem Beweis von Satz 16.8 mit Hilfe des Normalenvektors einer Hyperebene, die den Punkt $(x_0, f(x_0))$ eigentlich von $\operatorname{epi} f$ trennt.

Ein wichtiger Satz über das Subdifferential ist die folgende Summenregel, deren Beweis wiederum den eigentlichen Trennungssatz 15.32 verwendet.

Satz 16.9 (Summenregel für das Subdifferential, vgl. Rockafellar, 1970, Theorem 23.8)**.**
Es seien $f^{(1)}, f^{(2)} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ zwei konvexe Funktionen.

(*i*) Es gilt

$$\partial f^{(1)}(x_0) + \partial f^{(2)}(x_0) \subseteq \partial(f^{(1)} + f^{(2)})(x_0) \quad \text{für alle } x_0 \in \mathbb{R}^n. \tag{16.8}$$

(*ii*) Falls

$$(\operatorname{relint} \operatorname{dom} f^{(1)}) \cap (\operatorname{relint} \operatorname{dom} f^{(2)}) \neq \emptyset \tag{16.9}$$

erfüllt ist, dann gilt sogar

$$\partial f^{(1)}(x_0) + \partial f^{(2)}(x_0) = \partial(f^{(1)} + f^{(2)})(x_0) \quad \text{für alle } x_0 \in \mathbb{R}^n. \tag{16.10}$$

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 (16.8) 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

$$\begin{aligned} f^{(1)}(x) &\geqslant f^{(1)}(x_0) + s_1^\mathsf{T}(x - x_0), \\ f^{(2)}(x) &\geqslant f^{(2)}(x_0) + s_2^\mathsf{T}(x - x_0) \end{aligned}$$

für alle $x \in \mathbb{R}^n$. Die Addition der Ungleichungen ergibt (**Quizfrage 16.4:** Warum stimmt das auch dann, wenn hier nicht-endliche Funktionswerte vorkommen?)

$$(f^{(1)} + f^{(2)})(x) \geqslant (f^{(1)} + f^{(2)})(x_0) + (s_1 + s_2)^\mathsf{T}(x - x_0),$$

für alle $x \in \mathbb{R}^n$, d. h., wir haben $s = s_1 + s_2 \in \partial(f^{(1)} + f^{(2)})(x_0)$.

Aussage *(ii)*: Wir nehmen jetzt an, dass $(\operatorname{relint}\operatorname{dom} f^{(1)}) \cap (\operatorname{relint}\operatorname{dom} f^{(2)}) \neq \emptyset$ gilt, und zeigen die zu (16.8) 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 (16.8) 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 16.8 *(i)* folgt $x_0 \in \operatorname{dom}(f^{(1)} + f^{(2)})$, 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 $s_2 \in \partial f^{(2)}(x_0)$.

Schritt 1: Wir setzen

$$\begin{aligned} g_1(x) &:= f^{(1)}(x + x_0) - f^{(1)}(x_0) - s^\mathsf{T}x \\ g_2(x) &:= f^{(2)}(x + x_0) - f^{(2)}(x_0). \end{aligned}$$

Diese Funktionen 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 $(\operatorname{relint}\operatorname{dom} g_1) \cap (\operatorname{relint}\operatorname{dom} g_2) \neq \emptyset$, 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 16.5:** Details?)

Schritt 2: 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 wird unter Anwendung des eigentlichen Trennungssatzes 15.32 erfolgen. Wir betrachten dazu die konvexen Mengen[5]

$$\begin{aligned} C_1 &:= \{\tbinom{x}{\gamma} \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \geqslant g_1(x)\} &&= \operatorname{epi} g_1, \\ C_2 &:= \{\tbinom{x}{\gamma} \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \leqslant -g_2(x)\} &&= \operatorname{hypo} -g_2. \end{aligned}$$

(**Quizfrage 16.6:** Warum sind beide dieser Mengen nichtleer?) Lemma 16.7 zeigt

$$\begin{aligned} \operatorname{relint} C_1 &= \{\tbinom{x}{\gamma} \in \mathbb{R}^n \times \mathbb{R} \mid x \in \operatorname{relint}\operatorname{dom} g_1,\ \gamma > g_1(x)\}, \\ \operatorname{relint} C_2 &= \{\tbinom{x}{\gamma} \in \mathbb{R}^n \times \mathbb{R} \mid x \in \operatorname{relint}\operatorname{dom} g_2,\ \gamma < -g_2(x)\}. \end{aligned}$$

Wegen $0 \in \partial(g_1 + g_2)(\tilde{x}_0)$ ist $\tilde{x}_0 = 0$ nach Satz 16.3 *(ii)* ein globaler Minimierer für $g_1 + g_2$ mit Funktionswert $(g_1 + g_2)(0) = 0$. Daraus folgt, dass $\operatorname{relint} C_1 \cap \operatorname{relint} C_2 = \emptyset$ ist. (**Quizfrage 16.7:** Details?)

[5]Der **Hypograph** einer Funktion $f := \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ ist die Menge

$$\operatorname{hypo} f := \{\tbinom{x}{\gamma} \in \mathbb{R}^n \times \mathbb{R} \mid \gamma \leqslant f(x)\}.$$

Der Hypograph von f ist konvex genau dann, wenn f konkav ist.

Schritt 3: Nach dem eigentlichen Trennungssatz 15.32 lassen sich nun C_1 und C_2 durch eine Hyperebene $H(a,\beta)$ eigentlich trennen. Wir schreiben den Normalenvektor als $a = \left(\begin{smallmatrix}\tilde{s}\\ \sigma\end{smallmatrix}\right) \in \mathbb{R}^n \times \mathbb{R}$. Es gilt also

$$\begin{pmatrix}\tilde{s}\\ \sigma\end{pmatrix}^\mathsf{T}\begin{pmatrix}x_1\\ \gamma_1\end{pmatrix} \leqslant \begin{pmatrix}\tilde{s}\\ \sigma\end{pmatrix}^\mathsf{T}\begin{pmatrix}x_2\\ \gamma_2\end{pmatrix} \quad \text{für alle } \begin{pmatrix}x_1\\ \gamma_1\end{pmatrix} \in C_1 \text{ und } \begin{pmatrix}x_2\\ \gamma_2\end{pmatrix} \in C_2, \tag{16.11}$$

und es existieren $\left(\begin{smallmatrix}\overline{x}_1\\ \overline{\gamma}_1\end{smallmatrix}\right) \in C_1$ und $\left(\begin{smallmatrix}\overline{x}_2\\ \overline{\gamma}_2\end{smallmatrix}\right) \in C_2$, für die die Ungleichung strikt ist.

Die Komponente σ kann nicht gleich Null sein, denn sonst hätten wir $\operatorname{dom} g_1$ und $\operatorname{dom} g_2$ durch eine Hyperebene mit Normalenvektor $\tilde{s}$ eigentlich getrennt, was aufgrund der Voraussetzung $(\operatorname{relint}\operatorname{dom} g_1) \cap (\operatorname{relint}\operatorname{dom} g_2) \neq \emptyset$ und Satz 15.32 aber nicht möglich ist.

Schritt 4: Wegen $(g_1 + g_2)(0) = 0$ liegt $\left(\begin{smallmatrix}0\\ 0\end{smallmatrix}\right) \in C_1 \cap C_2$. Aus (16.11) folgt somit $\sigma \leqslant 0$. Da $\sigma \neq 0$ ist, können wir a so skalieren, dass $\sigma = -1$ wird. Außerdem gilt wegen $\left(\begin{smallmatrix}0\\ 0\end{smallmatrix}\right) \in C_1 \cap C_2$, dass die eigentlich trennende Hyperebene $H(a,\beta)$ den Offset $\beta = 0$ hat. Es folgt nun aus (16.11)

$$\begin{pmatrix}\tilde{s}\\ -1\end{pmatrix}^\mathsf{T}\begin{pmatrix}x_1\\ \gamma_1\end{pmatrix} \leqslant 0 \leqslant \begin{pmatrix}\tilde{s}\\ -1\end{pmatrix}^\mathsf{T}\begin{pmatrix}x_2\\ \gamma_2\end{pmatrix} \quad \text{für alle } \begin{pmatrix}x_1\\ \gamma_1\end{pmatrix} \in C_1 \text{ und } \begin{pmatrix}x_2\\ \gamma_2\end{pmatrix} \in C_2,$$

also

$$\begin{aligned}\tilde{s}^\mathsf{T}x &\leqslant \gamma \quad \text{für alle } \left(\begin{smallmatrix}x\\ \gamma\end{smallmatrix}\right) \in C_1,\\ \tilde{s}^\mathsf{T}x &\geqslant \gamma \quad \text{für alle } \left(\begin{smallmatrix}x\\ \gamma\end{smallmatrix}\right) \in C_2.\end{aligned}$$

Wegen $g_1(0) = g_2(0) = 0$ und der Definition von C_1 und C_2 heißt das aber

$$\begin{aligned}g_1(x) &\geqslant g_1(0) + \tilde{s}^\mathsf{T}(x-0) \quad &&\text{für alle } x \in \mathbb{R}^n,\\ g_2(x) &\geqslant g_2(0) + (-\tilde{s})^\mathsf{T}(x-0) \quad &&\text{für alle } x \in \mathbb{R}^n.\end{aligned}$$

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 Subdifferentiale (16.10) ohne eine Regularitätsbedingung wie $(\operatorname{relint}\operatorname{dom} f^{(1)}) \cap (\operatorname{relint}\operatorname{dom} f^{(2)}) \neq \emptyset$ im Allgemeinen falsch.

Beispiel 16.10 (Phelps, 1993, Bemerkung nach Theorem 3.16)**.**
Es seien $f^{(1)} := I_{C_1}$ und $f^{(2)} := I_{C_2}$ zwei Indikatorfunktionen auf $\mathbb{R}^2$, und zwar für die konvexen Mengen

$$\begin{aligned}C_1 &:= \{(x,\gamma) \mid \gamma \geqslant x^2\} = \operatorname{epi}(x \mapsto x^2)\\ C_2 &:= \{(x_1,x_2) \mid x_2 = 0\}.\end{aligned}$$

Dann ist $\partial f^{(1)}(0,0) = \{(s_1,s_2) \mid s_1 = 0,\ s_2 \leqslant 0\}$ 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 (16.9) ist nicht erfüllt, denn es gilt $\operatorname{dom} f^{(1)} \cap \operatorname{dom} f^{(2)} = C_1 \cap C_2 = \{(0,0)\}$, und dieser Punkt ist kein relativ innerer Punkt von C_1. △

Wir schließen diesen Abschnitt mit der Kettenregel, die wir hier ohne Beweis angeben.

Satz 16.11 (Kettenregel für das Subdifferential, vgl. Rockafellar, 1970, Theorem 23.9[6])**.**
Es sei $f\colon \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ konvex und $A \in \mathbb{R}^{m\times n}$. Weiter sei $g\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ definiert durch $g(x) := f(Ax)$.

[6] 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 (16.12) in diesem Fall $\mathbb{R}^n \subseteq \mathbb{R}^n$ lautet.

(*i*) Es gilt:

$$A^\top \partial f(A\,x_0) \subseteq \partial g(x_0) \quad \text{für alle } x_0 \in \mathbb{R}^n. \tag{16.12}$$

(*ii*) Falls

$$(\operatorname{Bild} A) \cap (\operatorname{relint} \operatorname{dom} f) \neq \emptyset \tag{16.13}$$

erfüllt ist, dann gilt sogar

$$A^\top \partial f(A\,x_0) = \partial g(x_0) \quad \text{für alle } x_0 \in \mathbb{R}^n. \tag{16.14}$$

Weitere Eigenschaften des Subdifferentials folgen in § 16.3.

Ende der Vorlesung 23

Ende der Woche 12

§ 16.2 Die Richtungsableitung

Frage: Gibt es einen Zusammenhang des Subdifferentials mit der Richtungsableitung?

Definition 16.12 ((Einseitige) Richtungsableitung)**.**
Es seien $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ eine (nicht notwendigerweise konvexe) Funktion, $x_0 \in \mathbb{R}^n$ ein Punkt, an dem $f(x_0)$ endlich ist, und $d \in \mathbb{R}^n$. Dann heißt der Grenzwert (sofern er in $\mathbb{R} \cup \{\pm\infty\}$ existiert)

$$f'(x_0; d) := \lim_{t \searrow 0} \frac{f(x_0 + t\,d) - f(x_0)}{t} \tag{16.15}$$

die **(einseitige) Richtungsableitung** der Funktion f im Punkt x_0 in Richtung d. △

Beachte: Da $f(x_0)$ endlich ist, ist der Differenzenquotient ist für alle $t > 0$ definiert und nimmt Werte in $\mathbb{R} \cup \{\pm\infty\}$ an.

Beispiel 16.13 (Beispiele zur Richtungsableitung)**.**
Wir betrachten als Beispiel für f wie in Beispiel 16.5 verschiedene Normen auf $\mathbb{R}^n$. Die jeweiligen Nachweise sind Gegenstand von Hausaufgabe 13.1.

(*i*) $f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$. Dann gilt

$$f'(x; d) = \sum_{\substack{i=1 \\ x_i > 0}}^n d_i - \sum_{\substack{i=1 \\ x_i < 0}}^n d_i + \sum_{\substack{i=1 \\ x_i = 0}}^n |d_i|.$$

(*ii*) $f(x) = \|x\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$. Dann gilt

$$f'(x; d) = \begin{cases} \frac{x^\top d}{\|x\|_2}, & \text{falls } x \neq 0, \\ \|d\|_2, & \text{falls } x = 0. \end{cases}$$

(*iii*) $f(x) = \|x\|_\infty = \max_{i=1,\dots,n} |x_i|$. Dann gilt für $x \neq 0$

$$f'(x;d) = \max\left\{(\operatorname{sgn} x_i)\, d_i \,\middle|\, i = 1,\dots,n,\ |x_i| = \|x\|_\infty\right\}$$

sowie

$$f'(0;d) = \|d\|_\infty. \qquad \triangle$$

Zur weiteren Untersuchung der Richtungsableitung führen wir die Abkürzung

$$q(t) := \frac{f(x_0 + t\,d) - f(x_0)}{t}, \quad t > 0 \tag{16.16}$$

für den Differenzenquotienten ein, wenn die Funktion f, der Punkt x_0 und die Richtung $d \in \mathbb{R}^n$ aus dem Kontext klar sind.

Beispiel 16.14 (Warum einseitige Richtungsableitungen?)**.**
Bei konvexen Funktionen arbeiten wir mit einseitigen Richtungsableitungen

$$\lim_{t \searrow 0} \frac{f(x_0 + t\,d) - f(x_0)}{t}$$

und nicht mit beidseitigen Richtungsableitungen

$$\lim_{t \to 0} \frac{f(x_0 + t\,d) - f(x_0)}{t},$$

weil letztere für viele Funktionen nicht existieren. Bereits die konvexe Funktion $f(x) = |x|$ mit $x \in \mathbb{R}$ an der Stelle $x_0 = 0$ ist ein Beispiel dafür. $\triangle$

Lemma 16.15 (Monotonie des Differenzenquotienten, vgl. Rockafellar, 1970, Theorem 23.1)**.**
Es seien $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex, $x_0 \in \mathbb{R}^n$ ein Punkt, an dem $f(x_0)$ endlich ist, und $d \in \mathbb{R}^n$. Dann ist $q(t)$ definiert und monoton wachsend auf $\mathbb{R}_{>0}$.

Beweis. Es sei $0 < t_1 < t_2$. Aus

$$x_0 + t_1\,d = \frac{t_1}{t_2}(x_0 + t_2\,d) + \left(1 - \frac{t_1}{t_2}\right) x_0$$

und der Konvexität von f folgt

$$f(x_0 + t_1\,d) = f\left(\frac{t_1}{t_2}(x_0 + t_2\,d) + \left(1 - \frac{t_1}{t_2}\right) x_0\right) \leqslant \frac{t_1}{t_2} f(x_0 + t_2\,d) + \left(1 - \frac{t_1}{t_2}\right) f(x_0).$$

Da $f(x_0)$ endlich ist, ist die rechte Seite definiert. Aus der obigen Ungleichung folgt weiter

$$f(x_0 + t_1\,d) - f(x_0) \leqslant \frac{t_1}{t_2} f(x_0 + t_2\,d) - \frac{t_1}{t_2} f(x_0)$$

und damit $q(t_1) \leqslant q(t_2)$. $\square$

Quizfrage 16.8: An welcher Stelle im Beweis geht die Voraussetzung ein, dass $f(x_0)$ endlich ist?

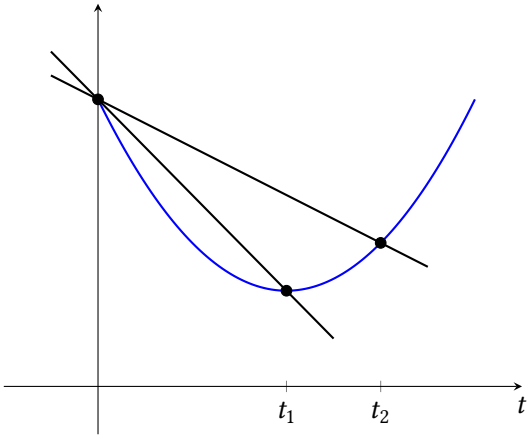


Abbildung 16.4.: Illustration der Monotonie des Differenzenquotienten q (Lemma 16.15).

Satz 16.16 (Existenz und elementare Eigenschaften der Richtungsableitung, vgl. Rockafellar, 1970, Theorem 23.1)**.**
Es seien $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex und $x_0 \in \mathbb{R}^n$ ein Punkt, an dem $f(x_0)$ endlich ist. Dann existiert die Richtungsableitung $f'(x_0; d)$ mit Werten in $\mathbb{R} \cup \{\pm\infty\}$ für alle Richtungen $d \in \mathbb{R}^n$. Es gilt

$$f(x) \geqslant f(x_0) + f'(x_0; x - x_0) \quad \text{für alle } x \in \mathbb{R}^n. \tag{16.17}$$

Die Richtungsableitung hat folgende Eigenschaften:

(*i*) Die Funktion

$$\mathbb{R}^n \ni d \mapsto f'(x_0; d) \in \mathbb{R} \cup \{\pm\infty\} \tag{16.18}$$

ist positiv homogen und konvex.

(*ii*) Die Funktion (16.18) ist **subadditiv** (englisch: *subadditive*):

$$f'(x_0; d_1 + d_2) \leqslant f'(x_0; d_1) + f'(x_0; d_2) \tag{16.19}$$

für alle $d_1, d_2 \in \mathbb{R}^n$, für die die rechte Seite definiert ist.

(*iii*) Es gilt $f'(x_0; 0) = 0$.

(*iv*) Es gilt

$$-f'(x_0; -d) \leqslant f'(x_0; d) \quad \text{für alle } d \in \mathbb{R}^n.$$

(*v*) Es sei U der Richtungsraum von aff dom f und $x_0 \in \operatorname{relint} \operatorname{dom} f$. Dann ist $f'(x_0; d) \in \mathbb{R}$ (also die Richtungsableitung endlich) für alle $d \in U$ und $f'(x_0; d) = \infty$ für alle $d \notin U$.

Beweis. Aus der Monotonie von q (Lemma 16.15) folgt

$$f'(x_0; d) = \lim_{t \searrow 0} q(t) = \inf_{t > 0} q(t) \in \mathbb{R} \cup \{\pm\infty\}.$$

Mit $x \in \mathbb{R}^n$ beliebig und $d := x - x_0$ folgt außerdem aus der Monotonie von q

$$f'(x_0; d) = \lim_{t \searrow 0} q(t) \leqslant q(1) = \frac{f(x_0 + 1\,d) - f(x_0)}{1} = f(x) - f(x_0)$$

und damit die Ungleichung (16.17).

Aussage (*i*): Für $\alpha > 0$ und $d \in \mathbb{R}^n$ gilt

$$f'(x_0; \alpha\, d) = \lim_{t \searrow 0} \frac{f(x_0 + t\,\alpha\, d) - f(x_0)}{t} = \alpha \lim_{t \searrow 0} \frac{f(x_0 + t\,\alpha\, d) - f(x_0)}{\alpha\, t} = \alpha\, f'(x_0; d),$$

d. h., $d \mapsto f'(x_0; d)$ ist positiv homogen. Um die Konvexität zu zeigen, seien $d_1, d_2 \in \mathbb{R}^n$ und $\alpha \in [0, 1]$. Wir müssen zeigen (vgl. (13.8)):

$$f'(x_0; \alpha\, d_1 + (1-\alpha)\, d_2) \leqslant \alpha\, f'(x_0; d_1) + (1-\alpha)\, f'(x_0; d_2), \tag{16.20}$$

sofern die rechte Seite definiert ist, wovon wir ab sofort ausgehen. Es gibt folgende Fälle.

Fall 1: $f'(x_0; d_1)$ und $f'(x_0; d_2)$ sind beide endlich. Aufgrund der Monotonie des Differenzenquotienten müssen dann auch $f(x_0 + t\, d_1)$ und $f(x_0 + t\, d_2)$ für $t \in (0, t_0)$ endlich sein. Für diese Werte von t folgt

$$\begin{aligned} &f\big(x_0 + t\,(\alpha\, d_1 + (1-\alpha)\, d_2)\big) - f(x_0) \\ &\quad = f\big(\alpha\,(x_0 + t\, d_1) + (1-\alpha)\,(x_0 + t\, d_2)\big) - f(x_0) \\ &\quad \leqslant \alpha\,\big[f(x_0 + t\, d_1) - f(x_0)\big] + (1-\alpha)\big[f(x_0 + t\, d_2) - f(x_0)\big] \end{aligned} \tag{16.21}$$

aus der Konvexität von f. Die Division durch $t > 0$ und der (monotone) Grenzübergang $\lim_{t \searrow 0}$ auf der linken Seite zeigen weiter

$$\begin{aligned} &f'(x_0; \alpha\, d_1 + (1-\alpha)\, d_2) \\ &\quad \leqslant \frac{f\big(x_0 + t\,(\alpha\, d_1 + (1-\alpha)\, d_2)\big) - f(x_0)}{t} \\ &\quad \leqslant \alpha\, \frac{f(x_0 + t\, d_1) - f(x_0)}{t} + (1-\alpha)\, \frac{f(x_0 + t\, d_2) - f(x_0)}{t} \end{aligned} \tag{16.22}$$

für $t \in (0, t_0)$. Durch Grenzübergang auf der rechten Seite der Ungleichung folgt nun die gewünschte Ungleichung (16.20).

Fall 2: Dasselbe Argument kann auf den Fall erweitert werden, dass $f'(x_0; d_1)$ und $f'(x_0; d_2)$ endlich oder ∞ sind. Dann sind die Werte von $f(x_0 + t\, d_1)$ und $f(x_0 + t\, d_2)$ für $t \in (0, t_0)$ endlich oder ∞.

Fall 3: Es verbleibt der Fall, in dem $f'(x_0; d_1) = -\infty$ und $f'(x_0; d_2)$ endlich oder ebenfalls $-\infty$ ist. (Der Beweis im umgekehrten Fall geht analog.) In diesem Fall sind die Werte von $f(x_0 + t\, d_1)$ und $f(x_0 + t\, d_2)$ für $t \in (0, t_0)$ endlich oder $-\infty$. Auch in diesem Fall gelten (16.21) und (16.22) für ein Intervall $(0, t_0)$, und der Grenzübergang zeigt (16.20).

Aussage (*ii*): Aus der positiven Homogenität und der Konvexität von $f'(x_0; \cdot)$ folgt

$$\begin{aligned} &f'(x_0; d_1 + d_2) \\ &\quad = 2\, f'(x_0; \tfrac{1}{2}(d_1 + d_2)) \qquad \text{(positive Homogenität)} \\ &\quad \leqslant 2\, \frac{1}{2} f'(x_0; d_1) + 2\, \frac{1}{2} f'(x_0; d_2), \end{aligned}$$

sofern die rechte Seite definiert ist.

Aussage (*iii*): Für $d = 0$ gilt

$$f'(x_0; d) = \lim_{t \searrow 0} \frac{f(x_0 + t\,0) - f(x_0)}{t} = 0.$$

Aussage (*iv*): Aus Aussage (*iii*) und Aussage (*ii*) folgt

$$0 = f'(x_0; d - d) \leqslant f'(x_0; d) + f'(x_0; -d),$$

sofern die rechte Seite definiert ist. Wenn das der Fall ist, dann sieht man die behauptete Ungleichung $-f'(x_0; -d) \leqslant f'(x_0; d)$ leicht ein, und zwar auch dann, wenn $f'(x_0; -d)$ oder $f'(x_0; d)$ oder beide gleich ∞ sind. Es verbleiben die Fälle $f'(x_0; d) = \infty$ und $f'(x_0; -d) = -\infty$ bzw. $f'(x_0; d) = -\infty$ und $f'(x_0; -d) = \infty$, für die ebenfalls die Ungleichung $-f'(x_0; -d) \leqslant f'(x_0; d)$ gilt.

Aussage (*v*): Es sei $x_0 \in \operatorname{relint} \operatorname{dom} f$ und $d \in U$, dem Richtungsraum von $\operatorname{aff} \operatorname{dom} f$. Dann ist $x_0 \pm \varepsilon\, d \in \operatorname{dom} f$ für hinreichend kleines $\varepsilon > 0$, siehe Lemma 15.19. Folglich sind die Ausdrücke

$$\frac{f(x_0 + t\,d) - f(x_0)}{t} \quad \text{und} \quad \frac{f(x_0 - t\,d) - f(x_0)}{t}$$

für hinreichend kleine $t > 0$ endlich, und aufgrund der Monotonie des Differenzenquotienten folgt $-\infty \leqslant f'(x_0; d) < \infty$ und $-\infty \leqslant f'(x_0; -d) < \infty$. Zusammen mit Aussage (*iv*) ergibt sich also

$$-\infty < -f'(x_0; -d) \leqslant f'(x_0; d) < \infty,$$

d. h. $f'(x_0; d)$ ist endlich.

Andererseits gehört, falls $d \notin U$ liegt, $x_0 + t\,d$ für *alle* $t > 0$ *nicht* zu $\operatorname{aff} \operatorname{dom} f$ und damit auch nicht zu $\operatorname{dom} f$. In diesem Fall ist also $f(x_0 + t\,d) = \infty$ für alle $t > 0$ und damit $q(t) \equiv \infty$, folglich $f'(x_0; d) = \infty$. □

Folgerung 16.17 (Endlichkeit der Richtungsableitung)**.**
Es seien $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex und $x_0 \in \mathbb{R}^n$ ein Punkt, an dem $f(x_0)$ endlich ist. Ist $x_0 \in \operatorname{int} \operatorname{dom} f$, dann ist die Richtungsableitung $f'(x_0; d)$ für alle $d \in \mathbb{R}^n$ endlich.

Beweis. Da $\operatorname{int} \operatorname{dom} f$ nichtleer ist, gilt nach Folgerung 15.17 $\dim \operatorname{dom} f = n$. Der Richtungsraum von $\operatorname{aff} \operatorname{dom} f$ ist damit $U = \mathbb{R}^n$. Die Behauptung folgt nun aus Satz 16.16 (*v*). □

§ 16.3 Zusammenhang zwischen Subdifferential und Richtungsableitung

Das wesentliche Resultat, in dem sich Subdifferential und Richtungsableitung einer konvexen Funktion gegenseitig charakterisieren, ist der folgende Satz.

Satz 16.18 (Zusammenhang zwischen Subdifferential und Richtungsableitung, vgl. Rockafellar, 1970, Theorem 23.2 und 23.4)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex.

(*i*) Für jedes $x_0 \in \mathbb{R}^n$, an dem $f(x_0)$ endlich ist, gilt:

$$\partial f(x_0) = \left\{ s \in \mathbb{R}^n \,\middle|\, s^\mathsf{T} d \leqslant f'(x_0; d) \text{ für alle } d \in \mathbb{R}^n \right\}. \tag{16.23}$$

(Möglicherweise sind beide Mengen leer.)

(*ii*) Für jedes $x_0 \in \operatorname{relint} \operatorname{dom} f$, an dem $f(x_0)$ endlich ist, und $d \in \mathbb{R}^n$ gilt:

$$f'(x_0; d) = \sup\left\{ s^\mathsf{T} d \,\middle|\, s \in \partial f(x_0) \right\} \in \mathbb{R} \cup \{\infty\}. \tag{16.24}$$

Genauer: Es sei U der Richtungsraum von $\operatorname{aff} \operatorname{dom} f$. Ist $d \in U$, dann sind beide Seiten in (16.24) endlich, und das Supremum ist ein Maximum. Ist $d \notin U$, dann sind beide Seiten in (16.24) gleich ∞.

Die Aussage (16.24) ist in der Literatur auch als **max formula** bekannt.

Beweis. Aussage (*i*): Es sei zunächst $s \in \partial f(x_0)$ und $d \in \mathbb{R}^n$ beliebig, aber fest. Aus der Subgradientenungleichung (16.1) folgt

$$f(x_0 + t\,d) \geqslant f(x_0) + s^\mathsf{T}(t\,d)$$

für alle $t > 0$. Da $f(x_0)$ nach Voraussetzung endlich ist, ergibt das Sortieren der Terme und Division durch t

$$\frac{f(x_0 + t\,d) - f(x_0)}{t} \geqslant s^\mathsf{T} d,$$

und durch Grenzübergang folgt $f'(x_0; d) \geqslant s^\mathsf{T} d$.

Nun sei $s \in \mathbb{R}^n$ ein Vektor mit der Eigenschaft $s^\mathsf{T} d \leqslant f'(x_0; d)$ für alle $d \in \mathbb{R}^n$. Weiter sei $x \in \mathbb{R}^n$ beliebig und $d := x - x_0$. Aus der Monotonie des Differenzenquotienten (Lemma 16.15) folgt insbesondere

$$s^\mathsf{T}(x - x_0) = s^\mathsf{T} d \leqslant f'(x_0; d) = \lim_{t \searrow 0} q(t) \leqslant q(1) = \frac{f(x_0 + 1\,d) - f(x_0)}{1},$$

d. h., es gilt

$$f(x) \geqslant f(x_0) + s^\mathsf{T}(x - x_0).$$

Da $x \in \mathbb{R}^n$ beliebig war, folgt $s \in \partial f(x_0)$.

Aussage (*ii*): Es sei $x_0 \in \operatorname{relint} \operatorname{dom} f$. Dann ist $\partial f(x_0) \neq \emptyset$ nach Satz 16.8. Es sei $d \in \mathbb{R}^n$ beliebig, aber fest. Die Richtungsableitung $f'(x_0; d)$ existiert (mit Werten in $\mathbb{R} \cup \{\pm\infty\}$) nach Satz 16.16. Nach Aussage (*i*) gilt $f'(x_0; d) \geqslant s^\mathsf{T} d$ für alle $s \in \partial f(x_0)$, also auch

$$f'(x_0; d) \geqslant \sup\left\{ s^\mathsf{T} d \,\middle|\, s \in \partial f(x_0) \right\}. \tag{16.25}$$

Die rechte Seite kann nicht gleich $-\infty$ sein. (**Quizfrage 16.9:** Warum?)

Wir müssen zeigen, dass in (16.25) Gleichheit gilt. Es sei dazu U der Richtungsraum von $\operatorname{aff} \operatorname{dom} f$. Wir unterscheiden zwei Fälle. Falls $d \notin U$ liegt, dann ist $f'(x_0; d) = \infty$ nach Satz 16.16 (*v*). Wir müssen zeigen, dass dann auch die rechte Seite in (16.25) gleich ∞ ist. Wir können d eindeutig darstellen als $d = d_1 + d_2$ mit $d_1 \in U$ und $d_2 \in U^\perp$. Nach Annahme ist $d_2 \neq 0$. Es sei $\bar{s}$ irgendein festes Element von $\partial f(x_0)$; es gilt nach Aussage (*i*) also $\bar{s}^\mathsf{T} y \leqslant f'(x_0; y)$ für alle $y \in \mathbb{R}^n$. Wir zeigen, dass $\bar{s} + \alpha\, d_2$ für alle $\alpha \in \mathbb{R}$ in $\partial f(x_0)$ liegt. Dazu zeigen wir:

$$(\bar{s} + \alpha\, d_2)^\mathsf{T} y \leqslant f'(x_0; y) \tag{16.26}$$

für alle $y \in \mathbb{R}^n$. Falls $y \notin U$ liegt, dann ist die rechte Seite in (16.26) gleich ∞, die linke Seite aber endlich. Falls dagegen $y \in U$ liegt, dann ist $d_2^\mathsf{T} y = 0$, also gilt (16.26) auch in diesem Fall. Für die rechte Seite in (16.25) erhalten wir nun

$$\sup\{s^\mathsf{T} d \,|\, s \in \partial f(x_0)\} \geqslant \sup\{(\bar{s} + \alpha\, d_2)^\mathsf{T} d \,|\, \alpha \in \mathbb{R}\} = \bar{s}^\mathsf{T} d + \sup\{\alpha\, \|d_2\|^2 \,|\, \alpha \in \mathbb{R}\} = \infty.$$

Damit ist in diesem Fall die Gleichheit in (16.25) gezeigt.

Im anderen Fall ist $d \in U$. Wir definieren die Funktion $g(d) := f'(x_0; d)$. Diese ist nach Satz 16.16 *(i)* konvex mit Werten in $\mathbb{R} \cup \{\infty\}$, da der Wert $-\infty$ oben durch (16.25) bereits ausgeschlossen wurde. Ihr eigentlicher Definitionsbereich ist $\operatorname{dom} g = U$, siehe Satz 16.16 *(v)*. Da U ein Unterraum von $\mathbb{R}^n$ ist, gilt $\operatorname{relint} U = U$. Nach Satz 16.8 *(ii)* existiert also ein Element $\bar{s} \in \partial g(d)$, d. h., es gilt

$$f'(x_0; y) \geqslant f'(x_0; d) + \bar{s}^\mathsf{T}(y - d) \quad \text{für alle } y \in \mathbb{R}^n. \tag{16.27}$$

Setzen wir speziell $y = 0$ ein, so folgt

$$0 \geqslant f'(x_0; d) - \bar{s}^\mathsf{T} d.$$

Setzen wir andererseits $y = 2\,d$ ein, so ergibt sich

$$f'(x_0; 2\,d) = 2\, f'(x_0; d) \geqslant f'(x_0; d) + \bar{s}^\mathsf{T}(2d - d),$$

d. h. $f'(x_0; d) \geqslant \bar{s}^\mathsf{T} d$. Durch beide Ungleichungen zusammen erhalten wir also $f'(x_0; d) = \bar{s}^\mathsf{T} d$. Es bleibt zu zeigen, dass $\bar{s}$ zu $\partial f(x_0)$ gehört. Aus (16.27) folgt nun aber $f'(x_0; y) \geqslant f'(x_0; d) + \bar{s}^\mathsf{T}(y - d) = \bar{s}^\mathsf{T} y$ für alle $y \in \mathbb{R}^n$. Nach Aussage *(i)* gilt damit tatsächlich $\bar{s} \in \partial f(x_0)$, und wir haben gezeigt:

$$f'(x_0; d) \geqslant \sup\{s^\mathsf{T} d \,|\, s \in \partial f(x_0)\} \geqslant \bar{s}^\mathsf{T} d = f'(x_0; d).$$

Insbesondere ist das Supremum ein Maximum. □

Expertenwissen: Zusammenhang des Subdifferentials mit der Richtungsableitung

Per Definition verwendet das Subdifferential $\partial f(x)$ Funktionswerte in ganz $\mathbb{R}^n$, siehe (16.1). Die Richtungsableitung (16.15) hingegen verwendet nur Funktionswerte in einer Umgebung von x. Es ist daher durchaus bemerkenswert, dass es überhaupt gelingt, diese beiden Konzepte zusammenzubringen. Das geht deshalb, weil man auch beim Subdifferential in Wirklichkeit mit Funktionswerten in einer Umgebung von x auskommt, wie wir (16.2) gesehen haben.

Folgerung 16.19 (Unbeschränktheit des Subdifferentials)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex. Weiter sei $f(x_0)$ endlich und $\partial f(x_0) \neq \emptyset$. Dann gehört mit jedem $s \in \partial f(x_0)$ auch $s + U^\perp$ zu $\partial f(x_0)$, wobei U der Richtungsraum von $\operatorname{aff} \operatorname{dom} f$ und $U^\perp$ sein orthogonales Komplement (bzgl. des Euklidischen Innenprodukts) ist. Wenn also $\dim \operatorname{dom} f < n$ ist, dann ist $\partial f(x_0)$ unbeschränkt. (**Quizfrage 16.10:** Wie kann man sich diese Tatsache grafisch vorstellen?)

Beweis. Es sei $s \in \partial f(x_0)$ und $\overline{d} \in U^\perp$. Nach Satz 16.18 Aussage *(i)* gilt $s^\mathsf{T} d \leqslant f'(x_0; d)$ für alle $d \in \mathbb{R}^n$. Wir zeigen, dass diese Ungleichung auch für $s + \overline{d}$ gilt. Es sei dazu $d \in \mathbb{R}^n$ beliebig. Wir können d

eindeutig darstellen als $d = d_1 + d_2$ mit $d_1 \in U$ und $d_2 \in U^\perp$. Falls $d_2 \neq 0$ gilt, dann ist gehört $x_0 + t\,\mathrm{d}$ für alle $t \neq 0$ nicht zu aff dom f und erst recht nicht zu dom f, also folgt $f'(x_0; d) = \infty$, womit

$$(s + \overline{d})^\mathsf{T} d \leqslant f'(x_0; d)$$

gezeigt ist. Andernfalls ist $d_2 = 0$, also $d \in U$, und daher gilt

$$(s + \overline{d})^\mathsf{T} d = s^\mathsf{T} d \leqslant f'(x_0; d)$$

nach Voraussetzung. Aus Satz 16.18 Aussage (i) folgt nun, dass auch $s + \overline{d}$ zu $\partial f(x_0)$ gehört. □

Beispiel 16.20 (Unbeschränktes Subdifferential)**.**

(i) Wir betrachten die Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x_1, x_2) = |x_1|.$$

Ihr Subdifferential im Ursprung ist beschränkt:

$$\partial f(0, 0) = [-1, 1] \times \{0\}.$$

(ii) Wir betrachten die Funktion $g\colon \mathbb{R}^2 \to \mathbb{R} \cup \{\infty\}$ mit

$$g(x_1, x_2) = \begin{cases} |x_1| & \text{falls } x_2 = 0 \\ \infty & \text{sonst.} \end{cases}$$

Ihr Subdifferential im Ursprung ist unbeschränkt:

$$\partial g(0, 0) = [-1, 1] \times \mathbb{R}.$$

Dies illustriert die Aussage aus Folgerung 16.19. Es ist dom $g = \mathbb{R} \times \{0\}$ mit Richtungsraum $U = \mathbb{R} \times \{0\}$ und orthogonalem Komplement $U^\perp = \{0\} \times \mathbb{R}$. △

Ende der Vorlesung 24

§ 16.4 Weitere Eigenschaften konvexer Funktionen

Lemma 16.21 (Lokale Beschränktheit nach oben impliziert lokale Lipschitz-Stetigkeit)**.**

Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und $x_0 \in \operatorname{dom} f$. Wenn f auf einer Kugel $B_r(x_0)$ nach oben beschränkt ist, dann gilt:

(i) f ist auf $B_r(x_0)$ auch nach unten beschränkt.

(ii) f ist auf $B_{r/2}(x_0)$ Lipschitz-stetig.

Beweis. Aussage (*i*): Es gelte $f \leqslant \widetilde{M}$ auf der Kugel $B_r(x_0)$ für ein $\widetilde{M} \in \mathbb{R}$. Es sei $x \in B_r(x_0)$ ein beliebiges Element und $y := 2\,x_0 - x = x_0 - (x - x_0)$ die Reflexion von x am Mittelpunkt x_0. Dann liegt y ebenfalls in $B_r(x_0)$, und es gilt aufgrund der Konvexität von f

$$f(x_0) = f\left(\frac{x+y}{2}\right) \leqslant \frac{1}{2}f(x) + \frac{1}{2}f(y).$$

(Der Fall $\infty - \infty$ auf der rechten Seite kann wegen $f \leqslant \widetilde{M}$ nicht auftreten.) Durch Umstellen ergibt sich $f(x) \geqslant 2\,f(x_0) - f(y) \geqslant 2\,f(x_0) - \widetilde{M}$, die gesuchte endliche untere Schranke.

Aussage (*ii*): Aufgrund von Aussage (*i*) ist

$$M := \sup\bigl\{|f(x)| \,\big|\, x \in B_r(x_0)\bigr\}$$

endlich. Es seien $x, y \in B_{r/2}(x_0)$ und $x \neq y$. Wir definieren $z := x + \frac{r}{2}\,\frac{(x-y)}{\|x-y\|}$. Dann ist $z \in x + \frac{r}{2}B_1(0) \subseteq x_0 + r\,B_1(0) = B_r(x_0)$. Wir setzen $\alpha := \frac{r}{2\,\|x-y\|}$. Also ist $z = x + \alpha\,(x-y)$ und daher[7]

$$x = \frac{1}{\alpha+1}z + \frac{\alpha}{\alpha+1}y.$$

Die Konvexität von f ergibt

$$f(x) \leqslant \frac{1}{\alpha+1}f(z) + \frac{\alpha}{\alpha+1}f(y)$$

und damit

$$f(x) - f(y) \leqslant \frac{1}{\alpha+1}\bigl[f(z) - f(y)\bigr] \leqslant \frac{2\,M}{\alpha+1} = 2\,M\,\frac{2\,\|x-y\|}{r + 2\,\|x-y\|} \leqslant \frac{4\,M}{r}\,\|x-y\|.$$

Durch Vertauschen den Rollen von x und y erhalten wir ganz analog $f(y) - f(x) \leqslant \frac{4\,M}{r}\,\|x-y\|$, also zusammen

$$|f(x) - f(y)| \leqslant \frac{4\,M}{r}\,\|x-y\| \tag{16.28}$$

für alle $x, y \in B_{r/2}(x_0)$. □

Beachte: Es sieht hier zunächst so aus, als würde die Lipschitz-Konstante in (16.28) für $r \searrow 0$ „explodieren“. Jedoch gilt natürlich eine Abschätzung (16.28) mit derselben Lipschitz-Konstanten auch auf allen Kugeln um x_0 mit kleinerem Radius.

Es stellt sich die Frage, in welchen Punkten $x_0 \in \operatorname{dom} f$ die Voraussetzung von Lemma 16.21 gilt, also dass $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ in einer Umgebung von x_0 nach oben beschränkt (und damit bereits in einer kleineren Umgebung von x_0 Lipschitz-stetig) ist. Dazu muss natürlich notwendigerweise $x_0 \in \operatorname{int}\operatorname{dom} f$ sein. Das ist aber auch bereits hinreichend, wie das folgende Resultat zeigt.

Satz 16.22 (Stetigkeit konvexer Funktionen)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex. Folgende Aussagen sind für $x_0 \in \mathbb{R}^n$ äquivalent:

(*i*) $x_0 \in \operatorname{int}\operatorname{dom} f$.

(*ii*) f ist in einer Umgebung von x_0 nach oben beschränkt.

[7] Es gilt $\alpha \in (1/2, \infty)$, wobei α groß ist für $y \approx x$ und $\alpha \approx 1/2$, falls $\|y - x\| \approx r$. Damit liegt $1/(\alpha+1)$ in $(0, 2/3)$ und $\alpha/(\alpha+1)$ in $(1/3, 1)$.

(*iii*) f ist in einer Umgebung von x_0 Lipschitz-stetig.

(*iv*) f ist im Punkt x_0 stetig.

Beachte: Konvexe *eigentliche* Funktionen sind also genau im (möglicherweise leeren) Inneren ihres eigentliches Definitionsbereiches **lokal beschränkt** und **lokal Lipschitz-stetig**.[8]

Beweis. Aussage (*ii*) ⇒ Aussage (*iii*): Das wurde in Lemma 16.21 gezeigt.

Aussage (*iii*) ⇒ Aussage (*iv*): Das ist offensichtlich.

Aussage (*iv*) ⇒ Aussage (*i*): Das folgt sofort aus der ε-δ-Definition der Stetigkeit.

Aussage (*i*) ⇒ Aussage (*ii*): Es sei $x_0 \in \operatorname{int}\operatorname{dom} f$. Wir können affin unabhängige Vektoren $v^{(0)}, \dots, v^{(n)}$ und $r > 0$ so wählen, dass mit $\Delta := \operatorname{conv}(\{v^{(0)}, \dots, v^{(n)}\})$ gilt: $B_r(x_0) \subseteq \Delta \subseteq \operatorname{dom} f$. (Zur Konstruktion siehe nachfolgendes Lemma 16.23.) Jeder Punkt $x \in B_r(x_0) \subseteq \Delta$ hat dann eine eindeutige Darstellung als Konvexkombination

$$x = \sum_{j=0}^{n} \alpha^{(j)}\, v^{(j)}$$

mit Koeffizienten $\alpha^{(j)} \geqslant 0$ und $\sum_{j=0}^{n} \alpha^{(j)} = 1$. Aus der Jensenschen Ungleichung (Hausaufgabe 9.5) für die konvexe Funktion f folgt nun

$$f(x) = f\Big(\sum_{j=0}^{n} \alpha^{(j)}\, v^{(j)}\Big) \leqslant \sum_{j=0}^{n} \alpha^{(j)} f(v^{(j)}) \leqslant \underbrace{\max_{j=0,\dots,n} f(v^{(j)})}_{=:M} \sum_{j=0}^{n} \alpha^{(j)} = M$$

für alle $x \in \Delta$ und insbesondere für alle $x \in B_r(x_0)$. □

Lemma 16.23 (Zwischen eine Menge und einen inneren Punkt passt immer ein Simplex[9])**.**

Es sei $M \subseteq \mathbb{R}^n$ eine Menge und $x_0 \in \operatorname{int} M$, sodass $B_R(x_0) \subseteq M$ liegt. Dann existieren affin unabhängige Punkte $v^{(0)}, \dots, v^{(n)} \in M$ und ein $r > 0$, sodass mit der konvexen Hülle $\Delta := \operatorname{conv}(\{v^{(0)}, \dots, v^{(n)}\})$ gilt:

$$B_r(x_0) \subseteq \Delta \subseteq B_R(x_0) \subseteq M.$$

Quizfrage 16.11: Wie kann diese Aussage veranschaulicht werden?

[8] Eine Funktion $f\colon M \to \mathbb{R} \cup \{\pm\infty\}$ heißt **lokal beschränkt** (englisch: *locally bounded*), wenn zu jedem Punkt $x_0 \in M$ eine Umgebung $U(x_0)$ und eine Konstante $C(x_0) \in \mathbb{R}$ existieren, sodass $|f(x)| \leqslant C(x_0)$ für alle $x \in U(x_0) \cap M$ gilt. Eine Funktion $f\colon M \to \mathbb{R} \cup \{\pm\infty\}$ heißt **lokal Lipschitz-stetig** (englisch: *locally Lipschitz continuous*), wenn zu jedem Punkt $x_0 \in M$ eine Umgebung $U(x_0)$ und eine Konstante $L(x_0) \in \mathbb{R}$ existieren, sodass $|f(x) - f(y)| \leqslant L(x_0)\, \|x - y\|$ für alle $x \in U(x_0) \cap M$ gilt. Man kann beide Begriffe auf Teilmengen $N \subseteq M$ einschränken und spricht dann von **lokaler Beschränktheit auf N** bzw.**lokaler Lipschitz-Stetigkeit auf N**.

[9] Ein **Simplex** (Plural: **Simplizes**, englisch: *simplex, simplices*) im $\mathbb{R}^n$ ist die konvexe Hülle affin unabhängiger Punkte im $\mathbb{R}^n$.

Beweis. Wir gehen zur Vereinfachung der Notation von $x_0 = 0$ und $R = 1$ aus, was wir immer erreichen können, indem wir M durch $(M - x_0)/R$ ersetzen.

Die gesuchten Punkte können wir nun beispielsweise wählen als

$$v^{(0)} := -\frac{1}{n+1}\mathbf{1} \quad \text{und} \quad v^{(j)} := e^{(j)} - \frac{1}{n+1}\mathbf{1}, \quad j = 1, \dots, n.$$

Hierbei ist $e^{(j)}$ der j-te Einheitsvektor im $\mathbb{R}^n$. Die affine Unabhängigkeit der $v^{(j)}$ sowie die Eigenschaft $\|v^{(j)}\| < 1$ für $j = 0, \dots, n$ prüft man leicht nach. Da $B_1(0)$ konvex ist, folgt bereits

$$\Delta \subseteq B_1(0) \subseteq M.$$

Aufgrund der affinen Unabhängigkeit lässt sich jedes $x \in \mathbb{R}^n$ eindeutig als Affinkombination $x = \sum_{j=0}^n \alpha^{(j)} v^{(j)}$ schreiben mit Koeffizienten, die $\sum_{j=0}^n \alpha^{(j)} = 1$ erfüllen. Der Ursprung ist gerade der Mittelpunkt von $v^{(0)}, \dots, v^{(n)}$, hat also den Koeffizientenvektor $\overline{\alpha} = 1/(n+1)\,\mathbf{1}$, denn es gilt

$$\sum_{j=0}^n \overline{\alpha}_j v^{(j)} = \frac{1}{n+1}\sum_{j=0}^n v^{(j)} = \frac{1}{n+1}\Big(-\frac{n+1}{n+1}\mathbf{1} + \sum_{j=1}^n e^{(j)}\Big) = \frac{1}{n+1}\Big(-\frac{n+1}{n+1}\mathbf{1} + \mathbf{1}\Big) = 0.$$

Insbesondere ist der Ursprung also eine echte Konvexkombination der Punkte $v^{(0)}, \dots, v^{(n)}$. Aufgrund der stetigen (linearen) Abhängigkeit der Koeffizienten α vom Punkt x gilt für eine ganze Umgebung $B_r(0)$ mit geeignetem Radius $r > 0$ die Eigenschaft $\alpha \geqslant 0$, d. h., $B_r(0) \subseteq \Delta$, was zu zeigen war.

Genauer: Wir können wie im Beweis von Satz 15.16 vorgehen. Das lineare Gleichungssystem, das die Koeffizienten der Darstellung eines beliebigen Punktes x als Affinkombination der Punkte $v^{(0)}, \dots, v^{(n)}$ ermittelt, vgl. (15.9), lautet

$$\underbrace{\begin{bmatrix} 1 & \cdots & 1 \\ | & & | \\ v^{(0)} & \cdots & v^{(n)} \\ | & & | \end{bmatrix}}_{=:B} \begin{pmatrix} \alpha^{(0)} \\ \vdots \\ \vdots \\ \alpha^{(n)} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ | \\ x \\ | \end{pmatrix}}_{=:b}.$$

Durch die Wahl $\varrho := 1/\big((n+1)\|B^{-1}\|_\infty\big)$ erreichen wir zunächst $\overline{B_\varrho^{\|\cdot\|_\infty}(\overline{x})} \subseteq \Delta$. (**Quizfrage 16.12:** Warum müssen wir hier, anders als im Beweis von Satz 15.16, nicht $\|(B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_\infty$ nehmen?) Die Äquivalenz der Normen $\|\cdot\|_\infty$ und $\|\cdot\|_2$ im $\mathbb{R}^n$, insbesondere $\|\cdot\|_\infty \leqslant \|\cdot\|_2$, zeigt schließlich, dass wir $r := \varrho$ wählen können, sodass (sogar für die abgeschlossene) r-Kugel gilt:

$$\overline{B_r(0)} \subseteq \overline{B_\varrho^{\|\cdot\|_\infty}(\overline{x})} \subseteq \Delta \subseteq B_1(0) \subseteq M. \qquad \square$$

Aus dem Satz 16.22 folgen weitere Konsequenzen für das Subdifferential und die Richtungsableitung konvexer Funktionen. Wir wissen bereits aus Satz 16.16 (*v*), dass in Punkten $x_0 \in \operatorname{int}\operatorname{dom} f$ die Richtungsableitung $f'(x_0; d)$ für alle Richtungen $d \in \mathbb{R}^n$ endlich ist. Es gilt jedoch mehr:

Lemma 16.24 (Lipschitz-Stetigkeit der Richtungsableitung als Funktion der Richtung)**.**
Es seien $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und $x_0 \in \operatorname{int}\operatorname{dom} f$. Weiter sei L eine Lipschitz-Konstante von f in einer Kugel $B_r(x_0)$, siehe Satz 16.22. Dann gilt

$$|f'(x_0; d_1) - f'(x_0; d_2)| \leqslant L\,\|d_1 - d_2\| \quad \text{für alle } d_1, d_2 \in \mathbb{R}^n \tag{16.29}$$

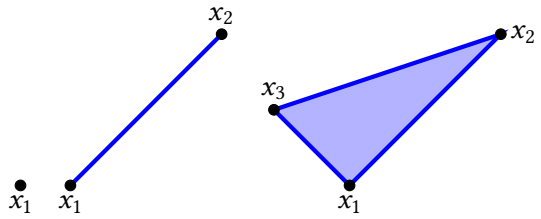


Abbildung 16.5.: Simplizes der Dimensionen 0, 1 und 2 im $\mathbb{R}^2$.

und insbesondere

$$|f'(x_0; d)| \leqslant L \, \|d\| \tag{16.30}$$

für alle $d \in \mathbb{R}^n$.

Beweis. Es sei $d \in \mathbb{R}^n$ beliebig. Für hinreichend kleines $t > 0$ gilt

$$|f(x_0 + t\, d) - f(x_0)| \leqslant L\, t \, \|d\|.$$

Die Division durch t und der Grenzübergang $t \searrow 0$ zeigt $|f'(x_0; d)| \leqslant L \, \|d\|$.

Um die Lipschitz-Stetigkeit von $f'(x_0; \cdot)$ zu zeigen, schätzen wir ab:

$$\begin{aligned}
&f'(x_0; d_1) - f'(x_0; d_2) \\
&\quad = f'(x_0; d_1) - f'(x_0; d_1 + (d_2 - d_1)) \\
&\quad \geqslant f'(x_0; d_1) - f'(x_0; d_1) - f'(x_0; d_2 - d_1) \quad \text{wegen der Subadditivität, siehe (16.19)} \\
&\quad \geqslant -L \, \|d_1 - d_2\|.
\end{aligned}$$

Beachte: Die Subadditivitätsungleichung (16.19) gilt, weil alle vorkommenden Werte $f'(x_0; \cdot)$ wegen (16.30) endlich sind.Analog gilt auch

$$\begin{aligned}
&f'(x_0; d_1) - f'(x_0; d_2) \\
&\quad = f'(x_0; d_2 + (d_1 - d_2)) - f'(x_0; d_2) \\
&\quad \leqslant f'(x_0; d_2) + f'(x_0; d_1 - d_2) - f'(x_0; d_2) \quad \text{wegen der Subadditivität, siehe (16.19)} \\
&\quad \leqslant L \, \|d_1 - d_2\|.
\end{aligned}$$

Das zeigt die Behauptung (16.29). □

Wir wissen bereits aus Satz 16.8 und Satz 16.4, dass das Subdifferential einer konvexen Funktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ in Punkten $x_0 \in \operatorname{relint} \operatorname{dom} f$ nichtleer, abgeschlossen und konvex ist. Im folgenden Satz wird das Subdifferential in inneren Punkten von $\operatorname{dom} f$ noch genauer charakterisiert.

Satz 16.25 (Kompaktheit des Subdifferentials)**.**

Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex. Es sei weiter $x_0 \in \operatorname{int} \operatorname{dom} f$ und L eine Lipschitz-Konstante von f in einer Kugel $B_r(x_0)$, siehe Satz 16.22. Dann ist $\partial f(x_0)$ kompakt, und es gilt $\|s\| \leqslant L$ für alle $s \in \partial f(x_0)$.

Beweis. Für $s \in \partial f(x_0)$ gilt nach Satz 16.18 Aussage (i) und Lemma 16.24:

$$s^\top d \leqslant f'(x_0; d) \leqslant L \|d\| \quad \text{für alle } d \in \mathbb{R}^n.$$

Mit der Wahl $d = s$ folgt $\|s\| \leqslant L$. □

Bemerkung 16.26 (Verallgemeinerung von Satz 16.25)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex. Wenn $x_0 \in \operatorname{relint} \operatorname{dom} f$ ist, dann gilt

$$\partial f(x_0) = U^\perp + \partial(f|_{\operatorname{aff}\operatorname{dom} f})(x_0). \tag{16.31}$$

Hier ist U der Richtungsraum von $\operatorname{aff}\operatorname{dom} f$ und $U^\perp$ sein orthogonales Komplement. Der Ausdruck $\partial(f|_{\operatorname{aff}\operatorname{dom} f})(x_0)$ ist das Subdifferential der Einschränkung von f auf $\operatorname{aff}\operatorname{dom} f$. Für diese Funktion ist x_0 ein innerer Punkt, und man kann zeigen, dass $\partial(f|_{\operatorname{aff}\operatorname{dom} f})(x_0)$ kompakt ist. △

Satz 16.27 (Subdifferential einer konvexen *diffbaren* Funktion, vgl. Rockafellar, 1970, Theorem 25.1)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ konvex.

(i) Wenn f an der Stelle x_0 diffbar ist, dann ist $\partial f(x_0) = \{\nabla f(x_0)\}$, und es gilt $x_0 \in \operatorname{int}\operatorname{dom} f$.

(ii) Wenn das Subdifferential $\partial f(x_0)$ einelementig ist, dann gilt $x_0 \in \operatorname{int}\operatorname{dom} f$, und f ist an der Stelle x_0 diffbar.

Beweis. Aussage (i): Es sei f an der Stelle x_0 diffbar, insbesondere ist also $f(x_0)$ endlich. Aufgrund der Diffbarkeit sind die Richtungsableitungen gegeben durch $f'(x_0; d) = \nabla f(x_0)^\top d$. Gemäß Satz 16.18 Aussage (i) ist $s \in \partial f(x_0)$ genau dann, wenn $s^\top d \leqslant f'(x_0; d)$, also

$$s^\top d \leqslant \nabla f(x_0)^\top d$$

für alle $d \in \mathbb{R}^n$ gilt. Diese Bedingung ist aber genau für $s = \nabla f(x_0)$ erfüllt. (**Quizfrage 16.13:** Warum?) Weiter folgt aus der Diffbarkeit von f an der Stelle x_0 auch die Stetigkeit von f an dieser Stelle. Daraus folgt insbesondere, dass f in einer Umgebung von x_0 nur endliche Werte annimmt, sodass $x_0 \in \operatorname{int}\operatorname{dom} f$ folgt.

Aussage (ii): Es sei $\partial f(x_0) = \{s\}$.

Schritt 1: Wiri zeigen zunächst, dass $f(x_0)$ endlich ist:

Wäre $f(x_0)$ nicht endlich, so würde aus der Subgradientenungleichung (16.1) folgen, dass entweder $\partial f(x_0) = \emptyset$ ist (im Fall $f(x_0) = \infty$ und $\operatorname{dom} f \neq \emptyset$) oder aber $\partial f(x_0) = \mathbb{R}^n$ (im Fall $f \equiv \infty$ oder im Fall $f(x_0) = -\infty$).

Schritt 2: Wir zeigen weiter, dass $x_0 \in \operatorname{int}\operatorname{dom} f$ folgt.

Aus Folgerung 16.19 ergibt sich zunächst, dass $\operatorname{dom} f$ notwendigerweise volle Dimension haben muss, da sonst $\partial f(x_0)$ unbeschränkt wäre. Angenommen, $x_0 \notin \operatorname{int}\operatorname{dom} f$, dann gilt auch $x_0 \notin \operatorname{core}\operatorname{dom} f$ nach Lemma 15.25. Es gibt also eine Richtung $\overline{d} \in \mathbb{R}^n$, $\overline{d} \neq 0$, sodass $x_0 \in \operatorname{dom} f$ liegt, aber $x_0 + \varepsilon\,\overline{d} \notin \operatorname{dom} f$ für alle $\varepsilon > 0$. Daraus folgt $f'(x_0; \overline{d}) = \infty$.

Wir zeigen unter Verwendung von Satz 16.18 Aussage (i), dass dann auch $s + \overline{d} \in \partial f(x_0)$ gilt, im Widerspruch zu Voraussetzung, dass $\partial f(x_0)$ einelementig ist. Es sei dazu $d \in \mathbb{R}^n$ beliebig. Wir können d eindeutig darstellen als $d = d_1 + d_2$ mit $d_1 \perp \overline{d}$ und $d_2 \in \operatorname{span}\{\overline{d}\}$,

also $d = d_1 + \alpha\,\overline{d}$. Wir machen eine Fallunterscheidung nach dem Vorzeichen von α. Wenn $\alpha > 0$ ist, dann gilt

$$f'(x_0; d) \leqslant f'(x_0; d_1) + \alpha f'(x_0; \overline{d}) = \infty$$

wegen Satz 16.16. Damit ist in diesem Fall $(s + \overline{d})^\mathsf{T} d \leqslant f'(x_0; d)$ klar. Andernfalls ist $\alpha \leqslant 0$ und daher

$$(s + \overline{d})^\mathsf{T} d = s^\mathsf{T} d + \overline{d}^\mathsf{T}(d_1 + \alpha\overline{d}) = s^\mathsf{T} d + \alpha \|\overline{d}\|^2 \leqslant s^\mathsf{T} d \leqslant f'(x_0; d),$$

wobei die letzte Ungleichung wieder aus Satz 16.18 Aussage (*i*) folgt. Wir haben also gezeigt, dass

$$(s + \overline{d})^\mathsf{T} d \leqslant f'(x_0; d) \quad \text{für alle } d \in \mathbb{R}^n$$

gilt. Aus Satz 16.18 Aussage (*i*) folgt damit $s + \overline{d} \in \partial f(x_0)$, Widerspruch. Folglich gilt notwendigerweise $x_0 \in \operatorname{int}\operatorname{dom} f$.

Schritt 3: Wir müssen noch zeigen, dass f an der Stelle x_0 diffbar ist. Dazu definieren wir die konvexe (Restglied-)Funktion $g\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$,

$$g(\delta x) := f(x_0 + \delta x) - f(x_0) - s^\mathsf{T}\delta x.$$

Für diese gilt dann $g(\delta x) \geqslant 0$ sowie $0 \in \operatorname{int}\operatorname{dom} g$ und $\partial g(0) = \{0\}$ (**Quizfrage 16.14:** Warum gelten diese Eigenschaften?). Wir müssen zeigen:

$$\lim_{\delta x \to 0} \frac{g(\delta x)}{\|\delta x\|} = 0. \tag{16.32}$$

Es sei dazu $R > 0$ so gewählt, dass $B_R(0) \subseteq \operatorname{dom} g$ liegt. Lemma 16.23 zeigt, dass es affin unabhängige Punkte $v^{(0)}, \ldots, v^{(n)}$ und ein $r > 0$ gibt, sodass mit $\Delta := \operatorname{conv}(\{v^{(0)}, \ldots, v^{(n)}\})$ gilt: $B_r(0) \subseteq \Delta \subseteq B_R(0) \subseteq \operatorname{dom} g$. Folglich gilt $\|v^{(j)}\| < R$ für alle $j = 0, \ldots, n$.

Jedes Element in Δ und insbesonderes jedes $y \in B_r(0)$ hat eine (eindeutige) Darstellung als Konvexkombination

$$y = \sum_{j=0}^{n} \alpha^{(j)}\, v^{(j)} \tag{16.33}$$

mit Koeffizienten $\alpha^{(j)} \geqslant 0$ und $\sum_{j=0}^n \alpha^{(j)} = 1$.

Es sei nun $\delta x \in \mathbb{R}^n$ beliebig, aber fest, mit $\|\delta x\| < \frac{r^2}{R}$ und $\delta x \neq 0$. Daraus folgt $\frac{\|\delta x\|}{r}\,\|v^{(j)}\| < r$.

Wir schätzen nun ab:

$$\begin{aligned}
g(\delta x) &= g\Big(\frac{\|\delta x\|}{r}\, r\, \frac{\delta x}{\|\delta x\|}\Big) \\
&= g\Big(\frac{\|\delta x\|}{r} \sum_{j=0}^{n} \alpha^{(j)}\, v^{(j)}\Big) \quad \text{wegen (16.33)} \\
&\leqslant \sum_{j=0}^{n} \alpha^{(j)}\, g\Big(\frac{\|\delta x\|}{r}\, v^{(j)}\Big) \quad \text{Jensensche Ungleichung (Hausaufgabe 9.5)} \\
&\leqslant \max_{j=0,\ldots,n} g\Big(\frac{\|\delta x\|}{r}\, v^{(j)}\Big).
\end{aligned} \tag{16.34}$$

Die Funktionswerte sind alle endlich, da, wie bereits gezeigt, $\frac{\|\delta x\|}{r} v^{(j)} \in B_r(0) \subseteq \operatorname{dom} g$ liegt.

Nach Satz 16.18 Aussage (ii) gilt wegen $g(0) = 0$ und $\partial g(0) = \{0\}$:

$$g'(0; d) = \max\{s^\mathsf{T} d \mid s \in \partial g(0)\} = 0$$

für alle $d \in \mathbb{R}^n$. Zusammen mit der Monotonie des Differenzenquotienten (Lemma 16.15) erhalten wir

$$0 = g'(0; d) = \lim_{t \searrow 0} \frac{g(t\,d) - g(0)}{t} = \lim_{t \searrow 0} \frac{g(t\,d)}{t} \leqslant \frac{g(t\,d)}{t} \tag{16.35}$$

für alle $t > 0$ und alle $d \in \mathbb{R}^n$. Wir können nun die Restgliedabschätzung (16.32) zeigen: Wegen (16.34) haben wir

$$0 \leqslant \frac{g(\delta x)}{\|\delta x\|} \leqslant \max_{j=0,\dots,n} \frac{g\left(\frac{\|\delta x\|}{r} v^{(j)}\right)}{\|\delta x\|},$$

und wegen (16.35) gilt für jeden Term unter dem Maximum auf der rechten Seite

$$\lim_{\|\delta x\| \searrow 0} \frac{g\left(\frac{\|\delta x\|}{r} v^{(j)}\right)}{\|\delta x\|} = \lim_{t \searrow 0} \frac{g\left(\frac{t}{r} v^{(j)}\right)}{t} = 0.$$

Somit erhalten wir

$$0 \leqslant \lim_{\|\delta x\| \searrow 0} \frac{g(\delta x)}{\|\delta x\|} = 0,$$

was zu zeigen war. □

Ende der Vorlesung 25

Ende der Woche 13

§ 17 Kegel

Literatur: Geiger, Kanzow, 2002, Kapitel 2.2.1

Definition 17.1 (Kegel)**.**
Eine Menge $K \subseteq \mathbb{R}^n$ heißt ein **Kegel** (englisch: *cone*), wenn $\beta\, x \in K$ gilt für $x \in K$ und alle $\beta > 0$. Kurz: $\beta\, K \subseteq K$ für alle $\beta > 0$. Der Kegel K heißt **spitz** (englisch: *pointed*), wenn $0 \in K$ ist, ansonsten **stumpf** (englisch: *stumpf*). △

Beachte: Mit jedem Punkt $x \in \mathbb{R}^n$ enthält ein Kegel bereits die gesamte Halbgerade $\{\beta\, x \mid \beta > 0\}$.

Beispiel 17.2 (Kegel)**.**
Beispiele für Kegel sind:

(i) offene Halbgeraden $\{\beta\, a \mid \beta > 0\}$ mit $a \in \mathbb{R}^n$, $a \neq 0$,

(ii) abgeschlossene Halbgeraden $\{\beta\, a \mid \beta \geqslant 0\}$ mit $a \in \mathbb{R}^n$, $a \neq 0$,

(*iii*) offene Orthanten $\mathbb{R}^n_{>0} = \{x \in \mathbb{R}^n \mid x > 0\}$,

(*iv*) abgeschlossene Orthanten $\mathbb{R}^n_{\geqslant 0} = \{x \in \mathbb{R}^n \mid x \geqslant 0\}$,

(*v*) der **Lorentzkegel** (englisch: *Lorentz cone*) $K = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} \mid \|x\| \leqslant t\}$,

(*vi*) die Menge der symmetrisch positiv definiten Matrizen S^n_{++} in $\mathbb{R}^{n\times n}$,

(*vii*) die Menge der symmetrisch positiv semidefiniten Matrizen S^n_{+} in $\mathbb{R}^{n\times n}$. △

Beachte: Kegel sind i. A. nicht konvex.

Satz 17.3 (Operationen auf Kegeln)**.**

(*i*) Es sei $\{K_j\}_{j\in J}$ eine in $\mathbb{R}^n$. Dann ist $\bigcap_{j\in J} K_j$ ein Kegel.

(*ii*) Es seien $K_i \subseteq \mathbb{R}^{n_i}$, $i = 1, \dots, k$ Kegel. Dann ist das kartesische Produkt $K_1 \times \cdots \times K_k$ ein Kegel in $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$.

(*iii*) Es sei $f\colon \mathbb{R}^n \to \mathbb{R}^m$ eine lineare Abbildung und $K \subseteq \mathbb{R}^n$ und $\widehat{K} \subseteq \mathbb{R}^m$ Kegel. Dann sind das Bild $f(K) \subseteq \mathbb{R}^m$ und das Urbild $f^{-1}(\widehat{K}) \subseteq \mathbb{R}^n$ Kegel. (**Quizfrage 17.1:** Gilt das auch, wenn f affin-linear ist?)

(*iv*) Sind $K_1, K_2 \subseteq \mathbb{R}^n$ Kegel, dann sind

$$\alpha K_1 = \{\alpha x_1 \mid x_1 \in K_1\} \quad \text{für } \alpha \in \mathbb{R}$$

sowie die Minkowski-Summe

$$K_1 + K_2$$

Kegel.

(*v*) Das Komplement $K^c = \mathbb{R}^n \setminus K$ eines Kegels $K \subseteq \mathbb{R}^n$ ist ein Kegel.

(*vi*) Ist $K \subseteq \mathbb{R}^n$ ein Kegel, dann sind das Innere $\operatorname{int}(K)$, der Abschluss $\overline{K}$ und der Rand ∂K wieder Kegel.

Beweis. Aussage (*i*): Für $x \in \bigcap_{j\in J} K_j$ ist $x \in K_j$ für alle $j \in J$. Da die K_j Kegel sind, ist $\beta x \in K_j$ für alle $j \in J$ alle $\beta > 0$. Entsprechend ist $\beta x \in \bigcap_{j\in J} K_j$ für alle $\beta > 0$.

Aussage (*ii*): Für $x \in K_1 \times \cdots \times K_k$ ist $x_i \in K_i \subseteq .\mathbb{R}^n_i$ für $i = 1, \dots, k$. Da die K_i Kegel sind ist sind damit auch $\beta x_i \in K_i$, für alle $i = 1, \dots, k$ und $\beta > 0$. Entsprechend ist $\beta x \in K_1 \times \cdots \times K_k$ für alle $\beta > 0$.

Aussage (*iii*): Es sei $\beta > 0$ und $y \in f(K)$ mit $y = f(x)$ für ein $x \in K$. Da K ein Kegel ist, ist $\beta x \in K$ und auf Grund der Linearität von f ist $\beta y = \beta f(x) = f(\beta x) \in f(K)$ und damit $f(K)$ ein Kegel.

Es sei nun weiter $x \in f^{-1}(\widehat{K})$ (also $f(x) \in \widehat{K}$) und $\beta > 0$. Da $\widehat{K}$ ein Kegel ist, ist $f(\beta x) = \beta f(x) \in \widehat{K}$ und damit $\beta x \in f^{-1}(\widehat{K})$.

Aussage (*iv*): Es seien $\alpha, \beta > 0$ und $x \in \alpha K_1$. Nach Definition ist $\frac{1}{\alpha} x \in K_1$, also, da K_1 ein Kegel ist, auch $\frac{\beta}{\alpha} x \in K_1$ und damit $\beta x \in \alpha K_1$.

Es sei nun $\beta > 0$ und $y \in K_1 + K_2$ mit $y = x_1 + x_2$ für $x_i \in K_i$, $i = 1, 2$. Dann ist $\beta y = \beta x_1 + \beta x_2$ und da die K_i Kegel sind, sind $\beta x_i \in K_i$ und damit $\beta y \in K_1 + K_2$.

Aussage (v): Es seien $\beta > 0$ und $x \in K^c$. Angenommen βx wäre nicht in K^c, sondern in K, dann wäre $\underbrace{\frac{1}{\beta}}_{>0}(\beta x) = x \in K$, was ein Widerspruch ist.

Aussage (vi): Es sei $\beta > 0$ und $x \in \operatorname{int}(K)$ mit $B_\varepsilon(x) \subseteq K$. Dann ist $B_{\beta\varepsilon}(\beta x) \subseteq K$, denn für $\tilde{x} \in B_{\beta\varepsilon}(\beta x)$ ist

$$\left\| \frac{1}{\beta}\tilde{x} - x \right\| = \frac{1}{\beta} \|\tilde{x} - \beta x\| < \frac{1}{\beta}\beta\varepsilon = \varepsilon,$$

also $\frac{1}{\beta}\tilde{x} \in B_\varepsilon(x) \subseteq K$, damit $\tilde{x} \in K$ und entsprechend $\beta x \in \operatorname{int}(K)$.

Mit Aussage (v) folgt sofort, dass $\overline{K} = \operatorname{int}(K^c)^c$ ein Kegel ist.

Mit Aussagen (i) und (v) ist $\partial K = \overline{K} \cap \overline{K^c}$ ebenfalls ein Kegel. □

Expertenwissen: Kegelhülle

Die **Kegelhülle** einer Menge $A \subseteq \mathbb{R}^n$ bzw. der **von A erzeugte Kegel** ist definiert als

$$\operatorname{cone}(A) := \bigcap \left\{ K \subseteq \mathbb{R}^n \,\middle|\, K \text{ ist Kegel und } A \subseteq K \right\}.$$

Man kann zeigen, dass

$$\operatorname{cone}(A) = \bigcup_{\beta > 0} \beta A$$

gilt.

Lemma 17.4 (Konvexe Kegel)**.**
Es sei $K \subseteq \mathbb{R}^n$.

(a) Folgende Aussagen sind äquivalent:
  (i) K ist ein konvexer Kegel.
  (ii) Es gilt $\alpha^{(1)} x_1 + \alpha^{(2)} x_2 \in K$ für alle $x_1, x_2 \in K$ und $\alpha^{(1)}, \alpha^{(2)} > 0$.

(b) Folgende Aussagen sind äquivalent:
  (i) K ist ein spitzer konvexer Kegel.
  (ii) Es gilt $\alpha^{(1)} x_1 + \alpha^{(2)} x_2 \in K$ für alle $x_1, x_2 \in K$ und $\alpha^{(1)}, \alpha^{(2)} \geqslant 0$.

Beweis. Aussage (a): Es seien zunächst $K \subseteq \mathbb{R}^n$ ein konvexer Kegel und $x_1, x_2 \in K$ sowie $\alpha^{(1)}, \alpha^{(2)} > 0$. Dann sind $\alpha^{(1)} x_1$ und $\alpha^{(2)} x_2$ in K und

$$\alpha^{(1)} x_1 + \alpha^{(2)} x_2 = 2 \overbrace{\left(\tfrac{1}{2}\underbrace{\alpha^{(1)} x_1}_{\in K} + \tfrac{1}{2}\underbrace{\alpha^{(2)} x_2}_{\in K} \right)}^{\in K \,(\text{Konvexität})} \in K.$$

Umgekehrt seien $x_1, x_2 \in K$ und $\alpha \in (0,1)$. Nach Voraussetzung ist $\alpha x_1 + (1-\alpha) x_2 \in K$, also K konvex. Es sei weiter $x \in K$ und $\beta > 0$. Wähle $\alpha^{(1)} = \alpha^{(2)} = \beta/2$ und $x_1 = x_2 = x$. Nach Voraussetzung ist $\beta x = \alpha x_1 + (1-\alpha) x_2 \in K$, also K ein Kegel.

Aussage (b): analog. □

Beachte: Der Rezessionskegel (6.13) eines Polyeders in Normalform sowie die konvexe Kegelhülle (6.14) $\operatorname{pos}\{b_1, \ldots, b_n\}$ einer endlichen Menge von Vektoren sind konvexe Kegel, die zudem abgeschlossen sind.

Für den Rest dieses Abschnitts werden wir uns mit Kegeln beschäftigen, die in der (insbesondere konvexen) Optimierung eine besondere Bedeutung haben.

§ 17.1 Radialkegel und Kegel zulässiger Richtungen

Definition 17.5 (Kegel zulässiger Richtungen)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge und $x \in M$. Dann heißt

$$\mathcal{F}_M(x) := \left\{ d \in \mathbb{R}^n \,\middle|\, \text{es gibt ein } \varepsilon > 0, \text{ sodass } x + t\,d \in M \text{ liegt für alle } t \in [0, \varepsilon] \right\} \tag{17.1}$$

der **Kegel der zulässigen Richtungen** (englisch: *cone of feasible directions*) der Menge M im Punkt x. Ein Vektor $d \in \mathcal{F}_M(x)$ heißt **zulässige Richtung** (englisch: *feasible direction*) von M im Punkt x. Man definiert $\mathcal{F}_M(x) := \emptyset$ für $x \notin M$. △

Definition 17.6 (Radialkegel)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge und $x \in M$. Dann heißt

$$\mathcal{K}_M(x) := \left\{ \beta\,(y - x) \,\middle|\, y \in M,\ \beta > 0 \right\} = \bigcup_{\beta > 0} \beta\,(M - x) \tag{17.2}$$

der **von $M - x$ erzeugte Kegel** oder der **Radialkegel** (englisch: *radial cone*) an die Menge M im Punkt x. Ein Vektor $d \in \mathcal{K}_M(x)$ heißt **radiale Richtung** (englisch: *radial direction*) von M im Punkt x. Man definiert $\mathcal{K}_M(x) := \emptyset$ für $x \notin M$. △

Quizfrage 17.2: Warum sind $\mathcal{F}_M(x)$ und $\mathcal{K}_M(x)$ Kegel?

Quizfrage 17.3: Was sind $\mathcal{F}_M(x)$ und $\mathcal{K}_M(x)$ im Fall $x \in \operatorname{int}(M)$?

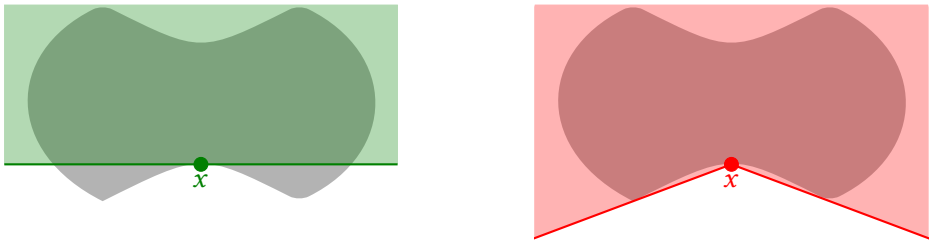


Abbildung 17.1.: Kegel der zulässigen Richtungen $\mathcal{F}_M(x)$ (grün) und Radialkegel $\mathcal{K}_M(x)$ (rot) einer nichtkonvexen Menge M in einem Punkt x. Dargestellt ist jeweils der verschobene Kegel $x + \mathcal{F}_M(x)$ bzw. $x + \mathcal{K}_M(x)$.

Satz 17.7 (Eigenschaften von $\mathcal{F}_M(x)$, $\mathcal{K}_M(x)$ und deren Zusammenhang)**.**
Es sei $M \subseteq \mathbb{R}^n$ beliebig und $x \in M$. Dann gilt:

(*i*) $\mathcal{F}_M(x)$ und $\mathcal{K}_M(x)$ sind spitze Kegel.

(*ii*) $\mathcal{F}_M(x) \subseteq \mathcal{K}_M(x)$.

(*iii*) $M \subseteq x + \mathcal{K}_M(x)$.

(*iv*) Es sei $C \subseteq \mathbb{R}^n$ konvex. Dann ist $\mathcal{K}_C(x)$ für jedes $x \in C$ ein spitzer konvexer Kegel, und es gilt $\mathcal{F}_C(x) = \mathcal{K}_C(x)$.

Beweis. Der Beweis ist Inhalt von Hausaufgabe 14.1. □

Beispiel 17.8 (Beispiele für Kegel der zulässigen Richtungen und Radialkegel)**.**

(*i*) Der Kegel der zulässigen Richtungen $\mathcal{F}_M(x)$ und der Radialkegel $\mathcal{K}_M(x)$ sind i. A. nicht abgeschlossen: $M = \overline{B_1(0)}$ und $x \in \partial M$. Dann ist $\mathcal{F}_M(x) = \mathcal{K}_M(x) =$ offener Halbraum $\cup \{0\}$ nicht abgeschlossen.

(*ii*) Der Kegel der zulässigen Richtungen $\mathcal{F}_M(x)$ und der Radialkegel $\mathcal{K}_M(x)$ sind i. A. nicht konvex: $M = \mathbb{R}_{\geqslant 0} \times \{0\} \cup \{0\} \times \mathbb{R}_{\geqslant 0}$ (L shape) und $x = 0$. Dann ist $\mathcal{F}_M(x) = \mathcal{K}_M(x) = \mathbb{R}_{\geqslant 0} \times \{0\} \cup \{0\} \times \mathbb{R}_{\geqslant 0}$ nicht konvex.

(*iii*) Der Kegel der zulässigen Richtungen $\mathcal{F}_M(x)$ und der Radialkegel $\mathcal{K}_M(x)$ sind i. A. nicht gleich: $M = \mathbb{R}_{\geqslant 0} \times \{0\} \cup \{0\} \times \mathbb{R}_{\geqslant 0}$ (L shape) und $x = \binom{1}{0}$. Dann ist $\mathcal{F}_M(x) = \mathbb{R} \times \{0\}$ konvex, aber $\mathcal{K}_M(x) = \mathbb{R} \times \{0\} \cup \mathbb{R}_{<0} \times \mathbb{R}_{\geqslant 0}$ nicht konvex. △

Quizfrage 17.4: Können Sie ein Beispiel für eine *nichtkonvexe* Menge M und einen Punkt $x \in M$ finden, sodass dennoch $\mathcal{F}_M(x) = \mathcal{K}_M(x)$ gilt?

Lemma 17.9 (Richtungsableitung der Indikatorfunktion)**.**
Es sei $C \subseteq \mathbb{R}^n$ eine nichtleere konvexe Menge und $x \in C$. Dann gilt:

$$I'_C(x; d) = \begin{cases} 0, & \text{falls } d \in \mathcal{F}_C(x) \text{ oder äquivalent: } d \in \mathcal{K}_C(x), \\ \infty & \text{sonst.} \end{cases} \tag{17.3}$$

Beweis. Der Differenzenquotient für $t > 0$

$$q(t) = \frac{I_C(x + t\,d) - I_C(x)}{t} = \frac{I_C(x + t\,d) - 0}{t}$$

wird entweder gleich null für ein $t_0 > 0$ (und dann wegen der Konvexität von C auch für alle $t \in [0, t_0]$), oder es ist $q(t) = \infty$ für alle $t > 0$. Der erste Fall ist genau der Fall $d \in \mathcal{F}_C(x)$, siehe (17.1). Die Gleichheit $\mathcal{F}_C(x) = \mathcal{K}_C(x)$ wurde in Satz 17.7 gezeigt. □

§ 17.2 Normalenkegel

Definition 17.10 (Normalenkegel)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge und $x \in M$. Dann heißt

$$\mathcal{N}_M(x) := \{s \in \mathbb{R}^n \mid s^\mathsf{T}(y - x) \leqslant 0 \text{ für alle } y \in M\} \tag{17.4}$$

der **Normalenkegel** (englisch: *normal cone*) von M im Punkt x. Ein Vektor $s \in \mathcal{N}_M(x)$ heißt **Normalenrichtung** (englisch: *normal direction*) von M im Punkt x. Man definiert $\mathcal{N}_M(x) := \emptyset$ für $x \notin M$. △

Beachte: s ist genau dann eine Normalenrichtung von M im Punkt $x \in M$, wenn M enthalten ist im Halbraum $H^-(s, \beta) = \{y \in \mathbb{R}^n \mid s^\mathsf{T} y \leqslant \beta\}$ mit Normalenvektor s und Offset $\beta := s^\mathsf{T} x$. Mit anderen Worten: Die Normalenrichtungen im Punkt x sind (bis auf $s = 0$) gerade die Normalenvektoren von Hyperebenen, die den Punkt x von M trennen, wobei M im negativen Halbraum liegt.

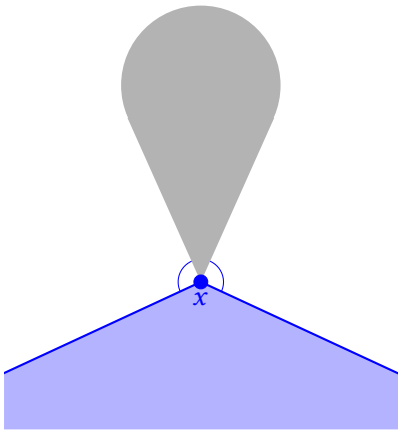


Abbildung 17.2.: Normalenkegel $\mathcal{N}_M(x)$ (blau) einer konvexen Menge M in einem Punkt x. Dargestellt ist der verschobene Kegel $x + \mathcal{N}_M(x)$.

Quizfrage 17.5: Was ist $\mathcal{N}_M(x)$ im Fall $x \in \operatorname{int}(M)$?

Lemma 17.11 (Eigenschaften des Normalenkegels)**.**
Es sei $M \subseteq \mathbb{R}^n$ eine beliebige Menge und $x \in M$. Dann gilt:

(*i*) Der Normalenkegel $\mathcal{N}_M(x)$ ist ein konvexer, abgeschlossener Kegel.

(*ii*) Es gilt

$$\mathcal{N}_M(x) = \mathcal{K}_M(x)^\circ := \{s \in \mathbb{R}^n \mid s^\mathsf{T} d \leqslant 0 \text{ für alle } d \in \mathcal{K}_M(x)\}.$$

Man sagt: Der Normalenkegel ist der **Polarkegel**[10] des Radialkegels.

Beweis. Der Beweis ist Inhalt von Hausaufgabe 14.2. □

Lemma 17.12 (Das Subdifferential der Indikatorfunktion ist der Normalenkegel)**.**
Es sei $C \subseteq \mathbb{R}^n$ eine nichtleere konvexe Menge. Dann gilt:

$$\partial I_C(x) = \mathcal{N}_C(x) \tag{17.5}$$

für alle $x \in \mathbb{R}^n$, d. h., das Subdifferential der Indikatorfunktion einer konvexen Menge ist gerade der Normalenkegel von C im Punkt x.

[10] Allgemein ist der **Polarkegel** einer beliebigen Menge $M \subseteq \mathbb{R}^n$ gegeben durch

$$M^\circ = \{s \in \mathbb{R}^n \mid s^\mathsf{T} y \leqslant 0 \text{ für alle } y \in M\}.$$

Beweis. Falls $x \notin C$ ist, dann ist $\partial I_C(x) = \emptyset$ (**Quizfrage 17.6:** Warum nochmal?) und $\mathcal{N}_C(x) = \emptyset$ per Definition. Im Fall $x \in C$ liegt $s \in \partial I_C(x)$ genau dann, wenn gilt:

$$I_C(y) \geqslant \underbrace{I_C(x)}_{=0} + s^\mathsf{T}(y - x) \quad \text{für alle } y \in \mathbb{R}^n.$$

Da diese Ungleichung für $y \notin C$ trivialerweise erfüllt ist, reicht es, sie für $y \in C$ zu fordern. Es ist also $s \in \partial I_C(x)$ genau dann, wenn

$$0 \geqslant 0 + s^\mathsf{T}(y - x) \quad \text{für alle } y \in C$$

gilt. Das ist aber gerade die Bedingung dafür, dass s zum Normalenkegel $\mathcal{N}_C(x)$ gehört, siehe (17.4). □

Quizfrage 17.7: Stimmt die Aussage von Lemma 17.12 auch noch im Fall $C = \emptyset$?

Ende der Vorlesung 26

§ 18 Optimalitätsbedingungen der konvexen Optimierung

Literatur: Rockafellar, 1970, Section 27

Wir betrachten wieder die konvexe Optimierungsaufgabe aus (14.1)

$$\text{Minimiere} \quad f(x) \quad \text{über } x \in \mathbb{R}^n \tag{18.1}$$

mit konvexer Zielfunktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$. Desweiteren nehmen wir in den Resultaten dieses Abschnitts f als eigentlich an. Das ist keine wesentliche Einschränkung, denn wenn $f \equiv \infty$ ist, dann ist die Aufgabe (18.1) nicht interessant. Auch wenn f irgendwo den Wert $-\infty$ annimmt, ist (18.1) als Optimierungsaufgabe nicht interessant, weil die globalen Minimierer dann einfach die Elemente von $f^{-1}(-\infty)$ sind.

Satz 18.1 (Notwendige und hinreichende Optimalitätsbedingungen)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und eigentlich. Dann sind die folgenden Aussagen für einen Punkt $x^* \in \mathbb{R}^n$ äquivalent:

(*i*) x^* ist ein (globaler) Minimierer für (18.1).

(*ii*) $f(x^*)$ ist endlich, und es gilt $f'(x^*; d) \geqslant 0$ für alle $d \in \mathbb{R}^n$.

(*iii*) Es gilt $0 \in \partial f(x^*)$.

Beweis. Die Äquivalenz von Aussage (*i*) und Aussage (*iii*) hatten wir bereits in Satz 16.3 bewiesen.

Aussage (*ii*) $\Rightarrow$ Aussage (*iii*): Da $f(x^*)$ nach Voraussetzung endlich ist, gilt nach Satz 16.18, dass $f'(x^*; d) \geqslant 0^\mathsf{T} d = 0$ für alle $d \in \mathbb{R}^n$ die gewünschte Eigenschaft $0 \in \partial f(x^*)$ impliziert.

Aussage (*iii*) $\Rightarrow$ Aussage (*ii*): $0 \in \partial f(x^*)$ besagt $f(x) \geqslant f(x^*) + 0^\mathsf{T}(x - x^*)$ für alle $x \in \mathbb{R}^n$. Da f eigentlich ist, folgt daraus, dass $f(x^*)$ endlich ist. Wiederum aus Satz 16.18 folgt, dass $0 \in \partial f(x^*)$ die gewünschte Eigenschaft $f'(x^*; d) \geqslant 0^\mathsf{T} d = 0$ für alle $d \in \mathbb{R}^n$ impliziert. □

Beachte: Dieses Resultat verallgemeinert die notwendigen Optimalitätsbedingungen der unrestringierten Optimierung aus Satz 3.1.

Beispiel 18.2 (Lösung von Proximal-Aufgaben mit Hilfe der Optimalitätsbedingungen).

(*i*) Wir betrachten die Aufgabe

$$\text{Minimiere} \quad \tau \|x\|_1 + \frac{1}{2}\|x - z\|_2^2$$

für gegebenes $\tau \geqslant 0$ und z, zunächst über $x \in \mathbb{R}$ und anschließend über $x \in \mathbb{R}^n$. Die eindeutige Lösung $z \mapsto x$ definiert die sogenannte **proximale Abbildung** (englisch: *proximal map*) von τf für die Funktion $f(x) = \|x\|_1$:

$$x^* = \operatorname{prox}_{\tau f}(z).$$

Zunächst in $\mathbb{R}$: Die Zielfunktion $\tau |x| + \frac{1}{2}|x - z|_2^2$ ist die Summe zweier konvexer, reellwertiger Funktionen. Die Regularitätsbedingung der Summenregel (Satz 16.9) ist damit erfüllt. Da die quadrierte Norm strikt (sogar stark) konvex ist, ist die Zielfunktion außerdem strikt konvex, der Minimierer, sofern er existiert, sind eindeutig.

Fall 1: Der Punkt $x^* = 0$ ist nach Satz 18.1 genau dann die Lösung der Aufgabe, wenn

$$0 \in \tau\,[-1, 1] + (0 - z)$$

gilt, also genau dann, wenn $|z| \leqslant \tau$ ist.

Fall 2: Andererseits ist $x^* \neq 0$ genau dann die Lösung, wenn

$$0 = \tau\,(\operatorname{sgn} x^*) + (x^* - z)$$

gilt. Im Fall $x^* > 0$ bedeutet das $0 = \tau + x^* - z$, d. h., $z = x^* + \tau > \tau$. Analog gilt im Fall $x^* < 0$ dann $z = x^* - \tau < -\tau$.

Wir können beide Fälle gemeinam in der Lösungformel

$$x^* = \max\bigl\{|z| - \tau,\ 0\bigr\}(\operatorname{sgn} z)$$

darstellen.

Analog können wir zeigen, dass sich diese Formel im vektorwertigen Fall wie folgt verallgemeinert:

$$x_i^* = \max\bigl\{|z_i| - \tau,\ 0\bigr\}(\operatorname{sgn} z_i). \tag{18.2}$$

Das liegt daran, dass die Zielfunktion eine Summe von Funktionen ist, die jeweils nur von einer der Variablen x_i abhängen. Diese Abbildung $z \mapsto x^*$ (also die proximale Abbildung von $\tau \|\cdot\|_1$) heißt auch **soft thresholding**.

(*ii*) Wir betrachten die Aufgabe

$$\text{Minimiere} \quad \tau \|x\|_2 + \frac{1}{2}\|x - z\|_2^2 \quad \text{über } x \in \mathbb{R}^n$$

für gegebenes $\tau \geqslant 0$ und z in $\mathbb{R}^n$.

Die eindeutige Lösung $z \mapsto x$ definiert die **proximale Abbildung** von τf mit $f(x) = \|x\|_2$:

$$x^* = \operatorname{prox}_{\tau f}(z).$$

Wie oben ist der Minimierer, sofern er existiert, eindeutig. Wir betrachten sofort den vektorwertigen Fall.

Fall 1: Der Punkt $x^* = 0$ ist nach Satz 18.1 genau dann die Lösung der Aufgabe, wenn

$$0 \in \tau\,\overline{B_1(0)} + (0 - z)$$

gilt, also genau dann, wenn $\|z\|_2 \leqslant \tau$ ist; siehe Beispiel 16.5.

Fall 2: Andererseits ist $x^* \neq 0$ genau dann die Lösung, wenn

$$0 = \tau\,\frac{x^*}{\|x^*\|_2} + (x^* - z)$$

gilt. Durch Umstellen erhalten wir

$$z = \tau\,\frac{x^*}{\|x^*\|_2} + x^*$$

und daher $\|x^*\|_2 = \|z\|_2 - \tau$. Somit ist

$$x^* = \frac{\|x^*\|_2}{\tau + \|x^*\|_2} z = \frac{\|z\|_2 - \tau}{\|z\|_2} z = \left(1 - \frac{\tau}{\|z\|_2}\right) z.$$

Wir können beide Fälle in der gemeinsamen Lösungformel

$$x^* = \max\left\{\|z\|_2 - \tau,\ 0\right\} \frac{z}{\|z\|_2} \tag{18.3}$$

darstellen. Im Fall $z = 0$ wird dieser Ausdruck als $x^* = 0$ gelesen. △

Folgerung 18.3 (Notwendige und hinreichende Optimalitätsbedingungen im diffbaren Fall)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und eigentlich. Dann sind die folgenden Aussagen für einen Punkt $x^* \in \mathbb{R}^n$, an dem f diffbar ist, äquivalent:

(*i*) x^* ist ein (globaler) Minimierer für (18.1).

(*ii*) Es gilt $f'(x^*)\,d = 0$ für alle $d \in \mathbb{R}^n$.

(*iii*) Es gilt $\nabla f(x^*) = 0$.

Beweis. Da f an der Stelle x^* diffbar ist, ist insbesondere $f(x^*)$ endlich. Das Resultat folgt sofort aus Satz 18.1 in Verbindung mit $\partial f(x^*) = \{\nabla f(x^*)\}$ (Satz 16.27 Aussage (*i*)) und der Tatsache, dass wegen der Diffbarkeit von f in x^* gilt: $f'(x^*; d) = f'(x^*)\,d = \nabla f(x^*)^\mathsf{T} d$. □

Da die Zielfunktion f eigentlich ist, kann die Aufgabe (18.1) bereits implizite Beschränkungen dadurch beinhalten, dass unzulässige Punkte durch den Funktionswert $f(x) = \infty$ effektiv ausgeschlossen werden. Jeder globale Minimierer x^* liegt nach Definition 14.1 notwendigerweise in $\operatorname{dom} f$.

Möchte man aber weitere Beschränkungen hinzufügen, so kann man dies durch Betrachtung der Aufgabe

$$\text{Minimiere} \quad f(x) + I_C(x) \quad \text{über } x \in \mathbb{R}^n \tag{18.4}$$

tun, wobei C eine nichtleere konvexe Menge ist. Wir sprechen hier von **abstrakten Nebenbedingungen** (englisch: *abstract constraints*) im Gegensatz zu Nebenbedingungen, die in Form von Gleichungen oder Ungleichungen gegeben sind, vgl. (1.1). Effektiv findet die Minimierung dann über $C \cap \operatorname{dom} f$ statt. Wir geben eine Version von Satz 18.1 für diese Aufgabe an. Zuvor benötigen wir jedoch eine Aussage über die Richtungsableitung von Funktionen der Bauart wie in (18.4).

Lemma 18.4 (Richtungsableitung von $f + I_C$)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und eigentlich, C eine nichtleere konvexe Menge sowie $x \in C$ ein Punkt, an dem $f(x)$ endlich ist. Dann gilt:

$$(f + I_C)'(x; d) = \begin{cases} f'(x; d) & \text{für } d \in \mathcal{F}_C(x) = \mathcal{K}_C(x), \\ \infty & \text{sonst.} \end{cases} \tag{18.5}$$

Beweis. Unter Beachtung von $I_C(x) = 0$ betrachten wir den Differenzenquotienten für $f + I_C$ an der Stelle x in Richtung d und mit $t > 0$:

$$\frac{f(x + t\,d) - f(x)}{t} + \frac{I_C(x + t\,d)}{t} =: q_1(t) + q_2(t).$$

Wir unterscheiden verschiedene Fälle. Wir kennen aus Lemma 17.9 die Darstellung der Richtungsableitung (17.3) $I_C'(x; d) = 0$ für $d \in \mathcal{F}_C(x) = \mathcal{K}_C(x)$ und $I_C'(x; d) = \infty$ sonst.

Fall 1: Falls $f'(x; d) \in \mathbb{R}$ oder $f'(x; d) = \infty$ ist, so können wir den Grenzübergang $t \searrow 0$ in $q_1(t)$ und $q_2(t)$ einzeln durchführen und erhalten die Behauptung (18.5).

Fall 2: Ist $f'(x; d) = -\infty$ und $I_C'(x; d) = 0$ (also $d \in \mathcal{K}_C(x)$), dann ist $q_2(t) = 0$ für alle hinreichend kleinen $t > 0$, und wir können ebenfalls den Grenzübergang einzeln durchführen, und es folgt die Behauptung (18.5).

Fall 3: Der verbleibende Fall $f'(x; d) = -\infty$ und $I_C'(x; d) = \infty$ (also $d \notin \mathcal{K}_C(x)$) bedeutet, dass $q_1(t)$ für alle hinreichend kleinen $t > 0$ endlich ist (**Quizfrage 18.1:** Warum?), aber $q_2(t) = \infty$. Damit ist $q_1(t) + q_2(t) = \infty$ für alle hinreichend kleinen $t > 0$, also auch der Grenzwert. □

Satz 18.5 (Notwendige und hinreichende Optimalitätsbedingungen unter abstrakten Nebenbedingungen)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und eigentlich sowie $C \subseteq \mathbb{R}^n$ konvex und nichtleer. Dann sind die folgenden Aussagen für einen Punkt $x^* \in C$ äquivalent:

(*i*) x^* ist ein (globaler) Minimierer für (18.4).

(*ii*) $f(x^*)$ ist endlich, und es gilt $f'(x^*; d) \geqslant 0$ für alle $d \in \mathcal{K}_C(x^*)$.

(*iii*) $f(x^*)$ ist endlich, und es gilt $f'(x^*; x - x^*) \geqslant 0$ für alle $x \in C$.

(*iv*) Es gilt $0 \in \partial(f + I_C)(x^*)$.

Ferner gilt: Die beiden folgenden Aussagen sind untereinander äquivalent, und jede der Aussagen ist hinreichend für jede der Aussagen (*i*) bis (*iv*).

(*v*) Es gilt $0 \in \partial f(x^*) + \mathcal{N}_C(x^*)$.

(*vi*) Es gibt ein $s \in \partial f(x^*)$ mit der Eigenschaft $s^\mathsf{T}(x - x^*) \geqslant 0$ für alle $x \in C$.

Falls die Regularitätsbedingung $\operatorname{relint} \operatorname{dom} f \cap \operatorname{relint} C \neq \emptyset$ gilt, dann ist jede der Aussagen (*v*) und (*vi*) auch notwendig für jede der Aussagen (*i*) bis (*iv*).

Beweis. Wir setzen $g := f + I_C$. Nach Satz 18.1 sind folgende Aussagen äquivalent:

- Aussage (*i*),
- $0 \in \partial g(x^*)$, also Aussage (*iv*),

- $g(x)$ ist endlich, und es gilt $g'(x;d) \geqslant 0$ für alle $d \in \mathbb{R}^n$, was nach (18.5) gleichbedeutend ist mit

$$f'(x;d) \geqslant 0 \quad \text{für } d \in \mathcal{K}_C(x), \tag{18.6}$$

  also Aussage *(ii)*.

Aus Aussage *(ii)* folgt aber sofort Aussage *(iii)*, weil

$$x - x^* \in \mathcal{K}_C(x^*) = \{\beta\,(x - x^*) \,|\, x \in C,\ \beta > 0\}$$

ist. Umgekehrt folgt aus $f'(x^*; x - x^*) \geqslant 0$ mit der positiven Homogenität der Richtungsableitung (Satz 16.16 Aussage *(i)*) auch $f'(x^*; \beta\,(x - x^*)) = \beta\, f'(x^*; x - x^*) \geqslant 0$ für alle $\beta > 0$, also impliziert Aussage *(iii)* auch Aussage *(ii)*. Die Äquivalenzen der Aussagen *(i)* bis *(iv)* sind damit gezeigt.

Aufgrund der Summenregel für das Subdifferential aus Satz 16.9 gilt $\partial f(x^*) + \partial I_C(x^*) \subseteq \partial(f + I_C)(x^*)$, wobei nach Lemma 17.12 wiederum $\partial I_C(x^*) = \mathcal{N}_C(x^*)$ ist. Wir haben damit gezeigt, dass Aussage *(v)* hinreichend für Aussage *(iv)* ist.

Aussage *(v)* bedeutet, dass ein $s \in \partial f(x^*)$ existiert mit der Eigenschaft $-s \in \mathcal{N}_C(x^*)$. Die Definition 17.10 des Normalenkegels zeigt sofort die Äquivalenz mit Aussage *(vi)*.

Aus Satz 16.9 folgt sogar $\partial f(x^*) + \partial I_C(x^*) = \partial(f + I_C)(x^*)$, also die Äquivalenz von Aussage *(iv)* und Aussage *(v)*, falls die Regularitätsbedingung (16.9), also $(\operatorname{relint} \operatorname{dom} f) \cap (\operatorname{relint} \operatorname{dom} I_C) = (\operatorname{relint} \operatorname{dom} f) \cap (\operatorname{relint} C)$ erfüllt ist. Damit ist alles gezeigt. □

Folgerung 18.6 (Orthogonale Projektionsaufgabe)**.**

Wir betrachten nochmals die orthogonale Projektionsaufgabe (15.2) aus Beispiel 15.1. Es sei also C eine nichtleere, abgeschlossene, konvexe Menge und $p \in \mathbb{R}^n$. Dann gilt nach Satz 18.5 *(i)* und *(iii)*: $x^* \in C$ ist genau dann gleich $\operatorname{proj}_C(p)$, also der eindeutige Minimierer von

$$\text{Minimiere} \quad f(x) := \frac{1}{2}\|x - p\|^2 + I_C(x) \quad \text{über } x \in \mathbb{R}^n,$$

wenn $f'(x^*; x - x^*) \geqslant 0$ gilt für alle $x \in C$, also

$$(x^* - p)^\mathsf{T}(x - x^*) \geqslant 0 \quad \text{für alle } x \in C.$$

Das ist genau die Bedingung, die wir bereits aus Satz 15.3 als notwendige und hinreichende Bedingung kennen, vgl. (15.4).

Quizfrage 18.2: Ist die Regularitätsbedingung aus Satz 18.5 für die Projektionsaufgabe erfüllt?

Quizfrage 18.3: Wofür benötigt man denn hier die Abgeschlossenheit von C?

Ende der Vorlesung 27

Ende der Woche 14

§ 19 Ausblick: Verfahren der konvexen Optimierung

Literatur: Geiger, Kanzow, 2002, Kapitel 6

Wir geben in diesem Abschnitt einen Ausblick auf Verfahren für *allgemeine* konvexe Optimierungsaufgaben ohne weitere Struktur. Genauer betrachten wir Aufgaben der Form

$$\text{Minimiere} \quad f(x) \quad \text{über } x \in \mathbb{R}^n \tag{19.1}$$

mit konvexer Zielfunktion $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Für Aufgaben mit mehr Struktur, etwa

$$\text{Minimiere} \quad g(Ax) + h(x), \tag{19.2}$$

wobei $g\colon \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ und $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvexe Funktionen sind sowie $A \in \mathbb{R}^{m\times n}$, gibt es geeignetere Verfahren, die diese Struktur ausnutzen. Diese verwenden dann i. d. R. eine andere, äquivalente Formulierung als Sattelpunktaufgabe

$$\text{Minimiere} \quad \max_{p\in\mathbb{R}^m} p^\mathsf{T} A x + h(x) - g^*(p)$$

bzw. die zugehörigen Optimalitätsbedingungen in der Form

$$\begin{aligned} Ax &\in \partial g^*(p), \\ -A^\mathsf{T} p &\in \partial h(x), \end{aligned}$$

wobei g^* die sogenannte **Fenchel-konjugierte Funktion** (auch: **konvex konjugierte Funktion**) (englisch: *Fenchel conjugate function, convex conjugate function*) zu g ist. Mehr zu Fenchel-konjugierten Funktionen, dualen Aufgaben und Sattelpunktaufgaben, deren Optimalitätsbedingungen und darauf aufbauende Lösungsverfahren erfährt man in Vorlesungen zur konvexen Optimierung.

Die Richtung des steilsten Abstiegs

Warum benötigen wir überhaupt spezielle Verfahren für Optimierungsaufgaben (19.1), in denen die Zielfunktion f konvex, aber i. A. nicht diffbar ist? Wir wollen dies an einem Beispiel motivieren. Dazu führen wir zunächst die **Richtung des steilsten Abstiegs** von f im Punkt x_0, äquivalent zu (4.9), als Lösung der Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & f'(x_0; d) \quad \text{über } d \in \mathbb{R}^n \\ \text{unter} \quad & \|d\| \leqslant 1 \end{aligned} \tag{19.3}$$

ein.[11] Da die Zielfunktion $d \mapsto f'(x_0; d)$ nach Lemma 16.24 stetig und $\overline{B_1(0)}$ kompakt ist, existiert nach dem Satz von Weierstraß bzw. Satz 1.9 eine globale Lösung von (19.3). Aufgrund der Konvexität der Zielfunktion (Satz 16.16) existieren keine lokalen Minimierer, die nicht gleichzeitig globale Minimierer sind.

Lemma 19.1 (Eindeutigkeit der Richtung des steilsten Abstiegs)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und $x_0 \in \mathbb{R}^n$ ein Punkt, an dem $f(x_0)$ endlich ist. Falls ein $d \in \mathbb{R}^n$ existiert, sodass $f'(x_0; d) < 0$ ist, dann ist die Lösung von (19.3) eindeutig.

[11] Statt der Euklidischen könnten wir wieder auch eine andere Norm betrachten.

Beweis. Der Beweis ist Inhalt von Hausaufgabe 14.4. □

Wir werden in Satz 19.3 die Richtung des steilsten Abstiegs bzgl. der Euklidischen Norm charakterisieren. Dazu benötigen wir folgendes Hilfsresultat.

Lemma 19.2 (**Minimax-Lemma** (englisch: *minimal lemma*))**.**
Es seien $M_1, M_2 \subseteq \mathbb{R}^n$ beliebige nichtleere Mengen. Dann gilt

$$\sup_{x\in M_1} \inf_{y\in M_2} x^\mathsf{T} y \leqslant \inf_{y\in M_2} \sup_{x\in M_1} x^\mathsf{T} y. \tag{19.4}$$

Beweis.

$$\begin{aligned}
& x^\mathsf{T} y \leqslant \sup_{\overline{x}\in M_1} \overline{x}^\mathsf{T} y && \text{für alle } x \in M_1,\ y \in M_2 \\
\Rightarrow\quad & \inf_{\overline{y}\in M_2} x^\mathsf{T} \overline{y} \leqslant \sup_{\overline{x}\in M_1} \overline{x}^\mathsf{T} y && \text{für alle } x \in M_1,\ y \in M_2 \\
\Rightarrow\quad & \inf_{\overline{y}\in M_2} x^\mathsf{T} \overline{y} \leqslant \inf_{\overline{y}\in M_2} \sup_{\overline{x}\in M_1} \overline{x}^\mathsf{T} \overline{y} && \text{für alle } x \in M_1 \\
\Rightarrow\quad & \sup_{\overline{x}\in M_1} \inf_{\overline{y}\in M_2} \overline{x}^\mathsf{T} \overline{y} \leqslant \inf_{\overline{y}\in M_2} \sup_{\overline{x}\in M_1} \overline{x}^\mathsf{T} \overline{y}. &&
\end{aligned}$$

□

Sogenannte **Minimax-Theoreme** (englisch: *minimal theorems*) beschäftigen sich mit der Frage, unter welchen Voraussetzungen in (19.4) die Gleichheit gilt.[12]

Satz 19.3 (Richtung des steilsten Abstiegs bzgl. der Euklidischen Norm)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex und $x_0 \in \mathbb{R}^n$ *kein* globaler Minimierer von f. Dann ist die eindeutige Lösung der Aufgabe (19.3) des steilsten Abstiegs gegeben durch

$$d := -\frac{g}{\|g\|}, \quad \text{wobei } g := \operatorname{proj}_{\partial f(x_0)}(0) \text{ ist.} \tag{19.5}$$

Es gilt $f'(x_0; d) = -\|g\|$.

Beweis. Wir zeigen zunächst die Eindeutigkeit der Lösung von (19.3). Da x_0 kein globaler und damit auch kein lokaler Minimierer von f ist, gibt es ein $d \in \mathbb{R}^n$ mit $f'(x_0; d) < 0$. (**Quizfrage 19.1:** Genaue Begründung?) Die Eindeutigkeit der Richtung des steilsten Abstiegs folgt nun aus Lemma 19.1.

[12] Minimax-Theoreme haben Anwendungen u. a. in der Spieltheorie. Ein klassisches Resultat ist das von von Neumann, 1928, das besagt: Wenn $C_1 \subseteq \mathbb{R}^n$ und $C_2 \subseteq \mathbb{R}^m$ beide konvex und kompakt sind und $f\colon C_1 \times C_2 \to \mathbb{R}$ stetig und **konkav-konvex** (englisch: *concave-convex*) ist, also

$$\begin{aligned}
f(\cdot, y)\colon C_1 \to \mathbb{R} \text{ ist konkav für alle } y \in C_2, \\
f(x, \cdot)\colon C_2 \to \mathbb{R} \text{ ist konvex für alle } x \in C_1,
\end{aligned}$$

dann gilt

$$\max_{x\in C_1} \min_{y\in C_2} f(x, y) = \min_{y\in C_2} \max_{x\in C_1} f(x, y).$$

Dieses Resultat ist mittlerweile in zahlreiche Richtungen verallgemeinert worden.

Da x_0 kein globaler Minimierer ist, gilt weiter $0 \notin \partial f(x_0)$ (Satz 18.1). Weil weiterhin $\partial f(x_0)$ nichtleer ist (Satz 16.8) sowie abgeschlossen und konvex (Satz 16.4), existiert $g := \operatorname{proj}_{\partial f(x_0)}(0) \neq 0$ nach Satz 15.3 und ist charakterisiert durch

$$(g-0)^\mathsf{T}(s-g) \geqslant 0 \quad \text{für alle } s \in \partial f(x_0).$$

Mit der Definition $d := -g/\|g\|$ folgt also

$$s^\mathsf{T} d \leqslant d^\mathsf{T} g = -\|g\| \quad \text{für alle } s \in \partial f(x_0).$$

Nach Satz 16.18 Aussage (ii) folgt also (**Quizfrage 19.2:** Warum ist (16.24) anwendbar, und warum ist das Supremum hier ein Maximum?)

$$f'(x_0; d) = \max\{s^\mathsf{T} d \mid s \in \partial f(x_0)\} \leqslant -\|g\|.$$

Wir können also den Infimalwert der Aufgabe (19.3) schreiben als

$$\min_{\|d\| \leqslant 1} f'(x_0; d) = \min_{\|d\| \leqslant 1} \max_{s \in \partial f(x_0)} s^\mathsf{T} d \leqslant -\|g\|. \tag{19.6}$$

Wir betrachten jetzt die Aufgabe nach Vertauschung von min und max, die den Infimalwert

$$\max_{s \in \partial f(x_0)} \min_{\|d\| \leqslant 1} s^\mathsf{T} d$$

besitzt. Zunächst ist noch unklar, ob wir hier überhaupt max bzw. min schreiben dürfen oder sup bzw. inf verwenden müssen. Die innere Aufgabe hat, für gegebenes $s \in \partial f(x_0)$, aber offenbar $d = -s/\|s\|$ als eindeutigen Minimierer mit Infimalwert $-\|s\|$. (**Quizfrage 19.3:** Warum ist $s \neq 0$?) Wir können die Aufgabe also auch schreiben als

$$\text{Maximiere} \quad -\|s\| \quad \text{über } s \in \partial f(x_0). \tag{19.7}$$

Da $\partial f(x_0)$ nichtleer und kompakt ist und $-\|s\|$ stetig, wird das Supremum als Maximum angenommen. Die Aufgabe ist weiter äquivalent zu

$$\text{Minimiere} \quad \|s - 0\| \quad \text{über } s \in \partial f(x_0),$$

deren (eindeutige) Lösung wir bereits kennen: $s = \operatorname{proj}_{\partial f(x_0)}(0) = g$. Der Infimalwert von (19.7) ist also $-\|g\|$.

Wir fassen zusammen und erhalten unter Zuhilfenahme des Minimax-Lemmas 19.2:

$$-\|g\| = \max_{s \in \partial f(x_0)} \min_{\|d\| \leqslant 1} s^\mathsf{T} d \leqslant \min_{\|d\| \leqslant 1} \max_{s \in \partial f(x_0)} s^\mathsf{T} d \leqslant -\|g\|.$$

Es gilt also überall Gleichheit, und aus (19.6) folgt wie behauptet $f'(x_0; d) = -\|g\|$. □

Beispiel 19.4 (Richtung des steilsten Abstiegs für die 1-Norm und die ∞-Norm)**.**

(i) Wir betrachten $f(x) := \|x\|_1$ auf $\mathbb{R}^2$. Das Subdifferential wurde in Beispiel 16.5 charakterisiert. Im Punkt $x^{(1)} = (1,1)^\mathsf{T}$ ist f diffbar und daher

$$g := \operatorname{proj}_{\partial f(x^{(1)})}(0) = \nabla f(x^{(1)}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{und} \quad d = -\frac{g}{\|g\|_2} = -\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

die Richtung des steilsten Abstiegs.

Im Punkt $x^{(2)} = (1,0)^\mathsf{T}$ gilt $\partial f(x^{(2)}) = \{1\} \times [-1,1]$. Es gilt also

$$g := \operatorname{proj}_{\partial f(x^{(2)})}(0) = (1,0)^\mathsf{T} \quad \text{und} \quad d = -\frac{g}{\|g\|_2} = -(1,0)^\mathsf{T}$$

für die Richtung des steilsten Abstiegs.

(*ii*) Wir betrachten nun $f(x) := \|x\|_\infty$ auf $\mathbb{R}^2$. Das Subdifferential wurde ebenfalls in Beispiel 16.5 charakterisiert.

Im Punkt $x^{(1)} = (1,1)^\mathsf{T}$ gilt $\partial f(x^{(1)}) = \Delta_2 = \{s \in \mathbb{R}^2 \mid s \geqslant 0,\ s_1 + s_2 = 1\}$. Es gilt also

$$g := \operatorname{proj}_{\partial f(x^{(1)})}(0) = (1/2, 1/2)^\mathsf{T} \quad \text{und} \quad d = -\frac{g}{\|g\|_2} = -\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}$$

für die Richtung des steilsten Abstiegs.

Im Punkt $x^{(2)} = (1,0)^\mathsf{T}$ ist die Funktion f diffbar, und daher gilt

$$g := \operatorname{proj}_{\partial f(x^{(2)})}(0) = \nabla f(x^{(2)}) = \begin{pmatrix}1\\0\end{pmatrix} \quad \text{und} \quad d = -\frac{g}{\|g\|_2} = -\begin{pmatrix}1\\0\end{pmatrix}$$

die Richtung des steilsten Abstiegs. △

Es folgt nun das eingangs erwähnte Beispiel, das zeigt, dass man unter Verwendung der Richtung des steilsten Abstiegs (19.5) selbst mit exakter Liniensuche ein Abstiegsverfahren erhält, das gegen „uninteressante“ Punkte konvergieren kann.

Beispiel 19.5 (aus Bonnans u. a., 2003, Example 9.1)**.**
Die konvexe Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ sei gegeben durch

$$f(x) := \max\{f^{(0)}(x), f^{(-1)}(x), f^{(-2)}(x), f^{(1)}(x), f^{(2)}(x)\}$$

mit

$$f^{(0)}(x) := -100, \quad f^{(\pm 1)}(x) := 3\,x_1 \pm 2\,x_2, \quad f^{(\pm 2)}(x) := 2\,x_1 \pm 5\,x_2.$$

Der Infimalwert von $f^* = -100$ wird auf der konvexen Menge der globalen Minimierer $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \leqslant -50,\ |x_2| \leqslant 0.4\,|x_1| + 20\}$ angenommen, vgl. Abbildung 19.1.

Man kann zeigen, dass das Verfahren des steilsten Abstiegs mit exakter Liniensuche, gestartet bei $x^{(0)} = (9,-3)^\mathsf{T}$, die Iterationsfolge

$$x^{(k)} = \begin{pmatrix} 3^{2-k} \\ (-1)^{k+1}\,3^{1-k} \end{pmatrix}$$

erzeugt. Das Subdifferential ist jeweils

$$\partial f(x^{(k)}) = \operatorname{conv}\left\{\begin{pmatrix}3\\(-1)^{k+1}2\end{pmatrix}, \begin{pmatrix}2\\(-1)^{k+1}5\end{pmatrix}\right\}.$$

Wir erhalten also

$$g^{(k)} = \operatorname{proj}_{\partial f(x^{(k)})}(0) = \begin{pmatrix}3\\(-1)^{k+1}2\end{pmatrix}$$

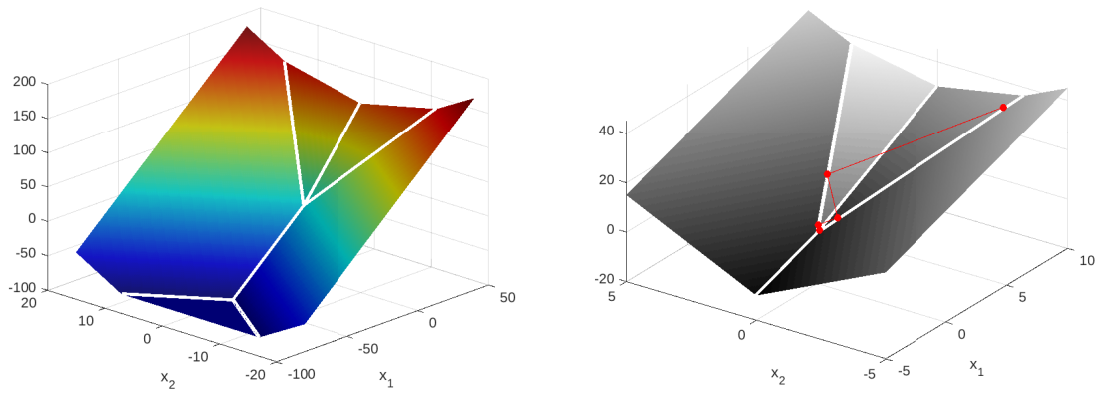


Abbildung 19.1.: Darstellung der Funktion aus Beispiel 19.5 mit Linien der Nichtdifferenzierbarkeit (links) und Ausschnitt mit den ersten Iterierten des Verfahrens des steilsten Abstiegs mit exakter Liniensuche, ausgehend von $x^{(0)} = (9, -3)^\mathsf{T}$ (rechts).

mit $\|g^{(k)}\| = \sqrt{13}$ und Suchrichtungen $d^{(k)} = -g^{(k)}/\|g^{(k)}\|$.

Die Funktionswerte $f(x^{(k)}) = 11 \cdot 3^{1-k}$ sind streng monoton fallend. Die Folge $x^{(k)}$ konvergiert (Q-linear) gegen den Punkt $x^* = (0, 0)^\mathsf{T}$ mit $f(x^*) = 0$, der ein nicht-optimaler, „nichtglatter“ Punkt der Zielfunktion f ist. △

Die Schwierigkeiten lassen sich auf folgende Beobachtungen für reellwertige, konvexe Funktionen f zurückführen:

(1) Das Subdifferential $\partial f\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ ist als mengenwertige Funktion zwar **außerhalbstetig** (englisch: *outer semicontinuous*), d. h.:

Für alle Folgen $x^{(k)} \to x$ und alle Folgen $s^{(k)} \in \partial f(x^{(k)})$ mit $s^{(k)} \to s$ gilt $s \in \partial f(x)$,

(**Quizfrage 19.4:** Beweis?) jedoch *nicht* **innerhalbstetig** (englisch: *inner semicontinuous*). Die Innerhalbstetigkeit würde bedeuten:

Für alle Folgen $x^{(k)} \to x$ und alle $s \in \partial f(x)$ gibt es eine Folge $s^{(k)} \in \partial f(x^{(k)})$ mit $s^{(k)} \to s$.

(**Quizfrage 19.5:** Beispiel, das die fehlende Innerhalbstetigkeit von ∂f zeigt?)

(2) Die fehlende Innerhalbstetigkeit von ∂f bedeutet, dass bereits kleine Änderungen in x große Änderungen in $\partial f(x)$ hervorrufen können. Da die Richtung des steilsten Abstiegs an einer Iterierten $x^{(k)}$ gemäß (19.5) aus $\partial f(x^{(k)})$ berechnet wird, ergibt sich kein stabiles Verfahren.

Weitere praktische Nachteile von Verfahren des steilsten Abstiegs für (19.1) sind:

(3) Die Bestimmung von $g^{(k)} = \mathrm{proj}_{\partial f(x^{(k)})}(0)$ und damit die Bestimmung der Suchrichtung $d^{(k)} = -g^{(k)}/\|g^{(k)}\|$ erfordert i. W. die Bestimmung des *gesamten* Subdifferentials.

Wir stellen im Anhang B einen einfachen Vertreter der Klasse der **Bundle-Verfahren** (englisch: *bundle methods*) vor, einer Familie leistungsfähiger Verfahren für *allgemeine* konvexe Optimierungsaufgaben (19.1), die die oben genannten Nachteile nicht aufweisen. Für Aufgaben mit mehr Struktur wie (19.2) würde man wie gesagt angepasste Verfahren benutzen, siehe Vorlesungen zur konvexen Optimierung.

Kapitel A Innere-Punkte-Verfahren für lineare Optimierungsaufgaben

Literatur: Geiger, Kanzow, 2002, Kapitel 4.1

§ 20 Innere-Punkte-Verfahren

Innere-Punkte-Verfahren (**IP-Verfahren**) (englisch: *interior point methods*) bewegen sich im Gegensatz zum Simplex-Verfahren im relativen Inneren des zulässigen Polyeders zu einer Lösung. Sie sind eine Alternative insbesondere für hochdimensionale LP und außerdem verallgemeinerbar auf allgemeine nichtlineare Optimierungsaufgaben. Obwohl es bereits frühere Ansätze gab, gelangten Innere-Punkte-Verfahren durch die Arbeiten Karmarkar, 1984a und Karmarkar, 1984b zu einem Durchbruch. Eines der Hauptargumente für Innere-Punkte-Verfahren ist, dass sie — im Gegensatz zu Simplex-Verfahren — lineare Programme garantiert in polynomialer Zeit (der Problemgröße) lösen.

Wir betrachten wieder das primale LP in Normalform

$$\left.\begin{aligned} \text{Minimiere} \quad & c^\mathsf{T} x \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} \quad & A x = b \\ \text{und} \quad & x \geqslant 0 \end{aligned}\right\} \tag{20.1}$$

mit dem dualen LP

$$\left.\begin{aligned} \text{Maximiere} \quad & b^\mathsf{T} \lambda \quad \text{über } (\lambda, s) \in \mathbb{R}^m \times \mathbb{R}^n \\ \text{sodass} \quad & A^\mathsf{T} \lambda + \mu = c \\ \text{und} \quad & \mu \geqslant 0 \end{aligned}\right\} \tag{20.2}$$

und den notwendigen und hinreichenden Optimalitätsbedingungen

$$\left.\begin{aligned} A^\mathsf{T}\lambda + \mu = c, \quad & \mu \geqslant 0 \quad && \text{duale Zulässigkeit} \\ A x = b, \quad & x \geqslant 0 && \text{primale Zulässigkeit} \\ x_i \mu_i = 0, \quad & i = 1, \dots, n && \text{Komplementarität.} \end{aligned}\right\} \tag{20.3}$$

Wir wissen nach Satz 8.5, dass eine Lösung (x, λ, μ) von (20.3) gleichzeitig Lösungen von (20.1) und (20.2) liefert. Wir konzentrieren uns daher nun auf sogenannte **primal-duale Innere-Punkte-Verfahren** zur Lösung von (20.3).

Wir betrachten dazu folgende Störung des Optimalitätssystems (20.3):

$$\left.\begin{aligned} A^\mathsf{T}\lambda + \mu = c, \quad & \mu > 0, \\ A x = b, \quad & x > 0, \\ x_i \mu_i = \tau, \quad & i = 1, \dots, n \end{aligned}\right\} \tag{20.4}$$

mit einem Parameter $\tau > 0$. Wir bezeichnen eine Lösung von (20.4) mit $(x^{(\tau)}, \lambda^{(\tau)}, \mu^{(\tau)})$. (Diese muss jedoch nicht immer existieren, dazu später mehr.)

Falls existent, so heißt die Abbildung

$$\tau \mapsto (x^{(\tau)}, \lambda^{(\tau)}, \mu^{(\tau)})$$

der **zentrale Pfad** (englisch: *central path*), und (20.4) heißen **Zentraler-Pfad-Bedingungen** (englisch: *central path conditions*). Die Idee der Innere-Punkte-Verfahren besteht darin, den Pfad für $\tau \searrow 0$ zu verfolgen.

Für eine alternativen Betrachtungsweise führen wir das zu (20.1) gehörige (primale) logarithmische **Barriere-Problem** (englisch: *barrier problem*) ein:

$$\left.\begin{aligned} \text{Minimiere} \quad & c^\mathsf{T} x - \tau \sum_{i=1}^{n} \ln(x_i) \quad \text{über } x \in \mathbb{R}^n \\ \text{sodass} \quad & A\,x = b \\ \text{und} \quad & x > 0. \end{aligned}\right\} \tag{20.5}$$

Das zum dualen Problem (20.2) gehörige Barriere-Problem lautet analog

$$\left.\begin{aligned} \text{Maximiere} \quad & b^\mathsf{T}\lambda + \tau \sum_{i=1}^{n} \ln(\mu_i) \quad \text{über } (\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^n \\ \text{sodass} \quad & A^\mathsf{T}\lambda + \mu = c \\ \text{und} \quad & \mu > 0. \end{aligned}\right\} \tag{20.6}$$

Bemerkung 20.1. Es handelt sich bei (20.5) und (20.6) um konvexe Optimierungsaufgaben. Nach dem Hauptsatz der konvexen Optimierung 14.2 ist also jeder lokale Minimierer von (20.5) bereits ein globaler Minimierer ist und jeder lokale Maximierer von (20.6) bereits ein globaler Maximierer. △

Den Zusammenhang zwischen (20.5), (20.6) und dem gestörten Optimalitätssystem (20.4) stellt der folgende Satz her, der ein Analogon zu Satz 8.5 ist.

Satz 20.2 (Notwendige und hinreichende Optimalitätsbedingungen)**.**
Es sei $\tau > 0$ gegeben.

- (i) Ist $x^{(\tau)}$ eine Lösung des primalen Barriere-Problems (20.5), dann existieren $(\lambda^{(\tau)}, \mu^{(\tau)})$, sodass $(x^{(\tau)}, \lambda^{(\tau)}, \mu^{(\tau)})$ das System (20.4) erfüllt.
- (ii) Ist $(\lambda^{(\tau)}, \mu^{(\tau)})$ eine Lösung des dualen Barriere-Problems (20.6), dann existiert $x^{(\tau)}$, sodass $(x^{(\tau)}, \lambda^{(\tau)}, \mu^{(\tau)})$ das System (20.4) erfüllt.
- (iii) Erfüllt $(x^{(\tau)}, \lambda^{(\tau)}, \mu^{(\tau)})$ das System (20.4), dann ist $x^{(\tau)}$ eine Lösung von (20.5), und $(\lambda^{(\tau)}, \mu^{(\tau)})$ ist eine Lösung von (20.6).

(**Quizfrage 20.1:** Beweis von Satz 20.2?)

Beachte: (20.4) ist also nicht nur die Störung des Optimalitätssystems (20.3), sondern seinerseits das notwendige und hinreichende Optimalitätssystem für die Barriere-Aufgaben (20.5) und (20.6).

Wir untersuchen jetzt, wann die Aufgaben (20.4)–(20.6) lösbar sind.

Beispiel 20.3 (Unlösbares Pfad-Problem)**.**
Wir betrachten die primale Aufgabe

$$\left.\begin{aligned} \text{Minimiere} \quad & x_1 + x_2 \quad \text{über } x \in \mathbb{R}^2 \\ \text{sodass} \quad & x_1 + x_2 = 0 \\ \text{und} \quad & x \geqslant 0 \end{aligned}\right\}$$

mit der eindeutigen Lösung $x^* = (0,0)^\mathsf{T}$. Im zugehörigen Barriere-Problem (20.4) widersprechen sich jedoch $x_1 + x_2 = 0$ und $x > 0$, sodass (20.4) keine Lösung besitzt. Wegen Satz 20.2 besitzen dann auch (20.5) und (20.6) keine Lösung. △

Definition 20.4 (Primal-dual zulässige Menge)**.**
Die Menge

$$F := \{(x, \lambda, \mu) \mid A\,x = b,\ A^\mathsf{T}\lambda + \mu = c,\ x \geqslant 0,\ \mu \geqslant 0\}$$

heißt die **primal-dual zulässige Menge** (englisch: *primal-dual feasible set*) und

$$F_0 := \{(x, \lambda, \mu) \mid A\,x = b,\ A^\mathsf{T}\lambda + \mu = c,\ x > 0,\ \mu > 0\}$$

die **primal-dual strikt zulässige Menge** (englisch: *primal-dual strictly feasible set*) für die Aufgaben (20.1)–(20.3). △

Aufgrund von Satz 20.2 ist $F_0 \neq \emptyset$ notwendig dafür, dass (20.4) und damit (20.5) und (20.6) Lösungen besitzen. Es ist aber auch hinreichend:

Satz 20.5 (Existenz einer Lösung für das primale Barriere-Problem)**.**
Die folgenden Aussagen sind äquivalent:

(*i*) Die primal-dual strikt zulässige Menge F_0 ist nichtleer.
(*ii*) Das primale Barriere-Problem (20.5) besitzt für jedes $\tau > 0$ eine (globale) Lösung $x^{(\tau)}$.
(*iii*) Das duale Barriere-Problem (20.6) besitzt für jedes $\tau > 0$ eine (globale) Lösung $(\lambda^{(\tau)}, \mu^{(\tau)})$.
(*iv*) Die Optimalitätsbedingungen (20.4) besitzen für jedes $\tau > 0$ eine Lösung $(x^{(\tau)}, \lambda^{(\tau)}, \mu^{(\tau)})$.

Beweis. Die Äquivalenz der Aussagen (*ii*) bis (*iv*) untereinander ist nach Satz 20.2 bereits bekannt.

Aussage (*iv*) ⇒ Aussage (*i*): Aus der Lösbarkeit von (20.4) folgt sofort, dass F_0 nichtleer ist.

Es bleibt Aussage (*i*) ⇒ Aussage (*ii*) zu zeigen. Es sei dafür $\tau > 0$ beliebig und ein $(x_0, \lambda_0, \mu_0) \in F_0$ gegeben, es gilt also

$$A^\mathsf{T}\lambda_0 + \mu_0 = c, \quad A\,x_0 = b, \quad x_0 > 0, \quad \mu_0 > 0. \tag{20.7}$$

Wir bezeichnen mit

$$f^{(\tau)}(x) := c^\mathsf{T}x - \tau \sum_{i=1}^{n} \ln(x_i)$$

die Zielfunktion in (20.5). Wir werden zeigen, dass die Sublevelmenge

$$L := \left\{x \in \mathbb{R}^n \,\middle|\, A\,x = b,\ x \geqslant 0,\ f^{(\tau)}(x) \leqslant f^{(\tau)}(x_0)\right\}$$

kompakt ist. (Nichtleer ist sie wegen $x_0 \in L$.) Das Barriere-Problem (20.5) ist daher äquivalent zur Minimierung der stetigen Funktion $f^{(\tau)}$ über der kompakten Menge L, besitzt also nach dem Satz von Weierstraß bzw. Satz 1.9 einen globalen Minimierer. Die eigentlich benötigte strengere Bedingung $x > 0$ in der Definition von L ergibt sich automatisch aus $f^{(\tau)}(x) \leqslant f^{(\tau)}(x_0)$.

Offenbar ist L abgeschlossen. (**Quizfrage 20.2:** Warum?) Für $x \in L$ folgt aus (20.7)

$$\begin{aligned} f^{(\tau)}(x) + \tau \sum_{i=1}^{n} \ln(x_i) &= c^\mathsf{T} x \\ &= c^\mathsf{T} x - \lambda_0^\mathsf{T}(A\,x - b) \\ &= c^\mathsf{T} x - x^\mathsf{T} A^\mathsf{T} \lambda_0 + b^\mathsf{T} \lambda_0 \\ &= c^\mathsf{T} x - x^\mathsf{T}(c - \mu_0) + b^\mathsf{T} \lambda_0 \\ &= x^\mathsf{T} \mu_0 + b^\mathsf{T} \lambda_0. \end{aligned}$$

Daher gilt für $x > 0$:

$$\begin{aligned} & f^{(\tau)}(x) \leqslant f^{(\tau)}(x_0) \\ \Leftrightarrow \quad & x^\mathsf{T}\mu_0 + b^\mathsf{T}\lambda_0 - \tau \sum_{i=1}^{n} \ln(x_i) \leqslant f^{(\tau)}(x_0) \\ \Leftrightarrow \quad & \sum_{i=1}^{n} \big[x_i \mu_{0,i} - \tau \ln(x_i)\big] \leqslant f^{(\tau)}(x_0) - b^\mathsf{T}\lambda_0 =: \text{const.} \end{aligned}$$

Die Funktionen

$$x_i \mapsto x_i\,\mu_{0,i} - \tau \ln(x_i)$$

sind auf $\mathbb{R}_{>0}$ wegen $\mu_{0,i} > 0$ nach unten beschränkt und konvergieren gegen ∞ für $x_i \to \infty$. Daher ist L auch beschränkt, also kompakt. □

Folgerung 20.6 (Existenz und Eigenschaften des zentralen Pfades)**.**
Die strikt zulässige Menge F_0 sei nichtleer. Dann besitzen die Zentraler-Pfad-Bedingungen (20.4) für jedes $\tau > 0$ eine Lösung $(x^{(\tau)}, \lambda^{(\tau)}, \mu^{(\tau)})$. Dabei sind die Komponente $x^{(\tau)}$ und $\mu^{(\tau)}$ eindeutig bestimmt. Besitzt A vollen Zeilenrang, gilt also $\operatorname{Rang}(A) = m$, so ist auch $\lambda^{(\tau)}$ eindeutig.

Beweis. Die Existenz von $(x^{(\tau)}, \lambda^{(\tau)}, \mu^{(\tau)})$ folgt aus Satz 20.5.

Zur Eindeutigkeit: Das primale Barriere-Problem (20.5) besitzt eine strikt konvexe Zielfunktion (Definition 13.9), und die zulässige Menge

$$\{x \in \mathbb{R}^n \mid A\,x = b,\ x > 0\}$$

ist konvex. Daher ist $x^{(\tau)}$ eindeutig bestimmt, siehe Satz 14.2. Aufgrund der Bedingungen $x_i^{(\tau)} \mu_i^{(\tau)} = \tau$ für $i = 1, \ldots, n$ ist damit auch $\mu^{(\tau)}$ eindeutig bestimmt.

Besitzt A vollen Zeilenrang, dann ist $AA^\mathsf{T} \in \mathbb{R}^{m \times m}$ invertierbar, und aus $A^\mathsf{T}\lambda^{(\tau)} + \mu^{(\tau)} = c$ folgt

$$\lambda^{(\tau)} = (AA^\mathsf{T})^{-1} A (c - \mu^{(\tau)}).$$

Damit ist in diesem Fall auch $\lambda^{(\tau)}$ eindeutig bestimmt. □

Definition 20.7 (Strikt komplementäre Lösung).
Eine Lösung (x^*, λ^*, μ^*) der Optimalitätsbedingungen (8.9) heißt **strikt komplementär** (englisch: *strictly complementary*), wenn für alle $i = 1, \dots, n$ entweder $x_i^* = 0$ oder $\mu_i^* = 0$ gilt. △

Satz 20.8 (Konvergenz für $\tau \searrow 0$).
Die strikt zulässige Menge F_0 sei nichtleer, und es gelte $\tau^{(k)} \searrow 0$. Es sei $(x^{(k)}, \mu^{(k)}, \lambda^{(k)})$ eine Lösung von (20.4) für $\tau = \tau^{(k)}$. Dann ist die Folge $\left(x^{(k)}, \mu^{(k)}\right)$ beschränkt und besitzt daher eine konvergente Teilfolge. Jeder Häufungspunkt (Grenzwert einer Teilfolge) gehört zu einer strikt komplementären Lösung (x^*, λ^*, μ^*) von (20.3).[1]

Beweis. Es sei $(x_0, \lambda_0, \mu_0) \in F_0$. Es gilt

$$(x^{(k)} - x_0)^\mathsf{T}(\mu^{(k)} - \mu_0) = (x^{(k)} - x_0)^\mathsf{T} A^\mathsf{T}(\lambda_0 - \lambda^{(k)}) = (b - b)^\mathsf{T}(\lambda_0 - \lambda^{(k)}) = 0$$

$$\Rightarrow \quad \underbrace{x_0^\mathsf{T}}_{>0}\mu^{(k)} + \left(x^{(k)}\right)^\mathsf{T}\underbrace{\mu_0}_{>0} = \underbrace{\left(x^{(k)}\right)^\mathsf{T}\mu^{(k)}}_{=\tau^{(k)}n} + \underbrace{x_0^\mathsf{T}\mu_0}_{=:c>0}$$

$$\Rightarrow \quad 0 \leqslant x_0^\mathsf{T}\mu^{(k)} + \left(x^{(k)}\right)^\mathsf{T}\mu_0 = \tau^{(k)}n + c \leqslant \overline{\tau}\, n + c \quad \text{für alle } k \in \mathbb{N}$$

$$\Rightarrow \quad \|x^{(k)}\| \text{ und } \|\mu^{(k)}\| \text{ sind beschränkt.}$$

Also existieren konvergente Teilfolgen

$$x^{(k^{(\ell)})} \to x^* \geqslant 0, \quad \mu^{(k^{(\ell)})} \to \mu^* \geqslant 0,$$

und es gilt $Ax^{(k^{(\ell)})} = b$, also auch $Ax^* = b$, sowie $\left(x^{(k^{(\ell)})}\right)^\mathsf{T}\mu^{(k^{(\ell)})} = \tau^{(k^{(\ell)})}\, n \searrow 0$, also auch $(x^*)^\mathsf{T}\mu^* = 0$.

Die Folge $\lambda^{(k^{(\ell)})}$ erfüllt $A^\mathsf{T}\lambda^{(k^{(\ell)})} = c - \mu^{(k^{(\ell)})}$, d. h., $c - \mu^{(k^{(\ell)})} \in \text{Bild}\,(A^\mathsf{T})$ für alle $m \in \mathbb{N}$. Da Bild (A^T) als Unterraum abgeschlossen ist, liegt auch der Grenzwert $c - \mu^* \in \text{Bild}\,(A^\mathsf{T})$, d. h., es existiert λ^* mit $A^\mathsf{T}\lambda^* + \mu^* = c$.[2] Mit anderen Worten: (x^*, λ^*, μ^*) erfüllt (8.9).

Zur strikten Komplementarität:

$$(x^{(k^{(\ell)})} - x^*)^\mathsf{T}(\mu^{(k^{(\ell)})} - \mu^*) = (x^{(k^{(\ell)})} - x^*)^\mathsf{T} A^\mathsf{T}(\lambda^* - \lambda^{(k^{(\ell)})}) = (b - b)^\mathsf{T}(\lambda^* - \lambda^{(k^{(\ell)})}) = 0$$

$$\Rightarrow \quad (x^*)^\mathsf{T}\mu^{(k^{(\ell)})} + \left(x^{(k^{(\ell)})}\right)^\mathsf{T}\mu^* = \underbrace{\left(x^{(k^{(\ell)})}\right)^\mathsf{T}\mu^{(k^{(\ell)})}}_{=\tau^{(k^{(\ell)})}n} + \underbrace{(x^*)^\mathsf{T}\mu^*}_{=0} = \tau^{(k^{(\ell)})}n$$

$$\Rightarrow \quad \sum_{i=1}^n \frac{x_i^*}{x_i^{(k^{(\ell)})}} + \sum_{i=1}^n \frac{\mu_i^*}{\mu_i^{(k^{(\ell)})}} = n \quad \text{wegen } \mu_i^{(k^{(\ell)})} = \frac{\tau^{(k^{(\ell)})}}{x_i^{(k^{(\ell)})}} \text{ und } x_i^{(k^{(\ell)})} = \frac{\tau^{(k^{(\ell)})}}{\mu_i^{(k^{(\ell)})}}.$$

Die $2n$ Quotienten in den Summen sind jeweils entweder $= 0$ oder konvergieren für $\ell \to \infty$ gegen 1. Daraus folgt entweder $x_i^* = 0$ oder $\mu_i^* = 0$ für alle $i = 1, \dots, n$. □

[1] Nur solche kann man also überhaupt durch primal-duale IP-Verfahren erreichen.

[2] D. h., λ^* ist nicht notwendig Grenzwert der Folge $\lambda^{(k^{(\ell)})}$, sondern wird konstruiert.

Kapitel B Bundle-Verfahren

Wir entwickeln im Folgenden einen einfachen Vertreter der Klasse der **Bundle-Verfahren** (englisch: *bundle methods*), einer Familie leistungsfähiger Verfahren für *allgemeine* konvexe Optimierungsaufgaben (19.1).

§ 21 Das Bundle-Teilproblem

Bundle-Verfahren basieren auf der Idee, Subgradienten $s^{(j)} \in \partial f(x^{(j)})$ einer Reihe von Punkten $x^{(j)}$, $j = 0, 1, \ldots, k$ zu sammeln (daher der Name **Bündel**, englisch: *bundle*) und daraus ein stückweise lineares, konvexes Modell der Zielfunktion f zu erstellen:

$$f^{\mathrm{CP}}(x) := \max\left\{ f(x^{(j)}) + (s^{(j)})^{\mathsf{T}}(x - x^{(j)}) \,\middle|\, j = 0, 1, \ldots, k \right\} \tag{21.1}$$

Dieses sogenannte **Schnittebenenmodell** (englisch: *cutting plane model*) hat folgende Eigenschaften:

Lemma 21.1 (Eigenschaften des Schnittebenenmodells)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex, $x^{(j)} \in \mathbb{R}^n$ und $s^{(j)} \in \partial f(x^{(j)})$ für $j = 0, 1, \ldots, k$. Dann ist $f^{\mathrm{CP}}\colon \mathbb{R}^n \to \mathbb{R}$ konvex, und es gilt $f^{\mathrm{CP}}(x) \leqslant f(x)$ für alle $x \in \mathbb{R}^n$ sowie $f^{\mathrm{CP}}(x^{(j)}) = f(x^{(j)})$ für $j = 0, 1, \ldots, k$.

Beweis. f^{CP} ist als Maximum konvexer (linearer) Funktionen konvex nach Satz 13.18. Nach der Subgradientenungleichung (16.1) gilt

$$f(x) \geqslant f(x^{(j)}) + (s^{(j)})^{\mathsf{T}}(x - x^{(j)}) \quad \text{für alle } x \in \mathbb{R}^n \text{ und alle } j = 0, 1, \ldots, k,$$

also auch

$$f(x) \geqslant \max\left\{ f(x^{(j)}) + (s^{(j)})^{\mathsf{T}}(x - x^{(j)}) \,\middle|\, j = 0, 1, \ldots, k \right\} = f^{\mathrm{CP}}(x) \quad \text{für alle } x \in \mathbb{R}^n.$$

Speziell für $x = x^{(i)}$ folgt

$$\begin{aligned} f(x^{(i)}) &\geqslant \max\left\{ f(x^{(j)}) + (s^{(j)})^{\mathsf{T}}(x^{(i)} - x^{(j)}) \,\middle|\, j = 0, 1, \ldots, k \right\} = f^{\mathrm{CP}}(x^{(i)}) \\ &\geqslant f(x^{(i)}) + (s^{(i)})^{\mathsf{T}}(x^{(i)} - x^{(i)}) \\ &= f(x^{(i)}), \end{aligned}$$

also die Gleichheit. □

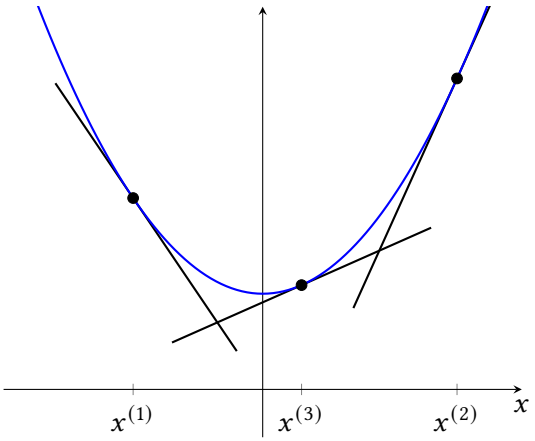


Abbildung 21.1.: Illustration des Schnittebenenmodells (21.1).

Das Schnittebenenmodell (21.1) und Varianten davon werden gleich als Zielfunktion eines Ersatzproblems für (19.1) verwendet. Zur Vereinfachung der Notation wählen wir einen Referenzpunkt $\overline{x}$ und ersetzen die Variable x durch die Richtungsvariable $d := x - \overline{x}$. Zur Abkürzung führen wir die **Linearisierungsfehler** (englisch: *linearization error*)

$$\overline{\alpha}^{(j)} := f(\overline{x}) - f(x^{(j)}) - (s^{(j)})^\mathsf{T}(\overline{x} - x^{(j)}) \geqslant 0, \quad j = 0, 1, \ldots, k \tag{21.2}$$

ein, die an der Stelle $\overline{x}$ die Differenz zwischen dem tatsächlichen Funktionswert und dem Wert des auf dem Subgradienten $s^{(j)}$ basierenden linearen Modells messen.

Beachte: Die Eigenschaft $\overline{\alpha}^{(j)} \geqslant 0$ folgt sofort aus Lemma 21.1.

Um das Modell (21.1) auf die Richtungsvariable d umzuschreiben, nutzen wir

$$\begin{aligned} f(x^{(j)}) + (s^{(j)})^\mathsf{T}(x - x^{(j)}) &= f(x^{(j)}) + (s^{(j)})^\mathsf{T}(x - \overline{x} + \overline{x} - x^{(j)}) \\ &= f(x^{(j)}) + (s^{(j)})^\mathsf{T}(d + \overline{x} - x^{(j)}) \\ &= (s^{(j)})^\mathsf{T} d - \overline{\alpha}^{(j)} + f(\overline{x}). \end{aligned}$$

Da bei der Minimierung bzgl. d die von j unabhängige Konstante $f(\overline{x})$ keine Rolle spielt, setzen wir als **Schnittebenenrichtungsmodell** (englisch: *cutting plane directional model*) jetzt die Funktion

$$m^{\mathrm{CP}}(d) := \max\left\{(s^{(j)})^\mathsf{T} d - \overline{\alpha}^{(j)} \,\middle|\, j = 0, 1, \ldots, k\right\} \tag{21.3}$$

an. Um die Minimierung dieser stückweise linearen, konvexen Funktion (21.3) durchzuführen, ist es günstig, zur sogenannten **Epigraph-Reformulierung** (englisch: *epigraph reformulation*) (21.4) überzugehen:

Lemma 21.2 (Epigraph-Reformulierung[1])**.**

[1]Wie sich aus dem Beweis ergibt, ist die Epigraph-Reformulierung immer möglich, wenn die Zielfunktion als das punktweise Maximum endlich vieler konvexer Funktionen definiert ist.

(*i*) Wenn $d \in \mathbb{R}^n$ ein (globaler) Minimierer von (21.3) ist, dann existiert ein $\xi \in \mathbb{R}$, sodass $(d, \xi) \in \mathbb{R}^n \times \mathbb{R}$ ein (globaler) Minimierer der Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & \xi \quad \text{über } (d, \xi) \in \mathbb{R}^n \times \mathbb{R} \\ \text{unter} \quad & (s^{(j)})^\mathsf{T} d - \overline{\alpha}^{(j)} \leqslant \xi, \quad j = 0, 1, \dots k \end{aligned} \tag{21.4}$$

ist.

(*ii*) Ist umgekehrt $(d, \xi) \in \mathbb{R}^n \times \mathbb{R}$ ein (globaler) Minimierer von (21.4), dann ist d ein (globaler) Minimierer von (21.3).

In beiden Fällen gilt $\xi = m^{\mathrm{CP}}(d)$.

Beweis. Der Beweis ist Inhalt von Aufgabe 0.65. □

Wir haben also die Nichtglattheit von (21.3) gegen Nebenbedingungen in (21.4) „eingetauscht“. Die Aufgabe (21.4) ist nun ein LP. Wir untersuchen jetzt das dazu duale LP.[2] Wir können leicht nachrechnen, dass dieses durch die (Minimierungs-)Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & \sum_{j=0}^{k} \overline{\alpha}^{(j)} \lambda_j \quad \text{über } \lambda \in \mathbb{R}^{k+1} \\ \text{unter} \quad & \lambda \geqslant 0 \text{ und } \sum_{j=0}^{k} \lambda_j = 1 \\ \text{sowie} \quad & \sum_{j=0}^{k} \lambda_j \, s^{(j)} = 0 \end{aligned} \tag{21.5}$$

gegeben ist.

Bemerkung 21.3 (Interpretation des dualen LPs (21.5))**.**

[2] Mit den Kenntnissen aus Kapitel 2 können wir leicht herleiten (siehe auch Hausaufgabe 6.1), dass das zu

$$\begin{aligned} \text{Minimiere} \quad & c^\mathsf{T} x \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & A x \leqslant b \end{aligned}$$

duale LP durch

$$\begin{aligned} \text{Maximiere} \quad & -b^\mathsf{T} \lambda \quad \text{über } \lambda \in \mathbb{R}^m \\ \text{unter} \quad & A^\mathsf{T} \lambda = -c \\ \text{und} \quad & \lambda \geqslant 0 \end{aligned}$$

gegeben ist. (**Quizfrage 21.1:** Begründung?) Die notwendigen und hinreichenden Optimalitätsbedingungen bestehen neben der primalen und der dualen Zulässigkeit aus der Komplementaritätsbedingung $\lambda^\mathsf{T}(A x - b) = 0$.

Die Zielfunktion des dualen LPs (21.5) lässt sich wie folgt interpretieren: Wir haben

$$\begin{aligned} 0 &\leqslant \sum_{j=0}^{k} \overline{\alpha}^{(j)} \lambda_j \\ &= \sum_{j=0}^{k} \lambda_j f(\overline{x}) - \sum_{j=0}^{k} \lambda_j f(x^{(j)}) - \sum_{j=0}^{k} \lambda_j \underbrace{(s^{(j)})^\mathsf{T}(\overline{x} - x^{(j)})}_{=0} \quad \text{nach Definition (21.2)} \\ &= f(\overline{x}) - \sum_{j=0}^{k} \lambda_j f(x^{(j)}). \end{aligned}$$

Durch die Minimierung dieses Ausdrucks wird also versucht, eine Konvexkombination der Funktionswerte $f(x^{(j)})$ zu finden, die dem Wert $f(\overline{x})$ möglichst nah kommt. △

Das LP (21.4) kann unbeschränkt sein, was genau dann der Fall ist, wenn (21.5) unzulässig ist (Satz 8.8). Letzteres ist genau dann der Fall, wenn sich der Nullvektor nicht aus den Subgradienten $s^{(j)}$ konvexkombinieren lässt, also insbesondere dann, wenn nur wenige Subgradienten verwendet werden.

Um diese Schwierigkeit zu umgehen, wollen wir das primale Problem (21.4) durch Hinzufügen eines Terms der Bauart $\|d\|^2$ in der Zielfunktion regularisieren. Die Aufgabe ist dann kein LP mehr, sondern ein konvexes **quadratisches Programm** (vgl. Definition 1.6). Auch für solche QPs gibt es eine **Dualitätstheorie** (englisch: *duality theory*), die wir hier aber nicht im Detail ausführen. Stattdessen stellen wir ohne Beweis oder Herleitung die primalen und dualen Aufgaben für (21.4) und zwei Varianten davon in Tabelle 21.1 zusammen. Dabei verwenden wir für die in den dualen Aufgaben stets vorkommenden Bedingungen $\lambda \geqslant 0$ und $\sum_{j=0}^{k} \lambda_j = 1$ die Abkürzung $\lambda \in \Delta$ (Einheitssimplex, vgl. Beispiel 13.2). Außerdem führen wir zur Vermeidung von Summensymbolen die Matrix

$$S := \begin{bmatrix} | & & | \\ s^{(0)} & \cdots & s^{(k)} \\ | & & | \end{bmatrix} \in \mathbb{R}^{n \times (k+1)}$$

sowie den Vektor

$$\overline{\alpha} := (\overline{\alpha}^{(0)}, \ldots, \overline{\alpha}^{(k)})^\mathsf{T} \in \mathbb{R}^{k+1}$$

ein. Die Parameter τ und ε sind positive Zahlen.

Beachte: Auch die dualen Aufgaben sind hier immer in Minimierungsform angegeben.

Bevor wir die Interpretationen und den Nutzen der verschiedenen Varianten angeben, benötigen wir weitere Informationen. Zunächst geben wir (ohne Beweis) ein bemerkenswertes Resultat für QPs, analog zum Existenzsatz für LPs 6.12 an.

Satz 21.4 (Frank-Wolfe lemma[3], Existenzsatz für QPs).
Es seien $Q \in \mathbb{R}^{n \times n}$, $c \in \mathbb{R}^n$, $\gamma \in \mathbb{R}$ sowie $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ mit $m \in \mathbb{N}_0$. Wir betrachten das QP

$$\begin{aligned} \text{Minimiere} \quad & f(x) := \frac{1}{2} x^\mathsf{T} Q\, x + c^\mathsf{T} x + \gamma \quad \text{über } x \in \mathbb{R}^n \\ \text{unter} \quad & A\, x \leqslant b \end{aligned} \tag{21.12}$$

[3]Dieses Resultat wurde ursprünglich in Frank, Wolfe, 1956 für konvexe QPs bewiesen. Ein direkter Beweis, der auch den nicht-konvexen Fall einschließt, findet sich in Blum, Oettli, 1972.

Minimiere	ξ	über $(d, \xi) \in \mathbb{R}^n \times \mathbb{R}$	Minimiere	$\overline{\alpha}^\mathsf{T} \lambda$	über $\lambda \in \mathbb{R}^{k+1}$
unter	$S^\mathsf{T} d - \overline{\alpha} \leqslant \xi \mathbf{1}$		unter	$\lambda \in \Delta$	
		(21.6)	und	$S\lambda = 0$	(21.7)

Komplementarität: $(S^\mathsf{T} d - \overline{\alpha} - \xi \mathbf{1})^\mathsf{T} \lambda = 0$

Minimiere	$\xi + \frac{\tau}{2}\|d\|^2$	über $(d, \xi) \in \mathbb{R}^n \times \mathbb{R}$	Minimiere	$\overline{\alpha}^\mathsf{T} \lambda + \frac{1}{2\tau}\|S\lambda\|^2$	über $\lambda \in \mathbb{R}^{k+1}$
unter	$S^\mathsf{T} d - \overline{\alpha} \leqslant \xi \mathbf{1}$	(21.8)	unter	$\lambda \in \Delta$	(21.9)

Komplementarität: $(S^\mathsf{T} d - \overline{\alpha} - \xi \mathbf{1})^\mathsf{T} \lambda = 0$

Minimiere	$\xi + \frac{\tau}{2}\|d\|^2 + \varepsilon\eta$	über $(d, \xi, \eta) \in \mathbb{R}^n \times \mathbb{R}^2$	Minimiere	$\frac{1}{2\tau}\|S\lambda\|^2$	über $\lambda \in \mathbb{R}^{k+1}$
unter	$S^\mathsf{T} d - \eta\,\overline{\alpha} \leqslant \xi \mathbf{1}$		unter	$\lambda \in \Delta$	
und	$\eta \geqslant 0$	(21.10)	und	$\overline{\alpha}^\mathsf{T} \lambda \leqslant \varepsilon$	(21.11)

Komplementarität: $(S^\mathsf{T} d - \overline{\alpha} - \xi \mathbf{1})^\mathsf{T} \lambda = 0$, $\eta^\mathsf{T}(\overline{\alpha}^\mathsf{T}\lambda - \varepsilon) = 0$

Tabelle 21.1.: Zusammenstellung primaler und dualer Varianten der Epigraph-Reformulierung (21.4) des Schnittebenenproblems.

mit zulässiger Menge F. Ist der Infimalwert

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

endlich, also die Aufgabe (21.12) weder unzulässig ($f^* = +\infty$) noch unbeschränkt ($f^* = -\infty$), so besitzt (21.12) mindestens einen globalen Minimierer.

Beachte: Der Satz gilt natürlich auch für QPs, die lineare Gleichungsnebenbedingungen enthalten, da man diese ja immer in der Form zweier Ungleichungen schreiben kann.

Bemerkung 21.5 (Existenz von Lösungen)**.**
Mit Satz 21.4 können wir zeigen, dass die Aufgaben (21.8) und (21.9) jeweils mindestens einen globalen Minimierer besitzen, während das für die LPs (21.6) und (21.7) ja nicht notwendigerweise der Fall war. Auch für (21.10) und (21.11) können wir die Existenz von Lösungen unter einer gewissen Bedingung zeigen. Da die Zielfunktionen und zulässigen Mengen jeweils konvex sind, gibt es jeweils keine lokalen Minimierer, die nicht bereits globale Minimierer sind.

(*i*) Für die duale Aufgabe (21.9) ist die Existenz klar, da die zulässige Menge nichtleer und kompakt ist.

(*ii*) Zur Existenz der primalen Aufgabe (21.8): Hier verwenden wir, dass jeder Funktionswert der dualen Zielfunktion eine untere Schranke für den primalen Funktionswert bildet (schwache Dualität). Das zeigen wir jetzt konkret für das Paar (21.8)–(21.9). Achtung, der duale Funktionswert ist das *Negative* der Zielfunktion in (21.9). Die Differenz von primalem und dualem Funktionswert ist daher

$$D := \xi + \frac{\tau}{2}\|d\|^2 + \overline{\alpha}^\mathsf{T}\lambda + \frac{1}{2\tau}\|S\,\lambda\|^2 = \frac{1}{2\tau}\|S\,\lambda + \tau\,d\|^2 - \lambda^\mathsf{T}S^\mathsf{T}d + \overline{\alpha}^\mathsf{T}\lambda + \xi.$$

Aus der primalen Zulässigkeit erhalten wir aber

$$-S^\mathsf{T}d + \overline{\alpha} + \xi\,\mathbf{1} \geqslant 0$$

und durch Multiplikation mit $\lambda \geqslant 0$ (duale Zulässigkeit):

$$-\lambda^\mathsf{T}S^\mathsf{T}d + \lambda^\mathsf{T}\overline{\alpha} + \xi\,\lambda^\mathsf{T}\mathbf{1} \geqslant 0.$$

Das zeigt $D \geqslant 0$. Damit erzeugt jeder dual zulässige Punkt für (21.11) eine untere Schranke an den primalen Funktionswert.

(*iii*) Dieselbe Technik verwenden wir für das Paar (21.10)–(21.11). Die Differenz von primalem und dualem Funktionswert ist hier

$$D := \xi + \frac{\tau}{2}\|d\|^2 + \varepsilon\,\eta + \frac{1}{2\tau}\|S\,\lambda\|^2 = \frac{1}{2\tau}\|S\,\lambda + \tau\,d\|^2 - \lambda^\mathsf{T}S^\mathsf{T}d + \xi + \varepsilon\,\eta.$$

Aus der primalen Zulässigkeit erhalten wir aber

$$-S^\mathsf{T}d + \eta\,\overline{\alpha} + \xi\,\mathbf{1} \geqslant 0$$

und durch Multiplikation mit $\lambda \geqslant 0$ (duale Zulässigkeit):

$$-\lambda^\mathsf{T}S^\mathsf{T}d + \eta\,\lambda^\mathsf{T}\overline{\alpha} + \xi\,\lambda^\mathsf{T}\mathbf{1} \geqslant 0.$$

Hier setzen wir $\mathbf{1}^\mathsf{T}\lambda = 1$ (duale Zulässigkeit) ein sowie $\lambda^\mathsf{T}\overline{\alpha} \leqslant \varepsilon$ (duale Zulässigkeit) und $\eta \geqslant 0$ (primale Zulässigkeit) und erhalten

$$-\lambda^\mathsf{T}S^\mathsf{T}d + \varepsilon\,\eta + \xi \geqslant 0.$$

Das zeigt $D \geqslant 0$. Damit erzeugt jeder dual zulässige Punkt für (21.11) eine untere Schranke an den primalen Funktionswert.

Aber gibt es überhaupt zulässige Punkte für (21.11)? Dafür ist offenbar notwendig und hinreichend, dass $\varepsilon \geqslant \min\{\overline{\alpha}^{(0)}, \dots, \overline{\alpha}^{(k)}\}$ gilt. Denn: Der Minimalwert, den $\overline{\alpha}^\mathsf{T}\lambda$ über Δ erreichen kann, ist gerade gleich $\min\{\overline{\alpha}^{(0)}, \dots, \overline{\alpha}^{(k)}\}$. **Beachte:** Es gilt $\overline{\alpha} \geqslant 0$.

An der Darstellung für D sieht man auch, dass das QP-Paar (21.8)–(21.9) und das QP-Paar (21.10)–(21.11) genau dann denselben Funktionswert haben, wenn die in Tabelle 21.1 behaupteten Komplementaritätsbedingungen gelten sowie $S\,\lambda + \tau\,d = 0$, also $d = -\frac{1}{\tau}S\,\lambda$, wie in (21.17) behauptet. △

Um die Aufgaben aus Tabelle 21.1 besser einordnen zu können, benötigen wir einige weitere Begriffe.

Definition 21.6 (ε-Subdifferential, vgl. Definition 16.1)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ eine konvexe Funktion und $\varepsilon \geqslant 0$.

(*i*) Ein Vektor $s \in \mathbb{R}^n$ heißt ein (Euklidischer) ε-**Subgradient** (englisch: *ε-subgradient*) von f im Punkt $x_0 \in \mathbb{R}^n$, wenn die ε-**Subgradientenungleichung** (englisch: *ε-subgradient inequality*) gilt:

$$f(x) \geqslant f(x_0) + s^\mathsf{T}(x - x_0) - \varepsilon \quad \text{für alle } x \in \mathbb{R}^n. \tag{21.13}$$

(*ii*) Die Menge $\partial_\varepsilon f(x_0)$ *aller* ε-Subgradienten im Punkt x_0 heißt das ε-**Subdifferential** (englisch: *ε-subdifferential*) von f in x_0.

(*iii*) f heißt ε-**subdifferenzierbar** (kurz: ε-**subdiffbar**, englisch: *ε-subdifferentiable*) im Punkt $x_0 \in \mathbb{R}^n$, wenn $\partial_\varepsilon f(x_0) \neq \emptyset$ ist. △

Das ε-Subdifferential hat u. a. folgende Eigenschaften:

Satz 21.7 (Eigenschaften des ε-Subdifferentials)**.**
Es sei $f\colon \mathbb{R}^n \to \mathbb{R}$ eine konvexe Funktion und $x_0 \in \mathbb{R}^n$. Dann gilt

(*i*) $\partial_0 f(x_0) = \partial f(x_0)$.

(*ii*) $\partial_\varepsilon f(x_0) \subseteq \partial_{\varepsilon'} f(x_0)$ für alle $0 \leqslant \varepsilon \leqslant \varepsilon'$ und insbesondere $\partial f(x_0) \subseteq \partial_\varepsilon f(x_0)$.

(*iii*) Für alle $\varepsilon \geqslant 0$ ist $\partial_\varepsilon f(x_0)$ konvex und kompakt.

(*iv*) Ist $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex und reellwertig, dann ist $\partial_\varepsilon f\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ für alle $\varepsilon > 0$ außerhalbstetig und innerhalbstetig.

Beweis. Der Beweis ist weitestgehend Inhalt von Aufgabe 0.67. Für den letzten Punkt siehe Hiriart-Urruty, Lemaréchal, 1993, Abschnitt XI.4.1. □

Mit dem ε-Subdifferential an Stelle des Subdifferentials könnte man im Prinzip ein Abstiegsverfahren mit Liniensuche aufbauen, wobei die Suchrichtung an einer Stelle $\overline{x}$ an Stelle von (19.5) durch

$$d := -\frac{g}{\|g\|} \quad \text{mit } g := \operatorname{proj}_{\partial_\varepsilon f(\overline{x})}(0) \tag{21.14}$$

bestimmt wird.[4] Jedoch muss man für (21.14) das gesamte ε-Subdifferential $\partial_\varepsilon f(\overline{x})$ berechnen können, was für viele Zielfunktionen unrealistisch ist.

Praktischer wäre es, eine implizit gegebene konvexe Teilmenge von $\partial_\varepsilon f(\overline{x})$ zu verwenden, auf die man einfach projizieren kann. Wir zeigen jetzt, dass die Nebenbedingungen in (21.11) genau eine solche Teilmenge

$$G_\varepsilon(\overline{x}) := \{S\lambda \mid \lambda \in \Delta \text{ und } \overline{\alpha}^\mathsf{T}\lambda \leqslant \varepsilon\} \subseteq \partial_\varepsilon f(\overline{x}) \tag{21.15}$$

[4] Die so festgelegte Richtung d minimiert dann gerade die ε-**Richtungsableitung** (englisch: *ε-directional derivative*)

$$f'_\varepsilon(x; d) := \lim_{t \searrow 0} \frac{f(x + t\,d) - f(x) + \varepsilon}{t},$$

vgl. (19.3).

beschreiben. Aufgrund von $s^{(j)} \in \partial f(x^{(j)})$ gilt nämlich

$$\begin{aligned} f(x) &\geqslant f(x^{(j)}) + (s^{(j)})^\mathsf{T}(x - x^{(j)}) \\ &= f(x^{(j)}) \pm f(\overline{x}) + (s^{(j)})^\mathsf{T}(x - \overline{x} + \overline{x} - x^{(j)}) \\ &= f(\overline{x}) - \overline{\alpha}^{(j)} + (s^{(j)})^\mathsf{T}(x - \overline{x}) \end{aligned}$$

für alle $x \in \mathbb{R}^n$, d. h., $s^{(j)}$ gehört zu $\partial_{\overline{\alpha}^{(j)}} f(\overline{x})$. Unter Berücksichtigung von $\lambda \in \Delta$ ergibt die Summation dieser mit λ_j gewichteten Ungleichungen:

$$f(x) \geqslant f(\overline{x}) - \overline{\alpha}^\mathsf{T}\lambda + (S\lambda)^\mathsf{T}(x - \overline{x}).$$

Die Nebenbedingung $\overline{\alpha}^\mathsf{T}\lambda \leqslant \varepsilon$ sichert also gerade

$$S\,\lambda \in \partial_\varepsilon f(\overline{x}). \tag{21.16}$$

Wir können nun unsere Interpretation der einzelnen Aufgaben (21.6)–(21.11) in einer Bemerkung festhalten:

Bemerkung 21.8 (zu den Aufgaben (21.6)–(21.9))**.**

(*i*) Die Aufgabe (21.8) entspricht einer Minimierung des Schnittebenenrichtungsmodells (21.3), wobei jedoch zur Zielfunktion der oft als **Proximalterm** (englisch: *proximal term*) $\frac{\tau}{2}\|d\|^2$ bezeichnete Term hinzugefügt wurde. Dieser bestraft Richtungen d mit großer Norm. Im ursprünglichen Schnittebenenmodell (21.1) mit der Variable x entspricht dies dem Hinzufügen des Terms $\frac{\tau}{2}\|x - \overline{x}\|^2$.

(*ii*) Die eindeutige (**Quizfrage 21.2:** Warum eindeutig?) Lösung (d, ξ) von (21.8) kann man aus einer Lösung λ des dualen QPs (21.9) erhalten, indem man

$$d := -\frac{1}{\tau}S\lambda \quad \text{und} \quad \xi := -\frac{1}{\tau}\|S\lambda\|^2 - \overline{\alpha}^\mathsf{T}\lambda = -\tau\|d\|^2 - \overline{\alpha}^\mathsf{T}\lambda \tag{21.17}$$

setzt.

Beachte: Während λ möglicherweise nicht eindeutig ist, ist es $S\lambda$ doch. (**Quizfrage 21.3:** Warum?)

(*iii*) Das duale QP (21.11) ist (und zwar für beliebiges $\tau > 0$) gerade die Aufgabe der orthogonalen Projektion der Null auf die kompakte Menge

$$G_\varepsilon(\overline{x}) \subseteq \partial_\varepsilon f(\overline{x}).$$

Dabei ist $G_\varepsilon(\overline{x})$ genau dann nichtleer, wenn $\varepsilon \geqslant \min\{\overline{\alpha}^{(0)}, \ldots, \overline{\alpha}^{(k)}\}$ gilt.

(*iv*) Die dualen Aufgaben (21.9) und (21.11) sind eng verwandt. Während in (21.11) die Nebenbedingung $\overline{\alpha}^\mathsf{T}\lambda \leqslant \varepsilon$ explizit gefordert wird, werden in (21.9) große Werte von $\overline{\alpha}^\mathsf{T}\lambda$ bestraft, und zwar umso mehr, je größer der Parameter τ ist.

Man kann zeigen, dass eine Lösung von (21.9) auch eine Lösung von (21.11) ist, wenn man $\varepsilon := \overline{\alpha}^\mathsf{T}\lambda$ wählt. Umgekehrt ist eine Lösung von (21.11) auch eine Lösung von (21.9) für geeignetes $\tau > 0$. △

Abschließend geben noch ein Ergebnis an, das für die Konvergenzanalysis von Bundle-Verfahren nützlich ist:

Satz 21.9 (Beschränktheit des Subdifferentials, vgl. Geiger, Kanzow, 2002, Lemma 6.19)**.**

Es sei $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ konvex und $B \subseteq \operatorname{int}\operatorname{dom} f$ eine kompakte Menge. Dann ist die Bildmenge

$$\partial f(B) = \bigcup_{x \in B} \partial f(x)$$

ebenfalls beschränkt.

Beweis. Wir betrachten die Überdeckung von B durch offene Kugeln

$$\bigcup_{x \in B} B_{r_x}(x),$$

wobei die Radien $r_x > 0$ so gewählt werden, dass $\overline{B_{2r_x}(x)} \subseteq \operatorname{int}\operatorname{dom} f$ bleibt. Aufgrund der Kompaktheit von B sind endlich viele Kugeln zur Überdeckung ausreichend, sagen wir

$$B \subseteq \bigcup_{j=1,\dots,m} B_{r_j}(x_j) \subseteq \bigcup_{j=1,\dots,m} \overline{B_{2r_j}(x_j)} \subseteq \operatorname{int}\operatorname{dom} f.$$

Nach Satz 16.22 ist f stetig auf $\operatorname{int}\operatorname{dom} f$ und damit auf jeder der endlich vielen kompakten Kugeln $\overline{B_{2r_j}(x_j)}$ beschränkt. Somit ist f auch auf der endlichen Vereinigung $\bigcup_{j=1,\dots,m} \overline{B_{2r_j}(x_j)}$ beschränkt.

Es sei nun $\overline{s} \in \partial f(\overline{x})$ für ein $\overline{x} \in B$. Es gilt also

$$f(x) \geqslant f(\overline{x}) + \overline{s}^{\mathsf{T}}(x - \overline{x}) \quad \text{für alle } x \in \mathbb{R}^n.$$

Der Punkt $\overline{x}$ gehört zu B, liegt also in einer der Kugeln, sagen wir in $B_{r_j}(x_j)$. Wir betrachten zunächst den Fall $\overline{s} \neq 0$. Der Punkt $x := \overline{x} + \frac{r_j}{\|\overline{s}\|}\,\overline{s}$ gehört zu $B_{2r_j}(x_j)$, denn es gilt

$$\|x - x_j\| \leqslant \|x - \overline{x}\| + \|\overline{x} - x_j\| < \left\| \frac{r_j}{\|\overline{s}\|}\,\overline{s} \right\| + r_j = 2r_j.$$

Wir setzen x in die Subgradientenungleichung oben ein und erhalten

$$f(x) \geqslant f(\overline{x}) + \overline{s}^{\mathsf{T}} \frac{r_j}{\|\overline{s}\|}\,\overline{s} = f(\overline{x}) + r_j \|\overline{s}\|,$$

also

$$\|\overline{s}\| \leqslant \frac{1}{r_j}\big[f(x) - f(\overline{x})\big].$$

Da x zu $\bigcup_{j=1,\dots,m} \overline{B_{2r_j}(x_j)}$ und $\overline{x}$ zu $\bigcup_{j=1,\dots,m} B_{r_j}(x_j) \subseteq \overline{B_{2r_j}(x_j)}$ gehört, wo f beschränkt ist, und $\overline{s} \in \partial f(B)$ bis auf die Annahme $\overline{s} \neq 0$ beliebig war, folgt, dass $\partial f(B) \setminus \{0\}$ beschränkt ist und damit auch $\partial f(B)$. □

§ 22 Ein Bundle-Verfahren

Wir werden uns mit einem Bundle-Verfahren für konvexe Funktionen $f\colon \mathbb{R}^n \to \mathbb{R}$ beschäftigen, das auf (21.9) basiert und damit gleichermaßen auf der Proximalpunkt-Regularisierung des Schnittebenenrichtungsmodells (21.3). Das Sammeln von Subgradienten $s^{(j)}$ an vorangegangen Iterierten dient also gleichzeitig dem Zweck, das Schnittebenenrichtungsmodell anzureichern, wie auch der besseren Ausschöpfung des ε-Subdifferentials $\partial_\varepsilon f(\overline{x})$ durch $G_\varepsilon(\overline{x})$ (für implizit festgelegtes ε). Gemeinsames Ziel ist es dabei, eine ausreichend gute Abstiegsrichtung zu erhalten.

Das Verfahren unterscheidet zwei Sorten von Iterierten, $x^{(k)}$ und $y^{(k)}$. Die sogenannten **wesentlichen Iterierten** (englisch: *serious iterates*) $x^{(k)}$ oder auch **Stabilitätszentren** (englisch: *stability centers*) dienen als Referenzpunkte und treten an die Position der bisher mit $\overline{x}$ bezeichneten Stelle. Ausgehend von der aktuellen wesentlichen Iterierten $x^{(k)}$ wird ein neuer Kandidat oder **Versuchspunkt** (englisch: *trial iterate*) $y^{(k+1)} := x^{(k)} + d$ bestimmt, basierend auf der Lösung von (21.8) bzw. gleichwertig von (21.9). Falls $y^{(k+1)}$ genügend Abstieg liefert, so wird dieser Punkt die nächste wesentliche Iterierte, also $x^{(k+1)} := y^{(k+1)}$ gesetzt. Das nennt man einen **wesentlichen Schritt** (englisch: *serious step*) des Verfahrens.

Andernfalls bleibt die wesentliche Iterierte unverändert, also $x^{(k+1)} := x^{(k)}$, aber die Subgradienteninformation an der Stelle $y^{(k+1)}$ fließt in das aktuelle Modell ein. Diesen Fall nennt man einen **Nullschritt** (englisch: *null step*).

Algorithmus 22.1 (Ein Bundle-Verfahren, vgl. Geiger, Kanzow, 2002, Algorithmus 6.72)**.**
Eingabe: Startschätzung $x^{(0)} \in \mathbb{R}^n$
Eingabe: Parameter $m \in (0,1)$
Ausgabe: ein globaler Minimierer von (19.1)

Setze $k := 0$
Setze $y^{(0)} := x^{(0)}$
Bestimme ein $s^{(0)} \in \partial f(y^{(0)})$
Setze $J^{(0)} := \{0\}$
Setze $S^{(0)} := \left[s^{(j)}\right]_{j\in J^{(0)}}$
Setze $\alpha^{(0)} := 0$
repeat
    Bestimme eine Lösung $\lambda^{(k)}$ der Aufgabe

$$\begin{aligned} \text{Minimiere} \quad & (\alpha^{(k)})^\mathsf{T}\lambda + \frac{1}{2}\|S^{(k)}\lambda\|^2 \quad \text{über } \lambda \in \mathbb{R}^{|J^{(k)}|} \\ \text{unter} \quad & \lambda \in \Delta \end{aligned} \tag{22.1}$$

    Setze $g^{(k)} := S^{(k)}\lambda^{(k)}$
    Setze $d^{(k)} := -g^{(k)}$
    Setze $\varepsilon^{(k)} := (\alpha^{(k)})^\mathsf{T}\lambda^{(k)}$
    Setze $\xi^{(k)} := -\|g^{(k)}\|^2 - \varepsilon^{(k)}$
    if $\xi^{(k)} < 0$ **then**
        Setze $y^{(k+1)} := x^{(k)} + d^{(k)}$
        if $f(y^{(k+1)}) \leqslant f(x^{(k)}) + m\,\xi^{(k)}$ **then**
            Setze $x^{(k+1)} := y^{(k+1)}$ // wesentlicher Schritt

17: **else**
18: Setze $x^{(k+1)} := x^{(k)}$ // Nullschritt
19: **end if**
20: Bestimme ein $s^{(k+1)} \in \partial f(y^{(k+1)})$
21: Setze $\overline{J}^{(k)} := \left\{ j \in J^{(k)} \,\middle|\, \lambda_j^{(k)} > 0 \right\}$
22: Setze $J^{(k+1)} := \overline{J}^{(k)} \cup \{k+1\}$
23: Setze $S^{(k+1)} := \left[s^{(j)} \right]_{j \in J^{(k+1)}}$
24: Setze $\alpha_j^{(k+1)} := f(x^{(k+1)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k+1)} - y^{(j)})$ für $j \in J^{(k+1)}$
25: Setze $k := k+1$
26: **end if**
27: **until** $\xi^{(k)} = 0$

Bemerkung 22.2 (zu Algorithmus 22.1)**.**

(*i*) Wie oben motiviert baut das Bundle-Verfahren implizit innere Approximationen $G_{\varepsilon^{(k)}}(x^{(k)})$ wie in (21.15) aus dem Bündel der aktuell verwendeten Subgradienten $\{s^{(j)} \mid j \in J^{(k)}\}$ auf.

(*ii*) Das Teilproblem (22.1) entspricht der oben besprochenen Aufgabe (21.9), d. h., der Projektion der Null auf die Menge $G_{\varepsilon^{(k)}}(x^{(k)})$ mit implizit gegebenem ε und $\tau = 1$.

(*iii*) In Zeile 22 wird der neue Subgradient ins Bündel aufgenommen, und die nicht verwendeten Subgradienten werden ein für alle Mal aus dem Bündel entfernt.

(*iv*) Für die Lösung des konvexen QPs (22.1) über dem Einheitssimplex gibt es maßgeschneiderte Lösungsverfahren. Die Dimension dieses QPs entspricht der Anzahl der Subgradienten im aktuellen Bündel. △

Zur Durchführung des Verfahrens aus Algorithmus 22.1 werden die folgenden problemspezifischen Routinen benötigt:

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

(2) Routine, die einen beliebigen Subgradienten $s \in \partial f(x)$ bestimmt.

Wir analysieren jetzt noch die Konvergenz des Verfahrens. Dies erfordert einige Hilfsresultate, bis wir mit Satz 22.10 schließlich das Hauptresultat erhalten. Es bezeichnen jeweils $\cdot^{(k)}$ die durch den Algorithmus 22.1 erzeugten Folgen.

Wir beginnen mit einem einfachen Resultat über die Zugehörigkeit zu gewissen ε-Subdifferentialen.

Lemma 22.3 (erzeugte ε-Subgradienten, vgl. Geiger, Kanzow, 2002, Lemma 6.73)**.**

(*i*) $\alpha_j^{(k)} \geqslant 0$ und $s^{(j)} \in \partial_{\alpha_j^{(k)}} f(x^{(k)})$ für alle $j \in J^{(k)}$ und $k \in \mathbb{N}_0$.

(*ii*) $\varepsilon^{(k)} \geqslant 0$ und $g^{(k)} \in \partial_{\varepsilon^{(k)}} f(x^{(k)})$ für alle $k \in \mathbb{N}_0$.

Beweis. Wir erinnern zunächst an die Definition des Linearisierungsfehlers aus Zeile 24:

$$\alpha_j^{(k)} := f(x^{(k)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k)} - y^{(j)}), \tag{22.2}$$

vgl. (21.2). Die Eigenschaft $\alpha_j^{(k)} \geqslant 0$ wurde in (21.2) gezeigt, wir wählen einfach den Referenzpunkt $\overline{x} = x^{(k)}$ und beachten, dass die Subgradienten $s^{(j)}$ zu den Punkten $y^{(j)}$ gehören.

Für $k = 0$ gilt

$$s^{(0)} \in \partial f(x^{(0)}) = \partial_0 f(x^{(0)}) = \partial_{\alpha_0^{(0)}} f(x^{(0)}),$$

d. h., die Aussage (i) gilt für $k = 0$, da $J^{(0)} = \{0\}$ ist.

Für $k \geqslant 0$ folgt aus der Definition (22.2)

$$\alpha_j^{(k+1)} := f(x^{(k+1)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k+1)} - y^{(j)})$$

und aus $s^{(j)} \in \partial f(y^{(j)})$ die Ungleichung

$$\begin{aligned} f(x) &\geqslant f(y^{(j)}) + (s^{(j)})^\mathsf{T}(x - y^{(j)}) \\ &= f(x^{(k+1)}) + (s^{(j)})^\mathsf{T}(x - x^{(k+1)}) - \left[f(x^{(k+1)}) - f(y^{(j)}) - (s^{(j)})^\mathsf{T}(x^{(k+1)} - y^{(j)})\right] \\ &= f(x^{(k+1)}) + (s^{(j)})^\mathsf{T}(x - x^{(k+1)}) - \alpha_j^{(k+1)} \end{aligned}$$

für alle $x \in \mathbb{R}^n$. Das heißt aber $s^{(j)} \in \partial_{\alpha_j^{(k+1)}} f(x^{(k+1)})$ für $j \in J^{(k+1)}$. Damit ist Aussage (i) gezeigt.

Aussage (ii): Wegen der Nebenbedingungen $\lambda^{(k)} \geqslant 0$, der gerade gezeigten Aussage $\alpha^{(k)} \geqslant 0$ sowie der Definition $\varepsilon^{(k)} := (\alpha^{(k)})^\mathsf{T}\lambda^{(k)}$ folgt $\varepsilon^{(k)} \geqslant 0$. Weiterhin folgt wegen $s^{(j)} \in \partial_{\alpha_j^{(k)}} f(x^{(k)})$ für $j \in J^{(k)}$:

$$f(x) \geqslant f(x^{(k)}) + (s^{(j)})^\mathsf{T}(x - x^{(k)}) - \alpha_j^{(k)} \quad \text{für alle } j \in J^{(k)}$$

und alle $x \in \mathbb{R}^n$ und $k \in \mathbb{N}_0$. Unter Verwendung der Definitionen von $g^{(k)}$ und $\varepsilon^{(k)}$ ergibt sich weiterhin

$$\begin{aligned} f(x) &= \sum_{j \in J^{(k)}} \lambda_j^{(k)} f(x) \\ &\geqslant \sum_{j \in J^{(k)}} \lambda_j^{(k)} \left[f(x^{(k)}) + (s^{(j)})^\mathsf{T}(x - x^{(k)}) - \alpha_j^{(k)}\right] \\ &= f(x^{(k)}) + \sum_{j \in J^{(k)}} \left[(\lambda_j^{(k)} s^{(j)})^\mathsf{T}(x - x^{(k)})\right] - \sum_{j \in J^{(k)}} \lambda_j^{(k)} \alpha_j^{(k)} \\ &= f(x^{(k)}) + (g^{(k)})^\mathsf{T}(x - x^{(k)}) - \varepsilon^{(k)} \end{aligned}$$

für alle $x \in \mathbb{R}^n$ und $k \in \mathbb{N}_0$. Das heißt aber $g^{(k)} \in \partial_{\varepsilon^{(k)}} f(x^{(k)})$. □

Die Abbruchbedingung $\xi^{(k)} = 0$ in Algorithmus 22.1 wird durch folgendes Resultat motiviert:

Lemma 22.4 (Interpretation der Abbruchbedingung, vgl. Geiger, Kanzow, 2002, Lemma 6.74)**.**

(i) $\xi^{(k)} \leqslant 0$ für alle $k \in \mathbb{N}_0$.

(ii) Ist $\xi^{(k)} = 0$, so ist $x^{(k)}$ ein Minimierer von (19.1).

Beweis. Die Aussage (i) folgt sofort aus der Definition $\xi^{(k)} := -\|g^{(k)}\|^2 - \varepsilon^{(k)}$ im Verfahren und $\varepsilon^{(k)} \geqslant 0$, siehe Lemma 22.3. Im Fall von $\xi^{(k)} = 0$ sind $g^{(k)} = 0$ und $\varepsilon^{(k)} = 0$. Aufgrund von Aussage (ii) in Lemma 22.3 gilt also $0 \in \partial_0 f(x^{(k)}) = \partial f(x^{(k)}$, d. h., $x^{(k)}$ ist ein globaler Minimierer von (19.1). □

Lemma 22.5 (Eigenschaften der Iterierten, vgl. Geiger, Kanzow, 2002, Lemma 6.75)**.**
Es gelte $f(x^{(k)}) \geqslant \underline{f}$ für alle $k \in \mathbb{N}_0$. Dann gelten:

(*i*)
$$f(x^{(k)}) - f(x^{(k+1)}) \to 0 \quad \text{für } k \to \infty.$$

(*ii*)
$$\sum_{k=0}^{\infty} t^{(k)} \left(\|g^{(k)}\|^2 + \varepsilon^{(k)} \right) \leqslant \left(f(x^{(0)}) - \underline{f} \right) / m.$$

Dabei ist $t^{(k)} := 0$ für Nullschritte und $t^{(k)} := 1$ für wesentliche Schritte.

(*iii*) Falls es unendlich viele wesentliche Schritt gibt und wir die entsprechende Teilfolge von Indizes mit $\left(k^{(\ell)}\right)$ bezeichnen, dann gilt $g^{(k^{(\ell)})} \to 0$ und $\varepsilon^{(k^{(\ell)})} \to 0$ für $\ell \to \infty$.

Beweis. Aussage (*i*): Per Konstruktion ist die Folge $f(x^{(k)})$ monoton fallend. (**Quizfrage 22.1:** Details?) Da sie nach Voraussetzung nach unten beschränkt ist, konvergiert sie. Damit konvergiert die Folge der Differenzen $f(x^{(k)}) - f(x^{(k+1)})$ gegen Null.

Aussage (*ii*): Aus Zeile 15 im Algorithmus 22.1, also der Entscheidung, ob ein wesentlicher oder ein Nullschritt durchgeführt wird, ergibt sich unter Berücksichtigung der Definition von $t^{(k)}$

$$f(x^{(k+1)}) \leqslant f(x^{(k)}) + m\, t^{(k)}\, \xi^{(k)}$$

für alle $k \in \mathbb{N}_0$, also

$$f(x^{(k)}) - f(x^{(k+1)}) \geqslant -m\, t^{(k)}\, \xi^{(k)}.$$

Durch Aufsummieren erhalten wir

$$f(x^{(0)}) - \underline{f} \geqslant f(x^{(0)}) - f(x^{(k)}) \geqslant -m \sum_{j=0}^{k-1} t^{(j)}\, \xi^{(j)}$$

und im Grenzübergang

$$f(x^{(0)}) - \underline{f} \geqslant -m \sum_{j=0}^{\infty} t^{(j)}\, \xi^{(j)} = m \sum_{j=0}^{\infty} t^{(j)} \left(\|g^{(j)}\|^2 + \varepsilon^{(j)} \right).$$

Die Division durch $m \in (0, 1)$ ergibt die Behauptung.

Aussage (*iii*): Die wesentlichen Schritte sind genau die mit $t^{(k)} = 1$. Ist dies für unendlich viele Indizes, die die Teilfolge $\left(k^{(\ell)}\right)$ bilden, der Fall, dann folgt aus der gerade gezeigten Summierbarkeit von

$$\sum_{k=0}^{\infty} t^{(k)} \left(\|g^{(k)}\|^2 + \varepsilon^{(k)} \right) = \sum_{\ell=0}^{\infty} \underbrace{t^{(k^{(\ell)})}}_{=1} \left(\|g^{(k^{(\ell)})}\|^2 + \varepsilon^{(k^{(\ell)})} \right),$$

dass notwendigerweise $g^{(k^{(\ell)})} \to 0$ und $\varepsilon^{(k^{(\ell)})} \to 0$ gelten. □

Lemma 22.6 (unendlich viele wesentliche Schritte, vgl. Geiger, Kanzow, 2002, Lemma 6.76)**.**
Es gebe unendlich viele wesentliche Schritte in der Folge $\left(x^{(k)}\right)$. Dann ist jeder Häufungspunkt ein Minimierer von (19.1).

Beweis. Es sei x^* ein Häufungspunkt der Folge $\left(x^{(k)}\right)$. Da $f(x^{(k)})$ monoton fallend ist und auf einer Teilfolge gegen $f(x^*)$ konvergiert, gilt $f(x^{(k)}) \geqslant f(x^*) =: f^*$ für alle $k \in \mathbb{N}_0$ und $f(x^{(k)}) \to f(x^*)$.

Nach Lemma 22.3 *(ii)* gilt weiter

$$g^{(k)} \in \partial_{\varepsilon^{(k)}} f(x^{(k)}). \tag{22.3}$$

Da x^* ein Häufungspunkt der Folge $\left(x^{(k)}\right)$ ist und sich $x^{(k)}$ in einem Nullschritt nicht ändert, ist x^* ebenfalls ein Häufungspunkt der Teilfolge $\left(x^{(k^{(\ell)})}\right)$ der wesentlichen Schritte, also der Grenzwert einer Teilfolge von $\left(x^{(k^{(\ell)})}\right)$. Um Dreifach-Indizierung zu vermeiden, bezeichnen wir diese einfach weiter mit $\left(x^{(k^{(\ell)})}\right)$. Es gilt also

$$x^{(k^{(\ell)})} \to x^* \quad \text{für } \ell \to \infty. \tag{22.4}$$

Wegen $f(x^{(k)}) \geqslant f^*$ für alle $k \in \mathbb{N}_0$ gilt nach Lemma 22.5 *(iii)*

$$g^{(k^{(\ell)})} \to 0 \quad \text{und} \quad \varepsilon^{(k^{(\ell)})} \to 0 \quad \text{für } \ell \to \infty. \tag{22.5}$$

Die ε-Subgradientenungleichung für (22.3) ergibt

$$f(x) \geqslant f(x^{(k^{(\ell)})}) + \left(g^{(k^{(\ell)})}\right)^{\mathsf{T}}(x - x^{(k^{(\ell)})}) - \varepsilon^{(k^{(\ell)})}$$

für alle $x \in \mathbb{R}^n$ und $\ell \in \mathbb{N}$. Der Grenzübergang $\ell \to \infty$ zusammen mit der Stetigkeit von f (Satz 16.22), (22.4) und (22.5) zeigt nun

$$f(x) \geqslant f(x^*) + 0 - 0$$

für alle $x \in \mathbb{R}^n$, d. h., x^* ist ein globaler Minimerer von (19.1). □

Lemma 22.7 (nur endliche viele wesentliche Schritte, vgl. Geiger, Kanzow, 2002, Lemma 6.77)**.**
Die in Algorithmus 22.1 erzeugte Folge beinhalte nur *endlich* viele wesentliche Schritte, also gilt $x^{(k)} = x^{(k^*)}$ für alle $k \geqslant k^*$. Dann ist $x^* := x^{(k^*)}$ ein Minimierer von (19.1).

Beweis. Nach Voraussetzung gilt $x^{(k+1)} = x^{(k)}$ für alle $k \geqslant k^*$. Wir haben also

$$\begin{aligned} \alpha_j^{(k+1)} &= f(x^{(k+1)}) - f(y^{(j)}) - (s^{(j)})^{\mathsf{T}}(x^{(k+1)} - y^{(j)}) \quad \text{für } j \in J^{(k+1)}, \\ \alpha_j^{(k)} &= f(x^{(k)}) - f(y^{(j)}) - (s^{(j)})^{\mathsf{T}}(x^{(k)} - y^{(j)}) \qquad \text{für } j \in J^{(k)}. \end{aligned}$$

Wegen $\overline{J}^{(k)} \subseteq J^{(k)}$ und $J^{(k+1)} = \overline{J}^{(k)} \cup \{k+1\}$ ist die Schnittmenge $J^{(k)} \cap J^{(k+1)}$ beider Indexmengen gerade $\overline{J}^{(k)}$. Es gilt also

$$\alpha_j^{(k+1)} = \alpha_j^{(k)} \quad \text{für } j \in \overline{J}^{(k)}.$$

Es folgt

$$\varepsilon^{(k)} = \sum_{j \in J^{(k)}} \alpha_j^{(k)} \lambda_j^{(k)} = \sum_{j \in \overline{J}^{(k)}} \alpha_j^{(k)} \lambda_j^{(k)} = \sum_{j \in \overline{J}^{(k)}} \alpha_j^{(k+1)} \lambda_j^{(k)} =: \sigma^{(k)} \tag{22.6}$$

für alle $k \geqslant k^*$.

Es sei nun $\mu \in [0,1]$ beliebig. Wir definieren den Vektor $\overline{\lambda}$ mit Komponenten $\overline{\lambda}_j$ für $j \in J^{(k)} = \overline{J}^{(k-1)} \cup \{k\}$ (disjunkte Vereinigung) gemäß

$$\overline{\lambda}_j := \begin{cases} \mu, & \text{falls } j = k, \\ (1-\mu)\,\lambda_j^{(k-1)}, & \text{falls } j \in \overline{J}^{(k-1)}. \end{cases}$$

Dabei sind die $\lambda_j^{(k-1)}$ für $j \in \overline{J}^{(k-1)}$ gerade die echt positiven Komponenten der Lösung $\lambda^{(k-1)}$ von (22.1) in der Iteration $k-1$. Dann ist $\overline{\lambda}$ zulässig für (22.1) in der Iteration k (**Quizfrage 22.2:** Begründung?)

Wir bezeichnen zur Abkürzung die Zielfunktion von (22.1) in Iteration k mit

$$q^{(k)}(\lambda) := (\alpha^{(k)})^\mathsf{T}\lambda + \frac{1}{2}\|S^{(k)}\lambda\|^2 = \sum_{j\in J^{(k)}} \alpha_j^{(k)}\lambda_j + \frac{1}{2}\Big\|\sum_{j\in J^{(k)}} \lambda_j\, s^{(j)}\Big\|^2.$$

Da $\lambda^{(k)}$ ein globaler Minimierer von (22.1) ist, gilt

$$q^{(k)}(\lambda^{(k)}) \leqslant q^{(k)}(\overline{\lambda}).$$

Wir werten nun die Terme in der rechten Seite aus:

$$\begin{aligned}
\sum_{j\in J^{(k)}} \overline{\lambda}_j\, s^{(j)} &= \overline{\lambda}_k\, s^{(k)} + \sum_{j\in\overline{J}^{(k-1)}} \overline{\lambda}_j\, s^{(j)}, && \text{da } J^{(k)} = \overline{J}^{(k-1)} \cup \{k\}\\
&= \mu\, s^{(k)} + \sum_{j\in\overline{J}^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)}\, s^{(j)} && \text{nach Definition von } \overline{\lambda}\\
&= \mu\, s^{(k)} + \sum_{j\in J^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)}\, s^{(j)}, && \text{da } \lambda_j^{(k-1)} = 0 \text{ für } j \in J^{(k-1)} \setminus \overline{J}^{(k-1)}\\
&= \mu\, s^{(k)} + (1-\mu)\, g^{(k-1)} && \text{nach Definition von } g^{(k-1)}
\end{aligned}$$

und

$$\begin{aligned}
\sum_{j\in J^{(k)}} \alpha_j^{(k)}\overline{\lambda}_j &= \alpha_k^{(k)}\,\overline{\lambda}_k + \sum_{j\in\overline{J}^{(k-1)}} \alpha_j^{(k)}\,\overline{\lambda}_j, && \text{da } J^{(k)} = \overline{J}^{(k-1)} \cup \{k\}\\
&= \mu\,\alpha_k^{(k)} + \sum_{j\in\overline{J}^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)}\,\alpha_j^{(k)} && \text{nach Definition von } \overline{\lambda}\\
&= \mu\,\alpha_k^{(k)} + \sum_{j\in J^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)}\,\alpha_j^{(k)}, && \text{da } \lambda_j^{(k-1)} = 0 \text{ für } j \in J^{(k-1)} \setminus \overline{J}^{(k-1)}\\
&= \mu\,\alpha_k^{(k)} + (1-\mu)\,\sigma^{(k-1)} && \text{nach Definition (22.6) von } \sigma^{(k-1)}.
\end{aligned}$$

Damit ergibt sich

$$q^{(k)}(\lambda^{(k)}) \leqslant q^{(k)}(\overline{\lambda}) = \mu\,\alpha_k^{(k)} + (1-\mu)\,\sigma^{(k-1)} + \frac{1}{2}\|\mu\, s^{(k)} + (1-\mu)\, g^{(k-1)}\|^2 =: h(\mu).$$

Da dies für alle $\mu \in [0,1]$ gilt, haben wir sogar

$$q^{(k)}(\lambda^{(k)}) \leqslant \min\{h(\mu) \mid \mu \in [0,1]\}.$$

Beachte: Es ist $h(\mu) \geqslant 0$ für alle $\mu \in [0,1]$, da $\alpha_k^{(k)} \geqslant 0$ und $\sigma^{(k-1)} = \varepsilon^{(k-1)} \geqslant 0$ nach (22.6) und Lemma 22.3 gilt.

Es sei $\mu^{(k)}$ die eindeutige Lösung dieser Aufgabe (**Quizfrage 22.3:** Warum ist diese eindeutig?) mit Infimalwert $\gamma^{(k)} := h(\mu^{(k)}) \geqslant 0$. Dann haben wir also

$$q^{(k)}(\lambda^{(k)}) \leqslant \gamma^{(k)} = h(\mu^{(k)}) \leqslant h(\mu) \quad \text{für alle } \mu \in [0,1].$$

Zusammen mit (22.6) erhalten wir für alle $k > k^*$:

$$\begin{aligned}
\gamma^{(k)} &\leqslant h(0) \\
&= \sigma^{(k-1)} + \frac{1}{2}\|g^{(k-1)}\|^2 \\
&= \sum_{j\in J^{(k-1)}} \alpha_j^{(k-1)}\lambda_j^{(k-1)} + \frac{1}{2}\Big\| \sum_{j\in J^{(k-1)}} \lambda_j^{(k-1)}\, s^{(j)}\Big\|^2 \qquad \text{nach (22.6) und Definition von } g^{(k-1)} \\
&= q^{(k-1)}(\lambda^{(k-1)}) \\
&\leqslant \gamma^{(k-1)}.
\end{aligned} \tag{22.7}$$

Es gilt also

$$0 \leqslant \gamma^{(k)} \leqslant \gamma^{(k-1)} \leqslant \gamma^{(k^*)} \quad \text{für alle } k > k^*$$

und somit

$$\frac{1}{2}\|g^{(k)}\|^2 \leqslant \gamma^{(k^*)} \quad \text{und} \quad \sigma^{(k)} \leqslant \gamma^{(k^*)} \quad \text{für alle } k \geqslant k^*. \tag{22.8}$$

Mit

$$\frac{1}{2}\|g^{(k-1)}\|^2 \leqslant \frac{1}{2}\|g^{(k-1)}\|^2 + \sigma^{(k-1)} \leqslant \gamma^{(k-1)}, \tag{22.9}$$

siehe (22.7), erhalten wir

$$\begin{aligned}
h(\mu) &= \mu\,\alpha_k^{(k)} + (1-\mu)\,\sigma^{(k-1)} + \frac{1}{2}\|\mu\, s^{(k)} + (1-\mu)\, g^{(k-1)}\|^2 && \text{nach Definition von } h \\
&= \mu\,\alpha_k^{(k)} + (1-\mu)\,\sigma^{(k-1)} \\
&\quad + \frac{\mu^2}{2}\|s^{(k)} - g^{(k-1)}\|^2 + \frac{2\mu}{2}\big(s^{(k)} - g^{(k-1)}\big)^{\mathsf T} g^{(k-1)} + \frac{1}{2}\|g^{(k-1)}\|^2 \\
&\leqslant \mu\,\big(\alpha_k^{(k)} - \sigma^{(k-1)}\big) + \gamma^{(k-1)} \\
&\quad + \frac{\mu^2}{2}\|s^{(k)} - g^{(k-1)}\|^2 + \mu\,\big(s^{(k)}\big)^{\mathsf T} g^{(k-1)} - \mu\,\|g^{(k-1)}\|^2 && \text{nach (22.9)}
\end{aligned}$$

für alle $k > k^*$.

Aus $x^{(k)} = x^{(k^*)}$ für alle $k \geqslant k^*$, der Beziehung

$$\begin{aligned}
\alpha_k^{(k)} &= f(x^{(k)}) - f(y^{(k)}) - (s^{(k)})^{\mathsf T}(x^{(k)} - y^{(k)}) \\
&= f(x^{(k)}) - f(y^{(k)}) - (s^{(k)})^{\mathsf T}(x^{(k)} - x^{(k-1)} - d^{(k)}) \\
&= f(x^{(k-1)}) - f(y^{(k)}) + (s^{(k)})^{\mathsf T} d^{(k)}
\end{aligned}$$

für alle $k > k^*$ sowie $f(y^{(k)}) > f(x^{(k-1)}) + m\,\xi^{(k-1)}$ für alle $k > k^*$ (der Bedingung für einen Nullschritt) folgt

$$\begin{aligned}
-\alpha_k^{(k)} + (s^{(k)})^{\mathsf T} d^{(k-1)} &= f(y^{(k)}) - f(x^{(k-1)}) \\
&> m\,\xi^{(k-1)} = -m\,\big(\|g^{(k-1)}\|^2 + \sigma^{(k-1)}\big) \quad \text{nach Definition von } \xi^{(k-1)} \text{ und (22.6)}
\end{aligned}$$

für alle $k > k^*$. Mit $d^{(k-1)} = -g^{(k-1)}$ können wir die Ungleichung umformen zu

$$(s^{(k)})^{\mathsf T} g^{(k-1)} < m\,\big(\|g^{(k-1)}\|^2 + \sigma^{(k-1)}\big) - \alpha_k^{(k)}.$$

Wir erhalten somit für alle $k > k^*$

$$\begin{aligned} h(\mu) &\leqslant \mu\left(\alpha_k^{(k)} - \sigma^{(k-1)}\right) + \gamma^{(k-1)} + \frac{\mu^2}{2}\|s^{(k)} - g^{(k-1)}\|^2 + \mu\left(s^{(k)}\right)^{\mathsf{T}} g^{(k-1)} - \mu\,\|g^{(k-1)}\|^2 \\ &\leqslant \mu\left(\alpha_k^{(k)} - \sigma^{(k-1)}\right) + \gamma^{(k-1)} + \frac{\mu^2}{2}\|s^{(k)} - g^{(k-1)}\|^2 + \mu\, m\left(\|g^{(k-1)}\|^2 + \sigma^{(k-1)}\right) - \mu\,\alpha_k^{(k)} - \mu\,\|g^{(k-1)}\|^2 \\ &= \gamma^{(k-1)} + \frac{\mu^2}{2}\|s^{(k)} - g^{(k-1)}\|^2 - \mu\,(1-m)\,\sigma^{(k-1)} - \mu\,(1-m)\,\|g^{(k-1)}\|^2. \end{aligned} \tag{22.10}$$

Nach (22.8) sind die Folgen $\sigma^{(k)}$ und $g^{(k)}$ beschränkt und daher auch $d^{(k)}$ und $y^{(k)}$. Wegen $s^{(k)} \in \partial f(y^{(k)})$ und Satz 21.9 ist auch $s^{(k)}$ beschränkt. Wir halten also fest, dass es eine Konstante c gibt, für die wir o. B. d. A. als $c \geqslant 1/2$ annehmen können, sodass

$$\|g^{(k)}\| \leqslant c, \quad \|s^{(k)}\| \leqslant c \quad \text{und} \quad \sigma^{(k)} \leqslant c$$

für alle $k \geqslant k^*$ gilt. Das zeigt

$$\|s^{(k)} - g^{(k-1)}\|^2 \leqslant \left(\|s^{(k)}\| + \|g^{(k-1)}\|\right)^2 \leqslant 4\,c^2,$$

und Einsetzen in (22.10) ergibt

$$\begin{aligned} h(\mu) &\leqslant \gamma^{(k-1)} + \frac{\mu^2}{2}\|s^{(k)} - g^{(k-1)}\|^2 - \mu\,(1-m)\,\sigma^{(k-1)} - \mu\,(1-m)\,\|g^{(k-1)}\|^2 \\ &\leqslant \gamma^{(k-1)} + 2\,c^2\mu^2 - \mu\,(1-m)\,\sigma^{(k-1)} - \mu\,(1-m)\,\|g^{(k-1)}\|^2 \\ &= 2\,c^2\mu^2 - (1-m)\left[\sigma^{(k-1)} + \|g^{(k-1)}\|^2\right]\mu + \gamma^{(k-1)} =: \theta(\mu) \end{aligned}$$

für alle $k > k^*$. Die Minimalwerte von θ über $\mathbb{R}$ und über $[0,1]$ stimmen überein und sind gleich (Minimalstelle ausrechnen) $\theta^* = \gamma^{(k-1)} - (1-m)^2\left[\sigma^{(k-1)} + \|g^{(k-1)}\|^2\right]^2/(8\,c^2)$.

Wir fassen zusammen:

$$\begin{aligned} \gamma^{(k)} &= h(\mu^{(k)}) && \text{nach Definition von } \gamma^{(k)} \\ &= \min\{h(\mu) \mid \mu \in [0,1]\} \\ &\leqslant \min\{\theta(\mu) \mid \mu \in [0,1]\} \\ &= \theta^* \\ &= \gamma^{(k-1)} - \frac{(1-m)^2}{8\,c^2}\left[\sigma^{(k-1)} + \|g^{(k-1)}\|^2\right]^2. \end{aligned}$$

Das Aufsummieren dieser Ungleichung für $j = k^* + 1, \ldots, k+1$ liefert

$$\frac{(1-m)^2}{8\,c^2}\sum_{j=k^*}^{k}\left[\sigma^{(k-1)} + \|g^{(k-1)}\|^2\right]^2 \leqslant \gamma^{(k^*)} - \gamma^{(k+1)}.$$

Mit $\gamma^{(k+1)} \geqslant 0$ folgt daraus

$$\sum_{j=k^*}^{\infty}\left[\sigma^{(k-1)} + \|g^{(k-1)}\|^2\right]^2 < \infty.$$

Daraus folgt nun schließlich $g^{(k)} \to 0$ und $\varepsilon^{(k)} = \sigma^{(k)} \to 0$ für $k \to \infty$. Eine solche Situation hatten wir schon im Beweis von Lemma 22.6. Wie dort folgt, dass x^* ein Minimierer von (19.1) ist. □

Aus Lemma 22.6 und Lemma 22.7 erhalten wir sofort die folgende vorläufige Konvergenzaussage:

Satz 22.8 (Häufungspunkte sind Minimierer, vgl. Geiger, Kanzow, 2002, Satz 6.78)**.**
Jeder Häufungspunkt einer von Algorithmus 22.1 erzeugten Folge $\left(x^{(k)}\right)$ ist ein globaler Minimierer von (19.1).

Um das Ergebnis noch zu verbessern, wollen wir zeigen, dass solche Häufungspunkte bereits unter einer schwachen Voraussetzung existieren.

Lemma 22.9 (Existenz von Häufungspunkten)**.**
Die Aufgabe (19.1) besitze mindestens einen globalen Minimierer. Ist x^* einer der Minimierer und $\left(x^{(k)}\right)$ eine von Algorithmus 22.1 erzeugte Folge, dann gelten:

(i)
$$\|x^{(k)} - x^*\|^2 \leqslant \|x^{(m)} - x^*\|^2 + \sum_{j=m}^{k-1} \left(\|x^{(j+1)} - x^{(j)}\|^2 + 2\, t^{(j)}\, \varepsilon^{(j)}\right)$$
für alle $m \in \mathbb{N}_0$ und alle $k \geqslant m$.

(ii)
$$\sum_{j=0}^{\infty} \left(\|x^{(j+1)} - x^{(j)}\|^2 + 2\, t^{(j)}\, \varepsilon^{(j)}\right) \text{ ist endlich.}$$

(iii) Die Folge $\left(x^{(k)}\right)$ ist beschränkt.

Beweis. Aussage (i): Wir nutzen $g^{(k)} \in \partial_{\varepsilon^{(k)}} f(x^{(k)})$ (Lemma 22.3) und die Voraussetzung $f(x^{(k)}) \geqslant f(x^*)$ und erhalten
$$0 \geqslant f(x^*) - f(x^{(k)}) \geqslant (g^{(k)})^{\mathsf{T}}(x^* - x^{(k)}) - \varepsilon^{(k)}$$
und somit
$$(g^{(k)})^{\mathsf{T}}(x^* - x^{(k)}) \leqslant \varepsilon^{(k)}.$$
Mit $x^{(k+1)} - x^{(k)} = t^{(k)} d^{(k)} = -t^{(k)} g^{(k)}$ und $t^{(k)} \geqslant 0$ erhalten wir
$$-(x^* - x^{(k)})^{\mathsf{T}}(x^{(k+1)} - x^{(k)}) \leqslant t^{(k)} \varepsilon^{(k)} \quad \text{für alle } k \in \mathbb{N}.$$
Das impliziert
$$\begin{aligned}\|x^* - x^{(k+1)}\|^2 &= \|x^* - x^{(k)} + x^{(k)} - x^{(k+1)}\|^2 \\ &= \|x^* - x^{(k)}\|^2 - 2\,(x^* - x^{(k)})^{\mathsf{T}}(x^{(k+1)} - x^{(k)}) + \|x^{(k+1)} - x^{(k)}\|^2 \\ &\leqslant \|x^* - x^{(k)}\|^2 + \|x^{(k+1)} - x^{(k)}\|^2 + 2\, t^{(k)} \varepsilon^{(k)}.\end{aligned}$$
Durch Summation erhalten wir die Aussage.

Aussage (ii): Die Aussage
$$\|x^{(k+1)} - x^{(k)}\|^2 = (t^{(k)})^2 \|d^{(k)}\|^2 = (t^{(k)})^2 \|g^{(k)}\|^2 \leqslant 2\,(t^{(k)})^2 \|g^{(k)}\|^2$$

und $t^{(k)} \in [0,1]$ ergibt zusammen mit Lemma 22.5

$$\begin{aligned}\sum_{j=0}^{\infty} \left(\|x^{(j+1)} - x^{(j)}\|^2 + 2\, t^{(j)}\, \varepsilon^{(j)} \right) &\leqslant 2 \sum_{j=0}^{\infty} \left(2\, (t^{(j)})^2 \|g^{(j)}\|^2 + t^{(j)}\, \varepsilon^{(j)} \right) \\ &\leqslant 2 \sum_{j=0}^{\infty} \left(t^{(j)} \|g^{(j)}\|^2 + t^{(j)}\, \varepsilon^{(j)} \right) \\ &< \infty.\end{aligned}$$

Aussage *(iii)*: Die Behauptung folgt aus Aussagen *(i)* und *(ii)*. □

Es folgt nun unser Hauptergebnis zur Konvergenz des Bundle-Verfahrens aus Algorithmus 22.1.

Satz 22.10 (Konvergenzsatz für Algorithmus 22.1)**.**
Die Aufgabe (19.1) besitze mindestens einen globalen Minimierer. Dann konvergiert jede von Algorithmus 22.1 erzeugte Folge $\left(x^{(k)}\right)$ gegen einen globalen Minimierer von (19.1).

Beweis. Nach Lemma 22.9 ist die Folge $\left(x^{(k)}\right)$ beschränkt. Es existiert also mindestens ein Häufungspunkt x^*. Nach Satz 22.8 ist dieser ein globaler Minimierer von f. Es bleibt zu zeigen, dass die gesamte Folge $\left(x^{(k)}\right)$ gegen x^* konvergiert.

Es sei $\varepsilon > 0$. Da $\left(x^{(k)}\right)$ auf einer Teilfolge gegen x^* konvergiert und die Reihe aus Lemma 22.9 *(ii)* konvergiert, gibt es ein $m \in \mathbb{N}$ mit

$$\|x^{(m)} - x^*\| \leqslant \frac{\varepsilon}{2} \quad \text{und} \quad \sum_{j=m}^{\infty} \left(\|x^{(j+1)} - x^{(j)}\|^2 + 2\, t^{(j)}\, \varepsilon^{(j)} \right) \leqslant \frac{\varepsilon}{2}.$$

Damit ergibt sich aus Lemma 22.9 *(i)*

$$\|x^{(k)} - x^*\|^2 \leqslant \|x^{(m)} - x^*\|^2 + \frac{\varepsilon}{2} \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

für alle $k \geqslant m$. Da $\varepsilon > 0$ beliebig war, folgt die Behauptung. □

Index

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.

Armijo, L. (1966). „Minimization of functions having Lipschitz continuous first partial derivatives“. *Pacific Journal of Mathematics* 16.1, S. 1–3. DOI: 10.2140/pjm.1966.16.1.

Blum, E.; W. Oettli (1972). „Direct proof of the existence theorem for quadratic programming“. *Operations Research* 20, S. 165–167. DOI: 10.1287/opre.20.1.165.

Bonnans, F.; C. Gilbert; C. Lemaréchal; C. Sagastizábal (2003). *Numerical Optimization*. 1. Aufl. Berlin: Springer. DOI: 10.1007/978-3-662-05078-1.

Cartan, H. (1967). *Calcul Différentiel*. Paris: Hermann.

Cohn, P. M. (1981). *Universal Algebra*. Second. Bd. 6. Mathematics and its Applications. Dordrecht-Boston: D. Reidel Publishing Co. DOI: 10.1007/978-94-009-8399-1.

Forsgren, A. (2008). *An elementary proof of optimality conditions for linear programming*. TRITA–MAT 2008–OS6. Department of Mathematics, Royal Institute of Technology (KTH) Stockholm.

Frank, M.; P. Wolfe (1956). „An algorithm for quadratic programming“. *Naval Research Logistics Quarterly* 3, S. 95–110. DOI: 10.1002/nav.3800030109.

Gass, S. I.; A. A. Assad (2005). *An Annotated Timeline of Operations Research: An Informal History*. Bd. 75. International Series in Operations Research & Management Science. Boston, MA: Kluwer Academic Publishers.

Geiger, C.; C. Kanzow (1999). *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. New York: Springer. DOI: 10.1007/978-3-642-58582-1.

– (2002). *Theorie und Numerik restringierter Optimierungsaufgaben*. New York: Springer. DOI: 10.1007/978-3-642-56004-0.

Gerdts, M.; F. Lempio (2011). *Mathematische Optimierungsverfahren des Operations Research*. de Gruyter. DOI: 10.1515/9783110249989.

Gill, P. E.; W. Murray; M. H. Wright (1981). *Practical Optimization*. London: Academic Press.

Hamacher, H.; B. Klamroth (2006). *Lineare Optimierung und Netzwerkoptimierung*. 2. Aufl. Vieweg. DOI: 10.1007/978-3-8348-9031-3.

Heuser, H. (2002). *Lehrbuch der Analysis. Teil 2*. 12. Aufl. Stuttgart: B.G.Teubner. DOI: 10.1007/978-3-322-96826-5.

– (2003). *Lehrbuch der Analysis. Teil 1*. 15. Aufl. Vieweg+Teubner Verlag. DOI: 10.1007/978-3-322-96828-9.

Hiriart-Urruty, J.-B.; C. Lemaréchal (1993). *Convex Analysis and Minimization Algorithms. II*. Bd. 306. Grundlehren der Mathematischen Wissenschaften. Advanced theory and bundle methods. Berlin: Springer. DOI: 10.1007/978-3-662-06409-2.

– (2001). *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer, Berlin. DOI: 10.1007/978-3-642-56468-0.

Horn, R. A.; C. R. Johnson (1990). *Matrix Analysis*. Corrected reprint of the 1985 original. Cambridge University Press, Cambridge.

Jarre, F.; J. Stoer (2004). *Optimierung*. Springer. DOI: 10.1007/978-3-642-18785-8.

Kager, W. (2023). „A short simple proof of closedness of convex cones and Farkas' lemma". *The American Mathematical Monthly* 131.1, S. 74–75. DOI: 10.1080/00029890.2023.2261816.

Karmarkar, N. (1984a). „A new polynomial-time algorithm for linear programming". *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing - STOC '84*. ACM Press. DOI: 10.1145/800057.808695.

– (1984b). „A new polynomial-time algorithm for linear programming". *Combinatorica* 4.4, S. 373–395. DOI: 10.1007/bf02579150.

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.

Nocedal, J.; S. J. Wright (2006). *Numerical Optimization*. 2. Aufl. New York: Springer. DOI: 10.1007/978-0-387-40065-5.

Phelps, R. R. (1993). *Convex Functions, Monotone Operators and Differentiability*. 2. Aufl. Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-46077-0.

Rockafellar, R. T. (1970). *Convex Analysis*. Bd. 28. Princeton Mathematical Series. Princeton, New Jersey: Princeton University Press. URL: https://www.jstor.org/stable/j.ctt14bs1ff.

Schrijver, A. (2003). *Combinatorial Optimization. Polyhedra and Efficiency. Volume A*. Bd. 24. Algorithms and Combinatorics. Paths, flows, matchings, Chapters 1–38. Springer, Berlin.

Vanderbei, R. J. (2008). *Linear Programming: Foundations and Extensions*. Ooperations Research, Management Science. New York, NY: Springer. DOI: 10.1007/978-0-387-74388-2.

Von Neumann, J. (1928). „Zur Theorie der Gesellschaftsspiele". *Mathematische Annalen* 100.1, S. 295–320. DOI: 10.1007/bf01448847.

Werner, J. (2007). *Vorlesung über Optimierung*. Lecture Notes, Department of Mathematics, University of Hamburg, Germany. URL: http://num.math.uni-goettingen.de/werner/.

Westermann, L. (1976). „On the hull operator". *Indagationes Mathematicae (Proceedings)* 79.2, S. 179–184. DOI: 10.1016/1385-7258(76)90065-2.