

**EINFÜHRUNG IN DIE MATHEMATIK
DES OPERATIONS RESEARCH**

Ulrich Faigle

Skriptum zur Vorlesung

Sommersemester 2009

Universität zu Köln

Universität zu Köln
Mathematisches Institut
Zentrum für Angewandte Informatik

Weyertal 80

faigle@zpr.uni-koeln.de

www.zaik.uni-koeln.de/AFS

Inhaltsverzeichnis

Notationen und Terminologie	3
1. Lineare Algebra	3
2. Ordnungsrelationen	5
3. Topologie	6
4. Mathematische Optimierungsprobleme	7
Kapitel 1. Konvexe Mengen und Funktionen	11
1. Konvexe Mengen	11
2. Konvexe Funktionen	19
3. Konvexe Optimierung	22
4. Dualität	29
Kapitel 2. Fundamentale Algorithmen	33
1. Zeilen- und Spaltenoperationen	33
2. Elimination nach Fourier-Motzkin	34
3. Die Ellipsoidmethode	40
4. Die Methode innerer Punkte	44
Kapitel 3. Struktur von Polyedern	49
1. Der Darstellungssatz von Weyl-Minkowski	49
2. Seitenflächen, Ecken und Facetten	55
3. Rationale lineare Systeme	61
4. Die Simplexmethode	65
Kapitel 4. Optimierung auf Netzwerken	77
1. Flüsse, Potentiale und Spannungen	77
2. Kürzeste Wege	81
3. Zirkulationen und das MAX-Flow-MIN-Cut-Theorem	83
4. Der Präfluss-Markierungsalgorithmus	86
Kapitel 5. Ganzzahlige lineare Programme	91
1. Unimodularität	91
2. Schnittebenen	94

Notationen und Terminologie

1. Lineare Algebra

Für beliebige Mengen R und N notiert man

$$R^N = \{f : N \rightarrow R\}.$$

Für $f \in R^N$ und $i \in N$ setzt man auch $f_i = f(i)$ und nennt f_i die *ite* Koordinate von f .

Besonders anwendungsrelevant sind die Skalarbereiche $R = \mathbb{N}$, $R = \mathbb{Z}$, $R = \mathbb{Q}$ oder $R = \mathbb{R}$, wo man die Elemente (Funktionen) in R^N koordinatenweise miteinander addieren und mit Skalaren multiplizieren kann.

Im Fall $N = \{1, \dots, n\}$ schreibt man oft kurz: $R^n = R^N$.

1.1. Vektoren und Matrizen. Die Elemente von \mathbb{R}^n heißen *n-dimensionale Parametervektoren*. Im Skriptum wird ein solches $\mathbf{x} \in \mathbb{R}^n$ **fett** notiert und typischerweise als *Spaltenvektor* verstanden:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad (x_i \in \mathbb{R}).$$

Als *Zeilenvektor* wird der Parametervektor meistens transponiert notiert:

$$\mathbf{x}^T = [x_1, \dots, x_n].$$

$\mathbf{0} = [0, \dots, 0]^T$ ist der *Nullvektor*. Wenn der formale Unterschied zwischen Spalten- und Zeilenvektor nicht so wichtig ist, wird ein Parametervektor auch mit runden Klammern notiert:

$$\mathbf{x} = (x_1, \dots, x_n).$$

$\mathbb{R}^{m \times n}$ ist die Menge aller $(m \times n)$ -Matrizen. Ein $A = [a_{ij}] \in \mathbb{R}^{m \times n}$ kann man entweder als n -Tupel von m -dimensionalen Spaltenvektoren A_j oder als m -Tupel von n -dimensionalen Zeilenvektoren A_i auffassen:

$$[A_1, \dots, A_n] = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix}$$

Für $A = [\mathbf{a}_1, \dots, \mathbf{a}_n] \in \mathbb{R}^{m \times n}$ und $\mathbf{x} = [x_1, \dots, x_n]^T \in \mathbb{R}^n$, notiert man die entsprechende Linearkombination der Spaltenvektoren:

$$A\mathbf{x} = x_1\mathbf{a}_1 + \dots + x_n\mathbf{a}_n = \sum_{j=1}^n x_j\mathbf{a}_j.$$

Für $\mathbf{y} \in \mathbb{R}^m$ ist $\mathbf{y}^T A = (A^T \mathbf{y})^T$ die analoge Linearkombination der Zeilenvektoren von A .

Ist $B = [\mathbf{b}_1, \dots, \mathbf{b}_k] \in \mathbb{R}^{n \times k}$ eine weitere Matrix, so kann man das folgende Matrixprodukt bilden:

$$AB = [A\mathbf{b}_1, \dots, A\mathbf{b}_k] \in \mathbb{R}^{m \times k}.$$

1.2. Analytische Geometrie. \mathbb{R}^n kann man auch als Menge der Koordinatenvektoren eines n -dimensionalen „Universums“ von „Punkten“ ansehen. Geometrische Punkte P, Q kann man „eigentlich“ nicht addieren oder subtrahieren. Die Differenz $Q - P$ der entsprechenden Koordinatenvektoren ist aber mathematisch sinnvoll. Man fasst

$$\overrightarrow{PQ} = Q - P$$

dann als einen Vektor auf, der eine „Wirkung“ beschreibt, die den Ortszustand P in den Ortszustand Q verändert.

1.3. Affine und lineare Teilräume. Ein *Hyperebene* in \mathbb{R}^n ist eine Teilmenge der Form

$$H = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T \mathbf{x} = b\} \quad (\mathbf{a} \in \mathbb{R}^n \setminus \{\mathbf{0}\}, b \in \mathbb{R}).$$

Ein *affiner* Teilraum \mathbb{A} ist ein Durchschnitt von Hyperebenen. Insbesondere ist $\emptyset \subseteq \mathbb{R}^n$ ein affiner Teilraum. Aus der linearen Algebra weiss man:

LEMMA 0.1. *Für eine beliebige nichtleere Teilmenge $S \subseteq \mathbb{R}^n$ sind die Aussagen äquivalent:*

- (0) S ist ein affiner Teilraum.
- (1) Es gibt ein $m \in \mathbb{N}$ und eine Matrix $A \in \mathbb{R}^{m \times n}$ und einen Vektor $\mathbf{b} \in \mathbb{R}^m$ so dass $S = \{x \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{b}\}$.
- (2) Es gibt Vektoren $\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$ so, dass $S = \{\mathbf{v}_0 + \sum_{i=1}^k \lambda_i \mathbf{v}_i \mid \lambda_i \in \mathbb{R}\}$.
- (3) Für beliebige $\mathbf{u}, \mathbf{v} \in S$ und Skalare $\alpha \in \mathbb{R}$ gilt: $\mathbf{z} = \alpha \mathbf{u} + (1 - \alpha) \mathbf{v} \in S$.

◇

Ein affiner Teilraum \mathbb{A} heisst *linear* im Fall $\mathbf{0} \in \mathbb{A}$.

2. Ordnungsrelationen

2.1. Koordinatenordnung. Für Vektoren $\mathbf{x} = [x_1, \dots, x_n]^T$ und $\mathbf{y} = [y_1, \dots, y_n]^T$ schreibt man

$$\mathbf{x} \leq \mathbf{y} \iff x_i \leq y_i \text{ für alle } i = 1, \dots, n$$

und

$$\mathbf{x} < \mathbf{y} \iff x_i < y_i \text{ für alle } i = 1, \dots, n.$$

NOTA BENE: Bei dieser Ordnungsrelation gibt es (im Fall $n \geq 2$) immer Vektoren $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, die nicht miteinander vergleichbar sind, d.h.

$$\mathbf{a} \not\leq \mathbf{b} \text{ und } \mathbf{b} \not\leq \mathbf{a}.$$

2.2. Lexikographische Ordnung. \mathbf{x} ist *lexikographisch kleiner* (Notation: $\mathbf{x} \prec \mathbf{y}$) als \mathbf{y} , wenn es einen Index $1 \leq \ell \leq n$ gibt mit der Eigenschaft

$$x_\ell < y_\ell \text{ und } x_j = y_j \text{ für alle } j < \ell.$$

LEMMA 0.2. Für beliebige $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ gilt genau eine der drei Aussagen:

- (0) $\mathbf{a} = \mathbf{b}$
- (1) $\mathbf{a} \prec \mathbf{b}$
- (2) $\mathbf{b} \prec \mathbf{a}$.

◇

2.3. Mengenoperationen.

2.3.1. *Minkowski-Summe.* Die *Minkowski-Summe* von $S, T \subseteq \mathbb{R}^n$ ist die Teilmenge

$$S + T = \{\mathbf{s} + \mathbf{t} \mid \mathbf{s} \in S, \mathbf{t} \in T\} \subseteq \mathbb{R}^n.$$

Im Spezialfall einer einelementigen Menge $T = \{\mathbf{t}\}$ erhält man die *Translation* von S um den Vektor \mathbf{t} :

$$S + \mathbf{t} = S + \{\mathbf{t}\} = \{\mathbf{s} + \mathbf{t} \mid \mathbf{s} \in S\}.$$

LEMMA 0.3. Die *Minkowski-Summe* zweier affiner Teilräume in \mathbb{R}^n ist selber ein affiner Teilraum.

◇

2.3.2. *Koordinatenprojektionen.* Sei $N = \{1, \dots, n\}$ und $\emptyset \neq I \subset N$. Für $\mathbf{x} \in \mathbb{R}^N$ bezeichnet \mathbf{x}_I die Restriktion von \mathbf{x} auf die Koordinaten in I .

In einer etwas lockeren (aber bequemen) Schreibweise haben wir dann:

$$\mathbf{x} = \mathbf{x}_N = \begin{bmatrix} \mathbf{x}_I \\ \mathbf{x}_J \end{bmatrix} \quad \text{mit } J = N \setminus I.$$

Diese Schreibweise ist auch vorteilhaft bei allgemeiner Matrixnotation:

$$A\mathbf{x} = A_N\mathbf{x}_N = A_I\mathbf{x}_I + A_{N \setminus I}\mathbf{x}_{N \setminus I}.$$

(Hier ist A_I natürlich die Restriktion von A auf die I entsprechenden Spalten.)

Für beliebiges $S \subseteq \mathbb{R}^n$ erhalten wir die *Projektion* $\pi_I(S)$ von S auf die Koordinatenmenge I als die Menge

$$\pi_I(S) = \{\mathbf{x}_I \mid \mathbf{x} \in S\} \subseteq \mathbb{R}^I.$$

Bsp. Sei $I = \{2, 3, \dots, n\}$. Dann gilt für $S \subseteq \mathbb{R}^n$:

$$\pi_I(S) = \{(x_2, x_3, \dots, x_n) \mid \exists x_1 \in \mathbb{R} : (x_1, x_2, x_3, \dots, x_n) \in S\}.$$

3. Topologie

Sei (\mathbf{x}^k) eine Folge von Vektoren $\mathbf{x}^k \in \mathbb{R}^n$. Wir schreiben

$$\mathbf{x}^k \rightarrow \mathbf{x} \quad \text{bzw.} \quad \mathbf{x} = \lim_{k \rightarrow \infty} \mathbf{x}^k,$$

wenn (\mathbf{x}^k) (komponentenweise) gegen $\mathbf{x} \in \mathbb{R}^n$ konvergiert. Bzgl. der *euklidischen Norm*

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{x_1^2 + \dots + x_n^2}$$

kann man das auch so ausdrücken:

$$\mathbf{x}^k \rightarrow \mathbf{x} \quad \iff \quad \|\mathbf{x}^k - \mathbf{x}\|^2 \rightarrow 0.$$

Eine Menge $S \subseteq \mathbb{R}^n$ heisst *abgeschlossen*, wenn für jede Folge (\mathbf{x}^k) mit $\mathbf{x}^k \in S$ gilt:

$$\mathbf{x}^k \rightarrow \mathbf{x} \quad \implies \quad \mathbf{x} \in S.$$

S ist *beschränkt*, wenn es eine Schranke $c > 0$ mit der Eigenschaft

$$\|\mathbf{x}\| \leq c \quad \forall \mathbf{x} \in S$$

gibt. Eine beschränkte und abgeschlossene Menge $S \subseteq \mathbb{R}^n$ ist *kompakt*.

3.1. Stetigkeit. Eine Funktion $f : S \rightarrow \mathbb{R}$ heisst *stetig*, wenn für alle $\mathbf{x} \in S$ und Folgen (\mathbf{x}^k) mit $\mathbf{x}^k \in S$ gilt:

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

Aus der Analysis weiss man:

LEMMA 0.4. Sei $\emptyset \neq S \subseteq \mathbb{R}^n$ kompakt und $f : S \rightarrow \mathbb{R}$ stetig. Dann existieren Punkte (Vektoren) $\mathbf{x}_{\min}, \mathbf{x}_{\max} \in S$ mit der Eigenschaft

$$f(\mathbf{x}_{\min}) \leq f(\mathbf{x}) \leq f(\mathbf{x}_{\max}) \quad \text{für alle } \mathbf{x} \in S.$$

◇

Quadratische und lineare Funktionen. Offenbar sind Summen und Produkte stetiger Funktionen wieder stetig. Also ist insbesondere jede *quadratische* Funktion $f : \mathbb{R}^n \rightarrow \mathbb{R}$, d.h. Funktion mit der Darstellung

$$f(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j - \sum_{k=1}^n c_k x_k$$

für geeignete skalare Koeffizienten a_{ij} und c_k , stetig. Im Fall $a_{ij} = 0$ für alle i, j heisst eine quadratische Funktion *linear*.

3.2. Gradienten. Sei $f : \mathbb{R}^n \rightarrow \mathbb{R}$ eine Funktion und $\mathbf{x}_0 \in \mathbb{R}^n$ ein Punkt, wo alle partiellen Ableitungen von f existieren. Dann bezeichnet man den (Zeilen-)Vektor der partiellen Ableitungen

$$\nabla f(\mathbf{x}_0) = \left[\frac{\partial f(\mathbf{x}_0)}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x}_0)}{\partial x_n} \right]$$

als den *Gradienten* von f an der Stelle \mathbf{x}_0 .

4. Mathematische Optimierungsprobleme

Ein „Optimierungsproblem“ ist im allgemeinen umgangssprachlich nicht so präzise formuliert, dass man es ohne weiteres mathematisch analysieren (und lösen) kann. Es muss zuallerst in ein „mathematisches“ Optimierungsproblem umformuliert werden.

Zu einem *mathematischen Optimierungsproblem* gehören:

- (1) eine Menge Ω (der sog. *Zulässigkeitsbereich*);
- (2) eine Menge W (der sog. *Wertebereich*) und ausserdem eine Funktion $f : \Omega \rightarrow W$ (die sog. *Zielfunktion*), welche die Elemente des Zulässigkeitsbereichs bewertet.

In dieser Vorlesung nehmen wir meist an:

- $W = \mathbb{R}$ und $\Omega \subseteq \mathbb{R}^n$ (für ein geeignetes n).

Die Optimierungsaufgabe ist dann so ausgedrückt:

$$\max_{\omega \in \Omega} f(\omega) \quad \text{oder} \quad \min_{\omega \in \Omega} f(\omega).$$

Um mit „ Ω “ überhaupt rechnerisch umgehen zu können, muss der Zulässigkeitsbereich numerisch spezifiziert werden. Oft sucht man dazu Funktionen $g_1, \dots, g_m : \mathbb{R}^n \rightarrow \mathbb{R}$ mit der Eigenschaft

$$\Omega = \{\mathbf{x} \in \mathbb{R}^n \mid g_1(\mathbf{x}) \leq 0, \dots, g_m(\mathbf{x}) \leq 0\}.$$

Die Funktionen $g_i(\mathbf{x})$ heissen dann *Restriktionsfunktionen* und das mathematische Optimierungsproblem wird dann z.B.

$$\max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{s.d.} \quad g_i(\mathbf{x}) \leq 0 \quad \forall i = 1, \dots, m.$$

Die Forderungen $g_i(\mathbf{x}) \leq 0$ sind die sog. *Nebenbedingungen* des Problems.

BEMERKUNG. Die Formulierung eines Optimierungsproblems aus dem Anwendungsbereich als mathematisches Optimierungsproblem ist im allgemeinen auf sehr viel verschiedene Arten möglich. Es ist nicht immer klar, welches „die beste“ ist.

Bsp. Es gibt n Objekte mit Gewichten a_1, \dots, a_n . Daraus sollen möglichst viele Objekte gewählt werden, sodass das Gesamtgewicht die gegebene Schranke b nicht überschreitet.

1. Formulierung: Repräsentiere die Objekte durch $(0, 1)$ -Variable x_i mit der Zielfunktion

$$f(x_1, \dots, x_n) = x_1 + \dots + x_n = \sum_{i=1}^n x_i$$

und erhalte

$$\max \sum_{i=1}^n x_i \quad \text{s.d.} \quad \sum_{i=1}^n a_i x_i \leq b$$

$$x_1, \dots, x_n \in \{0, 1\}.$$

2. Formulierung:

$$\max \sum_{i=1}^n x_i \quad \text{s.d.} \quad \sum_{i=1}^n a_i x_i \leq b$$

$$x_i(1 - x_i) = 0 \quad (i = 1, \dots, n).$$

In dieser Formulierung hat man $2n + 1$ Restriktionsfunktionen (und entsprechend viele Nebenbedingungen):

$$\begin{aligned}g_0(x_1, \dots, x_n) &= \left(\sum_{i=1}^n a_i x_i \right) - b \\g_i(x_1, \dots, x_n) &= +x_i(1 - x_i) \quad (i = 1, \dots, n) \\h_i(x_1, \dots, x_n) &= -x_i(1 - x_i) \quad (i = 1, \dots, n).\end{aligned}$$

KAPITEL 1

Konvexe Mengen und Funktionen

1. Konvexe Mengen

Eine (möglicherweise leere) Teilmenge $S \subseteq \mathbb{R}^n$ nennt man *konvex*, wenn sie – geometrisch gesprochen – mit je zwei Punkten auch deren Verbindungsgeradenstück enthält. Konkret heisst das:

$$\mathbf{x}, \mathbf{y} \in S \implies \mathbf{x} + \lambda(\mathbf{y} - \mathbf{x}) \in S \quad (0 \leq \lambda \leq 1).$$

Aus der Definition folgt sofort:

- \mathbb{R}^n ist konvex und beliebige Durchschnitte konvexer Mengen sind konvex.

Sei $X \subseteq \mathbb{R}^n$ eine beliebige Menge. Unter der *konvexen Hülle* von X verstehen wir die kleinste konvexe Menge $\text{conv}(X)$, die X enthält, d.h.

$$\text{conv}(X) = \bigcap \{S \subseteq \mathbb{R}^n \mid S \text{ konvex und } S \supseteq X\}.$$

LEMMA 1.1. $\text{conv}(X)$ besteht aus allen sog. konvexen Linearkombinationen

$$\mathbf{z} = \lambda_1 \mathbf{x}_1 + \dots + \lambda_k \mathbf{x}_k,$$

wobei $\mathbf{x}_1, \dots, \mathbf{x}_k \in X$ beliebig gewählt werden dürfen und die Koeffizienten λ_i eine sog. Wahrscheinlichkeitsverteilung bilden, d.h.

$$\lambda_1, \dots, \lambda_k \geq 0 \quad \text{und} \quad \lambda_1 + \dots + \lambda_k = 1.$$

Beweis. . Man rechnet ohne grosse Mühe nach, dass eine konvexe Menge auch die mit ihren Elementen gebildeten konvexen Linearkombinationen enthalten muss (s. Übungen).

Andererseits zeigt der Spezialfall $k = 2$, dass die Menge aller Konvexkombinationen einer Menge X eine konvexe Menge darstellt. Also muss sie genau die kleinste konvexe Menge sein, die X enthält.

◇

1.1. Konvexe Kegel. Eine nichtleere Teilmenge $K \subseteq \mathbb{R}^n$ ist ein *konvexer Kegel*, wenn gilt

$$\mathbf{x}, \mathbf{y} \in K \text{ und } \lambda_1, \lambda_2 \geq 0 \implies \lambda_1 \mathbf{x} + \lambda_2 \mathbf{y} \in K.$$

Man überzeugt sich leicht, dass der „konvexe Kegel“ K wirklich auch eine konvexe Menge ist. Ausserdem gilt offenbar immer $\mathbf{0} \in K$. Die Menge

$$\text{cone}(X) = \bigcap \{K \subseteq \mathbb{R}^n \mid K \text{ konvexer Kegel und } K \supseteq X\}$$

ist der kleinste konvexe Kegel, der die Teilmenge $X \subseteq \mathbb{R}^n$ enthält, und wird als die *konische Hülle* von S bezeichnet. Also haben wir z.B.

$$\text{cone}(\emptyset) = \{\mathbf{0}\}.$$

LEMMA 1.2. Sei $X \neq \emptyset$. Dann besteht $\text{cone}(X)$ aus allen konischen Linearkombinationen

$$\mathbf{z} = \lambda_1 \mathbf{x}_1 + \dots + \lambda_k \mathbf{x}_k,$$

von Vektoren $\mathbf{x}_1, \dots, \mathbf{x}_k \in X$ und nichtnegativen Koeffizienten $\lambda_i \geq 0$. ◇

BEMERKUNG. Allgemeiner ist ein *Kegel* eine nichtleere Menge K mit der Eigenschaft

$$\lambda \mathbf{x} \in K \quad \forall \mathbf{x} \in K, \lambda \geq 0.$$

Ein allgemeiner Kegel ist nicht unbedingt konvex. In dieser Vorlesung betrachten wir nur konvexe Kegel.

1.2. Polyeder. Ein *Polyeder* $P \subseteq \mathbb{R}^n$ ist eine Menge der Form

$$P = P(A, \mathbf{b}) = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} \leq \mathbf{b}\}$$

mit $A \in \mathbb{R}^{m \times n}$ für ein $m \in \mathbb{N}$ und $\mathbf{b} \in \mathbb{R}^m$. $P(A, \mathbf{b})$ ist die Lösungsmenge des *linearen Ungleichungssystems*

$$\begin{array}{rcccc} a_{11}x_1 & + & \dots & + & a_{1n} & \leq & b_1 \\ a_{21}x_1 & + & \dots & + & a_{2n} & \leq & b_2 \\ \vdots & & & & \vdots & & \\ a_{m1}x_1 & + & \dots & + & a_{mn} & \leq & b_m. \end{array}$$

Aus der Definition folgt z.B. sofort: \emptyset und \mathbb{R}^n sind Polyeder in \mathbb{R}^n (*Beweis?*).

Besteht A nur aus dem Zeilenvektor $\mathbf{a}^T = (a_1, \dots, a_n) \neq \mathbf{0}^T$, so ist

$$H = P(\mathbf{a}^T, b) = \{\mathbf{x} \in \mathbb{R}^n \mid a_1 x_1 + \dots + a_n x_n \leq b\}$$

ein *Halbraum*.

Seien $\mathbf{a}^T = (a_1, \dots, a_n) \neq \mathbf{0}^T$ und $b \in \mathbb{R}$ gegeben. Die Lösungsmenge

$$P(\mathbf{a}^T, b) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T \mathbf{x} \leq b\}$$

der linearen Ungleichung

$$a_1x_1 + \dots + a_nx_n \leq b$$

heißt *Halbraum*. Man sieht leicht, dass Halbräume konvexe (und abgeschlossene) Teilmengen des \mathbb{R}^n sind. Also findet man:

LEMMA 1.3. *Eine echte Teilmenge $P \subset \mathbb{R}^n$ ist ein Polyeder genau dann, wenn P ein Durchschnitt von endlich vielen Halbräumen ist.*

Beweis. $P(A, \mathbf{b})$ mit $A = [a_{ij}] \in \mathbb{R}^{m \times n}$ ist der Durchschnitt der den m Zeilen von $A\mathbf{x} \leq \mathbf{b}$ entsprechenden Halbräume

$$H_i = \{\mathbf{x} \in \mathbb{R}^n \mid a_{i1}x_1 + \dots + a_{in}x_n \leq b_i\} \quad (i = 1, \dots, m).$$

◇

Man erhält z.B. die *Hyperebene* $H = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T \mathbf{x} = b\}$ als Durchschnitt von zwei Halbräumen und somit als Polyeder:

$$H = P(\mathbf{a}^T, b) \cap P(-\mathbf{a}^T, -b).$$

BEMERKUNG. Auch die leere Menge $\emptyset \subseteq \mathbb{R}^n$ ist ein Polyeder, das man z.B. als Lösungsmenge von

$$\begin{aligned} x_1 + \dots + x_n &= +1 \\ x_1 + \dots + x_n &= -1 \end{aligned}$$

erhält. Man bemerke: *Die Darstellungsmatrix A und der Restriktionsvektor \mathbf{b} eines Polyeders P sind im allgemeinen nicht eindeutig bestimmt!*

Da Durchschnitte abgeschlossener Mengen abgeschlossen sind, ergibt sich:

- Polyeder sind konvexe und abgeschlossene Teilmengen des \mathbb{R}^n .

Ein *Polytop* ist ein beschränktes (und folglich kompaktes) Polyeder.

BEMERKUNG (POLYTOPE). Wir werden später beweisen, dass die konvexen Hüllen $\text{conv}(X)$ von endlichen(!) Mengen $X \subseteq \mathbb{R}^n$ genau die Polytope sind.

BEISPIEL 1.1. *Das sog. Standardsimplex ist das Polyeder*

$$\Delta_n = \{\mathbf{x} \in \mathbb{R}^n \mid x_1 + \dots + x_n = 1, x_i \geq 0\} = \text{conv}\{\mathbf{e}_1, \dots, \mathbf{e}_n\}.$$

Es besteht aus der Menge aller Wahrscheinlichkeitsverteilungen auf n Elementen, ist offensichtlich beschränkt (ist also ein Polytop) und wird erzeugt von den n Einheitsvektoren $\mathbf{e}_i \in \mathbb{R}^n$.

1.2.1. *Polyedrische Kegel.* Ein *polyedrischer Kegel* ist ein Polyeder, das zugleich ein Kegel ist.

LEMMA 1.4. *Polyedrische Kegel sind genau die Lösungsmengen von homogenen linearen Ungleichungssystemen. M.a.W.: Das Polyeder $P(A, \mathbf{b})$ ist ein polyedrischer Kegel genau dann, wenn*

$$P(A, \mathbf{b}) = P(A, \mathbf{0}) = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} \leq \mathbf{0}\}.$$

Beweis. Sei $P = P(A, \mathbf{b})$ ein Kegel. Wegen $\mathbf{0} \in P$ haben wir $\mathbf{0} = A\mathbf{0} \leq \mathbf{b}$, d.h. \mathbf{b} ist in jeder Komponente b_i nichtnegativ. Also finden wir

$$P(A, \mathbf{0}) = \{x \mid A\mathbf{x} \leq \mathbf{0}\} \subseteq \{x \mid A\mathbf{x} \leq \mathbf{b}\} = P(A, \mathbf{b}).$$

Angenommen, es gäbe ein $\mathbf{x} \in P(A, \mathbf{b}) \setminus P(A, \mathbf{0})$. Dann gibt es einen Zeilenvektor \mathbf{a}_i^T von A derart, dass

$$\mathbf{a}_i^T \mathbf{x} \leq b_i \quad \text{aber} \quad \mathbf{a}_i^T \mathbf{x} > 0.$$

Wir könnten also $\lambda > 0$ so gross wählen, dass

$$\mathbf{a}_i^T(\lambda\mathbf{x}) = \lambda(\mathbf{a}_i^T \mathbf{x}) > b_i.$$

Dann hätten wir aber $\lambda\mathbf{x} \notin P(A, \mathbf{b})$, was im Widerspruch zu der Kegeleigenschaft von $P(A, \mathbf{b})$ steht!

◇

1.2.2. *Geometrische Darstellungsweisen.* Sei $H = \{\mathbf{a}^T \mathbf{x} = b\}$ eine Hyperebene und $\mathbf{p} \in H$ ein beliebiger Punkt. Dann besteht H aus der Menge aller Punkte $\mathbf{x} \in \mathbb{R}^n$, bei denen der Differenzvektor $\mathbf{d} = \mathbf{x} - \mathbf{p}$ senkrecht auf dem sog. *Normalenvektor* \mathbf{a} steht:

$$H = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T(\mathbf{x} - \mathbf{p}) = 0\}.$$

Der Halbraum $P(\mathbf{a}^T, b)$ besteht aus den Punkten \mathbf{x} , bei denen der Differenzvektor $\mathbf{d} = \mathbf{x} - \mathbf{p}$ mit dem Normalenvektor \mathbf{y} einen stumpfen Winkel bildet:

$$P(\mathbf{a}^T, b) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T(\mathbf{x} - \mathbf{p}) \leq 0\}.$$

Der Halbraum $P(-\mathbf{a}^T, b)$ besteht aus den Punkten \mathbf{x} , bei denen der Differenzvektor $\mathbf{d} = \mathbf{x} - \mathbf{p}$ mit dem Normalenvektor \mathbf{a} einen spitzen Winkel bildet:

$$\begin{aligned} P(-\mathbf{a}^T, b) &= \{\mathbf{x} \in \mathbb{R}^n \mid -\mathbf{a}^T(\mathbf{x} - \mathbf{p}) \leq 0\} \\ &= \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T(\mathbf{x} - \mathbf{p}) \geq 0\}. \end{aligned}$$

BEMERKUNG. Ob man sich als Mensch überhaupt den Raum \mathbb{R}^n mit $n \geq 4$ geometrisch vorstellen kann, darf mit Fug und Recht bezweifelt werden. Der Philosoph KANT z.B. war davon überzeugt, dass die menschliche Vorstellungskraft bei $n = 3$ endet.

Für das Rechnen in Parameterräumen ist dieser Punkt jedoch irrelevant. In praktischen Anwendungen sind mathematische Modelle mit Hunderten oder Tausenden von Parametern (und damit entsprechend hohen Dimensionen) tagtägliches Brot.

1.3. Abgeschlossene konvexe Mengen. Sei $S \subseteq \mathbb{R}^n$ eine beliebige Teilmenge und $\mathbf{p} \in \mathbb{R}^n \setminus S$ ein festgewählter Punkt. Wir sagen, die Hyperebene $H = \{\mathbf{x} \mid \mathbf{a}^T \mathbf{x} = b\}$ trennt \mathbf{p} von S , wenn gilt

$$\mathbf{p} \notin P(\mathbf{a}^T, b) \quad \text{und} \quad S \subseteq P(\mathbf{a}^T, b).$$

Algebraisch ausgedrückt bedeutet dies:

$$\mathbf{a}^T \mathbf{p} > b \quad \text{und} \quad \mathbf{a}^T \mathbf{x} \leq b \quad \text{für alle } x \in S.$$

Offensichtlich kann man jeden Punkt \mathbf{p} in diesem Sinn von der leeren Menge \emptyset trennen. Andererseits findet man, dass man selbst konvexe Mengen und Punkte nicht immer trennen kann.

BEISPIEL 1.2. Sei $S = \{\mathbf{x} \in \mathbb{R}^2 \mid x_1^2 + x_2^2 < 1\}$ die offene Kreisscheibe. S ist konvex und $\mathbf{p} = (1, 0) \notin S$. Es gibt jedoch keine Hyperebene, die \mathbf{p} von S trennt.

Ein Hauptsatz über konvexe Mengen besagt allerdings, dass die Situation bei abgeschlossenen(!) konvexen Mengen günstiger ist: hier kann man immer trennen.

1.3.1. Der Hauptsatz. Sei $S \subseteq \mathbb{R}^n$ nichtleer, konvex und abgeschlossen und $\mathbf{p} \in \mathbb{R}^n \setminus S$. Sei $\mathbf{x}_0 \in S$ so, dass

$$\|\mathbf{x}_0 - \mathbf{p}\|^2 \leq \|\mathbf{x} - \mathbf{p}\|^2 \quad \forall \mathbf{x} \in S.$$

(Da S abgeschlossen und die Funktion $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{p}\|^2$ auf S stetig ist, wissen wir aus den Übungen, dass \mathbf{s}_0 existiert.) Wir setzen

$$\mathbf{a} = \mathbf{x}_0 - \mathbf{p} \quad (\neq \mathbf{0}).$$

Ein beliebiger Punkt $\mathbf{x} \in S$ kann in der Form $\mathbf{x} = \mathbf{x}_0 + \lambda \mathbf{d}$ mit $\|\mathbf{d}\| = 1$ geschrieben werden. Damit haben wir

$$\|\mathbf{a}\|^2 \leq f(\mathbf{x}) = (\mathbf{a} - \lambda \mathbf{d})^T (\mathbf{a} - \lambda \mathbf{d}) = \|\mathbf{a}\|^2 - 2\lambda \mathbf{a}^T \mathbf{d} + \lambda^2$$

Im Fall $\mathbf{x} \neq \mathbf{x}_0$ (d.h. $\lambda > 0$) bedeutet dies

$$\lambda/2 \geq \mathbf{a}^T \mathbf{d} \quad \text{und deshalb (mit } \lambda \rightarrow 0): \quad 0 \geq \mathbf{a}^T \mathbf{d}.$$

SATZ 1.1 (Hauptsatz über abgeschlossene konvexe Mengen). Sei $S \subseteq \mathbb{R}^n$ eine nichtleere konvexe und abgeschlossene Menge. Dann existiert zu jedem $\mathbf{p} \in \mathbb{R}^n \setminus S$ ein Vektor $\mathbf{a} \neq \mathbf{0}$ und ein Punkt $\mathbf{x}_0 \in S$ derart, dass gilt

$$(a) \quad \mathbf{a}^T \mathbf{x}_0 < \mathbf{a}^T \mathbf{p};$$

(b) $\mathbf{a}^T \mathbf{x}_0 \geq \mathbf{a}^T \mathbf{x}$ für alle $\mathbf{x} \in S$.

Beweis. Seien \mathbf{x}_0 und $\mathbf{a} = \mathbf{x}_0 - \mathbf{p}$ wie oben gewählt. Dann folgt (a) so:

$$\mathbf{a}^T \mathbf{x}_0 - \mathbf{a}^T \mathbf{p} = \mathbf{a}^T (\mathbf{x}_0 - \mathbf{p}) = \|\mathbf{a}\|^2 > 0.$$

Für $\mathbf{x} = \mathbf{x}_0 + \lambda \mathbf{d} \in S$ ergibt sich (b) analog:

$$\mathbf{a}^T \mathbf{x} - \mathbf{a}^T \mathbf{x}_0 = \mathbf{a}^T (\lambda \mathbf{d}) = \lambda \mathbf{a}^T \mathbf{d} \leq 0.$$

◇

Der Hauptsatz besagt, dass jeder Punkt $\mathbf{p} \in \mathbb{R}^n \setminus S$ durch eine Hyperebene von der abgeschlossenen konvexen Menge S getrennt werden kann. Folglich finden wir:

- Eine Menge $S \subseteq \mathbb{R}^n$ ist genau dann konvex und abgeschlossen, wenn sie sich als Durchschnitt von (möglicherweise unendlich vielen) Halbräumen darstellen lässt. (Dabei betrachten wir \mathbb{R}^n als „leeren Durchschnitt“.)

Algebraisch ausgedrückt können wir uns eine abgeschlossene konvexe Menge $S \in \mathbb{R}^n$ somit immer als die Lösungsmenge eines linearen Ungleichungssystems

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n \leq b_i \quad (i \in I)$$

vorstellen, wobei I eine (möglicherweise unendliche) Indexmenge ist.

BEISPIEL 1.3. Betrachten wir eine quadratische Matrix $X = [x_{ij}] \in \mathbb{R}^{n \times n}$ als Vektor mit n^2 Komponenten x_{ij} , können wir $\mathbb{R}^{n \times n}$ mit \mathbb{R}^{n^2} identifizieren. X ist genau dann symmetrisch, wenn das lineare Gleichungssystem

$$(1) \quad x_{ij} - x_{ji} = 0 \quad (1 \leq i < j \leq n)$$

erfüllt wird. Die Menge der symmetrischen Matrizen ist also ein konvexer und abgeschlossener polyedrischer Kegel (tatsächlich sogar ein linearer Teilraum), nämlich genau die Lösungsmenge des (endlichen) linearen Systems (1).

Eine symmetrische Matrix $X = [x_{ij}] \in \mathbb{R}^{n \times n}$ heisst positiv semidefinit, wenn für alle Parametervektoren $\mathbf{a} \in \mathbb{R}^n$ gilt:

$$(2) \quad \mathbf{a}^T X \mathbf{a} = \sum_{i=1}^n \sum_{j=1}^n a_i a_j x_{ij} \geq 0.$$

Auch die Menge aller positiv semidefiniten Matrizen ist ein konvexer und abgeschlossener Kegel als die Lösungsmenge des aus (1) und (2) zusammengesetzten unendlichen linearen Systems. (Man kann allerdings zeigen, dass dieser Kegel im allgemeinen nicht polyedrisch ist.)

1.3.2. *Stützhyperebenen.* Eine Hyperebene $H = \{\mathbf{x} \mid \mathbf{a}^T \mathbf{x} = b\}$ mit der Eigenschaft $S \subseteq P(\mathbf{a}^T, b)$ und $S \cap H \neq \emptyset$ heißt *Stützhyperebene* von S . Der Hauptsatz besagt somit, dass eine nichtleere abgeschlossene konvexe Menge Durchschnitt von Halbräumen ist, die zu Stützhyperebenen gehören.

Wir nennen einen Punkt $\mathbf{x}_0 \in S$ einen *Randpunkt* von S , wenn es eine Folge von Punkten $\mathbf{z}_k \in \mathbb{R}^n \setminus S$ gibt mit

$$\mathbf{z}_k \rightarrow \mathbf{x}_0, \quad \text{d.h.} \quad \lim_{k \rightarrow \infty} \|\mathbf{x}_0 - \mathbf{z}_k\| = 0.$$

SATZ 1.2. *Sei \mathbf{x}_0 Randpunkt der abgeschlossenen konvexen Menge $S \subseteq \mathbb{R}^n$. Dann existiert ein Vektor $\mathbf{c} \neq \mathbf{0}$ derart, dass*

$$\mathbf{c}^T \mathbf{x}_0 = \max_{\mathbf{x} \in S} \mathbf{c}^T \mathbf{x}.$$

Insbesondere ist $H = \{\mathbf{x} \mid \mathbf{c}^T \mathbf{x} = \mathbf{c}^T \mathbf{x}_0\}$ eine Stützhyperebene für S , die \mathbf{x}_0 enthält.

Beweis. Sei $\mathbf{z}_k \rightarrow \mathbf{x}_0$ mit $\mathbf{z}_k \notin S$. Dann gibt es nach dem Hauptsatz Vektoren $\mathbf{a}_k \neq \mathbf{0}$ und Zahlen b_k derart, dass für alle $\mathbf{x} \in S$ gilt:

$$\mathbf{a}_k^T \mathbf{x} \leq b_k < \mathbf{a}_k^T \mathbf{z}_k.$$

OBdA dürfen wir $\|\mathbf{a}_k\| = 1$ annehmen (sonst dividieren wir einfach die Ungleichungen jeweils durch $\|\mathbf{a}_k\|$). Die Folge der \mathbf{a}_k hat also in der kompakten Vollkugel $B_n = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| \leq 1\}$ einen Häufungspunkt \mathbf{c} . Also dürfen wir (möglicherweise durch Übergang auf eine konvergente Teilfolge) sogar oBdA annehmen:

$$\lim_{k \rightarrow \infty} \mathbf{a}_k = \mathbf{c}.$$

Daraus ergibt sich nun sogleich für $\mathbf{x} \in S$:

$$\mathbf{c}^T \mathbf{x} = \lim_{k \rightarrow \infty} \mathbf{a}_k^T \mathbf{x} \leq \lim_{k \rightarrow \infty} \mathbf{a}_k^T \mathbf{z}_k = \mathbf{c}^T \mathbf{x}_0.$$

◇

1.4. Gültige und implizierte Ungleichungen. Man sagt, dass eine lineare Ungleichung

$$\mathbf{c}^T \mathbf{x} \leq z \quad \iff \quad c_1 x_1 + \dots + c_n x_n \leq z$$

für die Menge $S \subseteq \mathbb{R}^n$ *gültig* ist, wenn sie von allen $\mathbf{s} \in S$ erfüllt wird. Geometrisch gesprochen bedeutet dies

$$S \subseteq P(\mathbf{c}^T, z).$$

Ist $A\mathbf{x} \leq \mathbf{b}$ ein lineares Ungleichungssystem (mit mindestens einer Lösung), so sagen wir, dass die Ungleichung $\mathbf{c}^T \mathbf{x} \leq z$ von $A\mathbf{x} \leq \mathbf{b}$ *impliziert* wird, wenn sie für das Polyeder $P(A, \mathbf{b})$ gültig ist, d.h. wenn für alle $\bar{\mathbf{x}} \in \mathbb{R}^n$ gilt:

$$A\bar{\mathbf{x}} \leq \mathbf{b} \quad \implies \quad \mathbf{c}^T \bar{\mathbf{x}} \leq z$$

SATZ 1.3 („Lemma von Farkas“). Sei $A \in \mathbb{R}^{m \times n}$ eine Matrix und $\mathbf{b} \in \mathbb{R}^m$ derart, dass das lineare Ungleichungssystem $A\mathbf{x} \leq \mathbf{b}$ mindestens einer Lösung $\mathbf{x}_0 \in P(A, \mathbf{b})$ besitzt. Sei ausserdem $\mathbf{c}^T \mathbf{x} \leq z$ eine beliebige lineare Ungleichung. Dann gilt genau eine der beiden Aussagen:

(I) Es gibt einen Parametervektor \mathbf{y} mit den Eigenschaften

$$\mathbf{y} \geq \mathbf{0}, \mathbf{c}^T = \mathbf{y}^T A \text{ und } \mathbf{y}^T \mathbf{b} \leq z.$$

(II) $\mathbf{c}^T \mathbf{x} \leq z$ ist nicht von $A\mathbf{x} \leq \mathbf{b}$ impliziert.

Die implizierten Ungleichungen sind also genau die Ungleichungen der Form (I).

Beweis. Wir betrachten den Punkt P mit Koordinatenvektor \mathbf{p} und den konvexen Kegel C , wobei

$$\mathbf{p} = \begin{bmatrix} \mathbf{c} \\ z \end{bmatrix} \quad \text{und} \quad C = \left\{ \begin{bmatrix} A^T \mathbf{y} \\ \zeta \end{bmatrix} \in \mathbb{R}^{n+1} \mid \mathbf{y} \geq \mathbf{0}, \zeta \geq \mathbf{b}^T \mathbf{y} \right\}.$$

Da die lineare Funktion $\mathbf{y} \mapsto A^T \mathbf{y}$ auf \mathbb{R}^m stetig ist, überzeugt man sich leicht, dass C sogar ein abgeschlossener konvexer Kegel ist (und insbesondere den Nullvektor $\mathbf{0}$ enthält).

I: Sei $P \in C$ und folglich $\mathbf{c}^T = \mathbf{y}^T A$ für ein geeignetes $\mathbf{y} \geq \mathbf{0}$. Dann gilt für alle $\bar{\mathbf{x}} \in P(A, \mathbf{b})$,

$$0 \leq \mathbf{y}^T (\mathbf{b} - A\bar{\mathbf{x}}) = \mathbf{y}^T \mathbf{b} - \mathbf{c}^T \bar{\mathbf{x}} \quad \text{d.h.} \quad \mathbf{c}^T \bar{\mathbf{x}} \leq \mathbf{y}^T \mathbf{b}.$$

Im Fall $P \in C$ ist $\mathbf{y}^T \mathbf{b} \leq z$ also $\mathbf{c}^T \mathbf{x} \leq z$ von $A\mathbf{x} \leq \mathbf{b}$ impliziert.

II: Sei $P \notin C$. Wir behaupten, dass dann $\mathbf{c}^T \mathbf{x} \leq z$ nicht für $P(A, \mathbf{b})$ gelten kann.

Wegen $P \notin C$ garantiert der Hauptsatz 1.1 ein $\mathbf{a}^T = (\mathbf{x}^T, a_{n+1})$ und ein $(\mathbf{y}_0^T A, \zeta_0)^T$ mit $\mathbf{y}_0 \geq \mathbf{0}$ und $\zeta_0 \geq \mathbf{b}^T \mathbf{y}_0$ so, dass gilt:

$$\begin{aligned} \mathbf{a}^T \mathbf{p} = \mathbf{p}^T \mathbf{a} &= \mathbf{c}^T \mathbf{x} + a_{n+1} z \\ &> \mathbf{y}_0^T A \mathbf{x} + a_{n+1} \zeta_0 \\ &\geq \mathbf{y}^T A \mathbf{x} + a_{n+1} \zeta \quad (\forall \mathbf{y} \geq \mathbf{0}, \zeta \geq \mathbf{b}^T \mathbf{y}). \end{aligned}$$

Falls $a_{n+1} = 0$, dann muss $A\mathbf{x} \leq \mathbf{0}$ erfüllt sein, wie aus der Beschränktheit folgt. Denn wir haben in diesem Fall

$$\mathbf{y}^T (A\mathbf{x}) \leq \mathbf{a}^T \mathbf{p} < \infty \quad \text{für alle } \mathbf{y} \geq \mathbf{0}.$$

Die Wahl $\mathbf{y} = \mathbf{0}$ zeigt zudem $\mathbf{c}^T \mathbf{x} > 0$. Wählen wir $\lambda > 0$ genügend gross, so erhalten wir

$$A(\mathbf{x}_0 + \lambda \mathbf{x}) = A\mathbf{x}_0 + \lambda A\mathbf{x} \leq \mathbf{b} \quad \text{und} \quad \mathbf{c}^T (\mathbf{x}_0 + \lambda \mathbf{x}) = \mathbf{c}^T \mathbf{x}_0 + \lambda \mathbf{c}^T \mathbf{x} > z.$$

Die Ungleichung gilt also nicht für alle Punkte in $P(A, \mathbf{b})$.

Im Fall $a_{n+1} \neq 0$ zeigt $\zeta \rightarrow \infty$ (wieder aus Gründen der Beschränktheit), dass $a_{n+1} < 0$ gelten muss. Mit $\mathbf{x}' := -\mathbf{x}/a_{n+1}$ haben wir

$$\mathbf{a}^T \mathbf{p} \geq \mathbf{y}^T (A\mathbf{x} + a_{n+1} \mathbf{b}) = (-a_{n+1}) \mathbf{y}^T (A\mathbf{x}' - \mathbf{b}) \quad \text{für alle } \mathbf{y} \geq \mathbf{0}.$$

Somit muss $A\mathbf{x}' \leq \mathbf{b}$ erfüllt sein. Ausserdem haben wir

$$\mathbf{c}^T \mathbf{x} + a_{n+1}z > 0 \quad \text{d.h.} \quad \mathbf{c}^T \mathbf{x} > -a_{n+1}z .$$

Daraus folgt (per Division durch $-a_{n+1} > 0$) die strikte Ungleichung $\mathbf{c}^T \mathbf{x}' > z$. Der Koordinatenvektor $\mathbf{x}' \in P(A, \mathbf{b})$ verletzt also die Ungleichung. \diamond

1.4.1. *Gleichungen.* Bei der Diskussion von linearen Ungleichungen und Ungleichungssystemen nehmen wir grundsätzlich die Form

$$A\mathbf{x} \leq \mathbf{b}$$

an. In diese Form können lineare Ungleichungen immer gebracht werden. Zum Beispiel sind folgende Ungleichungen äquivalent

$$a_1x_1 + \dots + a_nx_n \geq b \quad \longleftrightarrow \quad -a_1x_1 - \dots - a_nx_n \leq -b.$$

Auch bei von einem linearen System implizierten Ungleichungen können wir Gleichungen zulassen. Zum Beispiel ist eine *nichtnegative(!)* Linearkombination (mit Koeffizienten $y^+ \geq 0$ und $y^- \geq 0$) der Ungleichungen

$$\begin{aligned} y^+ : & a_1x_1 + \dots + a_nx_n \leq b \\ y^- : & -a_1x_1 - \dots - a_nx_n \leq -b \end{aligned}$$

äquivalent zu einer Multiplikation der Gleichung

$$y : a_1x_1 + \dots + a_nx_n = b$$

mit dem im Vorzeichen *unbeschränkten(!)* Skalar

$$y = y^+ - y^- , \quad \text{wobei} \quad y^+ = \max\{0, y\}, \quad y^- = \max\{0, -y\}$$

unterstellt werden darf.

2. Konvexe Funktionen

Sei $f : \mathcal{F} \rightarrow \mathbb{R}$ eine auf der Menge $\mathcal{F} \subseteq \mathbb{R}^n$ definierte reellwertige Funktion. Der *Epigraph* von f ist die Menge

$$\text{epi}(f) = \left\{ \begin{pmatrix} \mathbf{x} \\ z \end{pmatrix} \in \mathbb{R}^{n+1} \mid \mathbf{x} \in \mathcal{F}, z \geq f(\mathbf{x}) \right\}.$$

f heisst *konvex*, wenn der Epigraph $\text{epi}(f)$ eine konvexe Teilmenge in \mathbb{R}^{n+1} ergibt. Äquivalent kann man die Konvexität von f auch so definieren:

(KF1) \mathcal{F} muss eine konvexe Teilmenge von \mathbb{R}^n sein;

(KF2) Für alle $0 \leq \lambda \leq 1$ und $\mathbf{x}, \mathbf{y} \in \mathcal{F}$ muss gelten:

$$f(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})) \leq f(\mathbf{x}) + \lambda(f(\mathbf{y}) - f(\mathbf{x})).$$

Da wir uns in dieser Vorlesung nur auf konvexe Funktionen konzentrieren, von denen von vornherein klar ist, dass sie stetig sind, wird die nächste allgemeingültige Beobachtung nicht extra bewiesen¹.

LEMMA 1.5. *Jede auf einer offenen Menge $\mathcal{F} \subseteq \mathbb{R}^n$ definierte konvexe Funktion $f : \mathcal{F} \rightarrow \mathbb{R}$ ist stetig.*

◇

2.1. Subgradienten und Gradienten. Sei $f : \mathcal{F} \rightarrow \mathbb{R}$ eine auf der offenen Menge $\mathcal{F} \subseteq \mathbb{R}^n$ definierte Funktion und $\mathbf{x}_0 \in \mathcal{F}$. Ein Vektor \mathbf{d} ist ein sog. *Subgradient* von f an der Stelle \mathbf{x}_0 , wenn für alle $\mathbf{x} \in \mathcal{F}$ gilt

$$f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{d}^T(\mathbf{x} - \mathbf{x}_0).$$

$\partial f(\mathbf{x}_0)$ bezeichnet die (möglicherweise leere) Menge aller Subgradienten und heisst *Subdifferential*.

BEISPIEL 1.4. *Im Fall $\mathbf{0} \in \partial f(\mathbf{x}_0)$ haben wir für jedes $\mathbf{x} \in \mathcal{F}$:*

$$f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{0}^T(\mathbf{x} - \mathbf{x}_0) = 0 \quad \text{d.h.} \quad f(\mathbf{x}) \geq f(\mathbf{x}_0).$$

Das bedeutet: \mathbf{x}_0 minimiert die Funktion f über \mathcal{F} .

PROPOSITION 1.1. *Sei f an der Stelle \mathbf{x}_0 partiell differenzierbar. Dann gilt $\partial f(\mathbf{x}_0) = \emptyset$ oder $\partial f(\mathbf{x}_0) = \{\nabla f(\mathbf{x}_0)\}$, wobei $\nabla f(\mathbf{x}_0)$ der Gradient von f an der Stelle \mathbf{x}_0 ist:*

$$\nabla f(\mathbf{x}_0) = \left(\frac{\partial f(\mathbf{x}_0)}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x}_0)}{\partial x_n} \right).$$

Beweis. Sei $\mathbf{d} \in \partial f(\mathbf{x}_0)$. Dann finden wir z.B. für die i -te Komponente d_i :

$$\begin{aligned} \frac{\partial f(\mathbf{x}_0)}{\partial x_i} &= \lim_{t \downarrow 0} \frac{f(\mathbf{x}_0 + t\mathbf{e}_i) - f(\mathbf{x}_0)}{t} \geq \lim_{t \downarrow 0} \frac{\mathbf{d}^T(t\mathbf{e}_i)}{t} = d_i. \\ \frac{\partial f(\mathbf{x}_0)}{\partial x_i} &= \lim_{t \downarrow 0} \frac{f(\mathbf{x}_0 - t\mathbf{e}_i) - f(\mathbf{x}_0)}{-t} \leq \lim_{t \downarrow 0} \frac{\mathbf{d}^T(-t\mathbf{e}_i)}{-t} = d_i. \end{aligned}$$

◇

LEMMA 1.6. *Gilt $\partial f(\mathbf{x}_0) \neq \emptyset$ für alle $\mathbf{x}_0 \in \mathcal{F}$, dann ist $f : \mathcal{F} \rightarrow \mathbb{R}$ konvex.*

Beweis. Seien $\mathbf{x}, \mathbf{y} \in \mathcal{F}$ beliebig und $0 < \lambda < 1$. Wir setzen $\mathbf{z} = \mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})$ und wählen einen Subgradienten $\mathbf{d}_z \in \partial f(\mathbf{z})$. Dann gilt

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{z}) + \mathbf{d}_z^T(\mathbf{x} - \mathbf{z}) \\ f(\mathbf{y}) &\geq f(\mathbf{z}) + \mathbf{d}_z^T(\mathbf{y} - \mathbf{z}). \end{aligned}$$

¹Einen Beweis findet man z.B. in dem Buch FAIGLE/KERN/STILL: *Algorithmic Principles of Mathematical Programming* (Springer 2002)

Multiplizieren wir die erste Ungleichung mit $1 - \lambda > 0$ und die zweite mit $\lambda > 0$, so ergibt die Summe

$$f(\mathbf{x}) + \lambda(f(\mathbf{y}) - f(\mathbf{x})) \geq f(\mathbf{z}) + \mathbf{d}_z^T \mathbf{0} = f(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})).$$

◇

BEISPIEL 1.5. Sei die reelle Funktion $f : (a, b) \rightarrow \mathbb{R}$ differenzierbar und die Ableitungsfunktion $f' : (a, b) \rightarrow \mathbb{R}$ monoton steigend. Seien $a \leq x < y \leq b$ beliebig. Nach dem Mittelwertsatz existiert ein $x < \xi < y$ mit

$$f(y) - f(x) = f'(\xi)(y - x) \geq f'(x)(y - x).$$

Also finden wir $f'(x) \in \partial f(x)$. Im Fall $a < y < x < b$ schliessen wir (wegen $y - x < 0$):

$$f(y) - f(x) = f'(\xi)(y - x) \geq f'(x)(y - x).$$

Folglich ist $f'(x) \in \partial f(x)$ für alle x garantiert. Also ist f konvex.

SATZ 1.4. Die auf der offenen Menge \mathcal{F} definierte (stetige) Funktion $f : \mathcal{F} \rightarrow \mathbb{R}$ ist konvex genau dann, wenn $\partial f(\mathbf{x}) \neq \emptyset$ für jedes $\mathbf{x} \in \mathcal{F}$ garantiert ist.

Beweis. Nach Lemma 1.6 bleibt noch zu zeigen, dass im konvexen Fall $\partial f(\mathbf{x}) \neq \emptyset$ gewährleistet ist.

Der Vektor $\bar{\mathbf{x}} \in \text{epi}(f)$ liefert einen Rankpunkt, wobei

$$\bar{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ f(\mathbf{x}) \end{pmatrix} \quad \text{und} \quad \text{epi}(f) = \left\{ \begin{pmatrix} \mathbf{y} \\ z \end{pmatrix} \mid \mathbf{y} \in \mathcal{F}, z \geq f(\mathbf{y}) \right\}$$

Da $\text{epi}(f)$ konvex und abgeschlossen ist, existiert eine Stützhyperebene H für $\bar{\mathbf{x}}$, deren Koeffizientenvektor nun einen Vektor $\mathbf{c} \in \mathbb{R}^n$ und eine Zahl c_{n+1} liefert derart, dass $(\mathbf{c}, c_{n+1}) \neq (\mathbf{0}, 0)$ und für alle $\mathbf{y} \in \mathcal{F}$ und $z \geq f(\mathbf{y})$ gilt:

$$\mathbf{c}^T \mathbf{y} + c_{n+1} z \leq \mathbf{c}^T \mathbf{x} + c_{n+1} f(\mathbf{x}).$$

Die Überlegung $z \rightarrow +\infty$ zeigt $c_{n+1} \leq 0$. $c_{n+1} = 0$ ist unmöglich. Denn sonst hätten wir $\mathbf{c}^T (\mathbf{y} - \mathbf{x}) \leq 0$ für alle $\mathbf{y} \in \mathcal{F}$. Da \mathcal{F} offen ist, würde daraus $\mathbf{c} = \mathbf{0}$ folgen – im Widerspruch zur Wahl von $(\mathbf{c}, c_{n+1}) \neq (\mathbf{0}, 0)$.

Also wissen wir $c_{n+1} < 0$ und erhalten mit $\mathbf{d} = -\mathbf{c}/c_{n+1}$ und $z = f(\mathbf{y})$ die Subgradienteneigenschaft

$$f(\mathbf{y}) - \mathbf{d}^T \mathbf{y} \geq f(\mathbf{x}) - \mathbf{d}^T \mathbf{x} \quad \text{bzw.} \quad f(\mathbf{y}) - f(\mathbf{x}) \geq \mathbf{d}^T (\mathbf{x} - \mathbf{y}).$$

◇

3. Konvexe Optimierung

Wir verstehen unter einem *konvexen Optimierungsproblem* eine Aufgabe der Form

$$\min_{\omega \in \Omega} f(\omega),$$

wobei $f : \Omega \rightarrow \mathbb{R}$ eine konvexe Funktion ist. Die Standardformulierung beinhaltet also ein Minimierungsproblem. Natürlich kann man auch Maximierungsprobleme betrachten:

$$\max_{\omega \in \Omega} g(\omega).$$

Ein solches Problem ist dann aber nur dann ein „konvexes Optimierungsproblem“, wenn die Funktion $f(\omega) = -g(\omega)$ konvex ist. Denn man hat die Äquivalenz

$$\max_{\omega \in \Omega} g(\omega) \quad \longleftrightarrow \quad \min_{\omega \in \Omega} -g(\omega).$$

NOTA BENE. Eine auf einer konvexen Menge Ω definierte lineare Funktion $f(\omega)$ hat die bemerkenswerte Eigenschaft, dass sowohl $f(\omega)$ als auch $g(\omega) = -f(\omega)$ eine konvexe Funktion ist. Maximieren und Minimieren ist also im linearen Fall strukturell völlig äquivalent!

3.1. Konvexe Optimierung mit linearen Nebenbedingungen. Wir untersuchen das Problem, eine konvexe Funktion $f : \mathcal{F} \rightarrow \mathbb{R}$ über einem durch ein Ungleichungssystem $A\mathbf{x} \leq \mathbf{b}$ präsentiertes Polyeder zu minimieren:

$$(3) \quad \min_{\mathbf{x} \in \mathcal{F}} f(\mathbf{x}) \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}.$$

Mit dem Zulässigkeitsbereich $\Omega = \mathcal{F} \cap P(A, \mathbf{b})$ ist dies genau das mathematische Optimierungsproblem

$$\min_{\mathbf{x} \in \Omega} f(\mathbf{x}).$$

BEISPIEL 1.6 (Linear Programme). Seien die Matrix $A = [a_{ij}] \in \mathbb{R}^{m \times n}$ und die Vektoren $\mathbf{c} \in \mathbb{R}^n$ und $\mathbf{b} \in \mathbb{R}^m$ gegeben. Dann ist das Optimierungsproblem

$$(4) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}$$

ein sog. lineares Programm. Die konvexe Zielfunktion $f(\mathbf{x}) = \mathbf{c}^T \mathbf{x}$ ist hier auf $\mathcal{F} = \mathbb{R}^n$ definiert und besitzt den konstanten Gradienten $\mathbf{c}^T = \nabla f(\mathbf{x})$ (für alle $\mathbf{x} \in \mathbb{R}^n$).

Wir suchen Eigenschaften, die eine zulässige Lösung $\mathbf{x}^* \in P(A, \mathbf{b})$ des Optimierungsproblems (3) als optimal charakterisieren. $A = [a_{ij}] \in \mathbb{R}^{m \times n}$ und $\mathbf{b}^T = [b_1, \dots, b_m]$ werden als bekannte Parameter vorausgesetzt.

3.1.1. *Zulässige Richtungen.* Wir nehmen an, dass alle partiellen Ableitungen von f existieren und (in einer Umgebung von \mathbf{x}^*) stetig sind. Ein Vektor $\mathbf{d} \neq \mathbf{0}$ heisst *zulässige Richtung* für $\mathbf{x}^* \in P(A, \mathbf{b})$, wenn es ein $\varepsilon > 0$ gibt mit der Eigenschaft $\mathbf{x} = \mathbf{x}^* + \varepsilon \mathbf{d} \in P(A, \mathbf{b})$.

SATZ 1.5. $\mathbf{x}^* \in P(A, \mathbf{b})$ ist optimal für (3) genau dann, wenn

$$\nabla f(\mathbf{x}^*)\mathbf{d} \geq 0 \quad \text{für jede zulässige Richtung } \mathbf{d}.$$

Beweis. Sei \mathbf{x}^* optimal und \mathbf{d} eine zulässige Richtung. Wir setzen $\mathbf{x}_t = \mathbf{x}^* + t\mathbf{d}$ (für genügend kleines $t > 0$). Dann gilt

$$0 \geq f(\mathbf{x}^*) - f(\mathbf{x}_t) \geq \nabla f(\mathbf{x}_t)(\mathbf{x}^* - \mathbf{x}_t) = (-t)\nabla f(\mathbf{x}_t)\mathbf{d}.$$

und deshalb $\nabla f(\mathbf{x}_t)\mathbf{d} \geq 0$. Aus Stetigkeitsgründen ergibt sich daraus die Notwendigkeit der Bedingung:

$$\nabla f(\mathbf{x}^*)\mathbf{d} = \lim_{t \rightarrow 0} \nabla f(\mathbf{x}_t)\mathbf{d} \geq 0.$$

Dass die Bedingung hinreicht, um Optimalität zu garantieren, ist klar. Denn jedes $\mathbf{x} \in P(A, \mathbf{b})$ hat dann die Eigenschaft

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \geq 0 \quad \text{d.h.} \quad f(\mathbf{x}) \geq f(\mathbf{x}^*).$$

◇

3.1.2. *Die KKT-Bedingungen.* Man kann Satz 1.5 auch über die Charakterisierung implizierter Ungleichungen formulieren. Das ergibt dann die sog. *KKT-Bedingungen*². Dazu charakterisieren wir zuerst die zulässigen Richtungen als Lösungen eines linearen Ungleichungssystems.

LEMMA 1.7. \mathbf{d} ist eine zulässige Richtung (für \mathbf{x}^*) genau dann, wenn für alle Indizes $i = 1, \dots, m$ gilt:

$$\mathbf{a}_i^T \mathbf{x}^* = a_{i1}x_1^* + \dots + a_{in}x_n^* = b_i \quad \implies \quad \mathbf{a}_i^T \mathbf{d} \leq 0.$$

Beweis. Die Bedingung ist offenbar notwendig. Sei umgekehrt

$$J(\mathbf{x}^*) = \{1 \leq i \leq m \mid \mathbf{a}_i^T \mathbf{x}^* = b_i\}$$

die Menge der Zeilenindizes, wo \mathbf{x}^* das Ungleichungssystem mit Gleichheit erfüllt. Wir wählen

$$\varepsilon = \min_{i \notin J(\mathbf{x}^*)} \{t \mid \mathbf{a}_i^T \mathbf{x}^* + t\mathbf{a}_i^T \mathbf{d} = b_i\} > 0.$$

Dann folgt $A(\mathbf{x}^* + \varepsilon \mathbf{d}) = A\mathbf{x}^* + \varepsilon A\mathbf{d} \leq \mathbf{b}$, da keine der einzelnen Ungleichungen verletzt ist.

◇

Fassen wir die Zeilenvektoren \mathbf{a}_i^T von A mit $i \in J(\mathbf{x}^*)$ zu der Matrix $A_{J(\mathbf{x}^*)}$ zusammen, so sagt Satz 1.5:

$$A_{J(\mathbf{x}^*)}\mathbf{d} \leq \mathbf{0} \quad \implies \quad -\nabla f(\mathbf{x}^*)\mathbf{d} \leq 0.$$

²benannt nach KARUSH, KUHN und TUCKER

Aus dem Satz über implizierte Ungleichungen folgt, dass dies genau dann der Fall ist, wenn der negative Gradient $-\nabla f(\mathbf{x}^*)$ eine nichtnegative Linearkombination der Zeilen von $A_{J(\mathbf{x}^*)}$ ist. Anders ausgedrückt:

Es gibt Parameter $y_1^* \geq 0, \dots, y_m^* \geq 0$ derart dass

$$(KS) \quad y_i^* > 0 \implies \mathbf{a}_i^T \mathbf{x}^* = b_i;$$

$$(GB) \quad -\nabla f(\mathbf{x}^*) = \sum_{i=1}^m y_i^* \mathbf{a}_i^T = \sum_{j \in J(\mathbf{x}^*)} y_j^* \mathbf{a}_j^T.$$

BEMERKUNG. (KS) ist die Bedingung des sog. *komplementären Schlupfes*: Bei der Restriktion $\mathbf{a}_i^T \mathbf{x}^* \leq b_i$ darf kein(!) Schlupf bestehen, wenn der Multiplikator y_i^* nichttrivial ist.

Diese Überlegungen zusammenfassend erhalten wir

SATZ 1.6 (KKT-Bedingungen). *Der Vektor $\mathbf{x} \in \mathbb{R}^n$ ist genau dann optimal für das Problem (3), wenn gilt*

$$(P) \quad A\mathbf{x} \leq \mathbf{b};$$

$$(D) \quad \text{Es gibt ein } \mathbf{y}^T = (y_1, \dots, y_m) \geq \mathbf{0}^T \text{ derart, dass}$$

$$(D1) \quad \mathbf{y}^T (A\mathbf{x} - \mathbf{b}) = 0;$$

$$(D2) \quad \nabla f(\mathbf{x}) + \mathbf{y}^T A = \mathbf{0}^T.$$

◇

BEMERKUNG. (D1) ist die sog. *primale* und (D2) die *duale* Bedingung.

NOTA BENE: Das Optimierungsproblem (3) reduziert sich auf das Auffinden eines Lösungspaares $(\mathbf{x}^*, \mathbf{y}^*)$ des (im allgemeinen nichtlinearen!) Ungleichungssystems

$$(5) \quad \boxed{\begin{array}{rcl} A\mathbf{x} & \leq & \mathbf{b} \\ -\nabla f(\mathbf{x}) & = & \mathbf{y}^T A \\ \mathbf{y}^T A\mathbf{x} & = & \mathbf{y}^T \mathbf{b} \\ \mathbf{y} & \geq & \mathbf{0} \end{array}}$$

BEISPIEL 1.7 (Lineare Programme). *Im Fall des linearen Programms*

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}$$

ergibt sich wegen $\mathbf{c}^T = \nabla(\mathbf{c}^T \mathbf{x})$ das KKT-System als ein System linearer(!) Ungleichungen, das es (in den Variablen \mathbf{x} und \mathbf{y}) zu lösen gilt:

$$\boxed{\begin{array}{rcl} A\mathbf{x} & \leq & \mathbf{b} \\ & - & A^T \mathbf{y} = \mathbf{c} \\ \mathbf{c}^T \mathbf{x} & + & \mathbf{b}^T \mathbf{y} = 0 \\ & & \mathbf{y} \geq \mathbf{0} \end{array}}$$

3.1.3. *Gleichheitsrestriktionen.* Bei linearen Gleichheitsrestriktionen $A\mathbf{x} = \mathbf{b}$ vereinfachen sich auch die KKT-Bedingungen. Wir haben schon bei der Diskussion implizierter Ungleichungen bemerkt, dass eine Gleichheitsrestriktion einer Multiplikation mit einem im Vorzeichen unbeschränkten Skalar y entspricht.

Also erhalten wir für das Problem

$$\min_{\mathbf{x} \in \mathcal{F}} f(\mathbf{x}) \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b}$$

das KKT-System

$$\boxed{\begin{array}{l} A\mathbf{x} = \mathbf{b} \\ -\nabla f(\mathbf{x}) = \mathbf{y}^T A \end{array}}$$

3.1.4. *Ein quadratisches Problem.* Man sucht einen Punkt \mathbf{p} in dem Polyeder $P(A, \mathbf{b})$ mit dem kürzesten Abstand zum Ursprung:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{x}\|^2 \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}.$$

Die äquivalente Zielfunktion $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 = \frac{1}{2}(\mathbf{x}^T \mathbf{x})$ ist konvex und hat den Gradienten

$$\nabla f(\mathbf{x}) = (x_1, \dots, x_n) = \mathbf{x}^T$$

Nach den KKT-Bedingungen ist also das folgende (nichtlineare!) System zu lösen:

$$(6) \quad \left\{ \begin{array}{l} A\mathbf{x} \leq \mathbf{b} \\ -\mathbf{x}^T = \mathbf{y}^T A \\ \mathbf{y}^T A\mathbf{x} = \mathbf{y}^T \mathbf{b} \\ \mathbf{y} \geq \mathbf{0} \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{l} \bar{A}\mathbf{y} \leq \mathbf{b} \\ \mathbf{y}^T \bar{A}\mathbf{y} = \mathbf{b}^T \mathbf{y} \\ \mathbf{y} \geq \mathbf{0} \end{array} \right\}$$

wobei $\bar{A} = -AA^T$.

Auch hier ist die Situation bei Gleichheitsrestriktionen $A\mathbf{x} = \mathbf{b}$ wesentlich einfacher und führt auf die KKT-Bedingungen

$$(7) \quad \left\{ \begin{array}{l} A\mathbf{x} = \mathbf{b} \\ -\mathbf{x}^T = \mathbf{y}^T A \end{array} \right\} \longleftrightarrow \bar{A}\mathbf{y} = \mathbf{b}.$$

BEMERKUNG. (7) ist ein lineares Gleichungssystem und somit leicht zu lösen. Jedoch ist bei echten Ungleichungen die Lösung von (6) in der Praxis eine alles andere als triviale Aufgabe.

OPTIMIERUNG OHNE NEBENBEDINGUNGEN. Sei $f : \mathcal{F} \rightarrow \mathbb{R}$ auf der offenen Menge $\mathcal{F} \subseteq \mathbb{R}^n$ definiert und konvex. Dann ist bei jedem Punkt $\mathbf{x}^* \in \mathcal{F}$ jedes(!) $\mathbf{d} \in \mathbb{R}^n$ eine zulässige Richtung. Also folgern wir aus Satz 1.5:

KOROLLAR 1.1. *Unter den obigen Voraussetzungen gilt für jedes $\mathbf{x}^* \in \mathcal{F}$:*

$$f(\mathbf{x}^*) = \min_{\mathbf{x} \in \mathcal{F}} f(\mathbf{x}) \iff \nabla f(\mathbf{x}^*) = \mathbf{0}^T.$$

◇

3.1.5. *Das Regressionsproblem.* Man sucht die „beste“ Lösung des linearen Gleichungssystems

$$A\mathbf{x} = \mathbf{b}$$

zu bestimmen. Das soll heißen, man sucht eine Lösung des Problems

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - A\mathbf{x}\|^2 = \mathbf{b}^T \mathbf{b} - 2\mathbf{b}^T A\mathbf{x} + \mathbf{x}^T A^T A\mathbf{x}.$$

Setzen wir $\mathbf{c}^T = \mathbf{b}^T A$ und $Q = A^T A$, dann ist das Problem äquivalent mit

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T Q \mathbf{x} - \mathbf{c}^T \mathbf{x}.$$

$Q = A^T A$ ist positiv semidefinit und folglich f konvex. Also finden wir:

- $\mathbf{x} \in \mathbb{R}^n$ löst das Regressionsproblem genau dann, wenn gilt:

$$Q\mathbf{x} = \mathbf{c} \quad \text{bzw.} \quad A^T A\mathbf{x} = A^T \mathbf{b}.$$

Das Regressionsproblem reduziert sich also auf das Lösen des linearen Gleichungssystems $Q\mathbf{x} = \mathbf{c}$.

BEISPIEL 1.8 (Interpolation). Seien von $f : \mathbb{R} \rightarrow \mathbb{R}$ nur die Werte $y_i = f(t_i)$ bei den Stützstellen t_1, \dots, t_n bekannt. Man sucht eine Linearkombination

$$\hat{f}(t) = \sum_{j=1}^n a_j f_j(t)$$

von gegebenen Funktionen $f_1(t), \dots, f_m(t)$, die f an den Stützstellen bestmöglich interpoliert. D.h. man will die beste Lösung (in den Unbekannten a_1, \dots, a_n) des linearen Gleichungssystems

$$\begin{array}{ccccccc} a_1 f_1(t_1) & + & a_2 f_2(t_1) & + & \dots & + & a_n f_n(t_1) & = & y_1 \\ a_1 f_1(t_2) & + & a_2 f_2(t_2) & + & \dots & + & a_n f_n(t_2) & = & y_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_1 f_1(t_m) & + & a_2 f_2(t_m) & + & \dots & + & a_n f_n(t_m) & = & y_m \end{array}$$

Im Fall $\{f_1(t), f_2(t)\} = \{1, t\}$ spricht man auch von *linearer Regression* und nennt

$$\hat{f}(t) = a_1 + a_2 t$$

die *Regressionsgerade*. Im Fall $\{f_1(t), f_2(t), f_3(t)\} = \{1, t, t^2\}$ erhält man das *quadratische Regressionspolynom*

$$\hat{f}(t) = a_1 + a_2 t + a_3 t^2.$$

3.1.6. *Nichtkonvexe Zielfunktionen.* Sei $\mathcal{F} \subseteq \mathbb{R}^n$ offen und $f : \mathcal{F} \rightarrow \mathbb{R}$ stetig differenzierbar (aber nicht notwendigerweise konvex). Dann gilt die obige KKT-Analyse des Optimierungsproblems mit linearen Nebenbedingungen

$$(8) \quad \min_{\mathbf{x} \in \mathcal{F}} f(\mathbf{x}) \quad \text{s.d.} \quad \mathbf{Ax} \leq \mathbf{b}$$

nach wie vor – mit einer Ausnahme:

- Wenn Konvexität nicht garantiert ist, können wir nicht beweisen, dass die KKT-Bedingungen hinreichend für Optimalität sind.

Also:

SATZ 1.7. Wenn $\mathbf{x} \in \mathcal{F}$ eine Optimallösung von (8) ist, dann muss es notwendigerweise einen Parametervektor \mathbf{y} geben, sodass das Paar (\mathbf{x}, \mathbf{y}) die KKT-Bedingungen erfüllt.

◇

3.2. Der Optimierungsansatz von LAGRANGE. LAGRANGE betrachtet ein allgemeines Optimierungsproblem der Form

$$(9) \quad \min_{\mathbf{x} \in \mathcal{F}} f(\mathbf{x}) \quad \text{s.d.} \quad g_1(\mathbf{x}) \leq 0, \dots, g_m(\mathbf{x}) \leq 0,$$

wobei Funktionen $f, g_1, \dots, g_m : \mathcal{F} \rightarrow \mathbb{R}$ als gegeben vorausgesetzt werden. Die Restriktionsfunktionen g_i werden zur einfacheren Notation oft in die vektorwertige Funktion $\mathbf{g} : \mathcal{F} \rightarrow \mathbb{R}^m$ mit

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix}$$

zusammengefasst, sodass man (9) in der folgenden Form schreiben kann

$$\min_{\mathbf{x} \in \mathcal{F}} f(\mathbf{x}) \quad \text{s.d.} \quad \mathbf{g}(\mathbf{x}) \leq \mathbf{0}.$$

BEISPIEL 1.9. *Lineare Nebenbedingungen* $\mathbf{Ax} \leq \mathbf{b}$ ergeben im obigen Modell

$$\mathbf{g}(\mathbf{x}) = \mathbf{Ax} - \mathbf{b}.$$

LAGRANGE versucht, die Aufgabe (9) auf ein Optimierungsproblem ohne Nebenbedingungen zurückzuführen. Dazu führt er für jede Nebenbedingung $g_i(\mathbf{x}) \leq 0$ eine Straftgröße y_i (den sog. LAGRANGESchen Multiplikator) ein, die aktiv wird, wenn die Nebenbedingung verletzt wird.

Man betrachtet also die *Lagrangefunktion*

$$L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \sum_{i=1}^m y_i g_i(\mathbf{x}) \quad \text{mit} \quad \mathbf{y} \geq \mathbf{0}$$

und fragt, unter welchen Umständen das folgende gilt:

$$(10) \quad \min \{L(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in \mathbb{R}^n\} = \min \{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}.$$

3.3. Allgemeine KKT-Bedingungen. Sei $\mathbf{x}^* \in \mathbb{R}^n$ beliebig. Dann gilt (unter der Annahme, dass die Funktionen f, g_1, \dots, g_m stetig differenzierbar sind):

$$\begin{aligned} (1) \quad \mathbf{y}^T \mathbf{g}(\mathbf{x}^*) = 0 &\implies L(\mathbf{x}^*, \mathbf{y}) = f(\mathbf{x}^*). \\ (2) \quad L(\mathbf{x}^*, \mathbf{x}) = \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}) &\implies \nabla_{\mathbf{x}} L(\mathbf{x}^*, \mathbf{y}) = 0. \end{aligned}$$

Im Fall $\mathbf{g}(\mathbf{x}^*) \leq \mathbf{0}$ ist die Eigenschaft $\mathbf{y}^T \mathbf{g}(\mathbf{x}^*) = \sum_{i=1}^m y_i g_i(\mathbf{x}^*) = 0$ einfach die komplementäre Schlupfbedingung:

$$y_i > 0 \implies g_i(\mathbf{x}^*) = 0 \quad (i = 1, \dots, m).$$

Die zweite Bedingung ergibt die Gleichung

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \mathbf{y}) = \nabla f(\mathbf{x}^*) + \sum_{i=1}^m y_i \nabla g_i(\mathbf{x}^*) = 0.$$

Zusammengenommen erhalten wir die sog. *allgemeinen KKT-Bedingungen*:

$$(11) \quad \begin{array}{rcl} \mathbf{g}(\mathbf{x}) & \leq & \mathbf{0} \\ \mathbf{y}^T \mathbf{g}(\mathbf{x}) & = & 0 \\ -\nabla f(\mathbf{x}) & = & \mathbf{y}^T \nabla \mathbf{g}(\mathbf{x}) \\ \mathbf{y} & \geq & \mathbf{0} \end{array}$$

VORSICHT: Im Gegensatz zur konvexen Optimierung unter linearen Nebenbedingungen ist bei allgemeinen nichtlinearen Optimierungsproblemen das Erfülltsein der KKT-Bedingungen weder hinreichend noch notwendig für Optimalität!

TROTZDEM: Die Erfahrung zeigt, dass ein KKT-Punkt $(\mathbf{x}^*, \mathbf{y}^*)$ oft eine sehr gute Lösung ergibt. Deshalb ist die Strategie vieler Algorithmen für nichtlineare Probleme:

- Suche ein Punktepaar $(\mathbf{x}^*, \mathbf{y}^*)$, das die KKT-Bedingungen erfüllt.

BEISPIEL 1.10. Die Lagrangefunktion des linearen Programms

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}$$

ist (mit $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$):

$$L(\mathbf{x}, \mathbf{y}) = \mathbf{c}^T \mathbf{x} + \mathbf{y}^T (\mathbf{A}\mathbf{x} - \mathbf{b}) = \mathbf{c}^T \mathbf{x} + \mathbf{y}^T \mathbf{b} - \mathbf{y}^T \mathbf{A}\mathbf{x}.$$

Wegen $\nabla g(\mathbf{x}) = \mathbf{A}$ und $\nabla f(\mathbf{x}) = \mathbf{c}^T$ erhalten wir die uns für lineare Programme schon bekannten KKT-Bedingungen. (Nachrechnen!)

BEISPIEL 1.11. Die allgemeinen KKT-Bedingungen für das (nichtlineare und nicht-konvexe!) Optimierungsproblem

$$\min_{x, y, z \in \mathbb{R}} xyz \quad \text{s.d.} \quad x^2 + y^2 + z^2 = 1$$

ergeben sich (mit $f(x, y, z) = xyz$ und $h(x, y, z) = x^2 + y^2 + z^2 - 1$) als

$$\begin{aligned} h(x, y, z) &= 0 \\ \lambda h(x, y, z) &= 0 \\ \nabla f(x, y, z) + \lambda \nabla h(x, y, z) &= (0, 0, 0). \end{aligned}$$

λ ist hier ein (wegen der Gleichheitsrestriktion $h(x, y, z) = 0$) im Vorzeichen nicht beschränkter reeller Parameter.

4. Dualität

Seien X und Y beliebige Mengen und $L : X \times Y \rightarrow \mathbb{R}$ sei eine beliebige Funktion. Wir nehmen an, es gibt zwei Akteure P und D folgender Art:

(P) P will L durch Wahl eines $x \in X$ minimieren.

(D) D will L durch Wahl eines $y \in Y$ maximieren.

$(x^*, y^*) \in X \times Y$ heisst *Sattelpunkt* von L , wenn für alle $x \in X$ und $y \in Y$ gilt:

$$L(x^*, y) \leq L(x^*, y^*) \leq L(x, y^*).$$

BEMERKUNG. In der Sprache der ökonomischen Spieltheorie liegt hier ein sog. *Nullsummenspiel mit 2 Personen* vor. Ein Sattelpunkt (x^*, y^*) heisst dann *Nash-Gleichgewicht* mit folgender Interpretation: Wählt P die Aktion $x^* \in X$, dann kann D nichts besseres tun als zu $y^* \in Y$ zu greifen (und ebenso umgekehrt).

Die Suche nach einem Sattelpunkt (Nash-Gleichgewicht) stellt sich für den Optimierer so dar:

Wir definieren Funktionen $L_1 : Y \rightarrow \mathbb{R} \cup \{-\infty\}$ und $L_2 : X \rightarrow \mathbb{R} \cup \{+\infty\}$:

$$L_1(y) = \min_{x' \in X} L(x', y) \quad \text{und} \quad L_2(x) = \max_{y' \in Y} L(x, y').$$

Offensichtlich gilt immer

$$L_1(y) \leq L(x, y) \leq L_2(x).$$

Aus der Definition ergibt sich ausserdem sofort:

$$\boxed{(x^*, y^*) \text{ ist ein Sattelpunkt} \iff L_1(y^*) = L_2(x^*)}.$$

Primale und duale Probleme. Ist (x^*, y^*) Sattelpunkt, dann gilt

$$L_1(y^*) = \max_{y \in Y} L_1(y) \quad \text{und} \quad L_2(x^*) = \min_{x \in X} L_2(x).$$

Wir interessieren uns deshalb für das sog. *primale Problem*

$$(12) \quad \min_{x \in X} L_2(x) = \min_{x \in X} \max_{y \in Y} L(x, y).$$

Analog ist das zugehörige *duale Problem* definiert:

$$(13) \quad \max_{y \in Y} L_1(y) = \max_{y \in Y} \min_{x \in X} L(x, y).$$

Schwache und starke Dualität. Unter der *Dualitätslücke* versteht man die Differenz

$$\begin{aligned}\varepsilon &= \min_x L_2(x) - \max_y L_1(y) \\ &= \min_x \max_y L(x, y) - \max_y \min_x L(x, y) \geq 0.\end{aligned}$$

Die Eigenschaft $\varepsilon \geq 0$ der Dualitätslücke heisst auch *schwache Dualitätseigenschaft*. Im Fall $\varepsilon = 0$ spricht man von der *starken Dualitätseigenschaft*.

BEOBACHTUNG: Existiert ein Sattelpunkt, so ist die Dualitätslücke null.

4.1. Dualität bei Optimierungsproblemen. Die Lagrangefunktion des allgemeinen Optimierungsproblems (9),

$$L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x}),$$

ist eine spezielle Funktion $L : X \times Y \rightarrow \mathbb{R}$ mit $X = \mathcal{F}$ und $Y = \mathbb{R}_+^m$. Offenbar haben wir

$$L_2(\mathbf{x}) = \max_{\mathbf{y} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}) < +\infty \iff \mathbf{g}(\mathbf{x}) \leq \mathbf{0}.$$

Beim Minimieren von $L_2(\mathbf{x})$ genügt es also, sich auf solche $\mathbf{x} \in \mathcal{F}$ zu beschränken, die der Nebenbedingung $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$ genügen. Andererseits sieht man sofort:

$$\mathbf{g}(\mathbf{x}) \leq \mathbf{0} \implies L_2(\mathbf{x}) = L(\mathbf{x}, \mathbf{y}).$$

Folglich ergibt sich:

LEMMA 1.8. *Das Optimierungsproblem (9) hat entweder keine zulässige Lösung oder ist äquivalent zum primalen Problem:*

$$\min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{F}, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\} = \min_{\mathbf{x} \in \mathcal{F}} \max_{\mathbf{y} \geq \mathbf{0}} [f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x})].$$

◇

Das duale Problem lässt sich nicht so einfach auf das ursprüngliche Optimierungsproblem zurückführen. Es liefert aber (in der Praxis oft sehr nützliche!) Untergrenzen für den Minimalwert der Zielfunktion.

LEMMA 1.9 („Schwache Dualität“). *Sei $\mathbf{y} \geq \mathbf{0}$ beliebig. Dann gilt*

$$\min_{\mathbf{x} \in \mathcal{F}} [f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x})] = L_1(\mathbf{y}) \leq \min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{F}, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}.$$

◇

BEMERKUNG. Viele Optimierungsprobleme haben eine Dualitätslücke $\varepsilon > 0$ (d.h. keine starke Dualität und somit auch keine Sattelpunkte). *Lineare Programme bilden eine wichtige Ausnahme: Starke Dualität ist immer garantiert.*

4.2. Dualität bei linearen Programmen. Wir betrachten

$$L(\mathbf{x}, \mathbf{y}) = \mathbf{c}^T \mathbf{x} + \mathbf{y}^T (A\mathbf{x} - \mathbf{b}) = -\mathbf{y}^T \mathbf{b} + (\mathbf{y}^T A + \mathbf{c}^T) \mathbf{x}.$$

Das primale Problem ist hier äquivalent zu dem linearen Programm

$$(14) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}.$$

Bzgl. des dualen Problems überlegt man sich zunächst für jedes $\mathbf{y} \geq \mathbf{0}$:

$$L_1(\mathbf{y}) = \min_{\mathbf{x} \in \mathbb{R}^n} L(\mathbf{x}, \mathbf{y}) = -\infty \quad \text{wenn} \quad (\mathbf{y}^T A + \mathbf{c}^T) \neq \mathbf{0}^T.$$

Beim Maximieren von $L_1(\mathbf{y})$ darf man sich also auf solche \mathbf{y} beschränken, die $\mathbf{c}^T = -A^T \mathbf{y}$ ergeben. Somit erhält man

$$(15) \quad \max_{\mathbf{y} \geq \mathbf{0}} (-\mathbf{b})^T \mathbf{y} \quad \text{s.d.} \quad -A^T \mathbf{y} = \mathbf{c}.$$

Die *schwache Dualität* bedeutet hier für jedes $\mathbf{y} \geq \mathbf{0}$ und jedes $\mathbf{x} \in \mathbb{R}^n$ mit der Eigenschaft $A\mathbf{x} \leq \mathbf{b}$:

$$-A^T \mathbf{y} = \mathbf{c}^T \quad \implies \quad -\mathbf{b}^T \mathbf{y} \leq \mathbf{c}^T \mathbf{x}.$$

Wir wissen, dass Optimallösungen linearer Programme durch die KKT-Bedingungen charakterisiert sind. In der Sprache der Dualitätstheorie bedeutet dies:

SATZ 1.8 (Dualitätssatz der linearen Programmierung). *Genau dann ist eine zulässige Lösung \mathbf{x} des linearen Programms (14) optimal, wenn eine zulässige Lösung \mathbf{y} des dualen Problems (15) existiert mit der Eigenschaft*

$$-\mathbf{b}^T \mathbf{y} = \mathbf{c}^T \mathbf{x}.$$

◇

BEMERKUNG. Satz 1.8 ist auch unter der Bezeichnung „Hauptsatz der linearen Programmierung“ bekannt.

4.2.1. Das lineare Produktionsmodell. Wir gehen von einer Situation aus, wo Bedarfsgüter der Typen P_1, P_2, \dots, P_n produziert werden sollen. Dazu müssen Rohstoffe R_1, \dots, R_m verwendet werden. Die ökonomischen Produktionsparameter seien

$$\begin{aligned} a_{ij} &= \text{benötigte Menge von } R_i \text{ zur Produktion einer Einheit von } P_j \\ c_j &= \text{Gewinn pro Einheit bei Produktion von } P_j \\ b_i &= \text{Anzahl Einheiten von } R_i \text{ im Vorrat} \end{aligned}$$

Das Ziel der Gewinnmaximierung ergibt den optimalen Produktionsplan als Optimallösung des folgenden linearen Programms:

$$(16) \quad \begin{array}{rcll} \max & c_1 x_1 & + & \dots & + & c_n x_n \\ \text{s.d.} & a_{11} x_1 & + & \dots & + & a_{1n} x_n & \leq & b_1 \\ & \vdots & & & & & & \vdots \\ & a_{m1} x_1 & + & \dots & + & a_{mn} x_n & \leq & b_m \\ & & & x_1, \dots, x_n & & & \geq & 0 \end{array}$$

Welchen Marktwert haben die Rohstoffe R_i ? In einem stabilen Markt kann man unterstellen:

- Die Rohstoffpreise y_i haben die Eigenschaft, dass durch eine Umwandlung der Rohstoffe in Produkte P_j kein höherer Erlös als c_j erzielt werden kann.

(Sonst würde ja jedermann möglichst viele Rohstoffe kaufen und eine eigene Produktion errichten wollen. Dadurch würden wiederum die Rohstoffpreise explodieren.) Also erfüllen die y_i die linearen Restriktionen

$$\begin{array}{rcl} a_{11} y_1 & + & \dots & + & a_{m1} y_m & \geq & c_1 \\ \vdots & & & & \vdots & & \\ a_{1n} y_1 & + & \dots & + & a_{mn} y_m & \geq & c_n. \end{array}$$

Der Mindestmarktwert der Rohstoffe ergibt sich folglich als Optimallösung eines linearen Programms:

$$(17) \quad \begin{array}{rcll} \min & b_1 y_1 & + & \dots & + & b_m y_m \\ \text{s.d.} & a_{11} y_1 & + & \dots & + & a_{m1} y_m & \geq & c_1 \\ & \vdots & & & & & & \vdots \\ & a_{1n} y_1 & + & \dots & + & a_{mn} y_m & \geq & c_n \\ & & & y_1, \dots, y_m & & & \geq & 0 \end{array}$$

Problem (16) ist in Matrixschreibweise

$$\max \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad \begin{bmatrix} A \\ -I \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}$$

Das dazu duale Problem ist

$$\min \mathbf{y}^T \mathbf{b} \quad \text{s.d.} \quad \mathbf{y}^T A - \mathbf{z}^T I = \mathbf{c}^T, \quad \mathbf{y} \geq \mathbf{0}, \mathbf{z} \geq \mathbf{0}.$$

Wegen $\mathbf{z} \geq \mathbf{0}$ ist letzteres aber äquivalent zu (17):

$$\min \mathbf{b}^T \mathbf{y} \quad \text{s.d.} \quad A^T \mathbf{y} \geq \mathbf{c}, \mathbf{y} \geq \mathbf{0}.$$

BEMERKUNG. Die optimalen Parameter y_1^*, \dots, y_m^* , die sich aus (17) ergeben, sind die sog. *Schattenpreise* der Rohstoffe R_1, \dots, R_m .

KAPITEL 2

Fundamentale Algorithmen

Viele der über konvexe Mengen und Funktionen bewiesenen Aussagen sind reine Existenzaussagen und somit zunächst einmal nur theoretisch. Für den praktischen Umgang mit Optimierungsproblem braucht man Algorithmen, die einem erlauben, die theoretisch existierenden Grössen auch zu berechnen. Wir diskutieren hier einige für die mathematische Optimierung wichtige fundamentale Algorithmen.

Im Mittelpunkt stehen Algorithmen, die Lösungen linearer Ungleichungssysteme berechnen. Wie man von den KKT-Bedingung weiss, genügt dies, um Optimallösungen linearer Programme berechnen zu können.

MAN BEACHTE: *Ein lineares Ungleichungssystem $Ax \leq \mathbf{b}$ lässt sich typischerweise **nicht** mit dem Gauss'schen Algorithmus lösen!*

1. Zeilen- und Spaltenoperationen

Sei $A \in \mathbb{R}^{m \times n}$ eine Matrix. Wendet man die fundamentalen Operationen der linearen Algebra auf die Zeilenvektoren von A an, so spricht man von *elementare Zeilenoperation*. Sie sind:

- Multiplikation eines Zeilenvektors \mathbf{a}_i^T mit einem Skalar $y_i \neq 0$;
- Addition eines Zeilenvektors \mathbf{a}_i^T zu einem Zeilenvektor \mathbf{a}_j^T .

Bekanntlich lässt sich eine elementare Zeilenoperation algebraisch als Produkt PA mit einer (von links multiplizierten) invertierbaren Matrix P beschreiben. Das Produkt AP^T (mit der von rechts multiplizierten transponierten Matrix P^T) beschreibt die analoge *elementare Spaltenoperation*.

Unter einem (r, k) -Pivot verstehen wir die Folge von elementaren Zeilenoperationen:

- (1) Dividiere Zeile r durch a_{rk} ;
- (2) Subtrahiere jeweils das a_{ik} -fache der neuen Zeile r von den übrigen Zeilen $i \neq r$.

NOTA BENE: *Genau im Fall $a_{rk} \neq 0$ ist ein (r, k) -Pivot durchführbar.*

2. Elimination nach Fourier-Motzkin

Die Methode von Fourier-Motzkin zur Lösung linearer Ungleichungssysteme beruht auf folgender Beobachtung. Zwei Ungleichungen vom Typ

$$(18) \quad \begin{aligned} (+1)x_1 + a_{12}x_2 + \dots + a_{1n}x_n &\leq b_1 \\ (-1)x_1 + a_{22}x_2 + \dots + a_{2n}x_n &\leq b_2 \end{aligned}$$

sind äquivalent zu

$$(19) \quad -b_2 + \sum_{j=2}^n a_{2j}x_j \leq x_1 \leq b_1 - \sum_{j=2}^n a_{1j}x_j.$$

Ausserdem ist die Ungleichung

$$(20) \quad -b_2 + \sum_{j=2}^n a_{2j}x_j \leq b_1 - \sum_{j=2}^n a_{1j}x_j.$$

äquivalent zur Summe der Ungleichungen in (18):

$$(21) \quad \sum_{j=2}^n (a_{1j} + a_{2j})x_j \leq b_1 + b_2.$$

LEMMA 2.1. (18) Die Lösungen von (18) erhält man folgendermassen:

- Man bestimme eine Lösung (x_2, \dots, x_n) für (21) und ergänze diese mit einem x_1 , das (19) erfüllt.

Insbesondere ist (18) genau dann lösbar, wenn (21) lösbar ist.

Die Idee ist nun, nach der Variablen x_1 der Reihe nach die übrigen Variablen x_2, \dots, x_n zu eliminieren. Am Ende erweist sich dann das System entweder trivialerweise als unlösbar, weil man einen Widerspruch

$$0 \leq b' < 0$$

abgeleitet hat, oder man kann jede Lösung des Endsystems auf eine Lösung von (18) (wie in Lemma 2.1 beschrieben) der Reihe nach zurückrechnen.

2.0.2. *Das allgemeine Verfahren.* Wir betrachten das lineare Ungleichungssystem

$$(22) \quad \sum_{j=1}^n a_{ij}x_j \leq b_i \quad (i \in I)$$

mit endlicher Indexmenge I . Um z.B. x_1 zu eliminieren, teilen wir I in die Teilmengen I_+ , I_- und I_0 danach auf, ob der Koeffizient a_{i1} von x_1 positiv, negativ oder 0 ist.

Wir dividieren die Ungleichungen in $I_+ \cup I_-$ jeweils durch $|a_{i1}| > 0$. Damit erhalten wir das äquivalente System

$$(23) \quad \begin{aligned} (+1)x_1 + \sum_{j=2}^n a'_{sj}x_j &\leq b'_s \quad (s \in I_+) \\ (-1)x_1 + \sum_{j=2}^n a'_{tj}x_j &\leq b'_t \quad (t \in I_-) \\ \sum_{j=2}^n a_{ij}x_j &\leq b_i \quad (i \in I_0) \end{aligned}$$

und bemerken

$$(24) \quad \max_{t \in I_-} \left(-b'_t + \sum_{j=2}^n a'_{tj}x_j \right) \leq x_1 \leq \min_{s \in I_+} \left(b'_s - \sum_{j=2}^n a'_{sj}x_j \right)$$

Nun ersetzen wir die Ungleichungen in $I_+ \cup I_-$ durch alle Summen von Paaren und erhalten das System

$$(25) \quad \begin{aligned} \sum_{j=2}^n (a'_{sj} + a'_{tj})x_j &\leq b'_s + b'_t \quad (s \in I_+, t \in I_-) \\ \sum_{j=2}^n a_{ij}x_j &\leq b_i \quad (i \in I_0) \end{aligned}$$

SATZ 2.1. (x_1, \dots, x_n) ist genau dann eine Lösung von (22), wenn gilt

- (i) (x_2, \dots, x_n) löst das lineare System (25);
- (ii) x_1 genügt der Bedingung (24).

◇

BEMERKUNG. Die Bestimmung von x_1 aus einer Lösung (x_2, \dots, x_n) von (25) gemäss (24) heisst *Rücksstitution*.

Zur Lösung des Ungleichungssystems (22) kann man nun so vorgehen:

- (1) Man eliminiert der Reihe nach die Variablen x_1, \dots, x_n ;
- (2) Das Endsystem erkennt man entweder trivialerweise als unzulässig oder zulässig. Im zulässigen Fall gelangt man vom Endsystem der Reihe nach durch Rücksstitutionen zu einer Lösung von (22).

Mit der Methode von Fourier-Motzkin kann man im Prinzip jedes endliche lineare Ungleichungssystem in endlich vielen Schritten lösen. Allerdings ist das Verfahren in der Praxis oft sehr ineffizient. Denn:

- In einem Eliminationsschritt kann (beim Übergang von (23) zu (25)) die Anzahl der Ungleichungen sehr stark wachsen!

BEMERKUNG. Wie der Gauss'sche Algorithmus beruht auch das FM-Verfahren auf elementaren Zeilenoperationen: Addition von 2 Ungleichungen und Multiplikation einer Ungleichung mit einem Skalar. Allerdings werden bei der skalaren Multiplikation (im Gegensatz zum Gauss-Verfahren) nur *positive* Skalare zugelassen.

2.1. Das Lemma von Farkas (in alternativer Form). Nehmen wir an, wir hätten das FM-Verfahren auf das Ungleichungssystem

$$Ax \leq \mathbf{b}$$

angewandt und alle Variablen eliminiert. Dann haben wir insgesamt auf der linken Seite den Nullvektor als nichtnegative Linearkombination der Zeilen von A erzeugt. Ist $\mathbf{y} \geq 0$ der zugehörige Koeffizientenvektor, haben wir die Situation

$$\mathbf{y}^T Ax = \mathbf{0}^T x \leq \mathbf{y}^T \mathbf{b}.$$

Genau im Fall $\mathbf{y}^T \mathbf{b} < 0$ erweist sich $Ax \leq \mathbf{b}$ als unlösbar.

LEMMA 2.2 („Lemma von Farkas“). *Auf das lineare Ungleichungssystem $Ax \leq \mathbf{b}$ trifft genau eine der Aussagen zu:*

- (I) $Ax \leq \mathbf{b}$ besitzt eine zulässige Lösung $\bar{\mathbf{x}}$;
- (II) Es gibt einen Koeffizientenvektor \mathbf{y} mit den Eigenschaften

$$\mathbf{y} \geq \mathbf{0}, \mathbf{y}^T A = \mathbf{0}^T \text{ und } \mathbf{y}^T \mathbf{b} < 0.$$

(I) und (II) können nicht gleichzeitig zutreffen. Denn $\mathbf{y} \geq \mathbf{0}$ und $\mathbf{b} - A\bar{\mathbf{x}} \geq \mathbf{0}$ implizieren

$$0 \leq \mathbf{y}^T (\mathbf{b} - A\bar{\mathbf{x}}) = \mathbf{y}^T \mathbf{b} - (\mathbf{y}^T A)\bar{\mathbf{x}} = \mathbf{y}^T \mathbf{b} - 0 = \mathbf{y}^T \mathbf{b}.$$

Ist (I) falsch und somit $Ax \leq \mathbf{b}$ nicht lösbar, so folgt die Existenz eines $\mathbf{y} \geq \mathbf{0}$, das (II) erfüllt, aus dem FM-Verfahren.

◇

2.1.1. *Anwendung: Stochastische Matrizen.* In der Anwendungssimulation betrachtet man Systeme, die sich zu jedem (diskreten) Zeitpunkt in einem von n möglichen Zuständen $\{Z_1, \dots, Z_n\}$ befinden. Wir nehmen an, dass das System mit Wahrscheinlichkeit

$$m_{ij} = Pr(Z_j | Z_i)$$

in den Zustand Z_j übergeht, wenn es vorher im Zustand Z_i war. Die entsprechende *Übergangsmatrix*

$$M = [m_{ij}] \in \mathbb{R}^{n \times n}$$

hat nur nichtnegative Koeffizienten $m_{ij} \geq 0$ und Zeilensummen

$$\sum_{j=1}^n m_{ij} = 1.$$

D.h. M ist eine sog. *stochastische Matrix*: Alle Spalten sind Wahrscheinlichkeitsverteilungen.

Nehmen wir an, dass sich das System zum Zeitpunkt t mit der Wahrscheinlichkeit π_i im Zustand Z_i befindet ($i = 1, \dots, n$). Dann befindet es sich zum Zeitpunkt $t + 1$ mit der Wahrscheinlichkeit

$$\pi'_k = \sum_{i=1}^n m_{ik} \pi_i$$

im Zustand Z_k ($k = 1, \dots, n$). In Matrixnotation haben wir also:

$$\pi' = \pi M \quad (\pi = [\pi_1, \dots, \pi_n], \pi' = [\pi'_1, \dots, \pi'_n]).$$

π heisst *stationär*, wenn $\pi' = \pi$. Eine stationäre Wahrscheinlichkeitsverteilung ist als ein (linker) Eigenvektor von M zum Eigenwert $\lambda = 1$.

PROPOSITION 2.1. *Die stochastische Matrix $M = [m_{ij}]$ besitzt eine stationäre Wahrscheinlichkeitsverteilung.*

Beweis. Wir setzen $A = M^T - I$. Ist $\mathbf{1}$ der Vektor, der in jeder Komponente eine „1“ hat, dann genügt es zu zeigen, dass

$$A\mathbf{x} = \mathbf{0}, \mathbf{x} \geq \mathbf{0}, -\mathbf{1}^T \mathbf{x} < 0$$

eine Lösung hat. Diese können wir dann auf Koeffizientensumme 1 normieren und erhalten dann das gewünschte π .

Nehmen wir an, keine solche Lösung existierte. Dann hätte nach dem Lemma von Farkas das lineare System

$$A^T \mathbf{y} \leq -\mathbf{1} \quad \text{bzw.} \quad M\mathbf{y} \leq \mathbf{y} - \mathbf{1}$$

eine Lösung $\bar{\mathbf{y}}$. Das kann aber nicht sein. Denn für $\bar{y}_k = \min\{\bar{y}_1, \dots, \bar{y}_n\}$ käme man zu dem Widerspruch

$$\bar{y}_k - 1 \geq \sum_{j=1}^n m_{kj} \bar{y}_j \geq \sum_{i=1}^n m_{kj} \bar{y}_k = y_k.$$

◇

2.2. Das Projektionslemma. Sei $N = \{1, \dots, n\}$ die Menge der Indices des betrachteten Koordinatenraums und $S \subseteq N$ eine feste Teilmenge. Zu einem gegebenen $\mathbf{x} \in \mathbb{R}^N$ bezeichnen wir mit \mathbf{x}_S die Einschränkung von \mathbf{x} auf die Koordinaten in S .

Ist $X \subseteq \mathbb{R}^N$ eine beliebige Teilmenge, so nennen wir die Menge

$$X_S = \{\mathbf{x}_S \mid \mathbf{x} \in X\} \subseteq \mathbb{R}^S$$

die *Projektion* von X auf den Koordinatenraum \mathbb{R}^S .

LEMMA 2.3 („Projektionslemma“). *Die Projektion P_S eines beliebigen Polyeders $P \subseteq \mathbb{R}^N$ ist ein Polyeder.*

Beweis. Sei P die Lösungsmenge des Ungleichungssystems $Ax \leq \mathbf{b}$. Wir versuchen, dieses mit dem FM-Verfahren zu lösen und eliminieren zuerst die Variablen x_i mit Index $i \in N \setminus S$. Dann ist $P_S = \{\mathbf{x}_S \mid \mathbf{x} \in P(A, \mathbf{b})\}$ genau die Lösungsmenge des vom FM-Verfahren bis dahin berechneten Ungleichungssystems $\tilde{A}\tilde{\mathbf{x}} \leq \tilde{\mathbf{b}}$, d.h.

$$P_S = P(\tilde{A}, \tilde{\mathbf{b}}).$$

◇

2.3. Endlich erzeugte Kegel und Polytope. Wir zeigen, dass endlich erzeugte konvexe Kegel und konvexe Mengen immer Polyeder sind.

LEMMA 2.4. Sei $V = \{\mathbf{v}_1, \dots, \mathbf{v}_k\} \subseteq \mathbb{R}^n$ eine endliche Menge. Dann gilt

- (a) Die Menge $\text{cone}(V)$ aller konischen Linearkombinationen ist ein Polyeder.
- (b) Die Menge $\text{conv}(V)$ aller Konvexkombinationen ist ein Polyeder.

Beweis. Wir zeigen (a). (Die Behauptung (b) beweist man ganz analog.) Sei

$$P = \text{cone}(V) = \left\{ \sum_{i=1}^k \lambda_i \mathbf{v}_i \mid \lambda_1, \dots, \lambda_k \geq 0 \right\}.$$

Wir bezeichnen mit I die Einheitsmatrix und bilden die Matrix $V = [\mathbf{v}_1, \dots, \mathbf{v}_k]$ mit den Spaltenvektoren \mathbf{v}_i . Nun betrachten wir die Menge \bar{P} aller Vektoren $\begin{pmatrix} \mathbf{z} \\ \mathbf{x} \end{pmatrix} \in \mathbb{R}^{n+k}$ derart, dass

$$(26) \quad \begin{array}{rcl} I\mathbf{z} & - & V\mathbf{x} = \mathbf{0} \\ & & \mathbf{x} \geq \mathbf{0} \end{array}$$

\bar{P} ist Lösungsmenge eines linearen Systems und somit ein Polyeder. P ist die Projektion von \bar{P} auf die \mathbf{z} -Koordinaten und folglich auch ein Polyeder.

◇

NOTA BENE. Mit dem FM-Verfahren kann man eine Matrix B berechnen mit der Eigenschaft

$$\text{cone}(V) = P(B, \mathbf{0}),$$

indem man einfach die \mathbf{x} -Variablen aus dem System (26) eliminiert. Ganz analog ergibt sich aus dem FM-Verfahren eine Matrix C und ein Vektor \mathbf{b} mit der Eigenschaft

$$\text{conv}(V) = P(C, \mathbf{b}).$$

BEMERKUNG. Die Suche nach einer algorithmisch effizienteren Methode als dem FM-Verfahren zur Berechnung einer Darstellung $\text{cone}(V) = P(B, \mathbf{0})$ (bzw. $\text{conv}(V) = P(C, \mathbf{b})$) ist offen.

Mit Hilfe des Projektionslemmas lässt sich ebenso zeigen:

PROPOSITION 2.2. Die Minkowskissumme $S = P + Q$ zweier beliebiger Polyeder $P, Q \subseteq \mathbb{R}^n$ ist selber ein Polyeder in \mathbb{R}^n .

Beweis. Übung. ◇

2.4. Das Erfüllbarkeitsproblem. Wir rechnen über dem Zahlbereich $\{0, 1\}$ mit den Operationen

$$\begin{array}{c|c|c} \oplus & 0 & 1 \\ \hline 0 & 0 & 1 \\ \hline 1 & 1 & 1 \end{array} \quad \begin{array}{c|c|c} \odot & 0 & 1 \\ \hline 0 & 0 & 0 \\ \hline 1 & 0 & 1 \end{array} \quad \begin{array}{c|c|c} - & 0 & 1 \\ \hline 1 & 1 & 0 \end{array}$$

Eine *Boolesche Funktion* ist eine Funktion $\varphi : \{0, 1\}^n \rightarrow \{0, 1\}$. Es ist bekannt, dass eine Boolesche Funktion $\varphi(x_1, \dots, x_n)$ in einer sog. *konjunktiven Normalform* (KNF) dargestellt werden kann:

$$\varphi(x_1, \dots, x_n) = \prod C_i,$$

wobei die *Klauseln* C_i die Form haben

$$C_i = a_{i1}y_1 \oplus \dots \oplus a_{in}y_n \quad \text{mit } a_{ij} \in \{0, 1\} \text{ und } y_i \in \{x_i, \bar{x}_i\}.$$

BEISPIEL 2.1. $\varphi(x_1, x_2, x_3) = (x_1 \oplus x_2) \odot (\bar{x}_1 \oplus x_2 \oplus x_3) \odot \bar{x}_3$.

ERFÜLLBARKEITSPROBLEM: Man entscheide, ob die per KNF gegebene Boolesche Funktion φ den Wert 1 annehmen kann. Das heisst: Kann eine Belegung der Variablen gefunden werden derart, dass *jede Klausel* C_i den Wert 1 annimmt?

Das Problem kann man mit Ungleichungssystemen modellieren. In der K Klausel $C_i = a_{i1}y_1 + \dots + a_{in}y_n$ ersetzen wir \bar{x}_j durch $1 - x_j$ und haben dann das Problem: Gibt es eine Lösung mit ganzzahligen $x_j \in \{0, 1\}$ derart, dass

$$a_{i1}y_1 + \dots + a_{in}y_n \geq 1 ?$$

BEISPIEL 2.2. Sei $C = x_2 \oplus \bar{x}_5 \oplus x_7$. Dann ist C erfüllbar, wenn es eine ganzzahlige $(0, 1)$ -Lösung der Ungleichung

$$x_2 + (1 - x_5) + x_7 \geq 1 \quad \longleftrightarrow \quad -x_2 + x_5 - x_7 \leq 0$$

gibt.

Das Erfüllbarkeitsproblem fragt also nach einer ganzzahligen $(0, 1)$ -Lösung des aus allen Klauseln gebildeten Ungleichungssystems.

2-SAT: Das Erfüllbarkeitsproblem für Boolesche Funktionen in KNF, bei denen jede Klausel höchstens 2 Variablen enthält.

2-SAT kann mit dem FM-Verfahren effizient(!) gelöst werden. Um das einzusehen, betrachten wir das folgende typische Beispiel:

BEISPIEL 2.3 (Resolvente).

$$\begin{array}{rcl} C_1 & = & x_k \oplus x_s \\ C_2 & = & \overline{x_k} \oplus x_l \\ \hline C & = & x_s \oplus x_l \end{array} \longleftrightarrow \begin{array}{rcl} -x_k & - & x_s & \leq & -1 \\ x_k & & & - & x_l & \leq & 0 \\ & & & - & x_s & - & x_l & \leq & -1 \end{array}$$

C ist die sog. Resolvente der Klauseln C_1 und C_2 . Offensichtlich sind C_1 und C_2 genau dann gleichzeitig erfüllt, wenn ihre Resolvente C erfüllt ist. Im Ungleichungssystem entspricht C der Summe der aus C_1 und C_2 gewonnenen Ungleichungen.

MAN ERKENNT: Die Resolventenbildung resultiert in einer Klausel mit höchstens 2 Variablen. Insgesamt sind aber nur $2n^2$ solcher Klauseln überhaupt möglich.

PROPOSITION 2.3. Wendet man das FM-Verfahren auf ein 2-SAT-Problem mit n Variablen an, so werden insgesamt höchstens $2n^2$ verschiedene Ungleichungen erzeugt.

◇

BEMERKUNG. Für das allgemeine Erfüllbarkeitsproblem ist beim gegenwärtigen Stand der Wissenschaft kein effizienter Lösungsalgorithmus bekannt.

3. Die Ellipsoidmethode

Wir betrachten das Problem, einen Punkt \mathbf{x} einer abgeschlossenen konvexen Teilmenge $C \subseteq \mathbb{R}^n$ zu berechnen. Dazu nehmen wir an:

- Es steht eine Subroutine¹ *SEP* zur Verfügung, die bei Eingabe eines Parametervektors $\mathbf{x}_0 \in \mathbb{R}^n$ folgendes produziert:
 - (i) Im Fall $\mathbf{x}_0 \notin C$ eine Ungleichung $\mathbf{a}^T \mathbf{x} \leq b$, die für alle $\mathbf{x} \in C$ gilt aber von \mathbf{x}_0 verletzt wird (d.h. $\mathbf{a}^T \mathbf{x}_0 > b$);
 - (ii) Im Fall $\mathbf{x}_0 \in C$ die Bestätigung, dass \mathbf{x}_0 eine zulässige Lösung ist.

IDEE: Man beginnt mit einem Ellipsoid E , das so gross ist, dass es C enthält. Man testet den Mittelpunkt $\mathbf{t} \in E$ auf Zugehörigkeit zu C . Im Fall (i) (d.h. $\mathbf{a}^T \mathbf{t} > b$) ersetzt man E durch ein kleineres Ellipsoid, das die Menge

$$E(\mathbf{a}, b) = \{\mathbf{x} \in E \mid \mathbf{a}^T \mathbf{x} \leq b\} \quad (\supseteq C!)$$

enthält und wiederholt den Vorgang bis ein $\mathbf{x} \in C$ gefunden ist.

3.1. Ellipsoide. Ein Ellipsoid E ist definiert als das Bild der Einheitskugel $B_n \subseteq \mathbb{R}^n$ unter einer affinen Transformation $f(\mathbf{x}) = A\mathbf{x} + \mathbf{t}$, wobei $\mathbf{t} \in \mathbb{R}^n$ beliebig (aber fest) und $A \in \mathbb{R}^{n \times n}$ invertierbar ist:

$$E = \{\mathbf{t} + A\mathbf{x} \mid \mathbf{x}^T \mathbf{x} \leq 1\} = \{\mathbf{y} \in \mathbb{R}^n \mid (\mathbf{y} - \mathbf{t})^T Q^{-1}(\mathbf{y} - \mathbf{t}) \leq 1\},$$

wobei die Matrix $Q = AA^T$ positiv definit ist. $\mathbf{t} = f(\mathbf{0})$ ist der Mittelpunkt von E .

¹eine sog. Separationsroutine für C

LEMMA 2.5. Sei $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ eine beliebige affine Transformation und $S, T \subseteq \mathbb{R}^n$ beliebige Mengen. Dann gilt:

- (a) $f(S)$ ist genau dann ein Ellipsoid, wenn S ein Ellipsoid ist.
- (b) $S \subseteq T \iff f(S) \subseteq f(T)$.

Beweis. Übung.

3.1.1. *Das Löwner-Ellipsoid.* Das Löwner-Ellipsoid ist das Bild E_L der Einheitskugel unter der affinen Transformation $f_L(\mathbf{x}) = \mathbf{t}_L + A_L \mathbf{x}$ mit

$$\mathbf{t}_L^T = \left(\frac{1}{n+1}, 0, \dots, 0 \right), \quad A_L = \text{diag} \left(\frac{n}{n+1}, \frac{n}{\sqrt{n^2-1}}, \dots, \frac{n}{\sqrt{n^2-1}} \right).$$

LEMMA 2.6. Das Löwner-Ellipsoid E_L enthält alle Punkte $\mathbf{x} \in B_n$ mit nichtnegativer erster Koordinate $x_1 \geq 0$:

$$\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T \mathbf{x} \leq 1, \mathbf{e}_1^T \mathbf{x} = x_1 \geq 0\} \subseteq E_L.$$

Beweis. Übung.

BEMERKUNG. Man kann zeigen, dass E_L das kleinste Ellipsoid ist, das die Menge $\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T \mathbf{x} \leq 1, \mathbf{e}_1^T \mathbf{x} \geq 0\}$ enthält.

3.1.2. *Das allgemeine Löwner-Ellipsoid.* Wir betrachten ein allgemeines Ellipsoid

$$E = \{\mathbf{t} + A\mathbf{x} \mid \mathbf{x}^T \mathbf{x} \leq 1\} = \{\mathbf{y} \in \mathbb{R}^n \mid (\mathbf{y} - \mathbf{t})^T Q^{-1}(\mathbf{y} - \mathbf{t}) \leq 1\}$$

und einen Halbraum $H = P(\mathbf{a}^T, b_{\mathbf{t}}) \neq \mathbb{R}^n$ mit $b_{\mathbf{t}} = \mathbf{a}^T \mathbf{t}$ (d.h. der Mittelpunkt $\mathbf{t} \in E$ liegt auf der zugehörigen Hyperebene). Wir suchen ein kleines Ellipsoid E' , das die Menge

$$E(\mathbf{a}^T, b_{\mathbf{t}}) = \{\mathbf{y} \in E \mid \mathbf{a}^T \mathbf{y} \leq b_{\mathbf{t}}\}$$

enthält. Wir nehmen zuerst $\mathbf{t} = \mathbf{0}$ (und somit $b_{\mathbf{t}} = 0$) an und setzen $\bar{\mathbf{a}}^T = -\mathbf{a}^T A$. Für $\mathbf{y} = A\mathbf{x}$ hat man dann

$$\mathbf{a}^T \mathbf{y} \leq 0 \iff \bar{\mathbf{a}}^T \mathbf{x} \geq 0$$

Wir setzen ferner $\mathbf{r}_1 = \bar{\mathbf{a}}/\|\bar{\mathbf{a}}\|$ und ergänzen \mathbf{r}_1 zu einer Orthonormalbasis, die wir als Matrix $R = [\mathbf{r}_1, \dots, \mathbf{r}_n]$ auffassen. Für die Variablensubstitution $\mathbf{z} = R^T \mathbf{x}$ finden wir somit

$$\mathbf{a}^T \mathbf{y} \leq 0 \iff \bar{\mathbf{a}}^T \mathbf{x} \geq 0 \iff \mathbf{e}_1^T \mathbf{z} \geq 0.$$

Für das aus dem Löwner-Ellipsoid E_L abgeleitete Ellipsoid

$$E' = A(R(E_L))$$

haben wir (wegen $R^T = R^{-1}$ und $\mathbf{x} = R\mathbf{z}$) deshalb

$$(27) \quad \begin{aligned} E(\mathbf{a}^T, 0) &= \{A\mathbf{x} \mid \mathbf{x} \in B_n, \bar{\mathbf{a}}^T \mathbf{x} \geq 0\} \\ &= \{AR\mathbf{z} \mid \mathbf{z} \in B_n, \mathbf{e}_1^T \mathbf{z} \geq 0\} \subseteq E'. \end{aligned}$$

Im Fall $\mathbf{t} \neq \mathbf{0}$ leistet natürlich das Ellipsoid

$$E' = \mathbf{t} + A(R(E_L))$$

das Gewünschte.

3.1.3. *Aufdatierungsformeln.* Es stellt sich heraus, dass man zur Bestimmung von E' die Matrix R im vorigen Abschnitt gar nicht explizit berechnen muss: Die positiv definite Strukturmatrix Q' von E' kann direkt aus der Strukturmatrix Q von E und dem Vektor \mathbf{a} gewonnen werden. Um dies zu sehen, schreiben wir die Strukturmatrix des Löwner-Ellipsoids E_L als

$$\begin{aligned} Q_L &= \text{diag} \left(\frac{n^2}{(n+1)^2}, \frac{n^2}{n^2-1}, \dots, \frac{n^2}{n^2-1} \right) \\ &= \frac{n^2}{n^2-1} \left[I - \frac{2}{n+1} \mathbf{e}_1 \mathbf{e}_1^T \right]. \end{aligned}$$

Nun rechnet man

$$\begin{aligned} Q' &= (ARA_L)(ARA_L)^T = ARQ_LR^T A \\ &= \frac{n^2}{n^2-1} AR \left[I - \frac{2}{n+1} \mathbf{e}_1 \mathbf{e}_1^T \right] R^T AT \\ &= \frac{n^2}{n^2-1} \left[AA^T - \frac{2}{n+1} ARe_1 \mathbf{e}_1^T R^T A^T \right] \\ &= \frac{n^2}{n^2-1} \left[Q - \frac{2}{n+1} \mathbf{b} \mathbf{b}^T \right] \end{aligned}$$

mit

$$\mathbf{b} = ARe_1 = \frac{-1}{\|A^T \mathbf{a}\|} AA^T \mathbf{a} = \frac{-Q\mathbf{a}}{\sqrt{\mathbf{a}^T Q \mathbf{a}}}.$$

Analog berechnet man den Mittelpunkt von E' als

$$\mathbf{t}' = \mathbf{t} + \frac{1}{n+1} \mathbf{b}.$$

3.2. Volumen. Wir definieren das *Volumen* einer kompakten konvexen Menge $S \subseteq \mathbb{R}^n$ als

$$\text{vol}(S) = \max_{\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_n \in S} |\det(\mathbf{v}_1 - \mathbf{v}_0, \dots, \mathbf{v}_n - \mathbf{v}_0)|.$$

Da $\det(\mathbf{x}_1, \dots, \mathbf{x}_n)$ stetig ist, ist dieses Volumen für kompakte Mengen wohldefiniert (und endlich). Ausserdem macht man sich (aufgrund des Determinantenmultiplikationssatzes $\det(A) = (\det A)(\det B)$) sofort klar:

- (0) $S \subseteq T \implies \text{vol}(S) \leq \text{vol}(T)$;
- (i) Für jede affine Abbildung $f(\mathbf{x}) = \mathbf{t} + A\mathbf{x}$ gilt

$$\text{vol}(f(S)) = |\det A| \text{vol}(S).$$

BEMERKUNG. Dieser Volumenbegriff ist etwas schwächer als der in der Geometrie übliche. Er reicht aber für unsere Zwecke völlig aus.

BEISPIEL 2.4. Sei B_n die Einheitskugel. Dann ergibt sich aus der Hadamardschen Ungleichung

$$|\det(\mathbf{v}_1 - \mathbf{v}_0, \dots, \mathbf{v}_n - \mathbf{v}_0)| \leq \prod_{i=1}^n \|\mathbf{v}_i - \mathbf{v}_0\|$$

die Abschätzung

$$\text{vol}(B_n) \leq \max_{\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_n \in B_n} \prod_{i=1}^n \|\mathbf{v}_i - \mathbf{v}_0\| \leq 2^n .$$

Mit dem Volumenbegriff können wir abschätzen, wie klein das Löwner-Ellipsoid E' im Vergleich zu dem Ausgangsellipsoid E ist. Wegen $\det R = \pm 1$ erhalten wir

$$\frac{\text{vol}(E')}{\text{vol}(E)} = \frac{|(\det A)(\det R)(\det A_L)| \text{vol}(B_n)}{|\det A| \text{vol}(B_n)} = \det A_L,$$

wobei

$$\det A_L = \frac{n}{n+1} \left(\frac{n^2}{n^2-1} \right)^{(n-1)/2} = \left(1 - \frac{1}{n+1} \right) \left(1 + \frac{1}{n^2-1} \right)^{(n-1)/2}$$

Mit Hilfe der Ungleichung $1 - x \leq e^x$ finden wir nun

$$\det A_L \leq e^{-1/(n+1)} \cdot e^{(n-1)/2(n^2-1)} = e^{-1/2(n+1)} < 2^{-1/2(n+1)}$$

und fassen zusammen:

$$\boxed{\text{vol} E' \leq 2^{-1/2(n+1)} \text{vol} E}$$

3.3. Der Ellipsoid-Algorithmus. Wir nehmen an, dass wir eine Zahl $r > 0$ kennen mit der Eigenschaft

$$C \subseteq E_0 = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T \mathbf{x} \leq r^2\}.$$

Wir starten die Ellipsoidmethode mit dem Ellipsoid E_0 und iterieren folgendermassen:

- Verletzt der Mittelpunkt \mathbf{t}_k des aktuellen Ellipsoids E_k die für abgeschlossene konvexe Menge C gültige Ungleichung $\mathbf{a}^T \mathbf{x} \leq b$, dann setzen wir $b_{\mathbf{t}} = \mathbf{a}^T \mathbf{t}_k$ und wählen als E_{k+1} das entsprechende Löwnerellipsoid

$$E_{k+1} \supseteq E_k(\mathbf{a}, b_{\mathbf{t}}) \supseteq E(\mathbf{a}, b) \supseteq C.$$

PROPOSITION 2.4. Sei $\text{vol}(C) > 0$. Dann ist nach

$$K \leq 2(n+1)[n + n \log_2 r - \log_2 \text{vol}(C)]$$

Iterationen ein Ellipsoid E_K gefunden mit Mittelpunkt $\mathbf{t}_K \in C$.

Beweis. Ist nach K Iterationen noch kein zulässiges $\mathbf{t}_K \in C$ gefunden, so gilt wegen $C \subseteq E_K$:

$$0 < \text{vol}(C) \leq \text{vol}(E_K) \leq 2^{-K/2(n+1)} \text{vol}(E_0) \leq 2^{n \log_2 2r - K/(2(n+1))}$$

und folglich $\log_2 \text{vol}(C) \leq n \log_2(2r) - K/2(n+1)$, d.h.

$$K \leq 2(n+1)[n + n \log_2 r - \log_2 \text{vol}(C)].$$

◇

Die Endlichkeit des Ellipsoidverfahrens ist nur im Fall $\text{vol}(C) > 0$ von vornherein garantiert. Oft kann man im Fall $\text{vol}(C) = 0$ dieses Problem mit einem Kniff umgehen, der am Beispiel endlicher linearer Systeme mit rationalen Koeffizienten später demonstriert wird².

MAN BEACHTE: *Wir haben noch nicht diskutiert, wie man die Ellipsoidmethode überhaupt startet (d.h. wie man das erste Ellipsoid der Iterationsfolge wählen sollte!* (Das kommt später.)

4. Die Methode innerer Punkte

4.1. Vorbemerkung: Newtons Methode. Sei $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ eine beliebige Funktion. Newtons Methode versucht das Nullstellenproblem

$$F(\mathbf{x}) = \mathbf{0}$$

iterativ zu lösen. Dabei beginnt man mit einem $\mathbf{x}_0 \in \mathbb{R}^n$ und stoppt im Fall $F(\mathbf{x}_0) = \mathbf{0}$. Andernfalls sucht man nach einem Lösungskandidaten $\Delta \mathbf{x}$ für die Gleichung

$$F(\mathbf{x}_0 + \Delta \mathbf{x}) = \mathbf{0}.$$

Den bestimmt man dadurch, dass man das Gleichungssystem linear relaxiert. D.h. man wählt eine Matrix A_0 in der Hoffnung

$$F(\mathbf{x}_0 + \mathbf{h}) \approx F(\mathbf{x}_0) + A_0 \mathbf{h}$$

und löst das linearisierte System

$$F(\mathbf{x}_0) + A_0 \mathbf{h} = \mathbf{0} \quad \text{bzw.} \quad A_0 \mathbf{h} = -F(\mathbf{x}_0).$$

Ist \mathbf{h}_0 eine solche Lösung, so setzt man $\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{h}_0$ und verfährt nun mit \mathbf{x}_1 genauso wie eben mit \mathbf{x}_0 usw.

Auf diese Weise erzeugt man eine Folge $\mathbf{x}_0, \mathbf{x}_1, \dots$ von Vektoren. Man stoppt in Iteration K , wenn

$$F(\mathbf{x}_K) \approx \mathbf{0}.$$

BEMERKUNG. Obwohl man im allgemeinen (ohne starke Zusatzannahmen) keine Konvergenzgarantie geben kann, funktioniert die Methode in der Praxis überraschend gut.

²eine ausführliche Diskussion der Ellipsoidmethode im Fall $\text{vol}(C) = 0$ findet sich in dem Buch von Grötschel, Lovász und Schrijver: GEOMETRIC ALGORITHMS AND COMBINATORIAL OPTIMIZATION (Springer 1993)

BEISPIEL 2.5. Sei $f(x) = x^2 - 2 = 0$ in der Variablen $x \in \mathbb{R}$ zu lösen. Die Wahl $A_k = f'(x_k)$ ergibt

$$h_k = \frac{-x_k^2 + 2}{2x_k} \quad \text{und} \quad x_{k+1} = x_k + h_k = \frac{x_k}{2} + \frac{1}{x_k}.$$

4.2. Die eigentliche Methode innerer Punkte (IPM). Wir betrachten das lineare Programm

$$\min \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b} \quad \text{und} \quad \mathbf{x} \geq \mathbf{0}.$$

Mit $\mathbf{s} = \mathbf{c} - A^T \mathbf{y}$ erhalten wir bei $\mu = 0$ die KKT-Bedingung als

$$(28) \quad \begin{aligned} x_j s_j &= \mu \quad (j = 1, \dots, n) \\ A\mathbf{x} &= \mathbf{b} \\ A^T \mathbf{y} + \mathbf{s} &= \mathbf{c} \\ \mathbf{x}, \mathbf{s} &\geq \mathbf{0} \end{aligned}$$

Als ersten Ansatz relaxieren wir das System zu einem Gleichungssystem, das im Geiste Newtons (approximativ) gelöst werden soll. Dazu lassen wir einfach die Ungleichungen $\mathbf{x} \geq \mathbf{0}$ und $\mathbf{s} \geq \mathbf{0}$ weg und erhalten das System

$$(29) \quad \begin{aligned} x_j s_j &= \mu \quad (j = 1, \dots, n) \\ A\mathbf{x} &= \mathbf{b} \\ A^T \mathbf{y} + \mathbf{s} &= \mathbf{c}, \end{aligned}$$

das wir nun für ein gegebenes $\mu > 0$ lösen wollen. Seien Parametervektoren $\mathbf{x}, \mathbf{y}, \mathbf{s}$ gegeben, welche das folgende Gleichungssystem erfüllen:

$$(30) \quad \begin{aligned} A\mathbf{x} &= \mathbf{b} \\ A^T \mathbf{y} + \mathbf{s} &= \mathbf{c} \end{aligned}$$

Dann erfüllen $\mathbf{x}^+ = \mathbf{x} + \Delta \mathbf{x}$, $\mathbf{y}^+ = \mathbf{y} + \Delta \mathbf{y}$ und $\mathbf{s}^+ = \mathbf{s} + \Delta \mathbf{s}$ das System (29) genau dann, wenn das quadratische System

$$(31) \quad \begin{aligned} \Delta x_j \Delta s_j + s_j \Delta x_j + x_j \Delta s_j &= \mu - s_j x_j \quad (j = 1, \dots, n) \\ A \Delta \mathbf{x} &= \mathbf{0} \\ A^T \Delta \mathbf{y} + \Delta \mathbf{s} &= \mathbf{0} \end{aligned}$$

erfüllt ist. Als Linearisierung des Systems lassen wir einfach die quadratischen Terme weg und berechnen Vektoren $\Delta \mathbf{x}$, $\Delta \mathbf{y}$ und $\Delta \mathbf{s}$ als Lösung von

$$(32) \quad \begin{aligned} s_j \Delta x_j + x_j \Delta s_j &= \mu - s_j x_j \quad (j = 1, \dots, n) \\ A \Delta \mathbf{x} &= \mathbf{0} \\ A^T \Delta \mathbf{y} + \Delta \mathbf{s} &= \mathbf{0} \end{aligned}$$

Wir wollen nun Bedingungen ableiten, unter denen die Konvergenz zu einer Lösung der tatsächlichen KKT-Bedingungen garantiert werden kann.

BEMERKUNG. Der Name „innere Punkte“ kommt daher, dass wir nur mit Lösungen $\mathbf{x} > \mathbf{0}$ und $\mathbf{s} > \mathbf{0}$ arbeiten werden, bei denen alle(!) Komponenten s_j bzw. x_j echt positiv sind. (Die Punkte liegen somit immer im Inneren des Nichtnegativitätsbereichs \mathbb{R}_+^n .)

4.3. Konvergenzanalyse. Wir schreiben $\mathbf{u} > \mathbf{0}$, wenn alle Komponenten u_i echt grösser als 0 sind. $\sqrt{\mathbf{u}}$ ist der Vektor mit den Komponenten $\sqrt{u_i}$. Das Produkt $\mathbf{u}\mathbf{v}$ der Vektoren \mathbf{u} , und \mathbf{v} ist der Vektor mit den Komponenten $u_i v_i$. Wir nehmen im folgenden $\mathbf{x} > \mathbf{0}$ und \mathbf{y} mit $\mathbf{s} = \mathbf{b} - A\mathbf{x} > \mathbf{0}$ als gegeben an.

LEMMA 2.7. Für jedes $\mu > 0$ ist (32) lösbar.

Beweis. Wir setzen

$$\begin{aligned} U &= \{ \sqrt{\mathbf{s}\mathbf{x}^{-1}}\Delta\mathbf{x} \mid \Delta\mathbf{x} \in \ker A \} \\ V &= \{ \sqrt{\mathbf{x}\mathbf{s}^{-1}}\Delta\mathbf{s} \mid \Delta\mathbf{s} \in \text{lin}A \} \end{aligned}$$

und beobachten $\dim U = \dim(\ker A) = n - \text{rg}A$ und $\dim V = \dim(\text{lin}A) = \text{rg}A$. Da der Zeilenraum $\text{lin}A$ der Matrix A das orthogonale Komplement des Kerns $\ker A$ ist, sind auch die linearen Teilräume U und V paarweise orthogonale Komplemente. Denn

$$[\sqrt{\mathbf{s}\mathbf{x}^{-1}}\Delta\mathbf{x}]^T \sqrt{\mathbf{x}\mathbf{s}^{-1}}\Delta\mathbf{s} = (\Delta\mathbf{x})^T \Delta\mathbf{s} = 0.$$

Mit $\mathbf{1} = (1, 1, \dots, 1)^T$ gibt es wegen $U \oplus V = \mathbb{R}^n$ Vektoren $\mathbf{u} \in U$ und $\mathbf{v} \in V$ so, dass

$$\frac{\mu\mathbf{1} - \mathbf{x}\mathbf{s}}{\sqrt{\mathbf{x}\mathbf{s}}} = \mathbf{u} + \mathbf{v} = \sqrt{\mathbf{s}\mathbf{x}^{-1}}\Delta\mathbf{x} + \sqrt{\mathbf{x}\mathbf{s}^{-1}}\Delta\mathbf{s}$$

für geeignete $\Delta\mathbf{x} \in \ker A$ und $\Delta\mathbf{s} \in \text{lin}A$, d.h. $\Delta\mathbf{s} = A^T \Delta\mathbf{y}$ für einen geeigneten Vektor $\Delta\mathbf{y} \in \mathbb{R}^m$.

◇

Wir messen die Qualität einer Lösung mit dem Parameter

$$\delta = \delta(\mathbf{x}, \mathbf{s}, \mu) = \frac{1}{\mu} \left\| \frac{\mathbf{x}\mathbf{s} - \mu\mathbf{1}}{\sqrt{\mathbf{x}\mathbf{s}}} \right\|^2 = \frac{1}{\mu} \|\mathbf{u} + \mathbf{v}\|^2.$$

LEMMA 2.8. Für beliebige orthogonale Vektoren $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ gilt:

$$\|\mathbf{u}\mathbf{v}\| \leq \frac{1}{2} \|\mathbf{u} + \mathbf{v}\|^2.$$

Beweis. Wegen $\mathbf{u}^T \mathbf{v} = 0$ gilt $\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u} + \mathbf{v}\|^2$. Aus der allgemeinen Identität

$$4\alpha\beta = (\alpha + \beta)^2 - (\alpha - \beta)^2$$

folgt $4|u_j v_j| = |(u_j + v_j)^2 - (u_j - v_j)^2| \leq (u_j + v_j)^2 + (u_j - v_j)^2$ und somit

$$4\|\mathbf{u}\mathbf{v}\| \leq 4 \sum_j |u_j v_j| \leq \|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u} + \mathbf{v}\|^2.$$

◇

SATZ 2.2. Sei $\mu > 0$ und $\delta = \delta(\mathbf{x}, \mathbf{s}, \mu) \leq 1$. Dann gilt

- (a) $\mathbf{x}^+ = \mathbf{x} + \Delta\mathbf{x} > \mathbf{0}$ und $\mathbf{s}^+ = \mathbf{s} + \Delta\mathbf{s} > \mathbf{0}$.
- (b) $\delta^+ = \delta(\mathbf{x}^+, \mathbf{s}^+, \mu) \leq \frac{1}{2}\delta^2$.

Beweis. Nach Lemma 2.8 haben wir für jede Komponente j :

$$|\Delta x_j \Delta s_j| = |u_j v_j| \leq \|\mathbf{u}\mathbf{v}\| \leq \frac{1}{2} \|\mathbf{u} + \mathbf{v}\|^2 = \frac{1}{2} \mu \delta \leq \mu/2.$$

und folglich $x_j^+ s_j^+ = \mu + \Delta x_j \Delta s_j \geq \mu/2 > 0$. Daraus ergibt sich (b):

$$\delta^+ = \frac{1}{\mu} \left\| \frac{\mathbf{x}^+ \mathbf{s}^+ - \mu \mathbf{1}}{\sqrt{\mathbf{x}^+ \mathbf{s}^+}} \right\|^2 \leq \frac{1}{\mu} \left\| \frac{\mathbf{u}\mathbf{v}}{\sqrt{\mu/2}} \right\|^2 \leq \frac{2}{\mu^2} \cdot \frac{1}{4} \|\mathbf{u} + \mathbf{v}\|^4 = \frac{1}{2} \delta^2.$$

Aussage (a) sieht man so ein: Wäre $x_j^+ < 0$, dann auch $s_j^+ < 0$ (denn $x_j^+ s_j^+ > 0$). Das ergäbe aber wegen $x_j > 0$ und $s_j > 0$ den Widerspruch

$$\mu = x_j s_j + x_j \Delta s_j + s_j \Delta x_j = x_j s_j^+ - s_j x_j^+ < 0.$$

◇

Die Analyse zeigt bisher: Wenn man mit den Vektoren $\mathbf{x}_0 > \mathbf{0}$ und $\mathbf{s}_0 > \mathbf{0}$ beginnt und dann nach Newton iteriert, so gilt

$$\delta_0 = \delta(\mathbf{x}_0, \mathbf{s}_0, \mu) \leq 1 \quad \implies \quad \delta_k = \delta(\mathbf{x}_k, \mathbf{s}_k, \mu) \leq \frac{1}{2^k} \rightarrow 0$$

d.h. man hat Konvergenz $\mathbf{x}_k \mathbf{s}_k \rightarrow \mu \mathbf{1}$ und somit Konvergenz zu einer Lösung von (28).

4.4. Die erweiterte Methode IPM. Wir betrachten nun die Erweiterung der Methode, wo in jedem Iterationsschritt zusätzlich μ zu $\theta\mu$ mit einem geeigneten Faktor $0 < \theta < 1$ reduziert wird.

Wir nehmen $\delta(\mathbf{x}, \mathbf{s}, \mu) \leq 1$ an und beobachten zunächst

$$n\mu = (\mathbf{x}^+)^T \mathbf{s}^+ = \|\sqrt{\mathbf{x}^+ \mathbf{s}^+}\|^2.$$

Nun rechnet man

$$\begin{aligned} \delta(\mathbf{x}^+, \mathbf{s}^+, \theta) &= \frac{1}{\theta\mu} \left\| \frac{\mathbf{x}^+ \mathbf{s}^+ - \theta\mu \mathbf{1}}{\sqrt{\mathbf{x}^+ \mathbf{s}^+}} \right\|^2 \\ &= \frac{1}{\theta\mu} \left\| (1-\theta)\sqrt{\mathbf{x}^+ \mathbf{s}^+} + \theta \frac{\mathbf{x}^+ \mathbf{s}^+ - \theta\mu \mathbf{1}}{\sqrt{\mathbf{x}^+ \mathbf{s}^+}} \right\|^2 \\ &= \frac{1}{\theta\mu} \left\| (1-\theta)\sqrt{\mathbf{x}^+ \mathbf{s}^+} \right\|^2 + \frac{1}{\theta\mu} \left\| (1-\theta)\sqrt{\mathbf{x}^+ \mathbf{s}^+} + \theta \frac{\mathbf{x}^+ \mathbf{s}^+ - \theta\mu \mathbf{1}}{\sqrt{\mathbf{x}^+ \mathbf{s}^+}} \right\|^2 \\ &= \frac{(1-\theta)^2}{\theta} n + \theta \delta(\mathbf{x}^+, \mathbf{s}^+, \mu) \\ &\leq \frac{(1-\theta)^2}{\theta} n + \frac{\theta}{2} \quad (\text{Satz 2.2(b)}). \end{aligned}$$

BEMERKUNG. Die dritte Gleichheit folgt aus der Orthogonalität der Vektoren („Satz des Pythagoras“).

KOROLLAR 2.1. Sei $n \geq 2$. Dann gilt für $\theta^* = 1 - \frac{1}{n}$.

$$\delta(\mathbf{x}^+, \mathbf{s}^+, \theta^* \mu) \leq \frac{1}{n-1} + \frac{\theta^*}{2} \leq 1 :$$

◇

4.5. Konvergenz. Seien $\mathbf{x}_0, \mathbf{s}_0, \mu_0$ so, dass $\delta(\mathbf{x}_0, \mathbf{s}_0, \mu_0) \leq 1$ erfüllt ist, und $\varepsilon > 0$ fest gewählt. Dann ist nach K Iterationen des IPM-Algorithmus der momentane μ -Wert $(\theta^*)^K \mu_0$ und es gilt

$$\lim_{K \rightarrow \infty} (\theta^*)^K \mu_0 = 0.$$

Der momentane μ -Wert ist $\leq \varepsilon/n$ nach

$$K = \left\lceil n \cdot \ln \frac{\mu_0 n}{\varepsilon} \right\rceil \leq n \cdot \ln \frac{\mu_0 n}{\varepsilon}$$

Iterationen³. Für die entsprechende Lösung $(\mathbf{x}_K, \mathbf{y}_K)$ gilt dann

$$0 \leq \mathbf{c}^T \mathbf{x}_K - \mathbf{b}^T \mathbf{y}_K = \mathbf{x}_K^T \mathbf{s}_K \leq \varepsilon.$$

Mit anderen Worten:

- \mathbf{x}_K ist eine zulässige Lösung des betrachteten linearen Programms und erreicht dessen Optimalwert z^* bis auf ε genau.

MAN BEACHTE: Wir haben noch nicht diskutiert, wie man IPM überhaupt startet, d.h. wie geeignete Parametervektoren $\mathbf{x}_0 > \mathbf{0}$ und $\mathbf{s}_0 > \mathbf{0}$ und ein entsprechendes μ_0 gefunden werden können! (Das kommt später.)

³bei dieser Abschätzung benutzt man wieder die Ungleichung $(1 - 1/n)^n \leq 1/e$

KAPITEL 3

Struktur von Polyedern

1. Der Darstellungssatz von Weyl-Minkowski

Wir betrachten ein beliebiges Polyeder P , das sich als Lösungsmenge eines endlichen Systems von linearen Ungleichungen $\mathbf{a}_i^T \mathbf{x} \leq b_i$ (mit Indexmenge I) schreiben lässt:

$$P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}_i^T \mathbf{x} \leq b_i, i \in I\}.$$

Wir betrachten zuerst den Spezialfall

- $\mathbf{0} \in P$ und folglich $b_i \geq 0$ für alle $b_i \in I$.

Dividieren wir nun im Fall $b_i > 0$ die entsprechende Ungleichung durch b_i , so erhalten wir ein P definierendes System von Ungleichungen $\bar{\mathbf{a}}_i^T \mathbf{x} \leq \bar{b}_i$ mit $\bar{b}_i \in \{0, +1\}$ es gibt also Matrizen A, B derart, dass

$$P = \{\mathbf{x} \in \mathbb{R}^n \mid \begin{bmatrix} A \\ B \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}\},$$

wobei $\mathbf{1} = (1, 1, \dots, 1)^T$.

1.1. Die Polare. Unter der *Polaren* einer Menge $S \subseteq \mathbb{R}^n$ mit $\mathbf{0} \in S$ versteht man die Menge

$$S^{pol} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{s}^T \mathbf{x} \leq 1 \text{ für alle } \mathbf{s} \in S\}.$$

Ist S endlich und stellen wir uns S^T als die Matrix mit den Zeilenvektoren \mathbf{s}^T vor, dann erkennen wir die Polare als Polyeder

$$S^{pol} = P(S^T, \mathbf{1}).$$

Auf jeden Fall folgt direkt aus der Definition

$$S \subseteq (S^{pol})^{pol}.$$

LEMMA 3.1. Sei $S \subseteq \mathbb{R}^n$ konvex und abgeschlossen mit $\mathbf{0} \in S$. Dann gilt

$$S = (S^{pol})^{pol}.$$

Beweis. Sei $\bar{S} = (S^{pol})^{pol}$. Wir zeigen $\bar{S} \subseteq S$ per Widerspruch. Sei $\mathbf{x} \in \bar{S} \setminus S$. Dann existiert nach dem konvexen Trennungssatz ein Vektor \mathbf{a} und ein Skalar b mit der Eigenschaft

$$\mathbf{a}^T \mathbf{x} > b \geq \mathbf{a}^T \mathbf{s} \quad \text{für alle } \mathbf{s} \in S.$$

Wegen $\mathbf{0} \in S$ dürfen wir oBdA $b = 1$ annehmen. Also gilt $\mathbf{a} \in S^{pol}$. Aber dann folgt aus $\mathbf{x} \in \bar{S}$ auch $\mathbf{a}^T \mathbf{x} \leq 1$ und somit ein Widerspruch.

◇

LEMMA 3.2. Sei P ein Polyeder und A und B Matrizen mit der Eigenschaft $P = \{\mathbf{x} \mid A\mathbf{x} \leq \mathbf{1}, B\mathbf{x} \leq \mathbf{0}\}$. Dann ist die Polare von P die Minkowskisumme des von den Zeilenvektoren von A und dem Ursprung $\mathbf{0}$ bestimmten Polytops und des von den Zeilenvektoren von B erzeugten konvexen Kegels:

$$P^{pol} = \text{conv}(A^T, \mathbf{0}) + \text{cone}(B^T)$$

Insbesondere ist P^{pol} ein Polyeder (da die Minkowskisumme von Polyedern immer ein Polyeder ergibt).

Beweis. Ein Vektor \mathbf{c} liegt in P^{pol} genau dann, wenn die Ungleichung $\mathbf{c}^T \mathbf{x} \leq 1$ von dem linearen System

$$\begin{bmatrix} A \\ B \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}$$

impliziert wird. Das ist genau dann der Fall, wenn es Vektoren $\mathbf{y}, \mathbf{z} \geq \mathbf{0}$ gibt mit der Eigenschaft

$$\mathbf{c}^T = \mathbf{y}^T A + \mathbf{z}^T B \quad \text{und} \quad \mathbf{y}^T \mathbf{1} \leq 1.$$

Wegen $A^T \mathbf{y} \in \text{conv}(A^T, \mathbf{0})$ und $B^T \mathbf{z} \in \text{cone}(B^T)$ folgt dann

$$\mathbf{c} \in P^{pol} \iff \mathbf{c} \in \text{conv}(A^T, \mathbf{0}) + \text{cone}(B^T).$$

◇

1.2. Der Dekompositionssatz.

SATZ 3.1 (Weyl-Minkowski). Genau dann ist eine Teilmenge $P \subseteq \mathbb{R}^n$ ein Polyeder, wenn es endliche Mengen $V, W \subseteq \mathbb{R}^n$ gibt mit der Eigenschaft

$$(33) \quad P = \text{conv}(V) + \text{cone}(W).$$

Beweis. Da $\text{conv}(V)$ und $\text{cone}(W)$ Polyeder sind, ist deren Minkowskisumme ein Polyeder. Die Bedingung ist also hinreichend. Wir beweisen die Notwendigkeit und nehmen oBdA $P \neq \emptyset$ an.

Wir betrachten zuerst den Fall $\mathbf{0} \in P$. Dann kann P in der Form

$$P = \{\mathbf{x} \mid A\mathbf{x} \leq \mathbf{1}, B\mathbf{x} \leq \mathbf{0}\}$$

ausgedrückt werden. Nach Lemma 3.2 ist $Q = P^{pol}$ ein Polyeder und wir finden

$$P = (P^{pol})^{pol} = Q^{pol}.$$

Wiederum aus Lemma 3.2 schliessen wir nun, dass P als Minkowskisumme einer endlich erzeugten konvexen Menge und eines endlich erzeugten konvexen Kegels ausgedrückt werden kann.

Im Fall $\mathbf{0} \notin P$ wählen wir irgendein $\mathbf{t} \in P$ und betrachten die Translation (Minkowskisumme)

$$\bar{P} = P + \{-\mathbf{t}\}.$$

Wegen $\mathbf{0} \in \bar{P}$ gibt es endliche Mengen \bar{V} und \bar{W} derart, dass

$$\bar{P} = \text{conv}(\bar{V}) + \text{cone}(\bar{W}).$$

Nun verifiziert man leicht für $V = \bar{V} + \{\mathbf{t}\}$ und $W = \bar{W}$:

$$P = \text{conv}(V) + \text{cone}(W).$$

◇

Aus dem Dekompositionssatz folgt sofort eine wichtige Charakterisierung von Polytopen:

KOROLLAR 3.1. *Eine Teilmenge $P \subseteq \mathbb{R}^n$ ist genau dann ein Polytop, wenn es eine endliche Menge V gibt mit der Eigenschaft*

$$P = \text{conv}(V).$$

Beweis. Wir haben mit Hilfe des FM-Verfahrens schon die konvexen Hüllen von endlichen Mengen als Polytope erkannt. Sei umgekehrt P ein Polytop mit der Weyl-Minkowski-Dekomposition

$$P = \text{conv}(V) + \text{cone}(W).$$

Dann kann W keinen Vektor $\mathbf{w} \neq \mathbf{0}$ enthalten, da sonst P nicht beschränkt wäre. Also haben wir $P = \text{conv}(V)$.

◇

1.3. Dualität von Darstellungen. Der Satz von Weyl-Minkowski zeigt, dass ein Polyeder P zwei einander duale Sichtweisen erlaubt:

IMPLIZIT: P ist Lösungsmenge eines endlichen linearen Ungleichungssystems $A\mathbf{x} \leq \mathbf{b}$;

EXPLIZIT: P ist die Menge aller Vektoren (bzw. Punkte), die von den endlichen Mengen V und W gemäss (33) erzeugt werden.

Die Situation verallgemeinert damit die bei linearen oder affinen Teilräumen $\mathcal{A} \subseteq \mathbb{R}^n$ bekannte. Einerseits ist \mathcal{A} Lösungsmenge eines linearen Gleichungssystems $A\mathbf{x} = \mathbf{b}$. Andererseits gibt es eine endliche Menge $S = \mathbf{s}_1, \dots, \mathbf{s}_k$ derart, dass \mathcal{A} die Menge aller affinen Linearkombinationen

$$\mathbf{x} = \lambda_1 \mathbf{s}_1 + \dots + \lambda_k \mathbf{s}_k \quad \text{mit} \quad \sum_{i=1}^k \lambda_i = 1$$

ist. Die Umrechnung von einer Darstellung zur anderen ist im linearen/affinen Fall effizient möglich (z.B. mit dem Gauss-Verfahren).

NOTA BENE. Im linearen Fall sind alle minimalen Erzeugendensysteme (Basen) gleichmächtig. *Bei Ungleichungssystemen ist dies nicht mehr notwendigerweise so!*

Im allgemeinen Fall ist die Umrechnung nicht so einfach möglich. Wie der Beweis des Dekompositionssatzes zeigt, ist im Prinzip eine Umrechnung mit Hilfe des

Fourier-Motzkin-Verfahrens möglich. Diese Methode ist aber nicht effizient. Ein effizienter Algorithmus für das Umrechnungsproblem ist nicht bekannt.

1.4. Diskrete Optimierung und Polytope. Ein Grundproblem der diskreten Optimierung kann so formuliert werden. Gegeben ist eine Grundmenge E und eine Gewichtsfunktion $w : E \rightarrow \mathbb{R}$. Ausserdem sei eine Familie $\mathcal{F} \subseteq 2^E$ von Teilmengen spezifiziert. Man hat nun die Aufgabe

$$(34) \quad \max_{F \in \mathcal{F}} \sum_{e \in F} w(e).$$

Der Teilmengenfamilie $\mathcal{F} \subseteq 2^E$ ordnet man folgendermassen ein Polytop zu. Man repräsentiert jedes $F \in \mathcal{F}$ durch seinen Inzidenzvektor $\chi_F \in \mathbb{R}^E$, wobei

$$\chi_F(e) = \begin{cases} 1 & \text{wenn } e \in F \\ 0 & \text{wenn } e \notin F. \end{cases}$$

und definiert nun

$$\mathbb{P}(\mathcal{F}) = \text{conv}\{\chi_E \mid F \in \mathcal{F}\} \subseteq \mathbb{R}^E.$$

Das diskrete Optimierungsproblem (34) wird nun zu dem Problem, die lineare Funktion mit den Koeffizienten $w_e = w(e)$ über dem Polytop $\mathbb{P}(\mathcal{F})$ zu maximieren:

$$\max_{\mathbf{x} \in \mathbb{P}(\mathcal{F})} \sum_{e \in E} w_e x_e = \max_{F \in \mathcal{F}} \sum_{e \in E} w_e \chi_F(e) = \max_{F \in \mathcal{F}} \sum_{e \in F} w(e).$$

1.4.1. *Das Zuordnungs- und Heiratsproblem.* Wir gehen von endlichen und gleichmächtigen Mengen S und T (d.h. $|S| = |T|$) aus und betrachten die Menge aller Paare

$$S \times T = \{(s, t) \mid s \in S, t \in T\}$$

Eine *Zuordnung* (bzw. ein *perfektes Matching*) ist eine bijektive Abbildung $\pi : S \rightarrow T$. Wir stellen uns die Zuordnung als Menge von Paaren vor:

$$M = M(\pi) = \{(s, \pi(s)) \mid s \in S\} \subseteq S \times T.$$

\mathcal{M} sei die Menge aller Zuordnungen. Das *Zuordnungsproblem* bzgl. der Gewichtsfunktion $w : S \times T \rightarrow \mathbb{R}$ ist nun:

$$\max_{M \in \mathcal{M}} \sum_{(s,t) \in M} w(s, t)$$

Im Spezialfall $w : S \times T \rightarrow \{0, 1\}$ spricht man auch vom *Heiratsproblem*.

Das Zuordnungspolytop $\mathbb{P}(\mathcal{M})$ ist von der Menge \mathcal{M} aller Zuordnungen erzeugt. Wir suchen eine implizite Beschreibung, d.h. ein System linearer Ungleichungen dessen Lösungen \mathbf{x} gerade die Punkte in $\mathbb{P}(\mathcal{M})$ sind.

Sei $\mathbf{x} \in \mathbb{P}(\mathcal{M})$. Wir bezeichnen die Komponenten von \mathbf{x} mit x_{st} . Welche Ungleichungen muss \mathbf{x} erfüllen? Ist \mathbf{x} der Inzidenzvektor eines perfekten Matchings,

dann gilt sicherlich:

$$\begin{aligned} \text{(M0)} \quad x_{s,t} &\geq 0 \quad \text{für alle } (s,t) \in S \times T. \\ \text{(M1)} \quad \sum_{s \in S} x_{st} &= 1 \quad \text{für alle } s \in S. \\ \text{(M2)} \quad \sum_{t \in T} x_{st} &= 1 \quad \text{für alle } t \in T. \end{aligned}$$

Diese Ungleichungen gelten natürlich auch für alle Konvexkombinationen von Zuordnungen und deshalb für alle $\mathbf{x} \in \mathbb{P}(\mathbf{M})$. Es stellt sich heraus, dass sie $\mathbb{P}(\mathcal{M})$ schon vollständig bestimmen!

LEMMA 3.3. $\mathbb{P}(\mathcal{M}) = \{\mathbf{x} \in \mathbb{R}^{S \times T} \mid \mathbf{x} \text{ erfüllt (M0)-(M2)}\}$.

Das Lemma wird hier nicht bewiesen, da es aus sich später als Folgerung aus einem allgemeinen Optimierungsalgorithmus für sog. Flüsse in Netzwerken ergeben wird.

BEMERKUNG. Man bemerke, dass es $|S|!$ viele Zuordnungen gibt und diese alle Ecken von $\mathbb{P}(\mathcal{M})$ sind. Zur Beschreibung von $\mathcal{P}(\mathcal{M})$ genügen aber schon $2|S|$ Gleichungen und $|S|^2$ Ungleichungen.

1.4.2. *Das Rundreiseproblem.* Wir gehen von einer endlichen Menge S aus und betrachten die Menge $E = S \times S$. Eine *Rundreise* (oder aus *TSP-Tour*) ist eine Anordnung der Elemente von S :

$$\tau = s_0 s_1 \dots, s_n s_0,$$

bei der ausser s_0 kein Element zweimal auftritt. Wieder stellen wir τ als Menge von Paaren dar:

$$T = T(\tau) = \{(s_0, s_1), \dots, (s_n, s_0)\}$$

Gegeben eine *Distanzfunktion* $d : S \times S \rightarrow \mathbb{R}_+$, definiert man die *Länge* von T als

$$d(T) = \sum_{(s,t) \in T} d(s,t).$$

Sei \mathcal{T} die Menge aller Rundreisen. Das *Rundreiseproblem* (*TSP-Problem*) ist

$$\min_{T \in \mathcal{T}} d(T) \quad \longleftrightarrow \quad \max_{T \in \mathcal{T}} \sum_{(s,t) \in T} -d(s,t).$$

Auch hier kann man natürlich das entsprechende Rundreisepolyeder $\mathbb{P}(\mathcal{T})$ definieren, dessen Struktur allerdings zum grossen Teil noch ungeklärt ist. Obwohl sehr viele Klassen von gültigen Ungleichungen für $\mathbb{P}(\mathcal{T})$ bekannt sind, ist eine vollständige Beschreibung eines der grossen gegenwärtigen offenen Probleme der Berechenbarkeitstheorie der theoretischen Informatik.

1.5. Optimierung linearer Funktionen. Wir betrachten das lineare Programm

$$(35) \quad \max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b} \quad (A \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m).$$

Stellen wir den Lösungsraum $P = P(A, \mathbf{b})$ nach Weyl-Minkowski in der Form

$$P = \text{conv}(V) + \text{cone}(W)$$

dar, so sieht man sofort:

- (1) Gibt es ein $\mathbf{w} \in \text{cone}W$ mit $\mathbf{c}^T \mathbf{w} > 0$, so sind die Zielfunktionswerte nach oben unbeschränkt („ ∞ “).
- (2) Gilt $\mathbf{c}^T \mathbf{w} \leq 0$ für alle $\mathbf{w} \in \text{cone}(W)$ und ist $V \neq \emptyset$, dann ist

$$\max_{\mathbf{x} \in P} \mathbf{c}^T \mathbf{x} = \max_{\mathbf{v} \in V} \mathbf{c}^T \mathbf{v} < \infty.$$

Beweis von (2): Es gilt unter den angenommenen Umständen

$$\max_{\mathbf{x} \in P} \mathbf{c}^T \mathbf{x} = \max_{\mathbf{v} \in \text{conv}V} \mathbf{c}^T \mathbf{x}.$$

Sei $V = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ und $\mathbf{v} = \sum_{i=1}^k \mu_i \mathbf{v}_i \in \text{conv}V$ ein beliebiges Element. Dann finden wir wegen $\mu_i \geq 0$ und $\sum_i \mu_i = 1$:

$$\mathbf{c}^T \mathbf{v} = \sum_{i=1}^k \mu_i \mathbf{c}^T \mathbf{v}_i \leq \left(\max_{j=1, \dots, k} \mathbf{c}^T \mathbf{v}_j \right) \cdot \sum_{i=1}^k \mu_i = \max_{\mathbf{v}_j \in V} \mathbf{c}^T \mathbf{v}_j.$$

◇

Diese Beobachtungen zeigen, dass das lineare Optimierungsproblem in der Theorie darauf reduziert werden kann:

- (i) Stelle fest, ob die Ungleichungsrelation $\mathbf{c}^T \mathbf{w} \leq 0$ für alle $\mathbf{w} \in W$ gilt (d.h. ob $\mathbf{c} \in P(W^T, \mathbf{0}) = W^{\text{pol}}$ zutrifft);
- (ii) Wenn ja, löse das Problem $\max_{\mathbf{v} \in V} \mathbf{c}^T \mathbf{v}$.

1.6. Rezessionskegel und polyedrische Kegel. Wir nehmen

$$P = \text{conv}(V) + \text{cone}(W)$$

mit endlichen Mengen V und $W \neq \emptyset$ an. Dann heisst der konvexe Kegel

$$P_0 = \text{cone}(W)$$

der *Rezessionskegel* des Polyeders P . P_0 hängt nur von P ab.

LEMMA 3.4. *Seien A und \mathbf{b} beliebig mit der Eigenschaft $P = P(A, \mathbf{b})$ gewählt. Dann gilt*

$$\text{cone}(W) = P(A, \mathbf{0}).$$

Beweis. Im Fall $V = \emptyset$ ist die Behauptung trivial. Ansonsten wählen wir zu $\mathbf{w} \in \text{cone}(W)$ ein $\mathbf{v} \in V$ und betrachten die Punkte $\mathbf{p}_\lambda = \mathbf{v} + \lambda \mathbf{w}$ für alle $\lambda \geq 0$. Wegen $\mathbf{p}_\lambda \in P(A, \mathbf{b})$ finden wir

$$\lim_{\lambda \rightarrow \infty} \lambda(A\mathbf{w}) = \lim_{\lambda \rightarrow \infty} A(\lambda \mathbf{w}) \leq \mathbf{b} - A\mathbf{v}$$

und folglich $A\mathbf{w} \leq \mathbf{0}$. Also muss $\text{cone}(W) \subseteq P(A, \mathbf{0})$ gelten. Angenommen, die Enthaltenseinsrelation wäre strikt: $\text{cone}(W) \subset P(A, \mathbf{0})$. Dann hätten wir (nach Lemma 3.1) auch bei den Polaren die Beziehung

$$P(A, \mathbf{0})^{pol} \subset \text{cone}(W)^{pol}.$$

Es gibt also ein \mathbf{c} mit der Eigenschaft $\mathbf{c}^T \mathbf{w} \leq 1$ für alle $\mathbf{w} \in \text{cone}(W)$ aber $\mathbf{c}^T \mathbf{z} > 1$ für mindestes ein $\mathbf{z} \in P(A, \mathbf{0})$. Das kann aber nicht sein. Denn die erste Relation bedeutet, dass die lineare Funktion $f(\mathbf{x}) = \mathbf{c}^T \mathbf{x}$ auf P nach oben beschränkt ist:

$$\max_{\mathbf{x} \in P} \mathbf{c}^T \mathbf{x} \leq \max_{\mathbf{v} \in \text{conv}(V)} \mathbf{c}^T \mathbf{v} + \max_{\mathbf{w} \in \text{cone}(W)} \mathbf{c}^T \mathbf{w} < +\infty.$$

Die zweite Relation erweist $f(\mathbf{x})$ als auf P unbeschränkt:

$$\lim_{\lambda \rightarrow \infty} \mathbf{c}^T (\mathbf{v} + \lambda \mathbf{z}) = \mathbf{c}^T \mathbf{v} + \lim_{\lambda \rightarrow \infty} \lambda (\mathbf{c}^T \mathbf{z}) = +\infty.$$

◇

2. Seitenflächen, Ecken und Facetten

Wir betrachten ein Polyeder $P = P(A, \mathbf{b}) \subseteq \mathbb{R}^n$ und eine lineare Ungleichung $\mathbf{c}^T \mathbf{x} \leq z$. Wir nennen die Menge $F \subseteq \mathbb{R}^n$ eine *Seitenfläche* von P wenn gilt

$$F = \{\mathbf{x} \in P \mid \mathbf{c}^T \mathbf{x} = z\} \quad \text{und} \quad P \subseteq P(\mathbf{c}^T, z).$$

BEISPIEL 3.1 (Triviale Seitenflächen). Die Wahl $\mathbf{c} = \mathbf{0}$ und $z = 0$ zeigt, dass P selber eine Seitenfläche ist:

$$P = \{\mathbf{x} \in P \mid \mathbf{0}^T \mathbf{x} = 0\}.$$

Mit $\mathbf{c} = \mathbf{0}$ und $z = 1$ erhält man die leere Menge als Seitenfläche von P :

$$\emptyset = \{\mathbf{x} \in P \mid \mathbf{0}^T \mathbf{x} = 1\} \quad \text{und} \quad P \subseteq \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{0}^T \mathbf{x} \leq 1\}.$$

P und \emptyset sind die sog. trivialen Seitenflächen von P . Die übrigen Seitenflächen (sofern sie existieren) sind nichttrivial.

Der Punkt $\mathbf{v} \in P$ heisst *Ecke* (oder *Extrempunkt*) von P , Wenn $F = \{\mathbf{v}\}$ eine Seitenfläche von P ist.

BEMERKUNG. Die Seitenfläche F von P ist selber ein Polyeder, denn F ist die Lösungsmenge des linearen Systems

$$\begin{array}{rcl} -\mathbf{c}^T \mathbf{x} & \leq & -z \\ A\mathbf{x} & \leq & \mathbf{b} \end{array}$$

Man bemerke, dass aus der Sicht der Optimierung die obige Seitenfläche im Fall $F \neq \emptyset$ nichts anderes als die Menge der Optimallösungen des linearen Programms

$$(36) \quad \max \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}$$

ist.

2.1. Charakterisierung von Seitenflächen. Die nichttriviale Seitenfläche

$$F = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} \leq \mathbf{b} \text{ und } \mathbf{c}^T \mathbf{x} = z\}$$

des Polyeders $P = P(A, \mathbf{b})$ ist die Menge der Optimallösungen des linearen Programms

$$\max \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}.$$

Sei $\bar{\mathbf{y}} \geq \mathbf{0}$ eine Optimallösung des dazu dualen linearen Programms

$$\min \mathbf{y}^T \mathbf{b} \quad \text{s.d.} \quad \mathbf{y}^T A = \mathbf{c}^T, \mathbf{y} \geq \mathbf{0}.$$

Nach den Bedingungen des komplementären Schlupfes ist $\mathbf{x} \in P$ somit genau dann in F , wenn für alle Zeilenindices i gilt

$$\bar{y}_i > 0 \quad \implies \quad \mathbf{a}_i^T \mathbf{x} = b_i.$$

Sei $A_F \mathbf{x} \leq \mathbf{b}_F$ das Teilsystem aller Ungleichungen $\mathbf{a}_i^T \mathbf{x} \leq b_i$ mit $\bar{y}_i > 0$. Dann gilt also

$$F = \{x \in P \mid A_F \mathbf{x} = \mathbf{b}_F\}.$$

SATZ 3.2. *Eine nichtleere Teilmenge $F \subseteq P$ ist genau dann eine Seitenfläche des Polyeders $P = P(A, \mathbf{b})$, wenn es ein (möglicherweise leeres) Teilsystem $A_F \mathbf{x} \leq \mathbf{b}_F$ gibt mit der Eigenschaft*

$$F = \{\mathbf{x} \in P \mid A_F \mathbf{x} = \mathbf{b}_F\}.$$

Beweis. Es ist nur noch zu beweisen, dass die Bedingung hinreichend ist. Im Fall eines leeren Teilsystems liegt einfach $F = P$ als triviale Seitenfläche vor.

Ansonsten wählen wir \mathbf{c}^T als Summe der Zeilenvektoren von A_F und z als Summe der Komponenten von \mathbf{b}_F . Jedes $\mathbf{x} \in P$ erfüllt dann natürlich die Bedingung $\mathbf{c}^T \mathbf{x} \leq z$. Jedes $\mathbf{x} \in F$ ist somit optimal für das lineare Programm

$$\max \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}.$$

Jede Optimallösung \mathbf{x}^* muss $\mathbf{c}^T \mathbf{x}^* = z$ und somit $A_F \mathbf{x}^* = \mathbf{b}_F$ erfüllen. Also besteht F tatsächlich aus sämtlichen Optimallösungen und ist damit als Seitenfläche erkannt.

◇

Da $A\mathbf{x} \leq \mathbf{b}$ nur endlich viele Teilsysteme vom Typ $A_F \mathbf{x} \leq \mathbf{b}_F$ gestattet, schließen wir:

KOROLLAR 3.2. *Ein Polyeder P hat nur endlich viele Seitenflächen.*

◇

2.2. Dimension. Aus notationstechnischen Gründen nehmen wir nun an, dass das System $Ax \leq b$ die triviale Ungleichung $0^T x \leq 0$ enthält. Mit A_P bezeichnen wir die Teilmatrix aller Zeilenvektoren a_i^T für die gilt

$$a_i^T x = b_i \quad \text{für alle } x \in P(A, b)$$

und definieren den affinen Teilraum

$$\mathcal{A}_P = \begin{cases} \emptyset & \text{wenn } P = \emptyset \\ \{x \in \mathbb{R}^n \mid A_P x = b_P\} & \text{sonst.} \end{cases}$$

PROPOSITION 3.1. \mathcal{A}_P ist der (eindeutig bestimmte) kleinste affine Teilraum von \mathbb{R}^n , der das Polyeder $P = P(A, b)$ enthält.

Beweis. Sei oBdA $P \neq \emptyset$. Nach Definition gilt $P \subseteq \mathcal{A}_P$. Wir zeigen, dass jede lineare Gleichung, die von allen Punkten in P erfüllt wird, notwendigerweise eine Linearkombination von $A_P x = b_P$ ist. Also ist \mathcal{A}_P die kleinste Lösungsmenge eines linearen Gleichungssystems, die P enthält.

Sei z.B. $c^T x = z$ eine solche Gleichung. Dann ist jedes $x \in P$ Optimallösung des linearen Programms

$$\max c^T x \quad \text{s.d.} \quad Ax \leq b.$$

Sei $y \geq 0$ eine Optimallösung des zugehörigen dualen Programms. Dann gilt

$$y^T A = c^T.$$

Ausserdem gilt für alle Indizes i und alle $x \in P$ nach der komplementären Schlupf-eigenschaft:

$$y_i > 0 \implies a_i^T x = b_i \quad \text{und} \quad y^T b = z.$$

Also ist c^T sogar schon ein Linearkombination von A_P allein und z ist die entsprechende Linearkombination von b_P .

◇

Aus der linearen Algebra wissen wir, dass die (affine) Dimension von \mathcal{A}_P sich aus dem Rang von der Matrix A_P berechnen lässt:

$$\dim \mathcal{A}_P = \begin{cases} -1 & \text{wenn } \mathcal{A}_P = \emptyset \\ n - \text{rg} A_P & \text{sonst.} \end{cases}$$

Wir definieren die *Dimension* des Polyeders $P = P(A, b) \neq \emptyset$ als

$$\boxed{\dim P = \dim \mathcal{A}_P = n - \text{rg}(A_P)}$$

Eine nichtleere Seitenfläche $F = \{x \in P(A, b) \mid A_F x = b_F\}$ ist selber ein Polyeder und hat somit die Dimension:

$$\boxed{\dim F = n - \text{rg} \begin{bmatrix} A_P \\ A_F \end{bmatrix}}$$

2.3. Facetten. Eine *Facette* eines Polyeders $P = P(A, \mathbf{b})$ ist eine Seitenfläche F mit

$$\dim F = \dim P - 1.$$

SATZ 3.3. Sei F eine nichttriviale Seitenfläche des Polyeders $P = P(A, \mathbf{b})$. Dann ist F ein Durchschnitt von Facetten von P .

Beweis. Sei $F = \{\mathbf{x} \in P \mid A_F \mathbf{x} = \mathbf{b}_F\}$. OBdA können wir annehmen, dass keine der Gleichungen $\mathbf{a}_i \mathbf{x} = b_i$ des Systems $A_F \mathbf{x} = \mathbf{b}_F$ von sämtlichen $\mathbf{x} \in P$ erfüllt wird.

Für den entsprechenden affinen Teilraum gilt deshalb

$$\mathcal{A}_i = \{\mathbf{x} \in \mathbb{R}^n \mid A_P \mathbf{x} = \mathbf{b}_P, \mathbf{a}_i^T \mathbf{x} = b_i\} \neq \mathcal{A}_P$$

gilt folglich $\dim \mathcal{A}_i = \dim \mathcal{A}_P - 1 = \dim P - 1$. Also ist die Seitenfläche

$$F_i = \{\mathbf{x} \in P \mid \mathbf{a}_i^T \mathbf{x} = b_i\}$$

eine Facette. F ist folglich der Durchschnitt aller Facetten F_i von P , die man so aus $A_F \mathbf{x} = \mathbf{b}_F$ erhält.

◇

2.4. Ecken. Ein polyedrischer Kegel $P(A, \mathbf{0}) \subseteq \mathbb{R}^n$ besitzt nur den Nullvektor $\mathbf{0} \in P(A, \mathbf{0})$ als Kandidaten für eine Ecke. Genau im Fall $\text{rg} A = n$ ist $\mathbf{0}$ tatsächlich eine Ecke. Anders ausgedrückt:

- $P(A, \mathbf{0})$ besitzt genau dann *keine* Ecke, wenn $P(A, \mathbf{0})$ eine Gerade enthält.
- Wenn der Rezessionskegel $P(A, \mathbf{0})$ des Polyeders $P = P(A, \mathbf{b})$ nicht $\mathbf{0}$ als Ecke hat, dann besitzt auch P keine Ecke.

BEMERKUNG. Ein Kegel mit einer Ecke ist ein sog. *spitzer* Kegel.

SATZ 3.4 (Ecken von Polyedern). Sei V eine minimale Menge mit der Eigenschaft

$$P = P(A, \mathbf{b}) = \text{conv}(V) + P(A, \mathbf{0}).$$

Ist der Rezessionskegel $P(A, \mathbf{0})$ spitz, dann sind alle Punkte $\mathbf{v} \in V$ Ecken von P .

Beweis. Sei $\mathbf{v}^* \in V$ und $V' = V \setminus \{\mathbf{v}^*\}$. Wir setzen

$$P' = \text{conv}(V') + P(A, \mathbf{0}).$$

Aus der Minimalität von V folgt nun $P' \neq P$ und insbesondere $\mathbf{v}^* \notin P'$ (Beweis?). Der Hauptsatz über abgeschlossene konvexe Mengen garantiert somit eine Hyperebene, die \mathbf{v}^* von P' trennt. D.h. es gibt einen Parametervektor \mathbf{c} mit den Eigenschaften

- $\mathbf{c}^T \mathbf{x}' < \mathbf{c}^T \mathbf{v}^* < \infty$ für alle $\mathbf{x}' \in P'$ und folglich auch
- $\mathbf{c}^T \mathbf{w} \leq 0$ für alle $\mathbf{w} \in P(A, \mathbf{0})$.

Da $P(A, \mathbf{0})$ spitz ist, gibt es zudem einen Vektor $\bar{\mathbf{c}}$, mit der Eigenschaft

$$\bar{\mathbf{c}}^T \mathbf{w} < 0 \quad \text{für alle } \mathbf{w} \in P(A, \mathbf{0}) \setminus \{\mathbf{0}\}.$$

Nun wählen wir $\varepsilon > 0$ so klein, dass für jedes $\mathbf{v}' \in V'$ gilt:

$$(\mathbf{c} + \varepsilon \bar{\mathbf{c}})^T \mathbf{v}' = \mathbf{c}^T \mathbf{v}' + \varepsilon (\bar{\mathbf{c}}^T \mathbf{v}') < \mathbf{c}^T \mathbf{v}^* + \varepsilon (\bar{\mathbf{c}}^T \mathbf{v}^*).$$

Jetzt rechnet man leicht nach (Beweis?), dass \mathbf{v}^* die einzige Optimallösung des linearen Optimierungsproblems

$$\max (\mathbf{c}^T + \varepsilon \bar{\mathbf{c}})^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}.$$

ist. Damit ist \mathbf{v}^* als Ecke des Polyeders P erkannt. ◇

KOROLLAR 3.3. *Jedes nichtleere Polytop ist die konvexe Hülle seiner Ecken.*

Beweis. Ein nichtleeres Polytop $P = \text{conv}(V) = P(A, \mathbf{b})$ hat den Rezessionskegel $P(A, \mathbf{0}) = \{\mathbf{0}\}$, der trivialerweise spitz ist. ◇

2.4.1. *Basislösungen.* Wir interessieren uns für *nichtnegative* Lösungen linearer Gleichungssysteme d.h. für zulässige Lösungen von linearen Systemen der Form

$$(37) \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}$$

wobei $A \in \mathbb{R}^{m \times n}$ und $\mathbf{b} \in \mathbb{R}^m$. OBdA dürfen wir hier annehmen, dass A vollen Zeilenrang hat (d.h. $\text{rg}A = m$). Sei

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

das Polyeder aller zulässigen Lösungen des linearen Systems (37). Dann besteht der Rezessionskegel von P genau aus den zulässigen Lösungen von

$$A\mathbf{x} = \mathbf{0}, \mathbf{x} \geq \mathbf{0}.$$

Wegen $\mathbf{x} \geq \mathbf{0}$ ist klar, dass der Rezessionskegel keine Gerade enthalten kann (und folglich spitz ist). P ist also entweder leer oder enthält Ecken. Eine Ecke $\mathbf{v} \in P$ heisst auch *Basislösung* von (37).

Die Terminologie erklärt sich sofort so. \mathbf{v} ist eindeutige Lösung eines Teilsystems von (37) mit Rang n . Wegen $\text{rg}A = m$ muss die Basislösung \mathbf{v} also auf einer Teilmenge N von mindestens $|N| \geq n - m$ Indizes j die Bedingung

$$\mathbf{v}_j = 0 \quad (j \in N)$$

erfüllen. Sei B die Menge aller übrigen Indizes und A_B die Einschränkung von A auf die B entsprechenden Spalten. (Nach Indizes n in B bzw. N geordnet) ist \mathbf{v} die eindeutige Lösung von

$$A_B \mathbf{v}_B = \mathbf{b} \quad \text{und} \quad \mathbf{v}_N = \mathbf{0}_N.$$

Also muss A_B vollen Rang m besitzen. Mit anderen Worten:

- A_B ist eine Basis für den Spaltenraum von A .

TERMINOLOGIE: B ist die Menge der *Basisindizes* und N die Menge der *Nicht-Basisindizes* bzgl. der Ecke \mathbf{v} .

Wir halten fest:

SATZ 3.5. *Das lineare System (37) hat entweder keine zulässige Lösung oder (mindestens) eine zulässige Basislösung.*

◇

KOROLLAR 3.4 (Satz von Carathéodory). *Sei $X \subseteq \mathbb{R}^d$ eine beliebige nichtleere Menge von Vektoren und $\mathbf{z} \in \text{conv}(X)$. Dann lässt sich \mathbf{z} als Konvexkombination von höchstens $d + 1$ Vektoren aus X darstellen.*

Beweis. Übung. (Hinweis: Existenz von zulässigen Basislösungen.)

2.4.2. Konstruktion von guten Basislösungen.

PROPOSITION 3.2. *Sei $\mathbf{x}^{(0)} \geq \mathbf{0}$ eine Lösung von $A\mathbf{x} = \mathbf{b}$ und \mathbf{c} ein beliebiger Parametervektor. Dann gilt entweder*

$$(a) \min\{\mathbf{c}^T \mathbf{x} \mid A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} = -\infty$$

oder

(b) *man kann leicht (in weniger als n Schritten) eine zulässige Basislösung \mathbf{x}^* mit $\mathbf{c}^T \mathbf{x}^* \leq \mathbf{c}^T \mathbf{x}^{(0)}$ konstruieren.*

Beweis. OBdA dürfen wir bei allen Komponenten $x_j^{(0)} > 0$ annehmen (sonst ist die j te Spalte in $A\mathbf{x} = \mathbf{b}$ überflüssig.) Wir wollen \mathbf{x}^* iterativ in höchstens n Schritten konstruieren und lösen im ersten Schritt das lineare Gleichungssystem

$$\begin{aligned} \mathbf{c}^T \mathbf{d} &= -1 \\ A\mathbf{d} &= \mathbf{0} \end{aligned}$$

Falls eine Lösung \mathbf{d} existiert, dann gilt $A(\mathbf{x}^{(0)} + \lambda\mathbf{d}) = \mathbf{b}$ und

$$\mathbf{c}^T(\mathbf{x}^{(0)} + \lambda\mathbf{d}) = \mathbf{c}^T \mathbf{x}^{(0)} = -\lambda \quad \text{für alle } \lambda \geq 0.$$

Ist zusätzlich $\mathbf{x}^{(0)} + \lambda\mathbf{d} \geq \mathbf{0}$ immer garantiert, dann haben wir (a) vorliegen. Ansonsten gibt es einen Index j_1 und ein $\lambda_1 > 0$ mit

$$x_{j_1}^{(0)} + \lambda_1 d_{j_1} = 0 \quad \text{und} \quad x_j^{(0)} + \lambda_1 d_j \geq 0 \quad (j \neq j_1).$$

Wir setzen $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \lambda_1 \mathbf{d}$, entfernen Spalte j_1 aus $A\mathbf{x} = \mathbf{b}$ und verfahren mit $\mathbf{x}^{(1)}$ wie eben mit $\mathbf{x}^{(0)}$ usw.

Falls keine Lösung \mathbf{d} existiert, liegt \mathbf{c}^T im Zeilenraum der Matrix A . Folglich ist $\mathbf{c}^T \mathbf{x} = \mathbf{c}^T \mathbf{x}^{(0)}$ konstant auf der Menge der zulässigen Lösungen \mathbf{x} . Es genügt also, eine beliebige zulässige Basislösung \mathbf{x}^* zu konstruieren.

Wir wählen deshalb einfach ein $\mathbf{d} \in \ker A \setminus \{\mathbf{0}\}$ und berechnen $\mathbf{x}^{(1)}$ (entweder mit \mathbf{d} oder mit $-\mathbf{d}$) wie oben. Existiert ein solches \mathbf{d} nicht, dann sind alle Spalten von A linear unabhängig, d.h. $\mathbf{x}^{(0)}$ ist schon Basislösung.

◇

3. Rationale lineare Systeme

Ein Polyeder $P \subseteq \mathbb{R}^n$ heisst *rational*, wenn P mit rationalen Parametern präsentiert werden kann. Das ist auf zwei Arten möglich:

- Angabe einer Matrix $A \in \mathbb{Q}^{m \times n}$ und eines Vektors $\mathbf{b} \in \mathbb{Q}^m$ so, dass $P = P(A, \mathbf{b})$;
- Angabe von endlichen Mengen von Vektoren $V, W \subseteq \mathbb{Q}^n$ so, dass $P = \text{conv}(V) + \text{cone}(W)$.

Wir haben gesehen, dass alle Rechnungen (insbesondere Umrechnungen von Darstellungen) bei Polyedern im Prinzip mit dem Fourier-Motzkinschen Verfahren möglich sind. Das beruht nur auf den folgenden Operationen:

- Multiplikation mit einem positiven Skalar;
- (Komponentenweise) Addition von Vektoren.

Diese Operationen führen nie aus dem Skalarbereich \mathbb{Q} heraus. Kurz gesagt:

Sämtliche Kenngrößen rationaler Polyeder sind rational

Eine einfache (aber wichtige) Beobachtung ist die, dass bei einem rationalen Polyeder $P = P(A, \mathbf{b})$ sowohl die Koeffizienten a_{ij} von A als auch b_i von \mathbf{b} als ganzzahlig angenommen werden dürfen. Wenn man nämlich eine Ungleichung

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n \leq b_i$$

mit rationalen Koeffizienten $a_{ij}, b_i \in \mathbb{Q}$ mit deren gemeinsamem Hauptnenner multipliziert, erhält man eine ganzzahlige Ungleichung

$$a'_{i1}x_1 + a'_{i2}x_2 + \dots + a_{in}x_n \leq b'_i \quad (a'_{ij}, b_i \in \mathbb{Z})$$

mit derselben Lösungsmenge. Wir werden deshalb bei rationalen Polyedern $P = P(A, \mathbf{b})$ immer $A \in \mathbb{Z}^{m \times n}$ und $\mathbf{b} \in \mathbb{Z}^m$ annehmen.

3.1. Komplexität rationaler linearer Systeme. Sei $A = [a_{ij}] \in \mathbb{Z}^{m \times n}$ und $\mathbf{b} \in \mathbb{Z}^m$. Wir setzen

$$\gamma = \gamma(A, \mathbf{b}) = \min\{k \in \mathbb{N} \mid |a_{ij}| < 2^k, |b_i| < 2^k \text{ für alle } i, j\}$$

und nennen γ die (*numerische*) *Komplexität* des linearen Ungleichungssystems $A\mathbf{x} \leq \mathbf{b}$.

BEMERKUNG. $\gamma(A, \mathbf{b})$ ist in etwa die Anzahl der Stellen, die benötigt werden, um die Koeffizienten von A und von \mathbf{b} in Binärdarstellung auszudrücken.

Sei z.B. \mathbf{v} eine Ecke von $P(A, \mathbf{b})$. Dann existiert ein Teilsystem $\bar{A}\mathbf{x} \leq \bar{\mathbf{b}}$ aus n Ungleichungen derart, dass \mathbf{v} durch die Gleichung $\bar{A}\mathbf{v} = \bar{\mathbf{b}}$ eindeutig bestimmt ist. Nach der Cramerschen Regel ergeben sich die Komponenten als

$$v_i = \frac{\det(\bar{A}^{(i)})}{\det(\bar{A})}$$

wobei man $\bar{A}^{(i)}$ aus \bar{A} erhält, indem man die i te Spalte durch $\bar{\mathbf{b}}$ ersetzt. Ganzzahligkeit ergibt $|\det \bar{A}| \geq 1$. Mit $\gamma = \gamma(A, \mathbf{b})$ erhalten wir somit aus der Determinantenformel

$$|v_i| \leq |\det(\bar{A}^{(i)})| \leq \sum_{\pi} |\bar{a}_{1,\pi(1)}^{(i)} \cdots \bar{a}_{n,\pi(n)}^{(i)}| < n! 2^{\gamma n}$$

und daraus (wegen $n! \leq n^n$)

$$(38) \quad |v_i| < 2^{n \log_2 n} 2^{\gamma n} = 2^{n(\gamma + \log_2 n)}.$$

Insbesondere ist der euklidische Abstand von \mathbf{v} vom Ursprung höchstens

$$R = \sqrt{\sum_{i=1}^n |v_i|^2} < \sqrt{n(2^{n(\gamma + \log_2 n)})^2} = \sqrt{n} 2^{n(\gamma + \log_2 n)}$$

Ebenso erhält man auch eine Abschätzung von unten:

$$(39) \quad v_i = 0 \quad \text{oder} \quad |v_i| \geq \frac{1}{|\det A|} \geq \frac{1}{2^{n(\gamma + \log_2 n)}}.$$

3.2. Volumina rationaler Polytope. Wir nehmen an, $P = (A, \mathbf{b})$ ist ein Polytop mit $\text{vol}(P) > 0$. Es gibt deshalb $n + 1$ affin unabhängige Ecken $\mathbf{v}_0, \dots, \mathbf{v}_n$ mit der Eigenschaft

$$\text{vol}(P) \geq \left| \det \begin{bmatrix} 1 & \cdots & 1 \\ \mathbf{v}_0 & \cdots & \mathbf{v}_n \end{bmatrix} \right|$$

sei d_i der Nenner der Komponenten von \mathbf{v}_i (nach der Cramerschen Regel). Da die Zähler ganze Zahlen sind, gilt dann

$$\text{vol}(P) \geq \frac{1}{|d_0 d_1 \cdots d_n|} \geq \left(\frac{1}{n! 2^{\gamma n}} \right)^{n+1} > \frac{1}{2^{(n+1)n(\gamma + \log_2 n)}}$$

Wir halten fest:

- Für das rationale Polytop $P = P(A, \mathbf{b})$ gilt entweder $\text{vol}(P) = 0$ oder $\text{vol}(P) > 2^{-(n+1)n(\gamma + \log_2 n)}$.
- Im Fall $\text{vol}(P) \neq 0$ kann man mit der Ellipsoidmethode einen konkreten Vektor $\mathbf{t} \in P(A, \mathbf{b})$ bestimmen.

ÜBUNGSAUFGABE: Mit wievielen Iterationen kommt die Ellipsoidmethode im Fall $\text{vol}(P) \neq 0$ aus?

3.3. Lösbarkeit rationaler Systeme. Um den Fall $\text{vol}(P) = 0$ zu bewältigen, wählen wir eine grosse ganze Zahl $\Lambda > 0$ und betrachten das modifizierte System

$$(40) \quad \mathbf{Ax} \leq \mathbf{b} + \varepsilon \mathbf{1} \quad \text{bzw.} \quad \Lambda \mathbf{Ax} \leq \Lambda \mathbf{b} + \mathbf{1}$$

mit $\varepsilon = \Lambda^{-1}$ und $\mathbf{1} = (1, 1, \dots, 1)^T$.

LEMMA 3.5. *Sei $\mathbf{Ax} \leq \mathbf{b}$ ein lineares System mit ganzzahligen Koeffizienten. Dann kann man $\Lambda = \varepsilon^{-1} > 0$ so wählen, dass gilt:*

$$\mathbf{Ax} \leq \mathbf{b} \text{ lösbar} \iff \mathbf{Ax} \leq \mathbf{b} + \varepsilon \mathbf{1} \text{ lösbar.}$$

Beweis. Wir beweisen die nichttriviale Richtung der Behauptung und nehmen an, dass $\mathbf{Ax} \leq \mathbf{b}$ nicht lösbar ist. Nach dem Farkas-Lemma gibt es dann ein $\mathbf{y} \geq \mathbf{0}$ mit der Eigenschaft

$$\mathbf{y}^T \mathbf{A} = \mathbf{0}^T \quad \text{und} \quad \mathbf{y}^T \mathbf{b} = -1.$$

OBdA dürfen wir \mathbf{y} als Basislösung annehmen. Es gibt also höchstens $n + 1$ Komponenten $y_i \neq 0$. Für diese gilt

$$y_i < 2^{(n+1)(\gamma + \log_2(n+1))}.$$

Also schliessen wir

$$\mathbf{y}^T (\mathbf{b} + \varepsilon \mathbf{1}) < -1 + \frac{(n+1)2^{\gamma + \log_2(n+1)}}{\Lambda} < 0,$$

wenn Λ genügend gross ist. Wieder nach dem Farkas-Lemma erweist sich dann auch $\mathbf{Ax} \leq \mathbf{b} + \varepsilon \mathbf{1}$ als nicht lösbar. ◇

Offensichtlich gilt

$$\mathbf{Ax} \leq \mathbf{b} \text{ lösbar} \implies \text{vol}(P(\mathbf{A}, \mathbf{b} + \varepsilon \mathbf{1})) \neq 0.$$

Folgerung:

- Indem wir bei geeigneter Wahl von $\Lambda > 0$ die Ellipsoidmethode auf $\mathbf{Ax} \leq \mathbf{b} + \varepsilon \mathbf{1}$ anwenden, können wir zumindest testen, ob $\mathbf{Ax} \leq \mathbf{b}$ überhaupt eine Lösung besitzt.

ÜBUNGSAUFGABE: *Wieviele Iterationen der Ellipsoidmethode muss man höchstens ausführen, um sicher zu wissen, ob $\mathbf{Ax} \leq \mathbf{b}$ eine Lösung besitzt oder nicht?*

3.3.1. *Explizites Lösen rationaler Systeme.* Um ein rationales System explizit zu lösen, formen wir es zuerst in die Gestalt

$$\mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}$$

mit ganzzahligen Koeffizienten um und suchen nun folgendermassen nach einer zulässigen Basislösung:

- Entferne eine (beliebige) Spalte aus \mathbf{A} und teste, ob das Restsystem noch lösbar ist.
- (i) Wenn das Restsystem lösbar ist, fahren wir in gleicher Weise mit dem Restsystem fort.

- (ii) Wenn das Restsystem nicht lösbar ist, so tritt die eben betrachtete Spalte notwendigerweise in *jeder* zulässigen Basislösung auf. Wir markieren nun diese Spalte und wählen eine andere Spalte, um wie eben vorzugehen.
- Nach endlich vielen Schritten ist eine zulässige Spaltenbasis A_B gefunden. Die zugehörige Lösung kann man nun (z.B. mit dem Gaussverfahren) durch Lösen von

$$A_B \mathbf{x}_B = \mathbf{b}$$

ermitteln.

3.4. Methode der inneren Punkte (IPM). Wir wollen das System

$$(41) \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0} \quad (A \in \mathbb{Z}^{m \times n}, \mathbf{b} \in \mathbb{Z}^m)$$

lösen¹. Falls überhaupt eine Lösung existiert, dann auch eine zulässige Basislösung \mathbf{x}^* , für deren Komponenten $x_j^* \geq 0$ wir eine ganzzahlige obere Schranke $M \geq x_j^*$ ausrechnen können, die wir notfalls in die Restriktionen (durch Einführen einer neuen nichtnegativen Variablen x_M) als

$$x_1 + x_2 + \dots + x_n + x_M = nM$$

aufnehmen könnten, ohne das Problem wesentlich zu verändern. Also dürfen wir oBdA annehmen, dass der Lösungsbereich

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

von (41) beschränkt (und damit ein Polytop) ist.

Wir formulieren nun zu (41) das lineare Programm (LP):

$$(42) \quad \begin{array}{ll} \min & \lambda \\ \text{s.d.} & A\mathbf{x} - \mathbf{b}t + (\mathbf{b} - A\mathbf{1})\lambda = \mathbf{0} \\ & \mathbf{1}^T \mathbf{x} + t + \lambda = n + 2 \\ & (\mathbf{x}, t, \lambda) \geq \mathbf{0} \end{array}$$

mit der zulässigen und strikt positiven Startlösung

$$\mathbf{v}_0 = \begin{bmatrix} \mathbf{x}_0 \\ t_0 \\ \lambda_0 \end{bmatrix} = \begin{bmatrix} \mathbf{1} \\ 1 \\ 1 \end{bmatrix} \in \mathbb{R}^{n+2}.$$

Das dazu duale LP hat offenbar $\mathbf{y}_0^T = (\mathbf{0}^T, -1)$ als zulässige Lösung mit dem Schlupfvektor

$$\mathbf{s}_0^T = (\mathbf{1}^T, 1, 2) > \mathbf{0}^T.$$

Für $\mu_0 = 1$ erhalten wir $\delta(\mathbf{v}_0, \mathbf{s}_0, \mu_0) = \frac{1}{2} < 1$ und können folglich das LP mit dem IPM-Verfahren beliebig genau lösen.

LEMMA 3.6. (41) ist lösbar genau dann, wenn (42) eine zulässige Lösung (\mathbf{x}, t, λ) mit $\lambda = 0$ und $t > 0$ besitzt.

¹wobei wir oBdA $\text{rg}A = m < n$ annehmen dürfen – denn im Fall $\text{rg}A = n$ ist die Lösung $\mathbf{x} = A^{-1}\mathbf{b}$ ja schon eindeutig bestimmt

Beweis. Die Bedingung ist sicher notwendig. Andererseits ergibt sich aus einer Lösung (\mathbf{x}, t, λ) mit $\lambda = 0$ und $t > 0$ die Lösung $\bar{\mathbf{x}} = \mathbf{x}/t$ von (41). \diamond

Sei nun $\gamma = \gamma(A, \mathbf{b})$ die Komplexität von $A \in \mathbb{Z}^{m \times n}$ und $\mathbf{b} \in \mathbb{Z}^m$. Wir berechnen mit IPM eine Lösung (\mathbf{x}, t, λ) des LPs mit

$$0 \leq \lambda < \frac{1}{2^{n(\gamma + \log_2 n)}}$$

und dazu (wie in Abschnitt 2.4.2) eine zulässige Basislösung $(\mathbf{x}^*, t^*, \lambda^*)$ mit Wert $\lambda^* \leq \lambda$. Gemäss (39) muss dann $\lambda^* = 0$ gelten (sofern (41) überhaupt eine Lösung besitzt).

Gleichzeitig muss aber auch $t^* > 0$ gelten. Denn sonst hätten wir mit dem Vektor $\mathbf{x}^* \geq \mathbf{0}$ ein nichttriviales Element im Rezessionskegel

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

des Lösungsbereichs P gefunden. P wäre somit kein Polytop.

ÜBUNGSAUFGABE: Wieviele IPM-Iterationen sind nötig, bis man das (ganzzahlige) System (41) durch Übergang zu einer Basislösung von (42) lösen kann?

4. Die Simplexmethode

Wir gehen von einem Optimierungsproblem der Form

$$(43) \quad \min \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}$$

aus, wobei die Problemparameter $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$ und $A = [a_{ij}] \in \mathbb{R}^{m \times n}$ gegeben seien. OBdA nehmen wir an, dass A vollen Zeilenrang $m = \text{rg}A$ hat. Wir stellen uns vor, dass unser Optimierungsproblem gewisse „Kosten“ minimieren will, und beziehen uns deshalb auf \mathbf{c} als den Vektor von *Kostenkoeffizienten*.

Die KKT-Bedingungen lauten

$$\begin{array}{l} \mathbf{c}^T \mathbf{x} = \mathbf{b}^T \mathbf{y} \\ A\mathbf{x} = \mathbf{b} \\ A^T \mathbf{y} \leq \mathbf{c} \\ \mathbf{x} \geq \mathbf{0} \end{array}$$

Ein KKT-Paar $(\mathbf{x}^*, \mathbf{y}^*)$ liefert gleichzeitig eine optimale Lösung des Problems

$$(44) \quad \max \mathbf{y}^T \mathbf{b} \quad \text{s.d.} \quad \mathbf{y}^T A \leq \mathbf{c}^T.$$

Wir nennen (43) in diesem Zusammenhang das *primale* und (44) das *duale* Problem. Die zentrale Idee des Simplexalgorithmus ist, die Ecken (d.h. die Basislösungen) der beiden Zulässigkeitsbereiche

$$\begin{aligned} P &= \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \\ P^* &= \{\mathbf{y} \in \mathbb{R}^m \mid \mathbf{y}^T A \leq \mathbf{c}^T\} \end{aligned}$$

zu untersuchen. Sei also $B = \{r_1, \dots, r_m\} \subseteq \{1, \dots, n\}$ eine Indexmenge derart, dass die aus den entsprechenden Spalten von A gebildete $(m \times m)$ -Matrix A_B eine

Spaltenbasis von A (bzw. A_B^T eine Zeilenbasis von A^T) ist. Wir nennen oft kurz auch die Indexmenge B selber eine *Basis*.

Zur leichteren Notation setzen wir $N = \{1, \dots, n\} \setminus B$ und bezeichnen mit \mathbf{c}_B den Teilvektor von \mathbf{c} , der nur aus den B -Komponenten besteht *etc.* Wir ordnen B die folgenden Kandidaten für Optimallösungen zu:

$$\begin{aligned}\bar{\mathbf{x}} &\in \mathbb{R}^n && \text{mit } \bar{\mathbf{x}}_B = A_B^{-1}\mathbf{b}, \bar{\mathbf{x}}_N = \mathbf{0}_N. \\ \bar{\mathbf{y}} &\in \mathbb{R}^m && \text{mit } \bar{\mathbf{y}}^T = \mathbf{c}_B^T A_B^{-1}.\end{aligned}$$

Dann gilt auf jeden Fall die Gleichheit

$$\mathbf{c}^T \bar{\mathbf{x}} = \mathbf{c}_B^T \bar{\mathbf{x}}_B + \mathbf{c}_N^T \bar{\mathbf{x}}_N = (A_B^T \bar{\mathbf{y}})^T \bar{\mathbf{x}}_B = \bar{\mathbf{y}}^T A_B \bar{\mathbf{x}} = \bar{\mathbf{y}}^T \mathbf{b}.$$

Ausserdem beobachtet man

$$\begin{aligned}(\text{Z}) \quad \bar{\mathbf{x}} \in P & \iff \bar{\mathbf{x}}_B \geq \mathbf{0}_B. \\ (\text{Z}^*) \quad \bar{\mathbf{y}} \in P^* & \iff A_N^T \bar{\mathbf{y}} \leq \mathbf{c}_N.\end{aligned}$$

BEMERKUNG. Wegen $\text{rg} A_B = m$ ist klar, dass die B zugeordneten Punkte mit den Koordinaten $\bar{\mathbf{x}}$ und $\bar{\mathbf{y}}$ Ecken sind, sofern sie zu P bzw. P^* gehören.

Ziel ist es nun, eine Indexmenge B (bzw. eine Spaltenbasis A_B) zu bestimmen, die sowohl Zulässigkeit bzgl. P als auch Zulässigkeit bzgl. P^* zur Folge hat. Denn wir wissen von den KKT-Bedingungen:

- Im Fall (i) $\bar{\mathbf{x}} \in P$ (d.h. $\bar{\mathbf{x}}_B \geq \mathbf{0}$)
und (ii) $\bar{\mathbf{y}} \in P^*$ (d.h. $\bar{\mathbf{c}}^T = \mathbf{c}^T - \bar{\mathbf{y}}^T A \geq \mathbf{0}^T$)
ist $\bar{\mathbf{x}}$ eine primal optimale und $\bar{\mathbf{y}}$ eine dual optimale Lösung.

Die „Simplexmethode“ ist eine Sammlung von systematischen Verfahren, um zu einer (in diesem Sinn) optimalen Basis B zu gelangen, indem man sich am bekannten Gauss-Algorithmus orientiert.

TERMINLOGIE: Ist B eine Basis, $\bar{A} = A_B^{-1}A$ und $\mathbf{y}^T = \mathbf{c}_B^T \bar{A}$, dann ist der duale Schlupfvektor

$$\bar{\mathbf{c}}^T = \mathbf{c}^T - \mathbf{c}_B^T \bar{A} = \mathbf{c}^T - \bar{\mathbf{y}}^T A = \mathbf{c}^T - \mathbf{c}_B^T A_B^{-1} A$$

der Vektor der sog. *reduzierten Kosten*. Wir suchen also eine primal zulässige Basis B mit nichtnegativen reduzierten Kosten.

4.1. Das Simplextableau. Bezgl. einer Basis B fassen wir die Ausgangsdaten des linearen Programms in einem Matrixschema zusammen:

$$\begin{aligned}z &= \mathbf{c}^T \mathbf{x} \\ \mathbf{b} &= A\mathbf{x}\end{aligned} \iff \left[\begin{array}{c|cc} 0 & \mathbf{c}_B^T & \mathbf{c}_N^T \\ \mathbf{b} & A_B & A_N \end{array} \right]$$

Nun benutzen wir elementare Zeilenoperationen (genau wie beim Gauss-Verfahren der linearen Algebra!) um das Schema in die folgende Form zu bringen (wobei $I_B = A_B^{-1}A_B$ die Einheitsmatrix ist):

$$(45) \quad \left[\begin{array}{c|cc} -\bar{z} & \mathbf{0}_B^T & \bar{\mathbf{c}}_N^T \\ \bar{\mathbf{b}} & I_B & \bar{A}_N \end{array} \right] = \left[\begin{array}{c|c} -\bar{z} & \bar{\mathbf{c}}^T \\ \bar{\mathbf{b}} & \bar{A} \end{array} \right]$$

Das Schema (45) ist das *Simplextableau zur Basis B* . Aus der linearen Algebra (Gauss-Verfahren) ist nun klar:

$$(1) \bar{\mathbf{b}} = A^{-1}\mathbf{b} \text{ und } \bar{A} = A_B^{-1}A.$$

$$(2) \bar{z} = \mathbf{c}_B^T \bar{\mathbf{b}} \text{ und } \bar{\mathbf{c}}^T = \mathbf{c}^T - \mathbf{c}_B^T \bar{A}.$$

Im Fall $\bar{\mathbf{b}} \geq \mathbf{0}$ ist das Tableau *primal zulässig* und im Fall $\bar{\mathbf{c}}^T \geq \mathbf{0}^T$ *dual zulässig*. Das Tableau ist *optimal*, wenn es primal und dual zulässig ist.

MAN BEACHTE: *In der linken oberen Ecke des Simplextableaus steht der negative Zielfunktionswert bzgl. der gerade betrachteten Basisvektoren $\bar{\mathbf{x}}$ und $\bar{\mathbf{y}}$! Um diesen „Schönheitsfehler“ auszugleichen, findet man in der Literatur das Simplextableau oft mit den negativen Kostenkoeffizienten angegeben:*

$$\left[\begin{array}{c|cc} 0 & -\mathbf{c}_B^T & -\mathbf{c}_N^T \\ \mathbf{b} & A_B & A_N \end{array} \right] \longrightarrow \left[\begin{array}{c|cc} \bar{z} & \mathbf{0}_B^T & -\bar{\mathbf{c}}_N^T \\ \bar{\mathbf{b}} & I_B & \bar{A}_N \end{array} \right]$$

(Man kann diese zweite Version des Simplextableaus natürlich auch direkt von der Darstellung

$$\begin{array}{rcl} z & -\mathbf{c}^T \mathbf{x} & = 0 \\ & A\mathbf{x} & = \mathbf{b} \end{array}$$

her motivieren, wenn man möchte.)

TERMINOLOGIE. Die Variablen x_i mit Index $i \in B$ heissen *Basisvariablen* (bzgl. B). Die x_j mit $j \in N$ sind die *Nichtbasisvariablen*.

NOTA BENE. Bei einem Simplextableau gehen wir **immer** von einer Problemformulierung vom Typ (43) aus. Insbesondere unterstellen wir automatisch, dass sämtliche Variablen x_j nichtnegativ sein sollen.

4.1.1. *Zur Erinnerung: Elementare Operationen.* Unter einer *elementaren Zeilenoperation* versteht man eine der folgenden Operationen auf den Zeilenvektoren eines Matrixschemas:

- (1) Multiplikation (Division) einer Zeile mit einem Skalar $a_{kj} \neq 0$.
- (2) Addition (Subtraktion) einer Zeile zu einer anderen.
- (3) Permutation zweier Zeilen.

Hat man ein *Pivotelement* $a_{kj} \neq 0$ festgewählt, so kann man durch Zeilenoperationen vom Typ (1) und (2) das Matrixschema so umformen, dass in der Spalte j der Einheitsvektor (mit Eintrag 1 in der k entsprechenden Komponente) zu stehen kommt.

BEMERKUNG. Die Vertauschoperation (3) ist nur für das menschliche Auge. Hat zum Beispiel eine quadratische Matrix A_B vollen Rangs vorliegen, so kann man sie mit elementaren Zeilenoperationen in die entsprechende Einheitsmatrix I_B transformieren. Ohne (3) erreicht man nur eine Transformation, in der zwar die verschiedenen Einheitsvektoren auftreten – aber nicht unbedingt in der dem menschlichen Auge ästhetisch gefälligen Diagonalfom.

Dem Computer ist die menschliche Optik – und damit die Vertauschoperation (3) – völlig egal. Hauptsache: man weiss, in welcher Zeile (Spalte) welcher Einheitsvektor einer Pivotoperation auftritt.

4.2. Die primale Strategie. Wir betrachten die Zeilenvektoren des Simplextableaus (45):

$$\begin{aligned}\bar{\alpha}_0 &= (-\bar{z}, \bar{c}_{i1}, \dots, \bar{c}_{in}) \\ \bar{\alpha}_i &= (\bar{b}_i, \bar{a}_{i1}, \dots, \bar{a}_{in}) \quad (i = 1, \dots, m)\end{aligned}$$

Wir nennen das Tableau *lexikographisch positiv*, wenn jeder der Zeilenvektoren $\bar{\alpha}_i$ mit $i \geq 1$ lexikographisch positiv ist. Im lexikographisch positiven Fall gilt natürlich insbesondere $\bar{b}_i \geq 0$ für $i = 1, \dots, m$. Also:

- Ein lexikographisch positives Tableau ist primal zulässig.

Wir nehmen nun im weiteren an:

- Das Tableau ist lexikographisch positiv.
- Das Tableau ist nicht dual zulässig, d.h. es gibt eine Spalte j mit reduzierten Kosten $\bar{c}_j < 0$.

LEMMA 3.7. *Enthält die j te Spalte $\bar{A}_j = A_B^{-1}A_j$ (im primal zulässigen Simplextableau) kein positives Element, dann ist das lineare Programm unbegrenzt (und es ist folglich sinnlos, nach einer „Optimallösung“ zu suchen).*

Beweis. Im Fall $\bar{A}_j \leq \mathbf{0}$ betrachten wir ein beliebiges $\lambda > 0$ und setzen

$$\begin{aligned}x'_j &= \lambda \\ x'_i &= \bar{b}_i - \bar{a}_{ij}\lambda \quad (i \in B) \\ x'_\ell &= 0 \quad (\ell \in N \setminus \{j\}).\end{aligned}$$

Man macht sich leicht klar, dass \mathbf{x}' primal zulässig ist. Der Zielfunktionswert ist

$$\mathbf{c}^T \mathbf{x}' = \mathbf{c}_B^T \bar{\mathbf{b}} - \mathbf{c}_B^T \bar{A}_j \lambda + c_j \lambda = \mathbf{c}_B^T \bar{\mathbf{b}} + \bar{c}_j \lambda \rightarrow -\infty \quad (\text{wenn } \lambda \rightarrow +\infty).$$

◇

Im Fall, wo \bar{A}_j mindestens ein positives Element hat, wählen wir nun ein Pivotelement \bar{a}_{kj} nach der sog. *primale lexikographische Regel*, d.h. derart, dass gilt

$$(46) \quad \frac{\bar{\alpha}_k}{\bar{a}_{kj}} = \text{lexmin} \left\{ \frac{\bar{\alpha}_i}{\bar{a}_{ij}} \mid \bar{a}_{ij} > 0 \right\}.$$

ZUR ERINNERUNG: Das lexikographische Minimum einer Menge von Vektoren ist der eindeutig bestimmte kleinste Vektor bzgl. der lexikographischen Ordnung.

Wir pivotieren nun mit \bar{a}_{jk} das Simplextableau, d.h. wir berechnen ein neues Matrixschema mit den Zeilen:

$$\begin{aligned}\alpha'_k &= \bar{a}_{ij}^{-1} \cdot \bar{\alpha}_k \\ \alpha'_i &= \bar{\alpha}_i - \bar{a}_{ij} \cdot \alpha'_k \quad \text{für alle } i \neq k.\end{aligned}$$

Das neue Schema hat die Eigenschaften

- (1) In der Spalte j steht der Einheitsvektor mit $a'_{kj} = 1$.
- (2) In der Spalten $\ell \in B \setminus \{k\}$ der alten Basisindexmenge B stehen unverändert die alten Einheitsvektoren.

Damit sieht man:

- Das neue Schema ist genau das Simplextableau zur neuen Basis

$$B' = (B \setminus \{k\}) \cup \{j\}.$$

LEMMA 3.8. *Pivotiert man ein lexikographisches Simplextableau mit negativem reduzierten Kostenkoeffizienten $\bar{c}_j < 0$ nach der primalen lexikographischen Regel, dann gilt*

- (1) *Das neue Simplextableau ist wieder lexikographisch positiv.*
- (2) *Die Zeile α'_0 des neuen Tableaus ist lexikographisch echt grösser als die alte Zeile $\bar{\alpha}_0$.*

Beweis. Wegen $\bar{a}_{kj} > 0$ bleibt die Zeile k lexikographisch positiv:

$$\alpha'_k = \bar{a}_{ij}^{-1} \cdot \bar{\alpha}_k.$$

Ansonsten hat man für $i \neq k$:

$$\begin{aligned} \alpha'_i &= \bar{\alpha}_i - \frac{\bar{a}_{ij}}{\bar{a}_{kj}} \bar{\alpha}_k && \text{(lex. pos., falls } \bar{a}_{ij} \leq 0) \\ &= \bar{a}_{ij} \left[\frac{\bar{\alpha}_i}{\bar{a}_{ij}} - \frac{\bar{\alpha}_k}{\bar{a}_{kj}} \right] && \text{(lex. pos., falls } \bar{a}_{ij} > 0). \end{aligned}$$

Daraus folgt (1) und im Spezialfall $i = 0$ wegen $\alpha_{0j} = \bar{c}_j < 0$ auch (2). ◇

4.3. Der Algorithmus. Wenn einmal eine lexikographisch positive Basis zur Verfügung steht, dann iteriert man nach der primalen lexikographischen Regel solange, bis man zu einer Basis B gekommen ist, bei der alle reduzierten Kosten nichtnegativ sind. Eine Optimallösung ist dann gefunden mit

$$\bar{\mathbf{x}}_B = \bar{\mathbf{b}} \quad \text{und} \quad \bar{\mathbf{x}}_N = \mathbf{0}_N.$$

PROPOSITION 3.3. *Iteriert man nach der primalen lexikographischen Regel, dann terminiert der Algorithmus nach endlich vielen Schritten.*

Beweis. Findet man kein Pivotelement in Spalte \bar{A}_j , dann stoppt man das Iterationsverfahren, weil man weiss, dass überhaupt keine Optimallösung existiert.

Ansonsten erhält man eine neue lexikographisch positive Basis B' und ein Tableau mit einer lexikographisch echt grössere Zeile α'_0 . Da aber ein Tableau vollständig von der gerade betrachteten Basis bestimmt ist, bedeutet dies:

- Alle auftretenden Basen sind verschieden.

Nun gibt es aber nur endlich viele verschiedene mögliche Basen. Also führt man auch nur endlich viele Iterationen aus.

◇

Der Algorithmus in der Praxis. Um den Rechenaufwand zu reduzieren, wählt man das Pivotelement $\bar{a}_{kj} > 0$ oft nach der vereinfachten Regel

$$(47) \quad \frac{\bar{b}_k}{\bar{a}_{kj}} = \min\left\{\frac{\bar{b}_i}{\bar{a}_{ij}} \mid \bar{a}_{ij} > 0\right\}.$$

ABER VORSICHT: Bei dieser vereinfachten Regel *kann* es gelegentlich vorkommen, dass eine Basis wiederholt auftritt, der Algorithmus also „zykelt“. (In diesem Fall müsste man dann auch die lexikographische Regel umschalten, um den Algorithmus zu einem Ende zu führen.)

BEMERKUNG. Die lexikographische ist nicht die einzige Pivotregel, die Zykeln unterbindet.

4.4. Die 2-Phasen-Methode. Wie kommt man überhaupt zu einer lexikographisch positiven Basis B_0 , um den Simplexalgorithmus nach der primalen Pivotregel ablaufen zu lassen?

Oft ergibt sich eine solche aus der Problemstellung. Man betrachte z.B. ein LP der Form

$$\min \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}$$

mit nichtnegativem $\mathbf{b} \geq \mathbf{0}$.)Dieses ist noch nicht die vom Simplexalgorithmus erwartete Form!) Mit dem Schlupfvariablenvektor

$$\bar{\mathbf{x}} = \mathbf{b} - A\mathbf{x} \geq \mathbf{0}$$

erhalten wir nun das äquivalente Problem

$$\min \mathbf{c}^T \mathbf{x} + \mathbf{0}^T \bar{\mathbf{x}} \quad \text{s.d.} \quad I\bar{\mathbf{x}} + A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}, \bar{\mathbf{x}} \geq \mathbf{0},$$

das sofort auf das lexikographisch positive Simplextableau

$$\left[\begin{array}{c|cc} 0 & \mathbf{0}^T & \mathbf{c}^T \\ \mathbf{b} & I & A \end{array} \right]$$

mit Basismatrix I führt.

4.4.1. *Allgemeine Methode.* Wir betrachten nun ein LP der Form

$$(48) \quad \min \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}.$$

Dabei darf man oBdA $\mathbf{b} \geq \mathbf{0}$ annehmen. (Im Fall $b_i < 0$ kann man ja einfach die Zeile i mit (-1) multiplizieren, ohne den Lösungsbereich zu verändern.) Mit dem Vektor von *Hilfsvariablen*

$$\bar{\mathbf{x}} = \mathbf{b} - A\mathbf{x}$$

betrachten wir das *Hilfs-LP*

$$(49) \quad \min \mathbf{1}^T \bar{\mathbf{x}} \quad \text{s.d.} \quad I\bar{\mathbf{x}} + A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}, \bar{\mathbf{x}} \geq \mathbf{0}.$$

Bei diesem Hilfs-LP führt die Basis I auf ein lexikographisch positives Simplextableau. (49) kann also mit dem Simplexverfahren in endlicher Zeit gelöst werden. Offenbar gilt

LEMMA 3.9. *Das LP (48) besitzt eine zulässige Lösung genau dann, wenn das Hilfs-LP (49) den Optimalwert 0 aufweist.*

◇

Bei einer Optimallösung von (48) mit Zielfunktionswert 0 müssen alle Hilfsvariablen \bar{x}_i den Wert 0 annehmen. Die aus dem Simplexverfahren für (48) gewonnene optimale Basis B^* erlaubt dann die Konstruktion einer geeigneten Startbasis B_0 für das Ausgangsproblem (48) in folgender Weise:

- Entferne alle Spalten bzgl. B^* , die zu Hilfsvariablen gehören.
- Ergänze die restlichen Spalten in beliebiger Weise zu einer Spaltenbasis A_{B_0} von A .
- B_0 führt auf ein lexikographisch positives Simplextableau.

4.4.2. *Methode.* Wenn keine geeignete Startbasis sofort ersichtlich ist, durchlaufe man die *Phase I*:

(I) Ausgehend von der Form (48) löse man das zugeordnete Hilfs-LP (49).

Hat das Hilfs-LP keine keine Optimallösung mit Zielfunktionswert 0, dann STOPP. Andernfalls starte man nun *Phase II*:

(II) Konstruiere gemäss Lemma 3.9 eine lexikographisch positive Basis B_0 und löse das tatsächliche lineare Programm (48) mit der Simplexmethode.

Die 2-Phasen-Methode (+ lexikographische Pivotregel) zeigt:

- Das Simplexverfahren kann grundsätzlich so implementiert werden, dass es nach einer endlichen Anzahl von Pivotiterationen zu einem Ende gekommen ist.

4.5. Die duale Strategie. Wir nehmen nun an, die momentane Basis B im Simplexverfahren ist *dual zulässig*, d.h. die reduzierten Kosten \bar{c}_ℓ sind alle nicht-negativ (bzw. der Vektor \mathbf{y} ist eine Ecke von P^*).

Ist Optimalität noch nicht gegeben, existiert ein $\bar{b}_k < 0$. Wir suchen nun ein Pivotelement \bar{a}_{kj} (in der k -Zeile der gegenwärtigen Simplextableaus) derart, dass

$$B' = (B \setminus \{k\}) \cup \{\ell\}$$

wieder eine dual zulässige Basis ist. Wir wählen die Spalte j (und folglich das Pivotelement $a_{kj} < 0$) nach der Pivotregel

$$(50) \quad \frac{\bar{c}_j}{\bar{a}_{kj}} = \max_{\ell} \left\{ \frac{\bar{c}_\ell}{\bar{a}_{k\ell}} \mid \bar{a}_{k\ell} < 0 \right\}$$

LEMMA 3.10. Gilt $\bar{a}_{k\ell} \geq 0$ für $\ell = 1, \dots, n$, dann besitzt Problem (43) keine zulässige Lösung. Ist der Index j gemäss der Pivotregel (50) bestimmt, dann ist $B' = (B \cup \{j\}) \setminus \{k\}$ dual zulässig.

Beweis. Eine zulässige Lösung x_1, \dots, x_n ist nichtnegativ, also hätte man im Fall $\bar{a}_{k\ell} \geq 0$:

$$0 > \bar{b}_k = \sum_{\ell=1}^n \bar{a}_{k\ell} x_\ell \geq 0,$$

was nicht sein kann. Sei nun \bar{a}_{kj} gemäss (50) gewählt. Datieren wir das Simplextableau wieder so auf, dass die j te Spalte zum j ten Einheitsvektor wird, so ergibt sich die erste Zeile des aufdatierten Tableaus so:

- Subtrahiere \bar{c}_j mal die neue Zeile k vom reduzierten Kostenvektor $\bar{\mathbf{c}}^T \geq \mathbf{0}^T$.

Der neue Koeffizient in Spalte ℓ ist also

$$c'_\ell = \bar{c}_\ell - \bar{c}_j \bar{a}_{k\ell} / \bar{a}_{kj} \geq 0.$$

◇

BEMERKUNG. Um Endlichkeit des dualen Verfahrens zu garantieren, kann man wieder die Pivotregel durch eine lexikographische Ordnung verschärfen. Das geht völlig analog zur primalen lexikographischen Regel und wird deshalb hier nicht im Detail ausgeführt.

4.6. Die revidierte Simplexmethode. Es ist nicht notwendig, immer das gesamte Simplextableau zu berechnen. Zu der Basis $B \subseteq \{1, \dots, n\}$ kann man die für eine Iteration notwendige Information direkt aus den Ausgangsparametern gewinnen:

- Berechne $\bar{\mathbf{b}} = A_B^{-1} \mathbf{b}$ als Lösung der Gleichung $A_B \mathbf{x}' = \mathbf{b}$.
- Berechne $\mathbf{y}^T = \mathbf{c}_B^T A_B^{-1}$ als Lösung der Gleichung $A_B^T \mathbf{y}' = \mathbf{c}_B$.

Dann erhalten wir für die reduzierten Kosten

$$\bar{c}_j < 0 \iff c_j - \mathbf{y}^T A_j < 0 \quad \text{d.h.} \quad c_j < \sum_{i=1}^m y_i a_{ij}.$$

Nun kann z.B. bei der primalen Strategie sofort ein Pivotelement \bar{a}_{kj} nach der Regel (47) in der Spalte $\bar{A}_j = A_B^{-1} A_j$ gewonnen und die Basisaufdatierung

$$B \rightarrow B' = (B \cup \{j\}) \setminus \{k\}$$

durchgeführt werden. Diese frugale Version der Simplexmethode ist als *revidierte Simplexmethode* bekannt.

ANWENDUNGSBEISPIEL:

Das Schnittmusterproblem. Es seien Stoffballen der Länge ℓ gegeben. Davon sollen b_i Schnittstücke der Länge ℓ_i ($i = 1, \dots, m$) gewonnen werden, sodass insgesamt möglichst wenig Ballen angeschnitten werden.

Zur Modellierung des Problems definieren wir als *Schnittmuster* einen Vektor

$$\begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} \quad \text{mit} \quad a_i \in \mathbb{N} \quad \text{und} \quad \sum_{i=1}^m a_i \ell_i \leq \ell.$$

Sei A die Matrix, deren Spalten sämtliche möglichen Schnittmuster sind. Mit der Notation $\mathbf{1}^T = [1, 1, \dots, 1]$ ergibt sich das Schnittmusterproblem dann in Matrixform:

$$\min \mathbf{1}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0} \quad \text{und} \quad \text{ganzzahlig.}$$

Wegen der Ganzzahligkeitsrestriktion ist dieses Problem *kein* lineares Programm. Wir betrachten deshalb statt dessen die zugeordnete *LP-Relaxierung*

$$(51) \quad \min \mathbf{1}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0},$$

die zumindest eine Untergrenze für den gesuchten Optimalwert liefert. Die Lösung hat typischerweise Komponenten, die nicht ganzzahlig sind. Durch Runden erhält man daraus meist eine recht gute praktische Lösung.

Da A sehr viele Spalten haben kann, ist ein explizites Auflisten nicht erstrebenswert. Wir lösen (51) darum mit der revidierten Simplexmethode.

Haben wir schon eine Basis B von Schnittmustern gefunden und den entsprechenden Vektor \mathbf{y} berechnet, so ergibt die Suche nach einem Schnittmuster mit negativen reduzierten Kosten das Problem, $a_1, \dots, a_m \in \mathbb{N}$ zu ermitteln mit der Eigenschaft

$$(52) \quad 1 < \sum_{i=1}^m y_i a_i \quad \text{und} \quad \sum_{i=1}^m a_i \ell_i \leq \ell.$$

BEMERKUNG. Problem (52) ist (wegen der Ganzzahligkeitsbedingung) ein sog. *NP-schweres Problem*, also theoretisch schwierig. (In der Literatur ist es als *Knapsack-Problem* bekannt.) In der Praxis lässt sich dieses Problem aber sehr gut lösen.

4.7. Sensitivitätsanalyse. Sei B eine primal zulässige Basis für das lineare System

$$A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}.$$

FRAGE: Für welche Zielfunktionsparameter $\mathbf{c} \in \mathbb{R}^n$ ist die zu B gehörige Basislösung $\bar{\mathbf{x}}$ optimal?

Wir wissen: B ist optimal, wenn die reduzierten Kosten nichtnegativ sind:

$$\bar{\mathbf{c}}_N^T = \mathbf{c}_N^T - \mathbf{c}_B^T A_B^{-1} A_N \geq \mathbf{0}^T \quad \text{bzw.} \quad (A_B^{-1} A_N)^T \mathbf{c}_B - I_N \mathbf{c}_N \leq \mathbf{0}_N.$$

Die dieser Ungleichung genügenden \mathbf{c} sind genau die Elemente des polyedrischen Kegels

$$P([(A_B^{-1} A_N)^T, -I_N], \mathbf{0}).$$

Man kann daraus ablesen, welche Veränderungen der Zielfunktionsparameter zulässig sind, wenn man weiterhin die Optimalität einer gefundenen Lösung garantieren will.

Ist B dual zulässig bzgl. den festen Zielfunktionsparametern \mathbf{c} , so stellt sich dual die

FRAGE: Für welche Restriktionsparameter $\mathbf{b} \in \mathbb{R}^m$ ist die zu B gehörige Basislösung $\bar{\mathbf{x}}$ optimal?

Wir fragen also, wann $\bar{\mathbf{b}} = A_B^{-1}\mathbf{b} \geq \mathbf{0}$ gilt. Wieder erhalten wir die Elemente eines polyedrischen Kegels:

$$P(-A_B^{-1}, \mathbf{0}).$$

4.8. Die primal-duale Methode. Wir betrachten ein Paar dualer linearer Programme:

$$\begin{array}{ll} \min & \mathbf{c}^T \mathbf{x} \\ & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{array} \quad \longleftrightarrow \quad \begin{array}{ll} \max & \mathbf{b}^T \mathbf{y} \\ & A^T \mathbf{y} \leq \mathbf{c} \end{array}$$

Wir möchten nun das Konzept des komplementären Schlupfes ausnutzen, um zu einer Optimallösung zu kommen.

ANNAHME: Wir haben schon (irgend)eine dual zulässige Lösung \mathbf{y} zur Verfügung und kennen einen Vektor $\mathbf{x} = [x_1, \dots, x_n]^T \geq \mathbf{0}$ mit der Eigenschaft

$$x_j > 0 \implies \mathbf{y}^T A_j = \sum_{i=1}^m y_i a_{ij} = c_j \quad (j = 1, \dots, n).$$

(Z.B. hat $\mathbf{x} = \mathbf{0}$ trivialerweise diese Eigenschaft.) Dann wissen wir vom komplementären Schlupf: Im Fall $A\mathbf{x} = \mathbf{b}$ ist \mathbf{x} primal zulässig und folglich optimal.

Sei oBdA $\mathbf{b} \geq \mathbf{0}$. Im Fall $\mathbf{b} = \mathbf{0}$ ist $\mathbf{x} = \mathbf{0}$ optimal. Wir unterstellen deshalb

$$\sum_{i=1}^m b_i > 0.$$

Sei \tilde{A} die Teilmatrix aller Spalten A_j von A mit der Eigenschaft

$$\mathbf{y}^T A_j = \sum_{i=1}^m a_{ij} y_i = c_j.$$

Wir betrachten das zugeordnete linearer Optimierungsproblem

$$(53) \quad \begin{array}{ll} \min & \mathbf{1}^T \mathbf{z} \\ & \tilde{A}\mathbf{u} + I\mathbf{z} = \mathbf{b} \\ & \mathbf{u}, \mathbf{z} \geq \mathbf{0} \end{array}$$

Dieses Problem (53) hat immer I als primal zulässige Anfangsbasis und kann somit mit der primalen Strategie gelöst werden. Sei $(\mathbf{u}^*, \mathbf{z}^*)$ eine Optimallösung.

Im Fall $\zeta^* = \mathbf{1}^T \mathbf{z}^* = 0$, ist $\bar{\mathbf{x}} = (\mathbf{u}^*, \mathbf{z}^*) = (\mathbf{u}^*, \mathbf{0})$ primal zulässig für das Ausgangsproblem. Nach Wahl von \tilde{A} erfüllt $\bar{\mathbf{x}}$ die komplementären Schlupfbedingungen und ist folglich optimal.

Im Fall $\zeta^* > 0$ betrachten wir die dual optimale Lösung \mathbf{w} von (53). Diese erfüllt

$$\zeta^* = \mathbf{b}^T \mathbf{w} \quad \text{und} \quad \tilde{A}^T \mathbf{w} \leq \mathbf{0}, \mathbf{w} \leq \mathbf{1}.$$

Für jede Spalte A_ℓ der Ausgangsmatrix A , die *nicht* zu \tilde{A} gehört, haben wir nach Definition von \tilde{A} :

$$A_\ell^T \mathbf{y} < c_\ell.$$

Also erfüllt der Vektor $\mathbf{y}' = \mathbf{y} + \varepsilon \mathbf{w}$ die duale Zulässigkeitsbedingung

$$A^T \mathbf{y}' \leq \mathbf{c} \quad \text{für ein genügend kleines } \varepsilon > 0,$$

aber liefert einen besseren Zielfunktionswert als \mathbf{y} :

$$\mathbf{b}^t \mathbf{y}' = \mathbf{b}^t \mathbf{y} + \varepsilon \mathbf{b}^t \mathbf{w} = \mathbf{b}^t \mathbf{y} + \varepsilon \zeta^* > \mathbf{b}^t \mathbf{y}.$$

Wir können nun in gleicher Weise mit \mathbf{y}' anstelle von \mathbf{y} verfahren.

FAZIT: Eine Iteration der primal-dualen Methode liefert entweder eine Optimallösung (Fall $\zeta^* = 0$) oder eine dual zulässige Lösung mit echt besserem Zielfunktionswert (Fall $\zeta^* > 0$).

BEMERKUNG. Es bleibt die Frage, wie man zu Anfang der primal-dualen Methode ein dual zulässiges \mathbf{y} erhält. Das hängt von der Matrix A und dem Vektor \mathbf{c} ab. Im Fall $\mathbf{c} \geq \mathbf{0}$ kann man z.B. trivialerweise mit $\mathbf{y} = \mathbf{0}$ starten.

KAPITEL 4

Optimierung auf Netzwerken

Wir untersuchen hier spezielle lineare Programme, die eine zusätzliche kombinatorische (graphentheoretische) Struktur tragen. Nutzt man diese kombinatorische Struktur aus, kann man die Probleme in der Praxis oft schneller lösen.

TERMINOLOGIE: Es hat sich eingebürgert, einen gerichteten Graphen auch als *Netzwerk*¹ zu bezeichnen. Die Begriffsbildungen sind dabei oft entsprechenden Begriffen aus der Welt elektrischer Netze motiviert. Der Anwendungsbereich ist jedoch (sehr!) viel allgemeiner.

1. Flüsse, Potentiale und Spannungen

Sei $G = (V, E)$ ein gerichteter Graph mit Knotenmenge V und Kantenmenge $E \subseteq V \times V$. Wir repräsentieren G durch seine (*Knoten/Kanten*)-Inzidenzmatrix $A = [a_{v,e}] \in \mathbb{R}^{V \times E}$ mit den Koeffizienten

$$a_{v,e} = \begin{cases} +1 & \text{wenn } e = (w, v) \text{ mit } w \neq v \\ -1 & \text{wenn } e = (v, w) \text{ mit } w \neq v \\ 0 & \text{sonst.} \end{cases}$$

1.1. Flüsse. Ein *Fluss* auf G ist eine Kantenbewertung $x : E \rightarrow \mathbb{R}$. Der Fluss x bewirkt im Knoten v den (*Netto*-)Durchfluss

$$\delta_v(x) = \sum_{(w,v)} x_{(w,v)} - \sum_{(v,w)} x_{(v,w)} = \sum_{e \in E} a_{v,e} x_e.$$

In Matrixschreibweise ergibt sich also für den Vektor $\mathbf{b} \in \mathbb{R}^V$ mit den Komponenten $b_v = \delta_v(x)$:

$$\mathbf{b} = A\mathbf{x}.$$

Eine *Zirkulation* ist ein Fluss x , der bei jedem Knoten v den Nettodurchfluss $\delta(x)_v = 0$ bewirkt. Also:

$$\mathbf{x} \in \mathbb{R}^E \text{ Zirkulation} \iff \mathbf{x} \in \ker A.$$

¹früher kam man dazu mit der einen Silbe *Netz* aus

1.2. Potentiale und Spannungen. Ein *Potential* auf G ist eine Knotenbewertung $y : V \rightarrow \mathbb{R}$. Zum Beispiel bewirkt ein Fluss $x : E \rightarrow \mathbb{R}$ das *Fluss-Potential* $y : V \rightarrow \mathbb{R}$ mit

$$y_v = \delta_v(x) \quad (v \in V).$$

Umgekehrt induziert ein beliebiges Potential y einen Fluss $u : E \rightarrow \mathbb{R}$ vermöge der Potentialdifferenz

$$u(v, w) = y(w) - y(v) \quad \text{für alle } (v, w) \in E.$$

In Matrixschreibweise ergibt sich:

$$(54) \quad \mathbf{u}^T = \mathbf{y}^T A$$

Ein Fluss der Form (54) heisst *Spannung* auf G . Also:

- Die Spannungen \mathbf{u} bilden den Zeilenraum der Inzidenzmatrix A .
- Die Zirkulationen \mathbf{z} bilden den Kern der Inzidenzmatrix A .

Da Zeilenraum und Kern einer Matrix orthogonal komplementäre Unterräume bilden, folgt:

LEMMA 4.1. *Jeder Fluss $\mathbf{x} \in \mathbb{R}^E$ ist die eindeutige Summe einer Spannung \mathbf{u} und einer Zirkulation \mathbf{z} auf G : $\mathbf{x} = \mathbf{u} + \mathbf{z}$.*

◇

1.3. Das Flussproblem. Wir nehmen an, dass jeder Kante $(v, w) \in E$ eine nichtnegative *Kapazität* $c(v, w)$ zugeordnet ist. Die Knoten $v \in V$ seien mit b_v gewichtet. Wir suchen ein (bzgl. b) optimales Potential y so, dass die induzierte Spannung u die Kapazitätsschranke c einhält. D.h.

$$\max_y \sum_{v \in V} b_v y_v \quad \text{s.d.} \quad y(w) - y(v) \leq c(v, w) \quad \forall (v, w) \in E.$$

In Matrix und Vektorschreibweise ergibt sich

$$(55) \quad \max_{\mathbf{y}} \mathbf{y}^T \mathbf{b} \quad \text{s.d.} \quad \mathbf{y}^T A \leq \mathbf{c}.$$

Das lineare Programm (55) ist dual zum *Flussproblem*

$$(56) \quad \min \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}.$$

In (56) sucht man einen nichtnegativen Fluss $\mathbf{x} \geq \mathbf{0}$, der die Durchflussbedingung $A\mathbf{x} = \mathbf{b}$ einhält und dabei die „Kostenfunktion“ $f(\mathbf{x}) = \mathbf{c}^T \mathbf{x}$ minimiert.

1.3.1. *Lösungsstrategie nach der primal-dualen Simplexmethode.* Das Flussproblem (56) bzw. (55) kann z.B. primal-dual angegangen werden. Bei einem Iterationsschritt (zu einem gegebenen Potential y) entspricht die Matrix \tilde{A} genau den Kanten $e = (v, w)$ mit der Eigenschaft

$$y(w) - y(v) = c(v, w) \quad \text{bzw.} \quad y(w) = y(v) + c(v, w).$$

Das Hilfs-LP

$$\min \sum_{v \in V} z_v \quad \tilde{A}\mathbf{u} + \mathbf{z} = \mathbf{b}, \quad \mathbf{u}, \mathbf{z} \geq \mathbf{0}$$

will einen nichtnegativen Fluss \mathbf{u} auf dem \tilde{A} entsprechenden Teilgraphen bestimmen, der die vorgegebenen Knotendurchflusswerte b_v möglichst gut erreicht, aber nicht überschreitet. (Diese Idee wird im Algorithmus von Goldberg und Tarjan (siehe Abschnitt 4 in diesem Kapitel) aufgegriffen.)

NOTA BENE: Die Notationssymbole „ \mathbf{u} “ und „ \mathbf{z} “ beziehen sich im Hilfs-LP einfach auf die entsprechenden Variablen und bedeuten hier **nicht** allgemeine Spannungs- oder Zirkulationsvektoren!

1.4. Ganzzahligkeit. Wir beweisen zunächst eine allgemeinere Aussage und betrachten eine Matrix $A = [a_{ij}]$ mit den Eigenschaften

(UM₁) $a_{ij} \in \{-1, 0, +1\}$ für alle i, j .

(UM₂) Jede Spalte von A enthält höchstens einen Koeffizienten „ -1 “ und höchstens einen Koeffizienten „ $+1$ “.

LEMMA 4.2. *Erfüllt die quadratische Matrix A die Bedingungen (NM₁) und (NM₂), dann ist A unimodular, d.h. es gilt*

$$\det A \in \{-1, 0, +1\}.$$

Beweis. Wir argumentieren per Induktion über die Anzahl der Koeffizienten $\neq 0$. Gibt es in jeder Spalte von A genau zwei Koeffizienten $\neq 0$, so sind alle Spaltensummen 0. Folglich hat A nicht vollen Rang, d.h. $\det A = 0$.

Im Fall $\det A \neq 0$ dürfen wir somit annehmen, dass mindestens eine Spalte j genau einen Koeffizienten $a_{ij} \neq 0$ besitzt. Entwicklung der Determinante nach dieser Spalte j ergibt

$$|\det A| = |a_{ij}| \cdot |\det A'| \quad \text{mit} \quad |\det A'| \in \{0, 1\},$$

da wir die Behauptung für die Matrix A' , die aus A durch Streichen der Zeile i und Spalte j hervorgeht, per Induktion als richtig unterstellen.

◇

Als Folgerung erhalten wir:

SATZ 4.1. (a) *Ist im Flussproblem (56) der Durchflussvektor \mathbf{b} ganzzahlig, dann hat das Polyeder*

$$P = \{\mathbf{x} \in \mathbb{R}^E \mid A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$$

ganzzahlige Ecken.

(b) Ist im Potentialproblem (55) der Kapazitätsvektor \mathbf{c} ganzzahlig, dann hat das Polyeder

$$P^* = \{\mathbf{y} \in \mathbb{R}^V \mid \mathbf{y}^T A \leq \mathbf{c}^T\}$$

ganzzahlige Ecken.

Beweis. Wir wählen eine Zeilenbasis \bar{A} der Inzidenzmatrix A . Dann gilt weiterhin

$$P = \{\mathbf{x} \in \mathbb{R}^E \mid \bar{A}\mathbf{x} = \bar{\mathbf{b}}, \mathbf{x} \geq \mathbf{0}\}.$$

Jeder (Spalten-)Basismatrix \bar{A}_B von \bar{A} erfüllt die Bedingungen (UM₁) und (UM₂). Die Komponenten der zulässigen Basislösungen \mathbf{v} haben nach der Cramerschen Regel also die Form

$$v_i = \frac{\det \hat{A}^{(i)}}{\det \bar{A}_B} \in \mathbb{Z}.$$

Daraus folgt die Aussage (a). Die Aussage (b) beweist man ganz genauso. \diamond

KOROLLAR 4.1. Sind die Parametervektoren \mathbf{c} und \mathbf{b} ganzzahlig, so gestatten das Flussproblem (56) und das Potentialproblem (55) optimale Lösungen mit ganzzahligen Komponenten. \diamond

1.4.1. *Das Zuordnungspolytop.* Seien S und T disjunkte Mengen mit $|S| = |T|$. Wir betrachten den Graphen $G = (S \cup T, S \times T)$ mit der Inzidenzmatrix A .

Sei $\mathbf{b} \in \mathbb{R}^{S \cup T}$ der Vektor mit den Komponenten

$$b_v = \begin{cases} -1 & \text{wenn } v \in S \\ +1 & \text{wenn } v \in T. \end{cases}$$

Die nichtnegativen Flüsse mit Durchflussvektor \mathbf{b} sind genau die $\mathbf{x} \in \mathbb{R}^{S \times T}$, die das folgende Ungleichungssystem erfüllen:

$$(57) \quad \begin{aligned} \sum_{s \in S} x_{st} &= 1 && \text{für alle } t \in T \\ \sum_{t \in T} x_{st} &= 1 && \text{für alle } s \in S \\ x_{st} &\geq 0 && \text{für alle } (s, t) \in S \times T. \end{aligned}$$

Die Ecken des Polytops aller dieser Flüsse sind ganzzahlig und folglich $(0, 1)$ -Vektoren. Die $(0, 1)$ -Vektoren, die (57) erfüllen, sind aber genau die Inzidenzvektoren von Zuordnungen (Bijektionen) von S und T ! Also:

- Das lineare System (57) beschreibt genau das von den Zuordnungen erzeugte Polytop.

2. Kürzeste Wege

Seien $s, t \in V$ zwei festgewählte Knoten in G . Ein *gerichteter Weg* W von s nach t ist eine Folge von Kanten der Form

$$W = (s, v_1)(v_1, v_2) \cdots (v_k, t).$$

Wir geben nun einen speziellen Durchflussvektor \mathbf{b} vor mit den Komponenten

$$b_v = \begin{cases} -1 & \text{wenn } v = s \\ +1 & \text{wenn } v = t \\ 0 & \text{sonst.} \end{cases}$$

LEMMA 4.3. *Sei B eine zulässige Basis bzgl. $A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}$ und $\bar{\mathbf{x}}$ die zugehörige Basislösung. Dann enthält die Menge*

$$\text{tr}(\bar{\mathbf{x}}) = \{e \in E \mid \bar{x}_e > 0\}$$

einen gerichteten Weg von s nach t in G .

Beweis. Da der Durchfluss von s negativ ist, existiert eine Kante $(s, v_1) \in \text{tr}(\bar{\mathbf{x}})$. Ist $v_1 = t$, ist der Weg gefunden. Andernfalls folgt aus dem Durchfluss $\delta(\bar{\mathbf{x}})_{v_1} = 0$ die Existenz einer Kante $(v_1, v_2) \in \text{tr}(\bar{\mathbf{x}})$ usw.

Nach endlich vielen Schritten haben wir entweder t erreicht oder einen Knoten v_k , den wir schon einmal erreicht haben. Wir haben also einen Kreis durchlaufen. Letzteres ist aber unmöglich, da die Spalten von A , die zu einem Kreis gehören, linear abhängig sind (Warum?). Eine Basis besteht jedoch aus linear unabhängigen Spalten. \mathcal{P} kann also keinen Kreis enthalten. \diamond

PROPOSITION 4.1. *Sei der Durchflussvektor \mathbf{b} wie oben bzgl. $s, t \in V$ fest gewählt. Dann ist jede Ecke (bzw. zulässige Basislösung) $\bar{\mathbf{x}}$ von*

$$P = \{\mathbf{x} \in \mathbb{R}^E \mid A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$$

ein $(0, 1)$ -Vektor und $\text{tr}(\bar{\mathbf{x}}) = \{e \in E \mid \bar{x}_e = 1\}$ ein gerichteter Weg von s nach t .

Beweis. Sei $B \subseteq E$ die zu $\bar{\mathbf{x}}$ gehörige Basis. Da \mathbf{b} ein Vektor ist, mit genau einer Komponente „ -1 “ und einer Komponente „ $+1$ “ (und 0 sonst), erfüllt auch die Matrix im Zähler des Bruchs

$$\bar{x}_e = \frac{\det \hat{A}^{(e)}}{\det \bar{A}_B}$$

bei der Cramerschen Regel die Bedingungen (UM_1) und (UM_2) . Also schliessen wir auch $|\det \hat{A}^{(e)}| \in \{0, 1\}$ und folglich $\bar{x}_e \in \{0, 1\}$.

Sei $\tilde{\mathbf{x}}$ der $(0, 1)$ -Inzidenzvektor eines gerichteten Weges von s nach t in $\text{tr}(\bar{\mathbf{x}})$. Dann gilt

$$\bar{A}_B(\bar{\mathbf{x}}_B - \tilde{\mathbf{x}}_B) = \bar{A}_B\bar{\mathbf{x}}_B - \bar{A}_B\tilde{\mathbf{x}}_B = \mathbf{b} - \mathbf{b} = \mathbf{0}$$

Wegen $\ker \bar{A}_B = \{\mathbf{0}\}$ schliessen wir deshalb $\tilde{\mathbf{x}}_B = \bar{\mathbf{x}}_B$ und somit $\text{tr}(\bar{\mathbf{x}}) = \text{tr}(\tilde{\mathbf{x}})$.

◇

Sei $\mathbf{d} : E \rightarrow \mathbb{R}_+$ eine *Distanzfunktion* auf dem Graphen $G = (V, E)$. Eine optimale Basislösung des linearen Programms

$$(58) \quad \min \mathbf{d}^T \mathbf{x} = \sum_{e \in E} d_e x_e \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}$$

entspricht einem (bzgl. \mathbf{d}) *kürzesten Weg* von s nach t in G .

2.1. Dijkstras Algorithmus. Kürzeste (s, t) -Wege in $G = (V, E)$ kann man nicht nur über das lineare Programm (58) konstruieren. Der Algorithmus von Dijkstra löst das Problem, indem er vom zu (58) dualen linearen Programm ausgeht:

$$(59) \quad \max_{\mathbf{y}} \mathbf{y}^T \mathbf{b} \quad \text{s.d.} \quad \mathbf{y}^T A \leq \mathbf{d}^T \quad \longleftrightarrow \quad \max_{\mathbf{y}} y_t - y_s \quad \text{s.d.} \quad y_w - y_v \leq d_e \quad \forall (v, w) \in E$$

Man sucht also ein (*Knoten*)-*Potential* $y \in \mathbb{R}^V$, das die Potentialdifferenz $y_t - y_s$ maximiert, wobei die übrigen Potentialdifferenzen wie folgt beschränkt sind:

$$y_w - y_v \leq d_e \quad \text{bzw.} \quad y_w \leq y_v + d_e \quad \text{für alle } e = (v, w) \in E.$$

Der Einfachheit halber setzen wir $d_{vw} = +\infty$, wenn $(v, w) \notin E$. Der *Algorithmus von Dijkstra* baut nun ein optimales Potential sukzessive folgendermassen auf:

- (D0) Setze $y_s \leftarrow 0$, $U \leftarrow \{s\}$ und $y_v \leftarrow d_{sv}$ für alle $v \in V \setminus U$.
- (D1) Solange $U \neq V$ gilt, wähle ein $v \in V \setminus U$ mit minimalem Potential y_v , setze $U \leftarrow U \cup \{v\}$ und datiere auf:

$$y_w \leftarrow \min\{y_w, y_v + d_{vw}\} \quad \text{für alle } w \in V \setminus U.$$

Man sieht leicht per Induktion (über $|U|$):

LEMMA 4.4. *In jedem Stadium des Algorithmus von Dijkstra gilt:*

- (a) Die Potentiale y_u der Knoten $u \in U$ geben genau den minimalen Abstand von s nach u an.
- (b) Für jeden anderen Knoten $v \in V$ ist y_v zumindest eine Obergrenze für den Abstand von s .

◇

Man erkennt:

- Der Dijkstra-Algorithmus berechnet nicht nur den kürzesten Abstand von s zu t sondern von s zu allen Knoten $v \in V$ und löst so insbesondere das duale Problem (59).

Aus der Dijkstra-Lösung y lässt sich leicht ein kürzester Weg von s nach t durch „Zurückrechnen“ gewinnen:

Man beginnt bei t und sucht ein $v_1 \in V \setminus \{t\}$ mit

$$y_t = y_{v_1} + d_{v_1, t}.$$

Nun sucht man ein $v_2 \in V \setminus \{t, v_1\}$ mit

$$y_{v_1} = y_{v_2} + d_{v_2, v_1}$$

usw., bis man bei s gelangt ist.

3. Zirkulationen und das MAX-Flow-MIN-Cut-Theorem

Sei nun $f = (t, s) \in E$ eine festgewählte Kante und \mathbf{x} eine Zirkulation auf $G = (V, E)$. Stellen wir uns vor, dass die Kante f in G blockiert wird. Dann stellen die übrigen Kantenflusswerte x_e ($e \neq f$) einen Fluss auf $G_f = (V, E \setminus \{f\})$ dar, wo in s ein Nettoabfluss und bei t ein Nettozufluss in Höhe von x_f stattfindet. Mit anderen Worten:

- Die Einschränkung von \mathbf{x} auf G_f beschreibt den Transport eines „Gutes“ der Quantität x_f von der *Quelle* s zur *Senke* t entlang den Kanten von G_f , wobei bei keinem Knoten $v \neq s, t$ etwas verloren geht oder hinzugewonnen wird.

Unter der Annahme, dass jede Kante $e \in E \setminus \{f\}$ einer Kapazitätsschranke $c_e \geq 0$ unterliegt, sucht man im *Ford-Fulkerson-Problem* nach einer nichtnegativen Zirkulation \mathbf{x} , die den Transportwert x_f maximiert:

$$(60) \quad \begin{array}{ll} \max & x_f \\ \text{s.d.} & \delta_v(\mathbf{x}) = 0 \quad \text{für alle } v \in V \\ & 0 \leq x_e \leq c_e \quad \text{für alle } e \in E \setminus \{f\} \end{array}$$

Dazu ist das duale lineare Programm wie folgt:

$$(61) \quad \begin{array}{ll} \min & \sum_{e \in E} c_e z_e \\ \text{s.d.} & y_w - y_v + z_e \geq 0 \quad \text{für alle } e = (v, w) \in E \\ & y_s - y_t + z_f \geq 1 \quad \text{für } f = (t, s) \in E \\ & z_e \geq 0. \end{array}$$

Die Idee ist nun, eine schon gefundene nichtnegative Zirkulation \mathbf{x} nach Möglichkeit zu verbessern. Ford und Fulkerson haben vorgeschlagen, dies entlang von geeigneten Pfaden von s nach t zu tun.

Augmentierende Wege. Sei $\mathbf{x} \geq \mathbf{0}$ eine zulässige Lösung von (60). Wir definieren einen Hilfsgraphen $G(\mathbf{x})$ auf V mit Kanten

$$\begin{array}{ll} (v, w) & \text{wenn } e = (v, w) \in E \setminus \{f\} \text{ und } x_e < c_e \text{ („Vorwärtskante“)} \\ (w, v) & \text{wenn } e = (v, w) \in E \setminus \{f\} \text{ und } x_e > 0 \text{ („Rückwärtskante“)} \end{array}$$

Sei $\varepsilon > 0$ so, dass $x_e + \varepsilon \leq c_e$ gilt, wenn e eine Vorwärtskante ist, und $x_e - \varepsilon \geq 0$ auf den Rückwärtskanten e erfüllt ist. Dann ist klar:

- Existiert ein Weg \mathcal{P} von s nach t im Hilfsgraphen $G(\mathbf{x})$, dann kann der Fluss \mathbf{x} um mindestens $\varepsilon > 0$ verbessert werden.

Wir erhöhen nämlich einfach den Flusswert um ε auf den Vorwärtskanten von \mathcal{P} (und auf der Kante $f = (t, s)$) und erniedrigen ihn um ε auf den Rückwärtskanten von \mathcal{P} . Offensichtlich ist der resultierende Fluss nichtnegativ, respektiert die Kapazitätsgrenzen und genügt den Knotendurchflussbedingungen.

Sei S die Menge aller Knoten, die in $G(\mathbf{x})$ von s aus auf einem gerichteten Weg erreicht werden können. Dann wissen wir also: Im Fall $t \in S$ kann \mathbf{x} verbessert werden.

Schnitte. Sei allgemein $S \subseteq V$ eine Knotenmenge mit $s \in S$. Dann bestimmt S einen sog. s -Schnitt

$$[S : V \setminus S] = \{(v, w) \in E \mid v \in S, w \notin S\}$$

der Kapazität

$$\mathbf{c}[S : V \setminus S] = \sum_{e \in [S : V \setminus S]} c_e \geq 0.$$

LEMMA 4.5 (Schnittlemma). *Sei \mathbf{x} eine zulässige Zirkulation auf G und $[S : V \setminus S]$ ein beliebiger s -Schnitt. Dann gilt*

$$x_f \leq \mathbf{c}[S : V \setminus S].$$

Beweis. Wir setzen $y_v = 1$ für alle $v \in S$ und $y_v = 0$ für $v \notin S$. Ausserdem wählen wir $z_e = 1$ für $e \in [S : V \setminus S]$ und $z_e = 0$ sonst.

Dann erhalten wir eine zulässige Lösung des dualen Problems (61) mit Zielfunktionswert

$$\sum_{e \in E} c_e z_e = \mathbf{c}[S : V \setminus S].$$

Die schwache Dualität der linearen Programmierung impliziert damit die behauptete Ungleichung. ◇

Sei wie vorher $\mathbf{x} \geq \mathbf{0}$ eine zulässige Zirkulation und S die Menge aller von s in $G(\mathbf{x})$ erreichbaren Knoten. Im Fall $t \notin S$ ergibt sich nach Definition von $G(\mathbf{x})$ und der Knotenmenge S für eine Kante $e = (v, w) \in E \setminus \{f\}$:

$$x_e = \begin{cases} c_e & \text{wenn } v \in S \text{ und } w \in V \setminus S \\ 0 & \text{wenn } v \in V \setminus S \text{ und } w \in S. \end{cases}$$

Also schliessen wir, dass \mathbf{x} optimal ist. Denn

$$x_f = \sum_{e \in [S : V \setminus S]} x_e = \mathbf{c}[S : V \setminus S].$$

SATZ 4.2 (Ford-Fulkerson). *Eine zulässige Zirkulation \mathbf{x} ist optimal für (60) genau dann, wenn es im Hilfsgraphen $G(\mathbf{x})$ keinen augmentierenden Weg von s nach t gibt.* ◇

Das lineare Programm (60) hat auf jeden Fall $\mathbf{x} = \mathbf{0}$ als zulässige Lösung. Also erhalten wir unter den obigen Voraussetzungen eine kombinatorische (graphentheoretische) Form der LP-Dualität:

KOROLLAR 4.2 (MAX-Flow-MIN-Cut).

$$\max\{x_f \mid \mathbf{x} \text{ ist zulässig für (60)}\} = \min\{c[S : V \setminus V] \mid s \in S \subseteq V\}.$$

◇

3.1. Der Algorithmus von Ford-Fulkerson. Die vorangegangene Analyse des Ford-Fulkerson-Problems (60) legt folgenden Algorithmus nahe:

- (FF0) Beginne mit $\mathbf{x} = \mathbf{0}$ als Startlösung und suche im Hilfsgraphen $G(\mathbf{x})$ einen augmentierenden Weg \mathcal{P} von s nach t .
- (FF1) STOP, wenn \mathcal{P} nicht existiert: \mathbf{x} ist optimal.
- (FF2) Wenn \mathcal{P} existiert, modifiziere \mathbf{x} entlang \mathcal{P} zu einem verbesserten zulässigen Fluss \mathbf{x}' mit $x'_f = x_f + \varepsilon$ und iteriere nun mit \mathbf{x}' anstelle von \mathbf{x} .

Sind die Kapazitäten c_e ganzzahlig, so ist klar, dass der FF-Algorithmus nur ganzzahlige Lösungen \mathbf{x} produziert und in jeder Iteration der Flusswert x_f um ein ganzzahliges $\varepsilon \geq 1$ verbessert wird.

BEMERKUNG. Man kann zeigen, dass der FF-Algorithmus höchstens $|V| \cdot |E|$ Iterationen (Augmentierungen) erfordert, wenn man immer einen augmentierenden Weg \mathcal{P} mit so wenig Kanten wie möglich wählt (was z.B. automatisch der Fall ist, wenn man \mathcal{P} mit dem Dijkstra-Algorithmus berechnet). (Wir beweisen dies hier nicht, da wir später noch einen anderen Netzwerkfluss-Algorithmus analysieren werden.)

3.2. Das bipartite Matching- und Überdeckungsproblem. Wir gehen von endlichen disjunkten Mengen S und T und einer Teilmenge $E \subseteq S \times T$ aus und nennen den Graphen $G = (S \cup T, E)$ *bipartit*. Ein (nicht notwendigerweise perfektes) *Matching* ist eine Teilmenge $M \subseteq E$ von paarweise nichtinzidenten Kanten:

$$(s_1, t_1), (s_2, t_2) \in M \implies s_1 \neq s_2, t_1 \neq t_2.$$

Die Aufgabe, ein maximales Matching zu berechnen erweist sich als ein Spezialfall des FF-Problems. Dazu betrachten wir den Graphen $\bar{G} = (V, \bar{E})$, mit zwei neuen Knoten s_0, t_0 , d.h. $V = (S \cup T \cup \{s_0, t_0\})$, \bar{E} , und Kantenmenge

$$\bar{E} = E \cup \{(s_0, s) \mid s \in S\} \cup \{(t, t_0) \mid t \in T\} \cup \{(t_0, s_0)\}$$

Beschränken wir nun die Kapazität der Kanten vom Typ (s_0, s) und (t, t_0) auf 1 (und „ $+\infty$ “ sonst), so berechnet der FF-Algorithmus einen Vektor $\mathbf{x} \in \{0, 1\}^E$ mit maximalem Flusswert

$$x_{(t_0, s_0)} = \sum_{e \in E} x_e.$$

Folglich ist $M = \{e \in E \mid x_e = 1\}$ ein maximales Matching in G .

Unter einer (*Kanten-*)*Überdeckung* von $G = (S \cup T, E)$ versteht man eine Menge von Knoten(!) $C \subseteq S \cup T$ mit der Eigenschaft

$$(v, w) \in E \implies v \in C \text{ oder } w \in C.$$

C muss mindestens die Mächtigkeit eines beliebigen Matchings M haben, denn jede Kante aus M muss ja durch C abgedeckt sein:

$$|C| \geq |M|.$$

Sei andererseits M ein maximales Matching, das nach dem FF-Algorithmus konstruiert wurde und \overline{C} der Schnitt aller Knoten die von s_0 noch erreichbar sind. Wegen

$$\mathbf{c}[\overline{C} : V \setminus \overline{C}] = |M| < \infty,$$

kann es kein $e \in E$ geben, das von $S \cap \overline{C}$ nach $T \setminus \overline{C}$ verläuft. Also ist

$$C = (S \setminus \overline{C}) \cup (T \cap \overline{C})$$

eine Überdeckung und hat Mächtigkeit

$$|C| = |S \setminus \overline{C}| + |T \cap \overline{C}| = \mathbf{c}[\overline{C} : V \setminus \overline{C}] = |M|.$$

SATZ 4.3 (König). Sei $G = (S \cup T, E)$ bipartit. Dann gilt

$$\max\{|M| \mid M \text{ Matching}\} = \min\{|C| \mid C \text{ Überdeckung}\}$$

◇

BEMERKUNG. Das Matching- und Überdeckungsproblem kann sinnvoll auch im Fall $T = S$ (d.h. $E \subseteq S \times S$) formuliert werden. Während das Matchingproblem (mit etwas mehr Aufwand) noch effizient lösbar bleibt, ist in dieser Allgemeinheit kein effizienter Algorithmus für das analoge Überdeckungsproblem bekannt. Insbesondere gilt der „Satz von König“ in diesem Rahmen nicht mehr.

4. Der Präfluss-Markierungsalgorithmus

Wir betrachten wieder das Ford-Fulkerson-Problem in der Form

$$(62) \quad \begin{array}{ll} \max & x_f \\ \text{s.d.} & \delta_v(\mathbf{x}) = 0 \quad \text{für alle } v \in V \setminus \{s, t\} \\ & 0 \leq x_e \leq c_e \quad \text{für alle } E \setminus \{f\} \end{array}$$

(mit Kapazitäten $c_e \in \mathbb{R}_+ \cup \{+\infty\}$). Dabei nehmen wir oBdA an, dass $E = V \times V$ aus allen möglichen Kanten besteht und $f = (t, s)$ (mit $t \neq s$) eine speziell betrachtete Kante ist. (Eine „eigentlich nicht zur Verfügung stehende“ Kante (v, w) kann ja immer durch die Kapazitätsrestriktion $c_{vw} = 0$ simuliert werden.)

Wir wollen einen weiteren Typ von Algorithmus für dieses Problem diskutieren, der von Goldberg und Tarjan vorgeschlagen wurde.

Dazu bezeichnen wir einen Fluss $\mathbf{x} \in \mathbb{R}^E$ als *Präfluss*, wenn \mathbf{x} in jedem Knoten $v \neq s$ einen nichtnegativen Durchfluss bewirkt:

$$\delta_v(\mathbf{x}) = \sum_{z \neq v} x_{zv} - \sum_{w \neq v} x_{vw} \geq 0 \quad \text{für alle } v \in V \setminus \{s\}.$$

Im Fall $\delta_v(\mathbf{x}) > 0$ heisst der Knoten v *aktiv*. Der Präfluss \mathbf{x} wird *zulässig* genannt, wenn er den Kapazitätsschranken genügt:

$$0 \leq x_e \leq c_e \quad \text{für alle } e \in E \setminus \{f\}.$$

Die Sende-Operation. Sei \mathbf{x} ein zulässiger Präfluss und $G(\mathbf{x})$ der (wie im Ford-Fulkerson-Algorithmus definierte) entsprechende Hilfsgraph. (v, w) ist also eine Kante in $G(\mathbf{x})$ genau dann, wenn

$$x_{vw} < c_{vw} \quad \text{oder} \quad x_{wv} > 0.$$

Somit kann man bzgl. der $G(\mathbf{x})$ -Kante (v, w) zusätzliche Flusseinheiten in Höhe von

$$\tilde{c}_{vw} = c_{vw} - x_{vw} + x_{wv} > 0$$

von v nach w schicken ohne die Kapazitätsrestriktion zu verletzen. Ist v aktiv, so kann man also den Fluss von v nach w um

$$\varepsilon = \min\{\delta_v(\mathbf{x}), \tilde{c}_{vw}\} > 0$$

erhöhen und hat weiterhin einen zulässigen Präfluss. Die Grundidee des Algorithmus ist nun einfach:

- Man führe solange Sende-Operationen durch, bis eine Optimallösung vorliegt.

Um dies systematisch zu tun, benutzt man Knotenmarkierungen.

Zulässige Markierungen. Unter einer *zulässigen Markierung* bzgl. \mathbf{x} versteht man eine Bewertung $d : V \rightarrow \mathbb{Z}_+ \cup \{+\infty\}$ der Knoten derart, dass

- (ZM1) $d(v) \leq d(w) + 1$ für alle $(v, w) \in G(\mathbf{x})$;
- (ZM2) $d(s) = |V|$ und $d(t) = 0$.

SATZ 4.4 (Goldberg-Tarjan). Sei $\mathbf{x} \in \mathbb{R}^E$ ein zulässiger Präfluss mit einer zulässigen Markierung d , dann existiert eine Menge $S \subseteq V$ mit $s \in S$ und $t \notin S$ derart, dass für alle $(v, w) \in [S : V \setminus S]$ gilt:

$$x_{vw} = c_{vw} \quad \text{und} \quad x_{wv} = 0.$$

Beweis. Wegen $|V \setminus \{s, t\}| = |V| - 2$ muss es ein $0 < k < |V|$ geben mit $k \neq d(v)$ für alle $v \in V$. Sei

$$S = \{v \in V \mid d(v) > k\}.$$

Dann haben wir $s \in S$ und $t \notin S$. Ist $v \in S$ und $(v, w) \in G(\mathbf{x})$, dann haben wir

$$d(w) \geq d(v) - 1 \geq (k + 1) - 1 = k \quad \text{d.h.} \quad d(w) > k$$

und folglich $w \in S$. Keine Kante von $G(\mathbf{x})$ führt also aus S heraus, was die Behauptung impliziert. ◇

Aus dem Korollar des Satzes von Ford-Fulkerson schliessen wir somit:

KOROLLAR 4.3. *Jede zulässige Zirkulation \mathbf{x} , die eine zulässige Markierung gestattet, ist eine Optimallösung für (60).*

◇

Sei \mathbf{x} ein zulässiger Präfluss mit zulässiger Markierung d . Existiert bzgl. \mathbf{x} kein aktiver Knoten, dann ist \mathbf{x} eine Optimallösung von (62). Wir wollen nun zeigen, wie man im anderen Fall einen aktiven Knoten wählen und eine Sendeoperation durchführen kann, sodass man hinterher wieder einen zulässigen Präfluss mit zulässiger Markierung erhält.

Sei v ein beliebiger aktiver Knoten. Weil d zulässig ist, gilt $d(v) \leq d(w) + 1$ für jede Kante $(v, w) \in G(\mathbf{x})$. Wir unterscheiden zwei Fälle.

1. Fall: Es gibt eine Kante $(v, w) \in G(\mathbf{x})$ mit $d(v) = d(w) + 1$.

Wenn wir nun eine Sendeoperation entlang (v, w) durchführen, ist d auch für den neuen Präfluss \mathbf{x}' zulässig. Denn die einzige mögliche neue Kante in $G(\mathbf{x}')$ ist (w, v) , wofür ja schon $d(w) \leq d(v) + 1$ gilt.

2. Fall: Für alle Kanten $(v, w) \in G(\mathbf{x})$ gilt $d(v) \leq d(w)$.

Wir modifizieren nun die Markierung d zu d' :

$$d'(v) = \min\{d(w) + 1 \mid (v, w) \in G(\mathbf{x})\}.$$

Wegen $d'(v) > d(v)$ ist d' offenbar auch eine zulässige Markierung. Ausserdem befinden wir uns bzgl. d' wieder im 1. Fall!

Algorithmus. Man beginnt mit einem zulässigen Präfluss \mathbf{x} und einer zulässigen Markierung d . Zum Beispiel folgendermassen:

$$x_e = \begin{cases} c_e & \text{wenn } e \text{ von der Form } e = (s, v) \\ 0 & \text{sonst.} \end{cases}$$

$$d(v) = \begin{cases} |V| & \text{wenn } v = s \\ 0 & \text{sonst.} \end{cases}$$

Nun führt man Sendeoperationen durch, die zu zulässigen Präflüssen mit zulässigen Markierungen führen, bis ein zulässiger Fluss (gemäss Korollar 4.3) erreicht ist.

4.1. Laufzeitanalyse. Wieviele Sendeoperationen führt der Präfluss-Markierungsalgorithmus durch? Zur Analyse setzen wir $n = |V|$. Wir betrachten einen zulässigen Präfluss \mathbf{x} mit zulässiger Markierung d .

LEMMA 4.6. *Ist $v \in V$ aktiv bzgl. \mathbf{x} , dann existiert ein gerichteter Weg \mathcal{P} von v nach s in $G(\mathbf{x})$.*

Beweis. Sei R die Menge aller von v in $G(\mathbf{x})$ erreichbarer Knoten. Im Fall $s \notin R$ hätten wir einen echt positiven Nettozufluss aus $V \setminus R$ nach R :

$$\sum_{w \in R} \delta_w(\mathbf{x}) > 0.$$

Also muss es mindestens eine Kante $(z, w) \in [V \setminus R, R]$ existieren mit $x_{zw} > 0$. Aber dann wäre z von v in $G(\mathbf{x})$ erreichbar, ein Widerspruch!

◇

OBdA können wir annehmen, dass der Weg \mathcal{P} von v nach s in $G(\mathbf{x})$ höchstens $(n - 1)$ Kanten durchläuft. Weil sich entlang einer Kante in \mathcal{P} die d -Markierung um höchstens den Wert 1 ändert, finden wir

$$d(v) < d(s) + n = 2n$$

und schliessen:

LEMMA 4.7. *Im Algorithmus tritt der 2. Fall (Ummarkierung) bei einer Sendeoperation insgesamt weniger als $2n^2$ mal auf.*

Beweis. Der 2. Fall tritt nur bei aktiven Knoten v auf und bewirkt eine Erhöhung der Markierung, die aber den Wert $2n$ nie überschreitet. Also tritt dieser Fall bei jedem der n Knoten weniger als $2n$ mal ein.

◇

Um die Anzahl der Sendeoperation (1. Fall) abzuschätzen, unterscheiden wir zwischen einer *saturierenden* Sendung, d.h.

$$\varepsilon = \tilde{c}_{vw} = c_{vw} - x_{vw} + x_{wv}$$

und einer *nichtsaturierenden* Sendung, d.h.

$$\varepsilon = \delta_v(\mathbf{x}) < \tilde{c}_{vw},$$

nach welcher v inaktiv wird.

LEMMA 4.8. *Der Algorithmus führt weniger als $2n^3$ saturierende Sendungen aus.*

Beweis. Wir betrachten ein festes (v, w) mit $d(w) = d(v) + 1$. Wenn diese Kante nach einer Sendeoperation nach oben saturiert ist, steht sie nur dann wieder für eine v -Sendung zur Verfügung, wenn der Knoten w in der Zwischenzeit ummarkiert wurde. Dann ist die Markierung $d'(w)$ von w aber um mindestens 2 gewachsen. Der Fall kann also (wegen $d'(w) < 2n$) höchstens $(n - 1)$ mal eintreten. Analog argumentiert man bzgl. (w, v) und einer Saturierung nach unten. Also finden pro Knoten v höchstens $2(n - 1)^2$ saturierende Sendungen statt.

◇

LEMMA 4.9. *Der Algorithmus kann so ausgeführt werden, dass weniger als $2n^3$ nichtsaturierende Sendeoperationen entstehen.*

Beweis. Wir führen den Algorithmus so aus, dass immer der aktive Knoten v mit der höchsten Markierung $d(v)$ für die Sendeoperation benutzt wird.

Nach einer nichtsaturierenden Sendung ist v inaktiv. v kann nicht wieder aktiv werden, bevor eine Ummarkierung stattgefunden hat (da alle aktiven Knoten $d(w) \leq d(v)$ erfüllen). Gibt es n nichtsaturierende Sendungen ohne zwischenzeitliche Ummarkierungen, dann sind alle Knoten inaktiv, d.h. ein optimaler Fluss ist gefunden.

Also ist $(n - 1) \times$ (Anzahl der Ummarkierungen) eine Obergrenze für die Anzahl der nichtsättigenden Sendungen.

◇

KAPITEL 5

Ganzzahlige lineare Programme

Wir betrachten nun Optimierungsprobleme vom Typ

$$(63) \quad \min \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}, \mathbf{x} \text{ ganzzahlig,}$$

wobei die Matrix $A \in \mathbb{R}^{m \times n}$ und die Vektoren $\mathbf{c} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$ gegeben seien.

Wir setzen

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

Wir wissen: Ist A eine Inzidenzmatrix eines Netzwerkes und ist \mathbf{b} ganzzahlig, dann sind die Ecken von P ganzzahlig. Die Zusatzbedingung der Ganzzahligkeit in (63) ist also von Basislösungen automatisch erfüllt.

Im allgemeinen ist (63) jedoch *kein* lineares Programm im strengen Sinn, da der Zulässigkeitsbereich

$$\mathcal{F} = \{\mathbf{x} \in P \mid \mathbf{x} \in \mathbb{N}^n\} = P \cap \mathbb{Z}^n$$

eine diskrete Menge und (ausser im Trivialfall) kein Polyeder ist. Ist \mathcal{F} endlich und setzen wir $P_I = \text{conv}\mathcal{F}$, so ist (63) äquivalent zu dem Problem

$$\min \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad \mathbf{x} \in P_I.$$

Wäre eine Matrix A' und ein Vektor \mathbf{b}' mit $P_I = P(A', \mathbf{b}')$ bekannt, so könnte man jedoch das Ausgangsproblem (63) als lineares Programm formulieren:

$$\min \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A'\mathbf{x} \leq \mathbf{b}'.$$

VEREINBARUNG. In diesem Kapitel nehmen wir durchweg an, dass sämtliche Problemparameter $A, \mathbf{c}, \mathbf{b}$ rational sind. OBdA dürfen (und werden) wir bei der Problemanalyse deshalb sogar Ganzzahligkeit annehmen:

$$A \in \mathbb{Z}^{m \times n}, \mathbf{c} \in \mathbb{Z}^n, \mathbf{b} \in \mathbb{Z}^m.$$

1. Unimodularität

Wir verallgemeinern nun Netzwerkinzidenzmatrizen und geben eine Klasse von Restriktionsmatrizen $A \in \mathbb{Z}^{m \times n}$ an, die bei ganzzahligem \mathbf{b} auch die Ganzzahligkeit von Basislösungen garantieren. OBdA nehmen wir $m = \text{rg}A$ an.

Wir nennen A *unimodular*, wenn jede quadratische $(m \times m)$ -Basismatrix A_B von A die Eigenschaft $|\det A_B| = 1$ besitzt.

Ganz genau wie bei Netzwerkinzidenzmatrizen können wir nun schliessen:

PROPOSITION 5.1. Sei $A \in \mathbb{Z}^{m \times n}$ eine unimodulare Matrix vom Rang $m = \text{rg} A$ und $\mathbf{b} \in \mathbb{Z}^m$. Dann gilt für jedes $\mathbf{c} \in \mathbb{R}^n$: Entweder hat das lineare Programm

$$\min \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}$$

keine Optimallösung oder es existiert eine optimale Lösung \mathbf{x}^* mit ganzzahligen Komponenten x_j^* .

Der Beweis folgt aus der schon früher erkannten Tatsache, dass eine Optimallösung eines linearen Programms an einer Ecke des Lösungspolyeders erreicht wird.

◇

Für viele Anwendungen ist es geschickt, den Begriff der Unimodularität zu verschärfen. Wir nennen eine (allgemeine) Matrix A *total unimodular*, wenn jede quadratische Teilmatrix A' von A unimodular ist, d.h. wenn gilt:

$$\det A' \in \{-1, 0, +1\}.$$

Insbesondere gilt $a_{ij} \in \{-1, 0, +1\}$ für alle Koeffizienten der total unimodularen Matrix $A = [a_{ij}]$.

BEISPIEL 5.1. Jede Netzwerkinzidenzmatrix ist total unimodular. Die Matrix $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ ist unimodular aber nicht total unimodular.

Bevor wir Beispiele von total unimodularen Matrizen diskutieren, geben wir einige wichtige Matrixkonstruktionen an.

LEMMA 5.1. Sei $A \in \mathbb{Z}^{m \times n}$ total unimodular und $\mathbf{e} \in \mathbb{Z}^m$ ein Einheitsvektor. Dann gilt:

- (a) Wenn man eine Spalte von A mit 0 oder -1 multipliziert, erhält man wieder eine total unimodulare Matrix.
- (b) A^T total unimodular.
- (c) $\bar{A} = [A, \mathbf{e}]$ total unimodular.

Beweis. (a) folgt aus der Tatsache, dass sich die Skalarmultiplikation einer Spalte einer Matrix in der Skalarmultiplikation der Determinante auswirkt. (b) ergibt sich aus dem Transpositionssatz $\det C = \det C^T$.

Um (c) einzusehen, betrachten wir eine quadratische Untermatrix A' von $[A, \mathbf{e}]$. OBdA dürfen wir annehmen, dass die Spalte \mathbf{e} in A' auftaucht. Wir entwickeln die Determinante nach dieser Spalte \mathbf{e} und finden

$$\det A' = \pm 1 \cdot \det A'',$$

wobei A'' eine quadratische Untermatrix von A ist. Also gilt $\det A' \in \{-1, 0, +1\}$.

◇

PROPOSITION 5.2. Sei $A \in \mathbb{Z}^{m \times n}$ total unimodular, $\mathbf{b} \in \mathbb{Z}^m$ und $\mathbf{l}, \mathbf{u} \in \mathbb{Z}^n$ derart, dass

$$P = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} \leq \mathbf{b}, \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}\} \neq \emptyset.$$

Dann ist P ein Polytop mit ganzzahligen Ecken.

Beweis. P ist die Lösungsmenge des total unimodularen Ungleichungssystems

$$\begin{bmatrix} A \\ I \\ -I \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \\ -\mathbf{l} \end{bmatrix}$$

◇

Intervallmatrizen. Sei $M = \{1, \dots, m\}$. Unter einem *Intervall* versteht man eine Teilmenge $F \subseteq M$ derart, dass Elemente $i, j \in M$ existieren mit der Eigenschaft

$$F = \{k \in M \mid i \leq k \leq j\}.$$

Eine $(0, 1)$ -Matrix A heisst *Intervallmatrix*, wenn die Zeilen von A in einer solchen Reihenfolge angeordnet werden können, dass jede Spalte der $(0, 1)$ -Inzidenzvektor eines Intervalls der Zeilenindices ist.

Es ist klar, dass jede quadratische Untermatrix einer Intervallmatrix selber eine Intervallmatrix ist. Es gilt

LEMMA 5.2. Jede Intervallmatrix A ist total unimodular.

Beweis. OBdA sei $A = [a_{ij}]$ quadratisch und

$$a_{1j} = \begin{cases} 1 & \text{für } j = 1, \dots, k \\ 0 & \text{für } j = k + 1, \dots, n. \end{cases}$$

Ausserdem entspreche die erste Spalte von A dem kleinsten Intervall, das 1 enthält.

Im Fall $k = 1$ ist die erste Zeile ein Einheitsvektor. Entwicklung der Determinante nach der ersten Zeile liefert dann die Behauptung per Induktion wie bei Netzwerkmatrizen.

Im Fall $k \geq 2$ subtrahiert man die erste Spalte von den Spalten $2, \dots, k$. Die resultierende Matrix ist wieder eine Intervallmatrix und die Determinante hat sich nicht geändert. Auf die neue Matrix trifft aber der vorige Fall zu.

◇

BEMERKUNG. $(0, 1)$ -Inzidenzmatrizen von allgemeinen Familien \mathcal{F} von Teilmengen einer endlichen Grundmenge M sind typischerweise *nicht* total unimodular!

2. Schnittebenen

Es seien nun allgemein $A \in \mathbb{Z}^{m \times n}$ und $\mathbf{b} \in \mathbb{Z}^m$ gegeben und

$$P = P(A, \mathbf{b}) = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} \leq \mathbf{b}\}$$

das entsprechende rationale Polyeder. Wir interessieren uns für die Menge

$$P_I = \text{conv}\{\mathbf{x} \in P \mid \mathbf{x} \in \mathbb{Z}^n\}.$$

PROPOSITION 5.3. *Ist P ein rationales Polyeder, dann ist auch P_I ein rationales Polyeder.*

Beweis. Im Fall $P_I = \emptyset$ ist nichts zu beweisen. Sei also $P_I \neq \emptyset$. Nach dem Dekompositionssatz von Weyl-Minkowski existieren endliche Mengen $V, W \subseteq \mathbb{Q}^n$ so, dass

$$P = \text{conv}(V) + \text{cone}(W).$$

Stellen wir uns V und W als Matrizen mit den entsprechenden Vektoren als Spalten vor, so kann ein beliebiges $\mathbf{x} \in P$ in der Form

$$\mathbf{x} = V\mathbf{s} + W\mathbf{t} \quad \text{mit} \quad \mathbf{s}, \mathbf{t} \geq \mathbf{0}, \mathbf{1}^T \mathbf{s} = 1$$

dargestellt werden. Nach geeigneter Skalierung (Multiplikation mit dem Hauptnenner) dürfen wir dabei die rationalen Vektoren $\mathbf{w} \in W$ oBdA als ganzzahlig annehmen. Bezeichnen wir mit $\lfloor \mathbf{t} \rfloor$ den Vektor der ganzzahlig nach unten gerundeten Komponenten von \mathbf{t} und setzen $\bar{\mathbf{t}} = \mathbf{t} - \lfloor \mathbf{t} \rfloor \geq \mathbf{0}$, dann erhalten wir

$$\mathbf{x} = (V\mathbf{s} + W\bar{\mathbf{t}}) + W\lfloor \mathbf{t} \rfloor = \bar{\mathbf{x}} + W\lfloor \mathbf{t} \rfloor$$

mit dem ganzzahligen $\lfloor \mathbf{t} \rfloor$ und $\bar{\mathbf{x}}$ in

$$(64) \quad \bar{P} = \{V\mathbf{s} + W\bar{\mathbf{t}} \mid \mathbf{s} \geq \mathbf{0}, \mathbf{1}^T \mathbf{s} = 1, \mathbf{0} \leq \bar{\mathbf{t}} \leq \mathbf{1}\}.$$

Als Bild des Polytops $Q = \{\bar{\mathbf{t}} \mid \mathbf{0} \leq \bar{\mathbf{t}} \leq \mathbf{1}\}$ unter der linearen Abbildung W ist $W(Q)$ ein Polytop, Folglich ist auch

$$\bar{P} = \text{conv}(V) + W(Q)$$

als Minkowskische Summe von Polytopen ein Polytop. Wegen $W\lfloor \mathbf{t} \rfloor \in \mathbb{Z}^n$ finden wir

$$\mathbf{x} \in \mathbb{Z}^n \iff \bar{\mathbf{x}} \in \mathbb{Z}^n$$

und deshalb

$$P \cap \mathbb{Z}^n = \bar{P} \cap \mathbb{Z}^n + \{W\mathbf{z} \mid \mathbf{z} \geq \mathbf{0} \text{ ganzzahlig}\}.$$

Da \bar{P} ein Polytop (und somit beschränkt) ist, ist $\bar{P} \cap \mathbb{Z}^n$ eine endliche Menge. Wegen

$$(65) \quad P_I = \text{conv}(\bar{P} \cap \mathbb{Z}^n) + \text{cone}(W)$$

erkennen wir P_I somit als Polyeder. ◇

BEMERKUNG. Im Prinzip könnte man aus der Darstellung (65) (z.B. mit Fourier-Motzkin) eine lineare Beschreibung von P_I durch Ungleichungen ableiten. Für das

Optimierungsproblem (63) ist dies jedoch nicht interessant, da ein solches Vorgehen bedeutet, dass man ohnehin zuerst sämtliche ganzzahligen Vektoren in \overline{P} (darunter auch die Optimallösung) auflisten müsste.

2.1. Das Verfahren von Gomory. Um eine lineare Beschreibung von P_I zu erzielen, gehen wir von gültigen Ungleichungen für das Polyeder $P = P(A, \mathbf{b})$ aus. Gemäss dem Lemma von Farkas betrachten wir deshalb ein beliebiges rationales $\mathbf{y} \geq \mathbf{0}$ und $\mathbf{c}^T = \mathbf{y}^T A$. Dann ist $\mathbf{c}^T \mathbf{x} \leq z$ mit $z = \mathbf{y}^T \mathbf{b}$ eine gültige Ungleichung für P . Wir dürfen \mathbf{c} als ganzzahlig annehmen. Der springende Punkt ist dann die Beobachtung

- $\mathbf{c}^T = \mathbf{y}^T A$ ist ganzzahlig und

$$\boxed{\mathbf{c}^T \mathbf{x} \leq z' \quad \text{mit} \quad z' = \lfloor \mathbf{y}^T \mathbf{b} \rfloor \in \mathbb{Z}}$$

eine gültige Ungleichung für P_I , da sie von allen ganzzahligen Vektoren in P erfüllt wird.

Tatsächlich genügt es, sich dabei auf \mathbf{y} mit Komponenten $y_i \in [0, 1]$ zu beschränken. Denn bei einem allgemeinen rationalen $\mathbf{y}' \in \mathbb{Q}_+^n$ und $\mathbf{z} \in \mathbb{Z}_+^n$ mit der Eigenschaft

$$\mathbf{0} \leq \mathbf{y} = \mathbf{y}' - \mathbf{z} \leq \mathbf{1}$$

ist die Ungleichung

$$(\mathbf{z}^T A) \mathbf{x} \leq \mathbf{z}^T \mathbf{b} \in \mathbb{Z}$$

ja ohnehin schon von $A\mathbf{x} \leq \mathbf{b}$ impliziert. Für ganzzahliges $\mathbf{x} \in P(A, \mathbf{b})$ gilt darum

$$(\mathbf{y}')^T A \mathbf{x} \leq \lfloor (\mathbf{y}')^T \mathbf{b} \rfloor \iff \mathbf{y}^T A \mathbf{x} \leq \lfloor \mathbf{y}^T \mathbf{b} \rfloor.$$

Damit erhalten wir das *Gomory-Polyeder*

$$\boxed{P' = \{\mathbf{x} \in P \mid (\mathbf{y}^T A) \mathbf{x} \leq \lfloor \mathbf{y}^T \mathbf{b} \rfloor, \mathbf{y} \in [0, 1]^m, \mathbf{y}^T A \in \mathbb{Z}^n\}}$$

BEMERKUNG. P' ist tatsächlich ein Polyeder, denn es gibt nur endlich viele verschiedene Gomory-Schnitte. Das sieht man so: Die Menge $\{\mathbf{y}^T A \mid \mathbf{0} \leq \mathbf{y} \leq \mathbf{1}\}$ ist eine beschränkte Menge von Zeilenvektoren in \mathbb{R}^n und enthält deshalb nur endlich viele ganzzahlige Vektoren.

Iterieren wir diese Konstruktion, so ergibt sich die *Gomory-Folge*

$$P \supseteq P' \supseteq P'' \supseteq \dots \supseteq P_I.$$

Man bemerke, dass keine der Gomory-Ungleichungen einen ganzzahligen Punkt aus P „abschneidet“. Ausserdem gilt: Sobald bei der Gomory-Folge kein neues Polyeder konstruiert wird, hat man genügend viele Ungleichungen erzeugt, die P_I festlegen.

Ohne Beweis bemerken wir

SATZ 5.1 (Gomory). *Die Gomory-Folge eines rationalen Polyeders P hat endliche Länge und endet mit P_I .*

◇

Der Beweis ist nicht schwer aber etwas aufwendig. Deshalb sei hier darauf verzichtet. Wir beweisen nur:

LEMMA 5.3. *Sei P ein rationales Polytop mit $P = P'$. Dann gilt $P = P_I$.*

Beweis. Sei $P \neq P_I$. Dann besitzt P eine Ecke \mathbf{v} mit (mindestens) einer Komponente $v_j \notin \mathbb{Z}$. Ausserdem existiert ein $\mathbf{c} \in \mathbb{Z}^n$ derart, dass die Funktion $f(\mathbf{x}) = \mathbf{c}^T \mathbf{x}$ über P genau von \mathbf{v} maximiert wird. Seien $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(k)}$ die übrigen Ecken von P und

$$\begin{aligned} \min_{\ell=1, \dots, k} (\mathbf{c}^T \mathbf{v} - \mathbf{c}^T \mathbf{v}^{(\ell)}) &= \varepsilon > 0 \\ \max_{\mathbf{x} \in P} |x_1| + \dots + |x_n| &= M < \infty. \end{aligned}$$

Sei $K \in \mathbb{N}$ so gewählt, dass $K\varepsilon > 2M$ erfüllt ist. Dann maximiert \mathbf{v} auch die Funktion $\bar{f}(\mathbf{x}) = \bar{\mathbf{c}}^T \mathbf{x}$ über P , mit

$$\bar{\mathbf{c}}^T = [Kc_1, \dots, Kc_j + 1, \dots, Kc_n] = K\mathbf{c}^T + \mathbf{e}_j^T.$$

Denn für jede andere Ecke $\mathbf{v}^{(\ell)}$ von P gilt

$$K(\mathbf{c}^T \mathbf{v}^{(\ell)}) + v_j^{(\ell)} < K(\mathbf{c}^T \mathbf{v} - \varepsilon) + M < K\mathbf{c}^T \mathbf{v} - M < K\mathbf{c}^T \mathbf{v} + v_j.$$

Wegen $\bar{\mathbf{c}}^T \mathbf{v} - K\mathbf{c}^T \mathbf{v} = v_j \notin \mathbb{Z}$ ist entweder $\bar{\mathbf{c}}^T \mathbf{v}$ oder $\mathbf{c}^T \mathbf{v}$ keine ganze Zahl. Also ist

$$\bar{\mathbf{c}}^T \mathbf{x} \leq \lfloor \bar{\mathbf{c}}^T \mathbf{v} \rfloor \quad \text{oder} \quad \mathbf{c}^T \mathbf{x} \leq \lfloor \mathbf{c}^T \mathbf{v} \rfloor$$

eine Ungleichung, die zwar für P_I gültig ist aber nicht für P . D.h. $P \neq P'$, was im Widerspruch zu unseren Voraussetzungen stünde!

◇

Der Satz von Gomory führt zu einem endlichen Algorithmus zur Lösung von des ganzzahligen Optimierungsproblems

$$\max \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathbb{Z}^n.$$

Man löst die LP-Relaxierung

$$\max \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}.$$

Ist die gefundene Optimallösung \mathbf{x}^* ganzzahlig, dann ist nichts weiter zu tun. Andernfalls berechnet man das Gomory-Polyeder P' und löst

$$\max_{\mathbf{x} \in P'} \mathbf{c}^T \mathbf{x}$$

usw. bis das ganzzahlige Optimum gefunden ist. In der Praxis ist diese Vorgehensweise typischerweise jedoch hoffnungslos ineffizient.

2.2. Schnittebenenverfahren. Die Idee hinter *Schnittebenenverfahren* zur Lösung ganzzahliger linearer Programme ist wie die des Gomory-Verfahrens: Man löst die LP-Relaxierung des Problems. Ist die gefundene Optimallösung \mathbf{x}^* , so fügt man dem LP eine Ungleichung

$$\mathbf{a}^T \mathbf{x} \leq b$$

hinzu, die für alle $\mathbf{x} \in P \cap \mathbb{Z}^n$ gilt und von \mathbf{x}^* verletzt wird. Die entsprechende Hyperebene

$$H(\mathbf{a}, b) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T \mathbf{x} = b\}$$

heißt *Schnittebene* (bzgl. P und P_I). Unter der Ausnutzung der speziellen kombinatorischen Struktur, die das Optimierungsproblem haben mag, lassen sich in der Praxis oft gezielt Schnittebenen bestimmen, die zu effizienteren Algorithmen führen als das Allzweck-Gomoryverfahren. Ein *Schnittebenen-Verfahren* geht nach folgendem Prinzip zur Lösung des Problems

$$\max \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathbb{Z}^n$$

vor:

(SE0) Löse das relaxierte LP-Problem

$$\max \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}.$$

Ist die gefundene Optimallösung \mathbf{x}^* ganzzahlig, STOP.

(SE1) Bestimme im Fall $\mathbf{x}^* \notin \mathbb{Z}^n$ eine Schnittebenenungleichung $\mathbf{a}^T \mathbf{x} \leq b$ für P_I , die von \mathbf{x}^* verletzt wird (d.h. $\mathbf{a}^T \mathbf{x}^* > b$) und füge diese den Restriktionen hinzu. Löse nun

$$\max \mathbf{c}^T \mathbf{x} \quad \text{s.d.} \quad A\mathbf{x} \leq \mathbf{b}, \mathbf{a}^T \mathbf{x} \leq b.$$

Ist die gefundene Optimallösung $\bar{\mathbf{x}}$ ganzzahlig, STOP.

(SE2) Bestimme im Fall $\bar{\mathbf{x}} \notin \mathbb{Z}^n$ eine Schnittebenenungleichung $\bar{\mathbf{a}}^T \mathbf{x} \leq \bar{b}$ für P_I , die von $\bar{\mathbf{x}}$ verletzt wird (d.h. $\bar{\mathbf{a}}^T \bar{\mathbf{x}} > \bar{b}$) und füge diese den bisherigen Restriktionen hinzu usw.

2.2.1. Quadratische boolesche Optimierung. Als Beispiel betrachten wir zu gegebenen Parametern $q_{ij} \in \mathbb{R}$ das Problem

$$\max \sum_{i=1}^n \sum_{j=1}^n q_{ij} x_i x_j, \quad x_i \in \{0, 1\}.$$

Sei $V = \{1, \dots, n\}$ und E die Menge aller Paarmengen $\{i, j\}$. Zu $e = \{i, j\}$ setzen wir

$$d_i = q_{ii} \quad \text{und} \quad c_e = q_{ij} + q_{ji}.$$

Wegen $x_i^2 = x_i$ und $y_e = x_i x_j \in \{0, 1\}$ erhalten wir eine Formulierung als ganzzahliges LP:

$$(66) \quad \begin{array}{llll} \max & \sum_{i \in V} d_i x_i + \sum_{e \in E} c_e y_e & & \\ \text{s.d.} & y_e - x_i & \leq & 0 \quad e \in E, i \in e \\ & x_i + x_j - y_e & \leq & 1 \quad e = \{i, j\} \\ & x_i, y_e & \leq & 1 \\ & -x_i, -y_e & \leq & 0 \\ & x_i, y_e & & \text{ganzzahlig.} \end{array}$$

BEMERKUNG. Man kann sich dieses Problem vorstellen als die Aufgabe, im vollständigen Graphen K_n mit Knotenmenge V und Kantenmenge E einen vollständigen Untergraphen maximalen Gesamtgewichts zu wählen. Dabei sind die Knoten $i \in V$ mit d_i und die Kanten $e \in E$ mit c_e gewichtet.

Als Schnittebenen für das von den ganzzahligen Lösungen von (66) erzeugte Polytop P_I kommen alle Ungleichungen in Frage, die von den ganzzahligen Lösungsvektoren erfüllt werden. Beispiele sind etwa die *Dreiecksungleichungen*

$$x_i + x_j + x_k - y_e - y_f - y_g \leq 1$$

für jeweils drei Knoten $i, j, k \in V$ und die dazugehörigen Kanten $e, f, g \in E$ des entsprechenden „Dreiecks“ $\{i, j, k\}$.

Dieses Beispiel kann verallgemeinert werden. Dazu setzen für $S \subseteq V$ mit $|S| \geq 2$

$$\mathbf{x}(S) = \sum_{i \in S} x_i \quad \text{und} \quad \mathbf{y}(S) = \sum_{e \in E(S)} y_e,$$

wobei $E(S)$ die Menge aller Paarmengen $e = \{i, j\} \subseteq S$ ist. Zu $\alpha \in \mathbb{N}$ definieren wir die entsprechende *Cliquenungleichung* als

$$\alpha \mathbf{x}(S) - \mathbf{y}(S) \leq \alpha(\alpha + 1)/2.$$

LEMMA 5.4. *Jede zulässige $(0, 1)$ -Lösung (\mathbf{x}, \mathbf{y}) von (66) erfüllt jede Cliquenungleichung.*

Beweis. Sei $C = \{i \in S \mid x_i = 1\}$ und $s = |C| \leq |S|$. Dann gilt $\mathbf{x}(S) = s$ und $\mathbf{y}(S) = s(s - 1)/2$. Also finden wir

$$\begin{aligned} \alpha(\alpha + 1)/2 - \alpha \mathbf{x}(S) - \mathbf{y}(S) &= [\alpha(\alpha + 1) - 2\alpha s + s(s - 1)]/2 \\ &= (\alpha - s)(\alpha - s + 1)/2. \end{aligned}$$

Da α und s ganze Zahlen sind, ist der letzte Ausdruck immer nichtnegativ. \diamond

Es gibt allein schon $2^n - n - 1$ Cliquenungleichungen. Diese genügen noch nicht, um P_I vollständig zu beschreiben. Bei nicht zu grossen booleschen Problemen ($n \sim 40$) kommt man damit aber in der Praxis schon recht weit.