Mathematische Optimierung

Volker John

Sommersemester 2009

Inhaltsverzeichnis

Kapitel 1

Einführung

Die Optimierung oder Programmierung untersucht die Fragestellung: *Gesucht ist die optimale Lösung eines Problems unter irgendwelchen Bedingungen.* Die mathematische Formulierung ist: Gegeben seien Funktionen $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : S \to \mathbb{R}$, $i = 1, \ldots, m$, $S \subset \mathbb{R}^n$, suche

$$f(\mathbf{x}) \to \text{Extremum unter den Bedingungen } g_i(\mathbf{x}) \geq 0,\ i = 1, \ldots, m.$$

Sind alle Funktionen linear, so hat man ein Problem der linearen Optimierung oder linearen Programmierung, siehe Teil I.

Bei Optimierungsproblemen müssen folgende Fragestellungen untersucht werden:

- Wie lauten notwendige und hinreichende Bedingungen für die Existenz von Lösungen?
- Wie kann man Lösungen mit möglichst geringem Aufwand berechnen? Was sind die effizientesten Algorithmen?

In der Einführung werden einige typische Beispiele von Optimierungsproblemen angegeben.

Beispiel 1.1 Rundreiseproblem. Gegeben sind n verschiedene Orte O_i, $i = 1, \ldots, n$. Die Entfernung zwischen den Orten O_i und O_j sei a_{ij}. Anstelle der Entfernung können auch andere Parameter wie Kosten oder Zeit genommen werden. Man nimmt im allgemeinen auch $a_{ij} \neq a_{ji}$ an. Das Rundreiseproblem oder auch Traveling–Salesman–Problem kann nun wie folgt formuliert werden:
Ein Reisender, der in einem Ort startet, möchte alle restlichen Orte genau einmal besuchen und zum Ausgangsort zurückkehren. In welcher Reihenfolge hat er die Orte zu besuchen, damit die Gesamtlänge des Reiseweges minimal wird? □

Beispiel 1.2 Landwirtschaft, Anbauoptimierung. Es stehen 100 ha Ackerland zur Verfügung, die mit Kartoffeln x_1 *ha* und Getreide x_2 *ha* bestellt werden sollen. Ein Teil der Anbaufläche kann auch brach bleiben. Die Betriebskosten sind wie folgt (GE = Geldeinheit):

	Kartoffeln	Getreide	insgesamt verfügbar
Anbaukosten GE/ha	10	20	1100 GE
Arbeitstage/ha	1	4	160 Tage
Reingewinn GE/ha	40	120	

Bei welcher Bewirtschaftung erzielt man den größten Gewinn?

Die mathematische Formulierung des Problems ist wie folgt:

zu maximierende Funktion:	$z = 40x_1 + 120x_2$	$\rightarrow$	max,
zur Verfügung stehendes Geld:	$10x_1 + 20x_2$	$\leq$	1100,
zur Verfügung stehende Zeit:	$x_1 + 4x_2$	$\leq$	160,
zur Verfügung stehende Fläche:	$x_1 + x_2$	$\leq$	100,
keine negativen Anbauflächen:	x_1, x_2	$\geq$	0.

Diese Aufgabe kann man graphisch lösen.

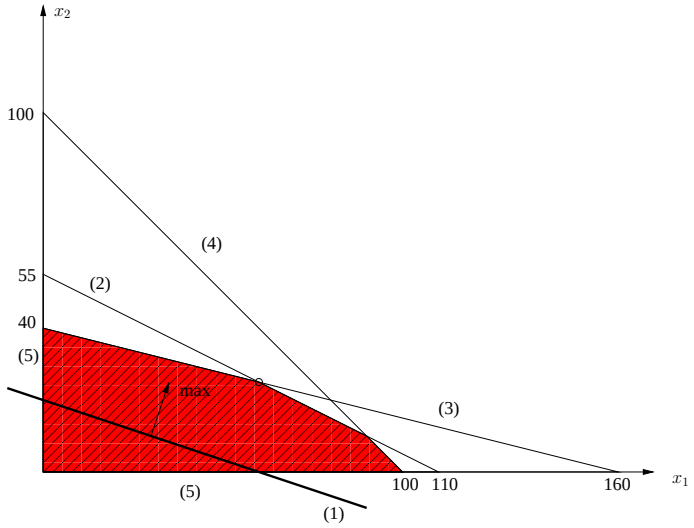


Abbildung 1.1: Darstellung der Nebenbedingungen und der Zielfunktion zum Beispiel 1.2.

Die Nebenbedingungen beschreiben Halbebenen und der Durchschnitt der Halbebenen ist gerade die Menge der Paare (x_1, x_2), in denen man das Maximum sucht. Zur graphischen Darstellung der Zielfunktion z wähle man sich einen beliebigen Punkt (x_1, x_2) und berechne die Gerade z durch diesen Punkt. In diesem Beispiel soll die Zielfunktion maximiert werden, das heißt, der Zielfunktionswert steigt, wenn man die Gerade orthogonal zu ihrem Anstieg nach oben verschiebt. Der letzte Punkt, der alle Nebenbedingungen erfüllt und der auf einer parallelen Geraden zur dargestellten Geraden liegt, ist der mit einem Kreis gekennzeichnete Punkt $(x_1, x_2) = (60, 25)$. Die Lösung dieses linearen Optimierungsproblems ist demzufolge $x_1 = 60$ ha, $x_2 = 25$ ha und der Gewinn ist 5400 GE. □

Beispiel 1.3 Ernährungsprogramm. Es stehen die folgenden drei Nahrungsmittel zur Verfügung (alle Angaben jeweils für 100 Gramm):

	Eiweiß	Fett	Kohlenhyd.	Wasser	Preis
1. Weißbrot	8	1	49	42	10
2. Wurst	12	20	0	68	80
3. Milch	3	3	5	89	7

Ein Ernährungsprogramm wird nur zugelassen, wenn es folgende Mindestanforderungen erfüllt: Eiweiß: 90 g, Fett: 80 g, Kohlenhydrate: 500 g, Wasser 2500 g. Ziel ist es, das kostengünstigste Ernährungsprogramm zu finden, welches diese Anforde-

rungen erfüllt. Das zugehörige Optimierungsproblem lautet:

$$\begin{aligned} z = 10x_1 + 80x_2 + 7x_3 &\rightarrow \min \\ 8x_1 + 12x_2 + 3x_3 &\geq 90 \\ x_1 + 20x_2 + 3x_3 &\geq 80 \\ 49x_1 + 5x_3 &\geq 500 \\ 42x_1 + 68x_2 + 89x_3 &\geq 2500 \\ x_1, x_2, x_3 &\geq 0, \end{aligned}$$

wobei die Maßeinheit für x_1, x_2, x_3 hier 100 g ist.

Die (gerundete) Lösung lautet: $x_1 = 7.71$, $x_2 = 0$, $x_3 = 24.45$, also 771 g Weißbrot und 2445 g Milch, das heißt vegetarisch. Die Kosten sind rund 248 GE. □

Beispiel 1.4 Rucksackproblem. Ein Wanderer kann in seinem Rucksack ein Gesamtgewicht von N tragen. Er hat n Gegenstände, die er mitnehmen möchte und jeder Gegenstand hat einen gewissen Nutzen n_i, $i = 1, \ldots, n$. Das Gesamtgewicht aller Gegenstände übersteigt das zulässige Maximalgewicht. Das Optimierungsproblem des Wanderers besteht nun darin, eine Teilmenge von Gegenständen mit maximalem Nutzen zu finden, so dass das Gesamtgewicht dieser Teilmenge höchstens N ist. Dabei kann als zusätzliche Nebenbedingung auftreten, dass gewisse Lösungskomponenten ganzzahlig sein müssen, zum Beispiel die Anzahl der Paar Schuhe, die er mitnehmen soll. □

Beispiel 1.5 Zuordnungsproblem. In einer Firma stehen zur Fertigung von n Produkten n Maschinen zur Verfügung. Jede Maschine eignet sich zur Herstellung jedes Produktes unterschiedlich gut. Es ergeben sich je nach Zuordnung verschiedene Arbeitszeiten. Jeder Maschine soll genau ein Produkt zugeordnet werden. Das Optimierungsproblem besteht darin, die Gesamtfertigungszeit der Produkte zu minimieren. □

Bemerkung 1.6 *Operations Research.* In der Fachliteratur werden Optimierungsaufgaben oft unter dem Begriff Operations Research (Optimalplanung) geführt. □

Literaturempfehlungen sind:

- Jarre und Stoer [JS04],
- Borgwardt [Bor01],
- Elster, Reinhardt, Schäuble, Donath [ERSD77],
- vor allem über das Gebiet der linearen Optimierung gibt es auch eine Reihe älterer Lehrbücher, die man verwenden kann.

Teil I

Lineare Optimierung

Kapitel 1

Grundlagen

Definition 1.1 Lineares Optimierungsproblem, lineares Programm. Eine Aufgabenstellung wird lineares Optimierungsproblem oder lineares Programm genannt, wenn das Extremum einer linearen Funktion

$$z = \sum_{i=1}^{n} c_i x_i = \mathbf{c}^T \mathbf{x} \tag{1.1}$$

zu bestimmen ist, über der durch das lineare Ungleichungssystem

$$\sum_{j=1}^{n} a_{ij} x_j \quad \leq (>) \quad b_i, \quad i = 1, \ldots, m \tag{1.2}$$

$$x_j \quad \geq \quad 0, \quad j = 1, \ldots, n, \tag{1.3}$$

definierten Punktmenge. □

Seien $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Dann wird die Bezeichnung

$$\mathbf{x} \geq (>) \mathbf{y} \quad \Longleftrightarrow \quad x_i \geq (>) y_i, \ \forall\ i = 1, \ldots, n,$$

verwendet.

Definition 1.2 Zulässiger Bereich. Die Menge $\mathcal{M}$ aller Punkte, die das Ungleichungssystem (1.2) – (1.3) erfüllen, heißt zulässiger Bereich. □

Beispiel 1.3 Der zulässige Bereich, der durch lineare Nebenbedingungen beschrieben ist, ist der Durchschnitt von Halbräumen. Für $n = 2$ sind das Halbebenen und ein Beispiel ist in Abbildung 1.1 zu sehen. □

Der zulässige Bereich ist nicht notwendig beschränkt. Er kann auch leer sein.

Definition 1.4 Konvexität. Eine Punktmenge $\mathcal{M}$ heißt konvex, wenn mit beliebigem $x_1, x_2 \in \mathcal{M}$ auch alle Punkte der Gestalt

$$\lambda x_1 + (1 - \lambda) x_2, \ 0 \leq \lambda \leq 1,$$

zu $\mathcal{M}$ gehören. □

Für den $\mathbb{R}^n$ bedeutet Konvexität, dass mit zwei Punkten $\mathbf{x}_1, \mathbf{x}_2$ aus $\mathcal{M}$ auch ihre Verbindungsstrecke in $\mathcal{M}$ liegt.

Satz 1.5 *Die durch das lineare Ungleichungssystem (1.2) – (1.3) definierte Punktmenge ist konvex.*

Beweis: Seien $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{M}$ gegeben. Dann gelten

$$A\mathbf{x}_1 \leq \mathbf{b},\ \mathbf{x}_1 \geq \mathbf{0},\ A\mathbf{x}_2 \leq \mathbf{b},\ \mathbf{x}_2 \geq \mathbf{0}.$$

Mit $\lambda \in [0,1]$ gelten

$$\lambda A\mathbf{x}_1 \leq \lambda\mathbf{b}, \quad (1-\lambda)A\mathbf{x}_2 \leq (1-\lambda)\mathbf{b}.$$

Addition und Linearität der Matrizenmultiplikation ergibt

$$A\left(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2\right) \leq \mathbf{b}.$$

Analog folgt

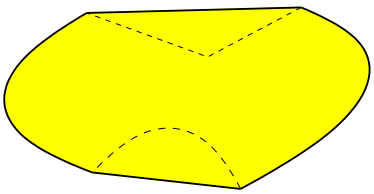
$$\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \geq \mathbf{0}.$$

■

Der Durchschnitt beliebig vieler konvexer Mengen ist wieder konvex. *Übungsaufgabe*

Definition 1.6 Konvexe Hülle. Die kleinstmögliche konvexe Menge $\overline{\mathcal{M}}$, die eine vorgegebene Punktmenge $\mathcal{P}$ enthält, heißt deren konvexe Hülle. □

Beispiel 1.7 Die dick umrandete Menge ist die konvexe Hülle der gestrichelt umrandeten Menge.



□

Beispiel 1.8 Die Menge

$$\mathcal{M} = \left\{ \frac{1}{n} \ : \ n \in \mathbb{N} \right\}$$

ist nicht konvex, da sie aus diskreten Punkten besteht. Ihre konvexe Hülle ist $(0,1]$. □

Definition 1.9 Konvexe Linearkombination. Gegeben seien q Punkte $\mathbf{x}_1, \ldots, \mathbf{x}_q$. Betrachtet werden alle Punkte der Gestalt

$$\mathbf{x} = \sum_{i=1}^{q} \lambda_i \mathbf{x}_i, \quad 0 \leq \lambda_i \leq 1, \quad \sum_{i=1}^{q} \lambda_i = 1. \tag{1.4}$$

Dann heißt die mit (1.4) erklärte Menge konvexe Linearkombination der Punkte $\mathbf{x}_1, \ldots, \mathbf{x}_q$. □

Welche Punkte einer konvexen Menge sollen ausgezeichnet werden?

Definition 1.10 Eckpunkt oder Extrempunkt einer konvexen Menge. Gegeben sei eine konvexe Menge $\mathcal{M}$. Der Punkt $\mathbf{x} \in \mathcal{M}$ heißt Eckpunkt oder Extrempunkt von $\mathcal{M}$, wenn aus $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$, $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{M}$, $0 < \lambda < 1$, folgt $\mathbf{x} = \mathbf{x}_1 = \mathbf{x}_2$. □

Man sagt, $\mathbf{x}$ lässt sich nicht als *echte* konvexe Linearkombination von Punkten $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{M}$ darstellen.

Beispiel 1.11 Bei einem Viereck im $\mathbb{R}^2$ sind die Eckpunkte gerade die vier Ecken. Bei einer Kugel im $\mathbb{R}^n$, $n \geq 1$, sind alle Randpunkte Eckpunkte. □

Definition 1.12 Konvexes Polyeder. Eine beschränkte konvexe Menge $\mathcal{M} \neq \emptyset$ mit endlich vielen Eckpunkten heißt konvexes Polyeder. □

Beispiel 1.13 Konvexe Polyeder in $\mathbb{R}^n, n = 1, 2, 3$. Ein konvexes Polyeder in $\mathbb{R}^1$ ist ein abgeschlossenes Intervall. In $\mathbb{R}^2$ und $\mathbb{R}^3$ kann man sich konvexe Polyeder noch gut vorstellen. Ein Beispiel in $\mathbb{R}^2$ findet man in Abbildung 1.1. □

Satz 1.14 *Sei $\mathcal{M}$ eine konvexe, abgeschlossene und beschränkte Menge, $\mathcal{P}$ sei die Menge der Eckpunkte von $\mathcal{M}$. Dann ist $\mathcal{M}$ die konvexe Hülle von $\mathcal{P}$.*

Beweis: Literatur. Beweisidee mit trennenden Hyperebenen siehe [ERSD77, Satz 2.48]. ■

Satz 1.15 *Ist der Lösungsbereich*

$$\mathcal{M} = \{\mathbf{x} \ : \ A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$$

beschränkt, so ist er ein konvexes Polyeder.

Beweis: Siehe später, Folgerung 3.7. ■

Kapitel 2

Geometrische Deutung des Linearen Programms

In diesem Abschnitt wird gezeigt, dass sich das anschauliche Merkmal des zweidimensionalen linearen Programms aus Beispiel 1.2, nämlich dass das Optimum auf dem Rand angenommen wird, verallgemeinern lässt.

Historie zur Untersuchung linearer Optimierungsprobleme:

- 1939 Leonid V. Kantorovitch (1912 – 1986); Methode der Auflösungskoeffizienten
- 1941 Frank L. Hitchcock, Transportproblem
- 1949 George Dantzig (1914 – 2005), Simplexmethode
- 1984 Narendra Karmarkar (geb. 1957), Innere–Punkt–Methoden für lineare Programme

Definition 2.1 Lineares Optimierungsproblem in 1. Normalform, lineares Programm in Normalform. Gesucht werden die Werte der n Variablen $x_1, \ldots, x_n$ so, dass die lineare Funktion

$$z = \sum_{j=1}^{n} c_j x_j = \mathbf{c}^T \mathbf{x} \tag{2.1}$$

die sogenannte Zielfunktion, unter den Nebenbedingungen

$$\sum_{j=1}^{n} a_{ij} x_j \quad \geq \quad b_i, \quad i = 1, \ldots, m, \quad (A\mathbf{x} \geq \mathbf{b}), \tag{2.2}$$

$$x_j \quad \geq \quad 0, \quad j = 1, \ldots, n, \quad (\mathbf{x} \geq \mathbf{0}), \tag{2.3}$$

ein Minimum annimmt. Alle Koeffizienten sind reell. Das System (2.1) – (2.3) heißt lineares Optimierungsproblem oder lineares Programm in 1. Normalform. □

Bemerkung 2.2

1. Ob (2.1) in min– oder max–Form benutzt wird, ist im allgemeinen ohne Belang, in [JS04] wird beispielsweise die max–Form verwendet.
2. Die Relationen $\geq$, $=$, $\leq$ im System der Nebenbedingungen sind im wesentlichen äquivalent.
3. Fehlt zum Beispiel für einen Index k die Bedingung $x_k \geq 0$, so setzt man $x_k := \overline{x}_k - \hat{x}_k$ mit $\overline{x}_k, \hat{x}_k \geq 0$. Man erhöht damit die Anzahl der Variablen um Eins.

□

Definition 2.3 Zulässiger Punkt, zulässiger Bereich. Ein Punkt $\mathbf{x} = (x_1, \ldots, x_n)^T$ heißt zulässig, wenn er die Nebenbedingungen (2.2), (2.3) erfüllt. Die Gesamtheit aller zulässigen Punkte heißt zulässiger Bereich. □

Für die Lösung von (2.1) – (2.3) kommen nur zulässige Punkte in Betracht. Der zulässige Bereich ist konvex. Ist er beschränkt, so ist er ein konvexes Polyeder. Ist der zulässige Bereich nicht beschränkt ist, dann gilt:

- entweder ist (2.1) über diesen Bereich selbst nicht beschränkt,
  Beispiel: Minimiere $-2x_1 - x_2$ im Bereich $\{(x_1, x_2) \ : \ x_1 \geq 0, x_2 \geq 0\}$,
- oder (2.1) ist über dem unbeschränkten Bereich beschränkt. Dann kann man Zusatzbedingungen an den zulässigen Bereich stellen, die das Optimum nicht ändern, so dass der neue zulässige Bereich beschränkt ist.
  Beispiel: Minimiere $2x_1 + x_2$ im Bereich $\{(x_1, x_2) \ : \ x_1 \geq 0, x_2 \geq 0\}$.

Weitere Beispiele findet man in Beispiel 2.6.

Wenn von einem konvexen Polyeder gesprochen wird, ist ab sofort immer ein abgeschlossenes konvexes Polyeder gemeint.

Satz 2.4 ***Extremwertannahme.*** *Eine auf einem konvexen Polyeder definierte lineare Funktion $z = f(\mathbf{x})$ nimmt ihren kleinsten Wert in (mindestens) einem Eckpunkt an.*

Beweis: Seien $\mathbf{x}_1, \ldots \mathbf{x}_p$ die Eckpunkte des konvexen Polyeders $\mathcal{M}$. Die Funktion $f(\mathbf{x})$ nehme ihr Minimum in $\mathbf{x}_0 \in \mathcal{M}$ an, das heißt

$$f(\mathbf{x}_0) \leq f(\mathbf{x}) \tag{2.4}$$

für alle Punkte $\mathbf{x}$ des konvexen Polyeders. Dass das Minimum angenommen wird, folgt nach dem Satz von Bolzano–Weierstraß (stetige Funktion in einem kompakten Gebiet nimmt ihre Extremwerte an). Ist $\mathbf{x}_0$ kein Eckpunkt, so existiert eine Darstellung (Satz 1.14)

$$\mathbf{x}_0 = \sum_{j=1}^{p} \lambda_j \mathbf{x}_j, \ 0 \leq \lambda_j \leq 1, \ \sum_{j=1}^{p} \lambda_j = 1.$$

Aus der Linearität von f folgt

$$f(\mathbf{x}_0) = f\left(\sum_{j=1}^{p} \lambda_j \mathbf{x}_j\right) = \sum_{j=1}^{p} \lambda_j f(\mathbf{x}_j).$$

Sei ein Index l definiert durch

$$f(\mathbf{x}_l) = \min_{j=1,\ldots,p} f(\mathbf{x}_j).$$

Dann folgt

$$f(\mathbf{x}_0) \geq f(\mathbf{x}_l) \sum_{j=1}^{p} \lambda_j = f(\mathbf{x}_l). \tag{2.5}$$

Wegen (2.4) und (2.5) wird das Minimum für $\mathbf{x}_l$ angenommen. ■

Folgerung 2.5 *Wird das Minimum in mehr als einem Eckpunkt des konvexen Polyeders angenommen, so wird es auf der konvexen Hülle dieser Eckpunkte angenommen.*

Beweis: Ohne Beschränkung der Allgemeinheit seien die Eckpunkte so numeriert, dass die Zielfunktion $f(\mathbf{x})$ ihr Minimum in den Eckpunkten $\mathbf{x}_1, \ldots, \mathbf{x}_q$ annehme. Die konvexe Hülle dieser Eckpunkte ist

$$\left\{ \tilde{\mathbf{x}} \ : \ \tilde{\mathbf{x}} = \sum_{i=1}^{q} \lambda_i \mathbf{x}_i, \ 0 \leq \lambda_i \leq 1, \sum_{i=1}^{q} \lambda_i = 1 \right\}.$$

Aus der Linearität von f folgt

$$f(\tilde{\mathbf{x}}) = f\left(\sum_{i=1}^{q} \lambda_i \mathbf{x}_i\right) = \sum_{i=1}^{q} \lambda_i f(\mathbf{x}_i) = f(\mathbf{x}_1) \sum_{i=1}^{q} \lambda_i = f(\mathbf{x}_1),$$

womit die Folgerung bewiesen ist. ■

Geometrische Interpretation

Die Gleichung $z = \mathbf{c}^T\mathbf{x} - d$ mit einer vorgegebenen Konstanten d ist die Gleichung einer Hyperebene in $\mathbb{R}^n$. Für $n = 3$, hat man beispielsweise die Normalform einer Ebenengleichung, wobei $\mathbf{c}$ ein Normalenvektor der Ebene ist.

Sei $z = \mathbf{c}^T\mathbf{x}$ die Zielfunktion. Es ist gerade

$$\mathbf{c} = \nabla z = \left(\frac{\partial z}{\partial x_1}, \ldots, \frac{\partial z}{\partial x_n}\right)^T.$$

Außerdem ist $\mathbf{c}$ orthogonal zu den Hyperebenen $\mathbf{c}^T\mathbf{x} = const.$ Sei nämlich ein beliebiger Vektor einer Hyperebene gegeben, etwa zwischen den Punkten $\tilde{\mathbf{x}}$ und $\hat{\mathbf{x}}$, dann gilt

$$\mathbf{c}^T\tilde{\mathbf{x}} = const,\ \mathbf{c}^T\hat{\mathbf{x}} = const \implies \mathbf{c}^T(\tilde{\mathbf{x}} - \hat{\mathbf{x}}) = 0.$$

Aus der Menge $\{\mathbf{c}^T\mathbf{x} = const\}$ wählen wir diejenige Hyperebene, die einen vorgegebenen zulässigen Punkt $\mathbf{x}_0 \in \mathcal{M}$, nicht notwendig einen Eckpunkt, enthält: $\mathbf{c}^T\mathbf{x} = \mathbf{c}^T\mathbf{x}_0$. Wir definieren

$$g := \{\mathbf{x} \ :\ \mathbf{x} = \mathbf{x}_0 + t\mathbf{c}, t \in \mathbb{R}\}.$$

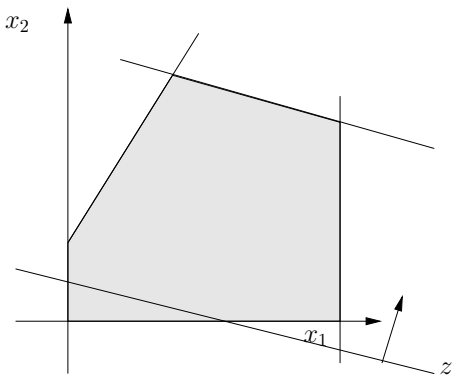
Diese Gerade enthält $\mathbf{x}_0$ und sie ist orthogonal zu $\mathbf{c}^T\mathbf{x} = const.$ Für alle $\mathbf{x} \in g$ gilt bezüglich der Zielfunktion

$$z = \mathbf{c}^T\mathbf{x} = \mathbf{c}^T(\mathbf{x}_0 + t\mathbf{c}) = \mathbf{c}^T\mathbf{x}_0 + t\,\|\mathbf{c}\|_2^2 =: z_0 + t\,\|\mathbf{c}\|_2^2,$$

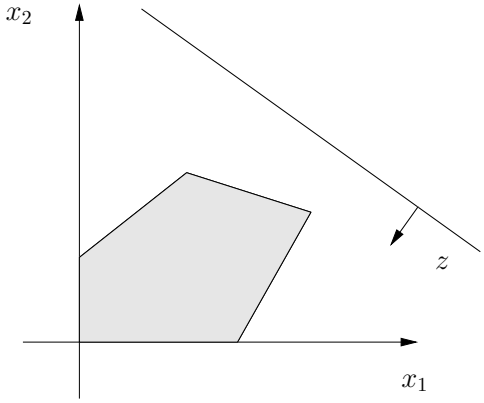
wobei z_0 der Startwert der Zielfunktion ist. Sei $t > 0$. Dann folgt $z > z_0$, das heißt, der Wert der Zielfunktion wächst streng monoton in Richtung von $\mathbf{c}$. Wenn man z zu maximieren hat, so verschiebe man die Hyperebene in Richtung ihres Gradienten. Also, ausgehend von $\mathbf{c}^T(\mathbf{x} - \mathbf{x}_0) = 0$ konstruiere man in Richtung von $\mathbf{c}$ eine Schar zu $\mathbf{c}^T(\mathbf{x} - \mathbf{x}_0) = 0$ paralleler Hyperebenen mit dem Ziel, diejenige Hyperebene aus der Schar zu finden, die $\{\mathbf{x} \ :\ A\mathbf{x} \le \mathbf{b},\ \mathbf{x} \ge \mathbf{0}\}$ berührt mit der Eigenschaft, dass $\{\mathbf{x} \ :\ A\mathbf{x} \le \mathbf{b},\ \mathbf{x} \ge \mathbf{0}\}$ ganz im negativen Halbraum der berührenden Hyperebene liegt. Berührung bedeutet, dass $\{\mathbf{x} \ :\ A\mathbf{x} \le \mathbf{b},\ \mathbf{x} \ge \mathbf{0}\} \cap \{\mathbf{c}^T\mathbf{x} = const\}$ eine Teilmenge des Randes des Polyeders ist, zum Beispiel ein Eckpunkt.

Beispiel 2.6 Beispiele für Situationen die in linearen Programmen auftreten können.

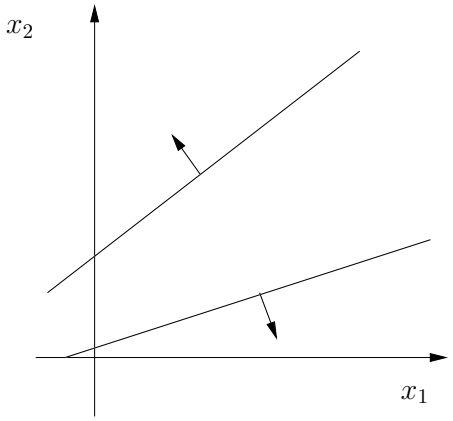
a) Es gibt unendlich viele Lösungen (eine gesamte Kante).



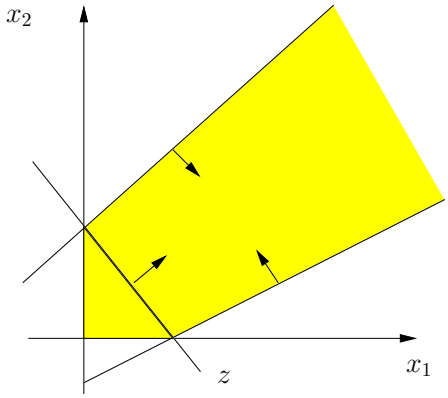
b) Es gibt überflüssig Nebenbedingungen. Die Zielfunktion nimmt ihren Extremwert in $(0, 0)$ an und die drei oberen Nebenbedingungen sind überflüssig.



c) Der zulässige Bereich ist leer.



d) Der Optimalwert ist nicht beschränkt.



Kapitel 3

Basislösungen eines linearen Programms in Normalform

Definition 3.1 Lineares Programm in 2. Normalform, einfache Normalform. Gegeben sei das lineare Programm

$$z = \mathbf{c}^T \mathbf{x} \rightarrow \min ! \tag{3.1}$$

unter den folgenden Bedingungen

$$A\mathbf{x} = \mathbf{b} \tag{3.2}$$

$$\mathbf{x} \geq \mathbf{0} \tag{3.3}$$

mit $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$ und $A \in \mathbb{R}^{m \times n}$. Dieses Problem wird lineares Programm in (2.) Normalform genannt. □

Bemerkung 3.2 Wenn man die lineare Ungleichung

$$\sum_{j=1}^{n'} a_{i^* j} x_j \leq b_{i^*}$$

gegeben hat, so kann man eine sogenannte Schlupfvariable einführen

$$\sum_{j=1}^{n'} a_{i^* j} x_j + x_{n'+1} = b_{i^*}, \quad x_{n'+1} \geq 0.$$

Mit Hilfe der Schlupfvariablen gelingt es aus dem linearen Programm in 1. Normalform ein lineares Programm in 2. Normalform zu machen. Diese sind äquivalent. Die Kosten der Einführung von Schlupfvariablen bestehen darin, dass man die Dimension des Lösungsvektors erhöht. □

Wir machen jetzt die folgenden Voraussetzungen:

1. $m < n$, das heißt weniger Nebenbedingungen als Unbekannte.
2. $\mathrm{rg}(A) = m$, das heißt, A hat vollen Zeilenrang, das heißt die Nebenbedingungen sind linear unabhängig.
3. $A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}$ sei widerspruchsfrei, das heißt, der zulässige Bereich ist nicht leer.

Definition 3.3 Basislösung. Basislösungen des linearen Programms (3.1) – (3.3) sollen die Lösungsvektoren $\mathbf{x} = (x_{i1}, \ldots, x_{im}, 0, \ldots, 0)^T$ heißen, für die die m Variablen $x_{i1}, \ldots, x_{im}$ eine nicht singuläre Koeffizientenmatrix

$$A_{m,m} = (\mathbf{a}_{i1}, \ldots, \mathbf{a}_{im}) \in \mathbb{R}^{m \times m},$$

besitzen, wobei $(\mathbf{a}_j)\,, j = 1, \ldots, n$, die Spaltenvektoren von A bezeichne. □

Die ersten m Variablen einer Basislösung können beliebige reelle Zahlen sein.

Definition 3.4 zulässige Basislösung, ausgeartete (entartete) Basislösung, Basisvariable, Basisvektor. Gilt für eine Basislösung $\mathbf{x} = (x_{i1}, \ldots, x_{im}, 0, \ldots, 0)^T$, dass $x_{ij} \geq 0$ für alle $j = 1, \ldots, m$, dann heißt sie zulässig. Verschwindet sie in mindestens einer der Variablen $x_{i1}, \ldots, x_{im}$, so heißt sie ausgeartet oder entartet.

Die Komponenten einer Basislösung werden Basisvariable genannt, die zugehörigen Spaltenvektoren heißen Basisvektoren. Entsprechend spricht man von Nichtbasisvariablen und Nichtbasisvektoren. □

Beispiel 3.5

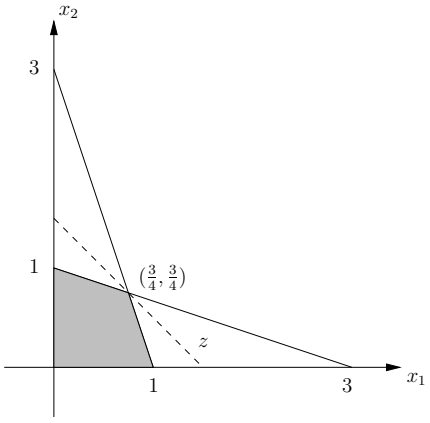
$$\begin{aligned} z = -x_1 - x_2 &\rightarrow \text{min !} \\ x_1 + 3x_2 + x_3 &= 3 \\ 3x_1 + x_2 + x_4 &= 3 \\ \mathbf{x} &\geq \mathbf{0}. \end{aligned}$$

Zulässige, nicht ausgeartete Basislösungen sind $(i1 = 3, i2 = 4)$

$$\mathbf{x} = (0, 0, 3, 3)^T,\ A_{2,2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ z = 0.$$

oder $(i1 = 1, i2 = 2)$

$$\mathbf{x} = (3/4, 3/4, 0, 0)^T,\ A_{2,2} \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}, z = -\frac{3}{2}.$$



Wir führen jetzt die weitere Nebenbedingung

$$x_1 + x_2 \leq \frac{3}{2}$$

ein. Die Nebenbedingungen des erweiterten linearen Programms haben die Gestalt

$$\begin{pmatrix} 1 & 3 & 1 & 0 & 0 \\ 3 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 1.5 \end{pmatrix}.$$

Dann ist eine ausgeartete zulässige Basislösung des erweiterten linearen Programms $(i1 = 1, i2 = 2, i3 = 5)$

$$(3/4, 3/4, 0, 0, 0)^T,\ A_{3,3} = \begin{pmatrix} 1 & 3 & 0 \\ 3 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix},\ z = -\frac{3}{2}.$$

Im Bild erkennt man, was Ausartung bedeutet. Die Ecke $(3/4, 3/4)$ des zulässigen Bereichs ist bereits durch die ersten beiden Nebenbedingungen bestimmt. Durch die neue Nebenbedingung ist diese Ecke nun wahlweise durch die ersten beiden, durch die erste und die dritte oder die zweite und die dritte Nebenbedingung bestimmt. Die Nebenbedingungen, die diese Ecke des zulässigen Bereichs bestimmen, sind nicht mehr eindeutig. □

Satz 3.6 *Ein Eckpunkt eines zulässigen Bereichs* $\mathcal{M} = \{\mathbf{x} \,:\, A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}\}$ *liegt genau dann vor, wenn seine Koordinaten eine zulässige Basislösung bilden.*

Beweis: *a) Aus Basislösung folgt Eckpunkt.* Sei $\mathbf{x} = (x_1, x_2, \ldots, x_m, 0, \ldots, 0)^T$ eine zulässige Basislösung, das heißt $x_i \geq 0$, $i = 1, \ldots m$,

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \ldots + \mathbf{a}_m x_m = \mathbf{b},$$

die Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_m$ sind linear unabhängig und die Nichtbasisvariablen $x_{m+1}, \ldots, x_n$ sind gleich Null.

Der Beweis wird indirekt geführt, indem angenommen wird, dass

$$\mathbf{x} = (x_1, \ldots, x_m, 0, \ldots, 0)$$

kein Eckpunkt ist. Dann gibt es Punkte $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{M}$ mit $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 = \mathbf{x}$, $\mathbf{x}_1 \neq \mathbf{x}_2$, $0 < \lambda < 1$. Da die letzten $n - m$ Komponenten von $\mathbf{x}$ verschwinden, muss das auch für entsprechenden Komponenten von $\mathbf{x}_1$ und $\mathbf{x}_2$ gelten, da alle Komponenten dieser Vektoren nichtnegativ sind. Seien nun

$$\mathbf{x}_1 = \left(x_1^{(1)}, x_2^{(1)}, \ldots, x_m^{(1)}, 0, \ldots, 0\right)^T, \quad \mathbf{x}_2 = \left(x_1^{(2)}, x_2^{(2)}, \ldots, x_m^{(2)}, 0, \ldots, 0\right)^T.$$

Da diese Punkte zulässig sind, folgt

$$\begin{aligned} \mathbf{a}_1 x_1^{(1)} + \mathbf{a}_2 x_2^{(1)} + \ldots + \mathbf{a}_m x_m^{(1)} &= \mathbf{b}, \\ \mathbf{a}_1 x_1^{(2)} + \mathbf{a}_2 x_2^{(2)} + \ldots + \mathbf{a}_m x_m^{(2)} &= \mathbf{b}. \end{aligned}$$

Wegen der linearen Unabhängigkeit der Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_m$ folgt daraus $\mathbf{x}_1 = \mathbf{x}_2$, was im Widerspruch zur Annahme steht. Also ist $\mathbf{x}$ ein Eckpunkt.

b) Aus Eckpunkt folgt Basislösung. Sei $\mathbf{x}$ ein Eckpunkt des zulässigen Bereichs mit den positiven Koordinaten $x_1, \ldots, x_k$, das heißt, es gilt

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \ldots + \mathbf{a}_k x_k = \mathbf{b} \text{ mit } x_j > 0,\ j = 1, \ldots, k \leq n.$$

Ohne Beschränkung der Allgemeinheit seien die verschwindenden Komponenten die hinteren. Es ist zu zeigen, dass $\mathbf{a}_1, \ldots, \mathbf{a}_k$ linear unabhängig sind.

Der Beweis ist wieder indirekt. Wir nehmen also an, dass es ein $\mathbf{y} \in \mathbb{R}^k$ gibt mit

$$\mathbf{a}_1 y_1 + \mathbf{a}_2 y_2 + \ldots + \mathbf{a}_k y_k = \mathbf{0}$$

und mindestens einem $y_i \neq 0$. Für jede reelle Zahl μ gelten damit

$$\sum_{j=1}^{k} \mathbf{a}_j \left(x_j + \mu y_j\right) = \mathbf{b} \quad \text{und} \quad \sum_{j=1}^{k} \mathbf{a}_j \left(x_j - \mu y_j\right) = \mathbf{b}.$$

Das bedeutet, die Punkte

$$\begin{aligned} \mathbf{x}_1 &= \left(x_1 + \mu y_1, \ldots, x_k + \mu y_k, 0, \ldots, 0\right), \\ \mathbf{x}_2 &= \left(x_1 - \mu y_1, \ldots, x_k - \mu y_k, 0, \ldots, 0\right) \end{aligned}$$

erfüllen die Nebenbedinungen (3.2). Falls man $\mu > 0$ hinreichend klein wählt, sind alle Komponenten dieser Punkte nichtnegativ und $\mathbf{x}_1, \mathbf{x}_2$ sind zulässig. Aus der Konstruktion von $\mathbf{x}_1, \mathbf{x}_2$ folgt, dass $\mathbf{x} = 0.5\mathbf{x}_1 + 0.5\mathbf{x}_2$ gilt. Das ist im Widerspruch zur Eckpunktannahme von $\mathbf{x}$. Diese Darstellung für den Eckpunkt $\mathbf{x}$ kann nur existieren, wenn $\mathbf{x}_1 = \mathbf{x}_2$. Da μ positiv ist, muss also $y_1 = \ldots = y_k = 0$ gelten. Also sind $\mathbf{a}_1, \ldots, \mathbf{a}_k$ linear unabhängig.

Die Basislösung verlangt jedoch m linear unabhängige Vektoren:

- Fall $k > m$. $m+1$ Vektoren des $\mathbb{R}^m$ sind stets linear abhänging. Dieser Fall kann also nicht eintreten.
- Fall $k = m$. In diesem Fall besitzt die zulässige Basislösung m positiven Komponenten, sie ist also nicht ausgeartet.
- Fall $k < m$. In diesem Fall hat man eine zulässige Basislösung mit weniger als m positiven Komponenten, also eine ausgeartete Basislösung. Aus den restlichen Spalten von A konstruiert man eine Menge von linear unabhängigen Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_m$, für welche offensichtlich

  $$\mathbf{a}_1 x_1 + \ldots + \mathbf{a}_k x_k + \mathbf{a}_{k+1} 0 + \ldots \mathbf{a}_m 0 = \mathbf{b}$$

  gilt. Diese Konstruktion ist möglich, da $\mathrm{rg}(A) = m$ ist.

■

Folgerung 3.7 ***Satz 1.15.*** *Ist der Lösungsbereich*

$$\mathcal{M} = \{\mathbf{x} \ : \ A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$$

beschränkt, so ist er ein konvexes Polyeder.

Beweis: Man kann nur endlich viele Mengen von m linear unabhängigen Spaltenvektoren der Matrix A bilden. (*Maximalanzahl ist Übungsaufgabe*) Mit dem eben bewiesenen Satz hat damit $\mathcal{M}$ nur endlich viele Ecken. ■

Eine weitere Folgerung des eben bewiesenen Satzes, zusammen mit Satz 2.4 ist wie folgt.

Folgerung 3.8 *Eine über einem konvexen Polyeder* $\mathcal{M} = \{\mathbf{x} \ : \ A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ *definierte lineare Funktion* $z = \mathbf{c}^T\mathbf{x}$ *nimmt ihr Minimum für wenigstens eine zulässige Basislösung an.*

Bemerkung 3.9 Naives Verfahren. Mit Hilfe der bisherigen Resultate können wir versuchen, ein Verfahren zur Lösung von

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ z = \mathbf{c}^T\mathbf{x} \ : \ \ A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0} \right\}$$

zu konstruieren:

1. Aufstellung der $\begin{pmatrix} n \\ m \end{pmatrix}$ linearen Gleichungssysteme der Dimension m aus $A\mathbf{x} = \mathbf{b}$.
2. Ist die so generierte Matrix $A_{m,m}$ regulär?
3. Angabe der Lösung für reguläre $A_{m,m}$.
4. Auswahl der zulässigen Lösungen
5. Bestimmung der Lösung(en), die das Minimum liefern.

Diese Herangehensweise ist jedoch schon bei relativ kleiner Anzahl von Unbekannten und Nebenbedingungen viel zu aufwendig. Zum Beispiel hätte man bei $n = 20$, $m = 10$ schon 184 756 Gleichungssysteme aufzustellen und diese zu untersuchen. □

Das Ziel wird nun sein, ein Verfahren zu finden, welches einen cleveren Weg zum Optimum findet, unter Nutzung von zulässigen Basislösungen.

Kapitel 4

Hauptsatz und Optimalitätskriterium der Simplexmethode

In diesem Abschnitt wird das wichtigste Verfahren zur Lösung linearer Optimierungsprobleme eingeführt – die Simplexmethode. Es existiere für

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ z = \mathbf{c}^T\mathbf{x} \ : \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0} \right\}$$

eine zulässige, nicht ausgeartete Basislösung.

Definition 4.1 Simplex. Ein Simplex ist die Menge aller Punkte $\mathbf{x} = (x_1, \ldots, x_n)^T$ mit

$$\sum_{i=1}^{n} x_i \leq 1, \quad x_i \geq 0, \ i = 1, \ldots, n.$$

□

Simplizes sind spezielle konvexe Polyeder. Für $n = 2$ ist das Dreieck mit den Eckpunkten $(0,0), (1,0), (0,1)$ ein Simplex.

Bemerkung 4.2 Geometrisches Prinzip der Simplexmethode. Das geometrische Prinzip der Simplexmethode ist wie folgt:

1. Man beginnt an einer Ecke des zulässigen Bereichs mit einer Startlösung $\mathbf{x}_1$ und dem Zielfunktionswert $z(\mathbf{x}_1)$.
2. Dann geht man entlang einer absteigenden Kante, dass heißt, bei welcher der Zielfunktionswert kleiner wird, $z(\mathbf{x}_1) > z(\mathbf{x}_2)$ zu einer sogenannten benachbarten Ecke $\mathbf{x}_2$.
3. Wiederhole Schritt 2 so lange, bis es keine absteigende Kante mehr gibt.

Dieses geometrische Prinzip muss in die algebraische Terminologie mit Basislösungen usw. transformiert werden. Wir werden später auch diskutieren, dass man Simplexschritte ausführen kann, bei denen der Zielfunktionswert gleich bleibt. In diesem Fall ist die Beschreibung des zweiten Schritts auch abzuändern, da man nicht zu einer benachbarten Ecke geht, sondern auf der gegebenen Ecke die Basis ändert. Diese Situation kann im Falle der Ausartung eintreten. Man nennt zwei Basislösungen benachbart, wenn sie sich nur in einem Basisvektor unterscheiden. □

Sei $\mathbf{x} = (x_1, \ldots, x_m, 0, \ldots, 0)^T$ eine erste zulässige Basislösung. Es gilt

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \ldots + \mathbf{a}_m x_m = \mathbf{b}. \tag{4.1}$$

Dabei sind $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ linear unabhängige Vektoren. Der Zielfunktionswert ist demzufolge

$$z_0 = c_1 x_1 + \ldots + c_m x_m. \tag{4.2}$$

Alle Nichtbasisvektoren $\mathbf{a}_{m+1}, \ldots, \mathbf{a}_n$ werden durch die Basis dargestellt

$$\mathbf{a}_j = x_{1j}\mathbf{a}_1 + \ldots + x_{mj}\mathbf{a}_m, \quad j = m+1, \ldots, n. \tag{4.3}$$

Mit diesen Darstellungskoeffizienten x_{ij}, die nichts mit den Werten $x_1, \ldots, x_m$ der Basislösung zu tun haben, werden die Hilfsgrößen

$$z_j = c_1 x_{1j} + c_2 x_{2j} + \ldots + c_m x_{mj}, \quad j = m+1, \ldots, n, \tag{4.4}$$

eingeführt.

Satz 4.3 ***Hauptsatz der Simplexmethode.*** *Sei z_0 der Wert der Zielfunktion für die zulässige Basislösung $\mathbf{x} = (x_1, \ldots, x_m, 0, \ldots, 0)^T$, $x_i > 0$, $i = 1, \ldots, m$. Gilt für ein k mit $m+1 \leq k \leq n$, dass $z_k - c_k > 0$, so existiert wenigstens eine zulässige Basislösung mit einem Zielfunktionswert kleiner als z_0.*

Beweis: Sei $\theta > 0$ vorerst beliebig gewählt. Man multipliziere (4.3) für $j = k$ mit θ und bilde (4.1)$-\theta$ (4.3):

$$\mathbf{a}_1 (x_1 - \theta x_{1k}) + \mathbf{a}_2 (x_2 - \theta x_{2k}) + \ldots + \mathbf{a}_m (x_m - \theta x_{mk}) + \theta \mathbf{a}_k = \mathbf{b}, \tag{4.5}$$

In der Gleichung (4.5) steht ein Vektor, der $A\mathbf{x} = \mathbf{b}$ erfüllt:

$$(x_1 - \theta x_{1k}, \ldots, x_m - \theta x_{mk}, 0, \ldots, \theta, \ldots, 0)^T. \tag{4.6}$$

Es wird in den Lemmata 4.4 und 4.7 gezeigt, dass man mit diesem Vektor eine Basislösung konstruieren kann. Für seinen Zielfunktionswert gilt

$$\begin{aligned} & c_1 (x_1 - \theta x_{1k}) + c_2 (x_2 - \theta x_{2k}) + \ldots + c_m (x_m - \theta x_{mk}) + \theta c_k \\ = \quad & z_0 - \theta z_k + \theta c_k = z_0 + \theta (c_k - z_k). \end{aligned} \tag{4.7}$$

Der zugehörige Zielfunktionswert ist kleiner als z_0 falls $\theta > 0$ und $z_k - c_k > 0$. ∎

Unter der Annahme, dass der Hauptsatz bereits vollständig bewiesen ist, haben wir ein hinreichendes Kriterium um zu entscheiden, ob es eine zulässige Basislösung mit einem kleineren Zielfunktionswert gibt. Man benötigt jetzt noch eine Methode zur Konstruktion dieser zulässigen Basislösung. Diese erfolgt mit Hilfe von θ.

Lemma 4.4 Wahl von θ. *Sei*

$$\theta = \min_{i=1,\ldots,m, x_{ik}>0} \frac{x_i}{x_{ik}} =: \frac{x_l}{x_{lk}}. \tag{4.8}$$

Dann ist die Lösung (4.6) zulässig.

Beweis: Man hat

$$\begin{aligned} & \mathbf{a}_1 \left(x_1 - \frac{x_l}{x_{lk}} x_{1k} \right) + \ldots + \mathbf{a}_{l-1} \left(x_{l-1} - \frac{x_l}{x_{lk}} x_{l-1,k} \right) + \mathbf{a}_l \underbrace{\left(x_l - \frac{x_l}{x_{lk}} x_{l,k} \right)}_{=0} \\ & + \mathbf{a}_{l+1} \left(x_{l+1} - \frac{x_l}{x_{lk}} x_{l+1,k} \right) + \ldots + \mathbf{a}_m \left(x_m - \frac{x_l}{x_{lk}} x_{mk} \right) + \frac{x_l}{x_{lk}} \mathbf{a}_k \quad = \quad \mathbf{b}, \end{aligned}$$

und die neue Lösung

$$\hat{x}_i = x_i - \frac{x_l}{x_{lk}} x_{ik}, \quad i = 1, \ldots, m, \ i \neq l, \quad \hat{x}_k = \frac{x_l}{x_{lk}}. \tag{4.9}$$

Alle Komponenten sind auf Grund der Konstruktion nichtnegativ und bei Ausschluss der Entartung sogar positiv. Damit hat man eine zulässige Lösung erhalten. Man hat also die Komponente x_l aus der Basisliste gestrichen und durch die Komponente x_k ersetzt. ∎

Bemerkung 4.5 Nicht nach unten beschränkte Zielfunktion. Damit die Wahl (4.8) von θ überhaupt funktioniert, brauchen wir ein $x_{ik} > 0$. Falls es kein solches x_{ik} gibt, dann folgt, dass die Zielfunktion nach unten unbeschränkt ist. Man kann nämlich in diesem Fall θ beliebig groß wählen, da stets $x_i - \theta x_{ik} \geq 0$. Aus (4.7) folgt dann, dass unter der Bedingung $c_k - z_k < 0$ die Zielfunktion unbeschränkt nach unten ist. Fazit: Falls für ein $z_k - c_k > 0$ alle $x_{ik} \leq 0$, dann ist die Zielfunktion nicht von unten beschränkt und man breche die Simplexmethode ab. □

Bemerkung 4.6 Basiszyklus. Wenn man Entartung ausschließt, dann wird das Minimum in (4.8) wird für genau einen Index l angenommen. Es gilt auch die Umkehrung, dass falls der Index l in (4.8) nicht eindeutig bestimmt ist, dann hat man Ausartung. Ausartung kann zur Folge haben, dass gilt $z(\mathbf{x}_i) = z(\mathbf{x}_{i+1}) = \ldots$. Das nennt man einen Basiszyklus. □

Es gilt also, (4.9) ist eine zulässige Lösung mit einem kleineren Zielfunktionswert als die ursprüngliche Lösung. Damit bleibt nur noch die Basiseigenschaft von $\{\mathbf{a}_1, \ldots, \mathbf{a}_{l-1}, \mathbf{a}_k, \mathbf{a}_{l+1}, \ldots, \mathbf{a}_m\}$ zu prüfen.

Lemma 4.7 *Sei $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ ein System linear unabhängiger Vektoren und sei*

$$\mathbf{w} = \sum_{i=1}^{m} \mu_i \mathbf{w}_i, \quad \mu_l \neq 0. \tag{4.10}$$

Dann ist auch $\{\mathbf{w}_1, \ldots, \mathbf{w}_{l-1}, \mathbf{w}, \mathbf{w}_{l+1}, \ldots \mathbf{w}_m\}$ ein System linear unabhängiger Vektoren.

Beweis: Indirekter Beweis. Sei $\{\mathbf{w}_1, \ldots, \mathbf{w}_{l-1}, \mathbf{w}, \mathbf{w}_{l+1}, \ldots \mathbf{w}_m\}$ kein System linear unabhängiger Vektoren. Dann gibt es Zahlen $\alpha_1, \ldots, \alpha_{l-1}, \alpha_{l+1}, \ldots, \alpha_m, \alpha$, von denen wenigstens eine ungleich Null ist, so dass

$$\sum_{i=1, i\neq l}^{m} \alpha_i \mathbf{w}_i + \alpha \mathbf{w} = \mathbf{0}.$$

In diese Gleichung wird (4.10) eingesetzt. Es folgt

$$\sum_{i=1, i\neq l}^{m} (\alpha_i + \alpha\mu_i)\, \mathbf{w}_i + \alpha\mu_l \mathbf{w}_l = \mathbf{0}.$$

Die Vektoren $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ sind linear unabhängig, das heißt, alle Koeffizienten in dieser Gleichung müssen Null sein. Wegen $\mu_l \neq 0$ folgt dann $\alpha = 0$ und daraus $\alpha_i = 0$ für alle i. Damit ist gezeigt, dass $\{\mathbf{w}_1, \ldots, \mathbf{w}_{l-1}, \mathbf{w}, \mathbf{w}_{l+1}, \ldots \mathbf{w}_m\}$ ein System linear unabhängiger Vektoren ist. ■

Folgerung 4.8 *Das im Hauptsatz der Simplexmethode, Satz 4.3, konstruierte System von Vektoren ist mit der Wahl von θ nach (4.8) ein System linear unabhängiger Vektoren.*

Beweis: Da in der Linearkombination (4.3) der Faktor vor $\mathbf{a}_l$ gleich $x_{lk} (= \mu_l)$ ist und wir $x_{lk} > 0$ bei der Definition von l vorausgesetzt haben, lässt sich Lemma 4.7 anwenden und die Basiseigenschaft von $\{\mathbf{a}_1, \ldots, \mathbf{a}_{l-1}, \mathbf{a}_k, \mathbf{a}_{l+1}, \ldots, \mathbf{a}_m\}$ ist gewährleistet. ■

Bemerkung 4.9 Im allgemeinen ist der Hauptsatz der Simplexmethode solange anzuwenden, wie noch wenigstens ein $z_k - c_k > 0$ ist. Dabei kann man im allgemeinen nicht erwarten, falls noch q Größen $z_j - c_j > 0$ existieren, dass man noch q Schritte auszuführen hat. Gilt für alle $z_j - c_j \leq 0$, j – Index von Nichtbasisvariablen, so ist man in dem Sinne fertig, dass der Hauptsatz nicht mehr anwendbar ist. Der Hauptsatz gibt aber bisher nur ein hinreichendes und kein notwendiges Kriterium für die Existenz einer Basislösung mit einem kleineren Zielfunktionswert. □

Im folgenden Satz wird gezeigt, dass das Kriterium auch notwendig ist.

Satz 4.10 ***Optimalitätskriterium.*** *Eine zulässige Basislösung* $\mathbf{x} \in \mathbb{R}^n$ *mit* $z_0 = \sum_{i=1}^{m} c_i x_i$ *ist optimale Basislösung, wenn für alle* $j = m+1, \ldots, n$, *gilt* $z_j - c_j \leq 0$.

Beweis: Sei $\mathbf{x} = (x_1, \ldots, x_m, 0, \ldots, 0)^T$. Des weiteren sei $\mathbf{y} = (y_1, \ldots, y_n)^T$ eine beliebige zulässige Lösung

$$\mathbf{a}_1 y_1 + \mathbf{a}_2 y_2 + \ldots + \mathbf{a}_n y_n = \mathbf{b}, \quad \mathbf{y} \geq \mathbf{0}, \tag{4.11}$$

$$\text{mit } z^* = \sum_{i=1}^{n} c_i y_i. \tag{4.12}$$

Zu zeigen ist, dass $z_0 \leq z^*$ für alle $\mathbf{y}$.

Durch (4.3) ist jeder Nichtbasisvektor mit Hilfe der Basis dargestellt. Jetzt wird diese Darstellung auf die Basisvektoren ausgedehnt

$$\mathbf{a}_j = x_{1j}\mathbf{a}_1 + \ldots + x_{mj}\mathbf{a}_m, \quad j = 1, \ldots, n,$$

wobei

$$x_{ij} = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases}, i = 1, \ldots, m. \tag{4.13}$$

Weiter gilt die Darstellung (4.4) für z_j, $j = m+1, \ldots, n$. Mit (4.13) hat man eine analoge Darstellung für $j = 1, \ldots, m$, die sich letztlich auf $z_j = c_j$ reduziert. Zusammen mit der Voraussetzung gilt jetzt $z_j \leq c_j$, $j = 1, \ldots, n$. Mit (4.12) folgt nun

$$\sum_{i=1}^{n} z_i y_i \leq z^*. \tag{4.14}$$

Nun wird in (4.11) die Darstellung aller Spaltenvektoren durch die ersten m Spaltenvektoren eingesetzt

$$y_1 \sum_{i=1}^{m} x_{i1}\mathbf{a}_i + y_2 \sum_{i=1}^{m} x_{i2}\mathbf{a}_i + \ldots + y_n \sum_{i=1}^{m} x_{in}\mathbf{a}_i = \mathbf{b}.$$

Durch Umordnung nach den Basisvektoren folgt

$$\mathbf{a}_1 \sum_{j=1}^{n} y_j x_{1j} + \mathbf{a}_2 \sum_{j=1}^{n} y_j x_{2j} + \ldots + \mathbf{a}_m \sum_{j=1}^{n} y_j x_{mj} = \mathbf{b}. \tag{4.15}$$

Analog schreibt man (4.14) mit Hilfe von (4.4) und der entsprechenden Darstellung für $j = 1, \ldots, m$, mit (4.13) ($z_j = c_j$, $j = 1, \ldots, m$)

$$c_1 \sum_{j=1}^{n} y_j x_{1j} + c_2 \sum_{j=1}^{n} y_j x_{2j} + \ldots + c_m \sum_{j=1}^{n} y_j x_{mj} \leq z^* \tag{4.16}$$

Der Vektor $\mathbf{x}$ ist eine zulässige Basislösung, das heißt, es gilt

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \ldots + \mathbf{a}_m x_m = \mathbf{b}. \tag{4.17}$$

Da $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ eine Basis ist, ist die Darstellung von $\mathbf{b}$ mit Hilfe dieser Vektoren eindeutig. Damit folgt aus (4.15) und (4.17)

$$x_i = \sum_{j=1}^{n} y_j x_{ij}, \; i = 1, \ldots, m.$$

Setzt man dies in (4.16), so erhält man

$$z_0 = \sum_{i=1}^{m} c_i x_i \leq z^*.$$

■

Kapitel 5

Die Simplexmethode

Bemerkung 5.1 Simplextabelle und Simplexmethode. Es werden folgende Bezeichnungen verwendet:

- das untersuchte Problem ist $\min_{\mathbf{x}\in\mathbb{R}^n}\left\{z=\mathbf{c}^T\mathbf{x} \;:\; A\mathbf{x}=\mathbf{b}, \mathbf{x}\geq\mathbf{0}\right\}$,
- die erste zulässige Basislösung sei $\mathbf{x}=(x_1,x_2,\ldots,x_m,0,\ldots,0)^T$, $\mathbf{x}\geq\mathbf{0}$, mit $z_0=\mathbf{c}^T\mathbf{x}$,
- die Basisvektoren sind $A_B=(\mathbf{a}_1,\ldots,\mathbf{a}_m)$,
- die Nichtbasisvektoren sind $A_N=(\mathbf{a}_{m+1},\ldots,\mathbf{a}_n)$,
- die Darstellung der Nichtbasisvektoren durch die Basis ist

$$\mathbf{a}_j = x_{1j}\mathbf{a}_1+\ldots+x_{mj}\mathbf{a}_m, \quad j=m+1,\ldots,n,$$

- die Hilfsgrößen z_j sind

$$z_j = c_1x_{1j}+c_2x_{2j}+\ldots+c_mx_{mj}, \quad j=m+1,\ldots,n.$$

Diese Größen werden in der sogenannten Simplextabelle eingetragen:

			$m+1$	$m+2$	$\ldots$	k	$\ldots$	n
i	c_i	x_i	c_{m+1}	c_{m+2}	$\ldots$	c_k	$\ldots$	c_n
1	c_1	x_1	$x_{1,m+1}$	$x_{1,m+2}$	$\ldots$	$x_{1,k}$	$\ldots$	$x_{1,n}$
2	c_2	x_2	$x_{2,m+1}$	$x_{2,m+2}$	$\ldots$	$x_{2,k}$	$\ldots$	$x_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
l	c_l	x_l	$x_{l,m+1}$	$x_{l,m+2}$	$\ldots$	$x_{l,k}$	$\ldots$	$x_{l,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	c_m	x_m	$x_{m,m+1}$	$x_{m,m+2}$	$\ldots$	$x_{m,k}$	$\ldots$	$x_{m,n}$
		z_0	$z_{m+1}-c_{m+1}$	$z_{m+2}-c_{m+2}$	$\ldots$	z_k-c_k	$\ldots$	z_n-c_n
Basisteil			Nichtbasisteil					

Bei der Simplexmethode folgt man jetzt im wesentlichen dem Beweis des Hauptsatzes. Sei $z_k-c_k>0$. Gilt für mehrere Indizes $j\in\{m+1,\ldots,n\}$, dass $z_j-c_j>0$, so nehme man zum Beispiel einen Index, bei dem die Differenz maximal ist

$$z_k-c_k := \max_{j=m+1,\ldots,n,} z_j-c_j.$$

Dann liegt x_k als Nichtbasisvariable vor, die in die Basis soll. Nun bestimmt man

$$\theta = \min_{i=1,\ldots,m,x_{ik}>0} \frac{x_i}{x_{ik}} =: \frac{x_l}{x_{lk}},$$

das heißt, x_l soll aus der Basis raus. □

Definition 5.2 Hauptspalte, Hauptzeile, Hauptelement, Pivotelement. Die Spalte k nennt man Hauptspalte, die Zeile l heißt Hauptzeile und das Element x_{lk} heißt Hauptelement oder Pivotelement. □

Bemerkung 5.3 Berechnung der Einträge der neuen Simplextabelle. Die neue Basislösung sei

$$(\hat{x}_1, \ldots, \hat{x}_{l-1}, \hat{x}_k, \hat{x}_{l+1}, \ldots, \hat{x}_m, 0, \ldots, 0)^T . \tag{5.1}$$

Nun müssen die Elemente der neuen Simplextabelle bestimmt werden:

1. Man benötigt insbesondere eine Darstellung von (5.1). Aus (4.9) erhält man

$$\hat{x}_i = x_i - \frac{x_l}{x_{lk}} x_{ik}, \quad i = 1, \ldots, m;\ i \neq l; \quad \hat{x}_k = \frac{x_l}{x_{lk}}. \tag{5.2}$$

2. Aus (4.3) folgt für $j = k$

$$\begin{aligned} \mathbf{a}_l &= \frac{1}{x_{lk}} \left(\mathbf{a}_k - x_{1k}\mathbf{a}_1 - \ldots - x_{l-1,k}\mathbf{a}_{l-1} - x_{l+1,k}\mathbf{a}_{l+1} - \ldots - x_{mk}\mathbf{a}_m \right) \\ &= -\frac{x_{1k}}{x_{lk}}\mathbf{a}_1 - \ldots - \frac{x_{l-1,k}}{x_{lk}}\mathbf{a}_{l-1} + \frac{\mathbf{a}_k}{x_{lk}} - \frac{x_{l+1,k}}{x_{lk}}\mathbf{a}_{l+1} - \ldots - \frac{x_{mk}}{x_{lk}}\mathbf{a}_m. \end{aligned} \tag{5.3}$$

   Damit haben wir eine Darstellung des neuen Nichtbasisvektors $\mathbf{a}_l$ durch die neue Basis und die neuen Elemente der alten Hauptspalte sind

$$\hat{x}_{kl} = \frac{1}{x_{lk}}, \quad \hat{x}_{il} = -\frac{x_{ik}}{x_{lk}},\ i = 1, \ldots, m,\ i \neq k. \tag{5.4}$$

3. Für den Rest erhält man, beispielhaft an $\mathbf{a}_n$ gezeigt, die folgende Darstellung, wobei man in der ersten Gleichung die alte Basisdarstellung (4.3) nutzt:

$$\begin{aligned} \mathbf{a}_n &= x_{1n}\mathbf{a}_1 + \ldots + x_{l-1,n}\mathbf{a}_{l-1} + x_{l+1,n}\mathbf{a}_{l+1} + \ldots + x_{mn}\mathbf{a}_n + x_{ln} \underbrace{\mathbf{a}_l}_{(5.3)} \\ &= \left(x_{1n} - \frac{x_{1k}x_{ln}}{x_{lk}} \right) \mathbf{a}_1 + \ldots + \left(x_{l-1,n} - \frac{x_{l-1,k}x_{ln}}{x_{lk}} \right) \mathbf{a}_{l-1} + \frac{x_{ln}}{x_{lk}}\mathbf{a}_k \\ &\quad + \ldots + \left(x_{mn} - \frac{x_{mk}x_{ln}}{x_{lk}} \right) \mathbf{a}_m. \end{aligned}$$

   Man erhält also die folgenden Koeffizienten für die neue Basisdarstellung

$$\hat{x}_{kj} = \frac{x_{lj}}{x_{lk}}, \quad j = m+1, \ldots, n,\ j \neq k, \tag{5.5}$$

$$\hat{x}_{ij} = x_{ij} - \underbrace{\frac{x_{lj}}{x_{lk}}}_{\hat{x}_{kj}} x_{ik}, \quad i = 1, \ldots, m,\ i \neq k,\ j = m+1, \ldots, n,\ j \neq l. \tag{5.6}$$

4. Die Elemente z_0, $z_{m+1} - c_{m+1}, \ldots, z_n - c_n$ transformieren sich ebenfalls nach den obigen Regeln. *Übungsaufgabe*

Damit sind alle Elemente der neuen Simplextabelle berechnet. Zur Berechnung von $\hat{x}_{ij}$ benötigt man die im Rechteck angeordneten Elemente x_{ij}, x_{lj}, x_{lk} und x_{ik} der alten Simplextabelle. Deshalb spricht man auch von der Rechteckregel. □

Bemerkung 5.4 Basisform der Simplexmethode.

1. Normalform des linearen Programms herstellen.
2. Erste zulässige Basislösung angeben.
3. Simplextabelle zu dieser Basislösung erstellen.
4. Existieren Bewertungen $z_j - c_j > 0$? Wenn ja, gehe zu 6.
5. Sind alle Bewertungen $z_j - c_j < 0$?
   - Wenn ja, einzige Optimallösung gefunden, *Simplexmethode beendet.*
   - Wenn nicht, dann gibt es außer negativen Bewertungen $z_j - c_j$ nur noch verschwindende. Das Optimum nicht eindeutig. Man hat ein Optimum gefunden, *beende Simplexmethode.*
6. Wähle die Hauptspalte, also die Spalte, zu der das größte $z_j - c_j > 0$, $j = k$ gehört.
7. Falls $x_{ik} \leq 0$ für alle $i = 1, \ldots, m$, so ist die Zielfunktion nach unten nicht beschränkt, *beende Simplexmethode.*
8. Bestimme θ zur Festlegung der Hauptzeile und des Pivotelements.
9. Basistransformation:
   9.1 Ersetze das Pivotelement durch seinen Kehrwert, siehe (5.4).
   9.2 Multipliziere die übrigen Elemente der Hauptzeile mit diesem Kehrwert, einschließlich x_l, siehe (5.2) und (5.5).
   9.3 Multipliziere die übrigen Elemente der Hauptspalte mit dem negativen Kehrwert, siehe (5.4).
   9.4 Vermindere die nicht in einer Hauptreihe stehenden Elemente, einschließlich der übrigen Werte von x_i und der letzten Zeile, um das Produkt der zugehörigen Hauptreihenelemente (Rechteckregel). Dabei nimmt man für das Pivotelement schon den neuen Wert und für die übrigen Elemente die alten Werte, siehe (5.2) und (5.6).
10. Gehe zu 4.

□

Beispiel 5.5 Wir betrachten das lineare Programm

$$z = -3x_1 - 2x_2 - 4x_3 - x_4 \quad \rightarrow \quad \min\ !$$

$$\begin{pmatrix} 2 & 2 & 3 & 0 & 1 & 0 & 0 \\ 1 & 3 & 0 & 2 & 0 & 1 & 0 \\ 1 & 1 & 5 & 2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} 700 \\ 400 \\ 500 \end{pmatrix}$$

$$\mathbf{x} \geq \mathbf{0}.$$

Bekannt sei eine erste zulässige Basislösung $x_1 = 350$, $x_4 = 25$, $x_7 = 100$, die den Zielfunktionswert $z = -1075$ besitzt. Die Basisvektoren sind demzufolge

$$\mathbf{a}_1 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{a}_4 = \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix}, \quad \mathbf{a}_7 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Gesucht ist nun die Darstellung der Nichtbasisvektoren durch die Basisvektoren. Setze $A_B = (\mathbf{a}_1, \mathbf{a}_4, \mathbf{a}_7)$ und $A_N = (\mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_5, \mathbf{a}_6)$. Dann ist die Matrix X der Simplexkoeffizienten gesucht, für die gilt

$$A_N = A_B X \quad \Longrightarrow \quad X = A_B^{-1} A_N.$$

Man erhält hier

$$X = \begin{pmatrix} 1 & 3/2 & 1/2 & 0 \\ 1 & -3/4 & -1/4 & 1/2 \\ -2 & 5 & 0 & -1 \end{pmatrix}.$$

Daraus ergibt sich

$$\begin{aligned} z_2 &= c_1x_{12} + c_4x_{42} + c_7x_{72} = (-3)1 + (-1)1 + 0(-2) = -4, \\ z_3 &= -9/2 + 3/4 + 0 = -15/4, \\ z_5 &= -3/2 + 1/4 + 0 = -5/4, \\ z_6 &= 0 - 1/2 + 0 = -1/2 \end{aligned}$$

und somit

$$z_2 - c_2 = -2, \quad z_3 - c_3 = 1/4, \quad z_5 - c_5 = -5/4, \quad z_6 - c_6 = -1/2.$$

Damit erhält man folgende Simplextabelle:

			2	3	5	6
i	c_i	x_i	-2	-4	0	0
1	-3	350	1	3/2	1/2	0
4	-1	25	1	-3/4	-1/4	1/2
7	0	100	-2	**5**	0	-1
		-1075	-2	**1/4**	-5/4	-1/2

Es gibt nur einen Index k mit $z_k - c_k > 0$, nämlich $k = 3$. Damit ist die Hauptspalte bestimmt (Schritt 6). Zur Bestimmung der Hauptzeile (Schritt 8) berechnet man θ:

$$\theta = \min_{x_{i3}>0, i\in\{1,4,7\}} \left(\frac{x_i}{x_{i3}}\right) = \min\left\{\frac{350}{3/2}, \frac{100}{5}\right\} = 20$$

für $i = 7$. Damit ist der Hauptzeilenindex $l = 7$ und das Pivotelement $x_{73} = 5$. Nun führt man die Basistransformation aus (Schritt 9):

			2	7	5	6
i	c_i	x_i	-2	0	0	0
1	-3	320	8/5	-3/10	1/2	3/10
4	-1	40	7/10	3/20	-1/4	7/20
3	-4	20	-2/5	1/5	0	-1/5
		-1080	-19/10	-1/20	-5/4	-9/20

Den neuen Wert für x_1 erhält man beispielsweise aus

$$x_1 = 350 - \frac{3}{2}100\frac{1}{5} = 350 - 30 = 320.$$

Da in der neuen Simplextabelle alle Werte $z_j - c_j < 0$, $j \in \{2, 5, 6, 7\}$, hat man die einzige Optimallösung bestimme: $\mathbf{x} = (320, 0, 20, 40, 0, 0, 0)^T$. □

Bemerkung 5.6 Angenommen, man hat in einer Simplextabelle mehrere $z_j - c_j > 0$. Zu einer dieser Spalten mögen nur Koeffizienten $x_{ij} \leq 0$ gehören. Dann ist die Zielfunktion unbeschränkt. □

Beispiel 5.7 Zur Ausartung. Wir betrachten das lineare Programm

$$\begin{aligned} z = -x_1 \quad &\rightarrow \quad \min\ ! \\ \begin{pmatrix} 1 & 1 & 1 & 0 \\ 4 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad &= \quad \begin{pmatrix} 1 \\ 4 \end{pmatrix} \\ \mathbf{x} \quad &\geq \quad \mathbf{0}. \end{aligned}$$

Eine zulässige Basislösung, die gleichzeitig ein Optimum ist, ist $\mathbf{x} = (1, 0, 0, 0)^T$. Wir nehmen als Basisvariablen x_1 und x_2. Da x_2 verschwindet, ist die Basislösung ausgeartet. Man hat

$$A_B = \begin{pmatrix} 1 & 1 \\ 4 & 1 \end{pmatrix}, \quad A_N = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

und erhält die Simplextabelle

			3	4
i	c_i	x_i	0	0
1	-1	1	-1/3	1/3
2	0	0	4/3	-1/3
		-1	1/3	-1/3

Die Simplexmethode sagt uns an dieser Stelle nicht, dass das Optimum bereits erreicht ist! Gemäß Simplexmethode muss x_3 in die Basis anstelle von x_2 eingeführt werden. Man erhält die Simplextabelle

			2	4
i	c_i	x_i	0	0
1	-1	1	1/4	1/4
3	0	0	3/4	-1/4
		-1	-1/4	-1/4

Damit ist das Optimalitätskriterium der Simplexmethode erfüllt und diese wird beendet. Man hat für das Optimum $\mathbf{x} = (1, 0, 0, 0)^T$ mit diesen beiden Simplextabellen zwei unterschiedliche Basisdarstellungen. Der Zielfunktionswert hat sich im Simplexschritt nicht verändert, es wurde lediglich die Basis gewechselt. □

Kapitel 6

Bestimmung einer ersten zulässigen Basislösung

Ein Problem, was man für die Durchführung der Simplexmethode lösen muss, ist die Bestimmung einer ersten zulässigen Basislösung. Wie gut das geht, hängt auch vom konkreten Problem ab.

Bemerkung 6.1 1. Fall. Liegt

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ z = \mathbf{c}^T\mathbf{x} \ : \quad A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0} \right\}$$

vor und gilt $\mathbf{b} \geq \mathbf{0}$. Dann führt man Schlupfvariablen ein und setzt $\mathbf{x} = (0, \ldots, 0, \mathbf{b}^T)^T$. □

Bemerkung 6.2 2. Fall, Engpassmethode. Liegt das lineare Programm in der Gestalt

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ z = \mathbf{c}^T\mathbf{x} \ : \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0} \right\}$$

vor mit $A = (a_{ij}), i = 1, \ldots, m; j = 1, \ldots, n$, $\mathbf{b} = (b_1, \ldots, b_m)^T$, $a_{ij} \geq 0$, $b_i \geq 0$ für alle i, j. Dann kann man mit einer sogenannten Engpassmethode zur ersten zulässigen Basislösung gelangen:

1. Ordne die Variablen nach wachsenden Zielfunktionskoeffizienten c_i, Beispiel

$$\begin{aligned} z = -10x_1 - 6x_2 - 4x_3 - 3x_4 - 5x_5 \quad &\rightarrow \quad \min\ ! \\ \begin{pmatrix} 2 & 0 & 4 & 0 & 2 & 1 & 0 & 0 & 0 \\ 2 & 3 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 5 & 0 & 0 & 2 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_9 \end{pmatrix} \quad &= \quad \begin{pmatrix} 8 \\ 6 \\ 20 \\ 8 \end{pmatrix} \\ \mathbf{x} \quad &\geq \quad \mathbf{0}. \end{aligned}$$

   Dann ist die Ordnung x_1, x_2, x_5, x_3, x_4.
2. In der festgelegten Reihenfolge werden die Variablen mit dem größtmöglichen Wert genommen, so dass die Nebenbedingungen erfüllt sind. Im Beispiel beginnt man mit $x_1 = 3$
3. Man setzt diesen Wert ein und entfernt die Variable damit aus den Nebenbedingungen. Im Beispiel ergibt sich

$$\begin{pmatrix} 0 & 4 & 0 & 2 & 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_9 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 5 \\ 8 \end{pmatrix}.$$

Aus der zweiten Gleichung folgt $x_2 = x_3 = x_7 = 0$, welche Werte man auch gleich einsetzen kann. Damit vereinfacht sich das System der Nebenbedingungen zu

$$\begin{pmatrix} 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ 3 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_4 \\ x_5 \\ x_6 \\ x_8 \\ x_9 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 5 \\ 8 \end{pmatrix}. \tag{6.1}$$

4. Gehe zu 2.

Im Beispiel betrachtet man als nächstes x_5, da ja bereits $x_2 = 0$ gilt. Der maximale Wert von x_5, so dass (6.1) erfüllt ist, beträgt $x_5 = 1$. Einsetzen ergibt

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 3 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_4 \\ x_6 \\ x_8 \\ x_9 \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \\ 8 \end{pmatrix}. \tag{6.2}$$

Damit folgt $x_6 = 0$. Da ja auch schon $x_3 = 0$ gilt, wird nun x_4 betrachtet. Der maximale Wert von x_4, so dass (6.2) erfüllt ist, ist $x_4 = 2$. Man erhält

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_8 \\ x_9 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Nun bestimmt man die letzten beiden Werte und erhält als erste zulässige Basislösung $\mathbf{x} = (3, 0, 0, 2, 1, 0, 0, 0, 2)^T$. □

Bemerkung 6.3 Hat man bei der Engpassmethode nicht genügend Variablen, dann führt man künstliche Variablen ein. □

Bemerkung 6.4 Anderes Ordnungsprinzip der Variablen im Fall, dass die Koeffizienten von unterschiedlicher Größenordnung sind. Wir betrachten das lineare Programm

$$\begin{aligned} z = 10x_1 + 20x_2 + 30x_3 + 40x_4 + 50x_5 \quad &\rightarrow \quad \min\ ! \\ \begin{pmatrix} 1 & 10 & 50 & 1 & 0 \\ 2 & 20 & 50 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_5 \end{pmatrix} \quad &= \quad \begin{pmatrix} 100 \\ 101 \end{pmatrix} \\ \mathbf{x} \quad &\geq \quad \mathbf{0}. \end{aligned}$$

Nach dem obigen Ordnungsprinzip hat man die Reihenfolge x_1, x_2, x_3, x_4, x_5 und erhält mit der Engpassmethode die erste zulässige Basislösung *Übungsaufgabe*

$$x_1 = \frac{101}{2},\ x_4 = \frac{99}{2} \quad \Longrightarrow \quad z = 2485.$$

Man erhält jedoch mit einer anderen Basislösung einen schon viel kleineren Zielfunktionswert

$$x_3 = 2,\ x_5 = 1 \quad \Longrightarrow \quad z = 110.$$

In diesem Fall ist das Ordnungsprinzip

$$\min_{j, c_j \neq 0} \left\{ c_j \min_{i, a_{ij} \neq 0} \left\{ \frac{b_i}{a_{ij}} \right\} \right\}$$

günstiger. Bei diesem Ordnungsprinzip wird auch die Größe der Matrixeinträge und der rechten Seite beachtet. Im Beispiel kann man x_3 wegen der großen Matrixeinträge nur relativ klein wählen, wenn man die Nebenbedingungen nicht verletzen will.

Im Gegensatz dazu kann man x_1 sehr groß wählen ohne die Nebenbedingungen zu verletzen. Der etwas höhere Koeffizient vor x_3 in der Zielfunktion führt wegen des viel kleineren möglichen Wertes dieser Variablen letztlich auf einen kleineren Beitrag als $10x_1$ mit großem x_1. □

Bemerkung 6.5 3. Fall, Bestimmung mit Hilfe der Simplexmethode. Die erste zulässige Basislösung soll jetzt

- ohne spezielle Voraussetzungen und
- mit Hilfe der Simplexmethode

bestimmt werden. Dazu betrachten wir

$$\sum_{j=1}^{n} c_j x_j = \mathbf{c}^T \mathbf{x} \quad \rightarrow \quad \min \; ! \tag{6.3}$$

$$A\mathbf{x} \quad = \quad \mathbf{b}, \tag{6.4}$$

$$\mathbf{x} \quad \geq \quad \mathbf{0}. \tag{6.5}$$

und nehmen $\mathbf{b} \geq \mathbf{0}$ an. Das kann immer durch Multiplikation der entsprechenden Gleichungen mit einer negativen Zahl erreicht werden. Dem Problem (6.3) – (6.5) wird die Hilfsaufgabe

$$\sum_{i=1}^{m} x_{n+i} \quad \rightarrow \quad \min \; ! \tag{6.6}$$

$$\sum_{j=1}^{n} a_{ij} x_j + x_{n+i} \quad = \quad b_i, \quad i = 1, \ldots, m, \tag{6.7}$$

$$x_j \quad \geq \quad 0, \quad j = 1, \ldots, m+n \tag{6.8}$$

zugeordnet. Die Variablen $x_{n+1}, \ldots, x_{n+m}$ heißen künstliche Variablen. Zur Bestimmung der ersten zulässigen Basislösung von (6.3) – (6.5) wird eine Zweiphasenmethode verwendet:

- *1. Phase.* Wähle als erste zulässige Basislösung für (6.6) – (6.8)

$$x_i = 0, \; i = 1, \ldots, n, \quad x_{n+i} = b_i, \; i = 1, \ldots, m.$$

- *2. Phase.* Löse (6.6) – (6.8) mit der Simplexmethode.

Es stellt sich nun die Frage, ob man auf diesem Wege schließlich eine erste zulässige Basislösung für (6.3) – (6.5) erhält. □

Im nächsten Satz wird gezeigt, dass eine Lösung von (6.6) – (6.8) existiert, falls man Ausartung ausschließt.

Lemma 6.6 *Unter der Annahme, dass (6.6) – (6.8) keine ausgearteten Basislösungen besitzt, liefert die Simplexmethode nach endlich vielen Schritten eine optimale Lösung des linearen Programms (6.6) – (6.8).*

Beweis: Da Ausartung per Annahme ausgeschlossen ist, kann kein Basiszyklus auftreten. Somit verringert die Simplexmethode in jedem Schritt den Zielfunktionswert. Es ist dann nur noch die Beschränktheit von unten der Zielfunktion (6.6) über (6.7) bis (6.8) zu zeigen. Das ist offensichtlich, da (6.6) eine Summe nichtnegativer reeller Zahlen ist, die durch Null nach unten beschränkt ist. ■

Nun wird eine Bedingung angegeben, mit welcher man aus dem Optimum des Hilfsproblems (6.6) – (6.8) eine erste zulässige Basislösung von (6.3) – (6.5) erhält.

Satz 6.7 *Sei $\tilde{\mathbf{x}}_0$ eine Optimallösung der künstlichen Aufgabe (6.6) – (6.8) mit dem zugehörigen Zielfunktionswert $\tilde{z}_0$. Gilt $\tilde{z}_0 = 0$, so sind die ersten n Komponenten von $\tilde{\mathbf{x}}_0$ eine zulässige Basislösung der Aufgabe (6.3) – (6.5). Gilt jedoch $\tilde{z}_0 > 0$, so besitzt (6.3) – (6.5) keine zulässige Basislösung.*

Beweis: Aus $\tilde{z}_0 = 0$ folgt $x_{n+i} = 0, i = 1, \ldots, m$, das heißt im Optimum verschwinden alle künstlichen Variablen. Also hat $\tilde{\mathbf{x}}_0$ die Gestalt

$$\tilde{\mathbf{x}}_0 = (x_1, \ldots, x_n, \underbrace{0, \ldots, 0}_{m})^T.$$

Da $\tilde{\mathbf{x}}_0$ mit der Simplexmethode konstruiert wurde, folgt dass $\tilde{\mathbf{x}}_0$ eine zulässige Basislösung von (6.3) – (6.5) ist.

Sei nun $\tilde{z}_0 > 0$. Der Beweis wird indirekt geführt, das heißt, wir nehmen an, dass (6.3) – (6.5) die zulässige Basislösung $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_n)^T$ besitzt. Dann besitzt jedoch (6.6) – (6.8) die zulässige Basislösung $(\bar{x}_1, \ldots, \bar{x}_n, 0, \ldots, 0)^T$ mit dem zugehörigen Zielfunktionswert (6.6) $\bar{z} = 0$. Das ist im Widerspruch zur Annahme dass $\tilde{z}_0$ der minimale Wert ist. ■

Bemerkung 6.8 4. Fall, M–Methode. Die M–Methode. Es wird das lineare Programm (6.3) – (6.5) betrachtet und diesem die folgende Hilfsaufgabe zugeordnet

$$\sum_{j=1}^{n} c_j x_j + M \sum_{i=1}^{m} x_{n+i} \quad \rightarrow \quad \min \ ! \tag{6.9}$$

$$\sum_{j=1}^{n} a_{ij} x_j + x_{n+i} \quad = \quad b_i \quad i = 1, \ldots, m \ , \tag{6.10}$$

$$\mathbf{x} \quad \geq \quad \mathbf{0}. \tag{6.11}$$

Bei dieser Aufgabe muss der Straffaktor $M > 0$ hinreichend groß gewählt werden, damit im Optimum die Variablen $x_{n+1}, \ldots, x_{n+m}$ verschwinden. Die Schwierigkeit besteht darin, dass man im allgemeinen nicht von vornherein festlegen kann, wie groß M zu wählen ist. □

Möglich sind Aussagen folgender Gestalt:

Satz 6.9 *Es existiert ein $M_0 > 0$, so dass für alle $M > M_0$ aus der Lösbarkeit von (6.3) – (6.5) die Lösbarkeit von (6.9) – (6.11) mit $x_{n+1} = \ldots = x_{n+m} = 0$ folgt.*

Beweis: Siehe Literatur. ■

Der Vorteil der M–Methode im Vergleich zur Herangehensweise von Fall 3 wird mit folgendem Satz beschrieben.

Satz 6.10 *Falls (6.9) – (6.11) eine Lösung*

$$\tilde{\mathbf{x}} = (x_1, \ldots, x_n, \underbrace{0, \ldots, 0}_{m})^T$$

besitzt, so ist $\mathbf{x} = (x_1, \ldots, x_n)^T$ bereits eine Optimallösung von (6.3) – (6.5).

Beweis: Das sieht man durch Einsetzen von $\tilde{\mathbf{x}}$ in (6.9) – (6.11). ■

Kapitel 7

Zur Ausartung

Bemerkung 7.1 Nach Definition 3.4 liegt Ausartung dann vor, wenn mindestens eine der Variablen $x_i, i = 1 \ldots, m$, einer zulässigen Basislösung verschwindet. Das dahinterliegende Problem ist, dass die Zuordnung Ecke – zulässige Basislösung nicht eindeutig ist. Eine Ecke des Polyeders kann Basislösung zu verschiedenen Basen sein. Das kann aber nur bei ausgearteten Basislösungen auftreten. □

Beispiel 7.2 Betrachte das lineare Programm mit

$$z = x_1 + x_2 - x_3 \to \min\ !, \quad A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Der einzige Extremalpunkt ist $\mathbf{x} = (0, 0, 1)^T$. Zulässige Basen sind

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ und } \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Der Grund für die Nichteindeutigkeit der Basis besteht darin, dass es zu viele Nebenbedingungen gibt, die den Extremalpunkt bestimmen. In diesem Beispiel ist er durch die beiden Gleichungen

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ und } \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

gleichermaßen gegeben. Das haben wir bereits in den Beispielen 3.5 (2. Teil) und 5.7 gesehen. □

Bemerkung 7.3

- In der Praxis stellt sich heraus, dass die meisten zu lösenden linearen Programme ausgeartet sind.
- In der Simplexmethode ist es möglich, dass im Falle der Ausartung der zulässigen Basislösung nur ein Basiswechsel stattfindet, siehe Beispiel 5.7. Das kann zu einem unendlichen Zyklus werden, einem sogenannten Basiszyklus. Es ist jedoch möglich, Ausartung prinzipiell auszuschließen, beispielsweise mit der Methode der ε–Störung, beziehungsweise einen Basiszyklus zu umgehen, mit der lexikographischen Simplexmethode.

□

7.1 Die Methode der ε–Störung

Bemerkung 7.4 Herangehensweise. Wir betrachten das lineare Programm

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \mathbf{c}^T \mathbf{x} \ : \ A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \right\}. \tag{7.1}$$

Sei $\mathbf{x} = (x_1, \ldots, x_m, 0, \ldots, 0)^T$ eine zulässige Basislösung mit den Basisvektoren $\mathbf{a}_1, \ldots, \mathbf{a}_m$:

$$\begin{aligned} \mathbf{a}_1 x_1 + \ldots + \mathbf{a}_m x_m &= \mathbf{b}, \\ \mathbf{a}_1 x_{1j} + \ldots + \mathbf{a}_m x_{mj} &= \mathbf{a}_j, \quad j = 1, \ldots, n. \end{aligned} \tag{7.2}$$

Sei $\varepsilon > 0$ vorgegeben und sei $A_B = (\mathbf{a}_1, \ldots, \mathbf{a}_m)$ die Matrix der Basisvektoren. Dann betrachtet man anstelle (7.1) ein lineares Programm mit gestörten Nebenbedingungen

$$\begin{aligned} \mathbf{c}^T \mathbf{x} &\rightarrow \text{min !} \\ A_B \mathbf{x} + \sum_{j=1}^{n} \varepsilon^j (7.2) &= \mathbf{b} \quad \Longrightarrow \\ A_B \mathbf{x} + \sum_{j=1}^{n} \varepsilon^j \mathbf{a}_j &= \mathbf{a}_1 \left(x_1 + \sum_{j=1}^{n} x_{1j} \varepsilon^j \right) + \ldots + \mathbf{a}_m \left(x_m + \sum_{j=1}^{n} x_{mj} \varepsilon^j \right) = \mathbf{b}. \end{aligned} \tag{7.3}$$

Mit den Nebenbedingungen (7.3) hat man für hinreichend kleines ε die zulässige Basislösung

$$x_i^{(\varepsilon)} := x_i + \sum_{j=1}^{n} x_{ij} \varepsilon^j = x_i + \varepsilon^i + \sum_{j=m+1}^{n} x_{ij} \varepsilon^j,$$

da für $i = 1, \ldots, m$, gilt $x_{ij} = \delta_{ij}$. Die Eigenschaft der Basislösung folgt daraus, dass die Basis nicht geändert wurde und die Nebenbedingung in (7.3) erfüllt ist. Die Zulässigkeit folgt für hinreichend kleines ε aus $\varepsilon^i > 0$ und $\varepsilon^i \gg \varepsilon^j$ für $i < j$

$$\varepsilon^i > \left| \sum_{j=m+1}^{n} x_{ij} \varepsilon^j \right| \quad \Longrightarrow \quad x_i^{(\varepsilon)} > 0.$$

Die Basislösung $\mathbf{x}^{(\varepsilon)}$ ist also für hinreichend kleine ε nicht ausgeartet. Der zugehörige Zielfunktionswert ist

$$\begin{aligned} z_0^{(\varepsilon)} &= \sum_{i=1}^{m} c_i x_i + \sum_{i=1}^{m} c_i \left(\sum_{j=1}^{n} x_{ij} \varepsilon^j \right) = \sum_{i=1}^{m} c_i x_i + \sum_{j=1}^{n} \left(\sum_{i=1}^{m} c_i x_{ij} \right) \varepsilon^j \\ &= \sum_{i=1}^{m} c_i x_i + \sum_{j=1}^{n} z_j \varepsilon^j. \end{aligned}$$

Im Bild 7.1 wird die Störung der Nebenbedingungen graphisch veranschaulicht. Im dicken Punkt schneiden sich drei Geraden. Das führt dazu, dass die Zuordnung Ecke – Basislösung nicht eindeutig ist. Man hat Ausartung. Durch die Störung der Nebenbedingungen (durchgezogene Geraden) erreicht man, dass es nur noch Schnittpunkte mit genau zwei Geraden gibt. □

Bemerkung 7.5 Berechnung von θ. In der Simplexmethode benötigt man die Größe θ, siehe (4.8). Sei $z_k - c_k > 0$. Dann berechnet sich θ in der Methode der ε–Störung durch

$$\theta = \min_{i=1,\ldots,m; x_{ik}>0} \frac{x_i^{(\varepsilon)}}{x_{ik}} = \min_{i=1,\ldots,m; x_{ik}>0} \frac{x_i + \varepsilon^i + \sum_{j=m+1}^{n} x_{ij} \varepsilon^j}{x_{ik}}. \tag{7.4}$$

□

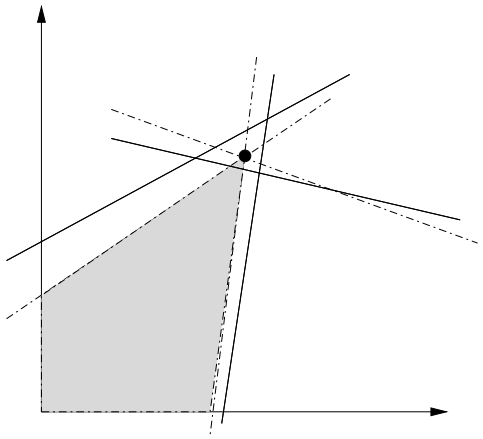


Abbildung 7.1: Veranschaulichung der Störung der Nebenbedingungen.

Satz 7.6 *Sei* $\mathbf{x} = (x_1, \ldots, x_m, 0, \ldots, 0)^T$ *eine zulässige Basislösung der Originalaufgabe (7.1). Falls*

$$\theta = \min_{i=1,\ldots,m; x_{ik}>0} \frac{x_i}{x_{ik}} = 0$$

gilt (Ausartung), dann gibt es ein $\bar{\varepsilon} > 0$ *dergestalt, dass*

$$\theta = \min_{i=1,\ldots,m; x_{ik}>0} \frac{x_i^{(\varepsilon)}}{x_{ik}} = \frac{x_l^{(\varepsilon)}}{x_{lk}} > 0 \quad \forall\, \varepsilon \in (0, \bar{\varepsilon}) \tag{7.5}$$

und der Index l *ist im gestörten Problem (7.3) eindeutig bestimmt.*

Beweis: Aus Bemerkung 7.4 folgt $x_i^{(\varepsilon)} > 0$, $i = 1, \ldots, m$, und damit $\theta > 0$ in (7.5) für $\varepsilon \in (0, \bar{\varepsilon})$.

Die Eindeutigkeit von l wird indirekt bewiesen. Sei l also nicht eindeutig bestimmt, das heißt es gibt zwei Indizes $l_1 \neq l_2$, $1 \leq l_1, l_2 \leq m$, mit

$$\frac{x_{l_1} + \varepsilon^{l_1} + \sum_{j=m+1}^{n} x_{l_1 j}\varepsilon^j}{x_{l_1 k}} = \frac{x_{l_2} + \varepsilon^{l_2} + \sum_{j=m+1}^{n} x_{l_2 j}\varepsilon^j}{x_{l_2 k}}$$

für alle $\varepsilon \in (0, \bar{\varepsilon})$. Die beiden Terme sind Polynome in ε. Diese sind genau dann gleich für alle $\varepsilon \in (0, \bar{\varepsilon})$, wenn sie in allen Koeffizienten übereinstimmen. Insbesondere müssen die Koeffizienten vor den Termen mit Potenz l_1 und l_2 gleich sein. Ist $l_1 \neq l_2$, so ist für den linken Term der Koeffizient vor ε^{l_1} ungleich Null und für den rechten Term gleich Null. Für den Koeffizienten vor ε^{l_2} gilt sinngemäß das gleiche. Das heißt, diese Koeffizienten können nur dann gleich sein, wenn $l_1 = l_2$, im Widerspruch zur Annahme. ■

Prinzipiell kann diese Manipulation in jedem Simplexschritt durchgeführt werden und man kann damit sichern, dass l stets eindeutig bestimmt ist. Diese Vorgehensweise ist für jeden Eckpunkt des zulässigen Bereichs ausgeführt zu denken. Da die Anzahl der Eckpunkte endlich ist, erhält man folgenden Satz.

Satz 7.7 *Zu jedem linearen Optimierungsproblem existiert bei geeigneter Wahl von* $\varepsilon \in (0, \varepsilon^*)$ *ein gestörtes lineares Optimierungsproblem, so dass dieses keine ausgeartete zulässige Basislösung besitzt. Für* $\varepsilon \to 0$ *konvergiert das Optimum des gestörten Problems (7.3) zum Optimum des Originalproblems (7.1).*

Bemerkung 7.8 Praktische Umsetzung der Methode der ε–Störung. Trotz dieser schönen Theorie macht man das alles bei praktischen Problemen nicht. Für

diese wird vorgeschlagen: Falls in einer zulässigen Basislösung wenigstens ein Wert $x_i = 0$ bestimmt wurde, so kann $\theta = 0$ sein. Wähle dann

$$l = \min_{x_{ik}>0} \{i \ : \ x_i = 0\},$$

wobei i über alle Basisvariablen läuft und k der Index der festgelegten Hauptspalte ist, und transformiere mit diesem Index l. Theoretisch besteht die Gefahr eines Zyklus, in der Praxis ist das aber eher unwahrscheinlich. □

7.2 Die lexikographische Simplexmethode

Bemerkung 7.9 Idee. Bei der lexikographischen Simplexmethode erfolgt die Auswahl der zu tauschenden Basisvektoren so, dass keine Wiederholungen auftreten können. Mit dieser Vorgehensweise wird nicht die Ausartung behoben, sondern es wird verhindert, dass Basiszyklen auftreten. □

Definition 7.10 Lexikopositiver Vektor. Ein Vektor $\mathbf{x} \in \mathbb{R}^n$ heißt lexikopositiv, falls $\mathbf{x} = (0, \ldots, 0, x_i, x_{i+1}, \ldots, x_n)^T$ mit $i \geq 1$ und $x_i > 0$. Das heißt, die erste von Null verschwindende Komponente ist positiv. Die Schreibweise ist

$$\mathbf{x} >_l \mathbf{0}.$$

Sei $\mathbf{y} \in \mathbb{R}^n$. Dann ist $\mathbf{y} >_l \mathbf{x}$ genau dann, wenn $\mathbf{y} - \mathbf{x} >_l \mathbf{0}$. □

Bemerkung 7.11 Die lexikographische Simplexmethode. Wir betrachten das lineare Programm (7.1) mit $\mathrm{rg}(A) = m$. Sei

$$\mathbf{x}_B = (x_1, \ldots, x_m, 0, \ldots, 0)^T$$

eine zulässige Basislösung. Die zugehörige Matrix der Basisvektoren sei $A_B \in \mathbb{R}^{m\times m}$ und die der Nichtbasisvektoren A_N. Dann sind die Zeilen der Matrix

$$(\bar{\mathbf{b}}, \bar{A}) := A_B^{-1}(\mathbf{b}, A) = A_B^{-1}\,(\mathbf{b}, A_B, A_N) \in \mathbb{R}^{m\times(n+1)}$$

lexikopositiv, da

$$A_B^{-1}(\mathbf{b}, A) = (\mathbf{x}_B, I_m, \bar{\mathbf{a}}_{m+1}, \ldots, \bar{\mathbf{a}}_n)\,,$$

$\mathbf{x}_B \geq \mathbf{0}$ und I_m die Einheitsmatrix des $\mathbb{R}^{m\times m}$ ist. Falls die Basisvariablen nicht die ersten m Variablen sind, dann ordnet man sie nach vorn.

Anstelle von (4.8) wird bei der lexikographischen Simplexmethode der Index l durch

$$\theta = \min_{>_l; i=1,\ldots,m, x_{ik}>0} \frac{\mathbf{e}_i^T(\bar{\mathbf{b}}, \bar{A})}{x_{ik}} =: \frac{\mathbf{e}_l^T(\bar{\mathbf{b}}, \bar{A})}{x_{lk}}$$

bestimmt, wobei $\mathbf{e}_i \in \mathbb{R}^m$ der Einheitsvektor ist, der in der i–ten Komponente eine Eins hat. Das heißt, das Minimum wird bezüglich der lexikographischen Ordnung genommen. Das obige Symbol bedeutet, dass man sich wie üblich alle Einträge mit $x_{ik} > 0$ ansieht, die zugehörigen Vektoren $\mathbf{e}_i^T(\bar{\mathbf{b}}, \bar{A})$ bildet, durch x_{ik} dividiert und von den so erhaltenen Vektoren den lexikographisch kleinsten nimmt, und von diesem die erste von Null verschiedene Komponente, um l zu bestimmen. Es gilt, siehe beispielsweise [JS04]:

- Falls l in der allgemeinen Simplexmethode (4.8) eindeutig bestimmt ist, erhält man bei der lexikographischen Simplexmethode den gleichen Index.
- Die lexikographische Simplexmethode definiert einen eindeutigen Index l. Man kann zeigen, dass eine Nichteindeutigkeit im Widerspruch zu $rg(A) = m$ steht.

- Das Ergebnis eines lexikographischen Simplexschrittes ist wiederum eine lexikopositive Basis. Die Basislösung ist also insbesondere zulässig.
- Bei der neuen Basislösung ist entweder der Zielfunktionswert kleiner oder die Differenz der Koeffizienten der Zielfunktion der neuen und der alten Basis ist lexikopositiv. Im ersten Fall hat man die Ecke verlassen. Im zweiten Fall kann es bei weiteren lexikographischen Simplexschritten nicht passieren, dass die alte Basis noch einmal verwendet wird. Ein Basiszyklus ist ausgeschlossen.

Bei der lexikographischen Simplexmethode werden also ausgehend von einer lexikopositiven Startlösung weitere lexikopositive Lösungen erzeugt. Dieses Verfahren ist endlich. Es bricht entweder mit einer Lösung des Optimierungsproblems ab, oder es wird gefunden, dass die Zielfunktion nicht nach unten beschränkt ist. Die Anzahl der Schritte kann $n!$ nicht übersteigen. Diese Schranke ist allerdings für größere Werte von n für die Praxis bedeutungslos. □

Kapitel 8

Zur Effizienz der Simplexmethode

Bemerkung 8.1 Fragestellung. Die Simplexmethode ist ein Verfahren zur Bestimmung der Lösung des linearen Programms

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \mathbf{c}^T\mathbf{x} \ : \ A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \right\}.$$

Für ihre praktische Anwendung muss untersucht werden, wie teuer die Berechnung des Optimums ist. Dazu unterscheidet man zwei Situationen:

- Analyse des schlimmsten Falls, der bei der Lösung eines linearen Programms mit der Simplexmethode auftreten kann, *worst case* Modell,
- Analyse des in der Praxis zu erwartenden normalen Falls, der bei der Lösung eines linearen Programms mit der Simplexmethode auftreten kann, *real world* Modell.

Die zweite Situation ist an sich interessanter. Das Problem besteht darin, ein real world Modell aufzustellen. Diese Frage ist bis heute teilweise ungeklärt. In der praktischen Anwendung der Simplexmethode hat man jedoch die Erfahrung gewonnen, dass sie im Normalfall hervorragend funktioniert. Wir werden hier nur das worst case Modell untersuchen. □

8.1 Maße für die Effizienz

Das Grundanliegen besteht darin, den Aufwand zur Abarbeitung eines numerischen Verfahrens in Abhängigkeit vom Umfang der Eingangsdaten abzuschätzen.

Definition 8.2 compl(A,B). Die Komplexität compl(A,B) eines Algorithmus A zur Lösung von Aufgaben B (eines Problemkreises P) ist die Anzahl der elementaren Operationen auf einem Computer oder die benötigte Rechenzeit in Abhängigkeit vom Umfang der Eingangsdaten. □

Der Wunsch ist natürlich, einen effizienten Algorithmus für jedes praxisrelevante Optimierungsproblem zu konstruieren. Das geht aber im allgemeinen nicht, da schwierige und auch unlösbare Probleme existieren.

Definition 8.3 P(d). Bezeichne P ein Problem und d den Umfang seiner Eingangsdaten. Dann beschreibt P(d) die Menge aller Aufgaben P mit gleichem Umfang d der Eingangsdaten. □

Definition 8.4 Worst case Komplexität eines Algorithmus. Die worst case Komplexität eines Algorithmus A zur Lösung eines Problems P ist gegeben durch

$$\text{w–compl}(A, P) = \sup_{B \in P(d)} \text{compl}(A, B).$$

Man kann entsprechend eine *average case Komplexität* von A bezüglich P erklären

$$\text{a–compl}(A, P) = \text{Erwartungswert}_{B \in P(d)} \text{compl}(A, B).$$

□

Definition 8.5 Worst case Komplexität eines Problems. Sei A_P die Menge aller Algorithmen zur Lösung eines Problems P. Die worst case Komplexität von P ist erklärt durch

$$\text{w–compl}(P) = \min_{A \in A_P} \text{w–compl}(A, P).$$

□

Die Komplexität wird im allgemeinen als Funktion der Menge der Eingangsdaten in der Form $\mathcal{O}(f(d))$ angegeben.

Beispiel 8.6 Matrizenmultiplikation. Gegeben seien zwei Matrizen $A, B \in \mathbb{R}^{n \times n}$ und gesucht ist das Produkt $C = AB$. Ein Eintrag von C berechnet sich wie folgt

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},$$

benötigt also n Multiplikationen und $(n-1)$ Additionen, dass heißt $\mathcal{O}(n)$ Operationen. Die Anzahl der zu berechnenden Einträge von C ist n^2. Somit hat man insgesamt $\mathcal{O}(n^3)$ Operationen durchzuführen.

Die Frage ist, ob der Aufwand von $\mathcal{O}(n^3)$ optimal ist. Da man insgesamt n^2 Größen zu berechnen hat, kann der minimale Aufwand der Matrizenmultiplikation nicht kleiner als $\mathcal{O}(n^2)$ sein. Man kennt heute Verfahren, deren Aufwand für große n wie $\mathcal{O}(n^{2.38})$ ist. (SIAM News 38, Vol. 9, 2005) □

Definition 8.7 Gutartiges Problem. Probleme mit polynomialer Komplexität $\mathcal{O}(d^z)$, $z \in \mathbb{R}$, heißen gutartig, Probleme mit exponentieller Komplexität $\mathcal{O}(z^d)$, $z \in \mathbb{R}, z > 1$, bösartig. □

8.2 Zur worst case Komplexität der Simplexmethode

Bemerkung 8.8 Konstruktion eines Beispielproblems. In diesem Abschnitt wird ein Beispiel konstruiert, bei welchem der Aufwand der Simplexmethode exponentiell wächst, wobei der Aufwand in der Anzahl der Schritte gemessen wird. Das bedeutet, dass die Simplexmethode theoretisch ein sehr ineffizientes Verfahren sein kann. Dieser Fall ist in der Praxis glücklicherweise faktisch nicht zu beobachten.

Bei der worst case Komplexität wird der schlechteste Fall betrachtet. Für die Simplexmethode bedeutet das, dass die schlechteste Wahl der Hauptspalte bezüglich der Anzahl der zulässigen Basislösungen betrachtet wird, die man beim Transformationsprozess erzeugt.

Wie betrachten den n–dimensionalen Einheitswürfel $[0, 1]^n$. Dieser hat 2^n Ecken. Die Koordinaten des Einheitswürfels werden nun gestört

$$\begin{aligned} \varepsilon &\leq x_1 \leq 1 \quad \text{mit } 0 < \varepsilon < 1/2, \\ \varepsilon x_{j-1} &\leq x_j \leq 1 - \varepsilon x_{j-1} \leq 1 \quad j = 2, \ldots, n. \end{aligned}$$

Über diesem gestörten Würfel wird folgendes lineares Programm definiert:

$$\begin{aligned} -x_n &\to \text{min !} \\ x_1 - r_1 &= \varepsilon \\ x_1 + s_1 &= 1 \\ x_j - \varepsilon x_{j-1} - r_j &= 0 \quad j = 2, \ldots, n, \\ x_j + \varepsilon x_{j-1} + s_j &= 1 \quad j = 2, \ldots, n, \\ x_j, r_j, s_j &\geq 0 \quad j = 1, \ldots, n. \end{aligned} \tag{8.1}$$

Ordnet man die Unbekannten gemäß $(x_1, \ldots, x_n, r_1, \ldots, r_n, s_1, \ldots, s_n)$, dann hat die Matrix der Nebenbedingungen die Gestalt

$$A = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & -1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ -\varepsilon & 1 & 0 & \ldots & 0 & 0 & -1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & -1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ \varepsilon & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \in \mathbb{R}^{2n \times 3n}.$$

□

Lemma 8.9 *Die Menge der zulässigen Basislösungen von (8.1) ist die Klasse der Untermengen von*

$$(x_1, \ldots, x_n, r_1, \ldots, r_n, s_1, \ldots, s_n),$$

bei denen alle $x_j > 0$, $j = 1, \ldots, n$, und entweder $r_j > 0$ oder $s_j > 0$ für jedes $j = 1, \ldots, n$. Alle Basislösungen sind nicht ausgeartet.

Beweis: Es wird erst die Zulässigkeit untersucht, dann die Basislösungseigenschaft.

Zuerst wird gezeigt, dass für zulässige Lösungen alle x_j, $j = 1, \ldots, n$ positiv sein müssen. Ist $x_1 = 0$, dann folgt aus der ersten Nebenbedingung $r_1 = -\varepsilon < 0$. Ein Vektor mit $x_1 = 0$ kann also nicht zulässig sein. Der Beweis erfolgt nun durch Induktion. Seien $x_j > 0$ für $j = 1, \ldots, k-1$ und $x_k = 0$. Dann folgt aus den Nebenbedingungen

$$0 = x_k = r_k + \underbrace{\varepsilon x_{k-1}}_{>0} \quad \Longrightarrow \quad r_k < 0,$$

im Widerspruch zur letzten Nebenbedingung. Also sind alle x_j positiv, sie können damit keine Nichtbasisvariablen sein.

Nun werden die r_j, s_j betrachtet. Gelte für ein j, dass $r_j = s_j = 0$. Für $j = 1$ folgt dann aus den ersten beiden Nebenbedingungen von (8.1) $x_1 = \varepsilon = 1$. Das steht im Widerspruch zur Wahl von ε. Sei $j > 1$. Dann gelten

$$\begin{aligned} x_j - \varepsilon x_{j-1} &= 0, \\ x_j + \varepsilon x_{j-1} &= 1. \end{aligned}$$

Subtraktion dieser Gleichungen ergibt

$$2\varepsilon x_{j-1} = 1.$$

Da jedoch $\varepsilon < 1/2$ und $x_{j-1} \leq 1$ ist die linke Seite echt kleiner als 1. Damit sind r_j oder s_j für jedes $j = 1, \ldots, n$, positiv. Da man genau $2n$ Basisvariablen hat und bereits n davon durch die x_j gegeben sind, ist entweder r_j oder s_j für jedes $j = 1, \ldots, n$, positiv.

Die Basiseigenschaft der soeben konstruierten Menge sieht man durch Umordnung der Zeilen der Matrix $A_{2n,2n}$. Für jeden Index j vertauscht man die Zeilen j und $j + n$ falls

$r_j > 0, s_j = 0$. Damit wird die Matrix auf Dreiecksform gebracht, wobei in der Diagonalen ± 1 stehen. Ihre Determinante ist somit ebenfalls ± 1.

Die Basislösungen sind nicht ausgeartet, weil die $2n$ Basisvariablen alle positiv sind. ■

Bemerkung 8.10 Die Menge der Indizes j für die die Basislösungen $r_j > 0$ erfüllen wird mit S bezeichnet. Sind dies beispielsweise die Indizes $1, 3, 7$, so ist $S = \{1, 3, 7\}$. Die zugehörigen Basislösungen werden als $\mathbf{x}^{(S)}$ geschrieben, im Beispiel $\mathbf{x}^{(\{1,3,7\})}$. Der Wert x_j in $\mathbf{x}^{(S)}$ wird mit $x_j^{(S)}$ bezeichnet. Wegen der Zielfunktion betrachten wir jetzt insbesondere den letzten Index n. □

Lemma 8.11 *Seien $n \in S$ und $n \notin S'$. Dann ist $x_n^{(S)} > x_n^{(S')}$. Falls außerdem $S' = S \setminus \{n\}$ gilt, folgt $x_n^{(S')} = 1 - x_n^{(S)}$.*

Beweis: Sei $n \in S$, das heißt $r_n > 0, s_n = 0$. Dann folgt aus

$$x_n^{(S)} + \varepsilon x_{n-1}^{(S)} + s_n = 1 \quad \Longrightarrow \quad x_n^{(S)} = 1 - \varepsilon x_{n-1}^{(S)} > 1/2$$

wegen $x_{n-1}^{(S)} \leq 1$, $\varepsilon < 1/2$.

Andererseits gilt für $n \notin S'$, dass $r_n = 0$. Mit denselben Argumenten folgt

$$x_n^{(S')} - \varepsilon x_{n-1}^{(S')} - r_n = 0 \quad \Longrightarrow \quad x_n^{(S')} = \varepsilon x_{n-1}^{(S')} < 1/2.$$

Die Mengen in der zweiten Aussage des Lemmas unterscheiden sich nur dadurch, dass in S gilt $r_n > 0, s_n = 0$ und in S' gilt $r_n = 0, s_n > 0$. Da alle anderen Indizes in S und S' gleich sind und die Nebenbedingungen für die Indizes kleiner n nicht von x_n, r_n, s_n abhängen, gilt insbesondere

$$x_{n-1}^{(S)} = x_{n-1}^{(S')}.$$

Da $r_n = 0$ für S' und $s_n = 0$ für S ist, folgt

$$x_n^{(S')} = \varepsilon x_{n-1}^{(S')} = 1 - \left(1 - \varepsilon x_{n-1}^{(S')}\right) = 1 - \left(1 - \varepsilon x_{n-1}^{(S)}\right) = 1 - x_n^{(S)}.$$

■

Lemma 8.12 *Die Untermengen von $\{1, 2, \ldots, n\}$ seien so geordnet, dass*

$$x_n^{(S_1)} \leq x_n^{(S_2)} \leq \ldots \leq x_n^{(S_{2^n})}$$

gilt. Diese Ungleichungen sind scharf, dass heißt es gilt $<$, und die zulässigen Basislösungen $x^{(S_j)}$ und $x^{(S_{j+1})}$ sind benachbart für $j = 1, \ldots, 2^n - 1$, das heißt, sie unterscheiden sich nur in einem Basisvektor.

Beweis: Der Beweis erfolgt durch Induktion über die Dimension n.

Induktionsanfang. $n = 1$. Man hat zwei Eckpunkte. Aus

$$x_1 - r_1 = \varepsilon, \quad x_1 + s_1 = 1$$

folgt

$$(x_1, r_1, s_1) = (\varepsilon, 0, 1 - \varepsilon) \vee (1, 1 - \varepsilon, 0).$$

Diese Punkte sind natürlich benachbart und die Schärfe der Ungleichung wurde bereits im letzten Lemma bewiesen ($x_n^{(S)} > x_n^{(S')}$).

Induktionsannahme. Für einen n–Würfel sei alles korrekt. Die entsprechende Numerierung sei $S_1, \ldots, S_{2^n}$.

Induktionsschritt. Man betrachtet jetzt $\{1, 2, \ldots, n+1\}$. Offensichtlich gelten $S_j \subset \{1, 2, \ldots, n+1\}$ und $n+1 \notin S_j$, $j = 1, \ldots, 2^n$. Damit ist $r_{n+1}^{(S_j)} = 0$. Aus der entsprechenden Nebenbedingung folgt

$$x_{n+1}^{(S_j)} = \varepsilon x_n^{(S_j)}.$$

Nach Induktionsannahme ist

$$x_n^{(S_1)} < x_n^{(S_2)} < \ldots < x_n^{(S_{2^n})},$$

woraus nun folgt (Durchmultiplizieren mit ε)

$$x_{n+1}^{(S_1)} < x_{n+1}^{(S_2)} < \ldots < x_{n+1}^{(S_{2^n})}. \tag{8.2}$$

Die Reihenfolge dieser Untermengen bleibt also erhalten.

Nun betrachten wir die Lösungen mit $r_{n+1}^{(S_j)} > 0$. Sei $S_j' = S_j \cup \{n+1\}$. Nach Lemma 8.11 ist

$$x_{n+1}^{(S_j)} = 1 - x_{n+1}^{(S_j')} \quad \Longrightarrow \quad x_{n+1}^{(S_j')} = 1 - x_{n+1}^{(S_j)} \quad j = 1, \ldots, 2^n, \tag{8.3}$$

und speziell

$$x_{n+1}^{(S_{2^n}')} > x_{n+1}^{(S_{2^n})}. \tag{8.4}$$

Aus (8.2), (8.3) und (8.4) resultiert

$$x_{n+1}^{(S_1)} < \ldots < x_{n+1}^{(S_{2^n})} < x_{n+1}^{(S_{2^n}')} < \ldots < x_{n+1}^{(S_1')}.$$

Nun muss noch die Nachbarschaft der Basislösungen bewiesen werden. Nach Induktionsannahme, sind $\mathbf{x}^{(S_1)}, \ldots, \mathbf{x}^{(S_{2^n})}$ in n Dimensionen benachbart. In der $(n+1)$–sten Dimension, erhalten diese Basislösungen alle noch den Spaltenvektor von s_{n+1}. Sie bleiben damit benachbart. Die Basislösungen $\mathbf{x}^{(S_{2^n})}$ und $\mathbf{x}^{(S_{2^n}')}$ unterscheiden sich nur im Basisvektor von s_{n+1} beziehungsweise r_{n+1}. Sie sind also auch benachbart. Die Basislösungen $\mathbf{x}^{(S_{2^n}')}, \ldots, \mathbf{x}^{(S_1')}$ sind benachbart, weil $\mathbf{x}^{(S_1)}, \ldots, \mathbf{x}^{(S_{2^n})}$ benachbart sind. ■

Satz 8.13 *Für jedes $n \geq 1$ existiert ein lineares Programm, bestehend aus $2n$ Gleichungen und $3n$ Variablen, so dass die Simplexmethode $2^n - 1$ Iterationsschritte braucht, um das Optimum zu bestimmen.*

Die Struktur dieses linearen Programms kann so eingerichtet werden, dass alle Koeffizienten (zum Beispiel) ≤ 4 sind.

Beweis: Der erste Teil des Satzes wird durch das angegebene Beispiel (8.1) bewiesen. In Lemma 8.12 sind die 2^n verschiedenen zulässigen Basislösungen streng geordnet. Die Simplexmethode wird so angewendet, dass mit jeder Transformation nur die jeweils geringste Verbesserung des Zielfunktionswertes erreicht wird. Bei $\mathbf{x}_n^{(1)}$ beginnend, sind somit $2^n - 1$ Transformationen nötig.

Für den zweiten Teil des Satzes wähle man $\varepsilon = 1/4$. ■

Beispiel 8.14 Wir betrachten das in diesem Abschnitt studierte Problem (8.1) für $n = 3$.

$$\begin{array}{rcl} -x_3 & \rightarrow & \min\ ! \\ \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -\varepsilon & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -\varepsilon & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \varepsilon & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & \varepsilon & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ r_1 \\ r_2 \\ r_3 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} & = & \begin{pmatrix} \varepsilon \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \\ \mathbf{x}, \mathbf{r}, \mathbf{s} & \geq & \mathbf{0}. \end{array}$$

Die Ecken des Würfels, des gestörten Würfels und die Zielfunktionswerte sind wie folgt

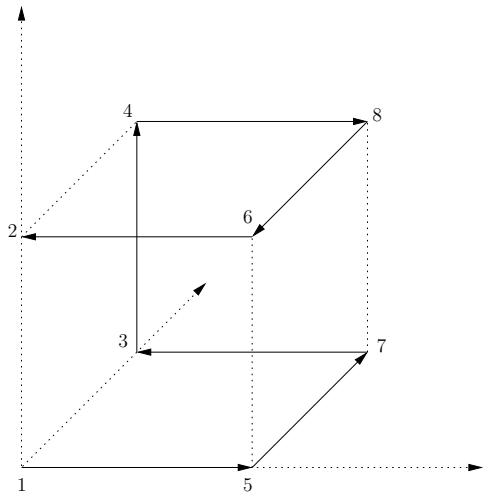
Nr.	Würfel	gest. Würfel	z	z für $\varepsilon = 1/4$	Reihenfolge
1	$(0,0,0)$	$(\varepsilon, \varepsilon^2, \varepsilon^3)$	$-\varepsilon^3$	-0.015625	8
2	$(0,0,1)$	$(\varepsilon, \varepsilon^2, 1-\varepsilon^3)$	$-1+\varepsilon^3$	-0.984375	1
3	$(0,1,0)$	$(\varepsilon, 1-\varepsilon^2, \varepsilon-\varepsilon^3)$	$-\varepsilon+\varepsilon^3$	-0.234375	5
4	$(0,1,1)$	$(\varepsilon, 1-\varepsilon^3, 1-\varepsilon+\varepsilon^3)$	$-1+\varepsilon-\varepsilon^3$	-0.765625	4
5	$(1,0,0)$	$(1, \varepsilon, \varepsilon^2)$	$-\varepsilon^2$	-0.0625	7
6	$(1,0,1)$	$(1, \varepsilon, 1-\varepsilon^2)$	$-1+\varepsilon^2$	-0.9375	2
7	$(1,1,0)$	$(1, 1-\varepsilon, \varepsilon-\varepsilon^2)$	$-\varepsilon+\varepsilon^2$	-0.1875	6
8	$(1,1,1)$	$(1, 1-\varepsilon, 1-\varepsilon+\varepsilon^2)$	$-1+\varepsilon-\varepsilon^2$	-0.8125	3

Man beginnt mit der Startlösung

$$\mathbf{x} = \begin{pmatrix} \varepsilon \\ \varepsilon^2 \\ \varepsilon^3 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{s} = \begin{pmatrix} 1-\varepsilon \\ 1-\varepsilon^2 \\ 1-\varepsilon^3 \end{pmatrix}.$$

Das Transformationsprinzip der Simplexmethode wählt jeweils die kleinste Verbesserung der Zielfunktion. Die Eckpunkte des gestörten Würfels werden in folgender Reihenfolge durchgegangen:

$$1 \ \Rightarrow \ 5 \ \Rightarrow \ 7 \ \Rightarrow \ 3 \ \Rightarrow \ 4 \ \Rightarrow \ 8 \ \Rightarrow \ 6 \ \Rightarrow \ 2.$$



□

Die Zusammenfassung dieses Kapitels ist wie folgt.

Satz 8.15 *Die Simplexmethode besitzt als worst case Komplexität mindestens* $\mathcal{O} = (2^n)$.

Bemerkung 8.16 Diese tritt aber praktisch nicht auf. In der Praxis hat die Simplexmethode eine polynomiale Komplexität. Erfahrungsgemäß liegt die Anzahl der Iterationen etwa bei $3n$. Im Kapitel 12 werden wir noch Verfahren kennenlernen, bei denen man beweisen kann, dass sie von polynomialer Komplexität sind, die sogenannten Innere–Punkt–Methoden. □

Kapitel 9

Dualitätssätze der linearen Optimierung

Definition 9.1 Duales lineares Programm. Sei

$$\begin{aligned} z = \mathbf{c}^T\mathbf{x} &\rightarrow \min ! \\ A\mathbf{x} &= \mathbf{b} \\ \mathbf{x} &\geq \mathbf{0} \end{aligned} \tag{9.1}$$

mit $\mathbf{c}, \mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$ ein lineares Programm.

Das lineare Programm

$$\begin{aligned} \tilde{z} = \mathbf{b}^T\mathbf{y} &\rightarrow \max ! \\ A^T\mathbf{y} &\leq \mathbf{c} \end{aligned} \tag{9.2}$$

mit $\mathbf{y} \in \mathbb{R}^m$ wird das zu (9.1) duale lineare Programm genannt. Man nennt (9.1) primal. □

Bemerkung 9.2 Ziel. Die Ziele dieses Abschnitts bestehen darin, die Existenz zulässiger Lösungen des dualen linearen Programms, die Relationen der zulässigen Lösungen des primalen und dualen linearen Programms, die Relationen zwischen den Optimallösungen und die Verbesserung der numerischen Verfahren zu untersuchen. Ein duales Analogon zur Simplexmethode soll entwickelt werden. □

Satz 9.3 *Ist* $\mathbf{x}$ *eine zulässige Lösung von (9.1) und ist* $\mathbf{y}$ *eine zulässige Lösung von (9.2), dann gilt* $z(\mathbf{x}) \geq \tilde{z}(\mathbf{y})$.

Beweis: Es gelten $\mathbf{x} \geq \mathbf{0}$ und $\mathbf{c}^T \geq \mathbf{y}^T A$. Damit folgt

$$z(\mathbf{x}) = \mathbf{c}^T\mathbf{x} \geq \mathbf{y}^T A\mathbf{x} = \mathbf{y}^T\mathbf{b} = \tilde{z}(\mathbf{y}).$$

■

Man nennt diesen Satz auch schwachen Dualitätssatz.

Folgerung 9.4 *Sind* $\mathbf{x}_0$ *eine zulässige Lösung von (9.1) und* $\mathbf{y}_0$ *eine zulässige Lösung von (9.2) und gilt* $z(\mathbf{x}_0) = \tilde{z}(\mathbf{y}_0)$*, dann ist* $\mathbf{x}_0$ *eine Optimallösung von (9.1) und* $\mathbf{y}_0$ *ist eine Optimallösung von (9.2).*

Satz 9.5 ***Starker Dualitätssatz.*** *Das primale Problem (9.1) besitzt genau dann eine endliche Optimallösung, wenn das duale Problem (9.2) eine endliche Optimallösung besitzt. In diesem Fall gilt* $z_{\min} = \tilde{z}_{\max}$.

Beweis: 1.) Es existiere ein endliches Minimum des primalen Problems (9.1) und dieses Minimum werde von $\mathbf{x}_0 = \left(x_1^{(0)}, \ldots, x_m^{(0)}, 0, \ldots, 0\right)^T$ angenommen. Dann sind erklärt:

- Die zugehörigen Basisvektoren seien $\mathbf{a}_1, \ldots, \mathbf{a}_m$.
- Mit $X = (x_{ij})_{i=1,\ldots,m,j=1,\ldots,n}$ werden die Darstellungskoeffizienten für alle Spalten von A bezüglich dieser Basisvektoren bezeichnet.
- $\mathbf{z} = (z_1, \ldots, z_n)^T$ sei der Vektor, der durch $z_j = \sum_{i=1}^m c_i x_{ij}$, $j = 1, \ldots, n$, erzeugt wird.
- $A_0 = (\mathbf{a}_1, \ldots, \mathbf{a}_m)$,
- $\mathbf{c}_0 = (c_1, \ldots, c_m)^T$.

Wegen des Optimalitätskriteriums der Simplexmethode gilt für $\mathbf{x}_0$

$$\mathbf{z} \leq \mathbf{c}, \tag{9.3}$$

wegen $z_k - c_k \leq 0$, $k = 1, \ldots, n$. Aus der Nebenbedingung und der Definition der Darstellungskoeffizienten folgt

$$A_0\mathbf{x}_0 = \mathbf{b}, \quad A_0 X = A.$$

Daraus ergibt sich

$$\mathbf{x}_0 = A_0^{-1}\mathbf{b}, \quad X = A_0^{-1}A. \tag{9.4}$$

Weiter erhalten wir aus der Definition von $\mathbf{z}$

$$\mathbf{c}_0^T X = \mathbf{z}^T \leq \mathbf{c}^T. \tag{9.5}$$

Jetzt setzen wir $\mathbf{y}_0 := A_0^{-T}\mathbf{c}_0$ und zeigen, dass $\mathbf{y}_0$ eine Optimallösung des dualen Problems (9.2) ist. Die Zulässigkeit von $\mathbf{y}_0$ folgt aus (9.4) und (9.5)

$$\mathbf{y}_0^T A = \mathbf{c}_0^T A_0^{-1} A = \mathbf{c}_0^T X \leq \mathbf{c}^T.$$

Die Lösung $\mathbf{y}_0$ ist optimal wegen

$$\tilde{z}(\mathbf{y}_0) = \mathbf{b}^T\mathbf{y}_0 = \mathbf{y}_0^T\mathbf{b} = \mathbf{c}_0^T A_0^{-1}\mathbf{b} = \mathbf{c}_0^T\mathbf{x}_0 = z(\mathbf{x}_0),$$

wobei (9.4) verwendet wurde. Damit liefert $\mathbf{y}_0$ einen Zielfunktionswert, der mit dem von $\mathbf{x}_0$ übereinstimmt. Da $\mathbf{x}_0$ Optimum des primalen Problems ist, ist $\mathbf{y}_0$ wegen Folgerung 9.4 Optimum des dualen Problems. Insbesondere gilt $z_{\min} = \tilde{z}_{\max}$.

2.) Das Ziel besteht darin, diesen Teil des Beweises auf den ersten Teil zurückzuführen, indem gezeigt wird, dass das duale Problem des dualen Problems (9.2) gerade das primale Problem (9.1) ist. Dazu wird das duale Problem so umgeformt, dass es die Gestalt eines primalen Problems annimmt. Bildet man dann aus dieser Form das duale Problem, erhält man die Behauptung.

Es existiere ein endliches Maximum $\tilde{z}_{\max}$ des dualen Problems (9.2). Wir setzen für einen beliebigen Vektor $\mathbf{y} \in \mathbb{R}^m$, der die Nebenbedingungen von (9.2) erfüllt, $\mathbf{y} = \mathbf{y}_1 - \mathbf{y}_2$, $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^m$, $\mathbf{y}_1, \mathbf{y}_2 \geq \mathbf{0}$. Aus den Nebenbedingungen von (9.2) erhält man

$$A^T(\mathbf{y}_1 - \mathbf{y}_2) + \mathbf{y}_3 = \mathbf{c} \quad \in \mathbb{R}^n,$$

mit den Schlupfvariablenvektor $\mathbf{y}_3 \in \mathbb{R}^n$, $\mathbf{y}_3 \geq \mathbf{0}$. Daraus bilden wir folgendes zu (9.2) äquivalentes Problem, wobei das Vorzeichen der Zielfunktion geändert wird

$$\begin{aligned} \mathbf{b}^T(\mathbf{y}_2 - \mathbf{y}_1) &\rightarrow \min\ ! \\ -A^T\mathbf{y}_1 + A^T\mathbf{y}_2 - \mathbf{y}_3 &= -\mathbf{c} \\ \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 &\geq \mathbf{0}. \end{aligned}$$

Setzt man

$$\begin{aligned} \mathbf{w} &:= \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{pmatrix} \in \mathbb{R}^{2m+n}, \ \mathbf{d} := \begin{pmatrix} -\mathbf{b} \\ \mathbf{b} \\ \mathbf{0}_n \end{pmatrix} \in \mathbb{R}^{2m+n}, \\ \mathcal{A} &:= \left(-A^T, A^T, -I_n\right) \in \mathbb{R}^{n\times(2m+n)}, \end{aligned}$$

so kann man dieses System in der Form

$$\begin{aligned} \mathbf{d}^T\mathbf{w} &\rightarrow \text{min !} \\ \mathcal{A}\mathbf{w} &= -\mathbf{c} \\ \mathbf{w} &\geq \mathbf{0} \end{aligned} \tag{9.6}$$

schreiben.

Einerseits ist dieses Problem zum dualen Problem (9.2) äquivalent. Damit besitzt die Zielfunktion von (9.6) nach Voraussetzung ein endliches Optimum $z_{\min}^{(\mathbf{w})}$, für welches gilt $z_{\min}^{(\mathbf{w})} = \tilde{z}_{\max}$.

Andererseits besitzt das Problem (9.6) die gleiche Gestalt wie das primale Problem (9.1). Die duale Aufgabe zu (9.6) hat nun folgende Gestalt: Gesucht ist $\mathbf{x} \in \mathbb{R}^n$ mit

$$-\mathbf{c}^T\mathbf{x} \rightarrow \max\ !, \quad \mathcal{A}^T\mathbf{x} \leq \mathbf{d},$$

das heißt

$$\begin{aligned} \mathbf{c}^T\mathbf{x} &\rightarrow \text{min !} \\ -A\mathbf{x} &\leq -\mathbf{b} \\ A\mathbf{x} &\leq \mathbf{b} \\ -\mathbf{x} &\leq \mathbf{0}. \end{aligned}$$

Aus den ersten beiden Nebenbedingungen folgt $A\mathbf{x} = \mathbf{b}$. Damit ist gezeigt, dass (9.1) das duale Problem zu (9.2) ist. Aus dem ersten Teil des Beweises wissen wir, dass die duale Aufgabe zu (9.6) ein endliches Maximum besitzt und das dieses Maximum mit $z_{\min}^{(\mathbf{w})} = \tilde{z}_{\max}$ übereinstimmt. ■

Jetzt wird der Fall betrachtet, dass die Zielfunktion der primalen Aufgabe nach unten nicht beschränkt ist.

Satz 9.6 *Ist die Zielfunktion $z = \mathbf{c}^T\mathbf{x}$ der primalen Aufgabe (9.1) auf der Menge der zulässigen Lösungen nach unten unbeschränkt, dann besitzt die zugehörige duale Aufgabe (9.2) keine zulässige Lösung. Analog gilt, dass im Falle dass die Zielfunktion der dualen Aufgabe auf der Menge der zulässigen Lösungen nicht nach oben beschränkt ist, die primale Aufgabe keine zulässige Lösung besitzt.*

Beweis: Indirekter Beweis. Sei die Zielfunktion des primalen Problems nicht nach unten beschränkt und sei $\mathbf{y}$ eine zulässige Lösung des dualen Problems, das heißt es gilt $A^T\mathbf{y} \leq \mathbf{c}^T$. Aus Satz 9.3 folgt dann aber $z(\mathbf{x}) \geq \tilde{z}(\mathbf{y})$ für alle zulässigen Lösungen $\mathbf{x}$ des primalen Problems und die Zielfunktion wäre nach unten beschränkt.

Die zweite Aussage folgt aus der ersten Aussage und daraus, dass das primale Problem (9.1) das duale Problem des dualen Problems (9.2) ist. ■

Folgerung 9.7 *Eine zulässige Lösung $\mathbf{x}_0$ des primalen Problems (9.1) ist genau dann optimal, wenn eine zulässige Lösung $\mathbf{y}_0$ des dualen Problems (9.2) existiert, mit $\mathbf{c}^T\mathbf{x}_0 = \mathbf{b}^T\mathbf{y}_0$. Eine analoge Aussage gilt, wenn man vom dualen Problem ausgeht.*

Satz 9.8 Komplementaritätssatz. *Es sei $\mathbf{x}_0 = (x_1^{(0)}, \ldots, x_m^{(0)}, 0, \ldots, 0)^T$ eine zulässige Basislösung des primalen Problems (9.1). Dann ist $\mathbf{x}_0$ genau dann optimal, wenn es eine zulässige Lösung $\mathbf{y}$ des dualen Problems (9.2) mit folgenden Eigenschaften gibt:*

1) für alle Indizes $i \in \{1, \ldots, m\}$ mit $x_i^{(0)} > 0$ gilt $\mathbf{a}_i^T\mathbf{y} = c_i$,

2) für alle Indizes $j \in \{1, \ldots, n\}$ mit $\mathbf{a}_j^T\mathbf{y} < c_j$ gilt $x_j^{(0)} = 0$.

Beweis: i) Sei $\mathbf{x}_0$ optimal. Nach Folgerung 9.7 gibt es dann ein $\mathbf{y}_0$ mit $\mathbf{c}^T\mathbf{x}_0 = \mathbf{b}^T\mathbf{y}_0$. Einsetzen von $A\mathbf{x}_0 = \mathbf{b}$ ergibt

$$\mathbf{c}^T\mathbf{x}_0 = \mathbf{b}^T\mathbf{y}_0 = (A\mathbf{x}_0)^T\,\mathbf{y}_0 = \underbrace{\mathbf{x}_0^T A^T \mathbf{y}_0}_{\in\mathbb{R}} = \mathbf{y}_0^T A\mathbf{x}_0 \quad \Longleftrightarrow \quad \left(\mathbf{c}^T - \mathbf{y}_0^T A\right)\mathbf{x}_0 = 0.$$

oder in Summenschreibweise

$$\sum_{j=1}^{n} \left(c_j - \mathbf{y}_0^T \mathbf{a}_j\right) x_j^{(0)} = 0.$$

Da $\mathbf{y}_0$ eine zulässige Lösung des dualen Problems ist und $\mathbf{x}^{(0)}$ eine zulässige Lösung des primalen Problems, sind alle Faktoren nichtnegativ. Damit die Summe Null wird, müssen alle Summanden verschwinden und wenigstens jeweils einer der Faktoren Null sein. Ist $x_j^{(0)} > 0$, muss $\mathbf{a}_j^T\mathbf{y}_0 = c_j$ sein. Ist $\mathbf{a}_j^T\mathbf{y}_0 < c_j$, so muss $x_j^{(0)} = 0$ sein. Die Optimallösung $\mathbf{y}_0$ des dualen Problems erfüllt also die Bedingungen 1) und 2).

ii) Es gibt einen Vektor $\mathbf{y}$ der die Bedingungen 1) und 2) erfüllt. Wir nehmen an, $\mathbf{x}_0$ sei nicht optimal. Dann gilt $\mathbf{c}^T\mathbf{x}_0 > \mathbf{b}^T\mathbf{y}$. Analog zum ersten Teil erhält man

$$\sum_{j=1}^{n} \left(c_j - \mathbf{y}^T \mathbf{a}_j\right) x_j^{(0)} > 0.$$

Aus den Bedinungen 1), 2) folgt jedoch, dass die Summe verschwindet. Damit ist die Annahme falsch und $\mathbf{x}_0$ ist optimal. ■

Bemerkung 9.9 Mit Hilfe der Dualität ist die Möglichkeit der Bestimmung von Schranken für eine zulässige (optimale) Lösung gegeben. Es gilt

$$z(\mathbf{x}) \geq z(\mathbf{x}_0) = \tilde{z}(\mathbf{y}_0) \geq \tilde{z}(\mathbf{y}),$$

wobei $\mathbf{x}$ eine zulässige Lösung des primalen Problems (9.1), $\mathbf{x}_0$ eine Optimallösung von (9.1), $\mathbf{y}$ eine zulässige Lösung des dualen Problems (9.2) und $\mathbf{y}_0$ eine Optimallösung des dualen Problems ist. □

Bemerkung 9.10 Symmetrisches duales Programm. Ein Spezialfall des dualen linearen Programms ist das symmetrische duale lineare Programm. Gegeben sei das lineare Programm

$$\begin{aligned} z = \mathbf{c}^T\mathbf{x} &\rightarrow \text{min !} \\ A\mathbf{x} &\geq \mathbf{b} \\ \mathbf{x} &\geq \mathbf{0}. \end{aligned} \tag{9.7}$$

Aus (9.7) wird ein lineares Programm

$$\begin{aligned} \tilde{z} = \mathbf{b}^T\mathbf{y} &\rightarrow \text{max !} \\ A^T\mathbf{y} &\leq \mathbf{c} \\ \mathbf{y} &\geq \mathbf{0}. \end{aligned} \tag{9.8}$$

konstruiert. □

Satz 9.11 *Die linearen Programm (9.7) und (9.8) sind duale lineare Programme im Sinne von Definition 9.1.*

Beweis: Aus (9.7) konstruieren wir das lineare Programm in Normalform

$$z = \mathbf{c}^T\mathbf{x} \rightarrow \text{min !}, \quad A\mathbf{x} - \mathbf{v} = \mathbf{b}, \quad \mathbf{x}, \mathbf{v} \geq \mathbf{0}.$$

Aus Definition 9.1 ergibt sich das folgende duale lineare Programm

$$\tilde{z} = \mathbf{b}^T\mathbf{y} \to \max\ !, \quad A^T\mathbf{y} \leq \mathbf{c}, \quad -I_m\mathbf{y} \leq \mathbf{0}.$$

Die letzte Bedingung ist äquivalent zu $\mathbf{y} \geq \mathbf{0}$. ■

Man hat nun eine Nichtnegativitätsbedingung an die zulässigen Lösungen des dualen Programms (9.8).

Satz 9.12 Komplementaritätssatz. *Sind $\mathbf{x}_0$ eine zulässige Lösung von (9.7) und $\mathbf{y}_0$ eine zulässige Lösung von (9.8), so sind sie genau dann optimal, wenn die folgenden Relationen erfüllt sind:*

$$\mathbf{y}_0^T\,(A\mathbf{x}_0 - \mathbf{b}) = 0, \tag{9.9}$$

$$\left(\mathbf{y}_0^T A - \mathbf{c}^T\right)\mathbf{x}_0 = 0. \tag{9.10}$$

Beweis: 1) Seien $\mathbf{x}_0$ und $\mathbf{y}_0$ optimal. Dann folgt aus der Nebenbedingung von (9.8), aus der Nichtnegativität von $\mathbf{x}_0$ und aus Folgerung 9.7

$$\mathbf{y}_0^T\,(A\mathbf{x}_0 - \mathbf{b}) = \mathbf{y}_0^T A\mathbf{x}_0 - \mathbf{y}_0^T\mathbf{b} \leq \mathbf{c}^T\mathbf{x}_0 - \mathbf{y}_0^T\mathbf{b} = 0.$$

Andererseits gilt mit $\mathbf{y}_0 \geq \mathbf{0}$ und der Nebenbedingung von (9.7)

$$\mathbf{y}_0^T\,(A\mathbf{x}_0 - \mathbf{b}) \geq 0.$$

Aus beiden Ungleichungen zusammen folgt (9.9). Die Beziehung (9.10) beweist man analog.

2) Gelten jetzt (9.9) und (9.10). Daraus folgt

$$\mathbf{c}^T\mathbf{x}_0 = \mathbf{y}_0^T A\mathbf{x}_0 = \mathbf{y}_0^T\mathbf{b}.$$

Nach Folgerung 9.7 sind $\mathbf{x}_0$ und $\mathbf{y}_0$ optimal. ■

Beispiel für Anwendungen des Dualitätsprinzips werden in den Übungsaufgaben behandelt.

Kapitel 10

Die duale Simplexmethode

Bemerkung 10.1 Motivation. Bei der dualen Simplexmethode ist eine Startlösung oftmals leichter angebbar als bei der Simplexmethode für das ursprüngliche lineare Programm, da man keine Nichtnegativitätsanforderungen zu erfüllen hat. Des weiteren ist die duale Simplexmethode ein wichtiges Verfahren zur Lösung von ganzzahligen linearen Programmen, das heißt, von linearen Programmen, bei denen die Lösung ganzzahlig sein soll.

Seien

$$\begin{array}{rcl} z = \mathbf{c}^T\mathbf{x} & \rightarrow & \min\,! \\ A\mathbf{x} & = & \mathbf{b} \\ \mathbf{x} & \geq & \mathbf{0} \end{array} \tag{10.1}$$

das primale Programm und

$$\begin{array}{rcl} \tilde{z} = \mathbf{b}^T\mathbf{y} & \rightarrow & \max\,! \\ A^T\mathbf{y} & \leq & \mathbf{c} \end{array} \tag{10.2}$$

das zugehörige duale Programm. Wir setzen voraus, dass $\tilde{z}$ endlich ist. □

Definition 10.2 Ecklösung. Ohne Beschränkung der Allgemeinheit sei $A_B = (\mathbf{a}_1, \ldots, \mathbf{a}_m)$ eine Basis. Ein Punkt $\mathbf{y} = (y_1, \ldots, y_m)^T$ heißt Ecklösung von (10.2), wenn

$$\begin{array}{rcll} \mathbf{a}_i^T\mathbf{y} & = & c_i & \text{für } i = 1, \ldots, m, \\ \mathbf{a}_i^T\mathbf{y} & < & c_i & \text{für } i = m+1, \ldots, n \end{array} \tag{10.3}$$

gelten. □

Bemerkung 10.3 Ecklösung bedeutet, dass die Nebenbedingungen, die durch die Basisvektoren von A_B gegeben sind, mit Gleichheit erfüllt sind und die Nebenbedingungen mit den Nichtbasisvektoren als echte Ungleichung.

Seien A_B^{-1} die Inverse von A_B mit der Darstellung

$$A_B^{-1} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_m \end{pmatrix}, \quad \text{dass heißt} \quad \mathbf{b}_i\mathbf{a}_j = \delta_{ij}, \quad i, j = 1, \ldots, m,$$

und $\mathbf{y} = y_1\mathbf{a}_1 + \ldots + y_m\mathbf{a}_m \in \mathbb{R}^m$ ein beliebiger Vektor. Dann folgt (man beachte, die $\mathbf{b}_i$ sind Zeilenvektoren)

$$\mathbf{b}_i\mathbf{y} = y_1\mathbf{b}_i\mathbf{a}_1 + \ldots + y_m\mathbf{b}_i\mathbf{a}_m = y_i, \quad i = 1, \ldots, m.$$

Daraus erhält man insbesondere die Darstellung

$$\mathbf{y} = (\mathbf{b}_1\mathbf{y})\,\mathbf{a}_1 + \ldots + (\mathbf{b}_m\mathbf{y})\,\mathbf{a}_m. \tag{10.4}$$

□

Die Ausartung in der dualen Simplexmethode, die im folgenden Satz ausgeschlossen ist, wird in seinem Beweis definiert, siehe auch Bemerkung 10.5.

Satz 10.4 Hauptsatz der dualen Simplexmethode. *Sei $\tilde{z}$ nach oben beschränkt und sei Ausartung ausgeschlossen. Ist $\mathbf{y} \in \mathbb{R}^m$ eine Ecklösung und gilt $\mathbf{b}_i\mathbf{b} < 0$ für wenigstens ein $i = 1, \ldots, m$, so existiert eine Ecklösung $\bar{\mathbf{y}}$ mit größerem Wert der Zielfunktion $\tilde{z}$.*

Beweis: Seien $\mathbf{y}$ eine Ecklösung, $\mathbf{b}_l\mathbf{b} < 0$ und $\theta > 0$ beliebig. Es wird ein $\bar{\mathbf{y}}$ konstruiert, welches die Bedingungen des Satzes erfüllt.

Man bildet

$$\bar{\mathbf{y}} = \mathbf{y} - \theta\mathbf{b}_l^T.$$

Aus der Eckpunkteigenschaft (10.3) folgt

$$\mathbf{a}_i^T\bar{\mathbf{y}} = \mathbf{a}_i^T\mathbf{y} - \theta\underbrace{\mathbf{a}_i^T\mathbf{b}_l^T}_{=0} = \mathbf{a}_i^T\mathbf{y} = c_i, \quad i = 1, \ldots, m,\ i \neq l, \tag{10.5}$$

$$\mathbf{a}_l^T\bar{\mathbf{y}} = \mathbf{a}_l^T\mathbf{y} - \theta\underbrace{\mathbf{a}_l^T\mathbf{b}_l^T}_{=1} = \mathbf{a}_l^T\mathbf{y} - \theta < c_l. \tag{10.6}$$

Damit man eine zulässige Lösung hat, muss auch

$$\mathbf{a}_i^T\bar{\mathbf{y}} = \mathbf{a}_i^T\mathbf{y} - \theta\mathbf{a}_i^T\mathbf{b}_l^T \leq c_i, \quad i = m+1, \ldots, n,$$

gelten. Für wenigstens einen Index i gilt $\mathbf{a}_i^T\mathbf{b}_l^T < 0$. Anderenfalls, falls also $\mathbf{a}_i^T\mathbf{b}_l^T \geq 0$ für $i = m+1, \ldots, n$, sind die Nebenbedingungen für beliebig großes θ erfüllt. Damit wäre

$$\tilde{z} = \mathbf{b}^T\bar{\mathbf{y}} = \mathbf{b}^T\mathbf{y} - \theta\underbrace{\mathbf{b}^T\mathbf{b}_l^T}_{<0 \text{ n.V.}}$$

unbeschränkt im Widerspruch zur Voraussetzung.

Wir wählen

$$\theta = \min_{i=m+1,\ldots,n;\mathbf{b}_l\mathbf{a}_i<0}\left(\frac{\mathbf{a}_i^T\mathbf{y} - c_i}{\mathbf{b}_l\mathbf{a}_i}\right) > 0. \tag{10.7}$$

Wir nehmen an, dass θ von genau einem Index $i = k$ bestimmt wird. Sonst hat man Ausartung. Es gelten:

1.) $\bar{\mathbf{y}}$ erfüllt (10.5) und (10.6), das heißt, die Basisvektoren die in der Basis verbleiben erfüllen die Nebenbedingung mit Gleichheit und die neue Nichtbasisvariable mit Index l als echte Ungleichung.

2.) Für den Index k, der in die Basis aufgenommen werden soll, gilt

$$\mathbf{a}_k^T\bar{\mathbf{y}} = \mathbf{a}_k^T\mathbf{y} - \frac{\mathbf{a}_k^T\mathbf{y} - c_k}{\mathbf{b}_l\mathbf{a}_k}\mathbf{b}_l\mathbf{a}_k = c_k,$$

3.) Aus der Wahl von θ folgt

$$\mathbf{a}_i^T\bar{\mathbf{y}} = \mathbf{a}_i^T\mathbf{y} - \theta\mathbf{a}_i^T\mathbf{b}_l^T < c_i, \quad i = m+1, \ldots, n,\ i \neq k.$$

Falls $\mathbf{a}_i^T\mathbf{b}_l^T$ nichtnegativ ist, ist das Erfülltsein dieser Bedingung klar. Ansonsten wurde bei der Wahl von θ gerade der Index k ausgewählt, der die k–te Nebenbedingung $\mathbf{a}_k^T\mathbf{b}_l^T < 0$ für $\bar{\mathbf{y}}$ zu einer Gleichung werden lässt, ohne dass die anderen Nebenbedingungen mit $\mathbf{a}_i^T\mathbf{b}_l^T < 0$ verletzt werden.

Aus 1) – 3) folgt, dass $\bar{\mathbf{y}}$ eine Ecklösung ist. Ferner gilt

$$\tilde{z} = \mathbf{b}^T\bar{\mathbf{y}} = \mathbf{b}^T\mathbf{y} - \theta\mathbf{b}^T\mathbf{b}_l^T > \mathbf{b}^T\mathbf{y}.$$

■

Bemerkung 10.5 Ausartung im dualen Programm. Ist der Index k bei der Wahl von θ in (10.7) nicht eindeutig, so liegt Ausartung vor. □

Bemerkung 10.6 Unbeschränktheit der Zielfunktion. Die Unbeschränktheit der Zielfunktion $\tilde{z}$ ist an $\mathbf{a}_j^T \mathbf{b}_l^T \geq 0$ für $j = m+1, \ldots, n$, zu erkennen. □

Satz 10.7 Optimalitätskriterium. *Sei* $\mathbf{y}$ *eine Ecklösung von (10.2) und gelte* $\mathbf{b}_i\mathbf{b} \geq 0$ *für alle* $i = 1, \ldots, m$. *Dann ist* $\mathbf{y}$ *die Optimallösung von (10.2) und die Größen* $x_i = \mathbf{b}_i\mathbf{b}$ *stellen die Basisvariablen der Optimallösung des primalen Problems (10.1) dar.*

Beweis: Es gilt mit (10.3)

$$z = \sum_{i=1}^{m} c_i x_i = \sum_{i=1}^{m} c_i \mathbf{b}_i \mathbf{b} = \sum_{i=1}^{m} \mathbf{y}^T \mathbf{a}_i \mathbf{b}_i \mathbf{b} = \mathbf{y}^T \underbrace{\left(\sum_{i=1}^{m} \mathbf{a}_i \mathbf{b}_i \right)}_{=I_m} \mathbf{b} = \mathbf{b}^T \mathbf{y} = \tilde{z}.$$

Nach dem starken Dualitätssatz, Satz 9.5, folgt die Aussage des Satzes.

Bemerkung zur Summe: Sei $\sum_{i=1}^{m} \mathbf{a}_i \mathbf{b}_i = C$ mit einer unbekannten Matrix C. Multiplikation diese Gleichung von rechts mit $\mathbf{a}_j$, $j = 1, \ldots, m$, ergibt

$$C\mathbf{a}_j = \sum_{i=1}^{m} \mathbf{a}_i \underbrace{\mathbf{b}_i \mathbf{a}_j}_{=\delta_{ij}} = \mathbf{a}_j$$

für alle $\mathbf{a}_j$. Da die $\{\mathbf{a}_j\}$ eine Basis des $\mathbb{R}^m$ bilden, gilt $C = I_m$. ■

Bemerkung 10.8 Duale Simplextabelle. Die duale Simplextabelle hat die Gestalt

			$m+1$	$\ldots$	k	$\ldots$	n
i	c_i	Lösung	c_{m+1}	$\ldots$	c_k	$\ldots$	c_n
1	c_1	$\mathbf{b}_1\mathbf{b}$	$\mathbf{b}_1\mathbf{a}_{m+1}$	$\ldots$	$\mathbf{b}_1\mathbf{a}_k$	$\ldots$	$\mathbf{b}_1\mathbf{a}_n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
l	c_l	$\mathbf{b}_l\mathbf{b}$	$\mathbf{b}_l\mathbf{a}_{m+1}$	$\ldots$	$\mathbf{b}_l\mathbf{a}_k$	$\ldots$	$\mathbf{b}_l\mathbf{a}_n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	c_m	$\mathbf{b}_m\mathbf{b}$	$\mathbf{b}_m\mathbf{a}_{m+1}$	$\ldots$	$\mathbf{b}_m\mathbf{a}_k$	$\ldots$	$\mathbf{b}_m\mathbf{a}_n$
		$\tilde{z}$	$\mathbf{a}_{m+1}^T\mathbf{y} - c_{m+1}$	$\ldots$	$\mathbf{a}_k^T\mathbf{y} - c_k$	$\ldots$	$\mathbf{a}_n^T\mathbf{y} - c_n$

Wie bei der Simplexmethode, wird die Zeile l Hauptzeile und die Spalte k Hauptspalte genannt. Das Pivotelement ist $\mathbf{b}_l\mathbf{a}_k$. Aus den Nebenbedingungen des dualen linearen Programms (10.2) folgt, dass die Einträge in der letzten Zeile im Nichtbasisteil nichtpositiv sind.

- In der Schlusszeile stehen die Größen, die man zur Berechnung von θ in (10.7) benötigt.
- Die Spalte *Lösung* enthält eine Basislösung des primalen Problems. Sei $\tilde{\mathbf{x}} \in \mathbb{R}^m$ der Vektor mit den Basisvariablen des primalen Problems. Aus den Nebenbedingungen des primalen Problems folgt

$$A_B\tilde{\mathbf{x}} = \mathbf{b} \quad \Longrightarrow \quad \tilde{\mathbf{x}} = A_B^{-1}\mathbf{b} = \begin{pmatrix} \mathbf{b}_1\mathbf{b} \\ \vdots \\ \mathbf{b}_m\mathbf{b} \end{pmatrix}.$$

  Diese Basislösung ist im allgemeinen nicht zulässig, da sie negative Komponenten besitzt. Gilt jedoch $\mathbf{b}_i\mathbf{b} \geq 0$ für alle $i = 1, \ldots, m$, dann ist sie primale Optimallösung, siehe Satz 10.7.

- Die Nichtbasisvektoren lassen sich als Linearkombination der Basisvektoren darstellen $\mathbf{a}_j = A_B\mathbf{a}$ mit einem unbekannten Koeffizientenvektor $\mathbf{a}$. Dieser lässt sich durch $\mathbf{a} = A_B^{-1}\mathbf{a}_j$ berechnen, was in Komponentenschreibweise zu

$$\mathbf{a}_j = \sum_{i=1}^{m} (\mathbf{b}_i\mathbf{a}_j)\,\mathbf{a}_i, \quad j = m+1,\ldots,n$$

  führt, siehe (10.4).
- Falls man das Optimum des dualen Problems mit der dualen Simplexmethode gefunden hat, ist die Tabelle der dualen Simplexmethode eine Optimaltabelle für die primale Aufgabe.

□

Bemerkung 10.9 Herleitung der Transformationsregeln. Sei $A_B = (\mathbf{a}_1,\ldots,\mathbf{a}_m)$ und sei $\hat{A}_B = (\mathbf{a}_1,\ldots,\mathbf{a}_{l-1},\mathbf{a}_k,\mathbf{a}_{l+1},\ldots,\mathbf{a}_m)$. Dann ist

$$\begin{aligned}
A_B^{-1}\hat{A}_B &= \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_m \end{pmatrix} (\mathbf{a}_1,\ldots,\mathbf{a}_{l-1},\mathbf{a}_k,\mathbf{a}_{l+1},\ldots,\mathbf{a}_m) \\
&= \begin{pmatrix} \mathbf{b}_1\mathbf{a}_1 & \mathbf{b}_1\mathbf{a}_2 & \cdots & \mathbf{b}_1\mathbf{a}_k & \cdots & \mathbf{b}_1\mathbf{a}_m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{b}_l\mathbf{a}_1 & \mathbf{b}_l\mathbf{a}_2 & \cdots & \mathbf{b}_l\mathbf{a}_k & \cdots & \mathbf{b}_l\mathbf{a}_m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{b}_m\mathbf{a}_1 & \mathbf{b}_m\mathbf{a}_2 & \cdots & \mathbf{b}_m\mathbf{a}_k & \cdots & \mathbf{b}_m\mathbf{a}_m \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 & \cdots & 0 & \mathbf{b}_1\mathbf{a}_k & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \mathbf{b}_2\mathbf{a}_k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \mathbf{b}_l\mathbf{a}_k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \mathbf{b}_m\mathbf{a}_k & 0 & \cdots & 1 \end{pmatrix} =: I_m^{(k)}.
\end{aligned}$$

Mit

$$\left(\hat{A}_B\right)^{-1} = \begin{pmatrix} \hat{\mathbf{b}}_1 \\ \vdots \\ \hat{\mathbf{b}}_m \end{pmatrix} \quad \Longrightarrow \quad A_B^{-1} = I_m^{(k)}\left(\hat{A}_B\right)^{-1},$$

oder

$$\begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_l \\ \vdots \\ \mathbf{b}_m \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{b}}_1 + (\mathbf{b}_1\mathbf{a}_k)\hat{\mathbf{b}}_k \\ \vdots \\ (\mathbf{b}_l\mathbf{a}_k)\hat{\mathbf{b}}_k \\ \vdots \\ \hat{\mathbf{b}}_m + (\mathbf{b}_m\mathbf{a}_k)\hat{\mathbf{b}}_k \end{pmatrix}.$$

Das Ziel besteht darin, die Vektoren $\mathbf{b}_i$ durhc die Vektoren $\hat{\mathbf{b}}_i$ zu ersetzen. Im Prinzip stehen nun die Transformationsregeln da. Für das Pivotelement gilt

$$\hat{\mathbf{b}}_k = \frac{\mathbf{b}_l}{\mathbf{b}_l\mathbf{a}_k} \quad \Longrightarrow \quad \hat{\mathbf{b}}_k\mathbf{a}_l = \frac{\mathbf{b}_l\mathbf{a}_l}{\mathbf{b}_l\mathbf{a}_k} = \frac{1}{\mathbf{b}_l\mathbf{a}_k}.$$

Für die Hauptspalte erhält man daraus

$$\hat{\mathbf{b}}_i = \mathbf{b}_i - (\mathbf{b}_i\mathbf{a}_k)\,\hat{\mathbf{b}}_k \quad \Longrightarrow \quad \hat{\mathbf{b}}_i\mathbf{a}_l = \mathbf{b}_i\mathbf{a}_l - (\mathbf{b}_i\mathbf{a}_k)\left(\hat{\mathbf{b}}_k\mathbf{a}_l\right) = 0 - \frac{\mathbf{b}_i\mathbf{a}_k}{\mathbf{b}_l\mathbf{a}_k} \quad i \neq l.$$

Für die Hauptzeile ergibt sich umittelbar, wobei $\mathbf{a}_0 := \mathbf{b}$ ist,

$$\hat{\mathbf{b}}_k \mathbf{a}_j = \frac{\mathbf{b}_l \mathbf{a}_j}{\mathbf{b}_l \mathbf{a}_k}, \qquad j = 0, m+1, \ldots, n,\ j \neq k.$$

Damit ergibt sich auch die Rechteckregel

$$\hat{\mathbf{b}}_i \mathbf{a}_j = \mathbf{b}_i \mathbf{a}_j - \mathbf{b}_i \mathbf{a}_k \hat{\mathbf{b}}_k \mathbf{a}_j = \mathbf{b}_i \mathbf{a}_j - \frac{\mathbf{b}_l \mathbf{a}_j}{\mathbf{b}_l \mathbf{a}_k} \mathbf{b}_i \mathbf{a}_k.$$

Zusammenfassung: Falls in der Spalte *Lösung* wenigstens ein $\mathbf{b}_i \mathbf{b} < 0$ steht, zum Beispiel für $i = l$, so transformiert man wie folgt:

- setze $\mathbf{a}_0 = \mathbf{b}$,
- vertausche die Indizes l und k,
- Pivotelement: $\hat{\mathbf{b}}_k \mathbf{a}_l = 1/(\mathbf{b}_l \mathbf{a}_k)$,
- Hauptspalte:
$$\hat{\mathbf{b}}_i \mathbf{a}_l = -\frac{\mathbf{b}_i \mathbf{a}_k}{\mathbf{b}_l \mathbf{a}_k}, \quad i = 1, \ldots, m,\ i \neq l,$$
- Hauptzeile:
$$\hat{\mathbf{b}}_k \mathbf{a}_j = \frac{\mathbf{b}_l \mathbf{a}_j}{\mathbf{b}_l \mathbf{a}_k}, \quad j = 0, m+1, \ldots, n,\ j \neq k,$$
- Rechteckregel:
$$\hat{\mathbf{b}}_i \mathbf{a}_j = \mathbf{b}_i \mathbf{a}_j - \frac{\mathbf{b}_l \mathbf{a}_j}{\mathbf{b}_l \mathbf{a}_k} \mathbf{b}_i \mathbf{a}_k, \quad i = 1, \ldots, m,\ i \neq l,\ j = 0, m+1, \ldots, n,\ j \neq k.$$

Die Rechteckregel wird auch auf die letzte Zeile angewandt *Übungsaufgabe.* Das sind dieselben Regeln wie im primalen Fall ! □

Beispiel 10.10 Wir betrachten noch einmal das Problem aus Beispiel 5.5:

$$\begin{array}{rcl} z = -3x_1 - 2x_2 - 4x_3 - x_4 & \rightarrow & \min\ ! \\ \begin{pmatrix} 2 & 2 & 3 & 0 & 1 & 0 & 0 \\ 1 & 3 & 0 & 2 & 0 & 1 & 0 \\ 1 & 1 & 5 & 2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_7 \end{pmatrix} & = & \begin{pmatrix} 700 \\ 400 \\ 500 \end{pmatrix} \\ \mathbf{x} & \geq & \mathbf{0}. \end{array}$$

In Beispiel 5.5 hatten wir die Optimallösung $\mathbf{x} = (320, 0, 20, 40, 0, 0, 0)^T$ mit $z = -1080$ erhalten. Das duale Problem zum obigen linearen Programm lautet

$$\begin{array}{rcl} \tilde{z} = 700y_1 + 400y_2 + 500y_3 & \rightarrow & \max\ ! \\ \begin{pmatrix} 2 & 1 & 1 \\ 2 & 3 & 1 \\ 3 & 0 & 5 \\ 0 & 2 & 2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} & \leq & \begin{pmatrix} -3 \\ -2 \\ -4 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \end{array}$$

Wir nehmen uns die erste, dritte und fünfte Nebenbedingung des dualen Problems her und betrachte diese Bedingungen als Gleichungen:

$$\begin{pmatrix} 2 & 1 & 1 \\ 3 & 0 & 5 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -3 \\ -4 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{y} = \begin{pmatrix} 0 \\ -11/5 \\ -4/5 \end{pmatrix}.$$

Durch Einsetzen in die anderen Nebenbedingungen verifiziert man, dass man damit eine Ecklösung des dualen Problems gefunden hat. Es ist

$$A_B = (\mathbf{a}_1, \mathbf{a}_3, \mathbf{a}_5) = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 0 & 0 \\ 1 & 5 & 0 \end{pmatrix},$$

$$A_B^{-1} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1/5 & 1/5 \\ 1 & -7/5 & -3/5 \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_3 \\ \mathbf{b}_5 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 700 \\ 400 \\ 500 \end{pmatrix}.$$

Damit kann man alle Größen für die duale Simplextabelle bestimmen:

			2	4	6	7
i	c_i	Lösung	-2	-1	0	0
1	-3	400	3	2	1	0
3	-4	20	-2/5	0	-1/5	1/5
5	0	-160	-14/5	-4	-7/5	-3/5
		-1280	-27/5	-5	-11/5	-4/5

Die Lösung ist nicht optimal, da $-160 < 0$. Damit ist $l = 5$ die Hauptzeile. Zur Bestimmung der Hauptspalte berechnet man

$$\theta = \min_{j \in \{2,4,6,7\},\ \mathbf{b}_l \mathbf{a}_j < 0} \left(\frac{\mathbf{a}_j^T \mathbf{y} - c_j}{\mathbf{b}_l \mathbf{a}_j} \right) = \min \left\{ \frac{27}{14}, \frac{5}{4}, \frac{11}{7}, \frac{4}{3} \right\} = \frac{5}{4}.$$

Damit ist die Hauptspalte $k = 4$. Mit den Transformationsregeln der Simplexmethode erhält man die neue duale Simplextabelle

			2	5	6	7
i	c_i	Lösung	-2	0	0	0
1	-3	320	8/5	1/2	3/10	-3/10
3	-4	20	-2/5	0	-1/5	1/5
4	-1	40	7/10	-1/4	7/20	3/20
		-1080	-19/10	-5/4	-9/20	-1/20

Damit ist das Optimum bestimmt. Die Optimallösung des primalen Problems findet man in der Spalte *Lösung* ebenso den zugehörigen Zielfunktionswert.

Die Lösung des dualen Problems $\mathbf{y}$ kann man im allgemeinen nicht direkt aus der dualen Simplextabelle ablesen. In der letzten Zeile steht nämlich $\mathbf{a}_j^T \mathbf{y} - c_j$. Das direkte Ablesen geht nur, wenn $c_j = 0$ und die $\mathbf{a}_j$ Einheitsvektoren sind. Das ist in diesem Beispiel gegeben, nämlich für die Indizes $5, 6, 7$. Die Lösung des dualen Problems ist also $\mathbf{y} = (-5/4, -9/20, -1/20)^T$ mit dem Zielfunktionswert $\tilde{z} = -1080$. Im allgemeinen muss man noch ein lineares Gleichungssystem lösen, um die Lösung des dualen Problems zu berechnen. □

Kapitel 11

Die duale Simplexmethode zur Lösung rein ganzzahliger linearer Programme

Wir betrachten folgendes Optimierungsproblem

$$\begin{aligned} z = \mathbf{c}^T\mathbf{x} &\rightarrow \text{min !} \\ A\mathbf{x} &= \mathbf{b} && (11.1)\\ \mathbf{x} &\geq \mathbf{0} && (11.2)\\ x_j &\text{ ganz für } j = 1,\ldots,n_1 \leq n, && (11.3)\\ a_{ij}, b_i &\text{ ganz für } i = 1,\ldots,m,\ j = 1,\ldots,n. && (11.4) \end{aligned}$$

Definition 11.1 Ganzzahliges lineares Programm. Das lineare Programm (11.1) – (11.4) heißt rein ganzzahliges lineares Programm falls $n_1 = n$. Ansonsten heißt es für $n_1 > 0$ gemischt ganzzahliges lineares Programm. □

Wir werden nur rein ganzzahlige lineare Programme betrachten.

Bemerkung 11.2 Häufig enthalten ganzzahlige lineare Programme Bedingungen der folgenden Art

$$0 \leq x_j \leq 1, \quad x_j \text{ ganz},$$

für gewisse Indizes j. □

Beispiel 11.3 Wir betrachten

$$\begin{aligned} z = -8x_1 - 4x_2 &\rightarrow \text{min !} \\ -2x_1 + 3x_2 &\leq 6 \\ 8x_1 + 3x_2 &\leq 20 \\ \mathbf{x} &\geq 0 \\ x_1, x_2 &\text{ ganz.} \end{aligned}$$

Das Problem ohne die Ganzheitsforderung kann man graphisch lösen, siehe Abbildung 11.1.

Das stetige Optimum ist

$$\mathbf{x} = \left(\frac{7}{5}, \frac{44}{15}\right), \quad z = -\frac{344}{15}.$$

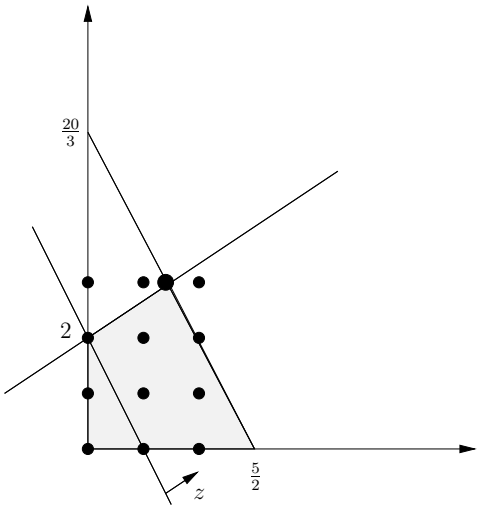


Abbildung 11.1: Illustration zu Beispiel 11.3.

Eine einfache Idee zur Bestimmung des ganzzahligen Optimums ist Runden. Man erhält damit $\mathbf{x} = (1, 3)^T$. Dieser Punkt ist jedoch nicht zulässig. Durch Abrunden folgt $\mathbf{x} = (1, 2)^T$. Man erhält mit diesem Punkt den Zielfunktionswert $z = -16$. Dieser ist jedoch nicht optimal, da man mit $\mathbf{x} = (2, 1)^T$ den Wert $z = -20$ erhält.

Runden ist also keine geeignete Lösungstechnik. □

Bemerkung 11.4 Motivation für das Schnittprinzip. Mit dem normalen Simplexverfahren kann man nicht arbeiten, da das Optimum ein innerer Punkt des zulässigen Bereiches ist. Das könnte man durch die Bestimmung der konvexen Hülle aller zulässigen Punkte ändern. Dieses Vorgehen ist aber im allgemeinen viel zu aufwendig. Stattdessen versucht man in der Nähe des stetigen Optimums durch Abschneiden den gesuchten ganzzahligen optimalen Punkt zu einem Eckpunkt zu machen. Dieses Schnittprinzip soll jetzt auf eine Art (nach Gomory (1957)) realisiert werden. □

Definition 11.5 Schnittbedingung. Die Nebenbedingung

$$\sum_{j=1}^{n} \beta_j x_j \leq \beta \tag{11.5}$$

heißt Schnittbedingung, wenn folgendes erfüllt ist:

i) Es sei $\mathbf{x}^{(0)}$ ein Optimum mit den Nebenbedingungen (11.1) – (11.2), aber nicht (11.3). Dann erfüllt $\mathbf{x}^{(0)}$ die Bedingung (11.5) nicht, das heißt

$$\sum_{j=1}^{n} \beta_j x_j^{(0)} > \beta.$$

ii) Jede Lösung, welche die Nebenbedingungen (11.1) – (11.3) erfüllt, erfüllt auch (11.5), das heißt

$$\{\mathbf{x} \ : \ \mathbf{x} \text{ erfüllt } \ (11.1) - (11.3)\} \subset \{\mathbf{x} \ : \ \mathbf{x} \text{ erfüllt } \ (11.5)\}.$$

□

Bemerkung 11.6 Vorgehen. Mit Hilfe der Schnittbedingung soll der zulässige Bereich verkleinert werden (eine Ecke wird abgeschnitten) ohne dass damit ganzzahlige Lösungen abgeschnitten werden.

Jetzt soll (11.1) – (11.4) durch die Einführung von Schnittbedingungen gelöst werden. Es sei dazu zuerst (11.1) – (11.2) mit Hilfe der Simplexmethode gelöst. Das Optimum sei $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_m^{(0)}, 0, \ldots, 0)^T$. Dabei sei wenigstens ein $x_i^{(0)}$, $i \in \{1, \ldots, m\}$, nicht ganz.

Sei $A_B = (\mathbf{a}_1, \ldots, \mathbf{a}_m)$ die Matrix der Basisvektoren. Die Auflösung von (11.1) nach den Basisvariablen liefert für jedes zulässige $\mathbf{x}$ die Darstellung

$$x_i = \alpha_i + \alpha_{i,m+1}(-x_{m+1}) + \ldots + \alpha_{i,n}(-x_n), \quad i = 1, \ldots, m. \tag{11.6}$$

□

Lemma 11.7 *Die Koeffizienten von (11.6) sind die Koeffizienten einer optimalen Simplextabelle.*

Beweis: Wir betrachten die Nebenbedingung (11.1) und zerlegen $A = (A_B|A_N)$ sowie $\mathbf{x} = (\mathbf{x}_B|\mathbf{x}_N)^T$. Wir wissen, dass $A_N = A_B X$, wobei X die Einträge der Simplextabelle sind. Aus

$$A_B\mathbf{x}_B + A_N\mathbf{x}_N = \mathbf{b}$$

folgt

$$\mathbf{x}_B = A_B^{-1}\mathbf{b} + A_B^{-1}A_N(-\mathbf{x}_N) = \underbrace{A_B^{-1}\mathbf{b}}_{\alpha_i} + \underbrace{X(-\mathbf{x}_N)}_{\text{Rest von (11.6)}} .$$

■

Bemerkung 11.8 Weiteres Vorgehen. Sei jetzt für $i = p$ die Variable $x_p^{(0)}$ nicht ganz, $i \in \{1, \ldots, m\}$. Falls mehrere $x_i^{(0)}$ nicht ganz sind, wähle man einen dieser Indizes. Welches der beste ist, ist im allgemeinen nicht zu beantworten.

Sei $a \in \mathbb{R}$. Dann bezeichnen wir

$$[a] = \mathrm{INT}(a), \quad \{a\} = a - [a]\,,$$

wobei $\mathrm{INT}(a)$ der größte ganzzahlige Bestandteil von a ist. Es gilt $\{a\} \in [0, 1)$. Insbesondere gilt $\left\{x_p^{(0)}\right\} > 0$. Da für das Optimum die Komponenten mit den Indizes $m+1, \ldots, n$ verschwinden und wegen (11.6) folgt damit

$$\alpha_p = x_p^{(0)} = \underbrace{\left[x_p^{(0)}\right]}_{\geq 0} + \underbrace{\left\{x_p^{(0)}\right\}}_{>0} > 0.$$

□

Satz 11.9 *Die Bedingung*

$$s_1 = -\{\alpha_p\} - \{\alpha_{p,m+1}\}(-x_{m+1}) - \ldots - \{\alpha_{p,n}\}(-x_n), \quad s_1 \geq 0 \tag{11.7}$$

stellt eine Schnittbedingung gemäß Definition 11.5 dar.

Beweis: Die Bedingungen von Definition 11.5 müssen geprüft werden. Wir fügen die Bedingung (11.7) zum System der Nebenbedingungen (11.1), (11.2) hinzu und setzen $\mathbf{x}^{(0)}$ ein. Aus (11.6) und wegen $x_{m+1}^{(0)} = \ldots = x_n^{(0)} = 0$ folgt

$$s_1 = -\alpha_p < 0,$$

also ist $\mathbf{x}^{(0)}$ nicht zulässig.

Nun ist zu zeigen, dass mit (11.7) kein bezüglich (11.1) – (11.3) zulässiger Punkt weggeschnitten wird. Sei $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_n)^T$ ein Punkt, der (11.1) – (11.3) erfüllt, also insbesondere ganzzahlige Komponenten besitzt. Dann folgt aus (11.7)

$$s_1 = -\underbrace{(\alpha_p - [\alpha_p])}_{\in[0,1)} - \sum_{j=m+1}^{n} \underbrace{(\alpha_{p,j} - [\alpha_{p,j}])}_{\in(0,1)} \underbrace{(-\tilde{x}_j)}_{\leq 0}.$$

Damit ist $s_1 > -1$. Andererseits gilt

$$s_1 = \underbrace{[\alpha_p] + \sum_{j=m+1}^{n} [\alpha_{p,j}]\,(-\tilde{x}_j)}_{\in\mathbb{Z}} \underbrace{-\alpha_p - \sum_{j=m+1}^{n} \alpha_{p,j}\,(-\tilde{x}_j)}_{(11.6)=-\tilde{x}_p\in\mathbb{Z}}.$$

Damit ist $s_1 \in \mathbb{Z}$. Da $s_1 > -1$ folgt $s_1 \geq 0$. ■

Bemerkung 11.10 Zusammenfassung. Man hat mit der Schnittbedingung (11.7) das Optimum bezüglich der Nebenbedingungen (11.1), (11.2) abgeschnitten, ohne dabei auch ganzzahlige Lösungen wegzuschneiden. Folgendes lineare Programm ist jetzt zu lösen:

$$\begin{aligned} z = \mathbf{c}^T\mathbf{x} &\rightarrow \text{min !} \\ A\mathbf{x} &= \mathbf{b} \\ \sum_{j=m+1}^{n} \{\alpha_{p,j}\}\,(-x_j) + s_1 &= -\{\alpha_p\} \qquad (11.8) \\ \mathbf{x} &\geq \mathbf{0} \\ s_1 &\geq 0. \end{aligned}$$

Zur Lösung von (11.8) berechnet man zuerst die optimale Lösung des linearen Programms ohne Ganzzahligkeitsbedingung mit Hilfe der dualen Simplexmethode. Hat man diese, und ist sie nicht ganzzahlig, betrachtet man im nächsten Schritt die duale Simplextabelle mit

$$x_i = \alpha_i, \quad i = 1, \ldots, m, \quad s_1 = -\alpha_p \qquad (11.9)$$

(beachte: in der Simplextabelle des dualen Problems steht $x_i = \mathbf{b}_i\mathbf{b} = \alpha_i$). Der Vektor (11.9) ist eine dual zulässige Lösung, das heißt, die Nebenbedingungen des dualen Problems sind erfüllt. *Übungsaufgabe* Eventuell ist die Einführung weiterer Schnittbedingungen nötig. Das oben beschriebene Vorgehen wird im nächsten Beispiel demonstriert. □

Beispiel 11.11 Wir betrachten das lineare Programm

$$\begin{aligned} z = -x_1 - 2x_2 &\rightarrow \text{min !} \\ \begin{pmatrix} 2 & 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_5 \end{pmatrix} &= \begin{pmatrix} 10 \\ 5 \\ 4 \end{pmatrix} \\ \mathbf{x} &\geq \mathbf{0} \\ \mathbf{x} &\text{ ganz.} \end{aligned}$$

Die optimale duale Simplextabelle lautet

			3	4
i	c_i	Lösung	0	0
1	-1	5/3	1/3	-1/3
2	-2	**20/3**	1/3	2/3
5	0	7/3	-1/3	1/3
		-15	-1	-1

Die Lösung ist optimal, aber nicht ganzzahlig. Heuristisch wählt man $\{\alpha_p\}$ möglichst groß, hier zum Beispiel $\{20/3\} = 2/3$. Es besteht allerdings die Gefahr, dass man an der falschen Stelle abschneidet. Nun verwendet man, dass die Tabelle der dualen Simplexmethode eine Optimaltabelle der primalen Aufgabe ist, Bemerkung 10.8. Damit kann man Lemma 11.7 für die Formulierung der Schnittbedingung nutzen, da die benötigten Koeffizienten $\alpha_{p,j}$ gerade im Nichtbasisteil der ausgewählten Zeile stehen:

$$s_1 = -\frac{2}{3} - \frac{1}{3}(-x_3) - \frac{2}{3}(-x_4) \geq 0.$$

Führt man die Variable s_1 als $m+1$–ste Eckvariable in die duale Simplextabelle ein, dann hat man die neue Matrix der Basisvektoren

$$\tilde{A}_B = \begin{pmatrix} \mathbf{a}_1 & \dots & \mathbf{a}_m & \mathbf{0} \\ 0 & \dots & 0 & 1 \end{pmatrix} \quad \Longrightarrow \quad \tilde{A}_B^{-1} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_m \\ \mathbf{e}_{m+1}^T \end{pmatrix},$$

wobei $\mathbf{e}_{m+1}$ der $m+1$–ste Einheitsvektor ist. Somit erhält man in der Spalte Lösung für s_1 : $\mathbf{e}_{m+1}^T \mathbf{b} = -\alpha_p = -2/3$. In den Nichteckspalten der Zeile von s_1 erhält man die Zahlen $-\{\alpha_{p,j}\}$, siehe (11.8). Man hat die neue duale Simplextabelle

			3	4
i	c_i	Lösung	0	0
1	-1	5/3	1/3	-1/3
2	-2	20/3	1/3	2/3
5	0	7/3	-1/3	1/3
s_1	0	-2/3	-1/3	**-2/3**
		-15	-1	-1

Die Hauptzeile ist die Zeile von s_1. Aus

$$\theta = \min_{j \in \{3,4\}} \left\{ \frac{-1}{-1/3}, \frac{-1}{-2/3} \right\} = \frac{3}{2}$$

folgt, dass $k = 4$ die Hauptspalte ist. Der Simplexschritt führt zu folgender Tabelle

			3	s_1
i	c_i	Lösung	0	0
1	-1	2	-1/2	-1/2
2	-2	6	0	1
5	0	2	-1/2	1/2
4	0	1	1/2	-3/2
		-14	-1/2	-3/2

Damit hat man die Optimallösung des ganzzahligen linearen Programms gefunden.

In der Praxis sind im allgemeinen mehr Schnittbedingungen nötig. Die Endlichkeit des Verfahrens ist nicht gesichert. □

Kapitel 12

Innere–Punkt–Verfahren

Innere–Punkt–Verfahren verfolgen die Idee, bei der Lösung von linearen Programmen durch das Innere des konvexen Polyeders zum Optimum zu gelangen. Damit unterscheiden sie sich grundsätzlich von der Simplexmethode.

Bemerkung 12.1 Historie von Innere–Punkt–Verfahren.

- Dikin 1967: hat bereits die Idee von Karmarkar (1984) umgesetzt, Arbeit wurde aber nicht wahrgenommen.
- Fiacco, McCormick 1968: Innere–Punkt–Verfahren für nichtlineare Optimierungsprobleme, waren aber nicht wettbewerbsfähig im Vergleich zu anderen Verfahren.
- Khachian 1979: Ellipsoidmethode. Polynomiale Komplexität konnte bewiesen werden, allerdings war das Verfahren wenig praxistauglich wegen numerischen Instabilitäten.
- Karmarkar 1984: projektive Methode, erste praxistaugliche Innere–Punkt–Methode.

□

Bemerkung 12.2 Motivationen. Die Motivation zur Entwicklung von Alternativen zur Simplexmethode liegt in deren schlechter theoretischer Komplexität begründet. Man kann nicht ausschließen, dass die Anzahl der Iterationen exponentiell mit der Problemgröße wächst, siehe Abschnitt 8.2. Das hat im wesentlichen zwei Gründe:

- Die Simplexmethode kennt als Abstiegsrichtungen nur die Kanten auf dem Rand des zulässigen Bereiches. Deren Anzahl wächst exponentiell mit der Dimension des Polyeders.
- Die Simplexmethode sucht die Abstiegsrichtung lokal aus. Für diese Richtung spielen nur die im aktuellen Eckpunkt aktiven Nebenbedingungen eine Rolle, nicht aber die Gesamtheit der Nebenbedingungen.

Alternativen zur Simplexmethode müssen an diesen beiden Schwachstellen angreifen: sie sollten Abstiegsrichtungen durch das Innere des Polyeders zulassen und bei der Bestimmung dieser Richtungen Informationen von allen Nebenbedingungen einbeziehen. □

Bemerkung 12.3 Ellipsoidmethoden. Die Idee von Khachians Ellipsoidmethode lässt sich grob wie folgt beschreiben: Man startet mit einem zulässigen Punkt $\overline{\mathbf{x}}$ und sucht einen möglichst zentral gelegenen Punkt $\tilde{\mathbf{x}}$ in einem Restpolyeder, welches nur Punkte $\mathbf{x}$ mit $\mathbf{c}^T\mathbf{x} \leq \mathbf{c}^T\overline{\mathbf{x}}$ enthält. Diesen Punkt findet man dadurch, dass man mittels einer Iteration das Restpolyeder möglichst gut in ein Ellipsoid einbettet und dessen Mittelpunkt $\tilde{\mathbf{x}}$ betrachtet. Ist $\tilde{\mathbf{x}}$ zulässig, setzt man $\overline{\mathbf{x}} = \tilde{\mathbf{x}}$ und startet die

Konstruktion von Neuem. Die Instabilitäten rüherten daher, dass die Ellipsoide im Laufe der Iteration immer schmaler wurden.

Das Verfahren von Karmarkar geht andersherum vor: hier wird ein Ellipsoid in den zulässigen Bereich eingebettet. Es funktioniert grob wie folgt: Man startet mit einem inneren Punkt des zulässigen Bereiches $\bar{\mathbf{x}}$ und konstruiert ein im zulässigen Bereich einbeschriebenes, möglichst großes Ellipsoid. Dann nimmt man anstelle des zulässigen Bereiches nur das Ellipsoid und minimiert darüber die Zielfunktion. Es ergibt sich ein Randpunkt $\tilde{\mathbf{x}}$, an dem das Minimum angenommen wird. Man setzt $\overline{\mathbf{x}} = \tilde{\mathbf{x}}$ und startet die Konstruktion von Neuem.

Das Verfahren von Karmarkar arbeitet praktisch wesentlich besser als die Ellipsoidmethode von Khachian. Man kann auch beweisen, dass die Komplexität des Verfahrens von Karmarkar besser ist. Die Komplexität ist also insbesondere polynomial. Für sehr große Problem ist dieses Verfahren oft schneller als die Simplexmethode. □

In diesem Kapitel soll jedoch nur eine einfache Innere–Punkt–Methode im Detail besprochen werden.

12.1 Das Newton–Verfahren

Die betrachtete Innere–Punkt–Methode beruht auf dem Newton–Verfahren. Hier werden nur kurz einige Fakten zu diesem Verfahren zusammgestellt.

Bemerkung 12.4 Herleitung. Sei $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ mit $\mathbf{f} \in C^1(\mathbb{R}^n)$ gegeben. Gesucht ist eine Nullstelle $\overline{\mathbf{x}}$ dieser Funktion $\mathbf{f}(\overline{\mathbf{x}}) = \mathbf{0}$. Sei $\mathbf{x}$ eine Näherung an $\overline{\mathbf{x}}$. Setzt man $\overline{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$, dann erhält man mit der Taylorentwicklung, abgebrochen nach dem linearen Term, die Approximation

$$\mathbf{0} = \mathbf{f}(\overline{\mathbf{x}}) = \mathbf{f}(\mathbf{x} + \Delta\mathbf{x}) \approx \mathbf{f}(\mathbf{x}) + D\mathbf{f}(\mathbf{x})\Delta\mathbf{x},$$

wobei $D\mathbf{f}(\mathbf{x})$ die Jacobi–Matrix von $\mathbf{f}(\mathbf{x})$ in $\mathbf{x}$ ist. Unter der Annahme, dass die Jacobi–Matrix regulär ist, erhält man

$$\Delta\mathbf{x} \approx -\left(D\mathbf{f}(\mathbf{x})\right)^{-1}\mathbf{f}(\mathbf{x}) \quad \Longrightarrow \quad \overline{\mathbf{x}} \approx \mathbf{x} - \left(D\mathbf{f}(\mathbf{x})\right)^{-1}\mathbf{f}(\mathbf{x}).$$

Diese Beziehung motiviert die Berechnung der nächsten Iterierten $\mathbf{x}^+$ des Newton–Verfahrens

$$\mathbf{x}^+ := \mathbf{x} - \left(D\mathbf{f}(\mathbf{x})\right)^{-1}\mathbf{f}(\mathbf{x}).$$

□

Satz 12.5 Konvergenzverhalten des Newton–Verfahrens. *Seien* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ *mit* $\mathbf{f} \in C^3(\mathbb{R}^n)$, $\overline{\mathbf{x}}$ *eine Nullstelle von* $\mathbf{f}(\mathbf{x})$ *und* $|D\mathbf{f}(\overline{\mathbf{x}})| \neq 0$ *(Determinante). Dann gibt es ein* $\varepsilon > 0$*, so dass das Newton–Verfahren für alle Startwerte* $\mathbf{x}^{(0)}$ *mit* $\left\|\overline{\mathbf{x}} - \mathbf{x}^{(0)}\right\|_2 \leq \varepsilon$ *quadratisch gegen* $\overline{\mathbf{x}}$ *konvergiert.*

Bemerkung 12.6 Interpretation. Der Satz besagt zum einen, dass das Newton–Verfahren lokal konvergent ist, das heißt, es konvergiert, wenn man nahe genug an der Lösung beginnt. Quadratische Konvergenz bedeutet, dass es Konstanten $c > 0$ und $k_0 > 0$ gibt, so dass

$$\left\|\overline{\mathbf{x}} - \mathbf{x}^{(k+1)}\right\|_2 \leq c\left\|\overline{\mathbf{x}} - \mathbf{x}^{(k)}\right\|_2^2$$

für alle $k \geq k_0$. □

12.2 Ein Kurz–Schritt–Algorithmus

Bemerkung 12.7 Überführung des Optimierungsproblems in ein äquivalentes nichtlineares Gleichungssystem. Wir betrachten das lineare Programm

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ z = \mathbf{c}^T\mathbf{x} \ : \quad A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0} \right\} \tag{12.1}$$

mit dem zugehörigen dualen Programm

$$\max_{\mathbf{y}\in\mathbb{R}^m} \left\{ \tilde{z} = \mathbf{b}^T\mathbf{y} \ : \ A^T\mathbf{y} + \mathbf{s} = \mathbf{c}, \ \mathbf{s} \geq \mathbf{0} \right\}. \tag{12.2}$$

Im dualen Programm haben wir hierbei die Schlupfvariablen $\mathbf{s}$ eingeführt, um Gleichungsnebenbedingungen zu erhalten.

Innere–Punkte–Verfahren erzeugen eine Folgen von Punkten $\{(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}, \mathbf{s}^{(k)})\}$, $k = 0, 1, \ldots$, mit $\mathbf{x}^{(k)} \geq \mathbf{0}$, $\mathbf{s}^{(k)} \geq \mathbf{0}$, deren Grenzwerte Optimallösungen von (12.1) und (12.2) liefern. Die Verfahren nennt man zulässige–Innere–Punkte–Verfahren, wenn alle Punkte $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}, \mathbf{s}^{(k)})$ zulässige innere Punkte von (12.1) und (12.2) sind, das heißt es gilt

$$A\mathbf{x}^{(k)} = \mathbf{b}, \mathbf{x}^{(k)} > \mathbf{0}, \quad A^T\mathbf{y}^{(k)} + \mathbf{s}^{(k)} = \mathbf{c}, \ \mathbf{s}^{(k)} > \mathbf{0}.$$

Um zulässige–Innere–Punkte–Verfahren verwenden zu können, müssen wir natürlich voraussetzen, dass die obigen Mengen nicht leer sind. Neben diesen Verfahren gibt es auch unzulässige–Innere–Punkte–Verfahren.

Aus der Nichtnegativität von $\mathbf{x}$ und $\mathbf{s}$ folgt

$$0 \leq \mathbf{x}^T\mathbf{s} = \mathbf{x}^T\left(\mathbf{c} - A^T\mathbf{y}\right) = \mathbf{c}^T\mathbf{x} - (A\mathbf{x})^T\mathbf{y} = \mathbf{c}^T\mathbf{x} - \mathbf{b}^T\mathbf{y} = z - \tilde{z}.$$

Von Satz 9.5 (Starker Dualitätssatz) wissen wir, dass für die Optimallösungen von (12.1) und (12.2) gilt $z = \tilde{z}$. Also folgt im Optimum $\mathbf{x}^T\mathbf{s} = 0$ und wegen Nichtnegativität dieser beiden Vektoren sogar $x_i s_i = 0$ für alle $i = 1, \ldots, n$. Zusammen mit den Nebenbedingungen von (12.1) und (12.2) sind die Optimallösungen von (12.1) und (12.2) Lösungen des nichtlinearen Systems

$$\Psi_0(\mathbf{x}, \mathbf{y}, \mathbf{s}) := \begin{pmatrix} A\mathbf{x} - \mathbf{b} \\ A^T\mathbf{y} + \mathbf{s} - \mathbf{c} \\ X\mathbf{s} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{x}, \mathbf{s} \geq \mathbf{0}, \tag{12.3}$$

wobei $X := \mathrm{diag}(\mathbf{x}) \in \mathbb{R}^{n\times n}$ ist. Die Nichtlinearität ist in der letzten Gleichung, welche ausgeschrieben $x_i s_i = 0$, $i = 1, \ldots, n$, bedeutet. Die Jacobi–Matrix von Ψ_0 ist gegeben durch

$$D\Psi_0(\mathbf{x}, \mathbf{y}, \mathbf{s}) = \begin{pmatrix} A & 0 & 0 \\ 0 & A^T & I_n \\ S & 0 & X \end{pmatrix} \in \mathbb{R}^{(2n+m)\times(2n+m)} \tag{12.4}$$

mit $S := \mathrm{diag}(\mathbf{s}) \in \mathbb{R}^{n\times n}$ und I_n der n–dimensionalen Einheitsmatrix. □

Lemma 12.8 Regularität der Jacobi–Matrix. *Unter den Voraussetzungen $rg(A) = m$ und $\mathbf{x} > \mathbf{0}$, $\mathbf{s} > \mathbf{0}$ ist die Jacobi–Matrix (12.4) regulär.*

Beweis: Indirekt. Sei $(\mathbf{u}^T, \mathbf{v}^T, \mathbf{w}^T)^T \neq \mathbf{0}$ ein Vektor mit

$$\begin{pmatrix} A & 0 & 0 \\ 0 & A^T & I_n \\ S & 0 & X \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

Aus der ersten Gleichung folgt $A\mathbf{u} = \mathbf{0}$. Diese Beziehung wird in die umgestellte zweite Gleichung eingesetzt

$$\mathbf{w} = -A^T\mathbf{v} \implies \mathbf{u}^T\mathbf{w} = -\mathbf{u}^T A^T\mathbf{v} = -\left(A\mathbf{u}\right)^T\mathbf{v} = 0.$$

Aus der dritten Gleichung folgt $\mathbf{u} = -S^{-1}X\mathbf{w}$. Die Invertierbarkeit von S folgt aus $\mathbf{s} > \mathbf{0}$. Mit der eben bewiesenen Beziehung und der Symmetrie der Diagonalmatrizen ergibt sich

$$0 = \mathbf{u}^T\mathbf{w} = -\mathbf{w}^T X S^{-1}\mathbf{w}.$$

Da $\mathbf{x}, \mathbf{s} > \mathbf{0}$, folgt daraus $\mathbf{w} = \mathbf{0}$. Aus der dritten Gleichung folgt damit $\mathbf{u} = \mathbf{0}$. Damit vereinfacht sich die zweite Gleichung zu $A^T\mathbf{v} = \mathbf{0}$. Wegen des vollen Spaltenrangs von A^T folgt daraus $\mathbf{v} = \mathbf{0}$. Damit ist der Widerspruch zur Annahme konstruiert. ■

Bemerkung 12.9 Das Verfahren. Auf Grund dieses Satzes liegt es nahe, die Lösung des Systems (12.3) mit dem Newton–Verfahren zu versuchen. Dabei gibt es allerdings einige Schwierigkeiten:

- Die Regularität der Jacobi–Matrix ist nur für innere Punkte ($\mathbf{x}, \mathbf{s} > \mathbf{0}$) bewiesen. Die gesuchte Lösung liegt aber nicht im Inneren, da aus der dritten Gleichung $x_i s_i = 0$, $i = 1, \ldots, n$, folgt, dass mindestens eine der Variablen x_i oder s_i im Optimum verschwindet.
- Es ist nicht klar, ob die in einem Newton–Schritt berechnete neue Iterierte überhaupt ein zulässiger Punkt ist. Falls nicht, ist die Regularität der Jacobi–Matrix nicht gesichert. Außerdem kann es passieren, dass das Newton–Verfahren gar nicht konvergiert oder zu einer Lösung konvergiert, welche die Nichtnegativitätsbedingung nicht erfüllt.

Um doch noch ein Newton–ähnliches Verfahren für (12.3) nutzen zu können, modifiziert man dieses System. Seien $\mathbf{e}_n = (1, \ldots, 1)^T \in \mathbb{R}^n$ und sei $\mu > 0$ ein Parameter. Anstelle von (12.3) betrachtet man nun

$$\Psi_\mu(\mathbf{x}, \mathbf{y}, \mathbf{s}) := \begin{pmatrix} A\mathbf{x} - \mathbf{b} \\ A^T\mathbf{y} + \mathbf{s} - \mathbf{c} \\ X\mathbf{s} - \mu\mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{x}, \mathbf{s} > \mathbf{0}. \tag{12.5}$$

Die Jacobi–Matrix dieses Systems ist (12.4).

Das schwerere Problem, eine Nullstelle von (12.3) zu finden, ersetzt man durch das einfachere Problem (12.5). Für festes $\mu > 0$ kann die Lösung nur im Inneren des zulässigen Bereichs liegen, da $x_i s_i = \mu$ für alle i. Da die Jacobi–Matrix nach Lemma 12.8 dort regulär ist, gibt es eine Umgebung der Lösung $(\mathbf{x}(\mu)^T, \mathbf{y}(\mu)^T, \mathbf{s}(\mu)^T)^T$, in der das Newton–Verfahren ohne Berücksichtigung der Ungleichungen $\mathbf{x}, \mathbf{s} > \mathbf{0}$ quadratisch konvergiert. In diesem Sinne hat man das Problem (12.3) mit den Nichtnegativitätsbedingungen durch ein System ohne Nebenbedingungen ersetzt.

Da das Optimum von (12.1) auf dem Rand angenommen wird ist zu erwarten, dass die Lösung $(\mathbf{x}(\mu)^T, \mathbf{y}(\mu)^T, \mathbf{s}(\mu)^T)^T$ umso weiter vom Rand entfernt liegt, je größer μ ist. Das Vorgehen ist:

- Man beginnt mit einem großen Wert μ_0, für den sich die Lösung leicht berechnen lässt. Für große μ_0 erwartet man ein großes Konvergenzgebiet.
- Dann wird μ sukzessive verkleinert und jedesmal wird die zugehörige Lösung bis auf eine gewisse Genauigkeit mit dem Newton–Verfahren approximiert.

Die Punktmenge

$$\left\{(\mathbf{x}(\mu)^T, \mathbf{y}(\mu)^T, \mathbf{s}(\mu)^T)^T \;:\; \mu > 0\right\}$$

wird zentraler Pfad genannt. Für seine Punkte existiert eine sogenannte Dualitätslücke:

$$\begin{aligned} 0 < \mu n &= \mathbf{x}(\mu)^T\mathbf{s}(\mu) = \mathbf{x}(\mu)^T\left(\mathbf{c} - A^T\mathbf{y}(\mu)\right) = \mathbf{c}^T\mathbf{x}(\mu) - \left(A\mathbf{x}(\mu)\right)^T\mathbf{y}(\mu) \\ &= \mathbf{c}^T\mathbf{x}(\mu) - \mathbf{b}^T\mathbf{y}(\mu) = z_\mu - \tilde{z}_\mu. \end{aligned} \tag{12.6}$$

□

Lemma 12.10 Lösbarkeit des modifizierten Systems. *Sei die dual zulässige Menge* $\{\mathbf{y} \,:\, A^T\mathbf{y} \leq \mathbf{c}\}$ *beschränkt und sei das Innere dieser Menge nichtleer. Dann besitzt das Problem (12.5) für jedes* $\mu > 0$ *eine eindeutige Lösung.*

Beweis: Literatur, zum Beispiel [JS04, S 75ff.]. ■

Bemerkung 12.11 Analyse eines Newton–Schrittes. Sei $(\mathbf{x}^T, \mathbf{y}^T, \mathbf{s}^T)^T$ die gegenwärtige Iterierte zur Lösung von (12.5) für festes $\mu > 0$. Die Korrektur für die nächste Iterierte des Newton–Verfahrens berechnet sich aus der Lösung von

$$\begin{pmatrix} A & 0 & 0 \\ 0 & A^T & I_n \\ S & 0 & X \end{pmatrix} \begin{pmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \\ \Delta\mathbf{s} \end{pmatrix} = - \begin{pmatrix} A\mathbf{x} - \mathbf{b} \\ A^T\mathbf{y} + \mathbf{s} - \mathbf{c} \\ X\mathbf{s} - \mu\mathbf{e} \end{pmatrix}. \tag{12.7}$$

Nach der Lösung dieses Systems folgt, dass die neue Iterierte $((\mathbf{x} + \Delta\mathbf{x})^T, (\mathbf{y} + \Delta\mathbf{y})^T, (\mathbf{s} + \Delta\mathbf{s})^T)^T$ die ersten zwei Gleichungen von (12.5) erfüllt, da diese linear sind. *Übungsaufgabe* Wir können annehmen, dass die gegenwärtige Iterierte bereits durch einen Newton–Schritt berechnet wurde, sie damit ebenfalls diese Eigenschaft besitzt und sich (12.7) vereinfacht zu

$$\begin{pmatrix} A & 0 & 0 \\ 0 & A^T & I_n \\ S & 0 & X \end{pmatrix} \begin{pmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \\ \Delta\mathbf{s} \end{pmatrix} = - \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ X\mathbf{s} - \mu\mathbf{e} \end{pmatrix} =: \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ -\mathbf{r} \end{pmatrix}. \tag{12.8}$$

Der Vektor $\mathbf{r}$ wird Residuum genannt. Das Residuum der neuen Newton–Iterierten ist, unter Verwendung der Beziehungen von (12.8),

$$\begin{aligned} \tilde{\mathbf{r}} &= (X + \Delta X)(\mathbf{s} + \Delta\mathbf{s}) - \mu\mathbf{e} = X\mathbf{s} + \underbrace{X\Delta\mathbf{s} + S\Delta\mathbf{x}}_{=-\mathbf{r}} + \Delta X\Delta\mathbf{s} - \mu\mathbf{e} \\ &= X\mathbf{s} - \mathbf{r} + \Delta X\Delta\mathbf{s} - \mu\mathbf{e} = \Delta X\Delta\mathbf{s}. \end{aligned} \tag{12.9}$$

Um den Newton–Schritt analysieren zu können, werden einige Bezeichungen eingeführt

$$D := \sqrt{XS^{-1}}, \quad \mathbf{q} := DX^{-1}\mathbf{r}, \quad R := \operatorname{diag}(\mathbf{r}).$$

Direktes Nachrechnen oder Einsetzen zeigt, dass man die Lösung von (12.8) wie folgt aufschreiben kann: *Übungsaufgabe*

$$\begin{aligned} \Delta\mathbf{y} &= (AD^2A^T)^{-1}AD\mathbf{q}, \\ \Delta\mathbf{x} &= D^2A^T\Delta\mathbf{y} - D\mathbf{q}, \\ \Delta\mathbf{s} &= -D^{-1}\mathbf{q} - D^{-2}\Delta\mathbf{x}. \end{aligned}$$

Die Matrix

$$DA^T(AD^2A^T)^{-1}AD$$

definiert die Orthogonalprojektion auf den Bildraum von DA^T. Diese Projektion wird Π_R genannt. Demzufolge ist die Projektion in den Nullraum von DA^T gegeben durch $\Pi_N := I - \Pi_R$, wobei I die identische Abbildung ist. Nun kann man zeigen (*Übungsaufgabe*), dass

$$\Delta\mathbf{x} = -D\Pi_N\mathbf{q}, \quad \Delta\mathbf{s} = -D^{-1}\Pi_R\mathbf{q}$$

gelten. Das sind die beiden Ausdrücke, die wir in (12.9) analysieren müssen.

Wir nehmen nun an, dass man für gegebene $\mathbf{x} > \mathbf{0}$, $\mathbf{s} > \mathbf{0}$ und $\mathbf{y}$ ein $\beta \in [0, 1/2]$ finden kann, für welches

$$\|\mathbf{r}\|_2 \leq \beta\mu \tag{12.10}$$

gilt. Das heißt, der gegebene Punkt ist in einer gewissen Umgebung des zentralen Pfades. Für $\beta = 0$ muss er direkt auf dem zentralen Pfad sein.

Wegen $\mathbf{r} = X\mathbf{s} - \mu\mathbf{e}$ gilt

$$DX^{-1} = \sqrt{XS^{-1}X^{-2}} = \sqrt{X^{-1}S^{-1}} = \sqrt{(R+\mu I)^{-1}} = \left(\sqrt{R+\mu I}\right)^{-1}.$$

Bei dieser Rechnung und auch bei folgenden Rechnungen wird stark ausgenutzt, dass hier die Matrizen D, R, S und X Diagonalmatrizen sind. Für Nichtdiagonalmatrizen funktionieren die Rechnungen nicht mehr. Es folgt

$$\left\|DX^{-1}\right\|_2^2 = \left\|(R+\mu I)^{-1}\right\|_2 = \max_{1\leq i\leq n} \frac{1}{|r_i+\mu|} \leq \frac{1}{\mu(1-\beta)}. \tag{12.11}$$

Hierbei wurde die Definition der Spektralnorm einer Matrix ausgenutzt, sowie dass wir es mit einer Diagonalmatrix zu tun haben. Die letzte Ungleichung folgt aus der Dreiecksungleichung und (12.10)

$$|r_i + \mu| \geq \mu - |r_i| \geq \mu - \beta\mu = \mu(1-\beta).$$

Aus der Verträglichkeit der Euklidischen Vektornorm und der Spektralnorm von Matrizen sowie (12.10) folgt damit

$$\tilde{\beta} := \|\mathbf{q}\|_2 = \left\|DX^{-1}\mathbf{r}\right\|_2 \leq \left\|DX^{-1}\right\|_2 \|\mathbf{r}\|_2 \leq \beta\sqrt{\frac{\mu}{(1-\beta)}}.$$

Sei θ der Winkel zwischen $\mathbf{q}$ und $\Pi_N\mathbf{q}$. Dann folgt mit der Orthogonalität der Projektionen

$$\cos(\theta) = \frac{(\mathbf{q}, \Pi_N\mathbf{q})}{\|\mathbf{q}\|_2 \|\Pi_N\mathbf{q}\|_2} = \frac{(\Pi_R\mathbf{q} + \Pi_N\mathbf{q}, \Pi_N\mathbf{q})}{\|\mathbf{q}\|_2 \|\Pi_N\mathbf{q}\|_2} = \frac{\|\Pi_N\mathbf{q}\|_2^2}{\|\mathbf{q}\|_2 \|\Pi_N\mathbf{q}\|_2} = \frac{\|\Pi_N\mathbf{q}\|_2}{\|\mathbf{q}\|_2}.$$

Es folgen

$$\begin{aligned} \left\|D^{-1}\Delta\mathbf{x}\right\|_2 &= \|\Pi_N\mathbf{q}\|_2 = \cos(\theta)\|\mathbf{q}\|_2 = \tilde{\beta}\cos(\theta), \\ \|D\Delta\mathbf{s}\|_2 &= \|\Pi_R\mathbf{q}\|_2 = \|(I-\Pi_N)\mathbf{q}\|_2 = \sin(\theta)\|\mathbf{q}\|_2 = \tilde{\beta}\sin(\theta). \end{aligned} \tag{12.12}$$

Diese Abschätzungen kann man nun in (12.9) einsetzen. Mit der Cauchy–Schwarz–Ungleichung und $\beta \leq 1/2$ erhält man

$$\begin{aligned} \|\tilde{\mathbf{r}}\|_2 &= \|\Delta X\Delta\mathbf{s}\|_2 \leq \left\|D^{-1}\Delta X\right\|_2 \|D\Delta\mathbf{s}\|_2 \leq \tilde{\beta}^2\cos(\theta)\sin(\theta) \\ &= \frac{\tilde{\beta}^2}{2}\sin(2\theta) \leq \frac{\tilde{\beta}^2}{2} \leq \beta^2\frac{\mu}{2(1-\beta)} \leq \beta^2\mu. \end{aligned} \tag{12.13}$$

Wir haben damit erhalten, dass unter der Voraussetzung (12.10) der relative Fehler $\|X\mathbf{s} - \mu\mathbf{e}\|_2 / \mu = \|\mathbf{r}\|_2 / \mu$ in jedem Schritt des Newton–Verfahrens quadriert wird.

Aus (12.12) folgt insbesondere $\left\|D^{-1}\Delta\mathbf{x}\right\|_2 \leq \tilde{\beta}$ und daraus, sowie mit (12.11), der Definition von $\tilde{\beta}$ und $\beta \leq 1/2$,

$$\begin{aligned} \left\|X^{-1}\Delta\mathbf{x}\right\|_2 &= \left\|X^{-1}DD^{-1}\Delta\mathbf{x}\right\|_2 \leq \left\|X^{-1}D\right\|_2 \left\|D^{-1}\Delta\mathbf{x}\right\|_2 \leq \frac{\tilde{\beta}}{\sqrt{\mu(1-\beta)}} \\ &= \frac{\beta}{1-\beta} \leq 1. \end{aligned}$$

Somit ist $(\Delta x_i)/x_i \leq 1$ woraus

$$x_i - |\Delta x_i| \geq x_i - x_i = 0, \quad i = 1, \ldots, n$$

folgt. Eine analoge Abschätzung kann man für $\mathbf{s}$ durchführen. Die Nichtnegativität der neuen Iterierten ist gesichert.

Es gilt sogar die Positivität. *Übungsaufgabe ab hier.* Mit den Abschätzungen (12.13) und (12.9) folgt

$$\|\tilde{\mathbf{r}}\|_2 < \mu \quad \Longrightarrow \quad \|\Delta X \Delta \mathbf{s}\|_2 < \mu \quad \Longrightarrow \quad |\Delta \mathbf{x}_i \Delta \mathbf{s}_i| < \mu$$

für alle $i = 1, \ldots, n$. Aus (12.9) erhält man ebenso

$$\|\tilde{\mathbf{r}}\|_2 = \|(X + \Delta X)(\mathbf{s} + \Delta \mathbf{s}) - \mu \mathbf{e}\|_2 < \mu,$$

was man abgekürzt und quadriert wie folgt schreiben kann

$$\sum_{i=1}^{n} (a_i - \mu)^2 < \mu^2,$$

mit $a_i = (X + \Delta X)(\mathbf{s} + \Delta \mathbf{s})$. Wir wissen bereits, dass beide Faktoren von a_i nichtnegativ sind. Sei für einen Index, zum Beispiel $i = j$, einer der Faktoren Null. Dann folgt

$$\mu^2 + \sum_{i=1, i \neq j}^{n} (a_i - \mu)^2 < \mu.$$

Diese Aussage kann nicht gelten, da alle Summanden der Summe nichtnegativ sind. Also sind alle Faktoren der a_i, $i = 1, \ldots, n$, positiv. Das heißt, es gelten $\mathbf{x} + \Delta \mathbf{x} > \mathbf{0}$, $\mathbf{s} + \Delta \mathbf{s} > \mathbf{0}$ und der Newton–Schritt liefert einen inneren Punkt. *Übungsaufgabe bis hier.* □

Algorithmus 12.12 Kurz–Schritt–Algorithmus. Es seien $\mathbf{x}^{(0)} > \mathbf{0}$, $\mathbf{y}^{(0)}$, $\mathbf{s}^{(0)} > \mathbf{0}$ und $\mu^{(0)}$ so gegeben, dass

$$\begin{aligned} A\mathbf{x}^{(0)} &= \mathbf{b}, \\ A^T \mathbf{y}^{(0)} + \mathbf{s}^{(0)} &= \mathbf{c}, \\ X^{(0)} \mathbf{s}^{(0)} - \mu^{(0)} \mathbf{e} &= \mathbf{r}^{(0)}, \\ \frac{\left\|\mathbf{r}^{(0)}\right\|_2}{\mu^{(0)}} &\leq \frac{1}{2} \end{aligned} \tag{12.14}$$

gelten. Es sei des weiteren eine gewünschte Genauigkeit $\varepsilon > 0$ vorgegeben. Setze $k = 0$.

1. Führe einen Newton–Schritt aus, das heißt löse (12.8). Setze $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta \mathbf{x}$, $\mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} + \Delta \mathbf{y}$ und $\mathbf{s}^{(k+1)} = \mathbf{s}^{(k)} + \Delta \mathbf{s}$.
2. Verkleinere den Parameter mittels der Vorschrift
$$\mu^{(k+1)} = \mu^{(k)} \left(1 - \frac{1}{6\sqrt{n}}\right).$$
3. Setze $k := k + 1$.
4. Falls $\mu^{(k)} \leq \varepsilon / n$ dann beende das Verfahren, ansonsten gehe zu Schritt 1.

□

Bemerkung 12.13 Die Bedingung (12.14) bedeutet, dass die Anfangsiterierte des Verfahrens sich relativ nahe am zentralen Pfad befinden muss. Die Bezeichnung „Kurz–Schritt–Algorithmus" hat sich eingebürgert, weil der Parameter μ in jedem Schritt nur etwa um den Faktor $1 - 1/\sqrt{n}$ verkleinert wird, der für große n nahe bei Eins liegt. □

Wir wollen nun die oben erarbeitete abstrakte Theorie auf den Kurz–Schritt–Algorithmus anwenden.

Lemma 12.14 *Der Kurz–Schritt–Algorithmus 12.12 erfüllt die Bedingung (12.10) mit $\beta = 1/2$.*

Beweis: Der Beweis erfolgt induktiv. Wegen der Forderung (12.14) an den Startpunkt, ist der Induktionsanfang gegeben. Gelte nun $\left\|\mathbf{r}^{(k)}\right\|_2 \leq \mu^{(k)}/2$ für ein $k \geq 0$. Aus (12.8), (12.9) (erste Gleichung) und der Definition des Parameters im Algorithmus 12.12 folgt

$$\mathbf{r}^{(k+1)} = X^{(k+1)}\mathbf{s}^{(k+1)} - \mu^{(k+1)}\mathbf{e} = \tilde{\mathbf{r}}^{(k)} + \mu^{(k)}\mathbf{e} - \mu^{(k+1)}\mathbf{e} = \tilde{\mathbf{r}}^{(k)} + \frac{\mu^{(k)}}{6\sqrt{n}}\mathbf{e}.$$

Aus (12.13) folgt $\left\|\tilde{\mathbf{r}}^{(k)}\right\|_2 \leq \mu^{(k)}/4$. Mit der Dreiecksungleichung, $\|\mathbf{e}\|_2 = \sqrt{n}$ und dem größtmöglichen Verhältnis von $\mu^{(k)}$ und $\mu^{(k+1)}$ (für $n = 1$) erhält man

$$\begin{aligned}\left\|\mathbf{r}^{(k+1)}\right\|_2 &\leq \left\|\tilde{\mathbf{r}}^{(k)}\right\|_2 + \frac{\mu^{(k)}}{6\sqrt{n}}\|\mathbf{e}\|_2 \leq \mu^{(k)}\left(\frac{1}{4} + \frac{1}{6}\right) = \frac{5}{12}\mu^{(k)} \leq \frac{5}{12}\frac{6}{5}\mu^{(k+1)} \\ &= \frac{1}{2}\mu^{(k+1)}.\end{aligned}$$

Das ist die Aussage des Lemmas. Die Induktionsvoraussetzung ist bei der Nutzung von (12.13) implizit verwendet wurden. ■

In einer Übungsaufgabe wurde gezeigt, dass die Iterierten bei der Ausführung des Newton–Schritts strikt zulässig bleiben ($\mathbf{x}, \mathbf{s} > \mathbf{0}$). Damit ist der Kurz–Schritt–Algorithmus 12.12 wohldefiniert.

Satz 12.15 Polynomiale Komplexität des Kurz–Schritt-Algorithmus. *Der Kurz–Schritt–Algorithmus 12.12 hält nach spätestens*

$$6\sqrt{n}\ln\left(\frac{n\mu^{(0)}}{\varepsilon}\right) \tag{12.15}$$

Iterationen mit Näherungslösungen $\mathbf{x} > \mathbf{0}$, $\mathbf{y}$, $\mathbf{s} > \mathbf{0}$ von (12.1) und (12.2), deren Dualitätslücke

$$\mathbf{c}^T\mathbf{x} - \mathbf{b}^T\mathbf{y} \leq 2\varepsilon$$

erfüllt.

Beweis: Man muss den Abbruchindex und den Abbruchfehler untersuchen. Die Abbruchbedingung in Schritt 4 ist erfüllt, wenn

$$\ln \mu^{(k)} \leq \ln \frac{\varepsilon}{n}.$$

Nach Konstruktion gilt

$$\mu^{(k)} = \left(1 - \frac{1}{6\sqrt{n}}\right)^k \mu^{(0)}.$$

Einsetzen liefert

$$k\ln\left(1 - \frac{1}{6\sqrt{n}}\right) + \ln\left(\mu^{(0)}\right) \leq \ln\frac{\varepsilon}{n}$$

oder

$$-k\ln\left(1 - \frac{1}{6\sqrt{n}}\right) \geq \ln\frac{n\mu^{(0)}}{\varepsilon} \quad \Longleftrightarrow \quad k\ln\left(1 + \frac{1}{6\sqrt{n} - 1}\right) \geq \ln\frac{n\mu^{(0)}}{\varepsilon}.$$

Wegen $\ln x \leq x - 1$ folgt daraus

$$k\left(1 - \frac{1}{6\sqrt{n} - 1} - 1\right) \geq \ln\frac{n\mu^{(0)}}{\varepsilon} \quad \Longleftrightarrow \quad k \geq \left(6\sqrt{n} - 1\right)\ln\frac{n\mu^{(0)}}{\varepsilon}.$$

Daraus folgt die erste Aussage des Satzes.

Die Abschätzung der Dualitätslücke erfolgt unter Nutzung von (12.6), der dritten Gleichung von (12.8), der Cauchy–Schwarz–Ungleichung, (12.10) und der Abbruchbedingung des Verfahrens

$$\begin{aligned} \mathbf{c}^T\mathbf{x} - \mathbf{b}^T\mathbf{y} &= \mathbf{x}^T\mathbf{s} = \mathbf{e}^T(X\mathbf{s}) = \mathbf{e}^T(\mu^{(k)}\mathbf{e} + \mathbf{r}) \leq n\mu^{(k)} + \|\mathbf{e}\|_2 \|\mathbf{r}\|_2 \\ &= n\mu^{(k)} + \sqrt{n}\beta\mu^{(k)} \leq 2n\mu^{(k)} \leq 2\varepsilon. \end{aligned}$$

■

Bemerkung 12.16 Man beachte, dass weder das Abbruchkriterium des Kurz–Schritt–Verfahrens noch die Dualitätslücke etwas darüber aussagen, wie groß der Abstand zwischen der mit dem Kurz–Schritt–Verfahren berechneten numerischen Lösung zur Lösung des linearen Programms (12.1) ist. □

Bemerkung 12.17 Zur Anzahl der Iterationen des Kurz–Schritt–Verfahrens. In vielen Anwendungen lässt sich ein Startpunkt zum Algorithmus 12.12 mit $n\mu^{(0)} \leq 10^{10}$ angeben. Die gewünschte Genauigkeit liegt im allgemeinen bei $\varepsilon = 10^{-8}$ bis $\varepsilon = 10^{-15}$. Für den letzten Wert ist der Faktor $6 \ln\left(\frac{n\mu^{(0)}}{\varepsilon}\right)$ in (12.15) ungefähr 345. Die Anzahl der Iterationen hängt im wesentlichen von $\sqrt{n}$ ab. □

Bemerkung 12.18 Innere–Punkt–Verfahren in der Praxis. In der Praxis nutzt man vor allem unzulässige Innere–Punkt–Verfahren, das heißt, mit Iterierten, die außerhalb des zulässigen Bereichs liegen können. Die analytische Untersuchung dieser Verfahren ist jedoch komplizierter. □

Teil II

Nichtlineare Optimierung

Kapitel 1

Einleitung

Bemerkung 1.1 Problemklasse. In diesem Abschnitt wird die Optimierung von Funktionen

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

betrachtet, wobei $\Omega \subset \mathbb{R}^n$ eine abgeschlossene Menge und $f : \Omega \to \mathbb{R}$ eine gegebene Funktion ist. Die Funktion f heißt Zielfunktion und Ω sei durch endlich viele Nebenbedingungen beschrieben. Bei den betrachteten Problemen können sowohl die Zielfunktion als auch die Nebenbedingungen nichtlinear sein. □

Beispiel 1.2 Ausgleichsrechnung. Ein konkretes Beispiel für eine Aufgabe der Nichtlinearen Optimierung ist die Ausgleichsrechnung. Zur mathematischen Formulierung von Gesetzmäßigkeiten, die ein technisches oder physikalisches Phänomen beschreiben, wird häufig eine Hypothese über den möglichen funktionalen Zusammenhang bekannter und beobachteter Variablen formuliert. Diese enthält im allgemeinen noch freie Parameter und hat etwa die Gestalt

$$z = g(\mathbf{x}, \boldsymbol{\lambda}), \quad \mathbf{x} \in \mathbb{R}^n, \ \boldsymbol{\lambda} \in \mathbb{R}^p.$$

In diesem Modell ist $\boldsymbol{\lambda}$ ein unbekannter Parametervektor mit p Komponenten und z ist eine Variable, deren Abhängigkeit modelliert werden soll.

Nach Auswertung von m Experimenten kennt man m Paare von Variablenwerten $(z_i, \mathbf{x}_i)$, die unter Berücksichtigung möglicher Beobachtungsfehler ε_i die funktionale Abhängigkeit annähernd erfüllen sollen. Das heißt, bei passenden Parametern muss gelten

$$z_i = g(\mathbf{x}_i, \boldsymbol{\lambda}) + \varepsilon_i, \quad i = 1, \ldots, m.$$

Um möglichst kleine Abweichungen ε_i zu erhalten, versucht man, die unbekannten Parameter $\boldsymbol{\lambda}$ optimal anzupassen, indem man entsprechende Optimierungsaufgaben löst, zum Beispiel

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^p} \sum_{i=1}^{m} \left(z_i - g(\mathbf{x}_i, \boldsymbol{\lambda})\right)^2$$

(Methode der kleinsten Quadrate) oder

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^p} \sum_{i=1}^{m} \omega_i \left|z_i - g(\mathbf{x}_i, \boldsymbol{\lambda})\right|^r,$$

wobei $r \geq 1$ eine ganze Zahl ist und $\boldsymbol{\omega} \geq \mathbf{0}$. □

Beispiel 1.3 Standortplanung. Eine Firma möchte n neue Lager der Kapazität a_i, $i = 1, \ldots, n$, errichten. Gegeben sind die Standorte der Abnehmer (α_j, β_j) sowie

der Bedarf b_j, $j = 1, \ldots, m$. In einigen Gebieten darf zudem nicht gebaut werden (ε_k–Umgebungen von (γ_k, δ_k), $k = 1, \ldots, p$).

Gesucht sind die Standorte der Lager (x_i, y_i), die den Lieferplan z optimieren. Mit den Bezeichnungen

- z_{ij} – Transportmenge vom Lager i zum Abnehmer j,
- d_{ij} – Entfernung Lager i und Abnehmer j,
- Δ_{ik} – Entfernung Lager i zum Mittelpunkt (γ_k, δ_k) der Verbotszone k,

lautet die Optimierungsaufgabe wie folgt:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} z_{ij} &\rightarrow \min\ ! \\
\sum_{j=1}^{m} z_{ij} &\leq a_i, \quad i = 1, \ldots, n \\
\sum_{i=1}^{n} z_{ij} &= b_j, \quad j = 1, \ldots, m \\
\Delta_{ik} &\geq \varepsilon_k, \quad i = 1, \ldots, n,\ k = 1, \ldots, p, \\
z_{ij} &\geq 0, \quad i = 1, \ldots, n,\ j = 1, \ldots, m.
\end{aligned}$$

Diese Aufgabe kann für verschiedene Entfernungsmaße gestellt werden, etwa für

$$d_{ij} = |x_i - \alpha_j| + |y_i - \beta_j|$$

oder die Euklidische Entfernung

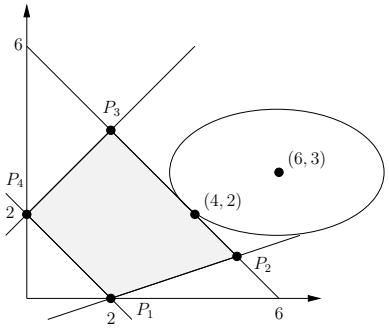
$$d_{ij} = \sqrt{(x_i - \alpha_j)^2 + (y_i - \beta_j)^2}.$$

Wird sowohl für d_{ij} als auch für Δ_{ik} die Euklidische Entfernung gewählt, so erhält man eine Optimierungsaufgabe mit quadratischer Zielfunktion und quadratischen Nebenbedingungen. □

Beispiel 1.4 Quadratische Zielfunktion über konvexem Polyeder. Wir betrachten ein nichtlineares Programm mit quadratischer Zielfunktion und linearen Nebenbedingungen:

$$\begin{aligned}
z = (x_1 - 6)^2 + 2\,(x_2 - 3)^2 &\rightarrow \min\ ! \\
x_1 + x_2 &\geq 2 \\
x_1 - x_2 &\geq -2 \\
x_1 + x_2 &\leq 6 \\
x_1 - 3x_2 &\leq 2 \\
\mathbf{x} &\geq \mathbf{0}.
\end{aligned}$$

Durch die Zielfunktion werden konzentrische Ellipsen mit Mittelpunkt $(6, 3)^T$ beschrieben. Die optimale Lösung ist ein Punkt auf dem Rand $\partial\Omega$ von Ω, aber kein Eckpunkt: $\mathbf{x}_{\text{opt}} = (4, 2)^T, z_{\text{opt}} = 6$.



Betrachtet man eine andere Zielfunktion

$$z = (x_1 - 2)^2 + 2\,(x_2 - 2)^2 \to \min\ !,$$

dann ist das Optimum offensichtlich $\mathbf{x}_{\text{opt}} = (2, 2)^T, z_{\text{opt}} = 0$. Damit liegt das Optimum im Inneren von Ω.

Eine andere quadratische Zielfunktion ist

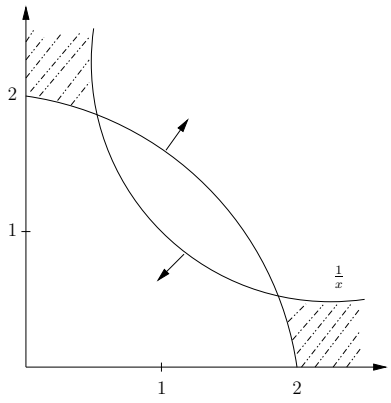
$$z = 2\,(x_1 - 2)^2 + (x_2 - 2)^2 \to \max\ !$$

Bei dieser Zielfunktion gibt es eine Ellipse, die gleichzeitig durch P_1 und P_3 geht. Beide Eckpunkte von Ω sind lokales Maximum mit $z = 4$. Ein weiteres lokales Maximum erhält man in P_4 mit $z = 8$. Das globale Maximum wird jedoch in P_2 mit $z = 19$ angenommen. Ein simplexartiges Vorgehen, wobei man von Eckpunkt zu Eckpunkt geht und in jedem Schritt den Zielfunktionswert verbessert (vergrößert), d.h. $P_1 \to P_4$ oder $P_3 \to P_4$ führt hier nicht zum Ziel. □

Beispiel 1.5 Unzusammenhängender zulässiger Bereich. Der zulässige Bereich Ω sei definiert durch

$$\begin{aligned} x_1^2 + x_2^2 &\geq 4 \\ x_1 x_2 &\leq 1 \\ \mathbf{x} &\geq \mathbf{0}. \end{aligned}$$

Der zulässige Bereich besteht aus zwei getrennt liegenden Teilen. Beide sind nicht konvex. Selbst bei einer linearen Zielfunktion können lokale Minima auftreten, die keine globalen sind.



□

Bemerkung 1.6 In der Vorlesung werden u.a. folgende Problemklassen innerhalb der nichtlinearen Programme nicht betrachtet:

- mehrere Zielfunktionen,
- unendlich viele Nebenbedingungen (semi–infinite Programme),
- stochastische Daten (stochastische Optimierung).

□

Kapitel 2

Nichtlineare Optimierung ohne Nebenbedingungen

2.1 Minimierung nichtglatter Funktionen in einer Variablen

Bemerkung 2.1 Motivation. Die Lösung nichtlinearer Optimierungsaufgaben in einer Dimension tritt als Teilproblem in vielen Verfahren zur Lösung nichtlinearer Optimierungsprobleme in höheren Dimensionen auf.

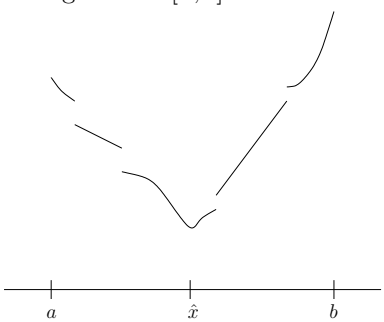
Seien $I \subset \mathbb{R}$ ein gegebenes abgeschlossenes Intervall und $f : I \to \mathbb{R}$ eine gegebene Funktion. Dann sucht man ein $\hat{x} \in I$, welches

$$f(\hat{x}) = \min_{x \in I} f(x) \tag{2.1}$$

erfüllt. Um ein solches Minimum mit Hilfe eines Verfahrens effizient bestimmen zu können, müssen einige einschränkende Bedingungen an $f(x)$ gestellt werden. Je nach Eigenschaften der Funktion, ist dann ein entsprechendes Verfahren zu wählen.

In diesem Abschnitt werden Verfahren zur Bestimmung des Minimums von nichtglatten Funktionen in einer Variablen im Detail vorgestellt, wobei der zulässige Bereich nicht durch Nebenbedingungen eingeschränkt ist. Die Minimierung eindimensionaler glatter Funktionen ohne Nebenbedingungen kann auf die Bestimmung von Nullstellen zurückgeführt werden. Verfahren zur Lösung dieser Aufgabe sind bereits aus der Grundvorlesung *Praktische Mathematik* bekannt. □

Definition 2.2 Unimodale Funktion. Eine Funktion $f : I \to \mathbb{R}$ heißt unimodal (einwellig), falls für ein $\hat{x} \in I = [a, b]$ gilt, dass $f(x)$ streng monoton fallend auf $[a, \hat{x}]$ und streng monoton steigend auf $[\hat{x}, b]$ ist. □



Bemerkung 2.3 Eigenschaften unimodaler Funktionen. Unimodale Funktionen erlauben nach Auswertung der Funktion an zwei Stellen $x < y$ mit $a < x <$

$y < b$ die Reduktion des Intervalls, in der man das Minimum zu suchen hat:

- 1. Fall: $f(a) \leq f(x)$, dann ist das Minimum in $[a, x]$.
- 2. Fall: $f(a) > f(x) > f(y)$. Dann kann das Minimum nicht in $[a, x]$ sein.
- 3. Fall: $f(a) > f(x)$ und $f(x) < f(y)$ und $f(y) < f(b)$. Dann kann das Minimum nicht in $[y, b]$ sein.
- 4. Fall: $f(y) \geq f(b)$. Dann ist das Minimum in $[y, b]$.

Alle anderen Fälle widersprechen der Definition einer unimodalen Funktion. □

Bemerkung 2.4 Der goldene Schnitt. Wir wollen jetzt durch die rekursive Anwendung dieser Beobachtung einen Algorithmus zur Approximation des Punktes konstruieren, an dem das Minimum angenommen wird. Dafür stellen wir zunächst einige Forderungen an den Algorithmus:

1.) Im ersten Schritt sollen zwei Funktionswertauswertungen, in allen weiteren Schritten nur eine weitere Funktionswertauswertung im Restintervall verwendet werden.

2.) Die Punkte x und y sind stets symmetrisch im Restintervall zu wählen.

3.) Der Reduktionsfaktor σ für die Intervalllänge sei konstant, $\sigma \in (0.5, 1)$. Das heißt, der Quotient der Länge des neuen Intervalls und der Länge des vorherigen Intervalls ist $\sigma = const.$

Diese Forderungen besitzen gewisse Ähnlichkeiten zu den Eigenschaften des Bisektions–Verfahrens zur Berechnung von Nullstellen. Auch dort wird die Nullstelle in einem Intervall eingeschachtelt, man hat am Anfang zwei Funktionswerte und später pro Schritt einen zu berechnen und der Reduktionsfaktor ist konstant $\sigma = 0.5$.

Durch die symmetrische Wahl von x, y im Restintervall wird der Reduktionsfaktor unabhängig von der konkreten Funktion $f(x)$. Verlangt man einen konstanten Reduktionsfaktor in allen Iterationen, so ist dieser eindeutig bestimmt.

Um diesen von $f(x)$ unabhängigen Reduktionsfaktor zu bestimmen, genügt es eine streng monoton wachsende Funktion $f : [0, 1] \to \mathbb{R}$ zu betrachten. Damit ist $\hat{x} = 0$. O.B.d.A. sei $0 < x < y < 1$. Wegen der verlangten Symmetrie folgt $y = \sigma, x = 1 - \sigma$.

Im ersten Schritt wird das Intervall $[0, 1]$ auf das Restintervall $[0, y]$ reduziert. In diesem Restintervall wird dann ein Punkt z symmetrisch zu x gewählt. Damit ist gesichert, dass im nächsten Schritt der Punkt x eine der neuen Stützstellen ist, man den Funktionswert $f(x)$ noch einmal verwenden kann und nur den Funktionswert $f(z)$ neu berechnen muss. Je nach Wahl von y gilt $z < x$ oder $z \geq x$. Diese Fälle entsprechen der Wahl von σ aus unterschiedlichen Intervallen.

1. Fall: $\sigma < 2/3$. Dann gilt $x = 1 - \sigma > 1/3$ und damit

$$z = y - x = \sigma - (1 - \sigma) < 1/3 < x.$$

Bei der ersten Gleichung wurde die Symmetrie im Restintervall verwendet. Konstante Reduktion bedeutet nun

$$\frac{x}{y} = \frac{y}{1} \quad \Longrightarrow \quad 0 = y^2 - x = \sigma^2 + \sigma - 1.$$

Die Lösung dieser quadratischen Gleichung ist

$$\sigma = \frac{1}{2}\left(\sqrt{5} - 1\right) \approx 0.618 < \frac{2}{3}.$$

2. Fall: $\sigma \geq 2/3$. Wird analog zum ersten Fall behandelt. Man erhält keine Lösung. *Übungsaufgabe*

Der gefundene Reduktionsfaktor ist also unter den obigen Annahmen der einzig mögliche und er legt den Algorithmus fest. □

Algorithmus 2.5 Goldener Schnitt.

1. *Initialisierung.*

   $\mathtt{i} := 0;\ \mathtt{a}_0 := \mathtt{a};\ \mathtt{b}_0 := \mathtt{b};$
   $\mathtt{x}_i := \mathtt{a}_i + (1-\sigma)(\mathtt{b}_i - \mathtt{a}_i);\ \mathtt{y}_i := \mathtt{a}_i + \sigma(\mathtt{b}_i - \mathtt{a}_i);$
   $\mathtt{fx} := \mathtt{f}(\mathtt{x}_i);\ \mathtt{fy} := \mathtt{f}(\mathtt{y}_i);$

2. *Iteration.*

   Falls $\mathtt{fx} < \mathtt{fy}$, **dann**
   $\mathtt{a}_{i+1} := \mathtt{a}_i;\ \mathtt{b}_{i+1} := \mathtt{y}_i;$
   $\mathtt{x}_{i+1} := \mathtt{a}_{i+1} + (1-\sigma)(\mathtt{b}_{i+1} - \mathtt{a}_{i+1});\ \mathtt{y}_{i+1} := \mathtt{x}_i;$
   $\mathtt{fy} := \mathtt{fx};\ \mathtt{fx} := \mathtt{f}(\mathtt{x}_{i+1});$
   sonst
   $\mathtt{a}_{i+1} := \mathtt{x}_i;\ \mathtt{b}_{i+1} := \mathtt{b}_i;$
   $\mathtt{x}_{i+1} := \mathtt{y}_i;\ \mathtt{y}_{i+1} := \mathtt{b}_{i+1} - (1-\sigma)(\mathtt{b}_{i+1} - \mathtt{a}_{i+1});$
   $\mathtt{fx} := \mathtt{fy};\ \mathtt{fy} := \mathtt{f}(\mathtt{y}_{i+1});$
   $\mathtt{i} := \mathtt{i} + 1;$

3. *Abbruch.*

   Falls $(\mathtt{b}_i - \mathtt{a}_i)/2 > \varepsilon$, **dann**
   gehe zu 2.
   sonst
   $\tilde{\mathtt{x}} := (\mathtt{a}_i + \mathtt{b}_i)/2;$
   stop

Man nimmt $\tilde{x}$ als Approximation an $\hat{x}$. □

Bemerkung 2.6 Zum Algorithmus des Goldenen Schnittes. In der Praxis führt man erst den Vergleich für den Abbruch durch und berechnet dann den neuen Funktionswert. Das spart im letzten Schritt eine Funktionswertberechnung. Diese Einsparung kann sich akkumulieren, falls man das Verfahren im Rahmen eines komplexen Problems oft aufruft.

Algorithmus 2.5 bricht ab, sobald der Funktionswert, für den das Minimum angenommen wird, in einem Intervall der Länge 2ε eingeschachtelt ist. Da $\tilde{x}$ der Mittelpunkt des Intervalls ist, gilt für den Fehlern $|\hat{x} - \tilde{x}| \leq \varepsilon$. Da nach $(n+1)$ Funktionswertauswertungen (n Iterationen) die Länge des Restintervalls $\sigma^n(b-a)$ ist, kann man n in Abhängigkeit von ε a priori abschätzen. □

Bemerkung 2.7 Die Fibonacci–Suche. Wir wollen jetzt auf die Konstanz der Reduktion der Intervalllänge verzichten und untersuchen, ob es Modifikationen von Algorithmus 2.5 gibt, die bei gleicher Anzahl von Funktionswertauswertungen ein kleineres Restintervall liefern. Sei L_n die maximale Länge eines Intervalls, welches man mittels n Funktionswertauswertungen in L_n auf $L_1 = 1$ reduzieren kann.

Beim Bisektions–Verfahren ist beispielsweise $L_5 = 8$. Mit fünf Funktionsauswertungen kann man vier Iterationen durchführen (in der ersten Iteration braucht man zwei Funktionswerte). Halbiert man ein Intervall der Länge 8, so erhält man ein Intervall der Länge 4. Halbiert man dieses, so ist die Intervalllänge 2 und die letzte Halbierung führt auf ein Intervall der Länge 1.

Seien $x < y$ die beiden Stützstellen in L_n. Durch jede dieser Stützstellen wird L_n in zwei Teilintervalle zerlegt. Betrachtet man die Stützstelle x, so enthält das linke Teilintervall höchstens $(n-2)$ Stützstellen des Gesamtprozesses (alle außer x und y) und das rechte Teilintervall höchstens $(n-1)$ der Stützstellen (alle außer x). Entsprechend sind die Längen dieser Teilintervalle höchstens L_{n-2} beziehungsweise L_{n-1}. Damit gilt

$$L_n \leq L_{n-1} + L_{n-2}, \quad n = 2, 3, \ldots. \tag{2.2}$$

Da man mit einer Funktionsauswertungen noch keine Intervallreduktion durchführen

kann, gelten $L_0 = L_1 = 1$. Mit der obigen Abschätzung hat man das größtmögliche Intervall L_n, falls eine Lösung der Ungleichung (2.2) diese sogar als Gleichung erfüllt.

Die Lösung der Differenzengleichung

$$F_0 = F_1 = 1, \quad F_n = F_{n-1} + F_{n-2}$$

ist bekannt. Es sind die sogenannten Fibonacci–Zahlen. Die Darstellung dieser Zahlen in geschlossener Form ist

$$F_i = \frac{1}{\sqrt{5}} \left(\left(\frac{1+\sqrt{5}}{2} \right)^{i+1} - \left(\frac{1-\sqrt{5}}{2} \right)^{i+1} \right).$$

Für den Algorithmus der Fibonacci–Suche setzen wir noch $F_{-2} := 1, F_{-1} := 0$.

Bei der Fibonacci–Suche wird die geforderte Länge des Intervalls, in dem $\hat{x}$ eingeschachtelt werden soll, vorgegeben. Die Iterationspunkte bei einem beliebigen Startintervall $[a, b]$ ergeben sich durch lineare Transformation mit dem Faktor

$$h = \frac{b-a}{F_n}. \tag{2.3}$$

Die Länge des Restintervalls in der Iteration $i = 0, 1, \ldots, n-1$ ist $F_{n-i}h$. Damit ist die Intervallreduktion

$$\frac{F_{n-i-1}h}{F_{n-i}h} = \frac{F_{n-i-1}}{F_{n-i}},$$

also der Quotient zweier aufeinanderfolgender Fibonacci–Zahlen. Außerdem besitzt das $(n-1)$–ste Intervall die Länge $F_1 h = h$ und das vorletzten Restintervall $[a_{n-2}, b_{n-2}]$ die Länge $F_2 h = 2h$. Die Endpunkte des $(n-1)$–sten Intervalls sind ein Endpunkt von $[a_{n-2}, b_{n-2}]$ und eine Stützstelle von $[a_{n-2}, b_{n-2}]$. Eine Stützstelle ist also der Mittelpunkt $a_{n-2} + h$ und wegen der Symmetrie der Stützstellen folgt $x_{n-2} = y_{n-2} = a_{n-2} + h$. Die Stützstellen fallen somit zusammen und die Iteration muss beendet werden. Für den Punkt $\hat{x}$, in dem das Minimum angenommen wird, gilt somit $\hat{x} \in [a_{n-2}, b_{n-2}]$, woraus $|\hat{x} - x_{n-2}| \leq h$ folgt. Damit ist durch (2.3) bei vorgegebenem h auch die Anzahl der Iterationen n gegeben. □

Algorithmus 2.8 Fibonacci–Suche.

1. *Initialisierung.*

   `gebe h vor, bestimme` n`, berechne die Fibonacci-Zahlen` $F_0, \ldots, F_n$

   $\mathtt{i} := 0;\ \mathtt{a_0} := \mathtt{a};\ \mathtt{b_0} := \mathtt{b};\ \mathtt{h} := (\mathtt{b} - \mathtt{a})/\mathtt{F_n};$

   $\mathtt{x_0} := \mathtt{a_0} + \mathtt{F_{n-2}h};\ \mathtt{y_0} := \mathtt{a_0} + \mathtt{F_{n-1}h};$

   $\mathtt{fx} := \mathtt{f(x_0)};\ \mathtt{fy} := \mathtt{f(y_0)};$

2. *Iteration.*

   `Falls` $\mathtt{fx} < \mathtt{fy}$`, dann`

   $\mathtt{a_{i+1}} := \mathtt{a_i};\ \mathtt{b_{i+1}} := \mathtt{y_i};$

   $\mathtt{x_{i+1}} := \mathtt{a_{i+1}} + \mathtt{F_{n-i-3}h};\ \mathtt{y_{i+1}} := \mathtt{x_i};$

   $\mathtt{fy} := \mathtt{fx};\ \mathtt{fx} := \mathtt{f(x_{i+1})};$

   `sonst`

   $\mathtt{a_{i+1}} := \mathtt{x_i};\ \mathtt{b_{i+1}} := \mathtt{b_i};$

   $\mathtt{x_{i+1}} := \mathtt{y_i};\ \mathtt{y_{i+1}} := \mathtt{b_{i+1}} - \mathtt{F_{n-i-3}h};$

   $\mathtt{fx} := \mathtt{fy};\ \mathtt{fy} := \mathtt{f(y_{i+1})};$

   $\mathtt{i} := \mathtt{i} + 1;$

3. *Abbruch.*

```
Falls i < n − 2, dann
      gehe zu 2.
sonst
      x̃ := x_i;
      stop
```

Man nimmt $\tilde{x}$ als Approximation an $\hat{x}$.

Analog wie beim Goldenen Schnitt kann man in der letzten Iteration eine Funktionswertberechnung sparen. □

Beispiel 2.9 Fibonacci–Suche. Wir betrachten die Funktion $f(x) = \sin(x-2)$ auf $[a,b] = [0,2]$. Auf diesem Intervall ist die Funktion $f(x)$ unimodal und sie nimmt ihr Minimum in $\hat{x} = 2 - \pi/2 \approx 0.4292037$ an. Wir wollen die Fibonacci–Suche mit $n = 6$ durchführen. Die Fibonacci–Zahlen $F_0, \ldots, F_6$ sind $1, 1, 2, 3, 5, 8, 13$. Daraus folgt, dass man ein Restintervall der Länge

$$h = \frac{2}{13} \approx 0.1538462$$

findet.

Die im Verfahren auftretenden Stützstellen und Intervallgrenzen sind alle von der Form $a + kh$ mit $k \in \{0, 1, 2, 3, 5, 8, 13\}$. Für eine Stützstelle $t \in [a,b]$ nennt man die entsprechende ganze Zahl $k(t) := (t-a)/h$ den Fibonacci–Index von t. Beim Start gilt $k(a_0) = 0$, $k(x_0) = F_{n-2} = F_4 = 5$, $k(y_0) = F_{n-1}$, $k(b_0) = F_n = F_6 = 13$.

Während der Iteration werden die Funktionswerte $f(x_i)$ und $f(y_i)$ verglichen. Die Variable mit dem kleinerem Funktionswert bleibt Stützstelle, die mit dem größeren Funktionswert wird neue Intervallgrenze. Der Fibonacci–Index der neuen Stützstelle ergibt sich wegen der symmetrischen Lage der Stützstellen im Restintervall aus $k(x_{i+1}) - k(a_{i+1}) = k(b_{i+1}) - k(y_{i+1})$. Ordnet man alle Werte einer Iteration zeilenweise in einer Tabelle an, so verschieben sich diese Werte beim Übergang zur nächsten Iteration nach rechts beziehungsweise nach links ab der Position der neuen Stützstelle.

i	$k(a_i)$	x_i	$k(x_i)$	y_i	$k(y_i)$	$k(b_i)$	$f(x_i)$	$f(y_i)$
0	0	0.7692308	5	1.2307692	8	13	- 0.9427456	- 0.6955828
1	0	0.4615385	3	0.7692308	5	8	- 0.9994773	- 0.9427456
2	0	0.3076923	2	0.4615385	3	5	- 0.9926266	- 0.9994773
3	2	0.4615385	3	0.6153846	4	5	- 0.9994773	- 0.9827183
4	2	0.4615385	3	0.4615385	3	4	- 0.9994773	- 0.9994773

Mit $x_4 = 0 + 3h$ bricht das Verfahren ab und für das Minimum von $f(x)$ gilt

$$\hat{x} \in [0.4615385 - h, 0.4615385 + h].$$

□

Bemerkung 2.10 Vergleich vom Goldenen Schnitt und Fibonacci–Suche, Anzahl der Iterationen. Aus

$$1 - \sigma = \sigma^2 \tag{2.4}$$

folgt induktiv (mit $F_{-2} = 1, F_{-1} = 0$)

$$\sigma^n = (-1)^n \left(F_{n-2} - F_{n-1}\sigma\right).$$

Übungsaufgabe Man rechnet direkt nach, dass

$$F_1 = 1 < \frac{1}{\sigma} < 2 = F_2, \quad F_2 < \frac{1}{\sigma^2} < F_3$$

gelten. Induktiv erhält man für $n > 1$ die Abschätzung

$$\frac{1}{F_n} > \sigma^n > \frac{1}{F_{n+1}}.$$

Außerdem findet man

$$\lim_{n\to\infty} \frac{F_n}{F_{n+1}} = \sigma.$$

Unter der Annahme, dass der rechte Grenzwert existiert, kann man diesen mit Hilfe von (2.4) berechnen. *Übungsaufgabe* Asymptotisch erhält man die gleiche Intervallreduktion und der Goldene Schnitt ist bei gleichem Aufwand demnach von kaum geringerer Genauigkeit. Auf der anderen Seite besitzt die Fibonacci–Suche den Nachteil, dass die gewünschte Genauigkeit a priori festgelegt werden muss, um n zu bestimmen.

Beide Verfahren reduzieren die Intervalllänge linear, dass heißt es gilt

$$\frac{|b_{i+1} - a_{i+1}|}{|b_i - a_i|} \leq \lambda \quad \text{mit } \lambda \in (0,1).$$

Dabei ist der Reduktionsfaktor λ beim Goldenen Schnitt durch σ gegeben und bei der Fibonacci–Suche durch eine Schranke für F_{n-1}/F_n, $n \geq 2$. Soll das n–te Intervall kleiner als ein gegebenes ε sein, so erhält man aus

$$\varepsilon \leq \lambda^n(b-a) = \lambda^n(b_0 - a_0)$$

die Abschätzung

$$n \geq \log\left(\frac{\varepsilon}{b-a}\right) / \log(\lambda).$$

□

2.2 Differenzierbare Funktionen in mehreren Dimensionen

Bemerkung 2.11 Notwendige Bedingung für ein Minimum, Berechnungsverfahren. Sei $f : \mathbb{R}^n \to \mathbb{R}$ mit $f \in C^1(\mathbb{R}^n)$ gegeben. Die notwendige Bedingung für ein lokales Minimum im Punkt $\hat{\mathbf{x}} \in \mathbb{R}^n$ ist

$$\nabla f(\hat{\mathbf{x}}) = \mathbf{0}.$$

Zur Berechnung von möglichen Werten für $\hat{\mathbf{x}}$ hat man damit ein nichtlineares Gleichungssystem zu lösen. Verfahren zur Lösung nichtlinearer Gleichungssysteme wurden bereits in der Vorlesung *Praktische Mathematik* behandelt:

- Bisektion falls $n = 1$,
- Abstiegsverfahren (Gradientenverfahren, Verfahren des steilsten Abstiegs),
- Fixpunktiteration, wobei die nachfolgenden Verfahren Spezialfälle sind,
- Newton–Verfahren,
- vereinfachtes und Quasi–Newton–Verfahren.

Auf Abstiegsverfahren soll hier etwas näher eingegangen werden. □

2.2.1 Abstiegsverfahren

Im folgenden wird vorausgesetzt, dass $f(\mathbf{x})$ zweimal stetig differenzierbar ist.

Definition 2.12 Abstiegsverfahren, Abstiegsrichtung, Schrittweite. Abstiegsverfahren zur Berechnung von $\mathbf{f}(\mathbf{x}) = 0$ besitzen folgende Gestalt:

- $\mathbf{x}^{(0)} \in \mathbb{R}^n$ sei gegeben,
- berechne für $k \geq 0$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}, \tag{2.5}$$

wobei $\mathbf{d}^{(k)}$ eine geeignete Suchrichtung, die sogenannte Abstiegsrichtung, ist und $\alpha_k > 0$ die Schrittweite.

Die Richtung $\mathbf{d}^{(k)}$ wird Abstiegsrichtung genannt, falls

$$\begin{array}{rcll} {\mathbf{d}^{(k)}}^T \nabla f\left(\mathbf{x}^{(k)}\right) & < & 0 & \text{falls } \nabla f\left(\mathbf{x}^{(k)}\right) \neq \mathbf{0}, \\ \mathbf{d}^{(k)} & = & \mathbf{0} & \text{falls } \nabla f\left(\mathbf{x}^{(k)}\right) = \mathbf{0}. \end{array} \tag{2.6}$$

□

Bemerkung 2.13 Abstiegseigenschaft. Unter der Bedingung (2.6) existieren hinreichend kleine $\alpha_k > 0$, so dass

$$f\left(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}\right) < f\left(\mathbf{x}^{(k)}\right).$$

Aus der Taylor–Entwicklung folgt nämlich für $\boldsymbol{\xi} = \mathbf{x}^{(k)} + \theta\alpha_k \mathbf{d}^{(k)}$ mit $\theta \in (0,1)$

$$f\left(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}\right) = f\left(\mathbf{x}^{(k)}\right) + \alpha_k {\mathbf{d}^{(k)}}^T \nabla f(\boldsymbol{\xi}).$$

Falls α_k hinreichend klein ist, dann ist $\boldsymbol{\xi}$ hinreichend nahe an $\mathbf{x}^{(k)}$ und dann ist das Vorzeichen von ${\mathbf{d}^{(k)}}^T \nabla f(\boldsymbol{\xi})$ das gleiche wie von ${\mathbf{d}^{(k)}}^T \nabla f\left(\mathbf{x}^{(k)}\right)$, woraus die Behauptung folgt. □

Beispiel 2.14 Abstiegsrichtungen. Einige Beispiele für Abstiegsrichtungen sind:

1. Newton–Verfahren

$$\mathbf{d}^{(k)} = -(H_f)^{-1}\left(\mathbf{x}^{(k)}\right) \nabla f\left(\mathbf{x}^{(k)}\right),$$

wobei $H_f\left(\mathbf{x}^{(k)}\right)$ die Hesse–Matrix von $f(\mathbf{x})$ an der Stelle $\mathbf{x}^{(k)}$ ist.

2. Quasi–Newton–Verfahren, inexaktes Newton–Verfahren, vereinfachtes Newton–Verfahren

$$\mathbf{d}^{(k)} = -B_k^{-1} \nabla f\left(\mathbf{x}^{(k)}\right),$$

wobei B_k eine geeignete Approximation von $H_f\left(\mathbf{x}^{(k)}\right)$ ist.

3. Gradienten–Verfahren, Verfahren des steilsten Abstieges

$$\mathbf{d}^{(k)} = -\nabla f\left(\mathbf{x}^{(k)}\right),$$

also das Quasi–Newton–Verfahren mit $B_k = I$. Es gilt

$${\mathbf{d}^{(k)}}^T \nabla f\left(\mathbf{x}^{(k)}\right) = -\left\|\nabla f\left(\mathbf{x}^{(k)}\right)\right\|_2^2 < 0 \quad \text{für } \nabla f\left(\mathbf{x}^{(k)}\right) \neq \mathbf{0}.$$

4. Verfahren der konjugierten Gradienten

$$\mathbf{d}^{(k)} = -\nabla f\left(\mathbf{x}^{(k)}\right) + \beta_k \mathbf{d}^{(k-1)},$$

wobei die β_k Parameter sind, die so gewählt werden, dass die Suchrichtungen $\{\mathbf{d}^{(k)}\}$ bezüglich eines geeigneten Skalarproduktes orthogonal sind, siehe Bemerkung 2.26.

□

Bemerkung 2.15 Liniensuche. Nachdem man eine Abstiegsrichtung gewählt hat, bleibt das Problem, wie man α_k wählen soll.

- Eine Möglichkeit besteht in der Lösung des nichtlinearen Minimierungsproblems in einer Dimension: finde α_k, so dass

$$\phi(\alpha_k) = f\left(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}\right)$$

  minimiert wird. *Übungsaufgabe Thm 7.3* Das geht nur in Spezialfällen mit vertretbarem Aufwand. Diese Spezialfälle sind aber durchaus wichtig, siehe Abschnitt 2.2.2.
- Man nutzt ein iteratives Verfahren, welches einen geeigneten Wert α_k bestimmt. Diese Herangehensweise wird Liniensuche genannt.

□

Beispiel 2.16 Armijo–Liniensuche. Seien $f(\mathbf{x})$ eine zu minimierende Funktion, $\mathbf{x}^{(k)}$ die gegenwärtige Iterierte, $\mathbf{d}^{(k)}$ eine irgendwie berechnete Abstiegsrichtung, $\mu \in (0,1)$ und $\gamma > 0$ eine Konstante. Man wählt $\alpha^{(0)} \geq \gamma \left\|\nabla f(\mathbf{x}^{(k)})\right\|_2$ und bestimme unter den Zahlen $\alpha^{(j)} = \beta^j \alpha^{(0)}$, zum Beispiel $\beta = 1/2$, die erste, für die

$$f\left(\mathbf{x}^{(k)} + \alpha^{(j)}\mathbf{d}^{(k)}\right) \leq f\left(\mathbf{x}^{(k)}\right) + \mu\alpha^{(j)}\nabla f\left(\mathbf{x}^{(k)}\right)^T \mathbf{d}^{(k)}$$

gilt. Die Motivation für diese Herangehensweise kommt von der nach dem linearen Glied abgebrochenen Taylor–Entwicklung

$$f\left(\mathbf{x}^{(k)} + \alpha^{(j)}\mathbf{d}^{(k)}\right) \approx f\left(\mathbf{x}^{(k)}\right) + \alpha^{(j)}\nabla f\left(\mathbf{x}^{(k)}\right)^T \mathbf{d}^{(k)}.$$

Ist $\mathbf{d}^{(k)}$ eine Abstiegsrichtung, dann findet man in dieser Richtung einen Funktionswert von $f(\mathbf{x})$, der kleiner als der Funktionswert $f\left(\mathbf{x}^{(k)}\right)$ ist, falls $\alpha^{(j)}$ nur hinreichend klein ist, siehe Bemerkung 2.13. Man fängt mit irgendeinem hinreichend großen $\alpha^{(0)}$ an. Dieses hängt von $\left\|\nabla f\left(\mathbf{x}^{(k)}\right)\right\|_2$ ab. Ist $\left\|\nabla f\left(\mathbf{x}^{(k)}\right)\right\|_2$ klein, kann man vermuten in der Nähe eines Minimums zu sein (besitzt $f(\mathbf{x})$ in $\mathbf{x}^{(k)}$ ein Minimum, so ist $\nabla f\left(\mathbf{x}^{(k)}\right) = \mathbf{0}$) und man braucht vielleicht nur einen kleinen Schritt. Man testet ob der Funktionswert sich verkleinert. Ist das nicht der Fall, wird der Parameter $\alpha^{(j)}$ sukzessive verkleinert, bis er klein genug ist. □

2.2.2 Abstiegsmethoden für quadratische Funktionen

Bemerkung 2.17 Quadratische Funktionen

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}, \tag{2.7}$$

mit einer symmetrischen und positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ und $\mathbf{b} \in \mathbb{R}^n$ sind ein wichtiger Spezialfall. Die notwendige Bedingung für ein Minimum in $\hat{\mathbf{x}} \in \mathbb{R}^n$ ist

$$\mathbf{0} = \nabla f(\hat{\mathbf{x}}) = A\hat{\mathbf{x}} - \mathbf{b}.$$

Man hat also das lineare Gleichungssystem $A\hat{\mathbf{x}} = \mathbf{b}$ zu lösen. Da $H_f(\hat{\mathbf{x}}) = A$ positiv definit ist, ist auch eine hinreichende Bedingung für ein Minimum erfüllt. □

Lemma 2.18 Optimale Schrittweite. *Das eindimensionale Optimierungsproblem finde α_k, so dass*

$$f\left(\mathbf{x}^{(k)} + \alpha_k\mathbf{d}^{(k)}\right)$$

minimiert wird, besitzt für die quadratische Funktion (2.7) die Lösung

$$\alpha_k = \frac{\mathbf{d}^{(k)^T}\mathbf{r}^{(k)}}{\mathbf{d}^{(k)^T}A\mathbf{d}^{(k)}}, \quad \mathbf{r}^{(k)} := \mathbf{b} - A\mathbf{x}^{(k)}, \tag{2.8}$$

$r^{(k)}$ *– Residuum.*

Beweis: Es gilt

$$\begin{aligned}
&\frac{d}{d\alpha_k} f\left(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}\right) \\
&= \frac{d}{d\alpha_k}\left[\frac{1}{2}\left(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}\right)^T A\left(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}\right) - \mathbf{b}^T\left(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}\right)\right] \\
&= \frac{d}{d\alpha_k}\left[\frac{\alpha_k^2}{2}\mathbf{d}^{(k)^T}A\mathbf{d}^{(k)} + \alpha_k\left(\frac{1}{2}\mathbf{d}^{(k)^T}A\mathbf{x}^{(k)} + \frac{1}{2}\mathbf{x}^{(k)^T}A\mathbf{d}^{(k)} - \mathbf{b}^T\mathbf{d}^{(k)}\right) + \ldots\right] \\
&= \alpha_k \mathbf{d}^{(k)^T}A\mathbf{d}^{(k)} + \left(\frac{1}{2}\mathbf{d}^{(k)^T}A\mathbf{x}^{(k)} + \frac{1}{2}\mathbf{x}^{(k)^T}A\mathbf{d}^{(k)} - \mathbf{b}^T\mathbf{d}^{(k)}\right) \\
&= \alpha_k \mathbf{d}^{(k)^T}A\mathbf{d}^{(k)} + \left(\mathbf{x}^{(k)^T}A\mathbf{d}^{(k)} - \mathbf{b}^T\mathbf{d}^{(k)}\right) = \alpha_k \mathbf{d}^{(k)^T}A\mathbf{d}^{(k)} - \mathbf{r}^{(k)^T}\mathbf{d}^{(k)}.
\end{aligned}$$

Dabei wurde die Symmetrie von A genutzt

$$\mathbf{d}^{(k)^T}A\mathbf{x}^{(k)} = \mathbf{x}^{(k)^T}A^T\mathbf{d}^{(k)} = \mathbf{x}^{(k)^T}A\mathbf{d}^{(k)}.$$

Die Ableitung muss im Optimum verschwinden, woraus durch Umstellen die Formel für α_k folgt. Die zweite Ableitung bezüglich α_k ist $\mathbf{d}^{(k)^T}A\mathbf{d}^{(k)}$. Wegen der positiven Definitheit von A ist dieser Ausdruck positiv für alle $\mathbf{d}^{(k)} \neq \mathbf{0}$, woraus folgt, dass man mit α_k ein Minimum bestimmt hat. ■

Bemerkung 2.19 Von A induziertes Skalarprodukt und Norm. Die positiv definite symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ induziert ein Skalarprodukt auf $\mathbb{R}^n$

$$(\mathbf{x}, \mathbf{y})_A := \mathbf{x}^T A\mathbf{y}$$

und damit eine Norm

$$\|\mathbf{x}\|_A := (\mathbf{x}, \mathbf{x})_A^{1/2}.$$

In dieser Norm betrachtet man die Konvergenz eines Abstiegsverfahrens. □

Lemma 2.20 Fehlerabschätzung eines Abstiegsverfahrens. *Für den Fehler eines Abstiegsverfahrens zur Minimierung der quadratischen Funktion (2.7) gilt*

$$\left\|\mathbf{x}^{(k+1)} - \hat{\mathbf{x}}\right\|_A = \rho_k \left\|\mathbf{x}^{(k)} - \hat{\mathbf{x}}\right\|_A$$

mit

$$\rho_k = \left(1 - \frac{\left(\mathbf{d}^{(k)^T}\mathbf{r}^{(k)}\right)^2}{\mathbf{d}^{(k)^T}A\mathbf{d}^{(k)}\ \mathbf{r}^{(k)^T}A^{-1}\mathbf{r}^{(k)}}\right)^{1/2}. \tag{2.9}$$

Beweis: Der Beweis geht wie folgt:

1. Schreibe $\left\|\mathbf{x}^{(k+1)} - \hat{\mathbf{x}}\right\|_A^2$ Term für Term.
2. Setze (2.8) ein.
3. Zeige $\left\|\mathbf{x}^{(k)} - \hat{\mathbf{x}}\right\|_A^2 = \mathbf{r}^{(k)^T}A^{-1}\mathbf{r}^{(k)}$ und setze ein.

Details sind *Übungsaufgabe.* ■

Satz 2.21 Lineare Konvergenz des Gradientenverfahrens. *Das Gradientenverfahren zur Lösung von (2.7) konvergiert linear mit*

$$\left\|\mathbf{x}^{(k+1)} - \hat{\mathbf{x}}\right\|_A \leq \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} \left\|\mathbf{x}^{(k)} - \hat{\mathbf{x}}\right\|_A .$$

Beweis: Für das Gradientenverfahren gilt $\mathbf{d}^{(k)} = \mathbf{r}^{(k)}$. Somit kann der Nenner in (2.9) außerhalb des Minimums nicht Null werden, da er die Gestalt

$$\mathbf{d}^{(k)^T} A \mathbf{d}^{(k)} \; \mathbf{r}^{(k)^T} A^{-1} \mathbf{r}^{(k)} = \left(\mathbf{d}^{(k)^T} A \mathbf{d}^{(k)}\right)^2 > 0$$

besitzt.

Zur Abschätzung des Konvergenzfaktors nutzt man die Ungleichung

$$\frac{(\mathbf{y}^T\mathbf{y})^2}{(\mathbf{y}^T A\mathbf{y})\,(\mathbf{y}^T A^{-1}\mathbf{y})} \geq \frac{4\lambda_{\max}(A)\lambda_{\min}(A)}{(\lambda_{\max}(A) + \lambda_{\min}(A))^2},$$

siehe Literatur. Einsetzen dieser Abschätzung und Nutzung von

$$\left(1 - \frac{4\lambda_{\max}(A)\lambda_{\min}(A)}{(\lambda_{\max}(A) + \lambda_{\min}(A))^2}\right)^{1/2} = \left(\frac{(\lambda_{\max}(A) + \lambda_{\min}(A))^2 - 4\lambda_{\max}(A)\lambda_{\min}(A)}{(\lambda_{\max}(A) + \lambda_{\min}(A))^2}\right)^{1/2},$$

sowie binomischer Formel beendet die Abschätzung. ■

Bemerkung 2.22 Zum Gradientenverfahren. Falls A schlecht konditioniert ist, das heißt

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

sehr groß ist, dann ist die Konvergenzrate

$$\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} = \frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}$$

nahe bei Eins und das Verfahren ist sehr langsam. Deshalb ist es nötig, bessere Verfahren zu entwickeln. □

Definition 2.23 A**–konjugiert.** Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Die Vektoren $\mathbf{y}_1, \ldots, \mathbf{y}_m \in \mathbb{R}^n$ heißen A–konjugiert (oder A–orthogonal), falls $\mathbf{y}_i \neq \mathbf{0}$ für $i = 1, \ldots, m$, und $(\mathbf{y}_i, \mathbf{y}_j)_A = 0$ für $i \neq j$ gelten. □

Lemma 2.24 *Seien $\mathbf{y}_1, \ldots, \mathbf{y}_m \in \mathbb{R}^n$ A–konjugierte Vektoren. Dann sind $\mathbf{y}_1, \ldots, \mathbf{y}_m$ linear unabhängig.*

Beweis: Indirekt. Sei

$$\mathbf{0} = \sum_{i=0}^{m} \alpha_i \mathbf{y}_i.$$

Durch Multiplikation von links mit $\mathbf{y}_k^T A$, $k = 1, \ldots, m$, folgt

$$0 = \mathbf{y}_k^T A \sum_{i=0}^{m} \alpha_i \mathbf{y}_i = \alpha_k \underbrace{\mathbf{y}_k^T A \mathbf{y}_k}_{>0}.$$

Somit muss $\alpha_k = 0$, $k = 1, \ldots, m$, gelten. ■

Satz 2.25 Endlichkeit des Abstiegsverfahrens für A**–konjugierter Abstiegsrichtungen.** *Verwendet man zur Minimierung von (2.7) A–konjugierte Abstiegsrichtungen $\{\mathbf{d}^{(k)}\}$ und wählt man α_k gemäß (2.8), dann gilt für beliebigen Startwert $\mathbf{x}^{(0)} \in \mathbb{R}^n$*

$$f\left(\mathbf{x}^{(n)}\right) = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}).$$

Beweis: Betrachte zunächst einen beliebigen Vektor $\mathbf{y} \in \mathbb{R}^n$. Da die Vektoren $\{\mathbf{d}^{(i)}\}_{i=0}^{n-1}$ linear unabhängig sind, bilden sie eine Basis des $\mathbb{R}^n$. Deshalb lässt sich $\mathbf{y}$ eindeutig in der Form

$$\mathbf{y} = \sum_{i=0}^{n-1} \beta_i \mathbf{d}^{(i)}$$

darstellen. Es folgt für jedes $k = 0, \ldots, n-1$,

$$\mathbf{d}^{(k)^T} A\mathbf{y} = \sum_{i=0}^{n-1} \beta_i \mathbf{d}^{(k)^T} A\mathbf{d}^{(i)} = \beta_k \mathbf{d}^{(k)^T} A\mathbf{d}^{(k)},$$

woraus sich die Darstellung

$$\mathbf{y} = \sum_{i=0}^{n-1} \frac{\mathbf{d}^{(i)^T} A\mathbf{y}}{\mathbf{d}^{(i)^T} A\mathbf{d}^{(i)}} \mathbf{d}^{(i)} \tag{2.10}$$

ergibt.

Nun ist

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)} = \mathbf{x}^{(k-1)} + \alpha_{k-1}\mathbf{d}^{(k-1)} + \alpha_k \mathbf{d}^{(k)} = \mathbf{x}^{(0)} + \sum_{i=0}^{k} \alpha_i \mathbf{d}^{(i)}.$$

Damit folgt für die Parameter α_k

$$\begin{aligned} \alpha_k &= \frac{\mathbf{d}^{(k)^T}\mathbf{r}^{(k)}}{\mathbf{d}^{(k)^T} A\mathbf{d}^{(k)}} = \frac{\mathbf{d}^{(k)^T}\left(\mathbf{b} - A\mathbf{x}^{(k)}\right)}{\mathbf{d}^{(k)^T} A\mathbf{d}^{(k)}} = \frac{\mathbf{d}^{(k)^T}\left(\mathbf{b} - A\left(\mathbf{x}^{(0)} + \sum_{i=0}^{k-1}\alpha_i \mathbf{d}^{(i)}\right)\right)}{\mathbf{d}^{(k)^T} A\mathbf{d}^{(k)}} \\ &= \frac{\mathbf{d}^{(k)^T}\left(\mathbf{b} - A\mathbf{x}^{(0)}\right) - \sum_{i=0}^{k-1}\alpha_i \mathbf{d}^{(k)^T} A\mathbf{d}^{(i)}}{\mathbf{d}^{(k)^T} A\mathbf{d}^{(k)}} = \frac{\mathbf{d}^{(k)^T}\left(\mathbf{b} - A\mathbf{x}^{(0)}\right)}{\mathbf{d}^{(k)^T} A\mathbf{d}^{(k)}}. \end{aligned}$$

Also gilt

$$\mathbf{x}^{(n)} = \mathbf{x}^{(0)} + \sum_{i=0}^{n-1} \frac{\mathbf{d}^{(i)^T}\left(\mathbf{b} - A\mathbf{x}^{(0)}\right)}{\mathbf{d}^{(i)^T} A\mathbf{d}^{(i)}} \mathbf{d}^{(i)} = \mathbf{x}^{(0)} + \sum_{i=0}^{n-1} \frac{\mathbf{d}^{(i)^T} A\left(A^{-1}\mathbf{b} - \mathbf{x}^{(0)}\right)}{\mathbf{d}^{(i)^T} A\mathbf{d}^{(i)}} \mathbf{d}^{(i)}.$$

Mit Darstellung (2.10) folgt

$$\mathbf{x}^{(n)} = \mathbf{x}^{(0)} + \left(A^{-1}\mathbf{b} - \mathbf{x}^{(0)}\right) = A^{-1}\mathbf{b}.$$

Das ist die Lösung von $A\mathbf{x} = \mathbf{b}$, die gesucht ist. ■

Bemerkung 2.26

- Die Methode endet nach spätestens n Abstiegen mit dem Optimum. Für große n ist diese Eigenschaft allerdings praktisch nicht bedeutsam.
- Eine Realisierung der Methode ist das Verfahren der konjugierten Gradienten

  $$\begin{aligned} \mathbf{d}^{(k+1)} &= \mathbf{r}^{(k)} + \beta_k \mathbf{d}^{(k)}, \\ \beta_k &= -\frac{\mathbf{r}^{(k+1)^T} A\mathbf{d}^{(k)}}{\mathbf{d}^{(k)^T} A\mathbf{d}^{(k)}} = \frac{\mathbf{r}^{(k+1)^T}\mathbf{r}^{(k+1)}}{\mathbf{r}^{(k)^T}\mathbf{r}^{(k)}}. \end{aligned}$$

  Man muss nicht alle A–konjugierten Abstiegsrichtungen speichern, sondern nur die drei Vektoren $\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)}, \mathbf{d}^{(k)}$. *Übungsaufgaben*
- Für das Verfahren der konjugierten Gradienten kann man die Fehlerabschätzung

  $$\left\|\mathbf{x}^{(k)} - \hat{\mathbf{x}}\right\|_A \leq 2\left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}\right)^k \left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_A$$

  beweisen, wobei $\kappa_2(A)$ die Spektralkonditionszahl von A ist.
- Das Verfahren der konjugierten Gradienten ist eines der beliebtesten Verfahren zur Lösung von linearen Gleichungssystemen mit symmetrischer und positiv definiter Matrix.

□

Kapitel 3

Konvexität

3.1 Konvexe Mengen

Der Begriff der konvexen Menge ist bereits aus Definition 1.4, Teil I, bekannt.

Definition 3.1 Konvexer Kegel. Eine Menge $\Omega \subset \mathbb{R}^n$ heißt konvexer Kegel, wenn mit $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$ auch $\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 \in \Omega$ für alle $\lambda_1, \lambda_2 \geq 0$. □

Beispiel 3.2 Konvexe Kegel.

- Für $\lambda_1 = \lambda_2 = 0$ folgt, dass insbesondere $\mathbf{0}$ in einem konvexen Kegel enthalten sein muss.
- Nicht jede konvexe Menge ist ein konvexer Kegel, zum Beispiel sind Kreise keine konvexen Kegel.
- Der Winkelraum mit Scheitel $\mathbf{0}$ ist ein konvexer Kegel. Der Winkelraum mit Scheitel $\neq \mathbf{0}$ ist kein konvexer Kegel.
- Sei $\mathbf{a} \in \mathbb{R}^n, \mathbf{a} \neq \mathbf{0}$, $\alpha \in \mathbb{R}$ und definiere die Hyperebene

  $$H(\mathbf{a}, \alpha) = \left\{ \mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{a}^T \mathbf{x} = \alpha \right\}.$$

  Diese Hyperebene ist eine konvexe Menge. *Übungsaufgabe* Ist $\alpha \neq 0$, dann ist $\mathbf{0} \notin H(\mathbf{a}, \alpha)$ und $H(\mathbf{a}, \alpha)$ ist kein konvexer Kegel. Im Falle $\alpha = 0$ ist $H(\mathbf{a}, \alpha)$ ein konvexer Kegel $H(\mathbf{a}, \alpha) = \mathbf{a}^{\perp}$. *Übungsaufgabe*
- Analog gilt für Halbräume

  $$H_+(\mathbf{a}, \alpha) = \left\{ \mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{a}^T \mathbf{x} \geq \alpha \right\}, \ H_-(\mathbf{a}, \alpha) = \left\{ \mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{a}^T \mathbf{x} \leq \alpha \right\},$$

  dass diese für $\alpha = 0$ konvexe Kegel sind und sonst nicht.
- Eine Menge, die durch Hyperebenen berandet ist, die allesamt den Punkt $\mathbf{0}$ enthalten, wird auch polyhedrischer Kegel genannt. Polyhedrische Kegel sind konvexe Kegel, insbesondere

  $$\mathbb{R}^n_+ = \left\{ \mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{x} \geq \mathbf{0} \right\}.$$

□

Definition 3.3 Trennung von Mengen. Die Hyperebene $H(\mathbf{a}, \alpha)$ trennt die nichtleeren Mengen $\Omega_1, \Omega_2 \subset \mathbb{R}^n$, wenn gilt

$$\mathbf{a}^T \mathbf{x} \leq \alpha \ \forall \mathbf{x} \in \Omega_1 \quad \text{und} \quad \mathbf{a}^T \mathbf{x} \geq \alpha \ \forall \mathbf{x} \in \Omega_2.$$

Falls in beiden Relationen die echten Ungleichungen stehen, dann wird die Trennung streng genannt. □

Lemma 3.4 *Sei $\Omega \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex und gelte $\mathbf{0} \notin \Omega$. Dann existiert eine Hyperebene $H(\mathbf{a}, \alpha)$ mit $\alpha > 0$, so dass $\mathbf{a}^T\mathbf{x} > \alpha$ für alle $\mathbf{x} \in \Omega$.*

Beweis: Wir werden eine konkrete Hyperebene angeben.

Seien $\mathbf{x}_1 \in \Omega$ und $B_1 = \{\mathbf{x} \in \mathbb{R}^n \ : \ \|\mathbf{x}\|_2 \leq \|\mathbf{x}_1\|_2\}$ eine Kugel mit Radius $\|\mathbf{x}_1\|_2$ und Mittelpunkt $\mathbf{0}$. Dann ist $\Omega \cap B_1 \neq \emptyset$ und der Durchschnitt ist kompakt (abgeschlossen und beschränkt). Damit nimmt die stetige Funktion $d(\mathbf{x}) = \|\mathbf{x}\|_2$ ihr Minimum über $\Omega \cap B_1$ in einem Punkt $\mathbf{x}_0 \in \Omega \cap B_1$ an (Satz von Bolzano–Weierstrass), dass heißt es gilt

$$\|\mathbf{x}\|_2 \geq \|\mathbf{x}_0\|_2 \quad \forall\ \mathbf{x} \in \Omega \cap B_1. \tag{3.1}$$

Wegen $\mathbf{0} \notin \Omega$ ist $\|\mathbf{x}_0\|_2 > 0$. Da die Elemente von Ω, die nicht in $\Omega \cap B_1$ liegen, eine größere Norm als $\|\mathbf{x}_1\|_2 \geq \|\mathbf{x}_0\|_2$ besitzen, gilt (3.1) für alle $\mathbf{x} \in \Omega$.

Wegen der Konvexität von Ω ist $\mathbf{x}_0 + \lambda\,(\mathbf{x} - \mathbf{x}_0) \in \Omega$ für alle $\mathbf{x} \in \Omega$ und $\lambda \in (0, 1]$ und es folgt

$$\|(1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}\|_2 = \|\mathbf{x}_0 + \lambda\,(\mathbf{x} - \mathbf{x}_0)\|_2 \geq \|\mathbf{x}_0\|_2 > 0.$$

Quadratur dieser Ungleichung und Ausschreiben der Skalarprodukte ergibt

$$2\lambda\,(\mathbf{x} - \mathbf{x}_0)^T\,\mathbf{x}_0 + \lambda^2\,(\mathbf{x} - \mathbf{x}_0)^T\,(\mathbf{x} - \mathbf{x}_0) \geq 0.$$

Division durch λ und dann $\lambda \to 0$ ergibt

$$\mathbf{x}^T\mathbf{x}_0 \geq \mathbf{x}_0^T\mathbf{x}_0 = \|\mathbf{x}_0\|_2^2 > \frac{\|\mathbf{x}_0\|_2^2}{2} > 0 \quad \forall\ \mathbf{x} \in \Omega.$$

Die Hyperebene mit $\mathbf{a} = \mathbf{x}_0$ und $\alpha = \|\mathbf{x}_0\|_2^2/2$ erfüllt nun die Bedingungen des Lemmas. ■

Eine Folgerung ist der nachfolgende Satz.

Satz 3.5 Erster Trennungssatz. *Seien $\Omega_1 \neq \emptyset$ abgeschlossen und konvex und Ω_2 kompakt und konvex, $\Omega_1, \Omega_2 \subset \mathbb{R}^n$, mit $\Omega_1 \cap \Omega_2 = \emptyset$. Dann existiert eine Hyperebene $H(\mathbf{a}, \alpha)$, die Ω_1 und Ω_2 streng trennt.*

Beweis: Seien

$$\begin{aligned} -\Omega_1 &= \{\mathbf{x} \ : \mathbf{x} = -\mathbf{y}, \mathbf{y} \in \Omega_1\}, \\ \Omega_2 - \Omega_1 &= \{\mathbf{x} \ : \ \exists \mathbf{x}_1 \in \Omega_1, \mathbf{x}_2 \in \Omega_2 \text{ mit } \mathbf{x} = \mathbf{x}_2 - \mathbf{x}_1\}. \end{aligned}$$

Diese Mengen sind konvex und abgeschlossen. Da Ω_1 und Ω_2 einen leeren Durchschnitt haben, gilt $\mathbf{0} \notin \Omega_2 - \Omega_1$. Nach Lemma 3.4 gibt es eine Hyperebene $H(\mathbf{a}, \alpha)$ mit $\alpha > 0$ und $\mathbf{a}^T\,(\mathbf{x}_2 - \mathbf{x}_1) > \alpha > 0$ für alle $\mathbf{x}_1 \in \Omega_1, \mathbf{x}_2 \in \Omega_2$. Durch Umstellen dieser Beziehung erhält man

$$\inf_{\mathbf{x}_2 \in \Omega_2} \mathbf{a}^T\mathbf{x}_2 \geq \sup_{\mathbf{x}_1 \in \Omega_1} \mathbf{a}^T\mathbf{x}_1 + \alpha > \sup_{\mathbf{x}_1 \in \Omega_1} \mathbf{a}^T\mathbf{x}_1.$$

Jede Hyperebene $H(\mathbf{a}, \beta)$ mit

$$\beta \in \left(\sup_{\mathbf{x}_1 \in \Omega_1} \mathbf{a}^T\mathbf{x}_1, \inf_{\mathbf{x}_2 \in \Omega_2} \mathbf{a}^T\mathbf{x}_2 \right)$$

trennt damit Ω_1 und Ω_2 streng. ■

Übungsaufgabe: Notwendigkeit der Kompaktheitsvoraussetzung für strenge Trennung an Beispiel zeigen.

Satz 3.6 *Jede abgeschlossene konvexe Menge $\Omega \subset \mathbb{R}^n$ ist der Durchschnitt aller abgeschlossenen Halbräume, die Ω enthalten.*

Beweis: Beweisskizze. Bezeichnen wir den Durchschnitt aller abgeschlossenen Halbräume, die Ω enthalten, mit D. Dann ist $\Omega \subset D$ klar (Durchschnitt beliebig vieler konvexer Mengen ist konvex). Zu zeigen bleibt $D \subset \Omega$. Das geschieht indirekt, wobei der Widerspruch mit einer trennenden Hyperebene hergeleitet wird (sogenannter zweiter Trennungssatz, Details siehe [ERSD77, Abschnitt 2.1.4]). ■

Bemerkung 3.7 Eckpunkt, Extrempunkt. Der Eckpunkt oder Extrempunkt einer konvexen Menge wurde bereits in Definition 1.10, Teil I, definiert als ein Punkt, für den es keine echte konvexe Linearkombination gibt. Dafür, dass $\mathbf{x}_0 \in \Omega$ ein Extrempunkt ist, reicht es bereits aus, dass man keine $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$ findet, so dass

$$\mathbf{x}_0 = \frac{1}{2}\mathbf{x}_1 + \frac{1}{2}\mathbf{x}_2, \quad \mathbf{x}_1, \mathbf{x}_2 \neq \mathbf{x}_0. \quad \textit{Übungsaufgabe}$$

□

Satz 3.8 Existenz eines Extrempunkts. *Jede nichtleere, kompakte, konvexe Menge $\Omega \subset \mathbb{R}^n$ besitzt mindestens einen Extrempunkt.*

Beweis: Da Ω kompakt ist, gibt es ein $\mathbf{x}_0 \in \Omega$ mit $\|\mathbf{x}_0\|_2 \geq \|\mathbf{x}\|_2$ für alle $\mathbf{x} \in \Omega$. Sei

$$\mathbf{x}_0 = \frac{1}{2}\mathbf{x}_1 + \frac{1}{2}\mathbf{x}_2 = \mathbf{x}_1 + \frac{1}{2}(\mathbf{x}_2 - \mathbf{x}_1) = \mathbf{x}_2 + \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_2)$$

für $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$, $\mathbf{x}_1, \mathbf{x}_2 \neq \mathbf{x}_0$. Damit folgt

$$\|\mathbf{x}_1\|_2^2 \leq \|\mathbf{x}_0\|_2^2 = \|\mathbf{x}_1\|_2^2 + \underbrace{\frac{1}{4}\|\mathbf{x}_2 - \mathbf{x}_1\|_2^2 + (\mathbf{x}_2 - \mathbf{x}_1)^T \mathbf{x}_1}_{\geq 0, \text{ da } \|\mathbf{x}_0\|_2 \geq \|\mathbf{x}_1\|_2}.$$

Das heißt

$$\|\mathbf{x}_2 - \mathbf{x}_1\|_2^2 \geq -4(\mathbf{x}_2 - \mathbf{x}_1)^T \mathbf{x}_1$$

und analog

$$\|\mathbf{x}_2 - \mathbf{x}_1\|_2^2 \geq -4(\mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{x}_2 = 4(\mathbf{x}_2 - \mathbf{x}_1)^T \mathbf{x}_2.$$

Addition der Ungleichungen ergibt

$$2\|\mathbf{x}_2 - \mathbf{x}_1\|_2^2 \geq 4(\mathbf{x}_2 - \mathbf{x}_1)^T (\mathbf{x}_2 - \mathbf{x}_1) = 4\|\mathbf{x}_2 - \mathbf{x}_1\|_2^2$$

und damit $\mathbf{x}_2 = \mathbf{x}_1 = \mathbf{x}_0$. Somit ist $\mathbf{x}_0$ ein Extrempunkt. ■

3.2 Konvexe und konkave Funktionen

Definition 3.9 Konvexe, konkave Funktion. Die Funktion $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ heißt konvex, wenn für beliebige $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$ gilt

$$f((1-\lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2) \leq (1-\lambda)f(\mathbf{x}_1) + \lambda f(\mathbf{x}_2) \quad \forall\ \lambda \in [0,1].$$

Sie heißt konkav, wenn

$$f((1-\lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2) \geq (1-\lambda)f(\mathbf{x}_1) + \lambda f(\mathbf{x}_2) \quad \forall\ \lambda \in [0,1].$$

Gilt die echte Ungleichung spricht man von strenger Konvexität (Konkavität). □

Beispiel 3.10 Die lineare Funktion $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ mit $\mathbf{c} \in \mathbb{R}^n$ ist konvex und konkav in $\mathbb{R}^n$, jedoch nicht streng. □

Beispiel 3.11 Jede über $\mathbb{R}^n$ positiv semidefinite Bilinearform $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ mit $A = A^T \in \mathbb{R}^{n\times n}$ ist konvex. Sei nämlich

$$\mathbf{x} = (1-\lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2 = \mathbf{x}_1 + \lambda(\mathbf{x}_2 - \mathbf{x}_1), \quad \mathbf{x}_1 \neq \mathbf{x}_2 \in \mathbb{R}^n,\ \lambda \in [0,1],$$

dann gilt

$$\begin{aligned}
\mathbf{x}^T A\mathbf{x} &= \mathbf{x}_1^T A\mathbf{x}_1 + 2\lambda(\mathbf{x}_2 - \mathbf{x}_1)^T A\mathbf{x}_1 + \underbrace{\lambda^2}_{\leq\lambda}\underbrace{(\mathbf{x}_2 - \mathbf{x}_1)^T A(\mathbf{x}_2 - \mathbf{x}_1)}_{\geq 0} \\
&\leq \mathbf{x}_1^T A\mathbf{x}_1 + 2\lambda(\mathbf{x}_2 - \mathbf{x}_1)^T A\mathbf{x}_1 + \lambda(\mathbf{x}_2 - \mathbf{x}_1)^T A(\mathbf{x}_2 - \mathbf{x}_1) \\
&= \mathbf{x}_1^T A\mathbf{x}_1 + \lambda(\mathbf{x}_2 - \mathbf{x}_1)^T A(\mathbf{x}_1 + \mathbf{x}_2) \\
&= (1-\lambda)\mathbf{x}_1^T A\mathbf{x}_1 + \lambda\mathbf{x}_2^T A\mathbf{x}_2.
\end{aligned}$$

Analog zeigt man, dass

- A ist positiv definit, dann ist $f(\mathbf{x})$ streng konvex,
- A ist negativ semidefinit, dann ist $f(\mathbf{x})$ konkav,
- A ist negativ definit, dann ist $f(\mathbf{x})$ streng konkav.

□

Beispiel 3.12 Sind die Funktionen $f_j(\mathbf{x}), j = 1, \ldots, k$, über $\Omega \subset \mathbb{R}^n$ konvex, dann ist auch die Linearkombination

$$f(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j f_j(\mathbf{x}), \quad \alpha_j \geq 0 \; \forall \; j$$

konvex. Dies zeigt man durch direktes Nachrechnen. □

Zur Stetigkeit von konvexen Funktionen gibt es folgende Aussage.

Satz 3.13 Stetigkeit konvexer Funktionen. *Seien $\Omega \subset \mathbb{R}^n$ konvex und das Innere der Menge,* $\mathrm{int}(\Omega)$*, nichtleer. Dann ist jede konvexe Funktion $f : \Omega \to \mathbb{R}$ stetig in* $\mathrm{int}(\Omega)$*.*

Beweis: Siehe Literatur, zum Beispiel [ERSD77, Satz 2.65]. Man prüft die ε–δ–Definition nach. ■

Beispiel 3.14 Nichtstetige konvexe Funktion über konvexer Menge. Stetigkeit bis auf den Rand muss bei einer konvexen Funktion im allgemeinen nicht vorliegen. Die konvexe Funktion $f : [1, 2] \to \mathbb{R}$ mit

$$f(x) = \begin{cases} 1/x & \text{für } x \in [1, 2) \\ 2 & \text{für } x = 2 \end{cases}$$

ist ein Beispiel. □

Lemma 3.15 Monotonieaussage für den Differenzenquotienten. *Seien $\Omega \subset \mathbb{R}^n$ eine offene konvexe Menge, $f : \Omega \to \mathbb{R}$ mit $f \in C^1(\Omega)$ eine konvexe Funktion, $\mathbf{x}_0 \in \Omega$, $\mathbf{h} \in \mathbb{R}^n$, $\mathbf{x}_0 - \nu\mathbf{h} \in \Omega$ und $\mathbf{x}_0 + \rho\mathbf{h} \in \Omega$ für gewisse $\nu, \rho > 0$. Dann gilt*

$$\begin{aligned} \frac{f(\mathbf{x}_0) - f(\mathbf{x}_0 - \nu\mathbf{h})}{\nu} &\leq \frac{f(\mathbf{x}_0) - f(\mathbf{x}_0 - \mu\mathbf{h})}{\mu} \\ &\leq \frac{f(\mathbf{x}_0 + \lambda\mathbf{h}) - f(\mathbf{x}_0)}{\lambda} \leq \frac{f(\mathbf{x}_0 + \rho\mathbf{h}) - f(\mathbf{x}_0)}{\rho} \end{aligned}$$

für alle $\mu \in (0, \nu], \lambda \in (0, \rho]$. Damit ist für konvexe Funktionen der Differenzenquotient im Punkt $\mathbf{x}_0$ in Richtung $\mathbf{h}$ eine monoton wachsende Funktion von λ.

Beweis: *Übungsaufgabe* ■

Jetzt werden einige Bedingung dafür gegeben, dass ein Funktion auf einer konvexen Menge konvex ist.

Satz 3.16 Hinreichende und notwendige Bedingung für Konvexität. *Seien $\Omega \subset \mathbb{R}^n$ eine offene konvexe Menge und $f : \Omega \to \mathbb{R}$ mit $f \in C^1(\Omega)$. Dann ist $f(\mathbf{x})$ auf Ω genau dann konvex, wenn für alle $\mathbf{x}_0, \mathbf{x}_1 \in \Omega$ mit $\mathbf{x}_0 \neq \mathbf{x}_1$ gilt*

$$f(\mathbf{x}_1) - f(\mathbf{x}_0) \geq (\mathbf{x}_1 - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0). \tag{3.2}$$

Beweis: 1) *Notwendigkeit von (3.2).* Da $f(\mathbf{x})$ differenzierbar ist, gilt für die Richtungsableitung in Richtung $\mathbf{h} \in \mathbb{R}^n$ mit $\|\mathbf{h}\|_2 = 1$

$$\frac{\partial f}{\partial \mathbf{h}}(\mathbf{x}_0) = \mathbf{h}^T \nabla f(\mathbf{x}_0) = \lim_{\lambda \to 0} \frac{f(\mathbf{x}_0 + \lambda \mathbf{h}) - f(\mathbf{x}_0)}{\lambda} = -\lim_{\lambda \to 0} \frac{f(\mathbf{x}_0 - \lambda \mathbf{h}) - f(\mathbf{x}_0)}{\lambda},$$

wobei bei der letzten Gleichung λ durch $-\lambda$ ersetzt wurde. Nach Lemma 3.15 gilt alle $\rho > 0$ mit $\mathbf{x}_0 + \rho \mathbf{h} \in \Omega$ (nehme $\lambda \to 0$ in Lemma 3.15)

$$\frac{f(\mathbf{x}_0 + \rho \mathbf{h}) - f(\mathbf{x}_0)}{\rho} \geq \mathbf{h}^T \nabla f(\mathbf{x}_0)$$

oder

$$f(\mathbf{x}_0 + \rho \mathbf{h}) - f(\mathbf{x}_0) \geq \rho \mathbf{h}^T \nabla f(\mathbf{x}_0).$$

Wählt man $\mathbf{x}_1 = \mathbf{x}_0 + \rho \mathbf{h}$, so ergibt sich (3.2).

2) *Hinlänglichkeit von (3.2).* Sei $\lambda_0 > 0$ beliebig gewählt, setzte $\lambda_1 = 1 - \lambda_0$. Dann gelten für $\mathbf{x}_2 = \lambda_0 \mathbf{x}_0 + \lambda_1 \mathbf{x}_1$ nach Voraussetzung

$$f(\mathbf{x}_0) - f(\mathbf{x}_2) \geq (\mathbf{x}_0 - \mathbf{x}_2)^T \nabla f(\mathbf{x}_2), \quad f(\mathbf{x}_1) - f(\mathbf{x}_2) \geq (\mathbf{x}_1 - \mathbf{x}_2)^T \nabla f(\mathbf{x}_2).$$

Damit folgt

$$\begin{aligned} \lambda_0 f(\mathbf{x}_0) + \lambda_1 f(\mathbf{x}_1) &\geq f(\mathbf{x}_2) + \Big(\lambda_0(\mathbf{x}_0 - \mathbf{x}_2) + \lambda_1(\mathbf{x}_1 - \mathbf{x}_2)\Big)^T \nabla f(\mathbf{x}_2) \\ &= f(\mathbf{x}_2) + \underbrace{\Big(\lambda_0 \mathbf{x}_0 + \lambda_1 \mathbf{x}_1 - \mathbf{x}_2\Big)^T}_{=\mathbf{0}} \nabla f(\mathbf{x}_2) \\ &= f(\mathbf{x}_2) = f(\lambda_0 \mathbf{x}_0 + \lambda_1 \mathbf{x}_1) \end{aligned}$$

für alle $\mathbf{x}_0, \mathbf{x}_1 \in \Omega$. ■

Bemerkung 3.17 Taylor–Entwicklung. Für eine zweimal stetig differenzierbare Funktion $f : \Omega \to \mathbb{R}, \Omega \subset \mathbb{R}^n$, offen, ist die Hesse–Matrix

$$H_f(\mathbf{x}) := \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{i,j=1,\ldots,n} (\mathbf{x})$$

symmetrisch (Satz von Schwarz).

Liegt zusammen mit $\mathbf{x}_0, \mathbf{x}_1$ die Strecke $[\mathbf{x}_0, \mathbf{x}_1]$ vollständig in Ω, so gilt für eine geeignete Zahl $\lambda \in (0,1)$ nach dem Satz von Taylor

$$f(\mathbf{x}_1) = f(\mathbf{x}_0) + (\mathbf{x}_1 - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_0)^T H_f(\mathbf{x}_0 + \lambda(\mathbf{x}_1 - \mathbf{x}_0))(\mathbf{x}_1 - \mathbf{x}_0). \quad (3.3)$$

□

Satz 3.18 Hinreichende und notwendige Bedingung für Konvexität mit Hesse–Matrix. *Seien $\Omega \subset \mathbb{R}^n$ eine offene konvexe Menge und $f : \Omega \to \mathbb{R}$ mit $f \in C^2(\Omega)$. Dann ist $f(\mathbf{x})$ auf Ω genau dann konvex, wenn $H_f(\mathbf{x})$ positiv semidefinit ist für alle $\mathbf{x} \in \Omega$.*

Beweis: 1.) Sei $H_f(\mathbf{x})$ positiv semidefinit. Dann folgt aus (3.3) für beliebige $\mathbf{x}_0, \mathbf{x}_1 \in \Omega$ und eine geeignete Zahl $\lambda \in (0,1)$

$$f(\mathbf{x}_1) - f(\mathbf{x}_0) - (\mathbf{x}_1 - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_0)^T H_f(\mathbf{x}_0 + \lambda(\mathbf{x}_1 - \mathbf{x}_0))(\mathbf{x}_1 - \mathbf{x}_0) \geq 0.$$

Aus Satz 3.16 folgt, dass $f(\mathbf{x})$ konvex ist.

2) Sei $f(\mathbf{x})$ konvex auf Ω. Seien $\mathbf{x}_0 \in \Omega$ und $\mathbf{h} \in \mathbb{R}^n$ beliebig gegeben. Da Ω offen ist, gibt es eine Zahl $\theta_0 > 0$ so dass $\mathbf{x}_0 + \theta \mathbf{h} \in \Omega$ für alle $\theta \in (0, \theta_0]$. Wählt man in (3.2) $\mathbf{x}_1 = \mathbf{x}_0 + \theta \mathbf{h}$, erhält man

$$f(\mathbf{x}_0 + \theta \mathbf{h}) - f(\mathbf{x}_0) - \theta \mathbf{h}^T \nabla f(\mathbf{x}_0) \geq 0 \quad \forall\, \theta \in (0, \theta_0].$$

Aus (3.3) folgt nun

$$\mathbf{h}^T H_f(\mathbf{x}_0 + \lambda\theta\mathbf{h})\mathbf{h} \geq 0 \quad \forall\, \theta \in (0, \theta_0]$$

mit $\lambda \in (0,1)$. Da die Hesse–Matrix nach Voraussetzung stetig ist, folgt für $\theta \to +0$

$$\mathbf{h}^T H_f(\mathbf{x}_0)\mathbf{h} \geq 0.$$

Da $\mathbf{x}_0 \in \Omega$ und $\mathbf{h} \in \mathbb{R}^n$ beliebig gewählt wurden, ist $H_f(\mathbf{x})$ in jedem Punkt $\mathbf{x} \in \Omega$ positiv semidefinit. ■

Beispiel 3.19 1D. Eine skalare Funktion einer skalaren Veränderlichen $f : (a,b) \to \mathbb{R}$ mit $f \in C^2(a,b)$ ist genau dann konvex, wenn $f''(x) \geq 0$ in (a,b), zum Beispiel die Funktion $f(x) = x^2$. □

Bemerkung 3.20 Es gibt Verallgemeinerungen des Begriffs der konvexen Funktion, zum Beispiel quasi–konvex und explizit konvex, siehe Literatur. □

3.3 Ungleichungen und konvexe Mengen

Satz 3.21 *Die Funktion $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ sei konvex, Ω sei konvex und sei $b \in \mathbb{R}$. Dann ist*

$$V := \{\mathbf{x} \in \Omega \;:\; f(\mathbf{x}) \leq b\}$$

konvex.

Beweis: Zu zeigen ist: $\mathbf{x}_1, \mathbf{x}_2 \in V \implies \mathbf{x} = (1-\lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2 \in V$ für alle $\lambda \in [0,1]$. Aus $f(\mathbf{x}_1), f(\mathbf{x}_2) \leq b$ folgt aus der Konvexität von f

$$f(\mathbf{x}) \leq (1-\lambda)f(\mathbf{x}_1) + \lambda f(\mathbf{x}_2) \leq b.$$

Damit ist $\mathbf{x} \in V$. ■

Bemerkung 3.22 Unterhalbmenge.

- Die Menge V aus Satz 3.21 heißt Unterhalbmenge von $f(\mathbf{x})$ in Ω zum Wert b.
- Die Niveaumenge $\{\mathbf{x} \in \Omega \;:\; f(\mathbf{x}) = b\}$ der konvexen Funktion $f(\mathbf{x})$ ist im allgemeinen nicht konvex. Zum Beispiel besteht die Niveaumenge zu $b = 1$ der konvexen Funktion $f(x) = x^2, x \in \mathbb{R}$, aus den Punkten -1 und 1.

□

Folgerung 3.23 *Sei $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ konkav, sei Ω konvex und sei $b \in \mathbb{R}$. Dann ist $U := \{\mathbf{x} \in \Omega \;:\; f(\mathbf{x}) \geq b\}$ konvex.*

Beweis: Nutze Satz 3.21 mit $-f(\mathbf{x})$, da diese Funktion konvex ist. ■

3.4 Extrema von konvexen Funktionen

Es wird die Frage untersucht, unter welchen Bedingungen eine konvexe Funktion ihr globales Minimum über einer gegebenen konvexen Menge Ω annimmt.

Satz 3.24 Lokales Minimum ist globales Minimum. *Sei $f : \Omega \to \mathbb{R}$ eine konvexe Funktion über einer konvexen Menge Ω. Dann ist ein lokales Minimum von $f(\mathbf{x})$ in Ω zugleich das globale Minimum über Ω.*

Beweis: Indirekt. Nehme $f(\mathbf{x})$ ein lokales Minimum in $\mathbf{x}_0 \in \Omega$ an. Sei weiter $\mathbf{x}^* \in \Omega$ mit $f(\mathbf{x}^*) < f(\mathbf{x}_0)$. Aus der Konvexität von $f(\mathbf{x})$ folgt

$$f((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}^*) \leq (1-\lambda)f(\mathbf{x}_0) + \lambda f(\mathbf{x}^*) = f(\mathbf{x}_0) + \lambda\left(f(\mathbf{x}^*) - f(\mathbf{x}_0)\right) < f(\mathbf{x}_0)$$

für alle $\lambda \in (0,1)$, da $f(\mathbf{x}^*) < f(\mathbf{x}_0)$. Aus der Konvexität von Ω folgt, dass der Funktionswert im Zwischenpunkt wohldefiniert ist. Wählt man λ beliebig klein, so ist der Zwischenpunkt beliebig nahe an $\mathbf{x}_0$ und die obige Ungleichung steht im Widerspruch dazu, dass $f(\mathbf{x})$ in $\mathbf{x}_0$ ein lokales Minimum annimmt. ■

Beispiel 3.25 Konvexe Funktion über konvexer Menge ohne Annahme eines lokalen Minimums. Der obige Satz kann nur angewendet werden, wenn es ein lokales Minimum gibt, das angenommen wird. Das muss nicht der Fall sein. Die konvexe Funktion aus Beispiel 3.14 nimmt ihr Minimum nicht an. □

Folgerung 3.26 *Die Menge Ω_0, über der eine konvexe Funktion $f : \Omega \to \mathbb{R}$ über einer konvexen Menge Ω das globale Minimum annimmt, ist konvex.*

Beweis: Folgt aus Satz 3.21. ■

Satz 3.27 Bedingung für globales Minimum für stetig differenzierbare konvexe Funktionen. *Sei $f(\mathbf{x}) \in C^1(\Omega)$ über einer konvexen Menge Ω konvex. Dann nimmt $f(\mathbf{x})$ in jedem $\mathbf{x}_1 \in \Omega$ mit $\nabla f(\mathbf{x}_1) = \mathbf{0}$ sein globales Minimum über Ω an.*

Beweis: Es ist zu zeigen, dass alle Punkte $\mathbf{x}_1 \in \Omega$ mit $\nabla f(\mathbf{x}_1) = \mathbf{0}$ wirklich Extremalpunkte sind und zudem Minima.

Sei $\mathbf{x}_1 \in \Omega$ mit $\nabla f(\mathbf{x}_1) = \mathbf{0}$. Da $f(\mathbf{x})$ konvex ist, gilt

$$f\left(\lambda\mathbf{x}_2 + (1-\lambda)\mathbf{x}_1\right) \leq (1-\lambda)f(\mathbf{x}_1) + \lambda f(\mathbf{x}_2) = f(\mathbf{x}_1) + \lambda\left(f(\mathbf{x}_2) - f(\mathbf{x}_1)\right)$$

für alle $\mathbf{x}_2 \in \Omega$ und $\lambda \in [0,1]$. Für $\lambda \in (0,1]$ erhält man daraus

$$\frac{f\left(\mathbf{x}_1 + \lambda(\mathbf{x}_2 - \mathbf{x}_1)\right) - f(\mathbf{x}_1)}{\lambda} \leq f(\mathbf{x}_2) - f(\mathbf{x}_1).$$

Nach dem Mittelwertsatz der Differentialrechnung gibt es ein $\theta \in [0,1]$ mit

$$(\mathbf{x}_2 - \mathbf{x}_1)^T \nabla f\left(\mathbf{x}_1 + \theta\lambda(\mathbf{x}_2 - \mathbf{x}_1)\right) \leq f(\mathbf{x}_2) - f(\mathbf{x}_1).$$

Für $\lambda \to 0$ folgt nun mit $\nabla f(\mathbf{x}_1) = \mathbf{0}$

$$0 = (\mathbf{x}_2 - \mathbf{x}_1)^T \nabla f\left(\mathbf{x}_1\right) \leq f(\mathbf{x}_2) - f(\mathbf{x}_1),$$

also $f(\mathbf{x}_1) \leq f(\mathbf{x}_2)$ für alle $\mathbf{x}_2 \in \Omega$. ■

Beispiel 3.28 Nichteindeutigkeit des globalen Minimums. Das globale Minimum kann an mehr als einem Punkt angenommen werden. Betrachte zum Beispiel die konvexe Funktion $f : [0.9, 2] \to \mathbb{R}$

$$f(x) = \begin{cases} \exp\left(-\dfrac{1}{1-x^2}\right) & \text{für } x \in [0.9, 1), \\ 0 & \text{für } x \in [1, 2]. \end{cases}$$

Es gilt $f \in C^\infty([0.9, 2])$. Das globale Minimum wird in $[1, 2]$ angenommen. In diesem Intervall gilt auch $f'(x) = 0$. □

Kapitel 4

Optimalitätskriterien

Bemerkung 4.1 Motivation. Als Optimalitätskriterien bezeichnet man notwendige oder hinreichende Bedingungen dafür, dass ein $\mathbf{x}_0 \in \Omega \subset \mathbb{R}^n$ Lösung eines Optimierungsproblems ist. Diese Kriterien besitzen sowohl unter theoretischen als auch unter numerischen Aspekten Bedeutung. In diesem Kapitel werden lokale und globale Optimalitätskriterien hergeleitet. □

4.1 Einleitung

Bemerkung 4.2 Methode der Lagrangeschen Multiplikatoren. Ein Gegenstand dieses Kapitels ist die Verallgemeinerung der Methode der Lagrangeschen Multiplikatoren auf Probleme mit Bedingungsungleichungen und beschränkten Variablen. Es wird der Zusammenhang zwischen der Lösung eines Optimierungsproblems und dem Sattelpunkt eines Lagrangeschen Funktionals hergestellt.

Wir betrachten zunächst die Lagrangesche Methode für ein Optimierungsproblem mit Nebenbedingungen in Gleichungsform

$$\begin{aligned} \{f(\mathbf{x}) \ : \ \mathbf{x} \in \Omega \subset \mathbb{R}^n\} \quad &\rightarrow \quad \min \ !, \\ \Omega \quad &= \quad \{\mathbf{x} \in \mathbb{R}^n \ : \ h_i(\mathbf{x}) = 0, \ i \in \{1, \ldots, p\}, \ p < n\}\,. \end{aligned} \tag{4.1}$$

Seien $f(\mathbf{x})$ in Ω und $h_i(\mathbf{x})$, $i = 1, \ldots, p$, in $\mathbb{R}^n$ differenzierbar. Für beliebiges festes $\mathbf{x} \in \Omega$ werden die Matrizen

$$H_0(\mathbf{x}) := (\nabla h_1(\mathbf{x}), \ldots, \nabla h_p(\mathbf{x}))^T \in \mathbb{R}^{p \times n}, \quad H(\mathbf{x}) := \begin{pmatrix} H_0(\mathbf{x}) \\ \nabla f(\mathbf{x})^T \end{pmatrix} \in \mathbb{R}^{(p+1) \times n}$$

definiert.

Für das Problem (4.1) wird die Lagrange–Funktion

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \lambda_0 f(\mathbf{x}) + \sum_{i=1}^{p} \lambda_i h_i(\mathbf{x}), \quad \mathbf{x} \in \Omega, \ \boldsymbol{\lambda} \in \mathbb{R}^{p+1},$$

eingeführt. Die Zahlen $\lambda_0, \ldots, \lambda_p$ nennt man Lagrangesche Multiplikatoren. □

Satz 4.3 Optimalitätskriterium für Problem (4.1). *Seien $\mathbf{x}_0 \in \Omega$ und $rg(H_0(\mathbf{x}_0)) = rg(H(\mathbf{x}_0))$. Dann gilt: besitzt $f(\mathbf{x})$ in $\mathbf{x}_0$ ein lokales Minimum bezüglich Ω, so existiert ein $\boldsymbol{\lambda}^{(0)} = \left(\lambda_0^{(0)}, \ldots, \lambda_p^{(0)}\right)^T \in \mathbb{R}^{p+1}$, mit $\lambda_0^{(0)} \neq 0$ und*

$$\nabla_{\mathbf{x}} L(\mathbf{x}_0, \boldsymbol{\lambda}^{(0)}) = \lambda_0^{(0)} \nabla f(\mathbf{x}_0) + \sum_{i=1}^{p} \lambda_i^{(0)} \nabla h_i(\mathbf{x}_0) = \mathbf{0}. \tag{4.2}$$

Beweis: Siehe Literatur. ■

Bemerkung 4.4 Fortsetzung: Methode der Lagrangeschen Multiplikatoren. Verallgemeinerungen dieser Aussage auf den Fall, dass die Matrizen unterschiedlichen Rang besitzen, sind möglich. Mit Hilfe des Gleichungssystems (4.2) hat man ein Kriterium zur Bestimmung von Punkten, in denen $f(\mathbf{x})$ ein lokales Minimum annehmen kann. Sofern dies möglich ist, berechnet man alle Lösungen $(\mathbf{x}, \boldsymbol{\lambda})$ von (4.2). Im allgemeinen sind jedoch nicht alle Lösungen auch lokale Extrempunkte.

Das heißt, man erweitert das gegebene Problem so, dass

- man für die Lösung des erweiterten Problems Standardkriterien, zum Beispiel dass Ableitungen verschwinden, für ein Minimum hat,
- die Lösung des erweiterten Problems Rückschlüsse auf die Lösung des gegebenen Problems zulässt.

□

4.2 Lokale Minima für Optimierungsprobleme ohne Einschränkungen an das zulässige Gebiet

Bemerkung 4.5 Ziel. Wir betrachten das Optimierungsproblem

$$z = \min\{f(\mathbf{x}) \ : \ \mathbf{x} \in \Omega\} \tag{4.3}$$

mit $f \in C^1(\mathbb{R}^n)$ (an sich reicht $f \in C^1(\tilde{\Omega})$ mit $\overline{\Omega} \subset \tilde{\Omega}$, $\tilde{\Omega}$ offen). Für diese Problem sollen im folgenden lokale Optimalitätskriterien hergeleitet werden. □

Zunächst werden spezielle konvergente Folgen betrachtet.

Definition 4.6 Konvergent gegen $\mathbf{x}_0$ in Richtung $\mathbf{y}$, gerichtet konvergent. Eine konvergente Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$, $\mathbf{x}_k \in \mathbb{R}^n$, mit $\mathbf{x}_k \to \mathbf{x}_0$ heißt konvergent gegen $\mathbf{x}_0$ in Richtung $\mathbf{y} \in \mathbb{R}^n$ oder gerichtet konvergent, wenn gilt

$$\lim_{k\to\infty} \frac{\mathbf{x}_k - \mathbf{x}_0}{\|\mathbf{x}_k - \mathbf{x}_0\|_2} = \mathbf{y}, \quad \|\mathbf{y}\|_2 = 1.$$

Die Schreibweise ist

$$\mathbf{x}_k \xrightarrow{\mathbf{y}} \mathbf{x}_0.$$

□

Beispiel 4.7 Sei $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ mit

$$\mathbf{x}_k = \left(\frac{1}{k}, \frac{1}{k}\right)^T.$$

Dann ist $\mathbf{x}_0 = (0,0)^T$ und es gilt

$$\lim_{k\to\infty} \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|_2} = \lim_{k\to\infty} \frac{k}{\sqrt{2}} \begin{pmatrix} \frac{1}{k} \\ \frac{1}{k} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \mathbf{y}.$$

□

Bemerkung 4.8 Gerichtet konvergente Folgen. Der Vektor $\mathbf{y}$ beschreibt die Richtung, aus der man sich mit der Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ gegen $\mathbf{x}_0$ annähert.

Man kann aus jeder gegen $\mathbf{x}_0$ konvergenten Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ eine gerichtet konvergente Teilfolge auswählen, sofern unendlich viele Elemente der Folge ungleich $\mathbf{x}_0$ sind. Sei $\mathbf{x}_k \neq \mathbf{x}_0$ für $k \geq k_0$, dann sind alle Glieder der Folge

$$\mathbf{y}_k := \frac{\mathbf{x}_k - \mathbf{x}_0}{\|\mathbf{x}_k - \mathbf{x}_0\|_2}, \quad k \geq k_0,$$

Elemente der kompakten Menge $\{\mathbf{x} \in \mathbb{R}^n \ : \ \|\mathbf{x}\|_2 = 1\}$. Damit besitzt $\{\mathbf{y}_k\}_{k\geq k_0}$ einen Häufungspunkt in dieser Menge und man findet in $\{\mathbf{y}_k\}_{k\geq k_0}$ eine konvergente Teilfolge. □

Die Aussage des folgenden Lemmas kann in gewisser Weise als Verallgemeinerung der Richtungsableitung aufgefasst werden.

Lemma 4.9 *Die Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ sei konvergent gegen $\mathbf{x}_0$ in Richtung $\mathbf{y}$. Dann gelten:*

1. *Ist $f \ : \ \mathbb{R}^n \to \mathbb{R}$ in $\mathbf{x}_0$ stetig differenzierbar, so folgt*

$$\lim_{k\to\infty} \frac{f(\mathbf{x}_k) - f(\mathbf{x}_0)}{\|\mathbf{x}_k - \mathbf{x}_0\|_2} = \mathbf{y}^T \nabla f(\mathbf{x}_0).$$

2. *Ist $f \ : \ \mathbb{R}^n \to \mathbb{R}$ in $\mathbf{x}_0$ zweimal stetig differenzierbar, so gilt*

$$\lim_{k\to\infty} \frac{f(\mathbf{x}_k) - f(\mathbf{x}_0) - (\mathbf{x}_k - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0)}{\|\mathbf{x}_k - \mathbf{x}_0\|_2^2} = \frac{1}{2}\mathbf{y}^T H_f(\mathbf{x}_0)\mathbf{y}.$$

Beweis: Es wird nur die erste Aussage bewiesen, der Beweis der zweiten Aussage ist analog.

Da $f(\mathbf{x})$ in $\mathbf{x}_0$ differenzierbar ist, also insbesondere alle Richtungsableitungen existieren, folgt unter Nutzung der Definition der Richtungsableitung, dass

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \left(\frac{f(\mathbf{x}) - f(\mathbf{x}_0)}{\|\mathbf{x} - \mathbf{x}_0\|_2} - \frac{(\mathbf{x} - \mathbf{x}_0)^T}{\|\mathbf{x} - \mathbf{x}_0\|_2} \nabla f(\mathbf{x}_0) \right) = 0.$$

Außerdem existiert nach Voraussetzung der Grenzwert

$$\lim_{\mathbf{x}_k\to\mathbf{x}_0} \frac{(\mathbf{x}_k - \mathbf{x}_0)^T}{\|\mathbf{x}_k - \mathbf{x}_0\|_2} \nabla f(\mathbf{x}_0) = \mathbf{y}^T \nabla f(\mathbf{x}_0).$$

Damit folgt

$$\lim_{\mathbf{x}_k\to\mathbf{x}_0} \frac{f(\mathbf{x}_k) - f(\mathbf{x}_0)}{\|\mathbf{x}_k - \mathbf{x}_0\|_2} = \lim_{\mathbf{x}_k\to\mathbf{x}_0} \frac{(\mathbf{x}_k - \mathbf{x}_0)^T}{\|\mathbf{x}_k - \mathbf{x}_0\|_2} \nabla f(\mathbf{x}_0) = \mathbf{y}^T \nabla f(\mathbf{x}_0).$$

■

Definition 4.10 Tangentenkegel. Der Tangentenkegel $T(\mathbf{x}_0)$ an die Menge Ω im Punkt $\mathbf{x}_0 \in \Omega$ ist gegeben durch

$$T(\mathbf{x}_0) := \left\{ \lambda\mathbf{y} \ : \ \|\mathbf{y}\|_2 = 1, \ \exists\{\mathbf{x}_k\}_{k\in\mathbb{N}} \in \Omega, \ \mathbf{x}_k \xrightarrow{\mathbf{y}} \mathbf{x}_0, \ \lambda \geq 0 \right\}.$$

□

Bemerkung 4.11 Tangentenkegel. Der Tangentenkegel $T(\mathbf{x}_0)$ hat nur etwas mit dem zulässigen Gebiet Ω zu tun und nicht mit der zu minimierenden Funktion $f(\mathbf{x})$. Er beschreibt die Richtungen, aus denen man sich $\mathbf{x}_0$ mit einer Folge annähern kann, deren Glieder alle in Ω liegen. Der Begriff des Tangentenkegels ist eine Verallgemeinerung der Tangetialhyperebene. Die Menge $T(\mathbf{x}_0)$ ist ein Kegel. Für jeden inneren Punkt $\mathbf{x}_0 \in \Omega$ gilt $T(\mathbf{x}_0) = \mathbb{R}^n$. Für isolierte Punkte setzen wir $T(\mathbf{x}_0) = \{\mathbf{0}\}$. □

Beispiel 4.12 Sei

$$\Omega = \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1^2 + x_2^2 = 1, \ x_1 \geq 0, \ 0 < x_2 < 2\}.$$

Für den Punkt $\mathbf{x}_0 = (0,1)^T \in \Omega$ erhält man, direkt aus der geometrischen Anschauung,

$$T(\mathbf{x}_0) = \left\{ \mathbf{y} \in \mathbb{R}^2 \ : \ \mathbf{y} = \lambda \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \lambda \geq 0 \right\}.$$

Analytisches Nachrechnen: Übungsaufgabe □

Lemma 4.13 *Für jeden Punkt $\mathbf{x}_0 \in \Omega$ ist der Tangentenkegel $T(\mathbf{x}_0)$ abgeschlossen.*

Beweis: Siehe Literatur. ■

Der folgende Satz gibt ein notwendiges Kriterium für die Existenz eines lokalen Minimums von $f(\mathbf{x})$ bezüglich Ω an.

Satz 4.14 Notwendiges Kriterium für Existenz eines lokalen Minimums. *Sei $f \ : \ \mathbb{R}^n \to \mathbb{R}$ in $\mathbf{x}_0 \in \Omega$ stetig differenzierbar. Besitzt $f(\mathbf{x})$ in $\mathbf{x}_0$ ein lokales Minimum bezüglich Ω, dann gilt*

$$\mathbf{y}^T \nabla f(\mathbf{x}_0) \geq 0 \quad \forall \ \mathbf{y} \in T(\mathbf{x}_0).$$

Beweis: Sei $\mathbf{y} \in T(\mathbf{x}_0)$. Für $\mathbf{y} = \mathbf{0}$ ist $\mathbf{y}^T \nabla f(\mathbf{x}_0) = 0$. Ansonsten existieren ein $\lambda \in \mathbb{R}, \lambda > 0$, und ein $\mathbf{y}^* \in T(\mathbf{x}_0)$, $\|\mathbf{y}^*\|_2 = 1$, mit $\mathbf{y} = \lambda \mathbf{y}^*$. Ferner sei $\{\mathbf{x}_k\}_{k \in \mathbb{N}} \subset \Omega$ eine Folge mit $\mathbf{x}_k \xrightarrow{\mathbf{y}^*} \mathbf{x}_0$.

Da $f(\mathbf{x})$ in $\mathbf{x}_0$ ein lokales Minimum bezüglich Ω annimmt, gilt für hinreichend große k die Ungleichung $f(\mathbf{x}_k) \geq f(\mathbf{x}_0)$. Mit dieser Beziehung folgt, unter Beachtung von Lemma 4.9, 1),

$$\lim_{k \to \infty} \frac{f(\mathbf{x}_k) - f(\mathbf{x}_0)}{\|\mathbf{x}_k - \mathbf{x}_0\|_2} = (\mathbf{y}^*)^T \nabla f(\mathbf{x}_0) \geq 0$$

und damit auch $\mathbf{y}^T \nabla f(\mathbf{x}_0) = \lambda \, (\mathbf{y}^*)^T \nabla f(\mathbf{x}_0) \geq 0$. ■

Bemerkung 4.15 Minimum auf Rand des Gebiets. Das lokale Minimum kann auch auf dem Rand von Ω liegen. Beachte, dass $f(\mathbf{x})$ nach Voraussetzung auf dem Rand von Ω stetig differenzierbar ist.

Aus diesem Satz folgt nicht, dass $f(\mathbf{x})$ in $\mathbf{x}_0$ ein lokales Minimum nur annehmen kann, wenn $\nabla f(\mathbf{x}_0) = \mathbf{0}$ gilt. Gilt jedoch $\nabla f(\mathbf{x}_0) = \mathbf{0}$, dann hat man ein zweites notwendiges Kriterium. □

Satz 4.16 Zweites notwendiges Kriterium. *Sei $f \ : \mathbb{R}^n \to \mathbb{R}$ in $\mathbf{x}_0 \in \Omega$ zweimal stetig differenzierbar. Besitzt $f(\mathbf{x})$ in $\mathbf{x}_0$ ein lokales Minimum bezüglich Ω und gilt $\nabla f(\mathbf{x}_0) = \mathbf{0}$, dann folgt*

$$\mathbf{y}^T H_f(\mathbf{x}_0) \mathbf{y} \geq 0 \quad \forall \ \mathbf{y} \in T(\mathbf{x}_0).$$

Beweis: Der Beweis ist analog zum Beweis von Satz 4.14, wobei jedoch jetzt Lemma 4.9, 2) verwendet wird. Aus $(\mathbf{x}_k - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0) = 0$ und $f(\mathbf{x}_k) \geq f(\mathbf{x}_0)$ für hinreichend große Indizes k folgt dann

$$\lim_{k \to \infty} \frac{f(\mathbf{x}_k) - f(\mathbf{x}_0) - (\mathbf{x}_k - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0)}{\|\mathbf{x}_k - \mathbf{x}_0\|_2^2} = \frac{1}{2} \, (\mathbf{y}^*)^T \, H_f(\mathbf{x}_0) \mathbf{y}^* \geq 0.$$

■

Der folgende Satz enthält eine hinreichende Bedingung für die Existenz eines isolierten lokalen Minimums von $f(\mathbf{x})$ bezüglich Ω.

Satz 4.17 Hinreichende Bedingung für die Existenz eines isolierten lokalen Minimums. *Sei* $f : \mathbb{R}^n \to \mathbb{R}$ *in* $\mathbf{x}_0 \in \Omega$ *zweimal stetig differenzierbar. Gelten* $\nabla f(\mathbf{x}_0) = \mathbf{0}$ *und* $\mathbf{y}^T H_f(\mathbf{x}_0)\mathbf{y} > 0$ *für alle* $\mathbf{y} \in T(\mathbf{x}_0), \mathbf{y} \neq \mathbf{0}$*, dann besitzt* $f(\mathbf{x})$ *in* $\mathbf{x}_0$ *ein isoliertes lokales Minimum bezüglich* Ω*.*

Beweis: Indirekter Beweis. Wir nehmen an, dass $f(\mathbf{x})$ in $\mathbf{x}_0$ kein isoliertes lokales Minimum bezüglich Ω besitzt. Dann existiert eine Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}} \subset \Omega$ mit $\mathbf{x}_k \to \mathbf{x}_0$, mit unendlich vielen Elementen, die ungleich $\mathbf{x}_0$ sind, und $f(\mathbf{x}_k) \leq f(\mathbf{x}_0), k \geq 1$. Es existiert eine Teilfolge von $\{\mathbf{x}_k\}$, die gegen ein $\tilde{\mathbf{y}}$ gerichtet konvergent ist. Ohne Beschränkung der Allgemeinheit kann man annehmen, dass die gesamte Folge gegen $\mathbf{x}_0$ aus Richtung $\tilde{\mathbf{y}}$ konvergiert: $\mathbf{x}_k \xrightarrow{\tilde{\mathbf{y}}} \mathbf{x}_0$. Dann folgt aus Lemma 4.9

$$\lim_{k\to\infty} \frac{f(\mathbf{x}_k) - f(\mathbf{x}_0)}{\|\mathbf{x}_k - \mathbf{x}_0\|_2^2} = \frac{1}{2}\tilde{\mathbf{y}}^T H_f(\mathbf{x}_0)\tilde{\mathbf{y}} \leq 0$$

mit $\tilde{\mathbf{y}} \in T(\mathbf{x}_0)$, im Widerspruch zur Voraussetzung. ■

Bemerkung 4.18 Bekannte Kriterien für innere Punkte. Für jeden inneren Punkt $\mathbf{x}_0 \in \Omega$ gilt $T(\mathbf{x}_0) = \mathbb{R}^n$ und aus den Aussagen der Sätze 4.14, 4.16 und 4.17 folgen die bekannten notwendigen und hinreichenden Bedingungen für die Existenz lokaler Minima einer Funktion $f : \mathbb{R}^n \to \mathbb{R}$. *Übungsaufgabe* □

4.3 Lokale Minima für Optimierungsprobleme, bei denen das zulässige Gebiet durch Ungleichungen gegeben ist

Bemerkung 4.19 Ziel. In diesem Abschnitt wird das Optimierungsproblem

$$z = \min\{f(\mathbf{x}) \ : \ \mathbf{x} \in \Omega\} \quad \text{mit } \Omega = \{\mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\} \tag{4.4}$$

mit $f \in C^1(\mathbb{R}^n)$ und $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{g} \in C^1(\mathbb{R}^n)$ untersucht. Unter Verwendung der Resultate aus Abschnitt 4.2 werden nun lokale Optimierungskriterien für (4.4) hergeleitet. Dazu wird eine lokale Theorie von Lagrange–Multiplikatoren entwickelt. Lokal bedeutet, dass das Gebiet lokal durch affine Bedingungen approximiert wird. □

Definition 4.20 Aktive Nebenbedingung. Eine Nebenbedingung $g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m$, wird im Punkt $\mathbf{x}_0 \in \Omega$ aktiv genannt, wenn gilt $g_i(\mathbf{x}_0) = 0$. Bezeichnung:

$$I_0 := \{i \in \{1, \ldots, m\} \ : \ g_i(\mathbf{x}_0) = 0\}.$$

□

Bemerkung 4.21 Sei $I_0 \neq \emptyset$. Die in $\mathbf{x}_0$ aktiven Nebenbedingungen werden nun durch affine Funktionen ersetzt

$$(\nabla g_i(\mathbf{x}_0))^T (\mathbf{x} - \mathbf{x}_0), \ \mathbf{x} \in \mathbb{R}^n, \ i \in I_0,$$

beziehungsweise in Matrixnotation

$$(\nabla \mathbf{g}_{I_0}(\mathbf{x}_0))^T (\mathbf{x} - \mathbf{x}_0), \quad \nabla \mathbf{g}_{I_0}(\mathbf{x}_0) \in \mathbb{R}^{n\times|I_0|}.$$

Der Gradient einer vektorwertigen Funktion $\mathbf{g}(\mathbf{x})$ ist wie folgt definiert:

$$\nabla \mathbf{g}(\mathbf{x}) = \begin{pmatrix} (g_1(\mathbf{x}))_{x_1} & \cdots & (g_m(\mathbf{x}))_{x_1} \\ \cdots & \cdots & \cdots \\ (g_1(\mathbf{x}))_{x_n} & \cdots & (g_m(\mathbf{x}))_{x_n} \end{pmatrix} \in \mathbb{R}^{n\times m}.$$

Ausgehend von der Anschauung, könnte man die Menge

$$\left\{\mathbf{x} \ : \ \left(\nabla g_{I_0}(\mathbf{x}_0)\right)^T (\mathbf{x}-\mathbf{x}_0) \leq \mathbf{0},\ \mathbf{g}_{I\setminus I_0}(\mathbf{x}_0) + \left(\nabla \mathbf{g}_{I\setminus I_0}(\mathbf{x}_0)\right)^T (\mathbf{x}-\mathbf{x}_0) \leq \mathbf{0}\right\}$$

als eine lineare Approximation der Menge Ω im Punkt $\mathbf{x}_0$ ansehen. Dies ist jedoch in gewissen ausgearteten Punkten nicht zutreffend, siehe Beispiel 4.25 □

Definition 4.22 Linearisierter Kegel. Die Menge

$$K(\mathbf{x}_0) := \{\mathbf{y} \in \mathbb{R}^n \ : \ \left(\nabla \mathbf{g}_{I_0}(\mathbf{x}_0)\right)^T \mathbf{y} \leq \mathbf{0}\}$$

heißt linearisierter Kegel von Ω im Punkt $\mathbf{x}_0 \in \Omega$. Für $I_0 = \emptyset$ setzen wir $K(\mathbf{x}_0) = \mathbb{R}^n$. □

Lemma 4.23 *Es gilt* $T(\mathbf{x}_0) \subseteq K(\mathbf{x}_0)$.

Beweis: Für $I_0 = \emptyset$ gilt die Behauptung trivialerweise. Sei also $I_0 \neq \emptyset$. Der Nullvektor gehört per Definition zu beiden Mengen. Sei $\mathbf{y} \in T(\mathbf{x}_0)$, $\mathbf{y} \neq \mathbf{0}$. Dann existieren ein $\mathbf{y}^* \in T(\mathbf{x}_0)$, $\|\mathbf{y}^*\|_2 = 1$ und ein $\lambda > 0$ mit $\mathbf{y} = \lambda \mathbf{y}^*$ und eine Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}} \subset \Omega$ mit $\mathbf{x}_k \xrightarrow{\mathbf{y}^*} \mathbf{x}_0$. Wegen $g_i(\mathbf{x}_k) \leq 0$ für alle i und $g_i(\mathbf{x}_0) = 0$ für alle $i \in I_0$ hat man

$$0 \geq \lim_{k\to\infty} \frac{g_i(\mathbf{x}_k) - g_i(\mathbf{x}_0)}{\|\mathbf{x}_k - \mathbf{x}_0\|_2} = \nabla g_i(\mathbf{x}_0)^T \mathbf{y}^* \quad \forall i \in I_0,$$

wobei Lemma 4.9, 1) verwendet wurde. Folglich gilt $\mathbf{y} \in K(\mathbf{x}_0)$. ■

Lemma 4.24 *Es gilt* $K_0(\mathbf{x}_0) := \{\mathbf{y} \in \mathbb{R}^n \ : \ \left(\nabla \mathbf{g}_{I_0}(\mathbf{x}_0)\right)^T \mathbf{y} < \mathbf{0}\} \subseteq T(\mathbf{x}_0)$.

Beweis: Siehe Literatur, zum Beispiel [ERSD77, S. 145]. ■

Beispiel 4.25 Tangentenkegel und linearisierter Kegel. Sei

$$\Omega = \{\mathbf{x} \in \mathbb{R}^2 \ : \ -x_1^3 + x_2 \leq 0, -x_2 \leq 0\},$$

das heißt die Menge Ω ist begrenzt von der positiven x_1–Achse und der Funktion x_1^3. Damit sind

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} -x_1^3 + x_2 \\ -x_2 \end{pmatrix}, \quad (\nabla \mathbf{g}(\mathbf{x}))^T = \begin{pmatrix} -3x_1^2 & 1 \\ 0 & -1 \end{pmatrix}.$$

Wir betrachten den Punkt $\mathbf{x}_0 = (0,0)^T$. In diesem Punkt sind beide Nebenbedingungen mit Gleichheit erfüllt, also $I_0 = \{1,2\}$. Es folgt

$$\left(\nabla \mathbf{g}_{I_0}(\mathbf{x}_0)\right)^T = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \quad \Longrightarrow \quad K(\mathbf{x}_0) = \begin{pmatrix} y_1 \\ 0 \end{pmatrix},\ y_1 \in \mathbb{R}.$$

Für den Tangentialkegel gilt $T(\mathbf{x}_0) = \{\mathbf{y} \in \mathbb{R}^2 \ : \ y_1 \geq 0, y_2 = 0\}$. Es ist also $T(\mathbf{x}_0) \subsetneq K(\mathbf{x}_0)$. Dieses Beispiel zeigt insbesondere, dass man für den Punkt $\mathbf{x}_0 = (0,0)^T$ mit $K(\mathbf{x}_0)$ keine lineare Approximation von Ω erhält. Der Grund ist die flache Spitze in $\mathbf{x}_0$. Falls die Spitze in $\mathbf{x}_0$ weniger flach wäre, wäre eine lineare Approximation möglich. □

Beispiel 4.26 Der linearisierte Kegel $K(\mathbf{x}_0)$ ist im Gegensatz zum Tangentenkegel $T(\mathbf{x}_0)$ von der analytischen Darstellung von Ω abhängig. Die Menge aus Beispiel 4.25 kann auch dargestellt werden durch

$$\Omega = \{\mathbf{x} \in \mathbb{R}^2 \ : \ -x_1^3 + x_2 \leq 0, -x_2^3 \leq 0\}$$

In diesem Fall hat man

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} -x_1^3 + x_2 \\ -x_2^3 \end{pmatrix} \quad (\nabla \mathbf{g}(\mathbf{x}))^T = \begin{pmatrix} -3x_1^2 & 1 \\ 0 & -3x_2^2 \end{pmatrix}$$

und man erhält für $\mathbf{x}_0 = (0,0)^T$, dass $I_0 = \{1,2\}$ und

$$(\nabla \mathbf{g}_{I_0}(\mathbf{x}_0))^T = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \Longrightarrow \quad K(\mathbf{x}_0) = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

mit $y_1 \in \mathbb{R}$ und $y_2 \leq 0$. □

Für den nächsten Beweis benötigen wir einen Hilfssatz.

Lemma 4.27 Alternativsatz von Gordan. *Sei $A \in \mathbb{R}^{m \times n}$, dann hat von den beiden Systemen*

$$A\mathbf{y} < \mathbf{0}$$

und

$$\mathbf{z}^T A = \mathbf{0},\ \mathbf{z} \geq \mathbf{0},\ \mathbf{z} \neq \mathbf{0},$$

genau eines eine Lösung.

Beweis: Literatur. ■

Bemerkung 4.28 Seien die Voraussetzungen von Satz 4.14 erfüllt und $I_0 \neq \emptyset$. Dann folgt wegen $K_0(\mathbf{x}_0) \subseteq T(\mathbf{x}_0)$

$$(\nabla \mathbf{g}_{I_0}(\mathbf{x}_0))^T \mathbf{y} < \mathbf{0} \quad \Longrightarrow \quad \mathbf{y}^T \nabla f(\mathbf{x}_0) \geq 0 \quad \forall\ \mathbf{y} \in K_0(\mathbf{x}_0).$$

Diese Beziehung ist äquivalent damit, dass das Ungleichungssystem

$$(\nabla \mathbf{g}_{I_0}(\mathbf{x}_0))^T \mathbf{y} < \mathbf{0}, \quad \mathbf{y}^T \nabla f(\mathbf{x}_0) < 0$$

keine Lösung $\mathbf{y} \in K_0(\mathbf{x}_0)$ besitzt. Nach Lemma 4.27 besitzt dieses System genau dann keine Lösung, wenn das System

$$\eta \nabla f(\mathbf{x}_0) + \nabla \mathbf{g}_{I_0}(\mathbf{x}_0) \mathbf{z}_{I_0} = \mathbf{0}, \quad \begin{pmatrix} \eta \\ \mathbf{z}_{I_0} \end{pmatrix} \geq \mathbf{0}, \quad \eta \in \mathbb{R}_+, \tag{4.5}$$

eine nichttriviale Lösung besitzt. Hierbei ist $\mathbf{z}_{I_0}$ ein $|I_0|$–dimensionaler Vektor, dessen Einträge durch die Indizes der aktiven Nebenbedingungen (unter Beibehaltung der natürlichen Reihenfolge) bestimmt sind. Setzt man $z_i = 0$ für $i \notin I_0$, so kann man (4.5) mit $\mathbf{z} \in \mathbb{R}^m$ wie folgt formulieren:

$$\eta \nabla f(\mathbf{x}_0) + \nabla \mathbf{g}(\mathbf{x}_0) \mathbf{z} = \mathbf{0},\ \mathbf{g}(\mathbf{x}_0) \leq \mathbf{0},\ \underbrace{\mathbf{z}^T}_{=0, i \notin I_0}\ \underbrace{\mathbf{g}(\mathbf{x}_0)}_{=0, i \in I_0} = 0, \begin{pmatrix} \eta \\ \mathbf{z} \end{pmatrix} \geq \mathbf{0},\ \begin{pmatrix} \eta \\ \mathbf{z} \end{pmatrix} \neq \mathbf{0}.$$

Sei $I_0 = \emptyset$ und hat $f(\mathbf{x})$ in $\mathbf{x}_0 \in \Omega$ ein lokales Minimum, das heißt es gelten $\mathbf{x}_0 \in \text{int}(\Omega)$, $\mathbf{z} = \mathbf{0}$ und $\nabla f(\mathbf{x}_0) = \mathbf{0}$, dann besitzt dieses System offensichtlich eine Lösung $\eta > 0, \mathbf{z} = \mathbf{0}$. Damit hat man auch für den Fall $I_0 = \emptyset$ ein sinnvolles Problem definiert. □

Bemerkung 4.29 Ziel. Wir betrachten jetzt das Problem: Gesucht ist ein Tripel $(\mathbf{x}_0, \mathbf{z}_0, \eta_0)^T \in \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}_+ (\geq 0)$, $(\eta_0, \mathbf{z}_0)^T \neq \mathbf{0}$, welches das System

$$\eta \nabla f(\mathbf{x}) + \nabla \mathbf{g}(\mathbf{x}) \mathbf{z} = \mathbf{0},\ \mathbf{g}(\mathbf{x}) \leq \mathbf{0},\ \mathbf{z}^T \mathbf{g}(\mathbf{x}) = 0, \begin{pmatrix} \eta \\ \mathbf{z} \end{pmatrix} \geq \mathbf{0} \tag{4.6}$$

löst. Mit dem Lagrange–Funktional von (4.4)

$$L_\eta(\mathbf{x}, \mathbf{z}) = \eta f(\mathbf{x}) + \mathbf{z}^T \mathbf{g}(\mathbf{x})$$

ergibt sich die äquivalente Formulierung

$$\nabla_{\mathbf{x}} L_\eta(\mathbf{x}, \mathbf{z}) = \mathbf{0},\ \nabla_{\mathbf{z}} L_\eta(\mathbf{x}, \mathbf{z}) \leq \mathbf{0},\ \mathbf{z}^T \nabla_{\mathbf{z}} L_\eta(\mathbf{x}, \mathbf{z}) = 0,\ \begin{pmatrix} \eta \\ \mathbf{z} \end{pmatrix} \geq \mathbf{0}. \tag{4.7}$$

□

Der Zusammenhang zwischen den Problemen (4.4) und (4.6) ist wie folgt.

Satz 4.30 Optimalitätskriterium für den Fall $\eta_0 \geq 0$. *Seien $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ differenzierbar und $\mathbf{x}_0 \in \Omega$ Stelle eines lokalen Minimums von $f(\mathbf{x})$ bezüglich Ω. Dann existieren ein $\mathbf{z}_0 \in \mathbb{R}^m_+$ und ein $\eta_0 \geq 0$, so dass $(\mathbf{x}_0, \mathbf{z}_0)^T$ und η_0 eine Lösung von (4.6) beziehungsweise von (4.7) sind.*

Beweis: Literatur, Fritz John (1948). ■

Bemerkung 4.31 Regularitätsbedingung. Damit haben wir für den Fall $\eta_0 > 0$ ein Optimalitätskriterium gefunden. Ist jedoch $\eta_0 = 0$, so kann der Funktionswert von $f(\mathbf{x})$ in (4.6) beliebig sein. Es wird jetzt eine Regularitätsbedingung eingeführt, die sichert, dass unter den Voraussetzungen von Satz 4.30 für jede Lösung von (4.6) $\eta_0 > 0$ gilt.

$$\text{Regularitätsbedingung:} \qquad K(\mathbf{x}_0) = T(\mathbf{x}_0). \tag{4.8}$$

□

Bemerkung 4.32 Zur Regularitätsbedingung.

- Für Optimierungsprobleme mit ausschließlich affinen Nebenbedingungen ist die Regularitätsbedingung stets erfüllt.
- Die Regularitätsbedingung stellt eine Bedingung an die analytische Darstellung der Menge Ω dar, vergleiche Beispiele 4.25 und 4.26.
- Man kann andere Regularitätsbedingungen formulieren, die die obige Bedingung implizieren, aber leichter zu überprüfen sind.

□

Bemerkung 4.33 Anstelle von (4.6) betrachten wir jetzt das Optimierungsproblem: Finde ein Paar $(\mathbf{x}_0, \mathbf{z}_0)^T \in \mathbb{R}^n \times \mathbb{R}^m_+$, welches das Problem

$$\nabla f(\mathbf{x}) + \nabla \mathbf{g}(\mathbf{x}) \mathbf{z} = \mathbf{0},\ \mathbf{g}(\mathbf{x}) \leq \mathbf{0},\ \mathbf{z}^T \mathbf{g}(\mathbf{x}) = 0, \mathbf{z} \geq \mathbf{0} \tag{4.9}$$

löst. Man bezeichnet (4.9) als lokale Kuhn–Tucker–Bedingung (Kuhn, Tucker (1951)). Unter Verwendung des Lagrange–Funktionals

$$L(\mathbf{x}, \mathbf{z}) := f(\mathbf{x}) + \mathbf{z}^T \mathbf{g}(\mathbf{x}), \quad (\mathbf{x}, \mathbf{z})^T \in \mathbb{R}^n \times \mathbb{R}^m_+$$

erhält man eine zu (4.9) äquivalente Formulierung

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{z}) = \mathbf{0},\ \nabla_{\mathbf{z}} L(\mathbf{x}, \mathbf{z}) \leq \mathbf{0},\ \mathbf{z}^T \nabla_{\mathbf{z}} L(\mathbf{x}, \mathbf{z}) = 0,\ \mathbf{z} \geq \mathbf{0}. \tag{4.10}$$

Die Optimierungsprobleme (4.9), (4.10) hat man aus (4.6), (4.7) durch die Festsetzung von $\eta = 1$ erhalten. □

Lemma 4.34 Satz von Farkas. *Seien $A \in \mathbb{R}^{m\times n}$ und $\mathbf{b} \in \mathbb{R}^n$. Dann hat von den beiden Systemen*

$$A\mathbf{x} \leq \mathbf{0}, \quad \mathbf{b}^T\mathbf{x} > 0,$$

und

$$\mathbf{y}^T A = \mathbf{b}^T, \quad \mathbf{y} \geq \mathbf{0}$$

genau eines eine Lösung.

Beweis: Siehe Literatur. ■

Satz 4.35 Satz von Kuhn/Tucker. *Seien $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ stetig differenzierbar und $\mathbf{x}_0 \in \Omega$ Stelle eines lokalen Minimums von $f(\mathbf{x})$ bezüglich Ω. Ist in $\mathbf{x}_0$ die Regularitätsbedingung (4.8) erfüllt, dann existiert ein $\mathbf{z}_0 \in \mathbb{R}^m_+$, so dass $(\mathbf{x}_0, \mathbf{z}_0)^T$ eine Lösung von (4.9) beziehungsweise von (4.10) ist.*

Beweis: Sei $I_0 \neq \emptyset$. Per Definition des linearisierten Kegels gilt

$$(\nabla \mathbf{g}_{I_0}(\mathbf{x}_0))^T \mathbf{y} \leq \mathbf{0} \quad \forall\, \mathbf{y} \in K(\mathbf{x}_0). \tag{4.11}$$

Da (4.8) erfüllt ist, folgt aus Satz 4.14 für $I_0 \neq \emptyset$

$$(\nabla f(\mathbf{x}_0))^T \mathbf{y} \geq 0 \iff (-\nabla f(\mathbf{x}_0))^T \mathbf{y} \leq 0 \quad \forall\, \mathbf{y} \in T(\mathbf{x}_0) = K(\mathbf{x}_0). \tag{4.12}$$

Aus Lemma 4.34 folgt, dass (4.11) und (4.12) genau dann gleichzeitig gelten, wenn das System

$$\nabla \mathbf{g}_{I_0}(\mathbf{x}_0)\mathbf{z}_{I_0} = -\nabla f(\mathbf{x}_0), \quad \mathbf{z}_{I_0} \geq \mathbf{0}$$

eine Lösung besitzt. Setzt man $z_i = 0$ für $i \notin I_0$, so folgt die erste Gleichung von (4.9) für $I_0 \neq \emptyset$. Die Beziehung $\mathbf{g}(\mathbf{x}_0) \leq \mathbf{0}$ gilt per Definition des Optimierungsproblems. Aus der Konstruktion von $\mathbf{z}$ folgt schließlich $\mathbf{z}^T\mathbf{g}(\mathbf{x}_0) = 0$.

Für $I_0 = \emptyset$ ist $(\mathbf{x}_0, \mathbf{0})$ eine Lösung von (4.9). ■

Bemerkung 4.36 Im Falle $\mathbf{x}_0 \in \text{int}(\Omega)$ gilt wegen der Regularitätsbedingung (4.8) $T(\mathbf{x}_0) = K(\mathbf{x}_0) = \mathbb{R}^n$. Aus der Definition von $K(\mathbf{x}_0)$ folgt dann, dass $\nabla\mathbf{g}(\mathbf{x}_0) = \mathbf{0}$ sein muss, da die Bedingung aus Satz 4.14 für jedes $\mathbf{y} \in \mathbb{R}^n$, insbesondere für $-\mathbf{y}$ gelten muss. Aus der ersten Gleichung der Kuhn–Tucker–Bedingung (4.9) folgt dann die bekannte notwendige Bedingung für ein lokales Minimum: $\nabla f(\mathbf{x}_0) = \mathbf{0}$.

Die Gültigkeit der Implikation von (4.11) nach (4.12) kann mit Hilfe eines linearen Programms überprüft werden. Diese Implikation gilt genau dann, wenn 0 der optimale Wert des Problem

$$\begin{aligned} z = \mathbf{z}^T\nabla f(\mathbf{x}_0) \quad &\to \quad \min\ ! \\ \mathbf{z}^T\nabla \mathbf{g}_{I_0}(\mathbf{x}_0) \quad &\leq \mathbf{0} \end{aligned}$$

ist.

Sind die Zielfunktion und die Nebenbedingungen zweimal differenzierbar, kann man weitere Kriterien mit Hilfe der Hesse–Matrix formulieren. □

4.4 Globale Theorie der Lagrange–Multiplikatoren

Bemerkung 4.37 Ziel. Wir betrachten das Optimierungsproblem

$$z = \min\{f(\mathbf{x}) \ :\ \mathbf{x} \in \Omega\} \quad \text{mit } \Omega = \{\mathbf{x} \in \mathbb{R}^n \ :\ \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}. \tag{4.13}$$

Im folgenden werden Optimalitätskriterien in Form von Sattelpunktaussagen formuliert. □

Definition 4.38 Sattelpunkt. Die Funktion $L(\mathbf{x}, \boldsymbol{\lambda})$ mit $\mathbf{x} \in \Omega$, $\boldsymbol{\lambda} \in D_{\boldsymbol{\lambda}}$ besitzt in $(\mathbf{x}_0, \boldsymbol{\lambda}_0)$ einen lokalen Sattelpunkt, wenn es ein $\varepsilon > 0$ gibt, so dass

$$L(\mathbf{x}_0, \boldsymbol{\lambda}) \leq L(\mathbf{x}_0, \boldsymbol{\lambda}_0) \leq L(\mathbf{x}, \boldsymbol{\lambda}_0) \quad \forall\ (\mathbf{x}, \boldsymbol{\lambda}) \in (\Omega \times D_{\boldsymbol{\lambda}}) \cap U_\varepsilon(\mathbf{x}_0, \boldsymbol{\lambda}_0), \tag{4.14}$$

wobei $U_\varepsilon(\mathbf{x}_0, \boldsymbol{\lambda}_0)$ eine ε–Umgebung von $(\mathbf{x}_0, \boldsymbol{\lambda}_0)$ ist. Gilt (4.14) für alle $(\mathbf{x}, \boldsymbol{\lambda}) \in \Omega \times D_{\boldsymbol{\lambda}}$, so hat $L(\mathbf{x}, \boldsymbol{\lambda})$ in $(\mathbf{x}_0, \boldsymbol{\lambda}_0)$ einen globalen Sattelpunkt. □

Bemerkung 4.39 Formulierung als Sattelpunktproblem. Zur Formulierung von globalen Optimalitätskriterien wird das folgende Sattelpunktproblem betrachtet: Gesucht sind $\mathbf{x}_0 \in \Omega$ und $(\eta_0, \mathbf{y}_0)^T \in \mathbb{R}^{m+1}_+$ mit $(\eta_0, \mathbf{y}_0)^T \neq \mathbf{0}$, so dass für das Lagrange–Funktional

$$L_\eta(\mathbf{x}, \mathbf{y}) := \eta f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x}), \quad (\mathbf{x}, \mathbf{y})^T \in \Omega \times \mathbb{R}^m_+, \tag{4.15}$$

gilt

$$L_{\eta_0}(\mathbf{x}_0, \mathbf{y}) \leq L_{\eta_0}(\mathbf{x}_0, \mathbf{y}_0) \leq L_{\eta_0}(\mathbf{x}, \mathbf{y}_0), \quad \forall\ (\mathbf{x}, \mathbf{y})^T \in \Omega \times \mathbb{R}^m_+.$$

□

Als nächstes soll ein notwendiges Optimalitätskriterium für (4.13) bewiesen werden. Dazu benötigen wir folgendes Lemma.

Lemma 4.40 *Seien $\Omega \subseteq \mathbb{R}^n$ eine konvexe Menge, $f : \Omega \to \mathbb{R}$, $\mathbf{g} : \Omega \to \mathbb{R}^q$ konvexe Funktionen und $\mathbf{h} : \Omega \to \mathbb{R}^r$ affine Funktionen. Das System*

$$f(\mathbf{x}) < 0, \quad \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \quad \mathbf{h}(\mathbf{x}) \leq \mathbf{0}, \quad \mathbf{x} \in \Omega$$

besitzt genau dann keine Lösung, wenn ein Vektor $(u, \mathbf{v}, \mathbf{w})^T \in \mathbb{R}_+ \times \mathbb{R}^q_+ \times \mathbb{R}^r_+$, $(u, \mathbf{v}, \mathbf{w})^T \neq \mathbf{0}$, existiert mit

$$uf(\mathbf{x}) + \mathbf{v}^T \mathbf{g}(\mathbf{x}) + \mathbf{w}^T \mathbf{h}(\mathbf{x}) \geq 0, \quad \forall\ \mathbf{x} \in \Omega.$$

Existiert ein $\bar{\mathbf{x}} \in \operatorname{int}(\Omega)$ welches zusätzlich $\mathbf{g}(\bar{\mathbf{x}}) < \mathbf{0}$ erfüllt, dann gilt $u \neq 0$.

Beweis: Siehe Literatur. ■

Satz 4.41 Notwendiges Optimalitätskriterium. *Seien $\Omega \subseteq \mathbb{R}^n$ eine konvexe Menge, $f : \Omega \to \mathbb{R}$ und $\mathbf{g} : \Omega \to \mathbb{R}^m$ konvexe Funktionen. Ist $\mathbf{x}_0$ eine Lösung von (4.13), so existieren ein $\eta_0 \in \mathbb{R}_+$ und ein $\mathbf{y}_0 \in \mathbb{R}^m_+$, so dass $(\mathbf{x}_0, \eta_0, \mathbf{y}_0)^T \in \Omega \times \mathbb{R}^{m+1}_+$ eine Lösung des Sattelpunktproblems (4.15) ist.*

Beweis: Sei $\mathbf{x}_0$ eine Lösung von (4.13). Dann besitzt das System

$$f(\mathbf{x}) - f(\mathbf{x}_0) < 0, \quad \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \quad \mathbf{x} \in \Omega$$

keine Lösung. Nach Lemma 4.40, mit $\mathbf{h}(\mathbf{x}) \equiv \mathbf{0}$, existiert dann ein Vektor $(\eta_0, \mathbf{y}_0)^T \in \mathbb{R}^{m+1}_+$, $(\eta_0, \mathbf{y}_0)^T \neq \mathbf{0}$, mit

$$L_{\eta_0}(\mathbf{x}, \mathbf{y}_0) = \eta_0 f(\mathbf{x}) + \mathbf{y}_0^T \mathbf{g}(\mathbf{x}) \geq \eta_0 f(\mathbf{x}_0) \quad \forall\ \mathbf{x} \in \Omega.$$

Wegen $\mathbf{g}(\mathbf{x}_0) \leq \mathbf{0}$ folgt für alle $\mathbf{y} \geq \mathbf{0}$ damit

$$L_{\eta_0}(\mathbf{x}, \mathbf{y}_0) = \eta_0 f(\mathbf{x}) + \mathbf{y}_0^T \mathbf{g}(\mathbf{x}) \geq \eta_0 f(\mathbf{x}_0) + \mathbf{y}^T \mathbf{g}(\mathbf{x}_0) = L_{\eta_0}(\mathbf{x}_0, \mathbf{y}) \quad \forall\ (\mathbf{x}, \mathbf{y}) \in \Omega \times \mathbb{R}^m_+. \tag{4.16}$$

Setzt man in (4.16) rechts $\mathbf{y} = \mathbf{y}_0$, ergibt sich

$$\eta_0 f(\mathbf{x}) + \mathbf{y}_0^T \mathbf{g}(\mathbf{x}) \geq \eta_0 f(\mathbf{x}_0) + \mathbf{y}_0^T \mathbf{g}(\mathbf{x}_0) = L_{\eta_0}(\mathbf{x}_0, \mathbf{y}_0) \quad \forall\ \mathbf{x} \in \Omega.$$

Setzt man in (4.16) links $\mathbf{x} = \mathbf{x}_0$, erhält man

$$\eta_0 f(\mathbf{x}_0) + \mathbf{y}_0^T \mathbf{g}(\mathbf{x}_0) \geq \eta_0 f(\mathbf{x}_0) + \mathbf{y}^T \mathbf{g}(\mathbf{x}_0) \quad \forall\ \mathbf{y} \in \mathbb{R}^m_+.$$

Aus den letzten beiden Ungleichungen folgt die Behauptung. ■

Bemerkung 4.42 Regularitätsbedingung. Die Aussage dieses Satzes hat in gewissem Sinne Ähnlichkeit mit der von Satz 4.30. Für einen gegebenen zulässigen Bereich ist die notwendige Bedingung (nichtnegative Lösung von (4.15)) für beliebige Werte der Zielfunktion erfüllt, falls $\eta_0 = 0$ ist. Analog wie bei den lokalen Lagrange–Multiplikatoren wird deshalb eine Regularitätsbedingung eingeführt, die $\eta_0 > 0$ sichert.

Regularitätsbedingung. Seien die Menge $\Omega \subseteq \mathbb{R}^n$ sowie die Funktionen $g_i : \Omega \to \mathbb{R}$, $i \in \{1, \ldots, m\} =: I$, konvex. Es existiere ein $\bar{\mathbf{x}} \in \text{int}(\Omega)$ mit

$$g_i(\bar{\mathbf{x}}) < 0 \;\text{ für } i \in I_N, \quad g_i(\bar{\mathbf{x}}) \leq 0 \;\text{ für } i \in I_L. \tag{4.17}$$

Hierbei ist I_L die Menge aller $i \in \{1, \ldots, m\}$, für die $g_i(\mathbf{x})$ eine affine Funktion ist und I_N die Menge aller Indizes, für die $g_i(\mathbf{x})$ keine affine Funktion ist.

Unter dieser Regularitätsbedingung wird das Sattelpunktproblem: Finde einen Sattelpunkt $(\mathbf{x}_0, \mathbf{y}_0) \in \Omega \times \mathbb{R}^m_+$ des Lagrange–Funktional

$$L(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x}), \quad (\mathbf{x}, \mathbf{y})^T \in \Omega \times \mathbb{R}^m_+, \tag{4.18}$$

betrachtet. □

Satz 4.43 Existenz eines Sattelpunktes. *Seien die Menge $\Omega \subseteq \mathbb{R}^n$ konvex, die Funktionen $f : \Omega \to \mathbb{R}$, $g_i : \Omega \to \mathbb{R}$, $i \in I$, konvex und die Regularitätsbedingung (4.17) erfüllt. Ist $\mathbf{x}_0 \in \Omega$ eine Lösung von (4.13), dann existiert ein $\mathbf{y}_0 \in \mathbb{R}^m_+$, so dass $(\mathbf{x}_0, \mathbf{y}_0)$ ein Sattelpunkt von (4.18) ist.*

Beweis: Da $\mathbf{x}_0$ eine Lösung von (4.13) ist, ist das System

$$f(\mathbf{x}) - f(\mathbf{x}_0) < 0, \quad \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \quad \mathbf{x} \in \Omega$$

nicht lösbar. Wegen der Regularitätsbedingung (4.17) besitzt jedoch das System

$$\mathbf{g}_{I_N}(\mathbf{x}) < \mathbf{0}, \; \mathbf{g}_{I_L}(\mathbf{x}) \leq \mathbf{0}, \quad \mathbf{x} \in \text{int}(\Omega)$$

eine Lösung. Nach Lemma 4.40 existiert ein $(\eta_0, \mathbf{y}_0)^T \in \mathbb{R}^{m+1}_+$, $\eta_0 \neq 0$, (das $\mathbf{g}_{I_N}(\mathbf{x})$ spielt die Rolle von $\mathbf{g}(\mathbf{x})$ aus Lemma 4.40) mit

$$\eta_0 \left(f(\mathbf{x}) - f(\mathbf{x}_0)\right) + \mathbf{y}_0^T \mathbf{g}(\mathbf{x}) \geq 0 \quad \forall\, \mathbf{x} \in \Omega.$$

Bei dieser Darstellung wurden die Vektoren $\mathbf{v}$ und $\mathbf{w}$ aus Lemma 4.40 zum Vektor $\mathbf{y}_0$ zusammengefasst. Man kann so skalieren, dass $\eta_0 = 1$ gilt. Dann folgt

$$L(\mathbf{x}, \mathbf{y}_0) = f(\mathbf{x}) + \mathbf{y}_0^T \mathbf{g}(\mathbf{x}) \geq f(\mathbf{x}_0) \quad \forall\, \mathbf{x} \in \Omega.$$

Wegen $\mathbf{g}(\mathbf{x}_0) \leq \mathbf{0}$ folgt für alle $\mathbf{y} \geq \mathbf{0}$ gilt

$$L(\mathbf{x}_0, \mathbf{y}) = f(\mathbf{x}_0) + \mathbf{y}^T \mathbf{g}(\mathbf{x}_0) \leq f(\mathbf{x}_0) \quad \forall\, \mathbf{y} \in \mathbb{R}^m_+.$$

Aus den letzten beiden Beziehungen folgt, indem man $\mathbf{x} = \mathbf{x}_0$ beziehungsweise $\mathbf{y} = \mathbf{y}_0$ setzt,

$$L(\mathbf{x}_0, \mathbf{y}_0) = f(\mathbf{x}_0).$$

Nach Definition ist $(\mathbf{x}_0, \mathbf{y}_0)^T$ somit ein Sattelpunkt des Lagrange–Funktionals (4.18). ■

Nun wird eine hinreichende Bedingung für eine Lösung von (4.13) formuliert. Dabei gilt eine gewisse Umkehrung von Satz 4.43, die ohne weitere Voraussetzungen an (4.13) gilt.

Satz 4.44 Hinreichendes Optimalitätskriterium. *Sei $(\mathbf{x}_0, \mathbf{y}_0)^T \in \Omega \times \mathbb{R}^m_+$ ein Sattelpunkt von (4.18). Dann ist $\mathbf{x}_0$ eine Lösung von (4.13).*

Beweis: Zunächst wird gezeigt, dass $\mathbf{x}_0$ die Nebenbedingungen erfüllt. Aus $L(\mathbf{x}_0, \mathbf{y}) \leq L(\mathbf{x}_0, \mathbf{y}_0)$ für alle $\mathbf{y} \in \mathbb{R}^m_+$ folgt

$$(\mathbf{y} - \mathbf{y}_0)^T \mathbf{g}(\mathbf{x}_0) \leq 0 \quad \forall\, \mathbf{y} \in \mathbb{R}^m_+. \tag{4.19}$$

Es gilt $\mathbb{R}^m_+ \subset \{\mathbf{y} - \mathbf{y}_0 \ : \ \mathbf{y}, \mathbf{y}_0 \in \mathbb{R}^m_+\}$. Somit gilt

$$\mathbf{y}^T \mathbf{g}(\mathbf{x}_0) \leq 0 \quad \forall\, \mathbf{y} \in \mathbb{R}^m_+.$$

Sei $g_i(\mathbf{x}_0) > 0$. Dann würde diese Beziehung nicht für den i–ten Einheitsvektor gelten, der aber zu $\mathbb{R}^m_+$ gehört. Damit folgt $\mathbf{g}(\mathbf{x}_0) \leq \mathbf{0}$.

Nun wird gezeigt, dass $f(\mathbf{x})$ in $\mathbf{x}_0$ ein globales Minimum annimmt. Aus (4.19) folgt für $\mathbf{y} = \mathbf{0}$ die Ungleichung $\mathbf{y}_0^T \mathbf{g}(\mathbf{x}_0) \geq 0$. Da $\mathbf{y}_0 \geq \mathbf{0}$ und $\mathbf{g}(\mathbf{x}_0) \leq \mathbf{0}$, kann das Skalarprodukt nur nichtpositiv sein, also gilt $\mathbf{y}_0^T \mathbf{g}(\mathbf{x}_0) = 0$. Mit dieser Beziehung und $L(\mathbf{x}_0, \mathbf{y}_0) \leq L(\mathbf{x}, \mathbf{y}_0)$ für alle $\mathbf{x} \in \Omega$ folgt

$$f(\mathbf{x}_0) \leq f(\mathbf{x}) + \mathbf{y}_0^T \mathbf{g}(\mathbf{x}) \quad \forall\, \mathbf{x} \in \Omega$$

und damit für alle $\mathbf{x} \in \Omega$

$$f(\mathbf{x}_0) \leq f(\mathbf{x}) + \underbrace{\mathbf{y}_0^T}_{\geq \mathbf{0}} \underbrace{\mathbf{g}(\mathbf{x})}_{\leq \mathbf{0}} \leq f(\mathbf{x}).$$

■

Bemerkung 4.45 Dualitätstheorie. Auch für die nichtlineare Optimierung gibt es eine Dualitätstheorie. Die Dualitätstheorie für lineare Programme ist darin als Spezialfall enthalten.

Gegeben sei das primale Problem

$$z = \min\{f(\mathbf{x}) \ : \ \mathbf{x} \in \Omega\} \quad \text{ mit } \Omega = \{\mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}, \tag{4.20}$$

wobei $\Omega \subseteq \mathbb{R}^n$ eine nichtleere Menge ist und $f \ : \ \Omega \to \mathbb{R}$, $\mathbf{g} \ : \ \Omega \to \mathbb{R}^m$ Abbildungen sind. Dem Problem (4.20) wird unter Verwendung des Lagrange–Funktionals das duale Problem

$$\tilde{z} = \max\{\phi(\mathbf{y}) \ : \ \mathbf{y} \in \mathbb{R}^m_+\}$$

zugeordnet, wobei

$$\phi(\mathbf{y}) \ : \ \mathbb{R}^m_+ \to \mathbb{R} \cup \{-\infty\}, \quad \phi(\mathbf{y}) = \inf_{\mathbf{x} \in \Omega} L(\mathbf{x}, \mathbf{y}) = \inf_{\mathbf{x} \in \Omega} \left(f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x}) \right).$$

Man sieht, dass hier die Nichtnegativitätsbedingung im dualen Problem steckt, anders als wir es bei linearen Programmen hatten. Da wir dort jedoch gezeigt hatten, dass das duale Problem des dualen Problems wieder das primale Problem ist, ist das kein Widerspruch dazu, dass die Theorie für lineare Programme als Spezialfall enthalten ist.

Man kann wieder schnell zeigen, dass der Maximalwert des dualen Problems, falls er existiert, kleiner oder gleich dem Minimalwert des primalen Problems ist. Gleichheit muss im allgemeinen jedoch nicht gelten. Man spricht dann von einer Dualitätslücke. Des weiteren sind die Fragen bezüglich der Lösbarkeit der beiden Probleme wesentlich komplizierter zu beantworten als bei linearen Programmen. □

Kapitel 5

Lösungsverfahren

Dieses Kapitel gibt einen Überblick über Lösungsverfahren für nichtlineare Optimierungsprobleme. Es wird vor allem auf die wesentlichen Ideen der Verfahren eingegangen und weniger auf Details.

5.1 Projektionsverfahren

Projektionsverfahren sind recht einfache Verfahren, die das Konzept von Abstiegsverfahren zur Minimierung von Funktionen ohne Nebenbedingungen auf Minimierungsprobleme mit konvexen Nebenbedingungen übertragen. Bei einem Abstiegsverfahren wird, ausgehend von einer Iterierten $\mathbf{x}^{(k)}$, die nächste Iterierte $\mathbf{x}^{(k+1)}$ so gewählt, dass sich der Wert der zu minimierenden Funktion verkleinert. Hat man ein Problem mit Nebenbedingungen, so muss man natürlich zusätzlich darauf achten, dass $\mathbf{x}^{(k+1)}$ zum zulässigen Bereich gehört. Anderenfalls kann es zum Beispiel vorkommen, dass die Zielfunktion gar nicht definiert ist. Projektionsverfahren projizieren die Abstiegsrichtung für die Zielfunktion in geeigneter Weise in die zulässige Menge.

Wir betrachten das Optimierungsproblem

$$z = \min\{f(\mathbf{x}) \ : \ \mathbf{x} \in \Omega\} \tag{5.1}$$

mit $f \in C^1(\mathbb{R}^n)$ und die zulässige Menge Ω sei nichtleer, konvex und abgeschlossen.

Von besonderer Bedeutung sind die Probleme, bei denen $f(\mathbf{x})$ quadratisch und

$$\Omega = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \leq \mathbf{b}\} \tag{5.2}$$

ein Polyeder ist, $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$. Solche Probleme treten als Teilprobleme bei den sogenannten SQP–Verfahren (sequential quadratic programming) auf, wo sie wiederholt mit unterschiedlichen Daten $A, \mathbf{b}, f$ gelöst werden müssen, siehe Abschnitt 5.4.

In anderen Anwendungen ist der zulässige Bereich sogar nur ein n–dimensionaler Quader („box constraints"):

$$\Omega = \times_{i=1}^{n}[l_i, u_i]. \tag{5.3}$$

Für $\mathbf{x} \in \mathbb{R}^n$ sei $P_\Omega(\mathbf{x})$ die Lösung von

$$\inf_{\mathbf{y}\in\Omega} \|\mathbf{y} - \mathbf{x}\|_2 ,$$

die Projektion von $\mathbf{x}$ auf Ω bezüglich der Euklidischen Norm. Da die Menge Ω nach Voraussetzung abgeschlossen ist, kann man das Infimum durch das Minimum ersetzen, also

$$P_\Omega(\mathbf{x}) = \arg\min_{\mathbf{y}\in\Omega} \|\mathbf{y} - \mathbf{x}\|_2 .$$

Beispiel 5.1 Falls Ω durch (5.2) gegeben ist, muss man zur Berechnung der Projektion $\overline{\mathbf{y}} = P_\Omega(\mathbf{x})$ das konvexe quadratische Programm

$$\overline{\mathbf{y}} = \arg\min_{\mathbf{y}\in\Omega}\{\|\mathbf{y}-\mathbf{x}\|_2^2 \ : \ A\mathbf{y} \le \mathbf{b}\} = \arg\min_{\mathbf{y}\in\Omega}\{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{x} \ : \ A\mathbf{y} \le \mathbf{b}\}$$

lösen. Die Menge Ω ist ein konvexes Polyeder. Die Zielfunktion $\|\mathbf{y}-\mathbf{x}\|_2^2$ ist eine konvexe Funktion.

Ist der zulässige Bereich ein n–dimensionaler Quader (5.3), kann man die Projektion komponentenweise berechnen:

$$(\overline{\mathbf{y}})_i = \begin{cases} x_i & \text{falls } x_i \in [l_i, u_i], \\ u_i & \text{falls } x_i > u_i, \\ l_i & \text{falls } x_i < l_i. \end{cases}$$

□

Definition 5.2 Stationärer Punkt. Der Punkt $\mathbf{x}_0$ wird stationärer Punkt des Problems (5.1) genannt, falls für alle Punkte des Tangentenkegels $\mathbf{y} \in T(\mathbf{x}_0)$ die Ungleichung

$$\mathbf{y}^T \nabla f(\mathbf{x}_0) \ge 0$$

gilt. □

Ein stationärer Punkt erfüllt also die im Satz 4.14 bewiesene notwendige Bedingung für ein lokales Minimum bezüglich Ω.

Algorithmus 5.3 Projektionsverfahren.

1. *Initialisierung.*
   Bestimme $\mathbf{x}^{(0)} \in \Omega$ und wähle drei reelle Parameter $0 < \beta$, $\mu < 1$ und $\gamma > 0$.
2. *Iteration.* $\mathtt{k} = 0, 1, 2, \ldots$
   Falls $\mathbf{x}^{(\mathtt{k})}$ ein stationärer Punkt ist, dann
   Stopp
   sonst
   Betrachte für $\alpha > 0$ den Pfad

$$\mathbf{x}^{(\mathtt{k})}(\alpha) := \mathtt{P}_\Omega\left(\mathbf{x}^{(\mathtt{k})} + \alpha\nabla\mathtt{f}\left(\mathbf{x}^{(\mathtt{k})}\right)\right)$$

   Setze

$$\mathbf{x}^{(\mathtt{k}+1)}(\alpha) := \mathbf{x}^{(\mathtt{k})}\left(\alpha^{(\mathtt{k})}\right),$$

   wobei $\alpha^{(k)} = \gamma\beta^{m^{(k)}}$ und $m^{(k)}$ die kleinste natürliche Zahl größer oder gleich Null ist mit

$$\mathtt{f}\left(\mathbf{x}^{(\mathtt{k}+1)}\right) \le \mathtt{f}\left(\mathbf{x}^{(\mathtt{k})}\right) + \mu\nabla\mathtt{f}\left(\mathbf{x}^{(\mathtt{k})}\right)^{\mathtt{T}}\left(\mathbf{x}^{(\mathtt{k}+1)} - \mathbf{x}^{(\mathtt{k})}\right)$$

   (Armijo–Liniensuche)

□

Im Algorithmus 5.3 hat man einen gekrümmten Pfad $\mathbf{x}^{(k)}(\alpha)$. Für jedes α hat man im allgemeinen ein konvexes, quadratisches Minimierungsproblem über Ω zu lösen. Das ist recht teuer.

Falls Ω ein Polyeder ist, kann man eine Startiterierte $\mathbf{x}^{(0)}$ mittels eines linearen Programms bestimmen.

Da der Algorithmus 5.3 im wesentlichen ein Gradientenverfahren ist, kann man im allgemeinen nur langsame Konvergenz erwarten. Außerdem ist die Berechnung von $P_\Omega\left(\mathbf{x}^{(k)} + \alpha\nabla f\left(\mathbf{x}^{(k)}\right)\right)$ für allgemeine Polyeder aufwendig. Wir betrachten jetzt noch eine Variante von Algorithmus 5.3, die zusätzlich leichter berechenbare Zwischenwerte $\mathbf{x}^{(k)} \in \Omega$ berechnet, bei denen der Funktionswert zumindest nicht ansteigt.

Algorithmus 5.4 Modifiziertes Projektionsverfahren.

1. *Initialisierung.*
   `Bestimme` $\mathbf{x}^{(0)} \in \Omega$ `und wähle Parameter falls nötig.`
2. *Iteration.* $\mathtt{k} = 0, 1, 2, \ldots$
   `Bestimme` $\mathbf{x}^{(\mathtt{k}+1)} = P_\Omega\left(\mathbf{x}^{(\mathtt{k})} + \alpha\nabla \mathtt{f}(\mathbf{x}^{(\mathtt{k})})\right)$ `entweder wie im Algorithmus 5.3 oder`
   `bestimme` $\mathbf{x}^{(\mathtt{k}+1)} \in \Omega$`, so dass` $\mathtt{f}\left(\mathbf{x}^{(\mathtt{k}+1)}\right) \leq \mathtt{f}\left(\mathbf{x}^{(\mathtt{k})}\right)$.

□

Die Umsetzung der zweiten Strategie hängt von der Art der Nebenbedingungen ab. Wir betrachten affine Nebenbedingungen (5.2). Bezeichne $\hat{\mathbf{a}}_i$ die Zeilen von A. Für $\mathbf{x} \in \Omega$ sei die Menge der aktiven Nebenbedingungen

$$I(\mathbf{x}) = \left\{i \in \{1, \ldots, m\} \ : \ \hat{\mathbf{a}}_i^T\mathbf{x} = b_i\right\}.$$

Dann wird die zweite Strategie häufig so realisiert, dass $I(\mathbf{x}^{(k)}) \subseteq I(\mathbf{x}^{(k+1)})$ gilt. Man wählt dazu in $\mathbf{x}^{(k)}$ eine Abstiegsrichtung

$$\mathbf{v}^{(k)} \in L\left(\mathbf{x}^{(k)}\right) := \{\mathbf{v} \ : \ A_{I(\mathbf{x}^{(k)})}\mathbf{v} = \mathbf{0}\},$$

wobei $A_{I(\mathbf{x}^{(k)})}$ die Teilmatrix von A mit den Zeilen ist, deren Indizes in $I(\mathbf{x}^{(k)})$ enthalten sind. Die Abstiegseigenschaft wird dann nur in schwacher Form

$$\nabla f\left(\mathbf{x}^{(k)}\right)^T \mathbf{v}^{(k)} \leq 0$$

verlangt. Die neue Iterierte besitzt die Gestalt $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha\mathbf{v}^{(k)}$ für ein geeignetes α. Aus der Wahl von $\mathbf{v}^{(k)}$ folgt

$$A\left(\mathbf{x}^{(k+1)}\right) = A\left(\mathbf{x}^{(k)} + \alpha\mathbf{v}^{(k)}\right) = A\mathbf{x}^{(k)} + \alpha A\mathbf{v}^{(k)}.$$

Für die aktiven Nebenbedingungen verschwindet der zweite Summand. Für die anderen Nebenbedingungen, $i \notin I(\mathbf{x}^{(k)})$, ist der erste Summand kleiner als b_i und man findet ein hinreichend kleines $\alpha > 0$, so dass die Summe der beiden Summanden kleiner oder gleich b_i bleibt. Mit

$$\overline{\alpha}^{(k)} := \sup\{\alpha \ : \ \mathbf{x}^{(k)} + \alpha\mathbf{v}^{(k)} \in \Omega\}$$

wird nun eine Liniensuche gestartet um einen Faktor $\alpha^{(k)}$ und damit ein Argument $\mathbf{x}^{(k+1)}$ zu finden, so dass $f\left(\mathbf{x}^{(k+1)}\right) \leq f\left(\mathbf{x}^{(k)}\right)$ gilt. Ist $\alpha^{(k)} < \overline{\alpha}^{(k)}$, dann ist $I(\mathbf{x}^{(k)}) = I(\mathbf{x}^{(k+1)})$, sonst $I(\mathbf{x}^{(k)}) \subset I(\mathbf{x}^{(k+1)})$.

5.2 Penalty–Verfahren (Strafverfahren)

Wir betrachten wieder das Optimierungsproblem

$$z = \min\{f(\mathbf{x}) \ : \ \mathbf{x} \in \Omega\}, \tag{5.4}$$

diesmal aber zunächst mit $f \in C^0(\mathbb{R}^n)$ und die Menge $\Omega \subset \mathbb{R}^n$ sei abgeschlossen. Um die Lösung von (5.4) mit einer Folge einfacherer Optimierungsprobleme ohne Nebenbedingungen zu approximieren, wird die Straffunktion

$$l : \mathbb{R}^n \to \mathbb{R}_+, \quad l(\mathbf{x}) \begin{cases} > 0 & \text{für } \mathbf{x} \notin \Omega, \\ = 0 & \text{für } \mathbf{x} \in \Omega \end{cases}$$

eingeführt. Diese Funktion bestraft Punkte, die nicht zum zulässigen Bereich gehören, mit positiven Funktionswerten.

Beispiel 5.5 Für

$$\Omega = \{\mathbf{x} \in \mathbb{R}^n \ : \ g_i(\mathbf{x}) \leq 0, \ i = 1, \ldots, p; \ g_i(\mathbf{x}) = 0, \ i = p+1, \ldots, m\}$$

ist eine mögliche Straffunktion

$$l(\mathbf{x}) = \sum_{i=1}^{p} \left(g_i^+(\mathbf{x})\right)^\alpha + \sum_{i=p+1}^{m} |g_i(\mathbf{x})|^\alpha$$

mit $\alpha > 0$, $g_i^+(\mathbf{x}) = \max\{0, g_i(\mathbf{x})\}$. □

Definition 5.6 Penalty–Funktion Die gewichtete Summe aus Zielfunktion und Straffunktion

$$p(\mathbf{x}, r) := f(\mathbf{x}) + rl(\mathbf{x}), \quad r \in \mathbb{R}_+, r > 0,$$

wird Penalty–Funktion genannt. Der Parameter r heißt Strafparameter. □

Für fest gewählte Parameter r werden jetzt die Minimierungsprobleme ohne Nebenbedingungen

$$\min_{\mathbf{x} \in \mathbb{R}^n} p(\mathbf{x}, r) \tag{5.5}$$

betrachtet. Der Strafterm belegt die Punkte, die nicht zum zulässigen Bereich gehören, mit positiven Werten, die für große Strafparameter r groß sind. Deswegen hofft man, dass die Minima von (5.5) für große Strafparameter im zulässigen Bereich liegen.

Algorithmus 5.7 Allgemeines Penalty–Verfahren.

1. *Initialisierung.*
   `Wähle` $\mathtt{r}^{(1)} > 0$.
2. *Iteration.* $\mathtt{k} = 0, 1, 2, \ldots$
   `Bestimme ein lokales Minimum` $\mathbf{x}^{(\mathtt{k})}$ `für` $\mathtt{p}(\mathbf{x}, \mathtt{r}^{(\mathtt{k})})$
   `Falls` $\mathbf{x}^{(\mathtt{k})} \in \Omega$`, dann Stopp`
   `sonst wähle` $\mathtt{r}^{(\mathtt{k}+1)} \geq 2\mathtt{r}^{(\mathtt{k})}$

□

Man kann zeigen, dass die $\mathbf{x}^{(k)}$ unter gewissen Voraussetzungen tatsächlich Näherungen eines lokalen Minimums von (5.4) sind.

Satz 5.8 *Seien $f : \mathbb{R}^n \to \mathbb{R}$ eine stetige Funktion, $\mathbf{x}_0$ ein striktes lokales Minimum und $l : \mathbb{R}^n \to \mathbb{R}_+$ eine stetige Straffunktion. Dann gibt es ein $r_0 > 0$ so, dass für alle $r > r_0$ die Penalty–Funktion $p(\mathbf{x}, r)$ ein lokales Minimum $\mathbf{x}(r)$ besitzt, dass für $r \to \infty$ gegen $\mathbf{x}_0$ konvergiert.*

Beweis: Literatur, [JS04, S. 294]. ■

Bemerkung 5.9 Zwei Eigenschaften sind für Penalty–Verfahren von Bedeutung:

1. In vielen Fällen ist die Zielfunktion $f(\mathbf{x})$ differenzierbar. Damit die Anwendung eines Verfahrens vom Newton–Typ zur Bestimmung des Minimums von (5.5) möglich ist, muss die Straffunktion $l(\mathbf{x})$ auch differenzierbar sein.
2. Damit das Verfahren nach endlich vielen Schritten abbricht, ist es wünschenswert, wenn es bereits einen endlichen Wert $\overline{r} > 0$ gibt, so dass ein lokales Minimum $\mathbf{x}_0$ von (5.4) auch lokales Minimum für jedes Problem ohne Nebenbedingungen (5.5) mit $r \geq \overline{r}$ ist. In diesem Fall nennt man die Penalty–Funktion exakt in $\mathbf{x}_0$.

Es stellt sich leider heraus, dass diese beiden wünschenswerten Eigenschaften in der Regel unvereinbar sind. Aus diesem Grunde werden Penalty–Verfahren in der Form von Algorithmus 5.7 praktisch nicht genutzt. Stattdessen betrachtet man modifizierte Penalty–Funktionen, die auf dem Konzept einer erweiterten Lagrange–Funktion (augmented Lagrange–Funktion) beruhen, siehe Literatur. □

5.3 Barrieremethoden

Wir betrachten das Problem

$$z = \min_{\mathbf{x}\in\Omega} f(\mathbf{x}) \tag{5.6}$$

mit den Nebenbedingungen

$$g_i(\mathbf{x}) \leq 0 \text{ für } 1 \leq i \leq p, \quad g_i(\mathbf{x}) = 0 \text{ für } p+1 \leq i \leq m. \tag{5.7}$$

Dabei seien $f, g_i \in C^2(\mathbb{R}^n)$, $i = 1, \ldots, m$, und wir nehmen an, dass (5.6) eine Optimallösung besitzt, die mit $\mathbf{x}_0$ bezeichnet wird.

Barrieremethoden sind eng verwandt mit den Penalty–Verfahren. Auch bei diesen Methoden betrachtet man eine Folge von Hilfsproblemen, bei denen die Zielfunktion $f(\mathbf{x})$ durch gewichtete Strafterme erweitert wird. Die Barriereverfahren erzeugen eine Folge von inneren Punkten, das heißt von Punkten welche die Ungleichungsrestriktionen sogar strikt erfüllen, $g_i(\mathbf{x}) < 0, i = 1, \ldots, p$, während die Gleichungsrestriktionen verletzt sein können.

Bezeichne

$$\begin{aligned} \hat{\Omega} &:= \{\mathbf{x} \in \mathbb{R}^n \ : \ g_i(\mathbf{x}) \leq 0, \ i = 1, \ldots p\}, \\ \hat{\Omega}_0 &:= \{\mathbf{x} \in \mathbb{R}^n \ : \ g_i(\mathbf{x}) < 0, \ i = 1, \ldots p\}. \end{aligned}$$

Man beachte, die Menge $\hat{\Omega}_0$ muss nicht notwendig die topologischen inneren Punkte von $\hat{\Omega}$ enthalten, wähle zum Beispiel $n = p = 1, g_1(x) = 0$. Dann sind $\hat{\Omega} = \mathbb{R}$ und $\hat{\Omega}_0 = \emptyset$.

Barriereverfahren bestrafen solche Punkte aus $\hat{\Omega}_0$, die sich dem Rand von $\hat{\Omega}_0$ nähern. Die Gleichheitsnebenbedingungen werden direkt mit Hilfe von Linearisierungen behandelt. Diese Gleichheitsnebenbedingungen sind grundsätzlich einfacher zu behandeln als Ungleichungsnebenbedingungen. Die Strafterme in den Barriereverfahren, die sogenannten Barriereterme, sind in $\hat{\Omega}_0$ endlich und wachsen zum Rand dieser Menge nach unendlich an. Außerhalb von $\hat{\Omega}$ besitzen sie den Wert ∞. Im Gegensatz zu den Penalty–Verfahren, bei denen die Strafterme sukzessive immer stärker gewichtet werden, siehe Algorithmus 5.7, muss bei den Barriere–Verfahren der Einfluss der Strafterme immer weiter abgeschwächt werden. Damit wird das Gewicht der Zielfunktion im Barriereproblem erhöht und man kann hoffen, dass die Minima der Barriereprobleme unter geeigneten Voraussetzungen gegen ein Minimum von $f(\mathbf{x})$ konvergieren. An der Eigenschaft, dass die Barriereterme außerhalb von $\hat{\Omega}$ unendlich sind, ändert man nichts. Somit ist garantiert, dass die Minima der Barriereprobleme immer in $\hat{\Omega}_0$ liegen.

Definition 5.10 Skalare Barrierefunktion. Eine skalare Barrierefunktion ist eine streng monoton fallende, glatte, konvexe Funktion $b : (0, \infty) \to \mathbb{R}$ mit

$$\lim_{t \to 0+0} b(t) = \infty \quad \text{und} \quad \lim_{t \to 0+0} b'(t) = -\infty.$$

□

Außerdem wird stets $b(t) = \infty$ für $t \leq 0$ gesetzt, so dass $b(t)$ formal eine auf $\mathbb{R}$ definierte konvexe Funktion ist, $b : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$.

Beispiel 5.11 Beispiele für Barrierefunktionen sind

$$b(t) = -\log t, \quad b(t) = \frac{1}{t^\alpha}, \ \alpha > 0.$$

Die logarithmische Barrierefunktion ist in gewisser Hinsicht optimal. □

Zur Konstruktion von Barriereverfahren zur Lösung von Problem (5.6) mit den Nebenbedingungen (5.7) werden nun Hilfsprobleme der Form

$$\inf_{\mathbf{x} \in \mathbb{R}^n} \left\{ f(\mathbf{x}) + \mu \sum_{i=1}^{p} b\left(d_i - g_i(\mathbf{x})\right) \ : \ g_i(\mathbf{x}) = 0, \ i = p+1, \ldots, m \right\} \tag{5.8}$$

betrachtet. In (5.8) ist $\mu > 0$ ein Gewicht für die Barriereterme und die Zahlen $d_i \geq 0, i = 1, \ldots, p$, sind Verschiebungen der Ungleichungsnebenbedingungen in (5.7), das heisst anstatt $g_i(\mathbf{x}) \leq 0$ ist nun $g_i(\mathbf{x}) \leq d_i$ erlaubt. Diese Verschiebungen gestatten es, dass man das Verfahren auch dann anwenden kann, wenn kein innerer Punkt für (5.6), (5.7) bekannt ist. Die Zielfunktion von (5.8) wird abkürzend mit

$$\Phi(\mathbf{x}; \mu, \mathbf{d}) = f(\mathbf{x}) + \mu \sum_{i=1}^{p} b\left(d_i - g_i(\mathbf{x})\right)$$

bezeichnet.

Wir nehmen an, dass (5.8) ein endliches lokales Minimum besitzt. Die gewichtete Summe der Barriereterme in der Zielfunktion garantiert, dass jedes $\mathbf{x}$ mit $\Phi(\mathbf{x}; \mu, \mathbf{d}) \in \mathbb{R}$ die abgeschwächten Nebenbedingungen $g_i(\mathbf{x}) < d_i$, $i = 1, \ldots, p$, erfüllt.

Lemma 5.12 *Falls $f(\mathbf{x})$ und $g_i(\mathbf{x})$, $i = 1, \ldots, p$ konvex sind, so ist auch $\Phi(\mathbf{x}; \mu, \mathbf{d})$ konvex.*

Beweis: Es ist bekannt, dass die Linearkombination konvexer Funktionen mit nichtnegativen Koeffizienten in der Linearkombination eine konvexe Funktion ist. Damit bleibt zu zeigen, dass die Funktionen $b\left(d_i - g_i(\mathbf{x})\right)$, $i = 1, \ldots, p$, konvex sind.

Da die Funktionen, $g_i(\mathbf{x})$ konvex sind, gilt für $\lambda \in [0, 1]$

$$\begin{aligned} d_i - g_i\left(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2\right) &\geq d_i - \left(\lambda g_i\left(\mathbf{x}_1\right) + (1-\lambda) g_i\left(\mathbf{x}_2\right)\right) \\ &= \lambda\left(d_i - g_i\left(\mathbf{x}_1\right)\right) + (1-\lambda)\left(d_i - g_i\left(\mathbf{x}_2\right)\right). \end{aligned}$$

Mit dieser Aussage, mit der Monotonie von $b(t)$ und der Konvexität von $b(t)$ folgt

$$\begin{aligned} b\left(d_i - g_i\left(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2\right)\right) &\leq b\left(\lambda\left(d_i - g_i\left(\mathbf{x}_1\right)\right) + (1-\lambda)\left(d_i - g_i\left(\mathbf{x}_2\right)\right)\right) \\ &\leq \lambda b\left(d_i - g_i\left(\mathbf{x}_1\right)\right) + (1-\lambda) b\left(d_i - g_i\left(\mathbf{x}_2\right)\right). \end{aligned}$$

■

Weiterhin gilt folgende stärkere Aussage.

Satz 5.13 *Gelten die Voraussetzungen von Lemma 5.12 sowie* $\lim_{t\to\infty} b'(t) = 0$. *Die Gleichheitsnebenbedingungen* $g_i(\mathbf{x})$, $i = p+1, \ldots, m$ *seien affin und die Menge der Optimallösungen von (5.6) sei nicht leer und beschränkt. Dann besitzt das Hilfsproblem (5.8) für jedes* $\mu > 0$ *eine Optimallösung und die Minima der Barriereprobleme nähern sich der Optimalmenge von (5.6).*

Beweis: Siehe Literatur. ■

Für Probleme (5.6), die die Bedingungen dieses Satzes erfüllen, kann man nun das folgende Verfahren konstruieren.

Algorithmus 5.14 Barrieremethode für konvexe Probleme.

1. *Initialisierung.*

   Bestimme $\mathbf{x}^{(0)} \in \mathbb{R}^n$ mit $g_i(\mathbf{x}^{(0)}) = 0$ für $i = p+1, \ldots, m$. Wähle $\mu^{(0)} > 0$ und $\mathbf{d}^{(0)} \geq \mathbf{0}$ so dass $d_i^{(0)} > g_i(\mathbf{x}^{(0)})$ für $i = 1, \ldots, p$.

2. *Iteration.* $k = 1, 2, \ldots$

   Wähle $\lambda^{(k)} \in (0,1)$ so, dass mit $\left(\mu^{(k)}, \mathbf{d}^{(k)}\right) := \lambda^{(k)} \left(\mu^{(k-1)}, \mathbf{d}^{(k-1)}\right)$ gilt

   $$g_i\left(\mathbf{x}^{(k-1)}\right) < d_i^{(k)}, \quad \text{für } i = 1, \ldots, p.$$

   Ausgehend von $\mathbf{x}^{(k-1)}$ führt man nun einige Schritte des Newton-Verfahrens (mit Liniensuche) zum lösen des Barriere-problems aus. Das Ergebnis ist $\mathbf{x}^{(k)}$.

□

Im ersten Schritt der Iteration werden sowohl das Gewicht als auch der Verschiebevektor verkleinert. Der Verkleinerungsfaktor wird so gewählt, dass mit dem neuen Verschiebevektor noch alle Ungleichungsnebenbedingungen erfüllt sind. Das Minimum $\mathbf{x}_0\left(\mu^{(k)}, \mathbf{d}^{(k)}\right)$ des Barriereproblems (5.8) zu den Parametern $\left(\mu^{(k)}, \mathbf{d}^{(k)}\right)$ wird im zweiten Schritt approximiert. Da die Barriereterme das Minimum vom Rand der Menge $\{\mathbf{x} : g_i(\mathbf{x}) \leq d_i^{(k)}\}$ abstoßen, kann man nach der Berechnung der Näherung $\mathbf{x}^{(k)}$ von $\mathbf{x}_0\left(\mu^{(k)}, \mathbf{d}^{(k)}\right)$ die Verschiebeparameter $d_i^{(k)}$ in der folgenden Iteration wieder etwas verkleinern.

Die Schwierigkeiten von Algorithmus 5.14 bestehen darin, dass das Newton–Verfahren für $\mu^{(k)} \to 0$ oft schlecht konvergiert. Deshalb wird diese Basisherangehensweise nicht genutzt. Man kann diese Herangehensweise durch Verfeinerung der Barrieremethode verbessern.

5.4 SQP–Verfahren

In diesem Abschnitt wird ein Zugang vorgestellt, der Punkte berechnet, die die notwendige Optimalitätsbedingung, die im Satz 4.35 (Kuhn/Tucker) formuliert ist, erfüllen, die sogenannten SQP–Verfahren (sequential quadratic programming). Es wird also eine Iteration durchgeführt, bei welcher in jedem Schritt ein quadratisches Optimierungsproblem gelöst wird.

Wir betrachten wieder das Optimierungsproblem (5.6) mit den Nebenbedingungen (5.7) und den gleichen Regularitätsvoraussetzungen wie im Abschnitt 5.3. Seien $\mathbf{x}_0$ ein lokales Minimum von (5.6), (5.7) und $\mathbf{z}_0$ der zugehörige Lagrange–Multiplikator zum Lagrange–Problem (4.10). Insgesamt erfüllen $(\mathbf{x}_0, \mathbf{z}_0)$ das Problem, siehe (4.10),

$$\Phi(\mathbf{x}_0, \mathbf{z}_0) = \begin{pmatrix} \nabla_{\mathbf{x}} L(\mathbf{x}_0, \mathbf{z}_0) \\ \mathbf{z}_0^T \mathbf{g}(\mathbf{x}_0) \end{pmatrix} = \begin{pmatrix} \nabla f(\mathbf{x}_0) + (\nabla \mathbf{g}(\mathbf{x}_0))^T \mathbf{z}_0 \\ \mathbf{z}_0^T \mathbf{g}(\mathbf{x}_0) \end{pmatrix} = \mathbf{0} \tag{5.9}$$

mit $\mathbf{z}_0 \geq \mathbf{0}$. Da die Gleichheitsbedingungen ohnehin verschwinden, kann man (5.9) sogar wie folgt schreiben

$$\Phi(\mathbf{x}_0, \mathbf{z}_0) = \begin{pmatrix} \nabla f(\mathbf{x}_0) + (\nabla \mathbf{g}(\mathbf{x}_0))^T \mathbf{z}_0 \\ z_{0,1} g_1(\mathbf{x}_0) \\ \vdots \\ z_{0,p} g_p(\mathbf{x}_0) \\ g_{p+1}(\mathbf{x}_0) \\ \vdots \\ g_m(\mathbf{x}_0) \end{pmatrix} = \mathbf{0}. \tag{5.10}$$

SQP–Verfahren wollen das nichlineare Problem (5.10) mit einem Verfahren vom Newton–Typ lösen. Dazu benötigt man die Jacobi–Matrix von $\Phi(\mathbf{x}, \mathbf{z})$, die durch

$$\begin{aligned} &\Psi\left(\mathbf{x}, \mathbf{z}, H_{L,\mathbf{x}}(\mathbf{x}, \mathbf{z})\right) \\ &= \begin{pmatrix} H_{L,\mathbf{x}}(\mathbf{x}, \mathbf{z}) & \nabla g_1(\mathbf{x}) & \cdots & \nabla g_p(\mathbf{x}) & \nabla g_{p+1}(\mathbf{x}) & \cdots & \nabla g_m(\mathbf{x}) \\ z_1(\nabla g_1(\mathbf{x}))^T & g_1(\mathbf{x}) & & & & & \\ \vdots & & \ddots & & & 0 & \\ z_p(\nabla g_p(\mathbf{x}))^T & & & g_p(\mathbf{x}) & & & \\ (\nabla g_{p+1}(\mathbf{x}))^T & & & & & & \\ \vdots & & & 0 & & 0 & \\ (\nabla g_m(\mathbf{x}))^T & & & & & & \end{pmatrix} \\ &\in \mathbb{R}^{(n+m)\times(n+m)} \end{aligned}$$

gegeben ist. Ein wesentliches Merkmal eines SQP–Verfahrens besteht darin, dass die teure Hesse–Matrix $H_{L,\mathbf{x}}(\mathbf{x}, \mathbf{z})$ in der Regel durch eine einfacher zu berechnende Matrix ersetzt wird.

Unter geeigneten Voraussetzungen kann man zeigen, dass die Jacobi–Matrix $\Psi\left(\mathbf{x}_0, \mathbf{z}_0, H_{L,\mathbf{x}}(\mathbf{x}_0, \mathbf{z}_0)\right)$ nichtsingulär ist und dass das Newton–Verfahren zur Nullstellenbestimmung von $\Phi(\mathbf{x}, \mathbf{z})$ quadratisch konvergiert. Sei eine aktuelle Iterierte $(\mathbf{x}^{(k)}, \mathbf{z}^{(k)})^T$ gegeben. Die Newton–Korrektur $(\Delta\mathbf{x}^{(k)}, \Delta\mathbf{z}^{(k)})^T$ berechnet sich als Lösung des Gleichungssystems

$$\Psi\left(\mathbf{x}^{(k)}, \mathbf{z}^{(k)}, H_{L,\mathbf{x}}(\mathbf{x}^{(k)}, \mathbf{z}^{(k)})\right)\left(\Delta\mathbf{x}^{(k)}, \Delta\mathbf{z}^{(k)}\right)^T = -\Phi(\mathbf{x}^{(k)}, \mathbf{z}^{(k)}).$$

Beim Newton–Verfahren kann man im allgemeinen jedoch nur lokale Konvergenz erwarten, dass heisst, der Startwert muss nahe genug an der (unbekannten) Lösung sein. Speziell für die Funktion $\Phi(\mathbf{x}, \mathbf{z})$ kann man auch nicht garantieren, dass alle Iterierten die Ungleichungen $g_i(\mathbf{x}^{(k)}) \leq 0$ und die Nichtnegativitätsbedingung $z_i^{(k)} \geq 0$ erfüllen. Damit ist es möglich, dass das Newton–Verfahren gegen eine nicht zulässige Lösung von $\Phi(\mathbf{x}, \mathbf{z}) = \mathbf{0}$ konvergiert, bei der Nebenbedingungen nicht erfüllt sind oder die Lagrange–Multiplikatoren negativ sind. Die Konvergenz gegen eine solche Lösung muss verhindert werden. Dazu wird anstelle des Newton–Verfahrens das System

$$\Psi\left(\mathbf{x}^{(k)}, \mathbf{z}^{(k+1)}, B^{(k)}\right)\left(\Delta\mathbf{x}^{(k)}, \Delta\mathbf{z}^{(k)}\right)^T = -\Phi(\mathbf{x}^{(k)}, \mathbf{z}^{(k)}) \tag{5.11}$$

betrachtet, wobei $\Delta\mathbf{x}^{(k)}$, $\Delta\mathbf{z}^{(k)}$ und $\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \Delta\mathbf{z}^{(k)}$ die zusätzlichen Forderungen

$$z_i^{(k+1)} \geq 0 \quad \text{für } i = 1, \ldots, p, \tag{5.12}$$

$$g_i\left(\mathbf{x}^{(k)}\right) + \nabla\left(g_i\left(\mathbf{x}^{(k)}\right)\right)^T \Delta\mathbf{x}^{(k)} \leq 0 \quad \text{für } i = 1, \ldots, p \tag{5.13}$$

erfüllen. Im Vergleich zum Newton–Verfahren ersetzt man $H_{L,\mathbf{x}}(\mathbf{x}^{(k)}, \mathbf{z}^{(k)})$ durch eien Matrix $B^{(k)}$, die in der Regel durch gewisse Quasi–Newton–Korrekturen (sogenannte Broyden–Verfahren) erzeugt wird. Des weiteren wird der Vektor $\mathbf{z}^{(k)}$ auf der linken Seite durch $\mathbf{z}^{(k+1)}$ ersetzt. Man erhält ein implizites Gleichungssystem, welches nicht mehr linear bezüglich der Lagrange–Multiplikatoren ist. Außerdem werden noch die linearen Ungleichungsbedingungen (5.12), (5.13) an $\Delta\mathbf{x}^{(k)}$ und $\Delta\mathbf{z}^{(k)}$ gestellt.

Ausgeschrieben besagt (5.11) – (5.13)

$$\begin{aligned} \nabla f\left(\mathbf{x}^{(k)}\right) + B^{(k)}\Delta\mathbf{x}^{(k)} + \sum_{i=1}^{m} z_i^{(k+1)} \nabla g_i\left(\mathbf{x}^{(k)}\right) &= 0, \\ z_i^{(k+1)}\left(g_i\left(\mathbf{x}^{(k)}\right) + \left(\nabla g_i\left(\mathbf{x}^{(k)}\right)\right)^T \Delta\mathbf{x}^{(k)}\right) &= 0, \quad i = 1, \ldots, p, \\ g_i\left(\mathbf{x}^{(k)}\right) + \left(\nabla g_i\left(\mathbf{x}^{(k)}\right)\right)^T \Delta\mathbf{x}^{(k)} &= 0, \quad i = p+1, \ldots, m. \end{aligned} \tag{5.14}$$

Wir betrachten das folgende quadratische Programm

$$z = \min_{\mathbf{s}\in\mathbb{R}^n} \left(\nabla f\left(\mathbf{x}^{(k)}\right)\right)^T \mathbf{s} + \frac{1}{2}\mathbf{s}^T B^{(k)} \mathbf{s} \tag{5.15}$$

unter den Nebenbedingungen

$$g_i\left(\mathbf{x}^{(k)}\right) + \left(\nabla g_i\left(\mathbf{x}^{(k)}\right)\right)^T \mathbf{s} \leq 0, \quad i = 1, \ldots, p, \tag{5.16}$$

$$g_i\left(\mathbf{x}^{(k)}\right) + \left(\nabla g_i\left(\mathbf{x}^{(k)}\right)\right)^T \mathbf{s} = 0, \quad i = p+1, \ldots, m. \tag{5.17}$$

Erfülle dieses Problem die Voraussetzungen des Satzes 4.35 (Kuhn/Tucker). Dann sind die notwendigen Bedingungen für ein Minimum gerade die Gleichungen (5.14). *Übungsaufgabe* Damit ergibt sich folgendes Verfahren:

Algorithmus 5.15 SQP–Algorithmus (Grundform).

1. *Initialisierung.*

   Wähle $\mathbf{x}^{(0)} \in \mathbb{R}^n, \mathrm{B}^{(0)} = \left(\mathrm{B}^{(0)}\right)^T \left(\approx \mathrm{H}_{\mathrm{L},\mathbf{x}}\left(\mathbf{x}^{(0)}, \mathbf{z}^{(0)}\right)\right)$ für ein $\mathbf{z}^{(0)} \in \mathbb{R}^m$ mit $\mathbf{z}_i^{(0)} > 0, i = 1, \ldots, p$

2. *Iteration.* $k = 0, 1, 2, \ldots$

   Bestimme die Lösung s von (5.15) - (5.17) und einen zugehörigen Lagrange-Multiplikator z. Setze

   $$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}, \quad \mathbf{z}^{(k+1)} = \mathbf{z}.$$

   Bestimme eine symmetrische Matrix

   $$\mathrm{B}^{(k+1)} \approx \mathrm{H}_{\mathrm{L},\mathbf{x}}\left(\mathbf{x}^{(k+1)}, \mathbf{z}^{(k+1)}\right).$$

□

Falls $B^{(k)}$ positiv semidefinit ist, dann ist (5.15) – (5.17) ein konvexes quadratisches Programm. In diesem Fall sind die Bedingungen des Satzes 4.35 notwendig und hinreichend für ein globales Minimum, siehe auch Satz 3.27. Zur Lösung kann man beispielsweise ein Projektionsverfahren nehmen, siehe Abschnitt 5.1.

Literaturverzeichnis

[Bor01] Karl Heinz Borgwardt. *Optimierung, Operations Research, Spieltheorie*. Birkhäuser Verlag Basel Boston Berlin, 2001.

[ERSD77] K.-H. Elster, R. Reinhardt, M. Schäuble, and G. Donath. *Einführung in die nichtlineare Optimierung*, volume 63 of *Mathematisch–Naturwissenschaftliche Bibliothek*. BSB B.G. Teubner Verlagsgesellschaft Leipzig, 1977.

[JS04] F. Jarre and J. Stoer. *Optimierung*. Springer–Verlag Berlin Heidelberg, 2004.

Index