Vorlesung über Optimierung

Jochen Werner

Wintersemester 1999/2000

Inhaltsverzeichnis

Kapitel 1

Einführung

Eine *Optimierungsaufgabe* (statt von einer Optimierungsaufgabe werden wir später auch oft von einem *Programm* sprechen) ist durch zwei Daten gegeben, nämlich durch die *Menge der zulässigen Lösungen* M und die *Zielfunktion* $f\colon M \longrightarrow \mathbb{R}$. Man kann sich M als eine Menge zugelassener Strategien zur Lösung einer Planungsaufgabe vorstellen. Jedem Element $x \in M$ sind hierdurch auftretende Kosten $f(x)$ zugeordnet, diese gilt es zu minimieren. Daher wird die Zielfunktion auch manchmal *Kostenfunktion* genannt. Die durch M und f gegebene Aufgabe schreiben wir in der Form

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M$$

und nennen $x^* \in M$ eine (globale) *Lösung* von (P), wenn $f(x^*) \leq f(x)$ für alle $x \in M$. Naheliegenderweise nennt man $x^* \in M$ eine *lokale Lösung* von (P), wenn es eine Umgebung U^* von x^* mit $f(x^*) \leq f(x)$ für alle $x \in M \cap U^*$ gibt. Eine triviale Bemerkung besteht darin, dass das Maximieren einer Funktion $g\colon M \longrightarrow \mathbb{R}$ auf M, wenn also jeder zulässigen Strategie ein hierdurch eintretender Gewinn zugeordnet ist, auf die Minimierungsaufgabe (P) durch Einführen von $f := -g$ zurückgeführt werden kann.

Wir werden uns darauf beschränken, *endlichdimensionale* bzw. *finite* Optimierungsaufgaben zu betrachten. Hier ist M eine Teilmenge des $\mathbb{R}^n$, die typischerweise durch endlich viele Ungleichungen und Gleichungen gegeben ist. Die Optimierungsaufgaben, die wir betrachten werden, haben daher i. Allg. die folgende Form:

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\, h(x) = 0\}.$$

Hier sind die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ gegeben, die $\leq$–Beziehung zwischen Vektoren ist stets komponentenweise zu verstehen. Im Gegensatz hierzu spricht man von *unendlichdimensionalen* bzw. *infiniten* Optimierungsaufgaben, wenn M Teilmenge eines (unendlichdimensionalen) linearen normierten Raumes ist bzw. der Ausgangsraum $\mathbb{R}^n$ (in dem eine Lösung gesucht wird) oder die Bildräume $\mathbb{R}^l$ bzw. $\mathbb{R}^m$ der Restriktionsabbildungen g bzw. h durch (unendlichdimensionale) lineare normierte Räume ersetzt sind[1]. Natürlich ist auch der Fall möglich, dass $M = \mathbb{R}^n$, also keine Restriktionen auferlegt werden. Man

[1] Infinite Optimierungsaufgaben werden z. B. ausführlich bei

Luenberger, D. G. (1969) *Optimization by Vector Space Methods.* John Wiley & Sons, New York-

spricht dann von einer *unrestringierten* Optimierungsaufgabe. Auf die numerische Behandlung unrestringierter Optimierungsaufgaben werden wir nicht eingehen[2]. Der Fall, dass M offen ist, ist (zumindestens theoretisch) nur unwesentlich schwieriger, da eine Aufgabe dieser Art sozusagen lokal unrestringiert ist.

1.1 Beispiele

Dadurch, dass wir eben erläuterten, was eine "allgemeine" Optimierungsaufgabe ist, haben wir gegen einen Rat verstoßen, den R. P. Boas[3] gegeben hat:

> Suppose that you want to teach the "cat" concept to a very young child. Do you explain that a cat is a relatively small, primarily carnivorous mammal[4] with retractile[5] claws, a distinctive sonic output, etc.? I'll bet not. You probably show the kid a lot of different cats, saying "kitty" each time, until it gets the idea. To put it more generally, generalizations are best made by abstractions from experience.

Wir geben daher gleich einige Beispiele von Optimierungsaufgaben an. Es mangelt natürlich nicht an Beispielen, denn eigentlich immer (in- und außerhalb der Mathematik) versucht man, etwas möglichst gut zu machen, wobei i. Allg. gewisse Restriktionen zu beachten sind.

Beispiel: Eine der ältesten Optimierungsaufgaben in der Geschichte der Mathematik findet sich in Euklid's Elementen, Buch VI, Theorem 27:

- Finde einen Punkt E auf der Seite $\overline{BC}$ eines Dreiecks $\triangle ABC$ derart, dass das Parallelogramm $ADEF$ mit Eckpunkten D bzw. F auf den Seiten $\overline{AB}$ bzw. $\overline{AC}$ maximalen Flächeninhalt besitzt.

Die Lösung ist offensichtlich dadurch gegeben, dass man E als Mittelpunkt von $\overline{BC}$ wählt. In Abbildung 1.1 wird dies verdeutlicht. Denn ist E beliebig auf $\overline{BC}$ und

$$x := \frac{\text{Länge}(\overline{BE})}{\text{Länge}(\overline{BC})},$$

London-Sydney-Toronto.

Werner, J. (1984) *Optimization. Theory and Applications.* Vieweg, Braunschweig-Wiesbaden.

Werner, J. (1989) *Optimierung.* Fernuniversität-Gesamthochschule Hagen.

Jahn, J. (1994) *Introduction to the theory of nonlinear optimization.* Springer, Berlin

untersucht.

[2]Siehe hierzu

Dennis, J. E. and R. B. Schnabel (1984) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, Englewood Cliffs.

Werner, J. (1992b) *Numerische Mathematik 2.* Vieweg, Braunschweig-Wiesbaden.

[3]R. P. Boas, Can we make mathematics intelligible? American Mathematical Monthly 88, 1981, 727–731.

[4]fleischfressendes Säugetier

[5]einziehbar

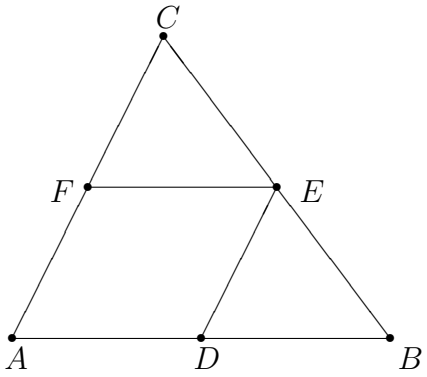

Abbildung 1.1: Die Lösung des ältesten Optimierungsproblems

so ist

$$g(x) := \text{Flächeninhalt}(ADEF) = 2x(1-x)\text{Flächeninhalt}(\triangle ABC),$$

und diese Funktion g wird offenbar auf $M := [0,1]$ maximal für $x^* := \frac{1}{2}$. □

Beispiel: Das folgende Problem scheint 1629 zum ersten Mal von Fermat formuliert worden zu sein:

- Gegeben seien drei Punkte in der Ebene. Man finde einen Punkt in der Ebene derart, dass die Summe der Abstände dieses Punktes zu den drei vorgegebenen Punkten minimal ist.

Die Verallgemeinerung auf m Punkte im $\mathbb{R}^n$ heißt das Fermat-Weber-Problem:

- Gegeben seien $m \geq 3$ paarweise verschiedene Punkte $a_1, \ldots, a_m \in \mathbb{R}^n$ und positive reelle Zahlen $w_1, \ldots, w_m$. Man bestimme eine Lösung $x^* \in \mathbb{R}^n$ von

  $$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{i=1}^{m} w_i \, \|x - a_i\| \quad \text{auf} \quad M := \mathbb{R}^n,$$

  wobei $\|\cdot\|$ in diesem Abschnitt die *euklidische Norm* auf dem $\mathbb{R}^n$ bedeutet.

Verglichen mit den später zu untersuchenden Optimierungsaufgaben ist das Fermat-Weber-Problem einfach in der Hinsicht, dass es sich hierbei um eine unrestringierte, konvexe Optimierungsaufgabe handelt. Schwierig ist es es vor allem deshalb, weil die Zielfunktion nicht überall differenzierbar ist.

Die ökonomische Interpretation (man spricht in den Wirtschaftswissenschaften auch von dem "Standortproblem") könnte die folgende sein: Eine Warenhauskette mit Filialen in $a_1, \ldots, a_k$ und Zulieferern in $a_{k+1}, \ldots, a_m$ will den Standort eines zusätzlichen Lagers bestimmen. Dieser soll so gewählt werden, dass eine gewichtete Summe der Abstände vom Lager zu den Filialen und von den Zulieferern zum Lager minimal wird.

Beim Fermat-Weber-Problem ist der Abstand zwischen zwei Punkten durch den euklidischen Abstand gegeben. Es liegt nun nicht nur an der bekannten Verallgemeinerungswut der Mathematiker, dass auch andere Abstandsbegriffe bzw. Normen in der

Literatur betrachtet wurden. Hierzu gehören insbesondere die 1-Norm, die ∞-Norm (Maximumnorm) und positive Linearkombinationen dieser beiden Normen als Spezialfälle sogenannter polyedrischer Normen (hier ist die Einheitskugel ein Polyeder).

Wir wollen hier auf das Fermat-Weber-Problem gar nicht weiter eingehen, sondern einen hübschen geometrischen Beweis dafür angeben, das beim eingangs genannten Fermat-Problem der gesuchte Punkt (auch Fermat- oder Torricelli-Punkt genannt) derjenige ist, von dem die drei Seiten des (spitzwinkligen) Dreiecks unter einem Winkel von 120° gesehen werden.

Gegeben sei ein spitzwinkliges Dreieck in der Ebene mit den Ecken A, B und C. In diesem Dreieck wähle man sich einen beliebigen Punkt P und verbinde ihn mit den Ecken. Das innere Dreieck $\triangle APB$ drehe man um 60° um B und erhalte das Dreieck $\triangle C'P'B$. In Abbildung 1.2 ist die Konstruktion angegeben. Dann sind $\triangle ABC'$ und

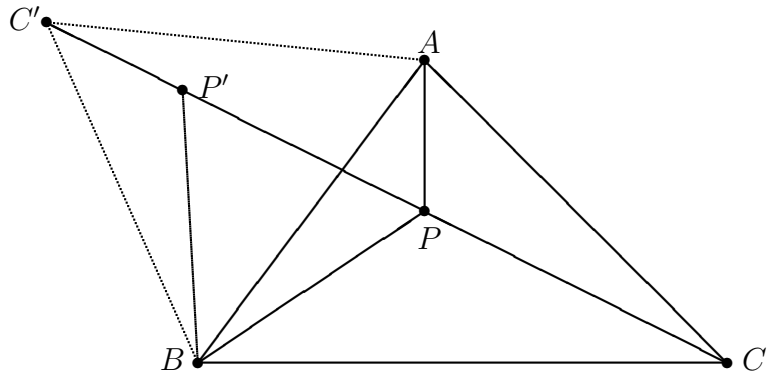

Abbildung 1.2: Konstruktion zum Fermat-Problem

$\triangle PBP'$ gleichseitig, die Winkel in diesen Dreiecken also jeweils 60°. Daher ist

$$AP + BP + CP = C'P' + P'P + PC,$$

und die rechtsstehende Summe ist die Länge eines i. Allg. gebrochenen Streckenzuges. Dieser ist minimal, wenn er ein Geradensegment ist. In diesem Falle ist

$$\angle BPC = 180° - \angle BPP' = 120°$$

und

$$\angle APB = \angle C'P'B = 180° - \angle PP'B = 120°.$$

Der gesuchte Punkt P, für den $AP + BP + CP$ minimal ist, ist also derjenige Punkt P, für den

$$\angle APB = \angle BPC = \angle CPA = 120°.$$

Diese Lösung des Fermat-Problems kann man bei H. S. M. Coxeter (1969, S. 21)[6] nachlesen. □

Beispiel: Ein von J. J. Sylvester (1857) gestelltes Problem lautet:

[6] H. S. M. Coxeter (1969) Introduction to Geometry. John Wiley & Sons, New York.

- It is required to find the least circle which shall contain a given system of points in a plane.

Nur leicht verallgemeinert bedeutet dies: Gegeben seien m Punkte $a_1, \ldots, a_m \in \mathbb{R}^n$, gesucht ist euklidische Kugel $B[x;r] := \{y \in \mathbb{R}^n : \|y-x\|_2 \leq r\}$ mit minimalem Radius r, welche die vorgegebenen Punkte enthält, für die also $\|a_i - x\|_2 \leq r$, $i = 1, \ldots, m$. Mit der Variablentransformation $r = \sqrt{2\delta}$ erhält man die Aufgabe:

$$\begin{cases} \text{Minimiere} \quad f(\delta, x) := \delta \quad \text{auf} \\ M := \{(\delta, x) \in \mathbb{R} \times \mathbb{R}^n : \frac{1}{2}\|x - a_i\|_2^2 \leq \delta,\ i = 1, \ldots, m\}. \end{cases}$$

Dies ist also eine Optimierungsaufgabe mit einer linearen Zielfunktion und (einfachen) quadratischen Ungleichungsnebenbedingungen. □

Das folgende Beispiel kann, zumindestens dann, wenn man es in eine unrestringierte Optimierungsaufgabe umwandelt, schon mit Methoden der Schulmathematik behandelt werden.

Beispiel: Man konstruiere eine möglichst billige Dose (mathematisch: Kreiszylinder) mit Radius r und Höhe h, welche ein vorgegebenes Volumen $V > 0$ besitzt. Die Kosten des Bodens und des Deckels seien c_1 Geldeinheiten (etwa Euro) pro Quadrateinheit (etwa cm^2), entsprechend die des Mantels c_2 Geldeinheiten. Die Gesamtkosten sind gegeben durch

$$f(r, h) := 2\pi r^2 c_1 + 2\pi r h c_2,$$

diese gilt es unter der Nebenbedingung

$$\pi r^2 h = V$$

(sowie $r > 0$, $h > 0$) zu minimieren. □

Beispiel: Auf *lineare Optimierungsaufgaben* (hier sind die Zielfunktion f sowie die Restriktionsabbildungen g und h affin linear) wollen wir nur als Spezialfall allgemeinerer Aufgabenstellungen eingehen. Trotzdem wollen wir hier ein (lineares) *Netzwerkflussproblem* schildern[7] und es als eine lineare Optimierungsaufgabe "entlarven".

Die zugrundeliegende Aufgabe kann man sich folgendermaßen vorstellen: Ein gewisses Gut, sagen wir Orangen, wird in gewissen Orten in einer bestimmten Menge angeboten und an anderen Orten verlangt. Schließlich gibt es Orte, die nichts anbieten und nichts verlangen, in denen aber umgeladen werden darf. Gewisse Orte sind miteinander durch Verkehrswege miteinander verbunden. Die Kosten für den Transport einer Mengeneinheit des Gutes längs eines Verkehrsweges sind bekannt, ferner ist die Kapazität eines jeden möglichen Transportweges vorgegeben. Diese gibt Unter- und Obergrenzen für die zu transportierende Menge auf dem Weg an. Gesucht ist ein kostenminimaler Transportplan.

[7] Als Literatur wird

D. P. Bertsekas (1998) Network Optimization: Continuous and Discrete Models. Athena Scientific, Belmont

empfohlen.

Wir werden zunächst einige Grundbegriffe klären. Ein *Netzwerk* (der Sprachgebrauch ist nicht ganz einheitlich: man spricht auch von einem gerichteten Graphen oder einem Digraphen) $(\mathcal{N}, \mathcal{A})$ besteht aus der endlichen Menge $\mathcal{N}$ der *Knoten* und der Menge $\mathcal{A}$ der *Pfeile* (häufig auch *Bögen* oder gelegentlich auch *gerichtete Kanten* genannt) [8], wobei $\mathcal{A} \subset \mathcal{N} \times \mathcal{N}$. Jeder Pfeil $(i, j) \in \mathcal{A}$ ist also ein geordnetes Paar von Knoten i und j. Hierbei heißt i der *Startknoten* und j der *Endknoten* des Pfeils (i, j).

Mit jedem Knoten $k \in \mathcal{N}$ ist eine Mengenangabe b_k des im Netzwerk zu transportierenden Gutes verbunden. Ist $b_k > 0$, so sind b_k Mengeneinheiten dieses Gutes im Knoten k vorhanden und Knoten k wird ein *Angebotsknoten* genannt. Ist dagegen $b_k < 0$, so werden dort $|b_k|$ Mengeneinheiten benötigt, man spricht von einem *Bedarfsknoten*. Im Fall $b_k = 0$ handelt es sich um einen *reinen Umladeknoten*. Weiter heißt ein Angebotsknoten *reiner Angebotsknoten*, wenn er nicht Endknoten eines Pfeils ist. Analog werden Bedarfsknoten ohne ausgehende Pfeile als *reine Bedarfsknoten* bezeichnet. Es wird angenommen, dass $\sum_{k\in\mathcal{N}} b_k = 0$, also das Gesamtangebot gleich dem Gesamtbedarf ist.

Zu jedem Pfeil $(i, j) \in \mathcal{A}$ des Netzwerks gehören die Kosten c_{ij} für den Fluss einer Mengeneinheit auf ihm. Mit x_{ij} wird der Fluss auf diesem Pfeil bezeichnet, die *Kapazitätsgrenzen* des Pfeils sind durch l_{ij} und u_{ij} angegeben. Gesucht wird ein Fluss im Netzwerk, der unter Berücksichtigung der Kapazitätsbeschränkungen die Angebote und "Bedarfe" mengenmäßig ausgleicht und die dafür erforderlichen Kosten minimiert. Dabei ist in jedem Knoten der Fluss zu erhalten. Dies bedeutet für den Knoten $k \in \mathcal{N}$, dass die Summe der Flüsse auf seinen eingehenden Pfeilen plus der in ihm verfügbaren (wenn k ein Angebotsknoten) beziehungsweise minus der von ihm benötigten (wenn k ein Bedarfsknoten) Menge $|b_k|$ gleich der Summe der Flüsse auf seinen ausgehenden Pfeilen ist. Die Flusserhaltungsbedingung für den Knoten k lautet daher

$$\sum_{i:(i,k)\in\mathcal{A}} x_{ik} + b_k = \sum_{j:(k,j)\in\mathcal{A}} x_{kj}.$$

Das kapazitierte lineare Netzwerkflussproblem (bzw. Minimum Cost Flow Problem) lässt sich daher wie folgt formulieren:

$$\begin{cases} \text{Minimiere} \quad \displaystyle\sum_{(i,j)\in\mathcal{A}} c_{ij}x_{ij} \\ \text{unter den Nebenbedingungen} \\ \displaystyle\sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik} = b_k \qquad (k \in \mathcal{N}), \qquad l_{ij} \le x_{ij} \le u_{ij} \qquad ((i,j) \in \mathcal{A}). \end{cases}$$

Diese Aufgabe wollen wir nun in Matrix-Vektorschreibweise formulieren. Dies kann folgendermaßen geschehen. Der Fluss $x = (x_{ij})$ hat soviele Komponenten wie es Pfeile gibt, ihre Anzahl sei $n := \#(\mathcal{A})$. Es liegt also nahe, $\mathcal{A}$ durchzunumerieren. Es sei etwa $\mathcal{A} = \{l_1, \ldots, l_n\}$ mit $l_p = (i_p, j_p)$, $p = 1, \ldots, n$. Dann kann $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ als Vektor $x = (x_1, \ldots, x_n)^T$ mit $x_p = x_{i_p j_p}$, $p = 1, \ldots, n$, geschrieben werden, entsprechendes gilt für die Kosten $c = (c_{ij})$ und Kapazitätsgrenzen $l = (l_{ij})$ und $u = (u_{ij})$. Ist ferner $m :=$

[8] Die Bezeichnungen $\mathcal{N}$ bzw. $\mathcal{A}$ stehen für $\mathcal{N}$odes bzw. $\mathcal{A}$rcs.

$\#(\mathcal{N})$ die Anzahl der Knoten, so kann man $(b_k)_{k\in\mathcal{N}}$ zu einem Vektor $b = (b_1, \ldots, b_m)^T$ zusammenfassen. Definiert man schließlich noch die *Knoten-Pfeil-Inzidenzmatrix* $A = (a_{kp}) \in \mathbb{R}^{m\times n}$ durch

$$a_{kp} := \begin{cases} +1, & \text{falls} \quad k = i_p, \\ -1, & \text{falls} \quad k = j_p, \\ 0 & \text{sonst}, \end{cases}$$

so erkennt man, dass obiges Netzwerkflussproblem in der Form

$$\text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : l \le x \le u,\ Ax = b\}$$

geschrieben werden kann. Im Gleichungssystem $Ax = b$ summieren die Gleichungen sich zu 0, daher kann z. B. die letzte Gleichung gestrichen werden. Als Beispiel betrachten wir das in Abbildung 1.3 angegebene Netzwerk mit 5 Knoten und 7 Pfeilen. Die Pfeile

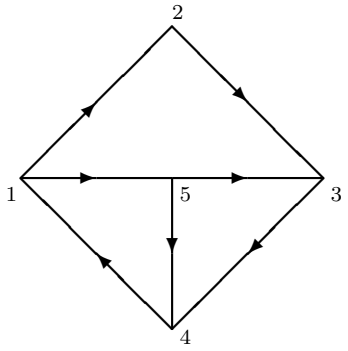

Abbildung 1.3: Ein Netzwerk mit 5 Knoten und 7 Pfeilen

seien in der folgenden Reihenfolge numeriert:

$$\mathcal{A} = \{(1,2), (2,3), 3, 4), (4,1), (1,5), (5,3), (5,4).$$

Der zugehörige Kostenvektor sei $c = (2,2,2,1,1,1,1)^T$, die Kapazitätsschranken

$$u = (0.5, 0.5, 0.1, 0.5, 1, 1, 1)^T, \qquad l = -u.$$

Schließlich sei der Vektor b durch $b = (1, 1, 0.5, 0.5, -3)^T$ gegeben. Als Knoten-Pfeil-Inzidenzmatrix erhält man

$$A = \begin{pmatrix} 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 & 1 \end{pmatrix}.$$

Als Lösung erhalten wir

$$x^* = \begin{pmatrix} -0.5 \\ 0.5 \\ 0.0 \\ -0.5 \\ 1.0 \\ -1.0 \\ -1.0 \end{pmatrix}.$$

Zum oben geschilderten allgemeinen Netzwerkflussproblem gibt es einige bekannte Spezialfälle. Im folgenden nehmen wir an, die unteren Kapazitätsschranken seien durch $l := 0$ gegeben. Ist z. B. jeder Knoten ein reiner Angebots- oder ein reiner Bedarfsknoten, so erhält man das *Transportproblem*. Beim *Maximalflussproblem* ist ein Netzwerk gegeben, in dem zwei Knoten q (Quelle, kein Pfeil ende in q) und s (Senke, kein Pfeil startet in s) ausgezeichnet sind. Längs der Pfeile sind wieder Kapazitäten festgelegt. Es wird angenommen, dass es eine die Quelle q und die Senke s verbindende Pfeilfolge gibt und nach dem maximalen Fluss von q nach s gefragt, also nach der maximalen Anzahl der Mengeneinheiten, die bei q losgeschickt werden können und in s ankommen, wobei natürlich die Kapazitätsbeschränkungen zu berücksichtigen sind (und alle Knoten Umladeknoten sind). Zur Einordnung in das allgemeine Netzwerkflussproblem nehmen wir an, dass die Kosten auf allen Pfeilen verschwinden und auf dem künstlichen Pfeil (s, q) die Kosten durch -1 gegeben sind. In der folgenden Abbildung geben wir ein Netzwerk mit 8 Knoten und 14 Pfeilen an, eingetragen sind ferner die Kapazitäten längs der Pfeile. Was ist der maximale Fluss? Klar ist, dass dieser nicht größer als 6 sein

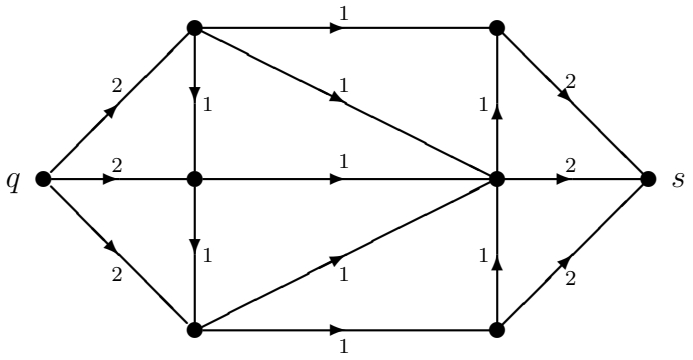

Abbildung 1.4: Ein Netzwerk mit 8 Knoten und 14 Pfeilen

kann, da die drei Wege weg von der Quelle nur eine Gesamtkapazität von 6 besitzen.

In der Abbildung 1.5 geben wir einen Fluss mit dem Wert 5 an. Gibt es auch einen mit dem Wert 6? □

Beispiel: Bei einer *quadratischen Optimierungsaufgabe* sind die Restriktionsabbildungen g und h affin linear, die Zielfunktion f aber quadratisch. Diese hat also die Form

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

mit vorgegebenen $c \in \mathbb{R}^n$ und (o. B. d. A. symmetrischer) Matrix $Q \in \mathbb{R}^{n \times n}$. Besonders angenehm ist hier der Fall, dass Q positiv semidefinit bzw. sogar positiv definit, weil dann f sogar konvex bzw. gleichmäßig konvex ist. Sucht man z. B. einen nichtnegativen Vektor x, für den das überbestimmte lineare Gleichungssystem $Ax = b$ (mit $A \in \mathbb{R}^{m \times n}$ und $m \geq n$, $b \in \mathbb{R}^m$) bezüglich der euklidischen Norm $\| \cdot \|_2$ einen minimalen Defekt besitzt, so hat man die (vorzeichenbeschränkte) quadratische Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \frac{1}{2} \|Ax - b\|_2^2, \quad x \geq 0,$$

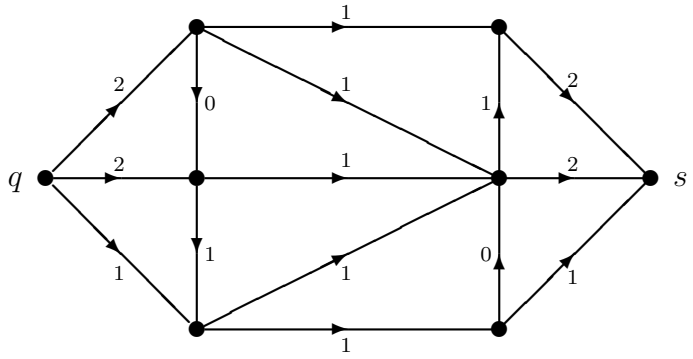

Abbildung 1.5: Ein Fluss mit dem Wert 5

zu lösen. Das *Portfolio Selection Problem* (sein "Erfinder", H. Markowitz, erhielt 1990 den Nobelpreis für Wirtschaftswissenschaften), bei dem es, lax gesagt, darum geht, ein vorhandenes Kapital so auf verschiedene Anlageformen zu verteilen, dass mit minimalem Risiko ein maximaler Ertrag erreicht wird (man erkennt, dass hier eigentlich zwei sich gegenseitig behindernde Ziele erreicht werden sollen), führt nach geeigneten Vereinfachungen ebenfalls auf ein quadratisches Programm. □

1.2 Problemstellungen der Optimierung

Wir wollen nun die wesentlichen Fragestellungen der Optimierung schildern und dabei gleichzeitig schon ein Gefühl für einige typische Vorgehensweisen vermitteln.

- Unter welchen Voraussetzungen besitzt (P) eine globale Lösung, wann ist diese eindeutig?

Viele (aber nicht[9] alle) Existenzbeweise beruhen auf einem Kompaktheitsschluss. Mit einem $x_0 \in M$ (ein solches Element existiert, wenn M nichtleer bzw. (P) *zulässig* ist) bilde man die sogenannte *Niveaumenge*

$$L_0 := M \cap \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}.$$

Außerhalb von L_0 braucht man offenbar nicht nach einer globalen Lösung von (P) zu suchen, weil Elemente aus dem Komplement von L_0 nicht zulässig sind oder größere Kosten als x_0 verursachen. Ist nun L_0 kompakt und f auf L_0 *nach unten halbstetig*, d. h. gilt die Implikation

$$\{x_k\} \subset L_0, \quad \lim_{k\to\infty} x_k = x \Longrightarrow f(x) \leq \liminf_{k\to\infty} f(x_k),$$

[9]Eine etwas andere Beweisanordnung ist die folgende: Zunächst zeigt man, dass $\inf(\mathrm{P}) := \inf_{x\in M} f(x) > -\infty$, die Zielfunktion also auf der Menge der zulässigen Lösungen nach unten beschränkt ist. Ist dies gelungen, so wähle man eine *Minimalfolge* $\{x_k\}$ aus, also eine Folge $\{x_k\} \subset M$ mit $f(x_k) \to \inf(\mathrm{P})$. Kann man zeigen, dass $\{x_k\}$ einen Häufungspunkt besitzt, so ist dieser i. Allg. (z. B. wenn M abgeschlossen und f auf M stetig ist) eine Lösung von (P).

Es sei aber ausdrücklich darauf hingewiesen, dass es auch Existenzsätze gibt, deren Beweis sich diesen Mustern entzieht.

so folgt aus einem bekannten Satz der Analysis, dass f auf L_0 sein Minimum annimmt, also (P) eine globale Lösung besitzt. Man sollte sich hier von dem Begriff der Halbstetigkeit nicht zu sehr abgeschreckt fühlen, da die Zielfunktion im weiteren stets mindestens stetig ist. Die *Eindeutigkeit* einer globalen Lösung wird man nur unter sehr einschneidenden Voraussetzungen an die Daten einer Optimierungsaufgabe erwarten können. Beherrschend wird hier, wie in vielen weiteren Bereichen der Optimierung, der Begriff der *Konvexität* sein. Hierauf werden wir ausführlich zurückkommen.

Beispiel: Mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ mit $m \geq n$ betrachte man das vorzeichenbeschränkte lineare Ausgleichsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \frac{1}{2}\|Ax - b\|_2^2, \quad x \geq 0.$$

Wir wollen uns überlegen, dass (P) eine globale Lösung besitzt, benutzen dabei aber schon die später zu beweisende Tatsache, dass die Menge

$$K := \{y = Ax : x \geq 0\}$$

abgeschlossen ist. Die Aufgabe (P) ist dann äquivalent zu

$$(\hat{\text{P}}) \qquad \text{Minimiere} \quad \hat{f}(y) := \frac{1}{2}\|y - b\|_2^2, \quad y \in K.$$

Hiermit meinen wir: Ist $x^* \geq 0$ eine Lösung von (P), so ist $y^* := Ax^*$ eine Lösung von $(\hat{\text{P}})$. Und umgekehrt: Ist $y^* \in K$, also $y^* = Ax^*$ mit $x^* \geq 0$, eine Lösung von $(\hat{\text{P}})$, so ist x^* eine Lösung von (P). Mit $y_0 := 0$ (oder einem beliebigen anderen Element von K) betrachte man nun die zu $(\hat{\text{P}})$ gehörende Niveaumenge

$$\hat{L}_0 := K \cap \{y \in \mathbb{R}^m : \hat{f}(y) \leq \hat{f}(y_0)\}.$$

Als Durchschnitt einer abgeschlossenen und einer kompakten Menge ist $\hat{L}_0$ kompakt. Da die Zielfunktion $\hat{f}$ trivialerweise stetig ist, besitzt $(\hat{\text{P}})$ und damit auch (P) eine Lösung. Ist Rang $(A) = n$, besitzt A also vollen Rang, so ist (P) *eindeutig* lösbar (Beweis?). □

Beispiel: Die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(\delta, x) := \delta \quad \text{auf} \\ M := \{(\delta, x) \in \mathbb{R} \times \mathbb{R}^n : \frac{1}{2}\|x - a_i\|_2^2 \leq \delta,\ i = 1, \ldots, m\} \end{cases}$$

(siehe das Sylvestersche Problem aus dem letzten Abschnitt) besitzt eine eindeutige Lösung (δ^*, x^*). Die Existenz folgt aus der Beobachtung, dass $M \neq \emptyset$ bzw. (P) zulässig und zugehörige Niveaumengen kompakt sind. Sind (δ_1^*, x_1^*) und (δ_2^*, x_2^*) zwei Lösungen von (P), so ist zunächst natürlich

$$\delta_1^* = \min_{(\delta,x)\in M} \delta = \delta_2^*.$$

Also ist $\delta^* := \delta_1^* = \delta_2^*$. Auch $(\delta^*, \frac{1}{2}(x_1^* + x_2^*))$ ist eine Lösung von (P). Da δ^* minimal, ist $\delta^* = \max_{j=1,\dots,m} \frac{1}{2}\|x^* - a_j\|_2^2$. Man wähle $i \in \{1, \dots, m\}$ mit

$$\frac{1}{2}\|x^* - a_i\|_2^2 = \max_{j=1,\dots,m} \frac{1}{2}\|x^* - a_j\|_2^2 = \delta^*.$$

Wir erhalten dann

$$\begin{aligned}
\delta^* &= \tfrac{1}{2}\|x^* - a_i\|_2^2 \\
&= \tfrac{1}{2}\|\tfrac{1}{2}(x_1^* - a_i) + \tfrac{1}{2}(x_2^* - a_i)\|_2^2 \\
&= \tfrac{1}{2}[\tfrac{1}{2}\|x_1^* - a_i\|_2^2 + \tfrac{1}{2}\|x_2^* - a_i\|_2^2 - \tfrac{1}{4}\|x_1^* - x_2^*\|_2^2] \\
&\qquad \text{(Anwendung der Parallelogrammgleichung)} \\
&\leq \tfrac{1}{2}[\delta_1^* + \delta_2^* - \tfrac{1}{4}\|x_1^* - x_2^*\|_2^2] \\
&= \delta^* - \tfrac{1}{8}\|x_1^* - x_2^*\|_2^2,
\end{aligned}$$

woraus $x_1^* = x_2^*$ und damit die Eindeutigkeit einer Lösung von (P) folgt. Hierbei besagt die einfach nachzuweisende Parallelogrammgleichung, dass $\|x + y\|_2^2 + \|x - y\|_2^2 = 2(\|x\|_2^2 + \|y\|_2^2)$. □

Eine weitere, für die Theorie außerordentlich wichtige, Fragestellung ist die folgende:

- Sei $x^* \in M$ eine lokale Lösung von (P). Welche Bedingungen müssen dann (unter geeigneten Glattheitsvoraussetzungen an die Zielfunktion f sowie die Restriktionsabbildungen g und h) *notwendigerweise* erfüllt sein, was sind also *notwendige Optimalitätsbedingungen*?

Ist (P) eine *unrestringierte Optimierungsaufgabe*, ist also $M = \mathbb{R}^n$ bzw. jeder Punkt des $\mathbb{R}^n$ zulässig für (P) (oder M offen), und ist f hinreichend glatt, so sind notwendige Optimalitätsbedingungen aus der Analysis wohlbekannt. Ist nämlich f in einem lokalen Extremum x^* partiell differenzierbar, existieren also die partiellen Ableitungen $(\partial f/\partial x_j)(x^*)$, $j = 1, \dots, n$, von f in x^*, so verschwindet notwendig der *Gradient* $\nabla f(x^*)$ von f in x^*, d. h. es ist

$$\nabla f(x^*) := \Big(\frac{\partial f}{\partial x_1}(x^*), \dots, \frac{\partial f}{\partial x_n}(x^*)\Big)^T = 0.$$

Ist f auf einer Umgebung eines lokalen Minimums x^* zweimal stetig partiell differenzierbar, existieren also auf einer Umgebung von x^* sämtliche partiellen Ableitungen $\partial^2 f/\partial x_i\,\partial x_j$, $1 \leq i, j \leq n$, und sind diese auf der Umgebung stetig, so gilt darüber hinaus. dass die *Hessesche* $\nabla^2 f(x^*)$ von f in x^*, also

$$\nabla^2 f(x^*) := \Big(\frac{\partial^2 f}{\partial x_i\,\partial x_j}\Big)_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n},$$

positiv semidefinit ist. Aber auch für restringierte Optimierungsaufgaben sind zumindestens notwendige Optimalitätsbedingungen erster Ordnung ebenfalls schon aus der Analysis bekannt. So sagt etwa die *Lagrangesche Multiplikatorenregel* aus:

∗ *Gegeben sei die Optimierungsaufgabe*

(P) *Minimiere* $f(x)$ *auf* $M := \{x \in \mathbb{R}^n : h(x) = 0\}$.

Ist $x^* \in M$ *eine lokale Lösung von* (P), *sind* $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ *und* $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *in* x^* *stetig partiell differenzierbar, sind ferner die Gradienten* $\nabla h_i(x^*)$, $i = 1, \ldots, m$, *der Komponenten* h_i *von* h *linear unabhängig, so existiert ein Vektor* $v^* = (v_i^*) \in \mathbb{R}^m$ *(die Komponenten heißen Lagrangesche Multiplikatoren) mit*

$$\nabla f(x^*) + h'(x^*)^T v^* = \nabla f(x^*) + \sum_{i=1}^{m} v_i^* \nabla h_i(x^*) = 0.$$

Hierbei ist

$$h'(x^*) := \Big(\frac{\partial h_i}{\partial x_j}(x^*)\Big)_{\substack{1\le i\le m\\ 1\le j\le n}} = \begin{pmatrix} \nabla h_1(x^*)^T \\ \vdots \\ \nabla h_m(x^*)^T \end{pmatrix} \in \mathbb{R}^{m\times n}$$

die Funktionalmatrix von h *in* x^*.

Beispiel: Das "optimale Dosenproblem" aus dem letzten Abschnitt lautet: Mit gegebenen positiven Werten c_1, c_2, V löse man die Aufgabe

$$\begin{cases} \text{Minimiere} \quad 2\pi r^2 c_1 + 2\pi r h c_2 \quad \text{unter den Nebenbedingungen} \\ \qquad\qquad \pi r^2 h = V, \quad r > 0, \quad h > 0. \end{cases}$$

Wir wollen die obige Lagrangesche Multiplikatorenregel anwenden und nehmen an, (r^*, h^*) sei eine lokale Lösung (die "offene" Nebenbedingung $r > 0$, $h > 0$ ist für die Anwendung der Lagrangeschen Multiplikatorenregel irrelevant). Hiernach existiert ein $v^* \in \mathbb{R}$ mit

$$(*) \qquad \begin{pmatrix} 4\pi r^* c_1 + 2\pi h^* c_2 \\ 2\pi r^* c_2 \end{pmatrix} + v^* \begin{pmatrix} 2\pi r^* h^* \\ \pi (r^*)^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Zusammen mit $\pi(r^*)^2 h^* = V$ hat man damit drei Gleichungen für die drei Unbekannten (r^*, h^*) und v^*. Mit Hilfe der zweiten Gleichung in $(*)$ erhält man

$$v^* = -\frac{2c_2}{r^*},$$

Einsetzen in die erste Gleichung liefert

$$h^* = \frac{2c_1}{c_2} r^*,$$

Einsetzen in die "Volumengleichung" ergibt die gesuchte Lösung

$$r^* = \Big(\frac{c_2 V}{2c_1 \pi}\Big)^{1/3}, \qquad h^* = \frac{2c_1}{c_2} r^*.$$

Natürlich hätte man dasselbe Ergebnis erhalten, wenn man die Höhe der gesuchten Dose durch $h = V/(\pi r^2)$ eliminiert und die unrestringierte Optimierungsaufgabe

$$\text{Minimiere} \quad f(r) := 2\pi r^2 c_1 + \frac{2c_2 V}{r}, \quad r > 0,$$

löst. Aus $f'(r^*) = 0$ (das meinten wir, als wir von Schulmathematik sprachen) erhalten wir wieder dieselbe Lösung. □

Einen Beweis der eben angegebenen Lagrangeschen Multiplikatorenregel findet man für $m = 1$, also einer Gleichung als Restriktion, z. B. bei O. Forster (1984, S. 78)[10]. Wir wollen hier schon versuchen, zugegebenermaßen verhältnismäßig unpräzise, einen Beweis für notwendige Optimalitätsbedingungen erster Ordnung, wie etwa die oben angegebene Lagrangesche Multiplikatorenregel bei Optimierungsaufgaben mit Gleichungen als Nebenbedingungen, anzudeuten. Sei hierzu die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$$

gegeben. Die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ seien in der lokalen Lösung $x^* \in M$ stetig partiell differenzierbar. Mit

$$I(x^*) := \{i \in \{1, \ldots, l\} : g_i(x^*) = 0\}$$

wird die Indexmenge der in x^* *aktiven Ungleichungsrestriktionen* bezeichnet. Die Menge $T(M; x^*)$ aller (Richtungen) $p \in \mathbb{R}^n$, zu denen es Folgen $\{t_k\} \subset \mathbb{R}_+$ und $\{r_k\} \subset \mathbb{R}^n$ mit $\{x^* + t_k p + r_k\} \subset M$ sowie $\lim_{k\to\infty} t_k = 0$ und $\lim_{k\to\infty} r_k/t_k = 0$ gibt, heißt *Tangentialkegel* an M in x^*. Unter einer (schwachen) Zusatzvoraussetzung, einer sogenannten *Constraint Qualification*, erwartet man, dass

$$(*) \quad L_0(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \leq 0 \quad (i \in I(x^*)),\ h'(x^*)p = 0\} \subset T(M; x^*)$$

gilt. Diese Aussage wollen wir uns für den Fall, dass keine Ungleichungen und nur eine Gleichung als Restriktion auftreten in Abbildung 1.6 verdeutlichen. Ist andererseits $p \in T(M; x^*)$ und sind $\{t_k\} \subset \mathbb{R}_+$ sowie $\{r_k\} \subset \mathbb{R}^n$ zugehörige Folgen, so ist $f(x^*) \leq f(x^* + t_k p + r_k)$ für alle hinreichend großen k, da $x^* \in M$ eine lokale Lösung von (P) ist. Folglich ist

$$\nabla f(x^*)^T p = \lim_{k\to\infty} \frac{f(x^* + t_k p + r_k) - f(x^*)}{t_k} \geq 0 \quad \text{für alle } p \in T(M; x^*).$$

Wegen $(*)$ ist $\nabla f(x^*)^T p \geq 0$ auch für alle $p \in L_0(M; x^*)$ und damit $p^* := 0$ eine Lösung der *linearen* Optimierungsaufgabe

$$\begin{cases} \text{Minimiere} \quad \nabla f(x^*)^T p \quad \text{auf} \\ L_0(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \leq 0 \quad (i \in I(x^*)),\ h'(x^*)p = 0\}. \end{cases}$$

[10] O. Forster (1984) Analysis 2. Vieweg-Verlag, Braunschweig-Wiesbaden.

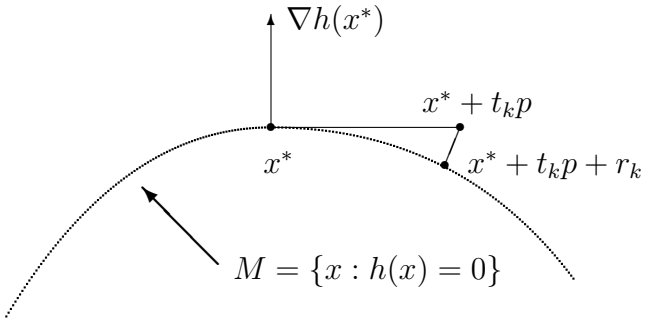

Abbildung 1.6: Eine Tangentialrichtung p in $x^* \in h^{-1}(0)$

Notwendige Optimalitätsbedingungen der linearen Optimierung (hierauf gehen wir im nächsten Kapitel ein) liefern dann die Existenz nichtnegativer, reeller Zahlen u_i^*, $i \in I(x^*)$, sowie eines Vektors $v^* \in \mathbb{R}^m$ mit

$$\nabla f(x^*) + \sum_{i \in I(x^*)} u_i^* \nabla g_i(x^*) + h'(x^*)^T v^* = 0.$$

Definiert man noch $u_i^* := 0$ für $i \in \{1, \ldots, l\} \setminus I(x^*)$, so erhält man in Vektor-Matrix-Schreibweise die Existenz eines Paares $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit:

(a) Es ist $u^* \geq 0$, d. h. die Multiplikatoren zu den Ungleichungsrestriktionen sind nichtnegativ.

(b) Es gilt die Lagrangesche Multiplikatorenregel

$$\nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0.$$

(c) Es gilt die sogenannte Gleichgewichtsbedingung $g(x^*)^T u^* = 0$, d. h. Multiplikatoren zu in x^* inaktiven Ungleichungsrestiktionen verschwinden.

Betont sei, dass der eben angedeutete „Beweis“ für diese Aussage zwei Lücken enthält. Zum einen ist die Gültigkeit von $(*)$ nicht gesichert. Im folgenden Beispiel zeigen wir, dass $(*)$ ohne eine Zusatzvoraussetzung nicht richtig ist.

Beispiel: Sei $M := \{x \in \mathbb{R}^2 : g_i(x) \leq 0 \ \ (i = 1, 2, 3)\}$, wobei die Restriktionabbildungen $g_i \colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ durch

$$g_1(x) := -x_2, \qquad g_2(x) := -x_1, \qquad g_3(x) := x_2 + (x_1 - 1)^3$$

definiert seien. Ferner sei $x^* := (1, 0)^T$. In Abbildung 1.7 skizzieren wir die Menge M. Offenbar ist $T(M; x^*) = \{p \in \mathbb{R}^2 : p_1 \leq 0, \ p_2 = 0\}$. In x^* sind die erste und die dritte Restriktion aktiv, so dass $I(x^*) = \{1, 3\}$ die Indexmenge der aktiven Restriktionen ist. Wegen

$$\nabla g_1(x^*) = (0, -1)^T, \qquad \nabla g_3(x^*) = (0, 1)^T$$

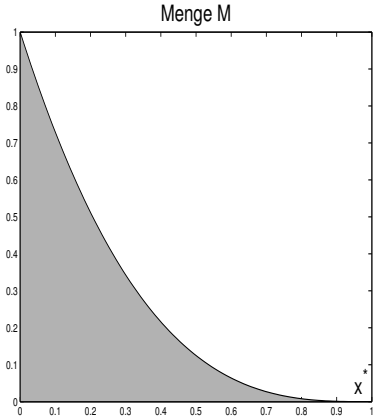

Abbildung 1.7: Die Menge $M := \{x \in \mathbb{R}^2 : g_i(x) \leq 0 \ \ (i = 1, 2, 3)\}$

ist daher

$$L_0(M; x^*) = \{p \in \mathbb{R}^2 : \nabla g_1(x^*)^T p \leq 0, \ \nabla g_3(x^*)^T p \leq 0\} = \{p \in \mathbb{R}^2 : p_2 = 0\}.$$

Damit ist in diesem Falle $L_0(M; x^*) \not\subset T(M; x^*)$ gezeigt. Ohne eine *Regularitätsbedingung* bzw. eine *Constraint Qualification* kann die Inklusion $L_0(M; x^*) \subset T(M; x^*)$ i. Allg. nicht bewiesen werden. □

Die zweite Lücke beim obigen „Beweis“ von notwendigen Optimalitätsbedingungen erster Ordnung bei einer nichtlinearen Optimierungsaufgabe besteht in der Anwendung einer (noch nicht bewiesenen) notwendigen Optimalitätsbedingung der *linearen* Optimierung. Auch diese wollen wir uns in einem Spezialfall schon einmal veranschaulichen. Um dies einfach im $\mathbb{R}^2$ zu verdeutlichen, nehmen wir an, es sei $p^* := (0, 0)^T$ eine Lösung der linearen Optimierungsaufgabe mit zwei (homogenen) Ungleichungen als Restriktionen:

$$\begin{cases} \text{Minimiere} \quad \nabla f(x^*)^T p \quad \text{auf} \\ L_0(M; x^*) := \{p \in \mathbb{R}^2 : \nabla g_1(x^*)^T p \leq 0, \ \nabla g_2(x^*)^T p \leq 0\}. \end{cases}$$

Durch die folgende Abbildung 1.8 wollen wir plausibel machen, dass dann notwendigerweise $-\nabla f(x^*)$ eine nichtnegative Linearkombination von $\nabla g_1(x^*)$ und $\nabla g_2(x^*)$ ist, also nichtnegative Zahlen u_1^* und u_2^* mit

$$\nabla f(x^*) + u_1^* \nabla g_1(x^*) + u_2^* \nabla g_2(x^*) = 0$$

existieren.

Weshalb sind notwendige Optimalitätsbedingungen in der Optimierung von besonderer Bedeutung? Die Antwort ist einfach. Ein zulässiger Punkt, in dem notwendige Optimalitätsbedingungen (erster oder gar zweiter Ordnung) erfüllt sind, ist zumindestens ein guter Kandidat für eine lokale Lösung der gegebenen Optimierungsaufgabe.

- Unter welchen Voraussetzungen ist eine lokale Lösung $x^* \in M$ der Optimierungsaufgabe

  (P) $\qquad$ Minimiere $\quad f(x) \quad$ auf $\quad M$

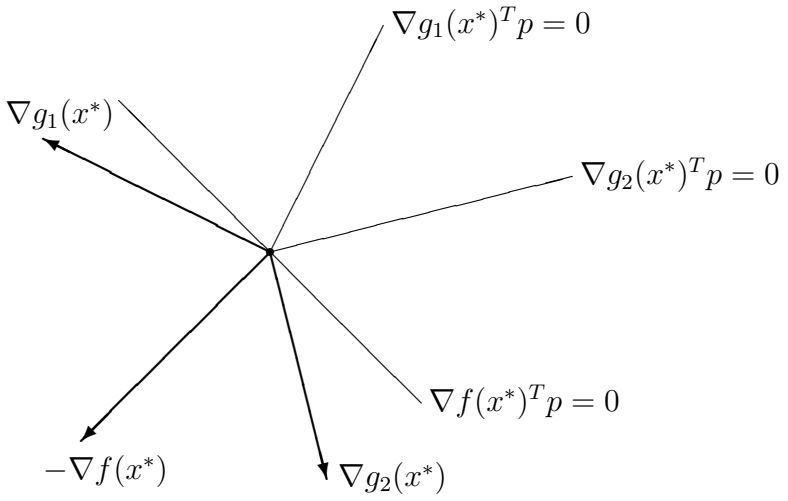

Abbildung 1.8: Eine (homogene) lineare Optimierungsaufgabe

sogar eine globale Lösung von (P)?

Eine sehr einfach zu beweisende, aber dennoch außerordentlich wichtige Antwort kann hierauf gegeben werden:

- *Ist die Menge $M \subset \mathbb{R}^n$ der zulässigen Lösungen konvex, ist ferner die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex auf M, so ist eine lokale Lösung $x^* \in M$ von* (P) *sogar eine globale Lösung.*

Zur Erinnerung: Eine Menge $M \subset \mathbb{R}^n$ heißt *konvex*, wenn

$$x,\, y \in M, \quad \lambda \in [0,1] \Longrightarrow (1-\lambda)x + \lambda y \in M,$$

wenn also mit je zwei Punkten aus M auch die gesamte Verbindungsstrecke zu M gehört. Entsprechend heißt eine reellwertige Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ *konvex* auf der konvexen Menge $M \subset \mathbb{R}^n$, wenn

$$x,\, y \in M, \quad \lambda \in [0,1] \Longrightarrow f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y).$$

Nun eine Begründung für die obige Aussage: Ist $x^* \in M$ eine lokale Lösung von (P), so existiert eine Umgebung U^* von x^* mit $f(x^*) \leq f(z)$ für alle $z \in M \cap U^*$. Zu einem beliebig vorgegebenem $x \in M$ existiert ein $\lambda \in (0,1]$ derart, dass $(1-\lambda)x^* + \lambda x \in M \cap U^*$, wobei die Konvexität von M ausgenutzt wurde. Da f auf M konvex ist, erhalten wir

$$f(x^*) \leq f((1-\lambda)x^* + \lambda x) \leq (1-\lambda)f(x^*) + \lambda f(x)$$

und hieraus $f(x^*) \leq f(x)$, so dass $x^* \in M$ sogar eine globale Lösung von (P) ist.

Ist die Menge M der zulässigen Lösungen durch ein System von Ungleichungen und Gleichungen in der Form

$$M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$$

gegeben, wobei $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$, so ist M konvex, wenn die Komponenten $g_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, l$, konvex (auf dem $\mathbb{R}^n$) sind, und h eine affine Abbildung ist, also durch $h(x) := Ax - b$ mit einer Matrix $A \in \mathbb{R}^{m\times n}$ und einem Vektor $b \in \mathbb{R}^m$ gegeben ist.

Konvexe Optimierungsaufgaben zeichnen sich nicht nur dadurch aus, dass lokale und globale Lösungen übereinstimmen, sondern auch dadurch, dass die oben angegebenen notwendigen Optimalitätsbedingungen sogar *hinreichend* für die Optimalität einer zulässigen Lösung $x^* \in M$ sind. Hilfsmittel zum Beweis ist die folgende einfache Aussage:

- *Die Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sei konvex und in $x^* \in \mathbb{R}^n$ stetig partiell differenzierbar. Dann ist*
$$\nabla f(x^*)^T(x - x^*) \le f(x) - f(x^*) \qquad \text{für alle } x \in \mathbb{R}^n.$$

Denn: Für alle $t \in (0, 1]$ ist

$$\frac{f(x^* + t(x - x^*)) - f(x^*)}{t} \le \frac{(1-t)f(x^*) + tf(x) - f(x^*)}{t} = f(x) - f(x^*),$$

mit $t \to 0+$ folgt die Behauptung. In Abbildung 1.9 veranschaulichen wir uns diese wichtige Eigenschaft konvexer Funktionen.

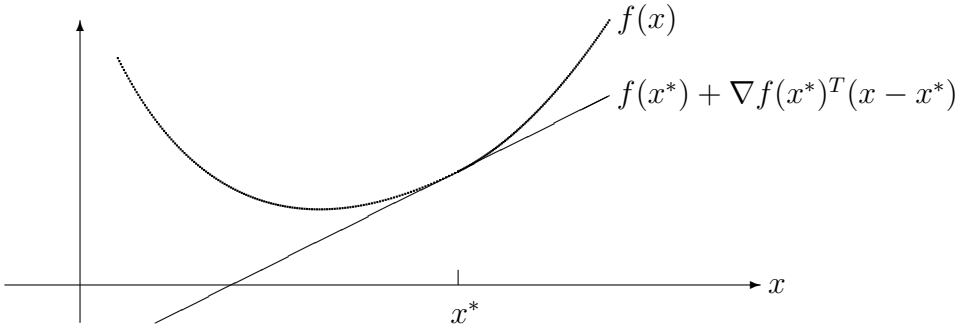

Abbildung 1.9: Eine konvexe Funktion

Nun die angekündigte Aussage:

- *Gegeben sei die Optimierungsaufgabe*

  (P) $\qquad$ *Minimiere* $\quad f(x) \quad$ *auf* $\quad M := \{x \in \mathbb{R}^n : g(x) \le 0,\ Ax = b\}.$

  Hierbei seien die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Komponenten $g_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ der Restriktionsabbildung $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ (auf dem $\mathbb{R}^n$) konvex und in $x^* \in M$ stetig partiell differenzierbar, $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$. Es existiere ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit:

$$u^* \ge 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + A^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

  Dann ist x^* eine (globale) Lösung von (P).

Denn: Ist $x \in M$ beliebig, so ist

$$\begin{aligned} f(x) - f(x^*) &\geq \nabla f(x^*)^T (x - x^*) \\ &= [g'(x^*)^T u^* + A^T v^*]^T (x^* - x) \\ &= \sum_{i=1}^{l} u_i^* \nabla g_i(x^*)^T (x^* - x) + [\underbrace{A(x^* - x)}_{=0}]^T v^* \\ &\geq \sum_{i=1}^{l} u_i^* [g_i(x^*) - g_i(x)] \\ &= \underbrace{g(x^*)^T u^*}_{=0} - \underbrace{g(x)^T u^*}_{\leq 0} \\ &\geq 0, \end{aligned}$$

womit die Behauptung bewiesen ist.

Beispiel: Wir betrachten die quadratische Optimierungsaufgabe

$$\left\{ \begin{array}{l} \text{Minimiere} \quad f(x) := \frac{1}{2} x^T \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} x + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}^T x \\ \qquad\qquad \text{unter den Nebenbedingungen} \\ \qquad\qquad x_1 + 2x_2 + x_3 \geq 4, \qquad x \geq 0. \end{array} \right.$$

Die Koeffizientenmatrix in der Zielfunktion ist positiv definit, ferner ist die Menge der zulässigen Lösungen nichtleer. Daher besitzt die Aufgabe genau eine Lösung x^*. Diese ist charakterisiert (wir benutzen hier schon, dass bei linearen Nebenbedingung keine Zusatzbedingung bzw. Constraint Qualification zur Aufstellung notwendiger Optimalitätbedingungen notwendig sind, wie wir später sehen werden) durch die Existenz von $u^* \in \mathbb{R}$ mit

$$u^* \geq 0, \qquad \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} x^* + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - u^* \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \geq 0$$

und den Gleichgewichtsbedingungen

$$u^*(x_1^* + 2x_2^* + x_3^* - 4) = 0$$

sowie

$$\left[\begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} x^* + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - u^* \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \right]^T x^* = 0.$$

Wir wollen zeigen, dass

$$x^* := \begin{pmatrix} \frac{7}{8} \\ \frac{11}{9} \\ \frac{7}{6} \end{pmatrix}$$

die Lösung ist, wobei der zugehörige Multiplikator durch $u^* := \frac{17}{18}$ gegeben ist. Denn es ist

$$\begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} x^* + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - u^* \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = 0$$

und

$$x_1^* + 2x_2^* + x_3^* - 4 = 0.$$

Alle Vorzeichenbedingungen sind also inaktiv, die Ungleichungsbedingung ist aktiv. □

Der *Wert* oder auch *Optimalwert* der Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M$$

ist definiert durch

$$\inf(\text{P}) := \begin{cases} \inf\{f(x) : x \in M\} & \text{für} \quad M \neq \emptyset, \\ +\infty & \text{für} \quad M = \emptyset. \end{cases}$$

Eine obere Schranke für $\inf(\text{P})$ erhält man trivialerweise, indem man die Zielfunktion in einer zulässigen Lösung auswertet. Schwieriger ist es, die folgende Frage zu beantworten:

- Wie erhält man untere Schranken für den Wert einer Minimierungsaufgabe?

Die Beantwortung dieser Frage kann interessant sein, um z. B. die mindestens zu erwartenden Kosten bei der Lösung einer Planungsaufgabe abzuschätzen. Wir nehmen an, die gegebene Optimierungsaufgabe habe die Form

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \in C,\ g(x) \leq 0,\ h(x) = 0\}.$$

Hier hat man sich die Nebenbedingung $x \in C$ als eine „einfache“ Restriktion vorzustellen. Z. B. ist $C = \{x \in \mathbb{R}^n : x \geq 0\}$ der sogenannte *nichtnegative Orthant* im $\mathbb{R}^n$, natürlich kann aber auch $C = \mathbb{R}^n$ sein. Wie bisher seien g und h Abbildungen des $\mathbb{R}^n$ in den $\mathbb{R}^l$ bzw. $\mathbb{R}^m$. Man definiere die *Lagrange-Funktion* $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ durch

$$L(x, u, v) := f(x) + g(x)^T u + h(x)^T v,$$

anschließend die zu (P) (Lagrange-) duale Optimierungsaufgabe (D) durch

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \phi(u, v) := \inf\limits_{x \in C} L(x, u, v) \quad \text{auf} \\ N := \{(u, v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \phi(u, v) > -\infty\}. \end{cases}$$

Trivialerweise gilt dann die Aussage des *schwachen Dualitätssatzes*:

- ∗ *Ist $x \in M$ zulässig für* (P) *und $(u, v) \in N$ zulässig für* (D), *so ist $\phi(u, v) \leq f(x)$ und damit* $\sup(\text{D}) \leq \inf(\text{P})$. *Sind ferner $x^* \in M$, $(u^*, v^*) \in N$ zulässig für* (P) *bzw.* (D) *und ist $\phi(u^*, v^*) = f(x^*)$, so ist x^* eine (globale) Lösung von* (P) *und (u^*, v^*) eine (globale) Lösung von* (D).

Denn: Für $x \in M$ und $(u, v) \in N$ ist

$$\phi(u,v) \le L(x,u,v) = f(x) + \underbrace{g(x)^T u}_{\le 0} + \underbrace{h(x)^T v}_{=0} \le f(x).$$

In jedem Falle (gleichgültig, ob (P) oder (D) zulässig sind) ist $\sup(\mathrm{D}) \le \inf(\mathrm{P})$, wobei der Wert $\sup(\mathrm{D})$ der Maximierungsaufgabe (D) naheliegenderweise durch

$$\sup(\mathrm{D}) := \begin{cases} \inf\{\phi(u,v) : (u,v) \in N\} & \text{für} \quad N \neq \emptyset, \\ -\infty & \text{für} \quad N = \emptyset \end{cases}$$

definiert ist. Für $x^* \in M$ und $(u^*, v^*) \in N$ ist daher

$$\phi(u^*, v^*) \le \sup(\mathrm{D}) \le \inf(\mathrm{P}) \le f(x^*).$$

Ist also $f(x^*) = \phi(u, v^*)$, so steht hier überall das Gleichheitszeichen, so dass x^* eine Lösung von (P) und (u^*, v^*) eine Lösung von (D) ist, ferner ist $\max(\mathrm{D}) = \min(\mathrm{P})$. Hierbei schreiben wir $\min(\mathrm{P})$ statt $\inf(\mathrm{P})$, wenn (P) eine Lösung besitzt, entsprechendes gilt für $\max(\mathrm{D})$.

I. Allg. tritt zwischen den beiden Optimierungsaufgaben (P) und (D) eine sogenannte *Dualitätslücke* auf, d. h. ohne weitere Voraussetzungen ist i. Allg. $\sup(\mathrm{D}) < \inf(\mathrm{P})$. Aussagen, die $\sup(\mathrm{D}) = \inf(\mathrm{P})$ (und eventuell die Lösbarkeit von (P) oder (D)) garantieren, nennt man *starke Dualitätssätze*.

Beispiel: Im letzten Abschnitt hatten wir das Sylvestersche Problem, zu m gegebenen Punkten $a_1, \ldots, a_m$ des $\mathbb{R}^n$ die kleinste sie enthaltende (euklidische) Kugel zu bestimmen, als Optimierungsaufgabe

$$(\mathrm{P}) \qquad \begin{cases} \text{Minimiere} \quad f(\delta, x) := \delta \quad \text{auf} \\ M := \{(\delta, x) \in \mathbb{R} \times \mathbb{R}^n : \frac{1}{2}\|x - a_i\|_2^2 \le \delta,\ i = 1, \ldots, m\} \end{cases}$$

formuliert. Wir wollen die hierzu duale Optimierungsaufgabe aufstellen. Die zugehörige Lagrange-Funktion ist

$$L(\delta, x, u) := \delta + \sum_{i=1}^{m} u_i (\tfrac{1}{2}\|x - a_i\|_2^2 - \delta).$$

Die duale Zielfunktion ist

$$\phi(u) := \inf_{(\delta,x) \in \mathbb{R} \times \mathbb{R}^n} L(\delta, x, u).$$

Bei gegebenem $u \ge 0$ ist offenbar $\phi(u) > -\infty$ genau dann, wenn $\sum_{i=1}^m u_i = 1$. Die Menge der dual zulässigen Lösungen ist also durch

$$N := \{u \in \mathbb{R}^m : u \ge 0,\ e^T u = 1\}$$

gegeben, wobei $e := (1, \ldots, 1)^T$. Für $u \in N$ ist

$$\begin{aligned} \phi(u) &= \inf_{x \in \mathbb{R}^n} \Big(\sum_{i=1}^m u_i \tfrac{1}{2} \|x - a_i\|_2^2 \Big) \\ &= \tfrac{1}{2} \sum_{i=1}^m u_i \Big\| \sum_{j=1}^m u_j a_j - a_i \Big\|_2^2 \\ &= \tfrac{1}{2} \sum_{i=1}^m u_i \|a_i\|_2^2 - \tfrac{1}{2} \Big\| \sum_{i=1}^m u_i a_i \Big\|_2^2. \end{aligned}$$

Das duale Problem lautet also

$$\text{(D)} \qquad \begin{cases} \text{Minimiere} \quad \phi(u) := \tfrac{1}{2} \displaystyle\sum_{i=1}^m u_i \|a_i\|_2^2 - \tfrac{1}{2} \Big\| \sum_{i=1}^m u_i a_i \Big\|_2^2 \quad \text{auf} \\ N := \{u \in \mathbb{R}^m : u \geq 0,\ e^T u = 1\}. \end{cases}$$

Dieses duale Problem ist insofern leichter als das Ausgangsproblem (P), als in ihm die Restriktionsmenge ein Simplex ist, insbesondere die Restriktionen also linear sind. Wir werden später sehen, dass man mit Hilfe einer Lösung u^* von (D) leicht die Lösung x^* von (P) erhält. Wird dieser Zusammenhang schon erraten? Wegen $\phi(u) \leq \min(\mathrm{P})$ für alle $u \in N$ erhält man weiter untere Schranken für den minimalen Radius $r^* = \sqrt{2 \min(\mathrm{P})}$ einer Umkugel zu den gegebenen Punkten $a_1, \ldots, a_m$. Setzt man z. B. $u := (1/m)e$ (hierbei ist e einmal wieder der Vektor, diesmal aus dem $\mathbb{R}^m$, dessen Komponenten alle gleich 1 sind), so erhält man

$$\phi(u) = \frac{1}{2m} \Big(\sum_{i=1}^m \|a_i\|_2^2 - \frac{1}{m} \Big\| \sum_{i=1}^m a_i \Big\|_2^2 \Big) \leq \min(\mathrm{P}) = \frac{(r^*)^2}{2}$$

bzw.

$$\frac{1}{m} \Big(\sum_{i=1}^m \|a_i\|_2^2 - \frac{1}{m} \Big\| \sum_{i=1}^m a_i \Big\|_2^2 \Big) \leq (r^*)^2.$$

Die linke Seite dieser Ungleichung ist zumindestens nichtnegativ, die Aussage also nicht ganz trivial, da mit einer Anwendung der Dreiecksungleichung und der Cauchy-Schwarzschen Ungleichung

$$\Big\| \sum_{i=1}^m a_i \Big\|_2^2 \leq \Big(\sum_{i=1}^m \|a_i\|_2 \Big)^2 \leq m \sum_{i=1}^m \|a_i\|_2^2.$$

Ist z. B. $m = 3$ und $n = 2$, ferner $a_1 = (-1, 0)$, $a_2 = (0, 1)$ und $a_3 = (1, 0)$ (siehe Abbildung 1.10)

Nach der obigen Formel erhalten wir $\sqrt{8}/3 \approx 0.9428 \leq r^*$. Dies wollen wir mit dem exakten Wert vergleichen und erinnern hierzu an einige aus der ebenen Geometrie (vielleicht) vorhandenen Vorkenntnisse. Bezeichnet man mit $\alpha = \angle(A)$ den Winkel bei A, entsprechend β und γ, so ist der Umkugelradius r^* gegeben durch

$$r^* = \frac{s}{4 \cos(\alpha/2) \cos(\beta/2) \cos(\gamma/2)},$$

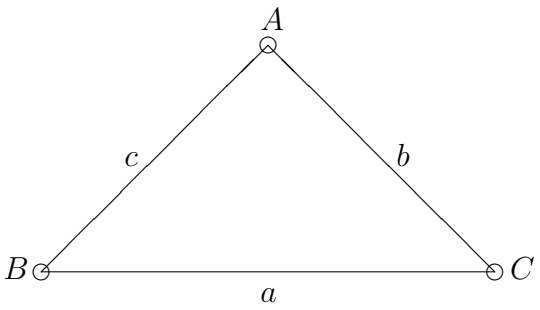

Abbildung 1.10: Radius der Umkugel?

wobei $s := \frac{1}{2}(a+b+c)$ den halben Umfang des Dreiecks $\triangle ABC$ angibt. Bekannt ist vielleicht auch noch die Aussage des Sinussatzes, dass nämlich

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma} = 2r^*.$$

In unserem Fall ist $\alpha = \pi/2$, weiter $\beta = \gamma = \pi/4$ sowie $a = 2$ und $b = c = \sqrt{2}$. Durch irgendeine dieser Formel erhält man für unser spezielles Dreieck den Umkugelradius $r^* = 1$. So ganz schlecht ist die obige Abschätzung also nicht. Gibt es eine Vermutung, wann die Abschätzung für $m = 3$ und $n = 2$ optimal ist? □

Eine weitere wichtige Fragestellung ist die folgende:

- Gegeben sei eine zulässige Lösung x^* einer Optimierungsaufgabe (P). Was sind *hinreichende* Bedingungen dafür, dass x^* eine lokale oder sogar globale Lösung von (P) ist? Hierbei sollten die hinreichenden Bedingungen "möglichst nahe" bei notwendigen Optimalitätbedingungen liegen.

Einige Antworten hierauf können wegen vorhandener Vorkenntnisse aus der Analysis oder der oben gemachten Bemerkungen jetzt schon leicht gegeben werden. Ist z. B. die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in einem Punkt $x^* \in \mathbb{R}^n$ zweimal stetig partiell differenzierbar, ist ferner $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv definit, so ist bekanntlich (siehe z. B. O. Forster (1984, S. 61)) x^* eine *isolierte lokale Lösung* der (unrestringierten) Optimierungsaufgabe, f auf dem $\mathbb{R}^n$ zu minimieren, d. h. es existiert eine Umgebung U^* von x^* mit $f(x^*) < f(x)$ für alle $x \in U^*$ mit $x \neq x^*$. Diese Aussage werden wir später auf restringierte Optimierungsaufgaben übertragen. Wie wir auf Seite 17 zeigten, sind unter Konvexitätsvoraussetzungen notwendige Optimalitätbedingungen auch hinreichend für globale Optimalität. Auch der schwache Dualitätssatz gibt eine hinreichende Optimalitätsbedingung an.

Beispiel: Es ist selten, dass für nichtkonvexe Optimierungsaufgaben notwendige und hinreichende Optimalitätsbedingungen übereinstimmen. Einen angenehmen Sonderfall nehmen hier die bei Trust-Region-Verfahren auftretenden Hilfsprobleme ein. Wie aus der unrestringierten Optimierung (vielleicht) bekannt ist, gilt nämlich die Aussage:

∗ *Gegeben sei die Aufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x, \quad \|x\|_2 \leq \Delta,$$

wobei $\Delta > 0$, $c \in \mathbb{R}^n$ und die symmetriche (nicht notwendig positiv semidefinite) Matrix $Q \in \mathbb{R}^{n\times n}$ gegeben sind. Dann ist ein $x^* \in \mathbb{R}^n$ mit $\|x^*\|_2 \leq \Delta$ genau dann eine globale Lösung von (P), wenn ein $\lambda^* \geq 0$ mit

(a) $(Q + \lambda^* I)x^* = -c$,

(b) $\lambda^*(\|x^*\|_2 - \Delta) = 0$,

(c) $Q + \lambda^* I$ *ist positiv semidefinit.*

Hierbei ist der "hinreichende Teil" einfach. Existiert nämlich zu einem für (P) zulässiges x^* ein $\lambda^* \geq 0$ mit den Eigenschaften (a)–(c), ist ferner $x \in \mathbb{R}^n$ ein beliebiger für (P) zulässiger Punkt, so ist

$$\begin{aligned}
f(x) - f(x^*) &= \nabla f(x^*)^T(x - x^*) + \frac{1}{2}(x - x^*)^T\nabla^2 f(x^*)(x - x^*) \\
&= \underbrace{(c + Qx^*)}_{=-\lambda^* x^*}{}^T(x - x^*) + \frac{1}{2}(x - x^*)^T Q(x - x^*) \\
&= -\lambda^*(x^*)^T(x - x^*) + \frac{1}{2}\underbrace{(x - x^*)^T(Q + \lambda^* I)(x - x^*)}_{\geq 0} - \frac{\lambda^*}{2}\|x - x^*\|_2^2 \\
&\geq -\lambda^*(x^*)^T(x - x^*) - \frac{\lambda^*}{2}\|x - x^*\|_2^2 \\
&= \frac{\lambda^*}{2}(\|x^*\|_2^2 - \|x\|_2^2) \\
&= \frac{\lambda^*}{2}(\Delta^2 - \|x\|_2^2) \\
&\geq 0,
\end{aligned}$$

also x^* eine Lösung von (P). Man kann (nicht nur bezogen auf den vorliegenden Fall) feststellen, dass der Beweis hinreichender Optimalitätsbedingungen eigentlich immer wesentlich einfacher als der von notwendigen Optimalitätsbedingungen ist. □

Die folgende Frage wird uns bei weitem am meisten beschäftigen.

- Gegeben sei die Optimierungsaufgabe

  (P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : g(x) \leq 0,\, h(x) = 0\}$.

  Wie berechnet man eine lokale oder globale Lösung von (P) bzw. wenigstens eine Näherung hierfür?

Die Wahl eines Verfahrens zur numerischen Lösung einer gegebenen Optimierungsaufgabe wird entscheidend von deren Struktur abhängen. Wie in anderen Bereichen der numerischen Mathematik kann man nicht hoffen, durch ein „Superverfahren" alle möglichen Aufgabenstellungen effizient zu lösen. So wird etwa die Dimension des Problems (Anzahl der Variablen und Nebenbedingungen) eine Rolle spielen, ferner ob Ableitungen analytisch zur Verfügung stehen. Weiter ist klar, dass bis auf einige Sonderfälle (lineare und geeignete quadratische Optimierungsaufgaben) eine Lösung nicht in endlich vielen Schritten berechnet werden kann.

1.3 Übersicht

Zunächst wollen wir sagen, worauf wir *nicht* eingehen werden. Am Anfang wurde schon betont, dass wir uns auch bei theoretischen Aussagen auf endlichdimensionale Optimierungsaufgaben beschränken werden, obwohl mit funktionalanalytischen Hilfmitteln weitgehende Übertragungen auf infinite Optimierungsaufgaben möglich sind. Nicht eingehen werden wir ferner auf die numerische Behandlung *unrestringierter Optimierungsaufgaben*, bei denen also die Menge M der zulässigen Lösungen der gesamte $\mathbb{R}^n$ (oder offen) ist. Auch die numerische Lösung *linearer Programme* (hier sind bekanntlich sowohl die Zielfunktion f als auch die Restriktionsabbildungen g und h affin linear) durch das bekannte *Simplexverfahren* wird nicht behandelt. Auch auf Fragen der *globalen Optimierung*, in der versucht wird, eine oder gar alle globalen Lösungen einer Optimierungsaufgabe zu berechnen, werden wir nicht eingehen können. Nicht behandeln werden wir ferner kombinatorische (und ganzzahlige) Optimierungsaufgaben, bei denen die Menge M der zulässigen Lösungen eine endliche (aber i. Allg. aus sehr vielen Elementen bestehende) Menge ist.

Stattdessen werden die folgenden Themen eine Rolle spielen:

- Theoretische Grundlagen.

Hier werden wir zunächst auf Trennungssätze für konvexe Mengen im $\mathbb{R}^n$ (u. a. wird das Farkas-Lemma bewiesen), dann auf Dualität bei konvexen Programmen eingehen. Es wird das Lagrange-duale Programm zu einem gegebenen konvexen Programm gebildet und gezeigt, dass man hierdurch untere Schranken für den Optimalwert gewinnen kann (schwache Dualität, diesen einfachen Sachverhalt schilderten wir schon im letzten Abschnitt), ferner wird untersucht, wann keine Dualitätslücke auftritt. Auf die Anwendung der allgemeinen Theorie auf lineare Programme werden wir nur sehr kurz eingehen, dafür aber ein Ergebnis beweisen, welches in nur wenigen Lehrbüchern über lineare Optimierung zu finden ist, dass nämlich zu einem linearen Programm ein strikt komplementäres optimales Paar existiert, wenn es überhaupt eine Lösung besitzt. Im letzten Abschnitt in diesem ersten Kapitel leiten wir die notwendigen und hinreichenden Optimalitätsbedingungen erster und zweiter Ordnung bei glatten, nichtlinearen Optimierungsaufgaben her. In der gesamten Vorlesung gehen wir i. Allg. davon aus, dass die auftretenden Zielfunktionen bzw. Restriktionsabbildungen glatt, also wenigstens stetig differenzierbar, sind. Daher werden auch konvexe (nicht notwendig glatte) Funktionen und ihre Eigenschaften (Existenz von Subgradienten usw.) kaum untersucht.

- Quadratische Optimierungsaufgaben.

Als einfachste (restringierte) nichtlineare Optimierungsaufgabe kann die Aufgabe angesehen werden, eine (konvexe) quadratische Funktion unter linearen Nebenbedingungen (bzw. auf einem Polyeder) zu minimieren. Im ersten Abschnitt wird relativ ausführlich das duale Verfahren von Goldfarb-Idnani angegeben und analysiert. Auch über die numerische Implementation werden wir einiges sagen. Anschließend werden primale Verfahren der "aktiven Mengen" untersucht, z. B. ein Verfahren von Fletcher. Bei einem primalen Verfahren wird eine Folge von zulässigen Näherungen mit monoton

abnehmenden (oder wenigstens nicht zunehmenden) Kosten bestimmt. Wie beim Simplexverfahren muss hier also eventuell eine Phase vorgeschaltet werden, in welcher eine zulässige Ausgangslösung bestimmt wird. Diese Verfahren sind nur für nicht zu hochdimensionale Probleme geeignet. Am Schluss wollen wir ein Verfahren von Coleman-Li schildern, das für hochdimensionale quadratische Programme mit sogenannten box-constraints (die Restriktionen haben hier die Form $l \leq x \leq u$) geeignet ist und z. B. in MATLAB implementiert ist.

- Linear restringierte Optimierungsaufgaben.

Nach den quadratischen Optimierungsaufgaben ist das Problem, eine nichtlineare (und i. Allg. auch nichtquadratische) Funktion auf einem Polyeder zu minimieren, die nächst schwierigere Aufgabe. Hier kann man nicht auf einen endlichen Algorithmus hoffen, d. h. wir werden Algorithmen haben, die eine nicht abbrechende Folge von Näherungslösungen erzeugen, von der wir zu zeigen haben, dass sie "angenehme" Konvergenzeigenschaften besitzt. Zunächst behandeln wir die Methode der aktiven Mengen, bei denen die Vorgehensweise der primalen Verfahren der quadratischen Optimierung simuliert und eine Folge von Optimierungsaufgaben mit linearen Gleichungen als Restriktionenmenge gelöst wird. Danach werden Verfahren der zulässigen Richtungen untersucht. Grob kann man hier sagen, dass man die aus der unrestringen Optimierung her bekannten Begriffe und Methoden (z. B. Abstiegsrichtung, Schrittweitenstrategie, Newton- und Quasi-Newton-Verfahren) auf die vorliegende Situation zu übertragen versucht.

- Nichtlinear restringierte Optimierungsaufgaben.

Die schwierigsten nichtlinearen Optimierungsaufgaben haben auch nichtlineare Restriktionen. Die Zulässigkeit von Näherungslösungen wird i. Allg. nicht gesichert werden können. Wir werden auch hier nur glatte Aufgaben betrachten und annehmen, dass zumindestens der Gradient der Zielfunktion und die Gradienten der Restriktionsabbildungen analytisch zur Verfügung stehen. Es wird auf quadratische und exakte, nichtdifferenzierbare Straffunktionen eingegangen, ferner wird die Idee der SQP (sequential quadratic programming) Verfahren angegeben.

- Innere-Punkt-Verfahren.

Wir wollen die Idee dieser zur Zeit sehr viel untersuchten Verfahren schildern und auf einige neuere Ergebnisse eingehen. Seit der bahnbrechenden Arbeit von Karmarkar erscheinen über dreißig Jahre alte Verfahren in einem neuen Licht. Hier werden wir uns aber auf die Untersuchung von konvexen, insbesondere konvexen, quadratisch restringierten quadratischen Programmen beschränken.

1.4 Aufgaben

1. Gegeben sei die konvexe Optimierungsaufgabe

   (P) $\qquad\qquad$ Minimiere $\quad f(x) \quad$ auf $\quad M,$

   d. h. die Menge $M \subset \mathbb{R}^n$ der zulässigen Lösungen von (P) sei konvex, die Zielfunktion $f\colon M \longrightarrow \mathbb{R}$ sei auf M konvex. Man zeige:

(a) Die Menge M_{opt} der (globalen) Lösungen von (P) ist konvex.

(b) Ist $f\colon M \longrightarrow \mathbb{R}$ auf M sogar *strikt konvex*, gilt also die Implikation

$$x, y \in M, \quad x \neq y, \quad \lambda \in (0,1) \Longrightarrow f((1-\lambda)x + \lambda y) < (1-\lambda)f(x) + \lambda f(y),$$

so besteht die Menge M_{opt} der Lösungen von (P) aus höchstens einem Punkt.

(c) Sei (P) zulässig (d. h. $M \neq \emptyset$), M abgeschlossen und f auf M stetig. Dann gilt:

i. Existiert ein $x_0 \in M$ derart, dass die Niveaumenge $L_0 := \{x \in M : f(x) \leq f(x_0)\}$ kompakt ist, so ist M_{opt} nichtleer und kompakt.

ii. Ist M_{opt} nichtleer und kompakt, so ist die Niveaumenge $L_0 := \{x \in M : f(x) \leq f(x_0)\}$ für jedes $x_0 \in M$ kompakt.

2. Sei $M \subset \mathbb{R}^n$ konvex und $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf einer offenen Obermenge von M stetig differenzierbar. Man zeige:

(a) f ist genau dann auf M konvex, wenn

$$\nabla f(x)^T (y - x) \leq f(y) - f(x) \qquad \text{für alle } x, y \in M.$$

(b) Ist f auf M konvex, so ist ein $x^* \in M$ genau dann eine Lösung der konvexen Optimierungsaufgabe, f auf M zu minimieren, wenn $\nabla f(x^*)^T (x - x^*) \geq 0$ für alle $x \in M$.

3. Sei $M \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex, $z \in \mathbb{R}^n$ vorgegeben. Dann besitzt die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad \|x - z\|_2 \quad \text{auf} \quad M$$

genau eine Lösung x^*. Ferner ist ein $x^* \in M$ genau dann eine Lösung von (P), wenn $(x^* - z)^T (x - x^*) \geq 0$ für alle $x \in M$.

Hinweis: Es handelt sich hier um den *Projektionssatz für konvexe Mengen*. Die Existenz einer Lösung zeige man mit Hilfe der Kompaktheit von Niveaumengen, die Eindeutigkeit durch die strikte Konvexität von $f(x) := \frac{1}{2}\|x - z\|_2^2$, schließlich führe man die Charakterisierung einer Lösung auf eine Aussage in Aufgabe 2 zurück.

4. Man betrachte die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{j=1}^{n} x_j \ln \frac{x_j}{p_j} \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : e^T x = 1,\ x \geq 0\}.$$

Hierbei sei $e := (1, \ldots, 1)^T \in \mathbb{R}^n$, die positiven reellen Zahlen $p_1, \ldots, p_n$ seien vorgegeben. Ferner ist natürlich $0 \ln 0$ durch 0 definiert. Man zeige, dass (P) eine eindeutige Lösung x^* besitzt. Anschließend überlege man sich, dass $x^* > 0$ bzw. x^* nur positive Komponenten besitzt. Mit Hilfe der Lagrangeschen Multiplikatorenregel berechne man x^*.

5. Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch $f(x) := c^T x + \frac{1}{2} x^T Q x$ mit symmetrischem $Q \in \mathbb{R}^{n \times n}$ definiert. Dann ist $\inf_{x \in \mathbb{R}^n} f(x) > -\infty$ genau dann, wenn Q positiv semidefinit ist und ein $x^* \in \mathbb{R}^n$ mit $\nabla f(x^*) = 0$ existiert.

6. Gegeben sei das zweiseitig quadratisch restringierte quadratische Programm

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := c_0^T x + \frac{1}{2} x^T Q_0 x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : \alpha_i \le g_i(x) := c_i^T x + \frac{1}{2} x^T Q_i x \le \beta_i,\ i = 1, \ldots, m\}. \end{cases}$$

   Hierbei seien $Q_0, Q_1, \ldots, Q_m \in \mathbb{R}^{n \times n}$ symmetrisch, $\alpha_i \le \beta_i$, $i = 1, \ldots, m$. Dann gilt:

   (a) Ist (P) zulässig und existieren $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ derart, dass $Q_0 + \sum_{i=1}^m \lambda_i Q_i$ positiv definit ist, so besitzt (P) eine Lösung.

   (b) Existiert zu $x^* \in M$ ein Vektor $\lambda^* = (\lambda_i^*) \in \mathbb{R}^m$ mit

   - $\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) = 0$,
   - $\lambda_i^*(\alpha_i - g_i(x^*)) \le 0 \le \lambda_i^*(g_i(x^*) - \beta_i)$, $i = 1, \ldots, m$,
   - $Q_0 + \sum_{i=1}^m \lambda_i^* Q_i$ ist positiv semidefinit,

   so ist x^* eine globale Lösung der (i. Allg. nichtkonvexen) Optimierungsaufgabe (P). Ist $Q_0 + \sum_{i=1}^m \lambda_i^* Q_i$ sogar positiv definit, so ist x^* eindeutige Lösung von (P).

   Hinweis: Sie beweisen eine Verallgemeinerung eines Teils von Theorem 2.1 bei R. J. Stern, H. Wolkowicz (1995))[11].

7. Gegeben seien $c \in \mathbb{R}^n \setminus \{0\}$, die symmetrische, positiv definite Matrix $Q \in \mathbb{R}^{n \times n}$ sowie $x_0 \in \mathbb{R}^n$. Hiermit betrachte man die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : (x - x_0)^T Q (x - x_0) \le 1\}.$$

   Man zeige, dass (P) eine eindeutige Lösung $x^* \in M$ besitzt und bestimme diese.

8. Beim Maximalflussproblem ist ein Netzwerk $(\mathcal{N}, \mathcal{A})$ mit zwei ausgezeichneten Knoten q (Quelle) und s (Senke) gegeben, ferner nichtnegative Kapazitäten u_{ij} auf den Pfeilen $(i,j) \in \mathcal{A}$. Ein *Fluss* $x = (x_{ij})_{(i,j) \in \mathcal{A}}$ heißt *zulässig*, wenn er den Kapazitätsbeschränkungen

$$0 \le x_{ij} \le u_{ij}, \qquad (i,j) \in \mathcal{A},$$

   und der Flussgleichung genügt. Diese besagt, dass in jedem Knoten außer der Quelle und Senke genau so viel Fluss ankommt wie auch wieder abtransportiert wird, also

$$\sum_{j:(k,j) \in \mathcal{A}} x_{kj} - \sum_{i:(i,k) \in \mathcal{A}} x_{ik} = 0, \qquad k \in \mathcal{N} \setminus \{q, s\},$$

   gilt. Unter diesen Bedingungen ist der Fluss $\sum_{j:(q,j) \in \mathcal{A}} x_{qj}$ zu maximieren. Ein *Schnitt* im Netzwerk eine Partition der Knotenmenge $\mathcal{N}$ (bzw. $\{1, \ldots, m\}$) in zwei (disjunkte) Mengen $\mathcal{N}_1$ und $\mathcal{N}_2$ mit $q \in \mathcal{N}_1$ und $s \in \mathcal{N}_2$. Zu einem Schnitt $(\mathcal{N}_1, \mathcal{N}_2)$ definieren wir die zugehörige *Kapazität* $C(\mathcal{N}_1, \mathcal{N}_2)$ als die Summe aller Kapazitätsschranken über Pfeilen, die in $\mathcal{N}_1$ starten und in $\mathcal{N}_2$ enden, also in der oben eingeführten Notation durch

$$C(\mathcal{N}_1, \mathcal{N}_2) := \sum_{\substack{(i,j) \in \mathcal{A} \\ i \in \mathcal{N}_1,\, j \in \mathcal{N}_2}} u_{ij}.$$

[11] R. J. Stern, H. Wolkowicz (1995) Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. SIAM J. Optim. 5, 286–313.

Man zeige: Ist $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ ein zulässiger Fluss und $(\mathcal{N}_1, \mathcal{N}_2)$ ein Schnitt mit zugehöriger Kapazität $C(\mathcal{N}_1, \mathcal{N}_2)$, so ist

$$\sum_{j:(q,j)\in\mathcal{A}} x_{qj} \leq C(\mathcal{N}_1, \mathcal{N}_2).$$

Besteht hier sogar Gleichheit, so ist x ein maximaler Fluss (und $(\mathcal{N}_1, \mathcal{N}_2)$ ein minimaler Schnitt). Mit dieser Aussage bestimme man in dem in der folgenden Abbildung angegebenen Netzwerk einen maximalen Fluss und einen minimalen Schnitt.

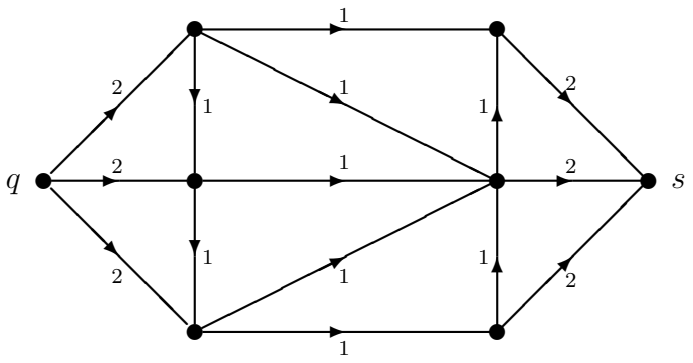

Abbildung 1.11: Maximaler Fluss, minimaler Schnitt?

9. Seien $a_1, \ldots, a_m \in \mathbb{R}^n$ mit $\|a_i\|_2 = 1$, $i = 1, \ldots, m$, und $b_1, \ldots, b_m \in \mathbb{R}$ gegeben. Die Menge

$$P := \{x \in \mathbb{R}^n : a_i^T x \leq b_i,\ (i = 1, \ldots, m)\}$$

sei nichtleer und beschränkt. Man zeige: Ist $(x^*, r^*) \in \mathbb{R}^n \times \mathbb{R}$ eine Lösung der linearen Optimierungsaufgabe

$$\text{Maximiere} \quad r \quad \text{auf} \quad M := \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : r \geq 0,\ a_i^T x + r \leq b_i\ (i = 1, \ldots, m)\},$$

so ist $B[x^*; r^*] := \{y \in \mathbb{R}^n : \|y - x^*\|_2 \leq r^*\}$ die größte (euklidische) Kugel (d. h. die Kugel mit maximalem Radius), die in P enthalten ist. Also kann man die Inkugel zu einem Polytop (kompakter Polyeder) durch Lösen eines linearen Programms bestimmen.

10. Gegeben seien m paarweise verschiedene Punkte $a_1, \ldots, a_m$ im $\mathbb{R}^n$, positive Gewichte $w_1, \ldots, w_m$ und eine nichtleere, konvexe und abgeschlossene Menge $M \subset \mathbb{R}^n$. Hiermit betrachte man das sogenannte *Fermat-Weber Problem*

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{i=1}^{m} w_i \|x - a_i\|_2 \quad \text{auf} \quad M,$$

wobei $\|\cdot\|_2$ natürlich die euklidische Norm auf dem $\mathbb{R}^n$ bedeutet. Man zeige:

   (a) Die Optimierungsaufgabe (P) besitzt mindestens eine (globale) Lösung.

   (b) Sind die gegebenen Punkte $a_1, \ldots, a_m$ nicht kollinear, liegen sie also nicht alle auf einer Geraden, so ist (P) sogar eindeutig lösbar.

11. Man löse das folgende, auf S. Lhulier (1782) zurückgehende geometrische Problem: Die Längen a_1 bzw. a_2 der Grundlinien zweier Dreiecke sowie die Summe l der Längen ihrer vier Schenkel seien gegeben, wobei natürlich $l > a_1 + a_2$ vorausgesetzt sei. Unter allen Paaren von Dreiecken mit diesen Eigenschaften bestimme man dasjenige, für welches die Summe der Flächeninhalte der beiden Dreiecke maximal ist. Für $a_1 = 1$, $a_2 = 2$ und $l = 5$ berechne man numerisch die Länge der gesuchten Schenkel.

Kapitel 2

Theoretische Grundlagen

In diesem Kapitel sollen die im weiteren benötigten Grundlagen bereitgestellt werden. Hier handelt es sich vor allem um Trennungssätze für konvexe Mengen, die Dualitätstheorie der konvexen, insbesondere der linearen Optimierung und notwendige (eventuell auch hinreichende) Optimalitätsbedingungen erster und zweiter Ordnung.

2.1 Trennung konvexer Mengen im $\mathbb{R}^n$

Die Standardreferenz für (fast) alle Ergebnisse der konvexen Analysis (Untersuchung konvexer Mengen und Funktionen) im $\mathbb{R}^n$ ist R. T. Rockafellar (1972)[1].

2.1.1 Definitionen, Projektionssatz, starker Trennungssatz

Im folgenden bedeute $\|\cdot\|$ stets die euklidische Norm im $\mathbb{R}^n$. Hyperebenen im $\mathbb{R}^n$ sind mit $(y, \gamma) \in (\mathbb{R}^n \setminus \{0\}) \times \mathbb{R}$ durch

$$H := \{x \in \mathbb{R}^n : y^T x = \gamma\}$$

gegeben. In Abbildung 2.1 wird dies mit einem $\gamma > 0$ veranschaulicht.

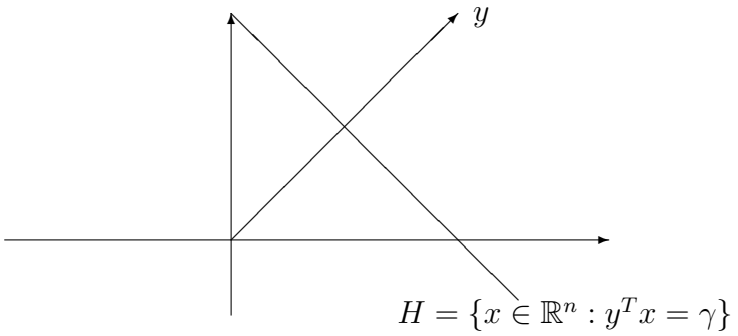

Abbildung 2.1: Hyperebene

Wir definieren:

[1] R. T. Rockafellar (1972) *Convex Analysis.* Princeton University Press, Princeton.

Definition 1.1 Seien $A, B \subset \mathbb{R}^n$ nichtleere Teilmengen.

(a) A und B heißen *trennbar*, wenn $y \in \mathbb{R}^n \setminus \{0\}$ mit

$$\sup_{a \in A} y^T a \leq \inf_{b \in B} y^T b$$

existiert.

(b) A und B heißen *echt trennbar*, wenn $y \in \mathbb{R}^n \setminus \{0\}$ mit

$$\sup_{a \in A} y^T a \leq \inf_{b \in B} y^T b, \qquad \inf_{a \in A} y^T a < \sup_{b \in B} y^T b$$

existiert.

(c) A und B heißen *strikt trennbar*, wenn $(y, \gamma) \in (\mathbb{R}^n \setminus \{0\}) \times \mathbb{R}$ mit

$$y^T a < \gamma < y^T b \qquad \text{für alle } a \in A,\, b \in B$$

existiert.

(d) A und B heißen *stark trennbar*, wenn $y \in \mathbb{R}^n \setminus \{0\}$ mit

$$\sup_{a \in A} y^T a < \inf_{b \in B} y^T b$$

existiert.

Bemerkung: In den Fällen (a), (b) und (d) definiere man

$$\gamma := \tfrac{1}{2}[\sup_{a \in A} y^T a + \inf_{b \in B} y^T b].$$

In allen vier Fällen sei die Hyperebene H durch

$$H := \{x \in \mathbb{R}^n : y^T x = \gamma\}$$

gegeben. Diese Hyperebene induziert zwei (abgeschlossene) Halbräume, nämlich

$$H^- := \{x \in \mathbb{R}^n : y^T x \leq \gamma\}, \qquad H^+ := \{x \in \mathbb{R}^n : y^T x \geq \gamma\}.$$

Diese Halbräume haben jeweils ein nichtleeres Inneres, nämlich

$$\operatorname{int}(H^-) = \{x \in \mathbb{R}^n : y^T x < \gamma\}, \qquad \operatorname{int}(H^+) = \{x \in \mathbb{R}^n : y^T x > \gamma\}.$$

Dann gelten die folgenden vier Aussagen, die man sich jeweils durch eine Skizze veranschaulichen sollte.

(a) Sind A und B trennbar, so existiert eine Hyperebene H mit $A \subset H^-$ und $B \subset H^+$.

(b) Sind A und B echt trennbar, so existiert eine Hyperebene H mit $A \subset H^-$, $B \subset H^+$ und $A \cup B \not\subset H$.

(c) Sind A und B strikt trennbar, so existiert eine Hyperebene H mit $A \subset \text{int}\,(H^-)$ und $B \subset \text{int}\,(H^+)$.

(d) Sind A und B stark trennbar, so existiert eine Hyperebene H und ein $\epsilon > 0$ mit $A + B[0;\epsilon] \subset H^-$ und $B + B[0;\epsilon] \subset H^+$. Hierbei bedeutet $B[0;\epsilon]$ die (euklidische) ϵ-Kugel um den Nullpunkt. Anschaulich bedeutet dies, dass man um A und B jeweils einen (eventuell) schmalen Schlauch legen kann und die so vergrößerten Mengen immer noch trennbar sind.

Hier ist wohl nur der Nachweis von (d) nicht ganz offensichtlich. Man definiere

$$\epsilon := \frac{1}{2\,\|y\|}[\inf_{b\in B} y^T b - \sup_{a\in A} y^T a].$$

Mit beliebigen $a \in A$ und $x \in B[0;\epsilon]$ ist

$$y^T(a+x) \le \sup_{a\in A} y^T a + \|y\|\,\epsilon = \gamma$$

bzw. $A + B[0;\epsilon] \subset H^-$. Entsprechend ist $B + B[0;\epsilon] \subset H^+$. □

Es folgt der bekannte *Projektionssatz für konvexe Mengen*, den wir ohne Beweis angeben (siehe auch Aufgabe 3 im einleitenden Kapitel).

Satz 1.2 (Projektionssatz) *Sei $M \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex, $z \in \mathbb{R}^n$. Dann besitzt die Aufgabe*

(P) $$\textit{Minimiere} \quad f(x) := \|x - z\| \quad \textit{auf} \quad M$$

genau eine Lösung x^*, die sogenannte Projektion von z auf M. Ferner ist ein $x^* \in M$ genau dann eine Lösung von (P), wenn $(x^* - z)^T(x - x^*) \ge 0$ für alle $x \in M$.

Es folgt der *starke Trennungssatz* für konvexe Mengen im $\mathbb{R}^n$.

Satz 1.3 *Seien $A, B \subset \mathbb{R}^n$ nichtleer, konvex und abgeschlossen. Sind dann A und B disjunkt und eine der beiden Mengen kompakt, so sind A und B stark trennbar.*

Beweis: Sei $M := B - A$, wobei die Differenz der beiden Mengen B und A natürlich durch

$$B - A := \{b - a : a \in A,\, b \in B\}$$

definiert ist. Dann ist M nichtleer, konvex und abgeschlossen (Beweis?), ferner $0 \notin M$, da $A \cap B = \emptyset$. Sei $x^* \in M$ die wegen des Projektionssatzes existierende Projektion von $z := 0$ auf M. Insbesondere ist $(x^*)^T x \ge \|x^*\|^2$ für alle $x \in M$ und $x^* \ne 0$. Definiert man daher $y := x^*$, so ist

$$y^T(b - a) \ge \|x^*\|^2 \qquad \text{für alle } a \in A,\, b \in B$$

und daher

$$\sup_{a\in A} y^T a < \sup_{a\in A} y^T a + \|x^*\|^2 \le \inf_{b\in B} y^T b.$$

Also sind A und B stark trennbar. □ □

Das folgende Korollar ist eine wichtige, unmittelbare Folgerung aus dem starken Trennungssatz.

Korollar 1.4 *Sei $K \subset \mathbb{R}^n$ nichtleer, konvex und abgeschlossen. Dann kann jedes $z \notin K$ von K stark getrennt werden, d. h. zu jedem $z \notin K$ existiert ein $y \in \mathbb{R}^n \setminus \{0\}$ mit $y^T z < \inf_{x \in K} y^T x$.*

2.1.2 Farkas-Lemma, Trennungssätze

Ziel in diesem Abschnitt ist es, das berühmte Farkas-Lemma (1902) und als Folgerung hieraus, zwei weitere Trennungssätze für konvexe Mengen zu beweisen. Wir werden den Beweis des Farkas-Lemmas so führen, dass wir zunächst die Abgeschlossenheit sogenannter *endlich erzeugter Kegel*[2] nachweisen und anschließend den starken Trennungssatz anwenden. Den hübschen Beweis des folgenden wichtigen Lemmas haben wir bei J. B. Hiriart-Urruty, C. Lemaréchal (1993, S. 130)[3] gefunden.

Lemma 1.5 *Sei $A = (\ a_1\ \cdots\ a_n\ ) \in \mathbb{R}^{m \times n}$. Dann ist der von $a_1, \ldots, a_n$ erzeugte Kegel $K := \{Ax : x \geq 0\}$ abgeschlossen.*

Beweis: Durch vollständige Induktion nach n zeigen wir, dass ein von n Elementen $a_1, \ldots, a_n \in \mathbb{R}^m$ erzeugter Kegel abgeschlossen ist. Dies ist für $n = 1$ offensichtlich richtig. Wir nehmen an, die Aussage sei für Kegel mit weniger als n Erzeugenden richtig. Weiter sei K ein von n Elementen $a_1, \ldots, a_n \in \mathbb{R}^m$ erzeugter konvexer Kegel, also

$$K = \Big\{ \sum_{j=1}^{n} x_j a_j : x_j \geq 0 \ \ (j = 1, \ldots, n) \Big\}.$$

Sind $a_1, \ldots, a_n$ linear unabhängig, so ist K offensichtlich abgeschlossen[4]. Daher können wir jetzt annehmen, dass ein $z \in \mathbb{R}^n \setminus \{0\}$ mit $\sum_{j=1}^{n} z_j a_j = 0$ existiert. O. B. d. A. existiert ein $j \in \{1, \ldots, n\}$ mit $z_j < 0$ (andernfalls gehe man zu $-z$ über). Wir wollen uns überlegen, dass

$$K = \bigcup_{j=1}^{n} \operatorname{cone} \{a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_n\},$$

dass sich K also als Vereinigung von Kegeln mit weniger als n Erzeugenden darstellen lässt. Aus der Induktionsannahme folgt dann die Behauptung. Zu zeigen ist offenbar nur, dass sich jedes Element aus K als nichtnegative Linearkombination von weniger als n der $a_1, \ldots, a_n$ darstellen lässt. Hierzu geben wir uns ein beliebiges $y = \sum_{j=1}^{n} x_j a_j \in K$, o. B. d. A. $x_j > 0$, $j = 1, \ldots, n$, vor. Sei

$$\min \Big\{ -\frac{x_j}{z_j} : z_j < 0 \Big\} = -\frac{x_{j(x)}}{z_{j(x)}} =: t^*(x).$$

[2] Unter einem *Kegel* versteht man eine Menge mit der Eigenschaft, dass eine Halbgerade (Strahl) vom Nullpunkt durch einen beliebigen Punkt der Menge ganz in der Menge liegt. Formal: Eine Menge K heißt *Kegel*, wenn aus $x \in K$ und $\lambda \geq 0$ folgt, dass $\lambda x \in K$.

[3] J. B. Hirriart-Urruty and C. Lemaréchal (1993) *Convex Analysis and Minimization Algorithms.* Springer-Verlag, Berlin.

[4] Denn ist $K = \{Ax : x \geq 0\}$, wobei $A \in \mathbb{R}^{m \times n}$ den vollen Rang n hat, so ist $A^T A \in \mathbb{R}^{n \times n}$ insbesondere nichtsingulär. Aus $\{Ax_k\} \subset K$ und $Ax_k \to y$ folgt daher $x_k \to (A^T A)^{-1} A^T y \geq 0$. Da weiter Bild (A) abgeschlossen ist, ist $y = Ax \in$ Bild (A), und folglich $x_k \to x \geq 0$. Also ist $y = Ax \in K$.

Mit

$$\hat{x}_j := x_j + t^*(x)z_j, \qquad j = 1, \ldots, n,$$

ist dann $\hat{x}_j \geq 0$, $j = 1, \ldots, n$, und $\hat{x}_{j(x)} = 0$ und daher

$$y = \sum_{j=1}^{n} x_j a_j = \sum_{j=1}^{n} (x_j + t^*(x)z_j)a_j = \sum_{\substack{j=1 \\ j \neq j(x)}}^{n} \hat{x}_j a_j.$$

Damit ist der Induktionsschluss vollständig und der Beweis der Abgeschlossenheit abgeschlossen. □
□

Nun ist es nicht schwierig, das Farkas-Lemma in seiner "Basis-Version" zu beweisen.

Lemma 1.6 (Farkas) *Seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^n$ gegeben. Dann besitzt das System*

$$\text{(I)} \qquad Ax = b, \qquad x \geq 0$$

genau dann keine Lösung, wenn das System

$$\text{(II)} \qquad A^T y \geq 0, \qquad b^T y < 0$$

eine Lösung besitzt.

Beweis: Wir nehmen zunächst an, (I) und (II) hätten Lösungen $x \in \mathbb{R}^n$ bzw. $y \in \mathbb{R}^m$. Dann wäre $0 > b^T y = (Ax)^T y = x^T A^T y \geq 0$, ein Widerspruch. Nun nehmen wir an, (I) sei nicht lösbar. Dann ist $b \notin K := \{Ax : x \geq 0\}$. Wegen des vorangegangenen Lemmas wissen wir, dass der (endlich erzeugte) Kegel K abgeschlossen ist. Der starke Trennungssatz (angewandt auf $\{b\}$ und K) liefert die Existenz eines $y \in \mathbb{R}^m \setminus \{0\}$ mit $b^T y < \inf_{x \geq 0} y^T Ax$. Hieraus folgt, dass y eine Lösung von (II) ist. □ □

Aus dem Farkas-Lemma erhält man leicht weitere sogenannte *Alternativsätze*. Ist z. B. das System

$$\text{(I)} \qquad Ax \leq b$$

nicht lösbar, so ist auch das System

$$\text{(I}') \qquad \begin{pmatrix} A & -A & I \end{pmatrix} \begin{pmatrix} x_+ \\ x_- \\ z \end{pmatrix} = b, \qquad \begin{pmatrix} x_+ \\ x_- \\ z \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

nicht lösbar. Das Farkas-Lemma liefert, dass

$$\text{(II}') \qquad \begin{pmatrix} A^T \\ -A^T \\ I \end{pmatrix} y \geq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad b^T y < 0$$

bzw.

$$\text{(II)} \qquad A^T y = 0, \qquad y \geq 0, \qquad b^T y < 0$$

lösbar ist. Die Idee hierbei war, das System (I) durch Einführung einer *Schlupfvariablen* und die Darstellung $x = x_+ - x_-$ mit nichtnegativen Vektoren x_+ sowie x_- auf das äquivalente System (I') zurückzuführen. Ähnlich kann man auch in anderen Situationen vorgehen. Es liegt daher nahe, nach einer Verallgemeinerung des Farkas-Lemmas zu fragen, welche alle diese Fälle enthält.

Definition 1.7 Eine Menge $P \subset \mathbb{R}^n$, die sich in der Form

$$P = \{x \in \mathbb{R}^n : Ax \leq b\}$$

mit $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ darstellen lässt, heißt ein *Polyeder*.

Eine Menge $C \subset \mathbb{R}^n$, die sich in der Form

$$C = \{x \in \mathbb{R}^n : U^T x \geq 0\}$$

mit einer Matrix $U \in \mathbb{R}^{n \times m}$ darstellen lässt, heißt ein *polyedrischer Kegel*.

Ist $C \subset \mathbb{R}^n$, so heißt

$$C^+ := \{z \in \mathbb{R}^n : z^T x \geq 0 \quad \text{für alle } x \in C\}$$

der zu C *duale Kegel*.

Die gesuchte Verallgemeinerung des Farkas-Lemmas geben wir nun an.

Lemma 1.8 *Seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ gegeben. Ferner seien $C \subset \mathbb{R}^n$ und $K \subset \mathbb{R}^m$ polyedrische Kegel, deren duale Kegel mit C^+ bzw. K^+ bezeichnet seien. Dann besitzt das System*

$$\text{(I)} \qquad b - Ax \in K, \qquad x \in C$$

genau dann keine Lösung, wenn das System

$$\text{(II)} \qquad A^T y \in C^+, \qquad y \in K^+, \qquad b^T y < 0$$

lösbar ist.

Beweis: Angenommen, (I) und (II) seien beide lösbar durch ein x bzw. ein y. Dann wäre

$$0 > b^T y = (\underbrace{b - Ax}_{\in K} + Ax)^T \underbrace{y}_{\in K^+} = \underbrace{y^T(b - Ax)}_{\geq 0} + \underbrace{(A^T y)^T x}_{\geq 0} \geq 0,$$

ein Widerspruch.

Da C und K polyedrische Kegel im R^n bzw. R^m sind, existieren Matrizen $U \in \mathbb{R}^{n \times k}$ und $V \in \mathbb{R}^{m \times l}$ mit

$$C = \{x \in \mathbb{R}^n : U^T x \geq 0\}, \qquad K = \{y \in \mathbb{R}^m : V^T y \geq 0\}.$$

Wir nehmen nun an, (I) besitze keine Lösung. Dies impliziert, dass

$$V^T(b - Ax) \geq 0, \qquad U^T x \geq 0$$

bzw.

$$\text{(I}'\text{)} \qquad \begin{pmatrix} V^TA & -V^TA & I & 0 \\ -U^T & U^T & 0 & I \end{pmatrix} \begin{pmatrix} x_+ \\ x_- \\ u \\ v \end{pmatrix} = \begin{pmatrix} V^Tb \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} x_+ \\ x_- \\ u \\ v \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

nicht lösbar ist. Das Farkas-Lemma 1.6 zeigt, dass

$$\text{(II}'\text{)} \qquad \begin{pmatrix} A^TV & -U \\ -A^TV & U \\ I & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} V^Tb \\ 0 \end{pmatrix}^T \begin{pmatrix} p \\ q \end{pmatrix} < 0$$

lösbar ist. Daher existieren nichtnegative p, q mit

$$A^T \underbrace{Vp}_{\in K^+} = \underbrace{Uq}_{\in C^+}, \qquad b^TVp < 0.$$

Setzt man also $y := Vp$, so ist y eine Lösung von (II). □ □

Zum Schluss dieses Unterabschnittes beweisen wir noch zwei Trennungssätze für konvexe Mengen im $\mathbb{R}^n$.

Satz 1.9 *Zwei nichtleere, disjunkte Polyeder im $\mathbb{R}^n$ sind stark trennbar.*

Beweis: Seien

$$P := \{x \in \mathbb{R}^n : Ax \leq b\}, \qquad Q := \{y \in \mathbb{R}^n : Cy \leq d\}$$

mit $A \in \mathbb{R}^{k\times n}$, $b \in \mathbb{R}^k$, $C \in \mathbb{R}^{m\times n}$ und $d \in \mathbb{R}^m$ zwei Polyeder. Wir zeigen, dass $P - Q$ abgeschlossen ist, woraus dann die Behauptung folgt (siehe den Beweis des starken Trennungssatzes). Hierzu sei $\{x_k\} \subset P$, $\{y_k\} \subset Q$ und $x_k - y_k \to z$. Angenommen, es wäre $z \notin P - Q$. Dann ist das Gleichungs-Ungleichungssystem

$$x - y = z, \qquad Ax \leq b, \qquad Cy \leq d$$

bzw.

$$\begin{pmatrix} z \\ b \\ d \end{pmatrix} - \begin{pmatrix} I & -I \\ A & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \in \{0\} \times \mathbb{R}^k_{\geq 0} \times \mathbb{R}^m_{\geq 0}$$

nicht lösbar. Das verallgemeinerte Farkas-Lemma liefert die Existenz von $(u, v, w) \in \mathbb{R}^n \times \mathbb{R}^k_{\geq 0} \times \mathbb{R}^m_{\geq 0}$ mit

$$\begin{pmatrix} I & A^T & 0 \\ -I & 0 & C^T \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} z \\ b \\ d \end{pmatrix}^T \begin{pmatrix} u \\ v \\ w \end{pmatrix} < 0$$

bzw.

$$u + A^Tv = 0, \quad -u + C^Tw = 0, \quad z^Tu + b^Tv + d^Tw < 0.$$

Wegen $Ax_k \leq b$, $Cy_k \leq d$ ist daher

$$\begin{aligned} 0 &> z^T u + b^T v + d^T w \\ &\geq z^T u + (Ax_k)^T v + (Cy_k)^T w \\ &= z^T u + x_k^T A^T v + y_k^T C^T w \\ &= z^T u - x_k^T u + y_k^T u \\ &= (z - (x_k - y_k))^T u \\ &\to 0 \quad \text{mit} \quad k \to \infty, \end{aligned}$$

ein Widerspruch. □ □

Der nächste Trennungssatz kann wesentlich verschärft werden (Anmerkungen hierzu machen wir im Anschluss). Der Satz sagt aus, dass sich zwei nichtleere, konvexe, disjunkte Teilmengen des $\mathbb{R}^n$ durch eine Hyperebene trennen lassen, schließt aber nicht aus, daß diese Hyperebene eine oder gar beide Mengen enthält. Den Beweis haben wir O. L. Mangasarian (1969, S. 47 ff.)[5] entnommen.

Satz 1.10 *Seien $A, B \subset \mathbb{R}^n$ nichtleer, konvex und disjunkt. Dann sind A und B trennbar.*

Beweis: Es ist $0 \notin C := B - A$, da A und B disjunkt sind, ferner ist C konvex. Wir zeigen die Existenz eines $y \in \mathbb{R}^n \setminus \{0\}$ mit $y^T x \geq 0$ für alle $x \in C$, woraus offenbar die Behauptung folgt.

Für $x \in C$ definieren wir

$$\Lambda_x := \{y \in \mathbb{R}^n : \|y\| = 1,\ y^T x \geq 0\},$$

eine nichtleere, abgeschlossene Teilmenge der kompakten Einheitssphäre. Wir wollen zeigen, dass $\bigcap_{x \in C} \Lambda_x \neq \emptyset$, denn ein Element aus diesem Durchschnitt ist der gesuchte Vektor y. Wegen der Kompaktheit der Einheitssphäre (sogenannte finite intersection property kompakter Mengen) genügt es zu zeigen: Sind $x_1, \ldots, x_m \in C$, so ist $\bigcap_{i=1}^m \Lambda_{x_i} \neq \emptyset$. Dies sieht man wiederum folgendermaßen ein. Angenommen, es wäre $\bigcap_{i=1}^m \Lambda_{x_i} = \emptyset$. Dann hätte das Ungleichungssystem $y^T x_i \geq 0$, $i = 1, \ldots, m$, keine nichttriviale Lösung. Mit $X := (\ x_1 \ \cdots \ x_m\ ) \in \mathbb{R}^{n \times m}$ und $e := (1, \ldots, 1)^T \in \mathbb{R}^m$ bedeutet dies, dass das Ungleichungssystem

$$X^T y \geq 0, \qquad (-Xe)^T y < 0$$

nicht lösbar ist. Das Farkas-Lemma 1.6 liefert die Existenz eines nichtnegativen Vektors $\lambda \in \mathbb{R}^m$ mit $X\lambda = -Xe$ bzw. $X(\lambda + e) = 0$. Also ist der Nullpunkt eine positive Linearkombination und dann auch ein Konvexkombination der Punkte $x_1, \ldots, x_m \in C$. Aus der Konvexität von C folgt $0 \in C$, was ein Widerspruch ist. □ □

Bemerkungen: Das *relative Innere* $\operatorname{ri}(A)$ einer Menge $A \subset \mathbb{R}^n$ ist definiert als

$$\operatorname{ri}(A) := \{x \in A : \text{Es existiert } \epsilon > 0 \text{ mit } B[x; \epsilon] \cap \operatorname{aff}(A) \subset A\}.$$

[5] O. L. Mangasarian (1969) *Nonlinear Programming.* McGraw-Hill Book Company, New York.

Hierbei ist (wie immer) $B[x;\epsilon]$ die abgeschlossene (euklidische) Kugel um x mit dem Radius ϵ, ferner bezeichnet $\operatorname{aff}(A)$ die *affine Hülle* von A, also den Durchschnitt aller affin linearen Teilräume des $\mathbb{R}^n$, die A enthalten. Dann gilt (siehe R. T. Rockafellar (1972, Theorem 11.3)):

- *Die nichtleeren, konvexen Mengen $A, B \subset \mathbb{R}^n$ sind genau dann echt trennbar, wenn* $\operatorname{ri}(A) \cap \operatorname{ri}(B) = \emptyset$.

Insbesondere erhält man als Verschärfung von Satz 1.10, dass zwei nichtleere, konvexe, disjunkte Teilmengen des $\mathbb{R}^n$ echt trennbar sind.

Ein weiterer interessanter (und nicht einfach zu beweisender) Trennungssatz ist das folgende Ergebnis (siehe R. T. Rockafellar (1972, Theorem 20.2)):

- *Seien $A, B \subset \mathbb{R}^n$ nichtleer, konvex, A sei sogar ein Polyeder. Dann sind A und B genau dann echt trennbar durch eine Hyperebene, die B nicht enthält, wenn* $A \cap \operatorname{ri}(B) = \emptyset$.

Nur bemerkt sei, dass das relative Innere $\operatorname{ri}(A)$ einer nichtleeren, konvexen Menge A selbst nichtleer und konvex ist. Ist ferner $A \subset \mathbb{R}^n$ konvex und $\operatorname{aff}(A) = \mathbb{R}^n$ (man sagt dann auch, die Menge A sei n-dimensional), so ist $\operatorname{ri}(A) = \operatorname{int}(A)$. □

2.1.3 Aufgaben

1. Sei $K \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex, ferner $P_K : \mathbb{R}^n \longrightarrow K \subset \mathbb{R}^n$ die zugehörige Projektionsabbildung. Man zeige:

   (a) Es ist
   $$\|P_K(x) - P_K(y)\| \leq \|x - y\| \qquad \text{für alle } x, y \in \mathbb{R}^n.$$

   (b) Ist $L \subset \mathbb{R}^n$ ein linearer Teilraum, so ist P_L eine lineare Abbildung und $x^T P_L(y) = P_L(x)^T y$ für alle $x, y \in \mathbb{R}^n$.

   (c) Ist $L := \operatorname{span}\{v_1, \ldots, v_p\}$ mit linear unabhängigen $v_1, \ldots, v_p \in \mathbb{R}^n$ und $V := \begin{pmatrix} v_1 & \cdots & v_p \end{pmatrix}$, so ist
   $$P_L(x) = V(V^T V)^{-1} V^T x \qquad \text{für alle } x \in \mathbb{R}^n.$$

2. Seien $l, u \in \mathbb{R}^n$ zwei Vektoren mit $l \leq u$. Hiermit definiere man den Quader
   $$Q := \{x \in \mathbb{R}^n : l \leq x \leq u\}.$$
   Man zeige, dass für $x \in \mathbb{R}^n$ die Projektion $P_Q(x)$ von x auf Q durch
   $$P_Q(x)_j = \begin{cases} l_j, & \text{falls} \quad x_j < l_j, \\ x_j, & \text{falls} \quad l_j \leq x_j \leq u_j, \\ u_j, & \text{falls} \quad u_j < x_j, \end{cases} \qquad j = 1, \ldots, n,$$
   gegeben ist.

3. Zwei nichtleere, konvexe Mengen $A, B \subset \mathbb{R}^n$ sind genau dann stark trennbar, wenn $0 \notin \operatorname{cl}(B - A)$.

4. Sei $C \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex mit nichtleerem Inneren $\operatorname{int}(C)$. Man zeige, dass es zu jedem $x^* \in C \setminus \operatorname{int}(C)$ ein $y \in \mathbb{R}^n \setminus \{0\}$ mit

$$C \subset \{x \in \mathbb{R}^n : y^T x \geq y^T x^*\}$$

gibt.

Hinweis: Man zeige, dass mit C auch $\operatorname{int}(C)$ konvex ist und wende auf $\{x^*\}$ und $\operatorname{int}(C)$ den Trennungssatz an. Anschließend zeige man, dass $C = \operatorname{cl}(\operatorname{int}(C))$.

5. Eine nichtleere, abgeschlossene, konvexe Menge $C \subset \mathbb{R}^n$ ist der Durchschnitt aller abgeschlossenen Halbräume, die C enthalten.

Hinweis: Man wende den starken Trennungssatz an.

6. Sei $C \subset \mathbb{R}^n$ ein nichtleerer, abgeschlossener, konvexer Kegel. Dann ist $(C^+)^+ = C$. Eine (dumme) Zusatzfrage: Kann Gleichheit auch gelten, wenn C nicht abgeschlossen, nicht konvex oder kein Kegel ist?

Hinweis: Man überzeuge sich davon, dass die Inklusion $C \subset (C^+)^+$ trivial ist. Mit Hilfe des starken Trennungssatzes zeige man anschließend, dass aus $z \notin C$ auch $z \notin (C^+)^+$ folgt.

7. Man zeige, dass jeder endlich erzeugte Kegel sich als dualer Kegel eines polyedrischen Kegels darstellen läßt. Genauer zeige man: Ist $U \in \mathbb{R}^{n \times m}$, so ist

$$\{Uy : y \geq 0\} = \{x \in \mathbb{R}^n : U^T x \geq 0\}^+.$$

8. Sei $A \in \mathbb{R}^{m \times n}$. Man beweise den Alternativsatz von Gordan: Genau eine der beiden Aussagen

(I) $\qquad Ax = 0, \quad x \geq 0, \quad x \neq 0 \qquad$ hat eine Lösung $x \in \mathbb{R}^n$

bzw.

(II) $\qquad A^T y > 0 \qquad$ hat eine Lösung $y \in \mathbb{R}^m$

ist richtig.

9. Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Man beweise den Alternativsatz von Gale: Genau eine der beiden Aussagen

(I) $\qquad Ax \leq b \qquad$ hat eine Lösung $x \in \mathbb{R}^n$

bzw.

(II) $\qquad A^T y = 0, \quad y \geq 0, \quad b^T y < 0 \qquad$ hat eine Lösung $y \in \mathbb{R}^m$

ist richtig.

10. Von A. Dax (1997)[6] stammt ein "elementarer" Beweis des Farkas-Lemmas. Wir wollen die Quintessenz dieses Arguments wiedergeben. Gegeben seien also wieder $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und hiermit die Systeme

(I) $\qquad Ax = b, \qquad x \geq 0$

[6] A. Dax (1997) An elementary proof of Farkas' Lemma. SIAM Rev. 39, 503–507.

und

(II) $$A^T y \geq 0, \qquad b^T y < 0.$$

Man zeige der Reihe nach:

(a) Die Optimierungsaufgabe

(P) $$\text{Minimiere} \quad f(x) := \frac{1}{2}\|Ax - b\|_2^2, \quad x \geq 0$$

besitzt eine Lösung x^*.

(b) Ist (I) nicht lösbar bzw. $y^* := Ax^* - b \neq 0$, so ist y^* eine Lösung von (II).

11. Man beweise den folgenden Satz von Fan-Glicksburg-Hoffman (siehe O. L. Mangasarian (1969, S. 63) und R. T. Rockafellar (1972, S. 186 ff.)):

Sei $C \subset \mathbb{R}^n$ nichtleer und konvex, die Abbildung $g\colon C \longrightarrow \mathbb{R}^l$ (komponentenweise) konvex, die Abbildung $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear. Besitzt dann

(I) $$x \in C, \quad g(x) < 0, \quad h(x) = 0$$

keine Lösung, so besitzt

(II) $$(u, v) \in \mathbb{R}^l \times \mathbb{R}^m \setminus \{(0,0)\}, \quad u \geq 0, \quad \inf_{x \in C}[u^T g(x) + v^T h(x)] \geq 0$$

eine Lösung.

Hinweis: Besitzt (I) keine Lösung, so ist

$$(0,0) \notin \{(g(x) + z, h(x)) \in \mathbb{R}^l \times \mathbb{R}^m : x \in C,\ z > 0\}.$$

Man überzeuge sich davon, dass die rechtsstehende Menge konvex ist und wende den Trennungsatz für konvexe Mengen an.

12. Man beweise die folgende Variante zum Satz von Fan-Glicksburg-Hoffman (siehe O. L. Mangasarian (1969, S. 65)):

Sei $C \subset \mathbb{R}^n$ nichtleer und konvex, die Abbildung $g\colon C \longrightarrow \mathbb{R}^l$ (komponentenweise) konvex. Dann ist genau eine der Aussagen

(I) $$\text{Es existiert } x \in C \text{ mit } g(x) < 0$$

bzw.

(II) $$\text{Es existiert } u \in \mathbb{R}^l \setminus \{0\} \text{ mit } u \geq 0 \text{ und } \inf_{x \in C} u^T g(x) \geq 0$$

richtig.

13. Man beweise: Ist $A \subset \mathbb{R}^n$ nichtleer und konvex, so ist $\operatorname{ri}(A) \neq \emptyset$ (siehe z. B. J.-B. Hirriart-Urruty, C. Lemaréchal (1993, S. 103) oder auch R. T. Rockafellar (1972, Theorem 6.2)).

2.2 Dualität bei konvexen Programmen

Dualität ist eines der wichtigsten Konzepte der Optimierung. Hierbei wird einem gegebenen (primalen) Minimierungs-Programm ein sogenanntes duales Programm zugeordnet. Dieses besteht darin, eine gewisse (duale) Zielfunktion auf der Menge der dual zulässigen Lösungen zu maximieren, wobei der duale Zielfunktionswert in einem beliebigen dual zulässigen Punkt nicht größer ist als der primale Zielfunktionswert in einer beliebigen primal zulässigen Lösung. Hierdurch kann der Optimalwert des Ausgangsproblems von unten angenähert werden. Im Idealfall ist dieses duale Programm (wenigstens in gewisser Hinsicht) einfacher als das Ausgangsproblem und hat die Eigenschaft, dass man aus einer Lösung eine des eigentlich interessierenden erhalten kann. Wir werden die Dualitätstheorie der linearen Optimierung nur sehr kurz streifen und z. B. auf ihre ökonomische Interpretation nicht näher eingehen. Ferner beschränken wir uns in diesem Abschnitt auf die Untersuchung des sogenannten *Lagrange-dualen* Programms und werden erst dann, wenn wir etwas über notwendige Optimalitätsbedingungen bei glatten, konvexen Optimierungsaufgaben wissen, auch auf das *Wolfe-duale* Programm wenigstens in den Aufgaben eingehen.

2.2.1 Definition des dualen Programms

Wir betrachten im folgenden eine Optimierungsaufgabe der Form

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \in C,\ g(x) \leq 0,\ h(x) = 0\}.$$

Hierbei wird i. Allg. vorausgesetzt:

(V) $C \subset \mathbb{R}^n$ ist nichtleer und konvex, $f\colon C \longrightarrow \mathbb{R}$ und $g\colon C \longrightarrow \mathbb{R}^l$ sind (komponentenweise) konvex, $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ist affin linear.

Unter der Voraussetzung (V) handelt es sich bei (P) um ein *konvexes Programm*, d. h. sowohl die Zielfunktion als auch die Menge der zulässigen Lösungen von (P) ist konvex.

Die zu (P) gehörende *Lagrange-Funktion* $L\colon C \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ ist durch

$$L(x, u, v) := f(x) + u^T g(x) + v^T h(x)$$

definiert. Schließlich ist das zu (P) Lagrange-duale Programm gegeben durch

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \phi(u, v) := \inf\limits_{x \in C} L(x, u, v) \quad \text{auf} \\ N := \{(u, v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \phi(u, v) > -\infty\}. \end{cases}$$

Bemerkung: Treten in (P) keine Gleichungen als Restriktionen auf, so werden in der Definition der Lagrange-Funktion bzw. des dualen Programms die entsprechenden Variablen bzw. Terme weggelassen. Auch jede Voraussetzung, die sich auf nichtvorhandene Gleichungen als Restriktionen bezieht, ist natürlich irrevalent. □

Beispiel: Man betrachte speziell ein lineares Programm in Normalform, also

$$\text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b\}.$$

Mit $f(x) := c^T x$, $C := \mathbb{R}^n_{\geq 0}$, $h(x) := b - Ax$ (implizite Ungleichungen bzw. entsprechende Terme in der Lagrange-Funktion treten nicht auf) ist die Zielfunktion des dualen Programms

$$\phi(v) = \inf_{x \geq 0}[b^T v + x^T(c - A^T v)] = \begin{cases} b^T v, & \text{falls} \quad A^T v \leq c, \\ -\infty, & \text{sonst.} \end{cases}$$

Als duales Programm erhält man also, wie aus der linearen Optimierung gewohnt, die Aufgabe

$$\text{Maximiere} \quad b^T v \quad \text{auf} \quad N := \{v \in \mathbb{R}^m : A^T v \leq c\}.$$

Das obige Dualitätskonzept ist also konsistent mit dem aus der linearen Optimierung bekannten. □

Der (Optimal) Wert des Programms (P) ist definiert durch

$$\inf(\mathrm{P}) := \begin{cases} \inf\limits_{x \in M} f(x), & \text{falls} \quad M \neq \emptyset, \\ +\infty, & \text{falls} \quad M = \emptyset. \end{cases}$$

Wir schreiben $\min(\mathrm{P})$ statt $\inf(\mathrm{P})$, falls (P) eine Lösung besitzt. Entsprechend ist der Wert des dualen Programms (D) durch

$$\sup(\mathrm{D}) := \begin{cases} \sup\limits_{(u,v) \in N} \phi(u, v), & \text{falls} \quad N \neq \emptyset, \\ -\infty, & \text{falls} \quad N = \emptyset \end{cases}$$

definiert. Entsprechend wie oben schreiben wir $\max(\mathrm{D})$ statt $\sup(\mathrm{D})$, wenn (D) lösbar ist.

Es folgt nun der (triviale) schwache Dualitätssatz, in dem die Konvexitätsvoraussetzung (V) noch keine Rolle spielt.

Satz 2.1 *Gegeben sei das Programm (P) und das dazu duale Programm (D). Dann gilt:*

1. *Ist $x \in M$ und $(u, v) \in N$, so ist $\phi(u, v) \leq f(x)$. Insbesondere ist* $\sup(\mathrm{D}) \leq \inf(\mathrm{P})$.
2. *Ist $x^* \in M$ und $(u^*, v^*) \in N$ mit $\phi(u^*, v^*) = f(x^*)$, so ist x^* eine Lösung von (P) und (u^*, v^*) eine Lösung von (D).*

Beweis: Für $x \in M$ und $(u, v) \in N$ ist

$$\phi(u, v) \leq L(x, u, v) = f(x) + \underbrace{u^T g(x)}_{\leq 0} + \underbrace{v^T h(x)}_{=0} \leq f(x),$$

womit der erste Teil des schwachen Dualitätssatzes bewiesen ist. Ist im zweiten Teil des Satzes $x^* \in M$, $(u^*, v^*) \in N$ und $\phi(u^*, v^*) = f(x^*)$, so ist

$$\phi(u^*, v^*) \leq \sup(\mathrm{D}) \leq \inf(\mathrm{P}) \leq f(x^*) = \phi(u^*, v^*),$$

also $f(x^*) = \inf(\mathrm{P})$ und $\phi(u^*, v^*) = \sup(\mathrm{D})$, womit die Behauptung bewiesen ist. □□

Der zweite Teil des schwachen Dualitätssatzes gibt eine *hinreichende Optimalitätsbedingung*: Gibt es zu einem $x^* \in M$ ein Paar $(u^*, v^*) \in N$ mit $f(x^*) = \phi(u^*, v^*)$, so ist x^* eine Lösung von (P). Natürlich sucht man nach notwendigen und hinreichenden Optimalitätsbedingungen, die möglichst wenig auseinander klaffen.

2.2.2 Starke Dualitätssätze für konvexe Programme

Wir werden hier zwei allgemeine starke Dualitätssätze formulieren und beweisen. Den gleich folgenden Satz werden wir später auf lineare, (konvexe) quadratische und quadratisch restringierte quadratische Programme anwenden.

Satz 2.2 *Gegeben sei das Programm (P), die Voraussetzung (V) sei erfüllt. Mit (D) wird das zu (P) duale Programm bezeichnet. Die Menge*

$$\Lambda := \{(f(x) + r, g(x) + z, h(x)) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m : x \in C,\ r \geq 0,\ z \geq 0\}$$

sei abgeschlossen. Dann gilt:

1. *Ist (P) zulässig und* $\inf(\mathrm{P}) > -\infty$*, so ist (P) lösbar, (D) zulässig und* $\sup(\mathrm{D}) = \min(\mathrm{P})$.

2. *Ist (D) zulässig und* $\sup(\mathrm{D}) < +\infty$*, so ist (P) zulässig und* $\inf(\mathrm{P}) > -\infty$.

Beweis: Sei (P) zulässig und $\inf(\mathrm{P}) > -\infty$. Um nachzuweisen, dass (P) lösbar ist, betrachte man eine Folge $\{x_k\} \subset M$ mit $f(x_k) \to \inf(\mathrm{P})$. Da die Folge $\{(f(x_k), 0, 0)\} \subset \Lambda$ gegen $(\inf(\mathrm{P}), 0, 0)$ konvergiert und Λ nach Voraussetzung abgeschlossen ist, ist $(\inf(\mathrm{P}), 0, 0) \in \Lambda$ und folglich (P) lösbar. Wir zeigen nun, dass (D) zulässig und $\sup(\mathrm{D}) = \min(\mathrm{P})$ ist. Hierzu sei $\alpha < \min(\mathrm{P})$ beliebig gewählt und damit $(\alpha, 0, 0) \notin \Lambda$, wobei wir notieren, dass Λ nichtleer, abgeschlossen und konvex (Beweis?) ist. Der starke Trennungssatz sichert die Existenz eines Tripels (q^*, u^*, v^*) und einer Zahl $\gamma \in \mathbb{R}$ mit

$$(*) \qquad \begin{cases} q^*\alpha < \gamma \leq q^*[f(x) + r] + (u^*)^T[g(x) + z] + (v^*)^T h(x) \\ \qquad\qquad \text{für alle } x \in C,\ r \geq 0,\ z \geq 0. \end{cases}$$

Mit einem Routineschluss folgt hieraus $q^* \geq 0$ und $u^* \geq 0$. Da $(\min(\mathrm{P}), 0, 0) \in \Lambda$, ist $q^*\alpha < \gamma \leq q^* \min(\mathrm{P})$, daher $q^* > 0$ und o. B. d. A. $q^* = 1$. Aus $(*)$ erhalten wir

$$\alpha < \gamma \leq f(x) + (u^*)^T g(x) + (v^*)^T h(x) \qquad \text{für alle } x \in C,$$

hieraus $(u^*, v^*) \in N$, so dass (D) zulässig ist, und $\alpha < \phi(u^*, v^*) \leq \sup(\mathrm{D})$. Da $\alpha < \min(\mathrm{P})$ beliebig ist, folgt $\min(\mathrm{P}) \leq \sup(\mathrm{D})$. Eine Anwendung des schwachen Dualitätssatzes schließt den Beweis des ersten Teiles des Satzes ab.

Zum Beweis des zweiten Teiles nehmen wir an, dass (D) zulässig und $\sup(\mathrm{D}) < +\infty$ ist. Wir zeigen $(\sup(\mathrm{D}), 0, 0) \in \Lambda$, woraus die Zulässigkeit von (P) und $\inf(\mathrm{P}) > -\infty$

folgt. Angenommen, es sei $(\sup(\mathrm{D}), 0, 0) \notin \Lambda$. Eine Anwendung des starken Trennungssatzes liefert die Existenz von (q^*, u^*, v^*) und $\gamma \in \mathbb{R}$ mit

$$\begin{cases} q^* \sup(\mathrm{D}) < \gamma \leq q^*[f(x) + r] + (u^*)^T[g(x) + z] + (v^*)^T h(x) \\ \text{für alle } x \in C,\ r \geq 0,\ z \geq 0. \end{cases}$$

Wie üblich folgt hieraus $q^* \geq 0$, $u^* \geq 0$ und

$$q^* \sup(\mathrm{D}) < \gamma \leq q^* f(x) + (u^*)^T g(x) + (v^*)^T h(x) \qquad \text{für alle } x \in C.$$

Ist $q^* > 0$, dann o. B. d. A. $q^* = 1$, folglich $(u^*, v^*) \in N$ und $\sup(\mathrm{D}) < \gamma \leq \phi(u^*, v^*)$, ein Widerspruch. Ist dagegen $q^* = 0$, so ist

$$0 < \gamma \leq (u^*)^T g(x) + (v^*)^T h(x) \qquad \text{für alle } x \in C.$$

Nach Voraussetzung ist (D) zulässig, d. h. es existiert $(u, v) \in N$. Für alle $t \geq 0$ ist $(u, v) + t(u^*, v^*) \in N$ und $\phi((u, v) + t(u^*, v^*)) \geq \phi(u, v) + t\gamma$, was wegen $\gamma > 0$ ein Widerspruch zu $\sup(\mathrm{D}) < +\infty$ ist. □ □

Es folgt ein zweiter starker Dualitätssatz, in dem durch eine Zusatzbedingung, eine sogenannte *Constraint Qualification*, auch die Lösbarkeit des dualen Programms gesichert werden kann.

Satz 2.3 *Gegeben sei (unter der Voraussetzung (V)) das konvexe Programm (P) und das hierzu duale Programm (D). Die sogenannte Slater'sche Constraint Qualification sei erfüllt, d. h. es gelte:*

(a) Es existiert ein $\hat{x} \in C$ mit $g(\hat{x}) < 0$ und $h(\hat{x}) = 0$,

(b) Es ist[7] *$h(C) = \mathbb{R}^m$.*

Ist dann $\inf(\mathrm{P}) > -\infty$, *so ist (D) lösbar und* $\max(\mathrm{D}) = \inf(\mathrm{P})$.

Beweis: Man definiere

$$\Lambda_+ := \{(f(x) + r, g(x) + z, h(x)) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m : x \in C,\ r > 0, z \geq 0\}.$$

Es ist leicht nachzuprüfen, dass Λ_+ konvex (und nichtleer) ist. Ferner ist $(\inf(\mathrm{P}), 0, 0) \notin \Lambda_+$, denn andernfalls gäbe es ein $x \in M$ mit $f(x) < \inf(\mathrm{P})$. Wegen des Trennungssatzes für konvexe Mengen läßt sich der Punkt $\{(\inf(\mathrm{P}), 0, 0)\}$ von der Menge Λ_+ trennen. Daher existiert ein Tripel $(q^*, u^*, v^*) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m \setminus \{(0, 0, 0)\}$ mit

$$\begin{gathered} q^* \inf(\mathrm{P}) \leq q^*[f(x) + r] + (u^*)^T[g(x) + z] + (v^*)^T h(x) \\ \text{für alle} \quad x \in C,\ r > 0,\ z \geq 0. \end{gathered}$$

Offenbar ist notwendigerweise $q^* \geq 0$ und auch $u^* \geq 0$. Wäre $q^* = 0$, so wäre

$$0 \leq (u^*)^T g(x) + (v^*)^T h(x) \qquad \text{für alle } x \in C.$$

[7] Ist $C = \mathbb{R}^n$ und $h(x) = b - Ax$ mit $A \in \mathbb{R}^{m \times n}$, so bedeutet dies, dass Rang $(A) = m$. Natürlich fällt diese Voraussetzung fort, wenn keine (affin linearen) Gleichungen im konvexen Programm (P) auftreten.

Mit (a) folgt $u^* = 0$, anschließend $v^* = 0$ mit (b). Dies ist ein Widerspruch zu $(q^*, u^*, v^*) \neq (0, 0, 0)$. O. B. d. A. können wir dann $q^* = 1$ annehmen und haben

$$\inf(\mathrm{P}) \leq f(x) + (u^*)^T g(x) + (v^*)^T h(x) = L(x, u^*, v^*) \qquad \text{für alle } x \in C.$$

Also ist $(u^*, v^*) \in N$ dual zulässig und $\inf(\mathrm{P}) \leq \phi(u^*, v^*)$. Aus dem schwachen Dualitätssatz folgt die Behauptung. □
□

Beispiel: Ohne eine Zusatzbedingung kann nicht die Lösbarkeit des dualen Programms gesichert werden. Hierzu betrachten wir ein triviales Beispiel:

$$(\mathrm{P}) \qquad \text{Minimiere} \quad x \quad \text{unter der Nebenbedingung} \quad \tfrac{1}{2}x^2 \leq 0.$$

Offenbar ist $x^* = 0$ die einzige zulässige Lösung, damit die Lösung und $\min(\mathrm{P}) = 0$. Die Lagrange-Funktion zu (P) ist $L(x, u) = x + \frac{1}{2}x^2 u$. Das zu (P) duale Problem ist (nach kurzer Rechnung)

$$(\mathrm{D}) \qquad \text{Maximiere} \quad -\frac{1}{2u} \quad \text{unter der Nebenbedingung} \quad u > 0.$$

Das duale Problem besitzt also keine Lösung, es ist aber $\min(\mathrm{P}) = \sup(\mathrm{D})$. □

Bemerkung: Im Anschluss an den Trennungssatz 1.10 zitierten wir ein Ergebnis, das man bei R. T. Rockafellar (1972, Theorem 20.2) finden kann. Insbesondere gilt:

- *Seien $A, B \subset \mathbb{R}^n$ nichtleer, konvex und disjunkt, A sei sogar ein Polyeder. Dann sind A und B durch eine Hyperebene, die B nicht enthält, echt trennbar.*

Mit Hilfe dieses Trennungssatzes kann man die Constraint Qualification im starken Dualitätssatz 2.3 abschwächen, wenn keine expliziten Restriktionen vorliegen (d. h. es ist $C = \mathbb{R}^n$) und ein Teil der Ungleichungsrestriktionen affin linear sind. Wie wir sehen werden, brauchen dann nur die Ungleichungsrestriktionen strikt erfüllbar zu sein, die nicht affin linear sind. Da wir affin lineare Gleichungen als zwei Ungleichungen schreiben können, nehmen wir an, dass in (P) keine Gleichungen auftreten. Wir gehen also jetzt aus von einem konvexen Programm der Form

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) \leq 0\}.$$

Hierbei setzen wir in dieser Bemerkung voraus:

(V) Die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ ist konvex, die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ bzw. $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ sind konvex bzw. affin linear.

Wieder wird die Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ durch

$$L(x, u, v) := f(x) + u^T g(x) + v^T h(x)$$

definiert. Das zu (P) duale Programm ist also

$$(\mathrm{D}) \qquad \begin{cases} \text{Maximiere} \quad \phi(u, v) := \inf\limits_{x \in \mathbb{R}^n} L(x, u, v) \quad \text{auf} \\ N := \{(u, v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ v \geq 0,\ \phi(u, v) > -\infty\}. \end{cases}$$

Wir wollen mit Hilfe des oben angegebenen Trennungssatzes den folgenden starken Dualitätssatz (siehe R. T. Rockafellar (1972, Theorem 28.2)) beweisen:

- *Das obige konvexe Programm (P) sei zulässig, ferner sei* $\inf(\mathrm{P}) > -\infty$. *Es existiere ein* $\hat{x} \in M$ *mit* $g(\hat{x}) < 0$. *Dann ist das duale Programm (D) lösbar und* $\max(\mathrm{D}) = \inf(\mathrm{P})$.

Denn: Wir definieren die beiden Mengen

$$\begin{aligned} A &:= \mathbb{R}_{\leq 0} \times \mathbb{R}^l_{\leq 0} \times \mathbb{R}^m_{\leq 0}, \\ B &:= \{(f(x) - \inf(\mathrm{P}) + r, g(x) + z, h(x)) : x \in \mathbb{R}^n,\ r > 0,\ z \geq 0\}. \end{aligned}$$

Dann sind A und B nichtleer, konvex (bei der Konvexität von B geht ein, dass h affin linear ist) und disjunkt (andernfalls existiert ein $x \in M$ mit $f(x) < \inf(\mathrm{P})$, was natürlich ein Widerspruch zur Definition von $\inf(\mathrm{P})$ ist). Ferner ist A ein Polyeder. Wir wenden den oben zitierten Trennungssatz an und erhalten die Existenz eines Tripels $(q^*, u^*, v^*) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m \setminus \{(0,0,0)\}$ und einer Zahl $\gamma \in \mathbb{R}$ mit

$$\begin{cases} q^* u_0 + (u^*)^T u + (v^*)^T v \leq \gamma \leq q^*[f(x) - \inf(\mathrm{P}) + r] + (u^*)^T[g(x) + z] + (v^*)^T h(x) \\ \qquad \text{für alle } (u_0, u, v) \in \mathbb{R}_{\leq 0} \times \mathbb{R}^l_{\leq 0} \times \mathbb{R}^m_{\leq 0},\ (x, r, z) \in \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^m_{\geq 0} \end{cases}$$

und

$$\begin{cases} \gamma < q^*[f(\tilde{x}) - \inf(\mathrm{P}) + \tilde{r}] + (u^*)^T[g(\tilde{x}) + \tilde{z}] + (v^*)^T h(\tilde{x}) \\ \qquad \text{für ein gewisses Tripel } (\tilde{x}, \tilde{r}, \tilde{z}) \in \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^m_{\geq 0}. \end{cases}$$

Aus der linken Ungleichung in der ersten Aussage schließen wir, dass $q^* \geq 0$, $u^* \geq 0$ und $v^* \geq 0$. Angenommen, es wäre $q^* = 0$. Dann ist

$$0 \leq \gamma \leq (u^*)^T g(x) + (v^*)^T h(x) \qquad \text{für alle } x \in \mathbb{R}^n.$$

Setzt man hier speziell $x := \hat{x}$, so erhält man

$$0 \leq \gamma \leq \underbrace{(u^*)^T g(\hat{x})}_{\leq 0} + \underbrace{(v^*)^T h(\hat{x})}_{\leq 0} \leq 0,$$

insbesondere $\gamma = 0$, $(u^*)^T g(\hat{x}) = 0$ und $(v^*)^T h(\hat{x}) = 0$. Wegen $g(\hat{x}) < 0$ folgt $u^* = 0$. Daher ist $(v^*)^T h(x) \geq 0$ für alle $x \in \mathbb{R}^n$. Eine affin lineare, nach unten beschränkte reellwertige Funktion ist konstant und daher $(v^*)^T h(x) = (v^*)^T h(\hat{x}) = 0$ für alle $x \in \mathbb{R}^n$. Andererseits ist wegen der zweiten durch den Trennungssatz gelieferten Aussage

$$0 < (v^*)^T h(\tilde{x})$$

mit einem gewissen $\tilde{x} \in \mathbb{R}^n$. Das ist ein Widerspruch, so dass wir o. B. d. A. $q^* = 1$ annehmen können. Daher ist

$$0 \leq f(x) - \inf(\mathrm{P}) + (u^*)^T g(x) + (v^*)^T h(x) \qquad \text{für alle } x \in \mathbb{R}^n,$$

folglich (u^*, v^*) dual zulässig und $\inf(\mathrm{P}) \leq \phi(u^*, v^*)$. Der schwache Dualitätssatz impliziert die Behauptung. □

2.2.3 Dualität in der linearen Optimierung

Bei einem *linearen Programm* sind in der obigen Formulierung eines konvexen Programms die Zielfunktion linear, die Restriktionsabbildungen g und h jeweils affin linear, ferner durch die Menge C sind Vorzeichenbedingungen für zulässige Lösungen gegeben. Typischerweise ist

$$C = \{x \in \mathbb{R}^n : x_j \geq 0,\ j = 1, \ldots, n_0\},$$

wobei $n_0 \in \{0, \ldots, n\}$. Es wäre nicht schwierig, diesen (scheinbar) allgemeineren Fall zu behandeln. Da man aber bekanntlich jedes lineare Programm (mit Hilfe von Schlupfvariablen und Darstellung nicht vorzeichenbeschränkter Variablen als Differenz nichtnegativer Variabler) auf äquivalente Normalform bringen kann, gehen wir bei der folgenden Formulierung des Existenzsatzes bzw. des starken Dualitätssatzes der linearen Optimierung gleich von einem Ausgangsproblem in Normalform aus, also von

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b\}.$$

Das hierzu duale lineare Programm ist

$$\text{(D)} \qquad \text{Maximiere} \quad b^T y \quad \text{auf} \quad N := \{y \in \mathbb{R}^m : A^T y \leq c\}.$$

Bekanntlich liefert das Dualisieren von (D) wieder das Ausgangsproblem (P).

Satz 2.4 *Das lineare Programm (P) sei zulässig und* $\inf(\mathrm{P}) > -\infty$. *Dann besitzt (P) eine Lösung.*

Beweis: Wir wollen die Existenzaussage in Satz 2.2 anwenden und haben hierzu zu zeigen, dass die Menge

$$\Lambda := \{(c^T x + r, b - Ax) : x \geq 0,\ r \geq 0\}$$

abgeschlossen ist. Nun ist aber

$$\Lambda = \begin{pmatrix} 0 \\ b \end{pmatrix} + \left\{ \begin{pmatrix} c^T & 1 \\ -A & 0 \end{pmatrix} \begin{pmatrix} x \\ r \end{pmatrix} : \begin{pmatrix} x \\ r \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\},$$

also Λ ein verschobener endlich erzeugter Kegel, nach Lemma 1.5 ist Λ abgeschlossen. Der Existenzsatz der linearen Optimierung ist damit bewiesen. □ □

Bemerkung: Wir können den Existenzsatz der linearen Optimierung auch folgendermaßen formulieren: Ist eine lineare Funktion auf einem nichtleeren Polyeder nach unten beschränkt, so nimmt sie auf dem Polyeder ihr Minimum an. □

Satz 2.5 *Gegeben sei das lineare Programm (P) und das dazu duale lineare Programm (D). Dann gilt:*

1. *Sind (P) und (D) zulässig, so sind (P) und (D) lösbar und es ist* $\max(\mathrm{D}) = \min(\mathrm{P})$.

2. *Ist (D) zulässig, aber (P) nicht zulässig, so ist* $\sup(\mathrm{D}) = +\infty$.

3. *Ist (P) zulässig, aber (D) nicht zulässig, so ist* $\inf(\mathrm{P}) = -\infty$.

Beweis: Da (P) und (D) zulässig sind, ist wegen des schwachen Dualitätssatzes

$$-\infty < \sup(\mathrm{D}) \leq \inf(\mathrm{P}) < +\infty.$$

Aus dem Existenzsatz folgt die Lösbarkeit von (P) und (D), wegen des starken Dualitätssatzes 2.2 (die Menge Λ ist abgeschlossen) ist $\max(\mathrm{D}) = \min(\mathrm{P})$. Auch die beiden weiteren Aussagen folgen direkt aus dem starken Dualitätssatz 2.2. □ □

Wir werden später sehen, dass sich die Existenzaussage vollständig und die Dualitätsaussage weitgehend auf konvexe quadratische, quadratisch restringierte Optimierungsaufgaben übertragen läßt. Lineare Programme zeichnen sich gegenüber diesen wesentlich allgemeineren Aufgaben dadurch aus, dass bei ihnen ein *strikt komplementäres, optimales Paar* existiert. Wir gehen weiter von dem linearen Programm (P) in Normalform und dem dazu dualen linearen Programm (D) aus. Mit M_{opt} bezeichnen wir die Menge der Lösungen von (P), entsprechend mit N_{opt} die Menge der Lösungen von (D). Ist dann $x^* \in M_{\mathrm{opt}}$ und $y^* \in N_{\mathrm{opt}}$, so ist

$$0 = \min(\mathrm{P}) - \max(\mathrm{D}) = c^T x^* - b^T y^* = (\underbrace{c - A^T y^*}_{\geq 0})^T \underbrace{x^*}_{\geq 0}$$

und daher

$$x_j^*(c - A^T y^*)_j = 0, \qquad j = 1, \ldots, n.$$

Hierdurch wird aber nicht ausgeschlossen, dass sowohl x_j^* als auch $(c - A^T y^*)_j$ für ein gewisses j verschwinden. Im folgenden Satz wird ausgesagt, dass es bei linearen Programmen wenigstens ein Paar von Lösungen gibt, für die das nicht der Fall ist. Siehe auch A. Schrijwer (1986, S. 95)[8].

Satz 2.6 *Die zueinander dualen linearen Programme*

(P) $$\textit{Minimiere} \quad c^T x \quad \textit{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b\}$$

und

(D) $$\textit{Maximiere} \quad b^T y \quad \textit{auf} \quad N := \{y \in \mathbb{R}^m : A^T y \leq c\}$$

seien zulässig. Mit M_{opt} bzw. N_{opt} seien die (nichtleeren) Lösungsmengen von (P) bzw. (D) bezeichnet. Dann existiert ein Paar $(x^*, y^*) \in M_{\mathrm{opt}} \times N_{\mathrm{opt}}$ mit $x^* + c - A^T y^* > 0$.

Beweis: Zunächst wollen wir uns überlegen, dass es genügt, die folgende Hilfsbehauptung zu beweisen:

- Sei $k \in \{1, \ldots, n\}$. Existiert kein $y \in N_{\mathrm{opt}}$ mit $(c - A^T y)_k > 0$, so existiert ein $x \in M_{\mathrm{opt}}$ mit $x_k > 0$.

[8] A. Schrijver (1986) *Theory of Linear and Integer Programming.* J. Wiley & Sons.

Ist diese Aussage bewiesen, so existieren für $k = 1, \ldots, n$ Paare $(x^{(k)}, y^{(k)}) \in M_{\text{opt}} \times N_{\text{opt}}$ mit $(x^{(k)} + c - A^T y^{(k)})_k > 0$. Durch

$$x^* := \frac{1}{n} \sum_{k=1}^{n} x^{(k)}, \qquad y^* := \frac{1}{n} \sum_{k=1}^{n} y^{(k)}$$

ist das gesuchte Paar (x^*, y^*) gefunden.

Es genügt also, die obige Hilfsbehauptung zu beweisen. Gibt es bei gegebenem $k \in \{1, \ldots, n\}$ kein $y \in N_{\text{opt}}$ mit $(c - A^T y)_k > 0$, so gilt die Implikation

$$y \in N, \quad b^T y \geq \max(\mathrm{D}) \Longrightarrow (-Ae_k)^T y \leq -c_k,$$

wobei e_k den k-ten Einheitsvektor im $\mathbb{R}^n$ bedeutet. Dann hat das lineare Programm

$$(\tilde{\mathrm{D}}) \quad \text{Maximiere} \quad (-Ae_k)^T y \quad \text{auf} \quad N_{\text{opt}} = \left\{ y \in \mathbb{R}^m : \begin{pmatrix} A^T \\ -b^T \end{pmatrix} y \leq \begin{pmatrix} c \\ -\max(\mathrm{D}) \end{pmatrix} \right\}$$

eine Lösung mit einem Wert $\max(\tilde{\mathrm{D}}) \leq -c_k$ ist. Der starke Dualitätssatz liefert, dass das dazu duale Programm

$$(\tilde{\mathrm{P}}) \qquad \begin{cases} \text{Minimiere} \quad c^T z - \lambda \max(\mathrm{D}) \quad \text{unter den Nebenbedingungen} \\ z \geq 0, \quad \lambda \geq 0, \quad Az - \lambda b = -Ae_k \end{cases}$$

eine Lösung (z^*, λ^*) mit dem gleichen Wert besitzt, so dass also

$$\min(\tilde{\mathrm{P}}) = c^T z^* - \lambda^* \max(\mathrm{D}) = \max(\tilde{\mathrm{D}}) \leq -c_k.$$

Ist $\lambda^* = 0$, so hat man ein z^* mit $z^* \geq 0$, $Az^* = -Ae_k$ und $c^T z^* \leq -c_k$ gefunden. Definiert man daher $x^{(k)} := x + z^* + e_k$ mit einem beliebigen $x \in M_{\text{opt}}$, so ist $x^{(k)} \in M$, $c^T x^{(k)} \leq c^T x$, folglich $x^{(k)} \in M_{\text{opt}}$, und $x_k^{(k)} \geq 1$. Ist dagegen $\lambda^* > 0$, so definiere man $x^{(k)} := (z^* + e_k)/\lambda^*$. Wieder ist $x^{(k)} \in M$, ferner $c^T x^{(k)} \leq \max(\mathrm{D}) = \min(\mathrm{P})$, also $x^{(k)} \in M_{\text{opt}}$. Wegen $x_k^{(k)} \geq 1/\lambda^*$ hat man auch in diesem Falle eine Lösung von (P) gefunden, deren k-te Komponente positiv ist. Insgesamt ist der Satz bewiesen. □ □

2.2.4 Quadratisch restringierte quadratische Programme

In diesem Unterabschnitt betrachten wir Aufgaben der Form

$$(\mathrm{P}) \qquad \begin{cases} \text{Minimiere} \quad f(x) := c_0^T x + \frac{1}{2} x^T Q_0 x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : g_i(x) := \beta_i + c_i^T x + \frac{1}{2} x^T Q_i x \leq 0, \ i = 1, \ldots, l, \ Ax = b\}. \end{cases}$$

Hierbei seien generell die Matrizen $Q_0, Q_1, \ldots, Q_l \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit, also (P) ein konvexes Programm. Ferner seien natürlich $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c_0, c_1, \ldots, c_l \in \mathbb{R}^n$ und $\beta_1, \ldots, \beta_l \in \mathbb{R}$ gegeben.

Das erste Ergebnis ist ein Existenzsatz, der vollständig dem Existenzsatz der linearen Optimierung entspricht. Er stammt von E. L. Peterson, J. G. Ecker (1969, 1970)[9]. Wir präsentieren allerdings einen wesentlich einfacheren Beweis.

Satz 2.7 *Das konvexe, quadratisch restringierte quadratische Programm (P) sei zulässig, ferner sei* $\inf(\mathrm{P}) > -\infty$. *Dann besitzt (P) eine Lösung.*

Beweis: Wir können offenbar annehmen, dass in (P) keine (linearen) Gleichungen als Restriktionen auftreten, da man ja eine Gleichung als zwei Ungleichungen schreiben kann. Wir gehen daher o. B. d. A. von der Aufgabe

$$(\mathrm{P}) \qquad \begin{cases} \text{Minimiere} \quad f(x) := c_0^T x + \frac{1}{2} x^T Q_0 x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : g_i(x) := \beta_i + c_i^T x + \frac{1}{2} x^T Q_i x \leq 0,\ i = 1, \ldots, l\} \end{cases}$$

aus. Wir definieren die konvexe, quadratische Funktion $g_0 \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch $g_0(x) := f(x) - \inf(\mathrm{P})$, weiter die konvexe Funktion $G \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$G(x) := \max_{i=0,\ldots,l} g_i(x).$$

Dann ist $\inf_{x \in \mathbb{R}^n} G(x) = 0$. Wir werden die Existenz eines $x^* \in \mathbb{R}^n$ mit $G(x^*) = 0$ zeigen. Offenbar ist dann $x^* \in M$ eine Lösung von (P).

Wir nennen eine (nicht notwendig nichtleere) Indexmenge $I \subset \{0, \ldots, l\}$ *kanonisch*, wenn die Implikation

$$p \in \mathbb{R}^n, \quad c_i^T p \leq 0, \quad Q_i p = 0 \quad (i \in I) \Longrightarrow c_i^T p = 0 \quad (i \in I)$$

gilt. In einem ersten Schritt zeigen wir:

- Ist $I \subset \{0, \ldots, m\}$ kanonisch, so existiert ein $x \in \mathbb{R}^n$ mit $g_i(x) \leq 0$, $i \in I$.

Denn: Die Aussage ist trivial, wenn $I = \emptyset$ oder $\inf_{x \in \mathbb{R}^n} \max_{i \in I} g_i(x) < 0$. Wir können also annehmen, daß $I \neq \emptyset$ und $\inf_{x \in \mathbb{R}^n} \max_{i \in I} g_i(x) = 0$. Mit $B[0; k]$ bezeichnen wir die euklidische Kugel um den Nullpunkt mit dem Radius $k \in \mathbb{N}$, ferner sei $x_k \in B[0; k]$ die Lösung minimaler euklidischer Norm der Optimierungsaufgabe

$$(\mathrm{P}_k) \qquad \text{Minimiere} \quad G_I(x) := \max_{i \in I} g_i(x), \quad x \in B[0; k].$$

Offenbar ist dann

$$\lim_{k \to \infty} G_I(x_k) = \lim_{k \to \infty} \min_{x \in B[0;k]} G_I(x) = \inf_{x \in \mathbb{R}^n} G_I(x) = 0.$$

[9] E. L. Peterson, J. G. Ecker (1970) "Geometric programming: Duality in quadratic programming and l_p-approximation I." In: *Proceedings of the Princeton Symposion on Mathematical Programming* (H. W. Kuhn, Ed.), 445–480. Princeton University Press, Princeton.

E. L. Peterson, J. G. Ecker (1969) "Geometric programming: Duality in quadratic programming and l_p-approximation II (canonical programs)." SIAM J. Appl. Math. 17, 317–340.

E. L. Peterson, J. G. Ecker (1970) "Geometric programming: Duality in quadratic programming and l_p-approximation III (degenerate programs)." J. Math. Anal. Appl. 29, 365–383.

Besitzt daher $\{x_k\}$ eine Häufungspunkt x, so ist $G_I(x) = 0$, also x der gesuchte Punkt. Andernfalls ist $\|x_k\| \to \infty$, o. B. d. A. gilt $x_k/\|x_k\| \to p$, wobei natürlich $\|p\| = 1$, insbesondere also $p \neq 0$. Wegen

$$g_i(x_k) = \beta_i + c_i^T x_k + \tfrac{1}{2} x_k^T Q_i x_k \leq G_I(x_k) \to 0, \qquad i \in I,$$

folgt

$$c_i^T p \leq 0, \quad Q_i p = 0 \quad (i \in I).$$

Da $I \subset \{0, \ldots, m\}$ nach Voraussetzung kanonisch ist, folgt $c_i^T p = 0$, $i \in I$. Für alle $t \in \mathbb{R}$ ist daher $g_i(x_k) = g_i(x_k - tp)$, $i \in I$, insbesondere $G_I(x_k) = G_I(x_k - tp)$ für alle $t \in \mathbb{R}$ und alle $k \in \mathbb{N}$. Andererseits ist

$$\lim_{t \to 0+} \frac{\|x_k - tp\|^2 - \|x_k\|^2}{t} = -2x_k^T p < 0$$

für alle hinreichend großen k. Für diese k und alle hinreichend kleinen $t > 0$ ist daher $x_k - tp$ eine Lösung von (P_k) mit einer kleineren euklidischen Norm als der von x_k, ein Widerspruch zu der Definition von x_k.

Nun kommen wir zum entscheidenden Schritt und zeigen:

- Sei $I^* \subset \{0, \ldots, m\}$ unter allen kanonischen Teilmengen von $\{0, \ldots, m\}$ maximal. Wegen der gerade eben bewiesenen Aussage existiert ein $x \in \mathbb{R}^n$ mit $g_i(x) \leq 0$, $i \in I^*$. Dann existiert ein $x^* \in \mathbb{R}^n$ mit $g_i(x^*) = g_i(x)$, $i \in I^*$, und $g_i(x^*) \leq 0$, $i \in \{0, \ldots, m\} \setminus I^*$. Dieses x^* ist eine Lösung von (P).

Denn: Wir können annehmen, daß I^* eine echte Teilmenge von $I_0 := \{0, \ldots, m\}$ ist, da man andernfalls $x^* := x$ wählen kann. Alle Teilmengen I von I_0, die I^* echt enthalten, sind nicht kanonisch, d. h. das Gleichungs-Ungleichungssystem

$$(I) \qquad c_i^T p \leq 0, \quad Q_i p = 0 \quad (i \in I), \quad \Big(\sum_{i \in I} c_i\Big)^T p < 0$$

besitzt eine Lösung. Auf die folgende Weise bestimmen wir strikt absteigende Indexmengen $I_0 \supset I_1 \supset \cdots \supset I_r \supset I^*$, welche mit der maximalen kanonischen Indexmenge I^* enden.

Für $k = 0, 1, \ldots$:

- Sei p_k eine Lösung von

  $$(I_k) \qquad c_i^T p \leq 0, \quad Q_i p = 0 \quad (i \in I_k), \quad \Big(\sum_{i \in I_k} c_i\Big)^T p < 0$$

  und definiere die (nichtleere) Indexmenge

  $$J_k := \{i \in I_k : c_i^T p_k < 0\}.$$

- Falls $I_k \setminus J_k = I^*$, dann: $r := k$, STOP.

– Andernfalls: Setze $I_{k+1} := I_k \setminus J_k$.

Nun setze man

$$x^* := x + \sum_{k=0}^{r} \alpha_k p_k$$

mit noch unbestimmten Konstanten $\alpha_0, \ldots, \alpha_r \geq 0$. Wegen

$$c_i^T p_k = 0, \quad Q_i p_k = 0 \quad (i \in I^*), \quad k = 0, \ldots, r,$$

ist

$$g_i(x^*) = g_i(x) \quad (i \in I^*).$$

Für $i \in J_r = I_r \setminus I^*$ ist

$$c_i^T p_r < 0, \quad c_i^T p_k \leq 0 \quad (k = 0, \ldots, r-1), \quad Q_i p_k = 0 \quad (k = 0, \ldots, r).$$

Nun wähle man $\alpha_r \geq 0$ so groß, dass (bei noch unbestimmten $\alpha_0, \ldots, \alpha_{r-1}$) gilt:

$$g_i(x^*) = g_i(x) + \alpha_r \underbrace{c_i^T p_r}_{<0} + \sum_{k=0}^{r-1} \alpha_k \underbrace{c_i^T p_k}_{\leq 0} \leq g_i(x) + \alpha_r c_i^T p_r \leq 0, \quad i \in J_r.$$

Für $i \in J_{r-1} = I_{r-1} \setminus I_r$ ist entsprechend

$$c_i^T p_{r-1} < 0, \quad c_i^T p_k \leq 0 \quad (k = 0, \ldots, r-2), \quad Q_i p_k = 0 \quad (k = 0, \ldots, r-1).$$

Durch Wahl eines hinreichend großen $\alpha_{r-1} \geq 0$ (bei noch unbestimmten $\alpha_0, \ldots, \alpha_{r-2}$) ist

$$g_i(x^*) \leq g_i(x + \alpha_r p_r) + \alpha_{r-1} c_i^T p_{r-1} \leq 0, \quad i \in J_{r-1}.$$

In dieser Weise kann man fortfahren. Nach endlich vielen Schritten hat man nichtnegative Zahlen $\alpha_r, \ldots, \alpha_0$ so bestimmt, dass für $x^* := x + \sum_{k=0}^{r} \alpha_k p_k$ nicht nur $g_i(x^*) = g_i(x) \leq 0$, $i \in I^*$, sondern auch $g_i(x^*) \leq 0$, $i \notin I^*$. Dann ist

$$0 = \inf_{z \in \mathbb{R}^n} G(z) \leq G(x^*) \leq 0,$$

also $G(x^*) = 0$ bzw. $x^* \in M$ und $f(x^*) = \inf(\mathrm{P})$. Damit ist die obige Behauptung und folglich der ganze Satz bewiesen. □ □

Bemerkung: Als Spezialfall von Satz 2.7 erhält man: Ist eine konvexe, quadratische Funktion auf einem nichtleeren Polyeder nach unten beschränkt, so nimmt sie auf diesem Polyeder ihr Minimum an. Dies ist ein Ergebnis, das zuerst von E. Barankin, R. Dorfman (1958)[10] bewiesen wurde. □

Unser Ziel ist es, den starken Dualitätssatz 2.2 auf das konvexe, quadratisch restringierte quadratische Programm (P) anzuwenden. Mit den durch

$$g(x) := \begin{pmatrix} g_1(x) \\ \vdots \\ g_l(x) \end{pmatrix}, \qquad h(x) := b - Ax$$

[10] Barankin, E. and R. Dorfman (1958) "On quadratic programming." University of California Publications in Statistics 2, 258–318.

definierten Abbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ bzw. $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ist hierzu zu zeigen, dass die Menge

$$\Lambda := \{(f(x) + r, g(x) + z, h(x)) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m : x \in \mathbb{R}^n,\ r \geq 0,\ z \geq 0\}$$

abgeschlossen ist. Dies wird eine verhältnismäßig einfache Folgerung aus dem nächsten Lemma sein, das man als eine Verallgemeinerung des Farkas-Lemmas auf (konvexe) quadratische Funktionen ansehen kann. Mit

$$g'(x) = \begin{pmatrix} (c_1 + Q_1 x)^T \\ \vdots \\ (c_l + Q_l x)^T \end{pmatrix}, \qquad h'(x) = -A$$

bezeichnen wir die Funktionalmatrizen zu g bzw. h. Dann gilt:

Lemma 2.8 *Seien $g_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, l$, konvex und quadratisch, also*

$$g_i(x) := \beta_i + c_i^T x + \tfrac{1}{2} x^T Q_i x, \qquad i = 1, \ldots, m,$$

mit symmetrischen, positiv semidefiniten Matrizen $Q_i \in \mathbb{R}^{n \times n}$, $i = 1, \ldots, l$. Ferner sei $h(x) := b - Ax$ mit $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Dann gilt genau eine der beiden folgenden Aussagen:

(I) Es existiert ein $x \in \mathbb{R}^n$ mit $g(x) \leq 0$, $h(x) = 0$.

(II) Es existiert ein Tripel $(u, v, z) \in \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^n$ mit

$$u \geq 0, \qquad g'(z)^T u + h'(z)^T v = 0, \qquad 0 < u^T g(z) + v^T h(z).$$

Beweis: Angenommen, (I) und (II) würden beide gelten. Dann wäre

$$0 < u^T g(z) + v^T h(z) \leq u^T g(x) + v^T h(x) - [u^T g'(z) + v^T h'(z)]^T (x - z) = u^T g(x) \leq 0,$$

ein Widerspruch. Nun nehmen wir an, (I) würde nicht gelten. Wir nehmen o. B. d. A. an, dass in (I) keine Gleichungen auftreten (andernfalls schreibe man $h(x) = 0$ als die beiden Ungleichungen $h(x) \leq 0$ und $-h(x) \leq 0$ und füge sie zu den übrigen Ungleichungen hinzu). Wir definieren die konvexe Funktion $G\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$G(x) := \max_{i=1,\ldots,l} g_i(x).$$

Dann ist $G(x) > 0$ für alle $x \in \mathbb{R}^n$ (da (I) nicht gilt). Die Aufgabe, $G(x)$ auf dem $\mathbb{R}^n$ zu minimieren, besitzt eine Lösung x^*, denn sie kann äquivalent in die quadratisch restringierte quadratische Aufgabe umformuliert werden, die Zielfunktion $f(x, s) := s$ unter der Nebenbedingung $g(x) - se \leq 0$ zu minimieren (e ist wieder der Vektor, dessen Komponenten alle gleich 1 sind). Dann ist

$$g(x) - G(x^*)e < 0$$

nicht lösbar, ferner ist $G(x^*) > 0$. Ein Satz von Fan-Glicksburg-Hoffman (siehe Aufgabe 12 in Abschnitt 2.1) liefert die Existenz von $u \in \mathbb{R}^l \setminus \{0\}$ mit $u \geq 0$ und

$$\inf_{x \in \mathbb{R}^n} u^T(g(x) - G(x^*)e) \geq 0.$$

Wegen $u \neq 0$ und $G(x^*) > 0$ ist

$$0 < u^T e\, G(x^*) \leq \inf_{x \in \mathbb{R}^n} u^T g(x).$$

Nun ist $u^T g(\cdot)$ eine konvexe, quadratische Funktion, die auf dem $\mathbb{R}^n$ nach unten (durch eine positive Zahl) beschränkt ist. Daher nimmt diese Funktion ihr Minimum in einen Punkt $z \in \mathbb{R}^n$ an, in dem der Gradient verschwindet. Damit ist nachgewiesen, dass (II) lösbar ist. □ □

Bemerkung: Angenommen, in Lemma 2.8 sei g affin linear, etwa $g(x) := \beta + Cx$ mit $\beta \in \mathbb{R}^l$ und $C \in \mathbb{R}^{l \times n}$. Ist wieder $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$, so sagt Lemma 2.8 in diesem Spezialfall aus: Es gilt genau eine der beiden folgenden Aussagen:

(I) Es existiert ein $x \in \mathbb{R}^n$ mit $\beta + Cx \leq 0$, $Ax = b$.

(II) Es existiert ein Tripel $(u, v, z) \in \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^n$ mit

$$u \geq 0, \qquad C^T u - A^T v = 0, \qquad 0 < u^T(\beta + Cz) + v^T(b - Az) = u^T\beta + v^T b$$

bzw. ein Paar $(u, v) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u \geq 0, \qquad C^T u - A^T v = 0, \qquad 0 < u^T\beta + v^T b.$$

Genau diese Aussage hätten wir auch aus dem verallgemeinerten Farkas-Lemma 1.8 erhalten. Man prüfe dies nach! □

Wir gehen weiter aus von dem konvexen, quadratisch restringierten quadratischen Programm

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad f(x) := c_0^T x + \frac{1}{2} x^T Q_0 x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : g_i(x) := \beta_i + c_i^T x + \frac{1}{2} x^T Q_i x \leq 0,\ i = 1, \ldots, l,\ Ax = b\}. \end{cases}$$

Mit $g(x) := (g_1(x), \ldots, g_l(x))^T$ und $h(x) := b - Ax$ ist die zugehörige Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ wie üblich durch

$$L(x, u, v) := f(x) + u^T g(x) + v^T h(x)$$

definiert. Ferner ist das zu (P) duale Problem durch

$$\text{(D)} \quad \begin{cases} \text{Maximiere} \quad \phi(u, v) := \inf_{x \in \mathbb{R}^n} L(x, u, v) \quad \text{auf} \\ N := \{(u, v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \phi(u, v) > -\infty\} \end{cases}$$

gegeben. Bei gegebenem $(u, v) \in \mathbb{R}^l \times \mathbb{R}^m$ mit $u \geq 0$ ist $L(\cdot, u, v)$ eine konvexe, quadratische Funktion. Es ist daher $L(\cdot, u, v)$ auf dem $\mathbb{R}^n$ nach unten beschränkt bzw.

$\phi(u,v) > -\infty$ genau dann, wenn ein $z \in \mathbb{R}^n$ mit $\nabla_x L(z,u,v) = 0$ existiert. Mit einem solchen z ist dann $\phi(u,v) = L(z,u,v)$.

Beispiel: Wir betrachten ein quadratisches Programm in Normalform, also die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x + \frac{1}{2} x^T Q x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b\}.$$

Hierbei sei $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$. Die zugehörige Lagrange-Funktion ist gegeben durch

$$L(x,u,v) := c^T x + \frac{1}{2} x^T Q x - u^T x + v^T (b - Ax).$$

Wie gerade eben erwähnt, ist $L(\cdot, u, v)$ genau dann auf dem $\mathbb{R}^n$ nach unten beschränkt, wenn ein $z \in \mathbb{R}^n$ mit

$$0 = \nabla_x L(z,u,v) = c + Qz - u - A^T v$$

existiert. Mit einem solchen z ist die duale Zielfunktion durch

$$\begin{aligned} \phi(u,v) &= L(z,u,v) \\ &= c^T z + \frac{1}{2} z^T Q z - u^T z + v^T (b - Az) \\ &= c^T z + \frac{1}{2} z^T Q z - (c + Qz - A^T v)^T z + v^T (b - Az) \\ &= b^T v - \frac{1}{2} z^T Q z. \end{aligned}$$

gegeben. Das zu (P) duale quadratische Programm ist also

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad b^T v - \frac{1}{2} z^T Q z \quad \text{auf} \\ N := \{(v,z) \in \mathbb{R}^m \times \mathbb{R}^n : c + Qz - A^T v \geq 0\}. \end{cases}$$

Man überlege sich, was für eine Optimierungsaufgabe man durch Dualisieren von (D) erhält. □

Beispiel: Wir wollen den Spezialfall betrachten, dass $Q_0 \in \mathbb{R}^{n \times n}$ sogar positiv definit ist. Es ist

$$L(x,u,v) = \sum_{i=1}^{l} u_i \beta_i + b^T v + \Big(c_0 + \sum_{i=1}^{l} u_i c_i - A^T v\Big)^T x + \tfrac{1}{2} x^T \Big(Q_0 + \sum_{i=1}^{l} u_i Q_i\Big) x,$$

wobei

$$Q(u) := Q_0 + \sum_{i=1}^{l} u_i Q_i$$

für $u \geq 0$ positiv definit ist. Für jedes Paar $(u, v) \in \mathbb{R}^l \times \mathbb{R}^m$ mit $u \geq 0$ nimmt also $L(\cdot, u, v)$ in genau einem Punkt z sein Minimum an, dieser Punkt ist durch

$$\nabla_x L(z, u, v) = c_0 + \sum_{i=1}^{l} u_i c_i - A^T v + \Big(Q_0 + \sum_{i=1}^{l} u_i Q_i\Big) z$$

festgelegt, also durch

$$z = -Q(u)^{-1}[c(u) - A^T v]$$

gegeben, wobei wir zur Abkürzung noch

$$c(u) := c_0 + \sum_{i=1}^{l} u_i c_i$$

gesetzt haben. Mit den eingeführten Abkürzungen $Q(u)$, $c(u)$ lautet in dem betrachteten Spezialfall das duale Programm

$$\begin{cases} \text{Maximiere} \;\; \phi(u, v) := \displaystyle\sum_{i=1}^{l} u_i \beta_i + b^T v - \tfrac{1}{2}[c(u) - A^T v] Q(u)^{-1} [c(u) - A^T v] \;\; \text{auf} \\ \qquad\qquad N := \{(u, v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0\}. \end{cases}$$

Wir haben hier also relativ einfache Nebenbedingungen, dafür eine kompliziertere Zielfunktion. □

Nun wenden wir den starken Dualitätssatz 2.2 auf das allgemeine quadratisch restringierte quadratische Programm an und erhalten:

Satz 2.9 *Gegeben sei das allgemeine konvexe, quadratisch restringierte quadratische Programm (P), mit (D) sei das hierzu duale Programm bezeichnet. Dann gilt:*

1. *Ist (P) zulässig und* $\inf(\mathrm{P}) > -\infty$*, so ist (P) lösbar, (D) zulässig und* $\sup(\mathrm{D}) = \min(\mathrm{P})$.
2. *Ist (D) zulässig und* $\sup(\mathrm{D}) < +\infty$*, so ist (P) zulässig und* $\inf(\mathrm{P}) > -\infty$.

Beweis: Die Aussage ist genau dieselbe wie die des starken Dualitätssatzes 2.2, so dass es darauf ankommt, dessen Voraussetzung nachzuprüfen. Hierzu müssen wir uns überlegen, dass die Menge

$$\Lambda := \{(f(x) + r, g(x) + z, h(x)) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m : x \in \mathbb{R}^n,\ r \geq 0,\ z \geq 0\}$$

abgeschlossen ist. Zum Nachweis geben wir uns eine Folge

$$\{(f(x_k) + r_k, g(x_k) + z_k, h(x_k))\} \subset \Lambda$$

vor, die gegen ein Tripel $(\hat{f}, \hat{g}, \hat{h}) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m$ konvergiert. Angenommen, es wäre $(\hat{f}, \hat{g}, \hat{h}) \notin \Lambda$. Dann hätte das System

$$f(x) - \hat{f} \leq 0, \quad g(x) - \hat{g} \leq 0, \quad h(x) - \hat{h} = 0$$

keine Lösung. Aus Lemma 2.8 folgt die Existenz von $(u_0, u, v, z) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^n$ mit

$$u_0 \geq 0, \quad u \geq 0, \quad u_0 \nabla f(z) + g'(z)^T u + h'(z)^T v = 0$$

und

$$0 < u_0[f(z) - \hat{f}] + u^T[g(z) - \hat{g}] + v^T[h(z) - \hat{h}].$$

Dann ist

$$\begin{aligned} u_0\hat{f} + u^T\hat{g} + v^T\hat{h} &< u_0 f(z) + u^T g(z) + v^T h(z) \\ &\leq u_0 f(x_k) + u^T g(x_k) + v^T h(x_k) \\ &\qquad \text{(denn } u_0 f(\cdot) + u^T g(\cdot) + v^T h(\cdot) \text{ ist in } z \text{ minimal)} \\ &\leq u_0[f(x_k) + r_k] + u^T[g(x_k) + z_k] + v^T h(x_k) \\ &\to u_0\hat{f} + u^T\hat{g} + v^T\hat{h}, \end{aligned}$$

offensichtlich ein Widerspruch. Damit ist der Satz bewiesen. □ □

2.2.5 Aufgaben

1. Gegeben sei das lineare Programm

   (P) Minimiere $c^T x$ auf $M := \{x : x \geq 0,\ b - Ax \leq 0\}$.

   Man stelle das zu (P) duale lineare Programm auf.

2. Gegeben sei das lineare Programm

   (P) Minimiere $c^T x$ auf $M := \{x \in \mathbb{R}^n : Gx \leq h,\ Ax = b\}$,

   wobei l Ungleichungen und m Gleichungen auftreten. Man stelle das zu (P) duale lineare Programm auf.

3. Seien $a_1, \ldots, a_l \in \mathbb{R}^n$ gegeben. Es sei die kleinste euklidische Kugel zu bestimmen, die $a_1, \ldots, a_l$ enthält. Man formuliere diese Aufgabe als eine Optimierungsaufgabe (P), bei der eine lineare Zielfunktion unter konvexen quadratischen Ungleichungsrestriktionen zu minimieren ist und stelle das zugehörige duale Programm (D) auf. Weiter zeige man, dass beide Probleme lösbar sind und $\max(\mathrm{D}) = \min(\mathrm{P})$ gilt.

4. Gegeben sei das konvexe Programm

   (P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : x \in C,\ g(x) \leq 0\}$.

   Hierbei wird vorausgesetzt:

   (V) $C \subset \mathbb{R}^n$ ist nichtleer und konvex, $f\colon C \longrightarrow \mathbb{R}$ und $g\colon C \longrightarrow \mathbb{R}^l$ sind (komponentenweise) konvex.

   Ferner sei die Slatersche Constraint Qualification erfüllt, d. h. es existiere ein $\hat{x} \in C$ mit $g(\hat{x}) < 0$. Man zeige: Ist (P) zulässig und $\inf(\mathrm{P}) > -\infty$, so ist die Menge N_{opt} der Lösungen des zu (P) dualen Programms

   (D) Maximiere $\phi(u) := \inf\limits_{x \in C} L(x, u)$ auf $N := \{u \in \mathbb{R}^l : u \geq 0,\ \phi(u) > -\infty\}$

   nichtleer und kompakt. Hierbei ist $L(x, u) := f(x) + u^T g(x)$ die zu (P) gehörende Lagrange-Funktion.

5. Gegeben sei die Aufgabe

   (P) Minimiere $f(x) := c^T x + \frac{1}{2} x^T Q x$ auf $M := \{x \in \mathbb{R}^n : \frac{1}{2}\|x\|_2^2 \leq \frac{1}{2}\Delta^2\}$,

   wobei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit ist und $\Delta > 0$. Man stelle das zu (P) duale Programm (D) auf und zeige, dass (P) und (D) lösbar sind und $\max(\mathrm{D}) = \min(\mathrm{P})$ gilt.

6. Unter der Voraussetzung

   (V) $C \subset \mathbb{R}^n$ ist nichtleer und konvex, $f\colon C \longrightarrow \mathbb{R}$ und $g\colon C \longrightarrow \mathbb{R}^l$ sind (komponentenweise) konvex, $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ist affin linear

   betrachte man das konvexe Programm

   (P) Minimiere $f(x)$ $M := \{x \in \mathbb{R}^n : x \in C,\ g(x) \leq 0,\ h(x) = 0\}$.

   Ein Tripel $(x^*, u^*, v^*) \in C \times \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ nennen wir einen *Sattelpunkt* der Lagrange-Funktion $L(x, u, v) := f(x) + u^T g(x) + v^T h(x)$, wenn

   $$\begin{cases} L(x^*, u, v) \leq L(x^*, u^*, v^*) \leq L(x, u^*, v^*) \\ \quad \text{für alle} \quad (x, u, v) \in C \times \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m. \end{cases}$$

   Man zeige:

   (a) Ist $x^* \in M$ eine Lösung von (P) und ist die Slatersche Constraint Qualification aus dem starken Dualitätssatz 2.3 erfüllt, so existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ derart, dass (x^*, u^*, v^*) ein Sattelpunkt von L ist.

   (b) Ist $(x^*, u^*, v^*) \in C \times \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ ein Sattelpunkt von L, so ist x^* eine Lösung von (P).

7. Seien $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$. Hiermit betrachte man die zueinander dualen linearen Programme

   (P) Minimiere $c^T x$ auf $M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b\}$

   und

   (D) Maximiere $b^T y$ auf $N := \{y \in \mathbb{R}^m : A^T y \leq c\}$.

   Es werde vorausgesetzt, daß

   $$M_0 := \{x \in \mathbb{R}^n : x > 0,\ Ax = b\} \neq \emptyset, \qquad N_0 := \{y \in \mathbb{R}^m : A^T y < c\} \neq \emptyset$$

   und Rang $(A) = m$. Man zeige, dass dann die Mengen M_{opt} und N_{opt} der optimalen Lösungen von (P) bzw. (D) nichtleer und kompakt sind.

8. Gegeben sei ein Vektor $x = (x_j) \in \mathbb{R}^n$ und $r \in \{1, \ldots, n\}$. Sei $p = \{p_1, \ldots, p_n\}$ eine Permutation von $\{1, \ldots, n\}$ mit $x_{p_1} \geq \cdots \geq x_{p_n}$. Man zeige, dass

   $$\sum_{j=1}^{r} x_{p_j} = \max\{x^T z : 0 \leq z \leq e,\ e^T z = r\},$$

   wobei e der Vektor im $\mathbb{R}^n$ ist, dessen Komponenten alle gleich 1 sind.

9. Gegeben sei das lineare Programm

   (P) $$\text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b\}.$$

   Hierbei seien $A \in \mathbb{R}^{m\times n}$ mit $m < n$ und Rang $(A) = m$ sowie $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ gegeben.

   (a) Man zeige, dass eine Matrix $B \in \mathbb{R}^{(n-m)\times n}$ mit Rang $(B) = n - m$ und $AB^T = 0$ existiert. Sind B_1 und B_2 zwei Matrizen mit diesen beiden Eigenschaften, so existiert eine nichtsinguläre Matrix $T \in \mathbb{R}^{(n-m)\times(n-m)}$ mit $B_1 = TB_2$.

   (b) Sei $B \in \mathbb{R}^{(n-m)\times n}$ wie in (a) gegeben, ferner sei $d := A^T(AA^T)^{-1}b$. Hiermit betrachte man das (von der Wahl von B unabhängige) lineare Programm

   (D) $$\text{Minimiere} \quad d^T y \quad \text{auf} \quad N := \{y \in \mathbb{R}^n : y \geq 0,\ By = Bc\}.$$

   Man begründe, weshalb (D) mit einigem Recht als zu (P) duales Programm bezeichnet werden kann, und beweise insbesondere einen schwachen und einen starken Dualitätssatz:

   (i) Sind $x \in M$ und $y \in N$, so ist $x^T y \geq 0$.

   (ii) Sind (P) und (D) zulässig, so besitzen beide Programme Lösungen $x^* \in M$ bzw. $y^* \in N$ und es ist $(x^*)^T y^* = 0$.

10. Gegeben seien symmetrische, positiv semidefinite Matrizen $A_1, \ldots, A_m \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n \setminus \{0\}$ und $v > 0$. Es wird vorausgesetzt, dass die Matrix $A(y) := \sum_{i=1}^m y_i A_i$ für jedes $y > 0$ positiv definit ist. Man betrachte die beiden Probleme

   (P_1)
   $$\left\{ \begin{array}{c} \text{Minimiere} \quad \dfrac{1}{2} c^T x \quad \text{auf} \\ P := \left\{ (x,y) \in \mathbb{R}^n \times \mathbb{R}^m : \displaystyle\sum_{i=1}^m y_i A_i x = c,\ e^T y = v,\ y \geq 0 \right\} \end{array} \right.$$

   und

   (P_2)
   $$\left\{ \begin{array}{c} \text{Minimiere} \quad \delta \quad \text{auf} \\ M := \left\{ (z,\delta) \in \mathbb{R}^n \times \mathbb{R} : \dfrac{v}{2} z^T A_i z - c^T z - \delta \leq 0,\ i = 1, \ldots, m \right\}. \end{array} \right.$$

   Man beachte, dass (P_2) eine konvexe, quadratisch restringierte Optimierungsaufgabe mit einer linearen Zielfunktion ist. Man zeige[11]:

   (a) Die beiden Optimierungsaufgaben (P_1) und (P_2) sind zulässig.

   (b) Sei $(x,y) \in P$ zulässig für (P_1) und $(z,\delta) \in M$ zulässig für (P_2). Dann ist $\delta \geq -\frac{1}{2} c^T x$. Hieraus schließe man, dass (P_2) lösbar ist und $\inf(\mathrm{P}_1) \geq -\min(\mathrm{P}_2)$ gilt.

   (c) Man zeige, dass die Slatersche Constraint Qualification für das Programm (P_2) erfüllt ist. Hieraus schließe man, dass das zu (P_2) duale Programm (D_2) lösbar ist und keine Dualitätslücke auftritt, also $\min(\mathrm{P}_2) = \max(\mathrm{D}_2)$ gilt.

[11] Ähnliche Aussagen werden bei

A. Ben-Tal, M. P. Bendsøe (1993) A new method for optimal truss topology design. SIAM J. Optim. 3, 322–358

gemacht.

(d) Sei u^* eine Lösung des zu (P_2) dualen Programms. Man setze $y^* := vu^*$ und zeige die Existenz eines $x^* \in \mathbb{R}^n$ mit der Eigenschaft, dass (x^*, y^*) eine Lösung von (P_1) ist.

2.3 Notwendige und hinreichende Optimalitätsbedingungen

2.3.1 Notwendige Optimalitätsbedingungen erster Ordnung

Wir betrachten die Optimierungsaufgabe

$$(P) \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\},$$

und wollen in diesem Unterabschnitt notwendige (Optimalitäts-) Bedingungen erster Ordnung (d. h. es treten nur Ableitungen erster Ordnung auf) dafür angeben, dass ein $x^* \in M$ eine lokale Lösung von (P) ist. Hierbei werden wir uns auf die Untersuchung glatter Probleme konzentrieren, also etwa voraussetzen, dass die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ in x^* stetig differenzierbar sind.

Von entscheidender Bedeutung bei der Herleitung notwendiger Optimalitätsbedingungen erster Ordnung ist der Begriff des Tangentialkegels.

Definition 3.1 Sei $M \subset \mathbb{R}^n$ und $x^* \in M$. Dann heißt

$$T(M; x^*) := \left\{ p \in \mathbb{R}^n : \begin{array}{c} \text{Es existieren Folgen } \{t_k\} \subset \mathbb{R}_+,\ \{r_k\} \subset \mathbb{R}^n \text{ mit} \\ x^* + t_k p + r_k \in M \text{ für alle } k,\ t_k \to 0,\ r_k/t_k \to 0. \end{array} \right\}$$

der *Tangentialkegel* an M in x^*. Ein Element $p \in T(M; x^*)$ heißt *Tangentialrichtung* an M in x^*.

Es ist leicht zu erklären, weshalb der Tangentialkegel, insbesondere bei nichtlinear restringierten Optimierungsaufgaben, von so großer Bedeutung bei der Gewinnung notwendiger Optimalitätsbedingungen ist. Denn sei $x^* \in M$ eine lokale Lösung von (P), so dass eine Umgebung U^* von x^* mit $f(x^*) \leq f(x)$ für alle $x \in U^* \cap M$ existiert. Ist $p \in T(M; x^*)$ und sind $\{t_k\} \subset \mathbb{R}_+$, $\{r_k\} \subset \mathbb{R}^n$ zugehörige Folgen, so ist $x^* + t_k p + r_k \in U^* \cap M$ für alle hinreichend großen k und daher $f(x^*) \leq f(x^* + t_k p + r_k)$ für alle hinreichend großen k. Wegen

$$\lim_{k \to \infty} \frac{f(x^* + t_k p + r_k) - f(x^*)}{t_k} = \nabla f(x^*)^T p$$

folgt die sehr allgemeine (und nicht von der speziellen Struktur von M abhängende) notwendige Optimalitätsbedingung erster Ordnung:

- *Sei $x^* \in M$ eine lokale Lösung der Optimierungsaufgabe*

  $$(P) \qquad \textit{Minimiere} \quad f(x), \quad x \in M.$$

  Die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sei in x^* stetig differenzierbar. Dann ist

  $$\nabla f(x^*)^T p \geq 0 \qquad \textit{für alle } p \in T(M; x^*).$$

Bemerkung: Ist $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ eine Abbildung, die auf einer Umgebung von $x^* \in \mathbb{R}^n$ definiert ist, so heißt eine Abbildung $f'(x^*;\cdot)\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ *Hadamard-Variation* von f in x^*, wenn für alle $p \in \mathbb{R}^n$ gilt: Sind $\{t_k\} \subset \mathbb{R}_+$ und $\{r_k\} \subset \mathbb{R}^n$ Folgen mit $t_k \to 0$, $r_k/t_k \to 0$, so ist

$$f'(x^*;p) = \lim_{k\to\infty} \frac{f(x^* + t_k p + r_k) - f(x^*)}{t_k}.$$

Ist f in x^* stetig (partiell) differenzierbar, so besitzt f in x^* eine Hadamard-Variation und es ist $f'(x^*;p) = \nabla f(x^*)^T p$ für alle $p \in \mathbb{R}^n$. Genau diese Aussage haben wir oben ausgenutzt. Allgemeiner folgt aus der Existenz des Fréchet-Differentials auch die der Hadamard-Variation (Beweis?). □

Neben dem Tangentialkegel spielt in der Optimierung ein weiterer Kegel, nämlich der *Kegel der zulässigen Richtungen*, eine wesentliche Rolle.

Definition 3.2 Sei $M \subset \mathbb{R}^n$ und $x^* \in M$. Dann heißt

$$F(M;x^*) := \left\{ p \in \mathbb{R}^n : \begin{array}{l} \text{Es existiert eine Folge } \{t_k\} \subset \mathbb{R}_+ \text{ mit} \\ t_k \to 0 \text{ und } x^* + t_k p \in M \text{ für alle } k \end{array} \right\}$$

der *Kegel der zulässigen Richtungen* an M in x^*.

Natürlich ist $F(M;x^*) \subset T(M;x^*)$. Treten aber im betrachteten Optimierungsproblem insbesondere nichtlineare Gleichungen als Restriktionen auf, so ist i. Allg. $F(M;x^*) = \{0\}$, der Kegel der zulässigen Richtungen also trivial.

In einem einfachen Lemma wollen wir einige Eigenschaften des Tangentialkegels einer Menge M in einem Punkt $x^* \in M$ zusammenstellen.

Lemma 3.3 *Sei $M \subset \mathbb{R}^n$ und $x^* \in M$. Mit $F(M;x^*)$ sei der Kegel der zulässigen Richtungen und mit $T(M;x^*)$ der Tangentialkegel an M in x^* bezeichnet. Dann gilt:*

1. *Es ist $T(M;x^*)$ ein nichtleerer, abgeschlossener Kegel, der $\operatorname{cl} F(M;x^*)$ enthält.*
2. *Ist M konvex, so ist $T(M;x^*) = \operatorname{cl} F(M;x^*)$ und*

$$F(M;x^*) = \{\lambda(x - x^*) : \lambda > 0,\ x \in M\}.$$

In diesem Falle ist der Tangentialkegel ebenfalls konvex.

Beweis: Natürlich werden $T(M;x^*)$ und $F(M;x^*)$ zu Recht als Kegel bezeichnet, denn mit einer Richtung gehört auch jedes nichtnegative Vielfache zu der entsprechenden Menge.

Wir zeigen, dass der Tangentialkegel $T(M;x^*)$ abgeschlossen ist. Wegen $F(M;x^*) \subset T(M;x^*)$ ist dann auch $\operatorname{cl} F(M;x^*) \subset T(M;x^*)$. Sei hierzu $\{p^{(j)}\} \subset T(M;x^*)$ eine gegen $p \in \mathbb{R}^n$ konvergente Folge. Nach Definition des Tangentialkegels existieren zu jedem $j \in \mathbb{N}$ Folgen $\{t_k^{(j)}\} \subset \mathbb{R}_+$ und $\{r_k^{(j)}\} \subset \mathbb{R}^n$ mit

$$x^* + t_k^{(j)} p^{(j)} + r_k^{(j)} \in M \quad \text{für alle } k$$

und

$$\lim_{k\to\infty} t_k^{(j)} = 0, \qquad \lim_{k\to\infty} \frac{r_k^{(j)}}{t_k^{(j)}} = 0.$$

Zu jedem $j \in \mathbb{N}$ existiert ein $k(j) \in \mathbb{N}$ mit

$$0 < t_k^{(j)} \le \frac{1}{j}, \qquad \frac{\|r_k^{(j)}\|}{t_k^{(j)}} \le \frac{1}{j} \qquad \text{für alle } k \ge k(j).$$

Nun definiere man die Folgen $\{t_j\} \subset \mathbb{R}_+$ und $\{r_j\} \subset \mathbb{R}^n$ durch

$$t_j := t_{k(j)}^{(j)}, \qquad r_j := r_{k(j)}^{(j)} + t_{k(j)}^{(j)}(p^{(j)} - p).$$

Dann ist

$$x^* + t_j p + r_j = x^* + t_{k(j)}^{(j)} p^{(j)} + r_{k(j)}^{(j)} \in M \qquad \text{für alle } j \in \mathbb{N}$$

und

$$t_j = t_{k(j)}^{(j)} \to 0, \qquad \frac{r_j}{t_j} = \underbrace{\frac{r_{k(j)}^{(j)}}{t_{k(j)}^{(j)}}}_{\to 0} + \underbrace{p^{(j)} - p}_{\to 0} \to 0.$$

Insgesamt ist damit $p \in T(M; x^*)$, die Abgeschlossenheit des Tangentialkegels $T(M; x^*)$ ist damit bewiesen.

Den Beweis des zweiten Teiles des Lemmas überlassen wir als Übungsaufgabe. □□

Allgemein (d. h. für nichtkonvexes M) kann man nicht hoffen, dass der Tangentialkegel $T(M; x^*)$ konvex ist. Daher sind wir daran interessiert, „möglichst große“ konvexe Teilmengen von $T(M; x^*)$ zu bestimmen. In der Einführung hatten wir schon eine Vermutung geäußert. Das entsprechende Resultat fassen wir in dem folgenden Satz zusammen, wobei wir insbesondere an die Restriktionsabbildung g der Ungleichungen unnötig starke Glattheitsanforderungen stellen.

Satz 3.4 *Seien $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ auf einer Umgebung von $x^* \in M := \{x \in \mathbb{R}^n : g(x) \le 0,\ h(x) = 0\}$ stetig differenzierbar. Mit*

$$I(x^*) := \{i \in \{1, \dots, l\} : g_i(x^*) = 0\}$$

werde die Indexmenge der aktiven Ungleichungsrestriktionen bezeichnet. Es werde vorausgesetzt:

(a) Es existiert ein $\hat{p} \in \mathbb{R}^n$ mit $\nabla g_i(x^*)^T \hat{p} < 0$ für alle $i \in I(x^*)$ und $h'(x^*)\hat{p} = 0$.

(b) Die Vektoren $\nabla h_1(x^*), \dots, \nabla h_m(x^*)$ sind linear unabhängig bzw. Rang $h'(x^*) = m$.

Dann ist

$$L_0(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \le 0 \ \ (i \in I(x^*)),\ h'(x^*)p = 0\} \subset T(M; x^*).$$

Beweis: Nach Voraussetzung ist

$$L_+(M;x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p < 0 \ \ (i \in I(x^*)),\ h'(x^*)p = 0\} \neq \emptyset.$$

Wir werden uns nun überlegen:

- Zu vorgegebenem $p \in L_+(M;x^*)$ existieren ein $\epsilon > 0$ und eine Abbildung $x\colon (-\epsilon,\epsilon) \longrightarrow \mathbb{R}^n$ mit $x(t) \in M$ für alle $t \in (0,\epsilon)$ und $\lim_{t\to 0}[x(t) - x^*]/t = p$.

Ist es gelungen, diese Hilfsbehauptung zu beweisen, so ist natürlich $L_+(M;x^*) \subset T(M;x^*)$, denn mit $r(t) := x(t) - x^* - tp$ ist $x^* + tp + r(t) \in M$ für alle $t \in (0,\epsilon)$ und $r(t) = o(t)$. Wegen der Abgeschlossenheit von Tangentialkegeln ist folglich

$$L_0(M;x^*) \subset \operatorname{cl} L_+(M;x^*) \subset T(M;x^*),$$

die Behauptung also bewiesen.

Zum Beweis der obigen Hilfsbehauptung geben wir uns ein $p \in L_+(M;x^*)$ vor, o. B. d. A. ist $\|p\|_2 = 1$. Zunächst berücksichtigen wir nur, dass $p \in \operatorname{Kern} h'(x^*)$. Wegen $\operatorname{Rang} h'(x^*) = m$ ist $\operatorname{Kern} h'(x^*)$ ein $(n-m)$-dimensionaler Teilraum des $\mathbb{R}^n$. Man ergänze p durch $b_1,\ldots,b_{n-m-1}$ zu einer orthonormalen Basis von $\operatorname{Kern} h'(x^*)$ und definiere $B \in \mathbb{R}^{n\times(n-m-1)}$ durch $B := (\ b_1 \ \cdots \ b_{n-m-1}\ )$ sowie die Abbildung $T\colon \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}^n$ durch

$$T(x,t) := \begin{pmatrix} h(x) \\ B^T x - B^T x^* \\ p^T x - p^T x^* - t \end{pmatrix}.$$

Dann ist $T(x^*,0) = 0$, ferner ist die Funktionalmatrix $T'_x(x,t)$ von T bezüglich x gegeben durch

$$T'_x(x,t) = \begin{pmatrix} h'(x) \\ B^T \\ p^T \end{pmatrix}.$$

Man stellt nun leicht fest, dass $T'_x(x^*,0) \in \mathbb{R}^{n\times n}$ nichtsingulär ist. Denn ist

$$T'_x(x^*,0)q = \begin{pmatrix} h'(x^*)q \\ B^T q \\ p^T q \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

so ist wegen der ersten Gleichung $q \in \operatorname{Kern} h'(x^*)$, andererseits wegen der letzten beiden Gleichungen $q \in [\operatorname{Kern} h'(x^*)]^\perp$, insgesamt also $q = 0$. Der Satz über implizite Funktionen liefert die Existenz eines $\epsilon > 0$ und einer auf $(-\epsilon,\epsilon)$ stetig differenzierbaren Abbildung $x\colon (-\epsilon,\epsilon) \longrightarrow \mathbb{R}^n$ mit $x(0) = x^*$ und $T(x(t),t) = 0$ für alle t mit $|t| < \epsilon$. Als Ableitung von $x(\cdot)$ in $t = 0$ berechnet man

$$x'(0) = -T'_x(x^*,0)^{-1}T'_t(x^*,0) = -\begin{pmatrix} h'(x^*) \\ B^T \\ p^T \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} = p.$$

Damit ist $x(t) = x^* + tp + r(t)$ mit $r(t) := x(t) - x^* - tp = o(t)$ und $h(x(t)) = 0$ für alle $t \in (-\epsilon,\epsilon)$. Indem man ϵ notfalls verkleinert, kann man erreichen, dass auch

$g(x(t)) \leq 0$ für alle $t \in (0, \epsilon)$. Um dies einzusehen, können die in x^* inaktiven Ungleichungsrestriktionen offensichtlich außer Acht gelassen werden. Sei daher $i \in I(x^*)$ eine in x^* aktive Ungleichungsrestriktion. Wegen

$$\lim_{t \to 0+} \frac{g_i(x^* + tp + r(t)) - g_i(x^*)}{t} = \nabla g_i(x^*)^T p < 0$$

und $g_i(x^*) = 0$ ist $g_i(x^* + tp + r(t)) \leq 0$ für alle hinreichend kleinen $t > 0$. Insgesamt ist die Hilfsbehauptung und damit der ganze Satz bewiesen. □ □

Bemerkung: Die Zusatzbedingung in Satz 3.4, also die Existenz eines $\hat{p} \in \mathbb{R}^n$ mit $\nabla g_i(x^*)^T \hat{p} < 0$ für alle $i \in I(x^*)$ sowie $h'(x^*)\hat{p} = 0$ und die lineare Unabhängigkeit von $\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ nennt man die *Arrow-Hurwicz-Uzawa* (oder auch *Mangasarian-Fromowitz*) Constraint Qualification. Hinreichend für die Gültigkeit der Arrow-Hurwicz-Uzawa Constraint Qualification ist offenbar, dass die Vektoren $\nabla g_i(x^*)$, $i \in I(x^*)$, $\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ linear unabhängig sind. Durch ein Beispiel hatten wir schon in der Einführung gezeigt, dass ohne eine Constraint Qualification i. Allg. $L_0(M; x^*) \not\subset T(M; x^*)$. □

Jetzt ist es einfach, die notwendigen Optimalitätsbedingungen erster Ordnung aufzustellen. Den folgenden Satz nennt man auch den Satz von Kuhn-Tucker (jetzt häufiger auch: Karush-Kuhn-Tucker).

Satz 3.5 *Sei x^* eine lokale Lösung von*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad \textit{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

Hierbei seien die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf einer Umgebung von x^* stetig differenzierbar. Es gelte die Arrow-Hurwicz-Uzawa Constraint Qualification, d. h. es existiere ein $\hat{p} \in \mathbb{R}^n$ mit $\nabla g_i(x^*)^T \hat{p} < 0$ für alle $i \in I(x^*)$ und $h'(x^*)\hat{p} = 0$, ferner sei Rang $h'(x^*) = m$. *Dann existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit*

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

Beweis: Wegen Satz 3.4 ist $L_0(M; x^*) \subset T(M; x^*)$, da außerdem $x^* \in M$ eine lokale Lösung von (P) ist, ist $\nabla f(x^*)^T p \geq 0$ für alle $p \in L_0(M; x^*)$. Mit anderen Worten ist $p^* := 0$ eine Lösung der linearen Optimierungsaufgabe

$$\text{(LP)} \qquad \begin{cases} \text{Minimiere} \quad \nabla f(x^*)^T p \quad \text{auf} \\ L_0(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \leq 0 \ \ (i \in I(x^*)),\ h'(x^*)p = 0\}. \end{cases}$$

Das hierzu duale Programm ist wegen des starken Dualitätssatzes der linearen Optimierung zulässig, so dass $u_i^* \in \mathbb{R}$, $i \in I(x^*)$, und $v^* \in \mathbb{R}^m$ mit

$$u_i^* \geq 0 \quad (i \in I(x^*)), \qquad \nabla f(x^*) + \sum_{i \in I(x^*)} u_i^* \nabla g_i(x^*) + h'(x^*)^T v^* = 0$$

existieren. Ergänzt man die u_i^*, $i \in I(x^*)$, noch zu einem Vektor $u^* \in \mathbb{R}^l$, indem man $u_i^* := 0$ für $i \notin I(x^*)$ setzt, so hat man das gesuchte Paar (u^*, v^*) gefunden. □ □

Bemerkung: Ist die Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ zur Optimierungsaufgabe (P) wie üblich durch

$$L(x,u,v) := f(x) + u^T g(x) + v^T h(x)$$

definiert, so sagt Satz 3.5 gerade aus, dass es zu einer lokalen Lösung x^* (bei erfüllter Constraint Qualification) ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$(*) \qquad u^* \geq 0, \qquad \nabla_x L(x^*, u^*, v^*) = 0, \qquad (u^*)^T g(x^*) = 0$$

gibt. Ein Tripel $(x^*, u^*, v^*) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ mit $x^* \in M$ und $(*)$ nennt man auch einen *Kuhn-Tucker-Punkt* (oder auch *Karush-Kuhn-Tucker-Punkt*) zu (P). □

Bemerkung: Eine fast triviale, aber gelegentlich nützliche Bemerkung ist die folgende: Ist eine Optimierungsaufgabe der Form

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\} \cap X_0$$

gegeben, wobei $X_0 \subset \mathbb{R}^n$ eine *offene* Menge ist, ist ferner x^* eine lokale Lösung von (P) und gelten sonst alle weiteren Voraussetzungen von Satz 3.5, so kann auch in diesem Fall die Existenz eines Paares (u^*, v^*) mit den angegebenen Eigenschaften garantiert werden. Eine offene Nebenbedingung spielt sozusagen lokal keine Rolle. □

Im Beweis von Satz 3.4 treten die einzigen Komplikationen durch die i. Allg. nichtlinearen Gleichungsnebenbedingungen auf. Hat man nur Ungleichungen (und eventuell noch lineare Gleichungen) als Nebenbedingungen, so ist die Analyse sehr viel einfacher, außerdem können die gestellten Glattheitsvoraussetzungen und die Constraint Qualification abgeschwächt werden. Hierauf wollen wir jetzt noch, zum Schluss der Untersuchungen zu notwendigen Optimalitätsbedingungen erster Ordnung, eingehen.

Satz 3.6 *Sei x^* eine lokale Lösung von*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad \textit{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

Hierbei sei $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear, $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ in x^* stetig differenzierbar. Die Indexmenge $I(x^*)$ der in x^* aktiven Ungleichungsrestriktionen wird zerlegt in

$$\begin{aligned} I_L(x^*) &:= \{i \in I(x^*) : g_i \textit{ ist affin linear}\}, \\ I_N(x^*) &:= \{i \in I(x^*) : g_i \textit{ ist nicht affin linear}\} \end{aligned}$$

und vorausgesetzt, daß

$$\hat{L}_+(M; x^*) := \left\{ p \in \mathbb{R}^n : \begin{array}{l} \nabla g_i(x^*)^T p \leq 0 \ \ (i \in I_L(x^*)), \\ \nabla g_i(x^*)^T p < 0 \ \ (i \in I_N(x^*)), \end{array} \ h'(x^*)p = 0 \right\} \neq \emptyset.$$

Dann existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

Beweis: Wir zeigen, dass $\hat{L}_+(M;x^*) \subset F(M;x^*)$. Ist dies gelungen, so ist mit

$$L_0(M;x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \leq 0 \ \ (i \in I(x^*)),\ h'(x^*)p = 0\}$$

wieder

$$L_0(M;x^*) \subset \operatorname{cl} \hat{L}_+(M;x^*) \subset \operatorname{cl} F(M;x^*) \subset T(M;x^*),$$

so dass man einen Ersatz für Satz 3.4 hat und im Beweis fortfahren kann wie im Beweis von Satz 3.5. Sei also $p \in \hat{L}(M;x^*)$. Wir zeigen, dass $x^* + tp \in M$ für alle hinreichend kleinen $t > 0$, woraus natürlich $p \in F(M;x^*)$ folgt. Für alle t ist

$$h(x^* + tp) = \underbrace{h(x^*)}_{=0} + t\,\underbrace{h'(x^*)p}_{=0} = 0,$$

wobei wir ausgenutzt haben, dass h affin linear ist. Bei den Ungleichungsrestriktionen brauchen wieder nur die in x^* aktiven betrachtet zu werden. Für $i \in I_L(x^*)$ ist

$$g_i(x^* + tp) = \underbrace{g_i(x^*)}_{=0} + t\,\underbrace{\nabla g_i(x^*)^T p}_{\leq 0} \leq 0,$$

während für $i \in I_N(x^*)$ wegen

$$\lim_{t \to 0+} \frac{g_i(x^* + tp) - g_i(x^*)}{t} = \nabla g_i(x^*)^T p < 0$$

unter Berücksichtigung von $g_i(x^*) = 0$ jedenfalls $g_i(x^* + tp) \leq 0$ für alle hinreichend kleinen $t > 0$ gilt. Damit ist der Satz bewiesen. □ □

In einem Korollar betrachten wir einen Spezialfall von Satz 3.6, in dem zusätzlich die Restriktionsabbildung $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ komponentenweise konvex ist.

Korollar 3.7 *Sei x^* eine lokale Lösung von*

(P) *Minimiere* $f(x)$ *auf* $M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$.

Hierbei sei $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear, $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ komponentenweise konvex und in x^* stetig differenzierbar. Es existiere ein $\hat{x} \in M$ mit $g_i(\hat{x}) < 0$ für alle $i \in \{1,\ldots,l\}$, für die g_i nicht affin linear ist. Dann existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

Beweis: Wir wenden Satz 3.6 an und benutzen die dort eingeführten Bezeichnungen. Wir zeigen, dass $\hat{x} - x^* \in \hat{L}_+(M;x^*)$. Unter Ausnutzung der Konvexität der g_i ist

$$\nabla g_i(x^*)^T(\hat{x} - x^*) \leq g_i(\hat{x}) - g_i(x^*) \begin{cases} \leq 0, & \text{falls} \quad i \in I_L(x^*), \\ < 0, & \text{falls} \quad i \in I_N(x^*), \end{cases}$$

ferner

$$h'(x^*)(\hat{x} - x^*) = \underbrace{h(\hat{x})}_{=0} - \underbrace{h(x^*)}_{=0} = 0.$$

Die Behauptung folgt dann aus Satz 3.6. □

Bemerkung: Satz 3.6 (oder auch sein Korollar) zeigen, dass bei linearen Restriktionen keine zusätzliche Constraint Qualification nötig ist, um die Existenz von Lagrange-Multiplikatoren (u^*, v^*) zu sichern. Dies ist eine ganz wichtige Bemerkung, die oft benutzt wird. □

Bemerkung: Wir betrachten das konvexe, quadratisch restringierte quadratische Programm

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad f(x) := c_0^T x + \frac{1}{2} x^T Q_0 x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : g_i(x) := \beta_i + c_i^T x + \frac{1}{2} x^T Q_i x \leq 0,\ i = 1, \ldots, l,\ Ax = b\}. \end{cases}$$

Hierbei seien $Q_0, \ldots, Q_l \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit, ferner Q_0 sogar positiv definit. Als duales Problem hierzu hatten wir schon früher erhalten:

$$\text{(D)} \quad \begin{cases} \text{Maximiere} \quad \phi(u, v) := \beta^T u + b^T v - \frac{1}{2}[c(u) - A^T v]^T Q(u)^{-1}[c(u) - A^T v] \quad \text{auf} \\ N := \{(u, v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0\}. \end{cases}$$

Hierbei ist zur Abkürzung

$$\beta := \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_l \end{pmatrix}, \qquad Q(u) := Q_0 + \sum_{i=1}^{l} u_i Q_i, \qquad c(u) := c_0 + \sum_{i=1}^{l} u_i c_i$$

gesetzt worden. Wir wollen uns nun überlegen:

- Ist $(u^*, v^*) \in N$ eine Lösung von (D), so ist durch

$$x^* := -Q(u^*)^{-1}[c(u^*) - A^T v^*]$$

die (notwendigerweise eindeutige) Lösung von (P) gegeben.

Denn: Zunächst berechnen wir

$$\nabla \phi(u, v) = \begin{pmatrix} \nabla_u \phi(u, v) \\ \nabla_v \phi(u, v) \end{pmatrix}$$

für $(u, v) \in N$. Zur Abkürzung setzen wir

$$x(u, v) := -Q(u)^{-1}[c(u) - A^T v], \qquad C := \begin{pmatrix} c_1 & \cdots & c_l \end{pmatrix} \in \mathbb{R}^{n \times l}$$

und

$$P(u, v) := \begin{pmatrix} Q_1 x(u, v) & \cdots & Q_l x(u, v) \end{pmatrix} \in \mathbb{R}^{n \times l}.$$

Dann erhält man nach einer einfachen (wenn man sich ungeschickt anstellt eventuell mühsamen) Rechnung, daß

$$\nabla_u \phi(u, v) = \beta + [C^T + \tfrac{1}{2} P(u, v)^T] x(u, v), \qquad \nabla_v \phi(u, v) = b - Ax(u, v).$$

Da $(u^*, v^*) \in N$ nach Voraussetzung eine Lösung von (D) ist, liefern die notwendigen Optimalitätsbedingungen, angewandt auf das duale Programm (D), dass (man beachte, dass hierzu nicht das Erfülltsein einer Constraint Qualification nötig ist, da es sich bei (D) um ein linear restringiertes Programm handelt)

$$\nabla_u \phi(u^*, v^*) \leq 0, \qquad (u^*)^T \nabla_u \phi(u^*, v^*) = 0, \qquad \nabla_v \phi(u^*, v^*) = 0.$$

Berücksichtigt man die oben angegebene Form des Gradienten $\nabla\phi(u, v)$ und $x^* = x(u^*, v^*)$, so erhält man

$$g(x^*) = \nabla_u \phi(u^*, v^*) \leq 0, \qquad b - Ax^* = \nabla_v \phi(u^*, v^*) = 0.$$

Also ist $x^* \in M$ zulässig für (P), ferner

$$(u^*)^T g(x^*) = (u^*)^T \nabla_u \phi(u^*, v^*) = 0.$$

Wegen $\nabla_x L(x^*, u^*, v^*) = 0$ (mit $L(x, u, v) := f(x) + u^T g(x) + v^T(b - Ax)$ wird wieder die Lagrange-Funktion zu (P) bezeichnet) ist schließlich

$$\phi(u^*, v^*) = L(x^*, u^*, v^*) = f(x^*) + \underbrace{(u^*)^T g(x^*)}_{=0} + (v^*)^T \underbrace{(b - Ax^*)}_{=0} = f(x^*).$$

Der schwache Dualitätssatz liefert damit, dass x^* Lösung von (P) ist. □

2.3.2 Notwendige Optimalitätsbedingungen zweiter Ordnung

Ist $x^* \in \mathbb{R}^n$ das lokale Minimum einer reellwertigen Funktion f, die in x^* zweimal stetig differenzierbar ist, so lauten die notwendigen Optimalitätsbedingungen zweiter Ordnung bekanntlich, dass zum einen $\nabla f(x^*) = 0$ (notwendige Optimalitätsbedingung erster Ordnung), zum anderen $\nabla^2 f(x^*)$ positiv semidefinit ist. Diesen Sachverhalt gilt es nun auf restringierte Optimierungsaufgaben zu übertragen. Im folgenden Satz geben wir ein entsprechendes Ergebnis an.

Satz 3.8 *Gegeben sei die Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad \textit{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

Seien $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ auf einer Umgebung der lokalen Lösung $x^* \in M$ von (P) *zweimal stetig differenzierbar. Mit $I(x^*)$ wird wieder die Indexmenge der in x^* aktiven Ungleichungsrestriktionen bezeichnet. Es werde vorausgesetzt, dass $\nabla g_i(x^*)$, $i \in I(x^*)$, $\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ linear unabhängig sind. Dann existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit*

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad g(x^*)^T u^* = 0$$

und der Eigenschaft, daß

$$p^T \Big(\nabla^2 f(x^*) + \sum_{i=1}^{l} u_i^* \nabla^2 g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla^2 h_i(x^*) \Big) p \geq 0 \qquad \textit{für alle } p \in L^0(M; x^*),$$

wobei

$$L^0(M;x^*) := \Big\{p \in \mathbb{R}^n : \begin{array}{l} \nabla g_i(x^*)^T p = 0 \ \ (i \in I_+(x^*)), \\ \nabla g_i(x^*)^T p \le 0 \ \ (i \in I(x^*) \setminus I_+(x^*)), \end{array} \ h'(x^*)p = 0\Big\}$$

mit

$$I_+(x^*) := \{i \in \{1,\dots,l\} : u_i^* > 0\}.$$

Beweis: Da die Vektoren $\nabla g_i(x^*)$, $i \in I(x^*)$, $\nabla h_1(x^*),\dots,\nabla h_m(x^*)$ nach Voraussetzung linear unabhängig sind, ist die Arrow-Hurwicz-Uzawa Constraint Qualification erfüllt. Denn z. B. existiert ein $\hat{p} \in \mathbb{R}^n$ mit $\nabla g_i(x^*)^T\hat{p} = -1$, $i \in I(x^*)$, $\nabla h_i(x^*)^T\hat{p} = 0$, $i = 1,\dots,m$. Wegen der notwendigen Bedingungen erster Ordnung (Satz 3.5) existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u^* \ge 0, \qquad \nabla f(x^*) + \sum_{i \in I(x^*)} u_i^* \nabla g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla h_i(x^*) = 0.$$

Sei $p \in L^0(M;x^*)$ beliebig vorgegeben. Wie im ersten Teil des Beweises von Satz 3.4 gezeigt wurde, existiert ein $\epsilon > 0$ und eine Abbildung $r\colon (-\epsilon,\epsilon) \longrightarrow \mathbb{R}^n$ mit $r(t) = o(t)$ und der Eigenschaft, dass $g_i(x^* + tp + r(t)) = 0$ für alle $i \in I(x^*)$ mit $\nabla g_i(x^*)^T p = 0$ (also insbesondere alle $i \in I_+(x^*)$) und $h(x^* + tp + r(t)) = 0$. Hieraus folgt aber, dass $x(t) := x^* + tp + r(t) \in M$ für alle hinreichend kleinen $t > 0$. Denn dies ist für alle in x^* inaktiven Ungleichungsrestriktionen selbstverständlich, während für alle $i \in I(x^*) \setminus I_+(x^*)$ mit $\nabla g_i(x^*)^T p \ne 0$ wegen $p \in L^0(M;x^*)$ sogar $\nabla g_i(x^*)^T p < 0$ gilt. Für das weitere beachten wir, dass

$$x(t) - x^* = t(p + r(t)/t) = O(t).$$

Da x^* eine lokale Lösung von (P) und $x(t) \in M$ für alle hinreichend kleinen $t > 0$, ist

$$\begin{aligned} 0 &\le f(x(t)) - f(x^*) \\ &= \nabla f(x^*)^T(x(t) - x^*) + \tfrac{1}{2}\,(x(t) - x^*)^T \nabla^2 f(x^*)(x(t) - x^*) + o(t^2). \end{aligned}$$

Für $i \in I_+(x^*)$ ist

$$\begin{aligned} 0 &= g_i(x(t)) \\ &= \underbrace{g_i(x^*)}_{=0} + \nabla g_i(x^*)^T(x(t) - x^*) + \tfrac{1}{2}\,(x(t) - x^*)^T \nabla^2 g_i(x^*)(x(t) - x^*) + o(t^2). \end{aligned}$$

Weiter ist für $i = 1,\dots,m$ ganz entsprechend

$$\begin{aligned} 0 &= h_i(x(t)) \\ &= \underbrace{h_i(x^*)}_{=0} + \nabla h_i(x^*)^T(x(t) - x^*) + \tfrac{1}{2}\,(x(t) - x^*)^T \nabla^2 h_i(x^*)(x(t) - x^*) + o(t^2). \end{aligned}$$

Multiplikation mit $u_i^* > 0$ bzw. v_i^* und Aufsummieren liefert

$$\begin{aligned} 0 \;\leq\; & \underbrace{\Big(\nabla f(x^*) + \sum_{i\in I_+(x^*)} u_i^* \nabla g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla h_i(x^*)\Big)^T (x(t)-x^*)}_{=0} \\ & + \tfrac{1}{2}\,(x(t)-x^*)^T \Big(\nabla^2 f(x^*) + \sum_{i\in I_+(x^*)} u_i^* \nabla^2 g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla^2 h_i(x^*)\Big) \\ & \times (x(t)-x^*) + o(t^2). \end{aligned}$$

Daher ist (es ist $u_i^* = 0$ für alle $\in \{1,\dots,l\} \setminus I_+(x^*)$)

$$0 \leq (x(t)-x^*)^T \Big(\nabla^2 f(x^*) + \sum_{i=1}^{l} u_i^* \nabla g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla h_i(x^*)\Big)(x(t)-x^*) + o(t^2).$$

Division durch t^2 ergibt unter Berücksichtigung von

$$\frac{x(t)-x^*}{t} = p + \frac{r(t)}{t} \to p \quad \text{mit} \quad t \to 0+$$

nach dem Grenzübergang $t \to 0+$ genau die Behauptung. □ □

Bemerkung: Definiert man die Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ durch

$$L(x,u,v) := f(x) + g(x)^T u + h(x)^T v,$$

so sagen die notwendigen Optimalitätsbedingungen zweiter Ordnung in Satz 3.8 aus: Ist $x^* \in M$ eine lokale Lösung von (P) und sind $\nabla g_i(x^*)$, $i \in I(x^*)$, $\nabla h_1(x^*), \dots, \nabla h_m(x^*)$ linear unabhängig, so existieren $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u^* \geq 0, \qquad \nabla_x L(x^*, u^*, v^*) = 0, \qquad g(x^*)^T u^* = 0$$

und der Eigenschaft, dass $\nabla^2_{xx} L(x^*, u^*, v^*)$ auf $L^0(M; x^*)$ positiv semidefinit ist. Insbesondere ist dann natürlich $\nabla^2_{xx} L(x^*, u^*, v^*)$ auf dem linearen Teilraum $\{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p = 0 \;\; (i \in I(x^*)),\; h'(x^*)p = 0\}$ positiv semidefinit. □

Beispiel: Wir betrachten das folgende Beispiel (siehe R. Fletcher (1987, S. 209)[12]):

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := \frac{1}{2}[(x_1-1)^2 + x_2^2] \quad \text{unter der Nebenbedingung} \\ \qquad\qquad\qquad\qquad h(x) := -x_1 + \beta x_2^2 = 0, \end{cases}$$

wobei β fest ist. Mit Hilfe der notwendigen Optimalitätsbedingungen zweiter Ordnung soll überprüft werden, für welche β durch $x^* := (0,0)$ eine lokale Lösung gegeben sein kann. Es ist

$$\nabla f(x^*) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \qquad \nabla h(x^*) = \begin{pmatrix} -1 \\ 0 \end{pmatrix},$$

[12] FLETCHER, R. (1987) *Practical Methods of Optimization. Second Edition.* John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.

so dass die notwendigen Optimalitätsbedingungen erster Ordnung mit $v^* := -1$ erfüllt sind. Weiter ist

$$L^0 := \{p \in \mathbb{R}^2 : \nabla h(x^*)^T p = 0\} = \{p = (p_1, p_2) \in \mathbb{R}^2 : p_1 = 0\},$$

und daher

$$\nabla^2 f(x^*) + v^* \nabla^2 h(x^*) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & 2\beta \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 - 2\beta \end{pmatrix}$$

auf L^0 genau dann positiv semidefinit, wenn $\beta \leq \frac{1}{2}$. Damit haben wir erhalten: Für $\beta > \frac{1}{2}$ ist $x^* = (0, 0)$ keine lokale Lösung von (P). Mit Hilfe hinreichender Optimalitätsbedingungen zweiter Ordnung werden wir zeigen können, dass $x^* = (0, 0)$ für $\beta < \frac{1}{2}$ eine lokale Lösung von (P) ist. Für $\beta = \frac{1}{2}$ müssten Bedingungen höherer Ordnung herangezogen werden. □

Zum Schluss dieses Abschnittes über notwendige Optimalitätsbedingungen zweiter Ordnung wollen wir den Spezialfall linearer Restriktionen noch etwas genauer betrachten. Bei linearen Restriktionen benötigt man zur Gewinnung notwendiger Optimalitätsbedingungen erster Ordnung keine Constraint Qualifikation und es ist zu hoffen, dass dies auch für notwendige Optimalitätsbedingungen zweiter Ordnung gilt. Das entsprechende Resultat formulieren wir im folgenden Satz.

Satz 3.9 *Gegeben sei die Optimierungsaufgabe*

(P) *Minimiere* $f(x)$ *auf* $M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$,

wobei $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ *und* $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *affin linear seien. Die Zielfunktion* f *sei in der lokalen Lösung* $x^* \in M$ *von (P) zweimal stetig differenzierbar. Dann existiert ein Paar* $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ *mit*

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad g(x^*)^T u^* = 0$$

und der Eigenschaft, daß

$$p^T \nabla^2 f(x^*) p \geq 0 \qquad \text{für alle } p \in L^0(M; x^*),$$

wobei

$$L^0(M; x^*) := \Big\{p \in \mathbb{R}^n : \begin{array}{l} \nabla g_i(x^*)^T p = 0 \ \ (i \in I_+(x^*)), \\ \nabla g_i(x^*)^T p \leq 0 \ \ (i \in I(x^*) \setminus I_+(x^*)), \end{array} \ h'(x^*)p = 0\Big\}$$

mit

$$I_+(x^*) := \{i \in \{1, \ldots, l\} : u_i^* > 0\}.$$

Beweis: Wegen Satz 3.6 existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

Zu zeigen bleibt, dass $\nabla^2 f(x^*)$ auf $L^0(M; x^*)$ positiv semidefinit ist. Sei daher $p \in L^0(M; x^*)$ beliebig vorgegeben. Da g und h affin linear sind, ist $L^0(M; x^*) \subset F(M; x^*)$

und damit p eine zulässige Richtung. Für alle hinreichend kleinen $t > 0$ ist daher mit $\theta(t) \in (0,1)$:

$$\begin{aligned}
0 &\leq \frac{f(x^*+tp)-f(x^*)}{t} \\
&= \nabla f(x^*)^T p + \tfrac{1}{2} t p^T \nabla^2 f(x^*+\theta(t)tp)p \\
&= \Big(-\sum_{i\in I_+(x^*)} u_i^* \nabla g_i(x^*) - h'(x^*)^T v^*\Big)^T p + \tfrac{1}{2} t p^T \nabla^2 f(x^*+\theta(t)tp)p \\
&= -\sum_{i\in I_+(x^*)} u_i^* \underbrace{\nabla g_i(x^*)^T p}_{=0} - (v^*)^T \underbrace{h'(x^*)p}_{=0} + \tfrac{1}{2} t p^T \nabla^2 f(x^*+\theta(t)tp)p \\
&= \tfrac{1}{2} t p^T \nabla^2 f(x^*+\theta(t)tp)p,
\end{aligned}$$

woraus nach Division durch t und Grenzübergang $t \to 0+$ die Behauptung folgt. □ □

2.3.3 Hinreichende Optimalitätsbedingungen

Bei *konvexen* Optimierungsaufgaben sind die durch den Satz von Kuhn-Tucker gegebenen notwendigen Optimalitätsbedingungen erster Ordnung auch hinreichend für Optimalität. Dieses einfache Ergebnis formulieren wir im nächsten Satz.

Satz 3.10 *Gegeben sei die Optimierungsaufgabe*

(P) *Minimiere* $f(x)$ *auf* $M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$.

Hierbei seien $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ *und* $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ *konvex sowie* $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *affin linear. Die Zielfunktion* f *und die Restriktionsabbildung* g *seien in* $x^* \in M$ *stetig differenzierbar. Existiert dann ein Paar* $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ *mit*

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad (u^*)^T g(x^*) = 0,$$

so ist x^* *eine (globale) Lösung von (P).*

Beweis: Sei $x \in M$ beliebig. Dann ist

$$\begin{aligned}
f(x) - f(x^*) &\geq \nabla f(x^*)^T (x - x^*) \\
&\qquad \text{(da } f(\cdot) \text{ konvex)} \\
&= [-g'(x^*)^T u^* - h'(x^*)^T v^*]^T (x - x^*) \\
&= -(u^*)^T g'(x^*)(x - x^*) - (v^*)^T \underbrace{h'(x^*)(x - x^*)}_{=0} \\
&\geq -(u^*)^T [g(x) - g(x^*)] \\
&\qquad \text{(da } (u^*)^T g(\cdot) \text{ konvex)} \\
&= -(u^*)^T g(x) \\
&\geq 0,
\end{aligned}$$

womit der Satz schon bewiesen ist. □ □

Beispiel: Die Zielfunktion f braucht natürlich nur dort glatt und konvex zu sein, "wo sich alles abspielt". Als Beispiel betrachten wir die Aufgabe

$$\text{(P)}\quad \text{Minimiere}\quad f(x) := \sum_{j=1}^{n}\Big(-\ln\frac{x_j}{p_j} + \frac{x_j}{p_j}\Big)\quad \text{auf}\quad M := \{x \in \mathbb{R}^n : x > 0,\ a^T x = \beta\}.$$

Hierbei sind $p = (p_j) > 0$, $a = (a_j) \neq 0$ und $\beta \in \mathbb{R}$ gegeben, ferner sei $M \neq \emptyset$. Offensichtlich ist die Zielfunktion f auf dem positiven Orthanten $\{x \in \mathbb{R}^n : x > 0\}$ konvex. Folglich ist ein $x^* \in M$ genau dann eine Lösung von (P), wenn ein $v^* \in \mathbb{R}$ mit

$$-\frac{1}{x_j^*} + \frac{1}{p_j} - v^* a_j = 0 \quad \text{bzw.} \quad x_j^* = \frac{p_j}{1 - v^* p_j a_j}, \qquad j = 1, \ldots, n,$$

existiert. Ist also v^* bekannt, so auch x^*, wobei x^* aber natürlich den Nebenbedingungen zu genügen hat. Das angegebene x^* ist genau dann positiv, wenn $v^* p_j a_j < 1$, $j = 1, \ldots, l$, bzw. $v^* \in (l, u)$ mit

$$l := \begin{cases} -\infty & \text{für} \quad a \geq 0, \\ \max\{1/(p_j a_j) : a_j < 0,\ j = 1, \ldots, n\} & \text{sonst} \end{cases}$$

und

$$u := \begin{cases} +\infty & \text{für} \quad a \leq 0, \\ \min\{1/(p_j a_j) : a_j > 0,\ j = 1, \ldots, n\} & \text{sonst.} \end{cases}$$

Nun stellt man durch eine einfache Diskussion fest, dass die Gleichung

$$g(v) := \sum_{j=1}^{n} \frac{p_j a_j}{1 - v p_j a_j} = \beta$$

in (l, u) genau eine Lösung besitzt (u. a. ist g auf (l, u) stetig und monoton wachsend). Daher besitzt (P) genau eine Lösung, deren Berechnung sich auf die Bestimmung von $v^* \in (l, u)$ mit $g(v^*) = \beta$ reduziert. □

Nun kommen wir zu den hinreichenden Optimalitätsbedingungen zweiter Ordnung. Diese verallgemeinern die aus der Analysis bzw. unrestringierten Optimierung her bekannte Tatsache, dass eine in einem Punkt $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbare Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ mit $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv definit in x^* ein isoliertes lokales Minimum besitzt. Den folgenden Satz findet man bei A. V. Fiacco, G. P. McCormick (1968, Theorem 4).

Satz 3.11 *Gegeben sei die Optimierungsaufgabe*

$$\text{(P)}\qquad \textit{Minimiere}\quad f(x)\quad \textit{auf}\quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

Die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sowie die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ seien in $x^* \in M$ zweimal stetig differenzierbar. Mit $I(x^*)$ sei die Indexmenge der in x^* aktiven Ungleichungsrestriktionen bezeichnet. Es existiere ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit:

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad (u^*)^T g(x^*) = 0$$

und

$$p^T\Big[\nabla^2 f(x^*) + \sum_{i=1}^{l} u_i^* \nabla^2 g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla^2 h_i(x^*)\Big]p > 0 \qquad \text{für alle } p \in L^0(M;x^*)\setminus\{0\},$$

wobei

$$L^0(M;x^*) := \Big\{p \in \mathbb{R}^n : \begin{array}{l} \nabla g_i(x^*)^T p = 0 \ \ (i \in I_+(x^*)), \\ \nabla g_i(x^*)^T p \le 0 \ \ (i \in I(x^*)\setminus I_+(x^*)), \end{array} \ h'(x^*)p = 0\Big\}$$

mit[13]

$$I_+(x^*) := \{i \in \{1,\dots,l\} : u_i^* > 0\}.$$

Dann ist x^* eine isolierte lokale Lösung von (P)*, d. h. es gibt eine Umgebung U^* von x^* mit $f(x^*) < f(x)$ für alle $x \in M \cap U^*$ mit $x \neq x^*$.*

Beweis: Im Widerspruch zur Behauptung nehmen wir an, es gäbe eine gegen x^* konvergente Folge $\{x_k\} \subset M$ mit $x_k \neq x^*$ und $f(x_k) \le f(x^*)$ für alle k. Es ist

$$x_k = x^* + t_k p_k \quad \text{mit} \quad t_k := \|x_k - x^*\|, \quad p_k := \frac{x_k - x^*}{\|x_k - x^*\|}.$$

Da wir notfalls zu einer Teilfolge übergehen können, kann die Konvergenz der Folge $\{p_k\}$ gegen ein $p \neq 0$ angenommen werden. Offenbar ist

$$\nabla f(x^*)^T p \le 0, \qquad \nabla g_i(x^*)^T p \le 0 \quad (i \in I(x^*)), \qquad h'(x^*)p = 0.$$

Es werden jetzt zwei Fälle betrachtet und gezeigt, dass sich jeweils ein Widerspruch ergibt.

Angenommen, es ist $\nabla g_i(x^*)^T p < 0$ für wenigstens ein $i \in I_+(x^*)$. Dann ist

$$0 \ge \nabla f(x^*)^T p = -\sum_{i\in I_+(x^*)} \underbrace{u_i^*}_{>0}\underbrace{\nabla g_i(x^*)^T p}_{\le 0} - (v^*)^T \underbrace{h'(x^*)p}_{=0} > 0,$$

ein Widerspruch.

Sei $\nabla g_i(x^*)^T p = 0$ für alle $i \in I_+(x^*)$. Durch eine Entwicklung nach Taylor erhält man

$$\begin{aligned} 0 &\ge f(x^* + t_k p_k) - f(x^*) \\ &= t_k \nabla f(x^*)^T p_k + \tfrac{1}{2}\, t_k^2 p_k^T \nabla^2 f(x_k^{(0)}) p_k \end{aligned}$$

und

$$\begin{aligned} 0 &\ge g_i(x^* + t_k p_k) \\ &= g_i(x^*) + t_k \nabla g_i(x^*)^T p_k + \tfrac{1}{2}\, t_k^2 p_k^T \nabla^2 g_i(x_{i,k}^{(1)}) p_k \\ &\qquad (i = 1,\dots,l) \end{aligned}$$

sowie

[13] Man beachte, dass $I_+(x^*)$ offenbar eine Teilmenge von $I(x^*)$, der Indexmenge der in x^* aktiven Ungleichungsrestriktionen ist.

$$\begin{aligned} 0 &= h_i(x^* + t_k p_k) \\ &= \underbrace{h_i(x^*)}_{=0} + t_k \nabla h_i(x^*)^T p_k + \tfrac{1}{2}\, t_k^2 p_k^T \nabla^2 h_i(x_{i,k}^{(2)}) p_k \\ &\qquad (i = 1, \ldots, m). \end{aligned}$$

Hierbei sind $\{x_k^{(0)}\}$, $\{x_{i,k}^{(1)}\}$ und $\{x_{i,k}^{(2)}\}$ mit $k \to \infty$ gegen x^* konvergente Folgen. Nach der Multiplikation der i-ten Ungleichungsrestrikion mit $u_i^* \geq 0$, der i-ten Gleichungsrestriktion mit v_i^*, Berücksichtigung der Gleichgewichtsbedingung $(u^*)^T g(x^*) = 0$ und anschließender Summation folgt

$$\begin{aligned} 0 \;\geq\; t_k \underbrace{\Big[\nabla f(x^*) + \sum_{i=1}^{l} u_i^* \nabla g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla h_i(x^*)\Big]^T}_{=0} p_k \\ + \tfrac{1}{2}\, t_k^2 p_k^T \Big[\nabla^2 f(x_k^{(0)}) + \sum_{i=1}^{l} u_i^* \nabla^2 g_i(x_{i,k}^{(1)}) + \sum_{i=1}^{m} v_i^* \nabla^2 h_i(x_{i,k}^{(2)})\Big] p_k. \end{aligned}$$

Folglich ist

$$0 \geq p_k^T \Big[\nabla^2 f(x_k^{(0)}) + \sum_{i=1}^{l} u_i^* \nabla^2 g_i(x_{i,k}^{(1)}) + \sum_{i=1}^{m} v_i^* \nabla^2 h_i(x_{i,k}^{(2)})\Big] p_k,$$

mit $k \to \infty$ hat man den gewünschten Widerspruch zur Voraussetzung erhalten. □ □

Beispiel: Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := x_2 + x_3 \quad \text{unter der Nebenbedingung} \\ \qquad\qquad h(x) := \begin{pmatrix} x_1 + x_2 + x_3 - 1 \\ x_1^2 + x_2^2 + x_3^2 - 1 \end{pmatrix}. \end{cases}$$

Wir wollen zunächst diejenigen zulässigen Punkte bestimmen, in denen die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind, anschließend mit Hilfe der hinreichenden Optimalitätsbedingungen zweiter Ordnung untersuchen, ob es sich hier wirklich um lokale Lösungen von (P) handelt.

Zunächst stellt sich also die Frage nach Lösungen $(x^*, v^*) \in \mathbb{R}^3 \times \mathbb{R}^2$ von

$$h(x) = 0, \qquad \nabla f(x) + h'(x)^T v = 0$$

bzw.

$$\begin{aligned} x_1 + x_2 + x_3 &= 1, \\ x_1^2 + x_2^2 + x_3^2 &= 1, \end{aligned} \qquad \begin{aligned} v_1 + 2x_1 v_2 &= 0, \\ 1 + v_1 + 2x_2 v_2 &= 0, \\ 1 + v_1 + 2x_3 v_2 &= 0. \end{aligned}$$

Aus den letzten Gleichungen erkennt man, dass notwendigerweise $v_2^* \neq 0$ und daher $x_2^* = x_3^*$. Zu lösen bleibt also das nichtlineare Gleichungssystem

$$\begin{aligned} x_1 + 2x_2 &= 1, \\ x_1^2 + 2x_2^2 &= 1, \end{aligned} \qquad \begin{aligned} v_1 + 2x_1 v_2 &= 0, \\ 1 + v_1 + 2x_2 v_2 &= 0. \end{aligned}$$

Aus den ersten beiden Gleichungen erhält man, dass das ursprüngliche nichtlineare Gleichungssystem genau zwei Lösungen (x^*, v^*) besitzt, nämlich

$$\hat{x}^* = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \hat{v}^* = \begin{pmatrix} -1 \\ \frac{1}{2} \end{pmatrix} \qquad \text{und} \qquad \tilde{x}^* = \begin{pmatrix} -\frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{pmatrix}, \quad \tilde{v}^* = \begin{pmatrix} -\frac{1}{3} \\ -\frac{1}{2} \end{pmatrix}.$$

Ferner ist

$$\nabla^2 f(\hat{x}^*) + \hat{v}_1^* \nabla^2 h_1(\hat{x}^*) + \hat{v}_2^* \nabla^2 h_2(\hat{x}^*) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

sogar auf dem $\mathbb{R}^3$ positiv definit, also $\hat{x}^*$ eine isolierte, lokale Lösung von (P). Andererseits ist

$$\nabla^2 f(\tilde{x}^*) + \tilde{v}_1^* \nabla^2 h_1(\tilde{x}^*) + \tilde{v}_2^* \nabla^2 h_2(\tilde{x}^*) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

auf

$$\operatorname{Kern} h'(\tilde{x}^*) = \{p = (p_1, p_2, p_3)^T \in \mathbb{R}^3 : p_1 = 0,\ p_2 = -p_3\}$$

noch nicht einmal positiv semidefinit. Daher ist $\tilde{x}^*$ keine lokale Lösung von (P). □

Beispiel: Das folgende Beispiel findet man bei R. Fletcher (1987, S. 228)[14] als Aufgabe. Sei $n > 2$, betrachte die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := -\sum_{j=1}^{n} x_j^3 \quad \text{unter der Nebenbedingung} \\ \qquad\qquad h(x) := \begin{pmatrix} e^T x \\ x^T x - n \end{pmatrix} = 0. \end{cases}$$

Wir wollen zunächst die Punkte finden, in denen die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind. Die Gleichung $\nabla f(x) + h'(x)^T v = 0$ führt auf

$$-3x_j^2 + v_1 + 2v_2 x_j = 0, \qquad j = 1, \ldots, n.$$

Aufsummieren liefert unter Berücksichtigung von $e^T x = 0$, $x^T x = n$, daß $v_1 = 3$. Mit noch unbekanntem v_2 ist daher x_j aus der quadratischen Gleichung

$$x_j^2 - \tfrac{2}{3} v_2 x_j - 1 = 0, \qquad j = 1, \ldots, n,$$

zu bestimmen. Hieraus folgt

$$x_j = \frac{v_2}{3} \pm \sqrt{1 + \left(\frac{v_2}{3}\right)^2}, \qquad j = 1, \ldots, n.$$

[14] FLETCHER, R. (1987) *Practical Methods of Optimization. Second Edition.* John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.

In einer möglichen Lösung haben die Komponenten also genau zwei verschiedene Werte (sie können nicht alle gleich sein, denn andernfalls wäre ihre Summe von Null verschieden). Auf die Reihenfolge der Komponenten kommt es nicht an, sei also etwa

$$(*) \qquad x_j = \begin{cases} \dfrac{v_2}{3} + \sqrt{1 + \left(\dfrac{v_2}{3}\right)^2}, & \text{für} \quad j = 1, \ldots, r, \\ \dfrac{v_2}{3} - \sqrt{1 + \left(\dfrac{v_2}{3}\right)^2}, & \text{für} \quad j = r+1, \ldots, n. \end{cases}$$

Dann ist bei gegebenem $r \in \{1, \ldots, n-1\}$ der Lagrange-Parameter v_2 aus

$$e^T x = n\frac{v_2}{3} + (2r - n)\sqrt{1 + \left(\frac{v_2}{3}\right)^2} = 0$$

zu bestimmen. Man erhält

$$\frac{v_2}{3} = \pm\frac{n - 2r}{2\sqrt{r(n-r)}}.$$

Bei gegebenem $r \in \{1, \ldots, n-1\}$ haben wir also zwei mögliche Lösungen, nämlich

$$x_j^{(1)} := \begin{cases} \sqrt{\dfrac{n-r}{r}}, & \text{für} \quad j = 1, \ldots, r, \\ -\sqrt{\dfrac{r}{n-r}}, & \text{für} \quad j = r+1, \ldots, n \end{cases}$$

und

$$x_j^{(2)} := \begin{cases} \sqrt{\dfrac{r}{n-r}}, & \text{für} \quad j = 1, \ldots, r, \\ -\sqrt{\dfrac{n-r}{r}}, & \text{für} \quad j = r+1, \ldots, n. \end{cases}$$

Bei festem $r \in \{1, \ldots, n-1\}$ berechnen wir die Funktionswerte

$$f(x^{(1)}) = -\frac{(n-r)^2 - r^2}{[r(n-r)]^{1/2}}, \qquad f(x^{(2)}) = \frac{(n-r)^4 - r^4}{[r(n-r)]^{3/2}}.$$

Für $2r \leq n$ ist offenbar $f(x^{(1)}) \leq 0 \leq f(x^{(2)})$, während $f(x^{(2)}) \leq 0 \leq f(x^{(1)})$ für $2r \geq n$. Für $2r \leq n$ erhält man den minimalen Funktionswert für $r = 1$, also

$$x_j^{(1)} = \begin{cases} \sqrt{n-1}, & \text{für} \quad j = 1, \\ -\dfrac{1}{\sqrt{n-1}}, & \text{für} \quad j = 2, \ldots, n, \end{cases}$$

der zugehörige Zielfunktionswert ist

$$f(x^{(1)}) = -\frac{(n-1)^2 - 1}{(n-1)^{1/2}}.$$

Ist dagegen $2r \geq n$, so erhält man den minimalen Funktionswert für $r = n - 1$, also

$$x_j^{(2)} = \begin{cases} \sqrt{n-1}, & \text{für} \quad j = 1, \ldots, n-1, \\ -\dfrac{1}{\sqrt{n-1}}, & \text{für} \quad j = n, \end{cases}$$

der zugehörige Zielfunktionswert ist

$$f(x^{(2)}) = -\frac{(n-1)^4 - 1}{(n-1)^{3/2}}.$$

Man prüft leicht nach, dass $f(x^{(2)}) < f(x^{(1)})$, so dass $x^{(2)}$ eine globale Lösung von (P) ist. Mit Hilfe der hinreichenden Optimalitätsbedingungen zweiter Ordnung können wir aber auch zeigen, dass durch $x^* := x^{(1)}$ mit zugehörigen Lagrange-Multiplikatoren

$$v_1^* := 3, \qquad v_2^* := \frac{3(n-2)}{2\sqrt{n-1}}$$

eine lokale Lösung von (P) gegeben ist. Entscheidend ist die Matrix

$$\nabla_{xx}^2 L(x^*, v^*) = \nabla^2 f(x^*) + v_1^* \nabla^2 h_1(x^*) + v_2^* \nabla^2 h_2(x^*) = \operatorname{diag}(-6x_j^* + 2v_2^*).$$

Es ist

$$-6x_j^* + 2v_2^* = \begin{cases} -\dfrac{3n}{\sqrt{n-1}}, & \text{für} \quad j = 1, \\ \dfrac{3n}{\sqrt{n-1}}, & \text{für} \quad j = 2, \ldots, n. \end{cases}$$

Also ist $\nabla_{xx}^2 L(x^*, v^*)$ eine Diagonalmatrix, die nur im ersten Diagonalelement einen negativen Eintrag besitzt, alle übrigen Diagonaleinträge sind positiv. Es ist zu zeigen, dass $\nabla_{xx}^2 L(x^*, v^*)$ auf $\{p \in \mathbb{R}^n : e^T p = 0,\ (x^*)^T p = 0\}$ positiv definit ist. Wegen

$$\{p \in \mathbb{R}^n : e^T p = 0,\ (x^*)^T p = 0\} = \Big\{p \in \mathbb{R}^n : p_1 = 0,\ \sum_{j=2}^n p_j = 0\Big\}$$

ist dies aber der Fall. □

2.3.4 Aufgaben

1. Man zeige, dass $x^* := (1, 1, 2)^T$ die Lösung von

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad -5x_2 + \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \quad \text{unter den Nebenbedingungen} \\ \begin{array}{rrrcr} -4x_1 & -3x_2 & & \geq & -8 \\ 2x_1 & +\ x_2 & & \geq & 2 \\ & -2x_2 & +x_3 & \geq & 0 \\ x_1 & -2x_2 & +x_3 & = & 1 \end{array} \end{cases}$$

   ist.

2. Für die Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := x_1^2 + 4x_2^2 + 16x_3^2 \quad \text{unter der Nebenbedingung} \\ \qquad\qquad\qquad\qquad h(x) := x_1 x_2 x_3 - 1 = 0 \end{cases}$$

bestimme man alle Punkte, in denen die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind und prüfe anschließend mit Optimalitätsbedingungen zweiter Ordnung, ob dies lokale Lösungen sind.

3. Gegeben sei die Optimierungsaufgabe[15]

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := -(x_1x_2 + x_2x_3 + x_1x_3) \quad \text{u. d. NB.} \\ \qquad\qquad\qquad h(x) := x_1 + x_2 + x_3 - 3 = 0. \end{cases}$$

Man bestimme den Punkt, in dem die notwendige Bedingung erster Ordnung erfüllt ist und prüfe anschließend mit einer hinreichenden Optimalitätsbedingung zweiter Ordnung, ob dies eine lokale Lösung ist.

4. Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := -\dfrac{1}{2}\sqrt{x_1} - \dfrac{1}{2}x_2 \quad \text{u. d. NB.} \\ g(x) := \begin{pmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} -0.1 \\ 0 \\ 1 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \end{cases}$$

Man bestimme den Punkt, in dem die notwendige Bedingung erster Ordnung erfüllt ist und prüfe anschließend mit einer hinreichenden Optimalitätsbedingung zweiter Ordnung, ob dies eine lokale Lösung ist.

5. Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{unter der Nebenbedingung} \quad x \geq 0.$$

Sei $x^* \geq 0$ eine lokale Lösung von (P) und die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in x^* stetig differenzierbar. Man stelle die notwendigen Optimalitätsbedingungen erster Ordnung auf.

6. Ganz ohne Constraint Qualification kann man immer noch den *Satz von F. John* beweisen:

Sei x^* eine lokale Lösung von

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

Hierbei seien die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ auf einer Umgebung von x^* stetig differenzierbar. Dann existiert ein von Null verschiedenes Tripel $(u_0^*, u^*, v^*) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m$ mit

$$(u_0^*, u^*) \geq (0, 0), \qquad u_0^* \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

Ist die Arrow-Hurwicz-Uzawa Constraint Qualification erfüllt, so ist hier notwendigerweise $u_0^* > 0$.

[15] Diese und die folgende Aufgabe findet man als Beispiel bei

W. Alt (2002) *Nichtlineare Optimierung. Eine Einführung in Theorie, Verfahren und Anwendungen.* Vieweg, Braunschweig-Wiesbaden.

7. Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

Hierbei seien $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ auf dem $\mathbb{R}^n$ konvex und stetig differenzierbar, $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear. Wie üblich sei die Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ zu (P) durch

$$L(x, u, v) := f(x) + u^T g(x) + v^T h(x)$$

definiert. Das zu (P) sogenannte *Wolfe-duale* Programm (siehe P. Wolfe (1961)[16]) ist dann durch

$$\text{(D)} \qquad \left\{ \begin{array}{c} \text{Maximiere} \quad L(z, u, v) \quad \text{auf} \\ N := \{(z, u, v) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \nabla_x L(z, u, v) = 0\} \end{array} \right.$$

gegeben. Man zeige:

(a) Ist $x \in M$ und $(z, u, v) \in N$, so ist $L(z, u, v) \leq f(x)$. Zwischen (P) und (D) gilt also ein schwacher Dualitätssatz.

(b) Die (schwache) Slatersche Constraint Qualification sei erfüllt, d. h. es existiere ein $\hat{x} \in M$ mit $g_i(\hat{x}) < 0$ für alle i, für die g_i nicht affin linear ist. Ist dann $x^* \in M$ eine Lösung von (P), so existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ derart, dass $(x^*, u^*, v^*) \in N$ und $f(x^*) = L(x^*, u^*, v^*)$. Ferner ist (u^*, v^*) eine Lösung des zu (P) Lagrange-dualen Programms.

8. Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : h(x) = 0\}.$$

Sei $x^* \in M$ eine lokale Lösung von (P) und $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sowie $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ auf einer Umgebung von x^* zweimal stetig differenzierbar. In x^* seien die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt, d. h. es existiere ein $v^* \in \mathbb{R}^m$ mit $\nabla f(x^*) + h'(x^*)^T v^* = 0$ und der Eigenschaft, dass

$$W^* := \nabla^2 f(x^*) + \sum_{i=1}^{m} v_i^* \nabla^2 h_i(x^*)$$

auf Kern $(h'(x^*))$ positiv definit ist. Schließlich sei Rang $(h'(x^*)) = m$. Man zeige, dass es ein $\sigma_0 > 0$ gibt derart, dass x^* für jedes $\sigma > \sigma_0$ eine isolierte, lokale Lösung der unrestringierten Optimierungsaufgabe

$$(\text{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + (v^*)^T h(x) + \tfrac{1}{2}\sigma \, \|h(x)\|^2, \quad x \in \mathbb{R}^n,$$

ist. Hierbei sei $\|\cdot\|$ die euklidische Norm.

Hinweis: Man zeige, dass $\nabla \Phi_\sigma(x^*) = 0$ für alle $\sigma > 0$ und $\nabla^2 \Phi_\sigma(x^*)$ für alle hinreichend großen $\sigma > 0$ positiv definit ist.

[16] Wolfe (1961) "A duality theorem for nonlinear programming." Quarterly of Applied Mathematics 19, 239–244.

9. Sei (x^*, v^*) ein Kuhn-Tucker-Paar zu der Optimierungsaufgabe

   (P) $$\text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : h(x) = 0\},$$

   also (x^*, v^*) eine Nullstelle der durch

   $$T(x, v) := \begin{pmatrix} \nabla f(x) + h'(x)^T v \\ h(x) \end{pmatrix}$$

   definierten Abbildung $T\colon \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n \times \mathbb{R}^m$. Man berechne die Funktionalmatrix von T in (x^*, v^*) und untersuche, unter welchen Voraussetzungen diese nichtsingulär ist. Hierbei sind natürlich $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ als zweimal stetig differenzierbar auf einer Umgebung von x^* vorausgesetzt.

10. Als Hoffman-Theorem (siehe A. J. Hoffman (1952)[17]) wollen wir die folgende Aussage verstehen (auch wenn sie nicht ganz mit der Originalversion übereinstimmt). Hierbei benutzen wir die folgende Bezeichnung: Für einen Vektor $y \in \mathbb{R}^l$ sei y_+ die Projektion von y auf den nichtnegativen Orthanten, also $(y_+)_i = \max(y_i, 0)$.

    Sei

    $$P := \{x \in \mathbb{R}^n : Ax \leq b,\ Cx = d\} \neq \emptyset.$$

    Hierbei seien $A \in \mathbb{R}^{l \times n}$, $b \in \mathbb{R}^l$, $C \in \mathbb{R}^{m \times n}$, $d \in \mathbb{R}^m$. Dann existiert eine Konstante $c_0 = c_0(A, C) > 0$ derart, daß

    $$\operatorname{dist}(z, P) := \inf_{x \in P} \|z - x\| \leq c_0 \left\| \begin{pmatrix} (Az - b)_+ \\ Cz - d \end{pmatrix} \right\| \qquad \text{für alle } z \in \mathbb{R}^n.$$

    Hierbei sei $\|\cdot\|$ jeweils die euklidische Norm auf dem entsprechenden Raum.

11. Mit Hilfe des Hoffman-Theorems zeige man: Ist $A \in \mathbb{R}^{m \times n}$, so existiert eine Konstante $c_0 = c_0(A) > 0$ derart, dass es zu jedem $b \in \operatorname{Bild}(A)$ ein $x^* \in \mathbb{R}^n$ mit $Ax^* = b$ und $\|x^*\| \leq c_0 \|b\|$ gibt.

12. Mit Hilfe des Hoffman-Theorems zeige man: Gegeben sei das lineare Programm

    (P) $$\text{Minimiere} \quad f(x) := c^T x, \quad x \in M.$$

    Hierbei sei $c \in \mathbb{R}^n$, $M \subset \mathbb{R}^n$ ein nichtleerer Polyeder und $\inf(\mathrm{P}) > -\infty$, daher die Menge M_{opt} der Lösungen von (P) nichtleer. Dann existiert eine Konstante $c_0 > 0$ derart, dass

    $$\operatorname{dist}(x, M_{\text{opt}}) \leq c_0[f(x) - \min(\mathrm{P})] \qquad \text{für alle } x \in M.$$

    Hinweis: Man beachte, dass $M_{\text{opt}} = M \cap \{x^* \in \mathbb{R}^n : c^T x^* - \min(\mathrm{P}) = 0\}$

13. Gegeben sei das quadratische Programm

    (P) $$\text{Minimiere} \quad f(x) := c^T x + \tfrac{1}{2} x^T Q x, \quad x \in M,$$

[17] Hoffman, A. J., "On approximate solutions of systems of linear inequalities." J. Res. Natl. Bur. Standards, 49 (1952), pp. 263–265.

wobei $c \in \mathbb{R}^n$, $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit, $M \subset \mathbb{R}^n$ ein nichtleerer Polyeder und $\inf(\mathrm{P}) > -\infty$. Die dann nichtleere Menge der Lösungen von (P) werde mit M_{opt} bezeichnet. Man zeige die Existenz einer Konstanten $c > 0$ mit

$$\operatorname{dist}(x, M_{\mathrm{opt}}) \le c\left[f(x) - \min(\mathrm{P}) + \sqrt{f(x) - \min(\mathrm{P})}\right] \qquad \text{für alle } x \in M.$$

Hinweis: Der Polyeder M habe die Darstellung $M = \{x \in \mathbb{R}^n : Ax \le b\}$, wobei $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$. Eine Lösung $x_0^* \in M$ von (P) ist charakterisiert durch die Existenz eines Vektors $u_0^* \in \mathbb{R}^m$ mit

$$u_0^* \ge 0, \qquad c + Qx_0^* + A^T u_0^* = 0, \qquad (u_0^*)^T(b - Ax_0^*) = 0.$$

Man zeige, dass die Menge M_{opt} der Lösungen von (P) sich darstellen lässt als

$$M_{\mathrm{opt}} = \{x^* \in \mathbb{R}^n : (b - Ax^*)^T u_0^* = 0,\ Qx^* = Qx_0^*,\ Ax^* \le b\}$$

und wende das Hoffman-Theorem an. (Ähnliche Ergebnisse findet man bei W. Li (1995)[18].).

14. Es sollen 400 m^3 Kies von einem Ort zu einem anderen transportiert werden. Dies geschehe in einer (nach oben!) offenen Box der Länge x_1, der Breite x_2 und der Höhe x_3 (jeweils in Metern gemessen). Der Boden und die beiden Längsseiten müssen aus einem Material hergestellt werden, das zwar nichts kostet, von dem aber nur 4 m^2 zur Verfügung steht. Das Material für die beiden Querseiten kostet 200 Euro pro m^2. Ein Transport der Box kostet 1 Euro. Wie hat man die Box zu konstruieren?

    Man stelle also die zugehörige Optimierungsaufgabe auf, bestimme die zulässigen Punkte, in denen die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind und überprüfe diese mit Hilfe der hinreichenden Optimalitätsbedingungen zweiter Ordnung auf Optimalität.

15. Man bestimme die Lösung von

$$(\mathrm{P}) \qquad \text{Maximiere} \quad f(x) := \prod_{j=1}^{n} x_j \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \ge 0,\ e^T x = 1\},$$

    wobei e einmal wieder den Vektor im $\mathbb{R}^n$ bezeichnet, dessen Komponenten sämtlich gleich 1 sind. Hiermit beweise man die Ungleichung vom geometrisch-arithmetischen Mittel, dass also für alle $x \in \mathbb{R}^n$ mit $x \ge 0$ gilt

$$\left(\prod_{j=1}^{n} x_j\right)^{1/n} \le \frac{1}{n}\sum_{j=1}^{n} x_j.$$

    Hierbei tritt Gleichheit genau dann ein, wenn $x = \alpha e$ mit $\alpha \ge 0$.

16. Bei gegebenem $\alpha \in (0,1)$ und $r := \sqrt{n/(n-1)}$ betrachte man die Optimierungsaufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) := \prod_{j=1}^{n} x_j \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : e^T x = n,\ \|x - e\|_2 \le \alpha r\}.$$

    Hierbei sei e wieder der Vektor des $\mathbb{R}^n$, dessen Komponenten alle gleich 1 sind. Man zeige:

[18] Li, W. (1995) "Error bounds for piecewise convex quadratic programs and applications." SIAM J. Control and Optimization 33, 1510– 1529.

(a) (P) besitzt eine Lösung x^* und es ist notwendig $x^* > 0$ und $\|x^* - e\|_2 = \alpha r$.

(b) Eine Lösung x^* von (P) besitzt genau zwei verschiedene Komponenten. Bis auf die Reihenfolge der Komponenten kommt als Lösungskandidat also nur ein Vektor $x^{(m)}$ in Frage, dessen erste m Komponenten übereinstimmen und kleiner sind als die restlichen (ebenfalls gleichen) $(n-m)$ Komponenten. Man zeige, dass $x^* = x^{(1)}$ bis auf die Reihenfolge der Komponenten die Lösung von (P) ist.

(c) Es ist

$$\prod_{j=1}^{n} x_j \geq (1-\alpha)\left(1 + \frac{\alpha}{n-1}\right)^{n-1} \qquad \text{für alle } x \in M.$$

Hinweis: Diese Aufgabe spielt im Zusammenhang mit der Konvergenzanalyse des Karmarkar-Verfahrens eine Rolle, siehe z. B. J. Werner (1992, S. 135 ff.).

Kapitel 3

Quadratische Optimierungsaufgaben

In diesem Kapitel betrachten wir die numerische Behandlung quadratischer Optimierungsaufgaben. Unter einer quadratischen Optimierungsaufgabe (bzw. quadratischem Programm) versteht man das Problem, eine quadratische Zielfunktion auf einem Polyeder im $\mathbb{R}^n$ zu minimieren. I. Allg. werden wir daher in diesem Kapitel die Aufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{l} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2}\, x^T Q x \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\} \end{array} \right.$$

betrachten. Hierbei seien $a_1, \ldots, a_m \in \mathbb{R}^n$, $b_1, \ldots, b_m \in \mathbb{R}$, $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n \times n}$ symmetrisch (und i. Allg. auch positiv (semi)definit). Zur Abkürzung setzen wir

$$A := \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix} \in \mathbb{R}^{m \times n}, \qquad b := \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m.$$

Unser Ziel wird es sein, numerische Verfahren zur Lösung quadratischer Programme anzugeben und zu analysieren. Im ersten Abschnitt gehen wir auf das primale Verfahren von Fletcher, in dem darauf folgenden Abschnitt auf das duale Verfahren von Goldfarb-Idnani ein. Schließlich werden im dritten Abschnitt Ansätze zur Behandlung von quadratischen Programmen mit sogenannten Box Constraints beschrieben.

3.1 Primale Verfahren

Wir beginnen mit primalen Verfahren bei quadratischen Programmen. Hier wird eine Folge zulässiger Lösungen mit monoton wachsenden oder zumindestens monoton nicht fallenden Zielfunktionswerten berechnet und abgebrochen, wenn eine notwendige (eventuell auch hinreichende) Optimalitätsbedingung erfüllt ist. Insbesondere muss beim Start eine zulässige Ausgangslösung bereitgestellt werden, welche notfalls (ähnlich wie beim Simplexverfahren) in einer Phase I berechnet werden muß. Natürlich ist der Fall, dass Q positiv definit ist, besonders angenehm. Denn da die Zulässigkeit von

(P) nach Angabe einer zulässigen Ausgangslösung gesichert ist, existiert in diesem Fall eine eindeutige Lösung von (P). Einige Bemerkungen sollen aber auch für den Fall gemacht werden, dass Q nur positiv semidefinit (dann ist (P) wenigstens noch ein konvexes Programm, lokale und globale Lösungen stimmen überein und die notwendigen Bedingungen erster Ordnung sind hinreichend für Optimalität) ist.

3.1.1 Das Verfahren von Fletcher

Gegeben sei obiges quadratisches Programm (P) mit der Menge M zulässiger Lösungen. In diesem Unterabschnitt werden wir ein Verfahren von R. Fletcher (1971)[1] schildern, welches zu den sogenannten *Methoden aktiver Mengen* gehört, die auch bei linear restringierten nichtlinearen Programmen eine wichtige Rolle spielen. Ein ähnliches Verfahren ist von D. Goldfarb (1972)[2] angegeben worden. Von P. E. Gill, W. Murray (1978)[3] stammen stabile Realisierungen (stabiles, effizientes Updaten der benötigten Matrizen) dieser Methoden, auch ihre Ausführungen werden in diesen Unterabschnitt einfließen. Hingewiesen sei schließlich noch auf das Kapitel über quadratische Programme bei R. Fletcher (1987)[4].

Grundlegend ist die Definition der Indexmenge *aktiver Restriktionen*. Diese ist für ein gegebenes $x \in M$, etwa einer aktuellen Näherung in einem primalen Verfahren, durch

$$I(x) := \{i \in \{1, \ldots, m\} : a_i^T x = b_i\}$$

definiert[5]. Daher enthält $I(x)$ insbesondere die Indizes aller Gleichungsrestriktionen, also $\{m_0 + 1, \ldots, m\}$. Der Einfachheit halber werden wir voraussetzen, dass die Vektoren $\{a_i\}_{i=m_0+1,\ldots,m} \subset \mathbb{R}^n$ linear unabhängig sind, was nach Streichen entsprechender Gleichungen natürlich o. B. d. A. angenommen werden kann.

Für eine Indexmenge $I \subset \{1, \ldots, m\}$ sei die Matrix $A_I \in \mathbb{R}^{q \times n}$ (es sei $q := \#(I)$ die Anzahl der Elemente von I) als die Matrix definiert, die gerade a_i^T für $i \in I$ als Zeilen (mit einer durch I festgelegten Reihenfolge) besitzt. Entsprechend werden wir die Bezeichnungen b_I und y_I usw. benutzen.

Nun können wir schon das Verfahren von Fletcher angeben, wobei wir zunächst voraussetzen werden, dass $Q \in \mathbb{R}^{n \times n}$ sogar positiv definit ist. Die Idee zu dem Verfahren ist einfach. Ist nämlich $x \in M$ eine aktuelle, zulässige Näherung und $I \subset \{1, \ldots, m\}$ eine Indexmenge, die alle Gleichungsrestriktionen enthält, mit $A_I x = b_I$ (d. h. es ist $I \subset I(x)$) und Rang $(A_I) = q$ mit $q := \#(I)$, so bestimme man x^+ als eindeutige

[1] Fletcher, R. (1971) " A general quadratic programming algorithm." Journal of the Institute of Mathematics and its Applications 7, 76–91.

[2] Goldfarb, D. (1972) "Extensions of Newton's method and simplex methods for solving quadratic programs." In: *Numerical Methods for Nonlinear Optimization.* (ed.: F. Lootsma), Academic Press, New York.

[3] Gill, P. E. and W. Murray (1978) "Numerically stable methods for quadratic programming." Mathematical Programming 14, 349–372.

[4] Fletcher, R. (1987) *Practical Methods of Optimization. Second Edition.* John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.

[5] Man beachte, dass wir hier die Bezeichnungen gegenüber dem letzten Kapitel geändert haben. Bisher wurde mit $I(x)$ die Menge der in x aktiven *Ungleichungs*restriktionen bezeichnet.

Lösung des durch lineare Gleichungen restringierten quadratischen Programms

$$\text{Minimiere} \quad f(z) := c^T z + \tfrac{1}{2}\, z^T Q z \quad \text{unter der Nebenbedingung} \quad A_I z = b_I.$$

Mit der Variablentransformation $z = x + p$ hat man also die Lösung p des Programms

$$\text{Minimiere} \quad (c + Qx)^T p + \tfrac{1}{2}\, p^T Q p \quad \text{unter der Nebenbedingung} \quad A_I p = 0$$

zu berechnen. Ist $x + p$ zulässig, so wird $x^+ := x + p$ als neue Näherung akzeptiert, wobei man sogar schon bei der Lösung von (P) angelangt ist, wenn alle zu I und Ungleichungen gehörenden Lagrange-Multiplikatoren nichtnegativ sind. Ist zwar $x + p$ zulässig, aber einer der zu I und Ungleichungen gehörenden Lagrange-Multiplikatoren negativ, so entferne man dessen Index aus I. Ist dagegen $x + p$ nicht zulässig, so ist man von x ausgehend zu weit in Richtung p gegangen. In diesem Falle gewinnt man $x^+ \in M$ dadurch, dass man von x aus so weit wie möglich in Richtung p geht, ohne die Zulässigkeit zu verletzen. Hierbei wird eine neue Restriktion aktiv, die in die Indexmenge I aufgenommen wird. Nur etwas genauer lautet das Verfahren von Fletcher folgendermaßen:

(0) Gegeben sei ein $x \in M$ und eine Indexmenge I mit $\{m_0 + 1, \ldots, m\} \subset I \subset I(x)$ und Rang $(A_I) = q$, wobei $q := \#(I)$.

(1) Berechne $p \in \mathbb{R}^n$ und $y_I = (y_i)_{i \in I}$ mit

$$c + Qx + Qp = A_I^T y_I, \qquad A_I p = 0,$$

d. h. bestimme die Lösung p und den zugehörigen Lagrange-Vektor y_I zu dem durch lineare Gleichungen restringierten quadratischen Programm

$$\text{Minimiere} \quad (c + Qx)^T p + \tfrac{1}{2}\, p^T Q p \quad \text{unter der Nebenbedingung} \quad A_I p = 0.$$

(2) Falls $x + p \in M$, dann:

Setze $x^+ := x + p$.

Bestimme $l \in I \cap \{1, \ldots, m_0\}$ mit $y_l = \min_{i \in I \cap \{1,\ldots,m_0\}} y_i$.

Falls $y_l \geq 0$, dann:

STOP, $x^* := x^+$ ist die Lösung von (P).

Andernfalls:

Setze $I^+ := I \setminus \{l\}$.

Andernfalls:

Berechne

$$s(x,p) := \min\Big\{ \frac{b_i - a_i^T x}{a_i^T p} : i \notin I,\ a_i^T p < 0 \Big\} = \frac{b_r - a_r^T x}{a_r^T p}.$$

Setze $x^+ := x + s(x,p)p$ und $I^+ := I \cup \{r\}$.

(3) Setze $(x, I) := (x^+, I^+)$, gehe nach (1).

Bemerkungen: Wir wollen uns davon überzeugen, dass das Verfahren von Fletcher *durchführbar* ist. Sei also (x, I) ein Paar, welches den in Schritt (0) angegebenen Bedingungen genügt. Das lineare Gleichungssystem

$$c + Qx + Qp = A_I^T y_I, \qquad A_I p = 0$$

bzw.

$$\begin{pmatrix} Q & A_I^T \\ A_I & 0 \end{pmatrix} \begin{pmatrix} p \\ -y_I \end{pmatrix} = - \begin{pmatrix} c + Qx \\ 0 \end{pmatrix}$$

ist eindeutig lösbar, da die Koeffizientenmatrix wegen Rang $(A_I) = q$ und der positiven Definitheit von Q nichtsingulär ist.

Ist nun $x + p \in M$ und $y_i \geq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$, so sind in $x^* := x + p$ die hinreichenden Optimalitätsbedingungen erfüllt, wenn man $y^* \in \mathbb{R}^m$ durch $y_i^* := y_i$ für $i \in I$ und $y_i^* := 0$ für $i \notin I$ definiert. Ist zwar $x + p \in M$, aber einer der zu einer Ungleichungsrestriktion $l \in I$ gehörenden Lagrange-Multiplikatoren negativ, so wird aus I der Index l entfernt, so dass auch $(x^+, I^+) := (x + p, I \setminus \{l\})$ den Bedingungen in Schritt (0) genügt.

Sei daher jetzt $x+p \notin M$. Wegen $\{m_0+1, \ldots, m\} \subset I \subset I(x)$ und $A_I p = 0$ existiert ein Index $i \in \{1, \ldots, m_0\}$ mit $i \notin I$, für den $a_i^T(x + p) < b_i$ bzw. $a_i^T p < b_i - a_i^T x \leq 0$. Daher ist $s(x, p) \in [0, 1)$ definiert und offenbar $x + tp \in M$ für alle alle $t \in [0, s(x, p)]$. Insbesondere ist $x^+ := x + s(x, p)p \in M$ zulässig, ferner wird die (oder genauer: eine) Ungleichungsrestriktion r, in welcher das Minimum bei der Berechnung von $s(x, p)$ angenommen wird, in x^+ aktiv. Für das neue Paar (x^+, I^+) ist daher offensichtlich $x^+ \in M$ und $\{m_0+1, \ldots, m\} \subset I^+ \subset I(x^+)$. Zu zeigen bleibt, dass Rang $(A_{I^+}) = q+1$ bzw. a_r von $\{a_i\}_{i \in I}$ linear unabhängig ist. Wegen $a_i^T p = 0$ für alle $i \in I$ und $a_r^T p < 0$ ist das aber trivialerweise der Fall, so dass auch hier (x^+, I^+) den Eingangsbedingungen (0) genügt. Insgesamt ist die Durchführbarkeit des Verfahrens von Fletcher bewiesen.

Eine triviale Bemerkung besteht noch darin, dass y_I nicht berechnet zu werden braucht, wenn $x + p \notin M$. Man berechnet also zunächst nur die erste „Komponente“ p der Lösung des Gleichungssystems in Schritt (1) (wie dies geschehen kann wird im Anschluss an diese Bemerkungen erläutert), testet anschließend, ob $x + p \in M$ (hierzu brauchen nur die Restriktionen mit einem Index $i \notin I$ überprüft zu werden), und berechnet y_I nur dann, wenn dieser Test erfolgreich verläuft.

Nun interessieren natürlich die Konvergenzeigenschaften dieses Verfahrens und hier insbesondere die Frage, ob das Fletcher-Verfahren nach endlich vielen Schritten abbricht. Hierzu überlegen wir uns, dass $f(x^+) \leq f(x)$ gilt. Denn wegen

$$\begin{aligned} f(x + tp) &= f(x) + t(c + Qx)^T p + \tfrac{1}{2} t^2 p^T Q p \\ &= f(x) + t(A_I^T y_I - Qp)^T p + \tfrac{1}{2} t^2 p^T Q p \\ &= f(x) + (\tfrac{1}{2} t^2 - t)\, p^T Q p \end{aligned}$$

ist $f(x + tp) \leq f(x)$ für alle $t \in [0, 2)$ und $f(x + tp)$ minimal für $t = 1$. Daher ist es vernünftig, die neue Näherung x^+ in der angegebenen Weise zu definieren, da

die Zielfunktion auf $\{x + tp : t \geq 0\} \cap M$ gerade in x^+ minimal wird. Jedenfalls ist $f(x^+) < f(x)$ außer in dem entarteten Fall $x^+ = x$.

Angenommen, das Fletcher-Verfahren breche nicht vorzeitig mit einer Lösung ab und erzeuge eine Folge $\{(x_k, I_k)\}_{k\in\mathbb{N}}$. Es gibt eine unendliche Indexmenge $K \subset \mathbb{N}$ mit $x_k + p_k \in M$ für alle $k \in K$. Denn ist $x_k + p_k \notin M$, so wird die Indexmenge I_k vergrößert und das kann nur endlich oft hintereinander geschehen. Da es ferner nur endlich viele Indexmengen $I \subset \{1, \ldots, m\}$ gibt, existiert unter den Indexmengen $\{I_k\}_{k\in K}$ eine, welche unendlich oft auftritt. O. B. d. A. ist daher $I_k = I$ für alle $k \in K$. Für diese k ist $x_k + p_k$, da nur von $I_k = I$ abhängig, selbst von k unabhängig. Da ferner $f(x_{k+1}) < f(x_k)$ für $x_{k+1} \neq x_k$, gibt es ein $k_0 \in \mathbb{N}$ mit $x_k = x_{k_0}$ für alle $k \geq k_0$. Solche Zyklen sind wie beim Simplexverfahren möglich, gelten aber als „extrem unwahrscheinlich“ und bei der praktischen Realisierung des Verfahrens wird i. Allg. davon ausgegangen, dass sie nicht auftreten. □

Beispiel: Wir betrachten die Aufgabe

$$\text{(P)} \quad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := \begin{pmatrix} -2 \\ -6 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \text{unter der Nebenbedingung} \\ \begin{pmatrix} -1 & -1 \\ 1 & -2 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq \begin{pmatrix} -2 \\ -2 \\ -3 \end{pmatrix}. \end{array} \right.$$

Wir starten mit

$$x^0 := \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

die Kosten sind $f(x^0) = 0$. Dann ist keine der Ungleichungsrestriktionen aktiv, also $I^0 := \emptyset$. Es ist p^0 zu berechnen aus

$$\begin{pmatrix} -2 \\ -6 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} p_1^0 \\ p_2^0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

was auf

$$p^0 = \begin{pmatrix} 10 \\ 8 \end{pmatrix}$$

führt. Da $x^0 + p^0$ nicht zulässig ist, ist die maximale Schrittweite $s(x^0, p^0)$ zu berechnen. Es ist

$$s(x^0, p^0) = \min(\tfrac{2}{18}, \tfrac{2}{6}, \tfrac{3}{28}), \qquad x^1 = \tfrac{1}{14} \begin{pmatrix} 15 \\ 12 \end{pmatrix}, \qquad I^1 = I^0 \cup \{3\} = \{3\}.$$

Die zugehörigen Kosten sind

$$f(x^1) = -6.89540816326531.$$

Im nächsten Schritt sind p^1 und $y_{I^1} = y_3$ zu berechnen aus dem linearen Gleichungssystem

$$\begin{pmatrix} 1 & -1 & 2 \\ -1 & 2 & 1 \\ -2 & -1 & 0 \end{pmatrix} \begin{pmatrix} p_1^1 \\ p_2^1 \\ y_3 \end{pmatrix} = \tfrac{1}{14} \begin{pmatrix} 25 \\ 75 \\ 0 \end{pmatrix}.$$

Wir erhalten

$$\begin{pmatrix} p_1^1 \\ p_2^1 \\ y_3 \end{pmatrix} = \begin{pmatrix} -2.40384615384615 \\ 4.80769230769231 \\ 6.73076923076923 \end{pmatrix}.$$

Nun ist $x^1 + p^1$ nicht zulässig, daher ist wieder die maximale Schrittweite zu berechnen. Man erhält

$$s(x^1, p^1) = 0.02971428571429, \qquad x^2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad I^2 = \{3, 1\}.$$

Die zugehörigen Kosten sind

$$f(x^2) = -7.5.$$

Im nächsten Schritt sind p^2 und

$$y_{I^2} = \begin{pmatrix} y_3 \\ y_1 \end{pmatrix}$$

zu bestimmen aus dem linearen Gleichungssystem

$$\begin{pmatrix} 1 & -1 & 2 & 1 \\ -1 & 2 & 1 & 1 \\ -2 & -1 & 0 & 0 \\ -1 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} p_1^2 \\ p_2^2 \\ y_3 \\ y_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 0 \\ 0 \end{pmatrix}.$$

Als Lösung erhält man

$$p^2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad y_{I^2} = \begin{pmatrix} -3 \\ 8 \end{pmatrix}.$$

Wir setzen daher $x^3 := x^2$ und $I^3 := I^2 \setminus \{3\} = \{1\}$. Jetzt sind p^3 und $y_{I^3} = (y_1)$ aus dem linearen Gleichungssystem

$$\begin{pmatrix} 1 & -1 & 1 \\ -1 & 2 & 1 \\ -1 & -1 & 0 \end{pmatrix} \begin{pmatrix} p_1^3 \\ p_2^3 \\ y_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 0 \end{pmatrix}$$

zu berechnen. Dies ergibt

$$p^3 = \begin{pmatrix} -0.6 \\ 0.6 \end{pmatrix}, \qquad y_{I^3} = (3.2).$$

Da $x^3 + p^3$ die zweite Restriktion verletzt, ist die maximale Schrittweite $s(x^3, p^3)$ zu berechnen. Man erhält

$$s(x^3, p^3) = 0.55555555555556,$$

damit

$$x^4 = \begin{pmatrix} 0.66666666666667 \\ 1.33333333333333 \end{pmatrix}, \qquad I^4 = I^3 \cup \{2\} = \{1, 2\}.$$

Jetzt müssen p^4 und $y_{I^4} = (y_1, y_2)^T$ aus

$$\begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 2 & 1 & 2 \\ -1 & -1 & 0 & 0 \\ 1 & -2 & 0 & 0 \end{pmatrix} \begin{pmatrix} p_1^4 \\ p_2^4 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \frac{8}{3} \\ 4 \\ 0 \\ 0 \end{pmatrix}$$

berechnet werden, was auf

$$p^4 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad y_{I^4} = \begin{pmatrix} 3.11111111111111 \\ 0.44444444444444 \end{pmatrix}$$

führt. Da $x^5 := x^4 + p^4 = x^4$ zulässig ist und die zugehörigen Multiplikatoren sogar positiv sind, ist die Lösung mit zugehörigen Multiplikatoren durch

$$x^* := \begin{pmatrix} 0.66666666666667 \\ 1.33333333333333 \end{pmatrix}, \qquad y^* := \begin{pmatrix} 3.11111111111111 \\ 0.44444444444444 \\ 0.00000000000000 \end{pmatrix}$$

gegeben. Die zugehörigen Kosten sind

$$f(x^*) = -8.22222222222222.$$

□

Die Hauptarbeit im Verfahren von Fletcher besteht in der Berechnung der Lösung $(p, y_I) \in \mathbb{R}^n \times \mathbb{R}^q$ des linearen Gleichungssystems

$$(*) \qquad \begin{pmatrix} Q & A_I^T \\ A_I & 0 \end{pmatrix} \begin{pmatrix} p \\ -y_I \end{pmatrix} = - \begin{pmatrix} c + Qx \\ 0 \end{pmatrix}.$$

Nach wie vor gehen wir davon aus, dass $Q \in \mathbb{R}^{n \times n}$ positiv definit und Rang $(A_I) = q$ ist. Dann ist auch $A_I Q^{-1} A_I^T \in \mathbb{R}^{q \times q}$ positiv definit und es ist einfach, die Lösung von $(*)$ geschlossen anzugeben. Denn aus $Qp - A_I^T y_I = -(c + Qx)$ erhält man nach Multiplikation von links mit $A_I Q^{-1}$ unter Berücksichtigung von $A_I p = 0$, daß

$$y_I = (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} (c + Qx).$$

Aus $p = Q^{-1} A_I^T y_I - Q^{-1}(c + Qx)$ folgt damit

$$p = -Q^{-1}(I - A_I^T (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1})(c + Qx).$$

Definiert man also (genau wie in dem später zu beschreibenden Verfahren von Goldfarb-Idnani) die Matrizen $N_I \in \mathbb{R}^{q \times n}$ und $H_I \in \mathbb{R}^{n \times n}$ durch

$$N_I := (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1}, \qquad H_I := Q^{-1}(I - A_I^T N_I),$$

so ist die Lösung (p, y_I) von $(*)$ gegeben durch (siehe auch Aufgabe 1)

$$p = -H_I(c + Qx), \qquad y_I = N_I(c + Qx).$$

Natürlich könnte man in jedem Iterationsschritt N_I und H_I mit Hilfe der angegebenen Formeln neu berechnen, was aber jeweils einen Aufwand von $O(n^3)$ flops bedeuten würde. Schon besser ist es, so vorzugehen, wie R. Fletcher (1971) und D. Goldfarb (1972) vorgeschlagen haben, nämlich zu berücksichtigen, dass sich die Indexmenge I von Schritt zu Schritt nur um ein Element verändert. Die entsprechenden Update-Formeln zur Berechnung von N_{I^+} und H_{I^+} sind in der Aufgabe 3 (hier ist $I^+ := I \cup \{r\}$) bzw. der Aufgabe 4 (hier ist $I^+ := I \setminus \{l\}$) angegeben worden. Besser ist es, geeignete Zerlegungen von N_I und H_I upzudaten. Eine Möglichkeit (jedenfalls für den Fall, dass Q positiv definit ist) besteht darin, folgendermaßen vorzugehen. Sei $I \subset \{1, \ldots, m\}$ (wieder sei $q := \#(I)$ die Anzahl der Elemente von I) eine Indexmenge mit der Eigenschaft, dass Rang $(A_I) = q$. Gegeben sei $Z_I \in \mathbb{R}^{n\times n}$ derart, daß

$$Z_I Z_I^T = Q^{-1}, \qquad Z_I^T A_I^T = \begin{pmatrix} R_I \\ 0 \end{pmatrix} \begin{matrix} \} \, q \\ \} \, n-q \end{matrix}$$

mit einer oberen Dreiecksmatrix $R_I \in \mathbb{R}^{q\times q}$ ist. Zerlegt man Z_I in

$$Z_I = (\underbrace{Z_I^{(1)}}_{q} \quad \underbrace{Z_I^{(2)}}_{n-q}),$$

so ist

$$N_I = R_I^{-1} Z_I^{(1)T}, \qquad H_I = Z_I^{(2)} Z_I^{(2)T}.$$

In nächsten Abschnitt werden wir ausführlich schildern, wie man Z_I und R_I bei Hinzunahme (oder Wegfall eines Index) mit Hilfe von Givens-Rotationen updaten kann. Man berechnet die Lösung (p, y_I) des linearen Gleichungssystems $(*)$, indem man zunächst

$$Z_I^T(c + Qx) = \begin{pmatrix} Z_I^{(1)T}(c + Qx) \\ Z_I^{(2)T}(c + Qx) \end{pmatrix} = \begin{pmatrix} d_I^{(1)} \\ d_I^{(2)} \end{pmatrix}$$

bestimmt und anschließend $p := -Z_I^{(2)} d_I^{(2)}$ und $y_I := R_I^{-1} d_I^{(1)}$ durch Rückwärtseinsetzen. Ein kleiner Unterschied zum Verfahren von Goldfarb-Idnani besteht im Start. Während bei Goldfarb-Idnani am Anfang $I := \emptyset$ gesetzt wird, ist im Verfahren von Fletcher beim Start i. Allg. $I \neq \emptyset$. Z. B. müssen die Indizes zu Gleichungsrestriktionen in I enthalten sein. Kommen in (P) aber keine Gleichungsrestriktionen vor, so kann natürlich auch beim Fletcher-Verfahren zum Start $I := \emptyset$ gesetzt werden. Ist aber am Anfang $I \neq \emptyset$ (und Rang $(A_I) = q$), so berechnet man zunächst eine obere Dreiecksmatrix Z mit $ZZ^T = Q^{-1}$, setzt $Z_\emptyset := Z$ und erhält hieraus (und der „leeren“ Matrix $R_\emptyset$) durch sukzessive Hinzunahme der Indizes aus I die Matrizen Z_I und R_I.

Nun setzen wir nicht mehr voraus, dass $Q \in \mathbb{R}^{n\times n}$ positiv definit ist. Natürlich ist dann weder die Existenz (es ist $\inf(\mathrm{P}) = -\infty$ möglich) noch die Eindeutigkeit einer Lösung von (P) gesichert. Ferner wird man sich für indefinites Q damit begnügen müssen, in einer stationären Lösung von (P) zu enden, also einer zulässigen Lösung, in der die notwendigen Bedingungen erster Ordnung erfüllt sind.

Ist (x, I) ein Paar, das den Bedingungen in Schritt (0) des obigen Verfahrens von Fletcher genügt, ist also $x \in M$ und I eine Indexmenge mit $\{m_0+1, \ldots, m\} \subset I \subset I(x)$

und Rang $(A_I) = q$, so unterscheiden wir zwischen zwei Fällen. Ist Q positiv definit auf Kern (A_I), ist also $p^TQp > 0$ für alle $p \in \mathbb{R}^n \setminus \{0\}$ mit $A_I p = 0$, so kann, falls man nicht mit einer gefundenen stationären Lösung das Verfahren beendet, wie im obigen Fletcher-Verfahren ein neues Paar (x^+, I^+) berechnet werden, da dann die Matrix

$$K_I := \begin{pmatrix} Q & A_I^T \\ A_I & 0 \end{pmatrix} \in \mathbb{R}^{(q+n)\times(q+n)}$$

nichtsingulär ist (siehe Aufgabe 2). Andernfalls bestimme man ein $p \in \text{Kern}\,(A_I) \setminus \{0\}$ mit $p^TQp \le 0$ und $(c+Qx)^Tp \le 0$. Dann ist

$$f(x+tp) = f(x) + t(c+Qx)^Tp + \frac{1}{2}\,t^2p^TQp \le f(x)$$

für alle $t \ge 0$. Ist $a_i^Tp \ge 0$ für alle $i \notin I$, so ist

$$x + tp \in L(x) := \{z \in \mathbb{R}^n : f(z) \le f(x)\} \cap M$$

für alle $t \ge 0$. Die Niveaumenge $L(x)$ wäre nicht beschränkt, für $(c+Qx)^Tp < 0$ oder $p^TQp < 0$ würde sogar $\inf(\mathrm{P}) = -\infty$ folgen. Es liegt daher nahe, in diesem Falle mit einer entsprechenden Meldung auszusteigen. Existiert dagegen ein $i \notin I$ mit $a_i^Tp < 0$, so berechne man die maximale Schrittweite

$$s(x,p) := \min\Big\{\frac{b_i - a_i^Tx}{a_i^Tp} : i \notin I,\ a_i^Tp < 0\Big\} = \frac{b_r - a_r^Tx}{a_r^Tp}$$

und setze anschließend $x^+ := x + s(x,p)p$ sowie $I^+ := I \cup \{r\}$. Das in diesem Schritt gefundene neue Paar (x^+, I^+) genügt wiederum den Bedingungen in Schritt (0), womit das Verfahren von Fletcher dem Prinzip nach auch im indefiniten Fall beschrieben ist.

Natürlich sind hier noch viele Fragen offen geblieben. Vor allem interessiert, wie ein Iterationsschritt effizient und stabil durchgeführt werden kann. Hierauf wollen wir nur sehr kurz eingehen, näheres findet man z. B. bei P. E. Gill, W. Murray (1978), M. J. Best (1984)[6] und R. Fletcher (1987). Damit in diesem Unterabschnitt wenigstens ein Lemma steht, formulieren wir:

Lemma 1.1 *Sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch, $A \in \mathbb{R}^{m\times n}$ und* Rang $(A) = m$. *Gegeben seien ferner eine orthogonale Matrix*

$$Z = (\underbrace{Z^{(1)}}_{m} \quad \underbrace{Z^{(2)}}_{n-m}) \in \mathbb{R}^{n\times n},$$

eine obere Dreiecksmatrix $R \in \mathbb{R}^{m\times m}$, eine untere Dreiecksmatrix $L \in \mathbb{R}^{(n-m)\times(n-m)}$ mit Einsen in der Diagonalen und eine Diagonalmatrix $D \in \mathbb{R}^{(n-m)\times(n-m)}$ mit

$$(*) \qquad Z^{(1)T}A^T = R, \qquad AZ^{(2)} = 0, \qquad Z^{(2)T}QZ^{(2)} = LDL^T.$$

Hiermit gilt:

[6] Best, M. J. (1984) "Equivalence of some quadratic programming algorithms." Mathematical Programming 30, 71–87.

(a) *Es ist* $\text{Kern}\,(A) = \text{Bild}\,(Z^{(2)})$.

(b) *Die Matrix* Q *ist auf* $\text{Kern}\,(A)$ *genau dann positiv definit, wenn* $Z^{(2)T}QZ^{(2)}$ *positiv definit ist bzw. die Diagonalmatrix* D *nur positive Elemente enthält.*

(c) *Ist* Q *positiv definit auf* $\text{Kern}\,(A)$*, so ist bei gegebenem* $g \in \mathbb{R}^n$ *die eindeutige Lösung* (p, y) *von*

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -y \end{pmatrix} = - \begin{pmatrix} g \\ 0 \end{pmatrix}$$

gegeben durch

$$p := -Z^{(2)}(LDL^T)^{-1}Z^{(2)T}g, \qquad y := R^{-1}Z^{(1)T}(g + Qp).$$

Man kann also p *und* y *dadurch berechnen, dass man zunächst* $w \in \mathbb{R}^{n-m}$ *aus* $LDL^T w = -Z^{(2)T}g$ *durch Rückwarts- und Vorwärtseinsetzen und anschließend* $p = Z^{(2)}w$ *erhält bzw.* $g^+ := g + Qp$ *und dann* y *aus* $Ry = Z^{(1)T}g^+$ *durch Rückwärtseinsetzen bestimmt.*

Beweis: Wegen $AZ^{(2)} = 0$ ist offensichtlich $\text{Bild}\,(Z^{(2)}) \subset \text{Kern}\,(A)$. Die $n - m$ Spalten von $Z^{(2)}$ sind linear unabhängig, so dass $\text{Bild}\,(Z^{(2)})$ ein $(n-m)$-dimensionaler linearer Teilraum des $\mathbb{R}^n$ ist. Ferner folgt aus $\text{Rang}\,(A) = m$, dass auch $\text{Kern}\,(A)$ ein $(n-m)$-dimensionaler linearer Teilraum des $\mathbb{R}^n$ ist. Hiermit ist (a) bewiesen.

Sei $Z^{(2)T}QZ^{(2)}$ positiv definit. Ist dann $p \in \text{Kern}\,(A) \setminus \{0\}$, so existiert wegen (a) ein $w \in \mathbb{R}^{n-m} \setminus \{0\}$ mit $p = Z^{(2)}w$. Dann ist aber

$$0 < w^T Z^{(2)T} Q Z^{(2)} w = (Z^{(2)}w)^T Q (Z^{(2)}w) = p^T Q p,$$

d. h. Q ist auf $\text{Kern}\,(A)$ positiv definit. Die Umkehrung folgt genauso.

Der Beweis von (c) ist einfach, er bleibt dem Leser überlassen. □ □

Bemerkung: Die folgende Bemerkung soll klären, unter welchen Voraussetzungen Matrizen Z, R, L und D mit den in Lemma 1.1 angegebenen Eigenschaften existieren.

Die Matrix $A^T \in \mathbb{R}^{n\times m}$ besitzt eine QR-Zerlegung

$$A^T = Z \begin{pmatrix} R \\ 0 \end{pmatrix} \begin{matrix} \} \, m \\ \} \, n-m \end{matrix}$$

mit einer orthogonalen Matrix $Z \in \mathbb{R}^{n\times n}$und einer (wegen $\text{Rang}\,(A^T) = m$) nichtsingulären oberen Dreiecksmatrix. Mit der angegebenen Partitionierung

$$Z = (\underbrace{Z^{(1)}}_{m} \quad \underbrace{Z^{(2)}}_{n-m})$$

sind die ersten beiden Gleichungen in $(*)$ erfüllt. Jede symmetrische Matrix $B \in \mathbb{R}^{k\times k}$, deren Hauptabschnittsdeterminanten sämtlich von Null verschieden sind, besitzt eine eindeutige LDL^T-Zerlegung, es existiert also eine eindeutige Darstellung $B = LDL^T$ mit einer unteren Dreiecksmatrix $L \in \mathbb{R}^{k\times k}$, die nur Einsen in der Diagonalen enthält, und einer Diagonalmatrix $D \in \mathbb{R}^{k\times k}$. Dieses einfache Ergebnis folgt aus der eindeutigen Existenz einer LR-Zerlegung einer Matrix, deren Hauptabschnittsdeterminanten

sämtlich ungleich Null sind. Die in Lemma 1.1 auftretenden Matrizen existieren also, wenn sämtliche Hauptabschnittsdeterminanten von $Z^{(2)T}QZ^{(2)}$ nicht verschwinden, was insbesondere dann der Fall ist, wenn Q auf Kern (A) positiv definit ist. □

Sei (x, I) ein Paar mit $x \in M$ sowie $\{m_0 + 1, \ldots, m\} \subset I \subset I(x)$ (mit $q := \#(I)$ Elementen) und Rang $(A_I) = q$. Entsprechend den Voraussetzungen in Lemma 1.1 seien eine orthogonale Matrix

$$Z_I = (\underbrace{Z_I^{(1)}}_{q} \quad \underbrace{Z_I^{(2)}}_{n-q}) \in \mathbb{R}^{n \times n},$$

eine obere Dreicksmatrix $R_I \in \mathbb{R}^{q \times q}$, eine untere Dreiecksmatrix mit Einsen in der Diagonalen $L_I \in \mathbb{R}^{(n-q)\times(n-q)}$ und eine Diagonalmatrix $D_I \in \mathbb{R}^{(n-q)\times(n-q)}$ mit

$$Z_I^{(1)T} A_I^T = R_I, \qquad A_I Z_I^{(2)} = 0, \qquad Z_I^{(2)T} Q Z_I^{(2)} = L_I D_I L_I^T.$$

Nun interessieren selbstverständlich Update-Formeln für diese Matrizen, wobei die beiden Fälle zu unterscheiden sind, ob $I^+ := I \cup \{r\}$ oder $I^+ := I \setminus \{l\}$ gilt. Der erste Fall ist der angenehmere, weil wegen Kern $(A_{I\cup\{r\}}) \subset$ Kern (A_I) aus der positiven Definitheit von Q auf Kern (A_I) auch die von Q auf Kern $(A_{I\cup\{r\}})$ folgt. Die entsprechende Aussage ist natürlich beim Wegfall einer Restriktion i. Allg. nicht mehr richtig. Auf Einzelheiten zu diesen Update-Formeln wollen wir hier verzichten und verweisen auf P. E. Gill, W. Murray (1978) und die ausführlichen Hinweise zu den Aufgaben 6 und 7.

Zum Schluss dieses Unterabschnittes über das Verfahren von Fletcher (und ähnliche Verfahren) wollen wir noch auf einige Punkte hinweisen, auf die nicht eingegangen wird. So kann es z. B. sinnvoll sein, spezielle Restriktionen gesondert zu behandeln. Sehr oft treten sogenannte Box-Constraints auf, also Restriktionen der Form $l \le x \le u$, wobei gewisse Komponenten von l und u auch gleich $-\infty$ bzw. gleich $+\infty$ sein können. Es ist nicht schwierig, höchstens etwas mühsam, die Methode der aktiven Restriktionen adäquat zu modifizieren. Ferner weisen wir darauf hin, dass weder das Verfahren von Goldfarb-Idnani noch das Verfahren von Fletcher in der vorgestellten Form für hochdimensionale und dann i. Allg. speziell strukturierte quadratische Programme geeignet sind. Hinweise zu geeigneten Modifikationen findet man z. B. bei N. I. M. Gould (1991)[7].

3.1.2 Aufgaben

1. Sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und $A \in \mathbb{R}^{m\times n}$ eine Matrix mit Rang $(A) = m$. Man zeige, dass dann die Matrix

$$K := \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix}$$

nichtsingulär ist. Ferner zeige man, dass mit

$$N := (AQ^{-1}A^T)^{-1}AQ^{-1}, \qquad H := Q^{-1}(I - A^T N)$$

[7] Gould, N. I. M (1991) "An algorithm for large-scale quadratic programming." IMA Journal of Numerical Analysis 11, 299–324.

die Inverse K^{-1} gegeben ist durch

$$K^{-1} = \begin{pmatrix} H & N^T \\ N & -NQN^T \end{pmatrix}.$$

Hinweis: Diese Aussage findet man schon bei R. Fletcher (1971).

2. Sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch, die Matrix $A \in \mathbb{R}^{m\times n}$ habe vollen Zeilenrang, d. h. es sei Rang $(A) = m$. Hiermit definiere man die Matrix

$$K := \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}$$

und zeige:

   (a) Ist Q auf Kern (A) positiv definit. ist also $p^TQp > 0$ für alle $p \in \mathbb{R}^n \setminus \{0\}$ mit $Ap = 0$, so ist K nichtsingulär.

   (b) Ist Q positiv semidefinit und K nichtsingulär, so ist Q auf Kern (A) positiv definit.

3. Sei $I \subset \{1,\ldots,m\}$ mit $q := \#(I)$ eine (nichtleere) Indexmenge, $r \in \{1,\ldots,m\} \setminus I$ und $\{a_i\}_{i\in I\cup\{r\}}$ linear unabhängig. Die Matrizen

$$N_I := (A_IQ^{-1}A_I^T)^{-1}A_IQ^{-1} \in \mathbb{R}^{q\times n}, \qquad H_I := Q^{-1}(I - A_I^TN_I) \in \mathbb{R}^{n\times n}$$

und die Vektoren

$$z := H_Ia_r \in \mathbb{R}^n, \qquad r_I := N_Ia_r \in \mathbb{R}^q$$

seien bekannt. Man zeige, daß

$$N_{I\cup\{r\}} = \begin{pmatrix} N_I - \dfrac{r_Iz^T}{a_r^Tz} \\ \dfrac{z^T}{a_r^Tz} \end{pmatrix}, \qquad H_{I\cup\{r\}} = H_I - \frac{zz^T}{a_r^Tz}.$$

4. Sei $I \subset \{1,\ldots,m\}$ (wieder sei $q := \#(I)$) eine nichtleere Indexmenge mit der Eigenschaft, dass die Vektoren $\{a_i\}_{i\in I} \subset \mathbb{R}^n$ linear unabhängig sind. Insbesondere sei also $1 \le q \le n$ und Rang $(A_I) = q$. Bekannt seien die Matrizen

$$N_I := (A_IQ^{-1}A_I^T)^{-1}A_IQ^{-1} \in \mathbb{R}^{q\times n}, \qquad H_I := Q^{-1}(I - A_I^TN_I) \in \mathbb{R}^{n\times n}.$$

Ferner sei $l \in I$ vorgegeben. Man überlege sich, wie man auf effiziente Weise die analog definierten Matrizen $N_{I\setminus\{l\}}$ und $H_{I\setminus\{l\}}$ berechnen kann.

5. Gegeben sei das (übliche) quadratische Programm (P) mit der symmetrischen, positiv definiten Matrix $Q \in \mathbb{R}^{n\times n}$. Das Paar (x, I) genüge den Bedingungen in Schritt (0) des Fletcher-Verfahrens. Sei (p, y_I) die eindeutige Lösung des linearen Gleichungssystems in Schritt (1). Es sei $x+p \in M$ und $y_l < 0$ für ein $l \in I \cap \{1,\ldots,m_0\}$. Wie im Verfahren von Fletcher setze man $x^+ := x + p$ und $I^+ := I \setminus \{l\}$. Ist dann p^+ die Lösung von

$$\text{Minimiere} \quad (c + Qx^+)^Tz + \tfrac{1}{2}\,z^TQz \quad \text{unter der Nebenbedingung} \quad A_{I^+}z = 0,$$

so ist $a_l^Tp^+ > 0$ und $(c + Qx^+)^Tp^+ = -(p^+)^TQp^+ < 0$, insbesondere also $p^+ \neq 0$.

6. Sei $I \subset \{1,\ldots,m\}$ (mit $q := \#(I)$) und $r \in \{1,\ldots,m\} \setminus I$. Die wie üblich definierten Matrizen A_I und $A_{I\cup\{r\}}$ mögen maximalen Zeilenrang q bzw. $q+1$ besitzen, die Matrix $Q \in \mathbb{R}^{n\times n}$ sei symmetrisch und auf Kern (A_I) positiv definit. Bekannt seien die orthogonale Matrix

$$Z_I = (\underbrace{Z_I^{(1)}}_{q} \quad \underbrace{Z_I^{(2)}}_{n-q}) \in \mathbb{R}^{n\times n},$$

die obere Dreiecksmatrix R_I, die untere Dreiecksmatrix mit Einsen in der Diagonalen L_I sowie die (positiv definite) Diagonalmatrix D_I mit

$$Z_I^{(1)T} A_I^T = R_I, \qquad A_I Z_I^{(2)} = 0, \qquad Z_I^{(2)T} Q Z_I^{(2)} = L_I D_I L_I^T.$$

Man setze $I^+ := I \cup \{r\}$ und entwickle ein effizientes, stabiles Verfahren zur Berechnung der Matrizen Z_{I^+}, R_{I^+}, L_{I^+} und D_{I^+} mit den zu Z_I, R_I, L_I bzw. D_I analogen Eigenschaften.

7. Sei $I \subset \{1,\ldots,m\}$ eine Indexmenge mit q Elementen, $l \in I$ und $I^+ := I\setminus\{l\}$. Die Matrix A_I habe maximalen Zeilenrang, es sei also Rang $(A_I) = q$, ferner sei die symmetrische Matrix $Q \in \mathbb{R}^{n\times n}$ auf Kern (A_I) positiv definit. Bekannt seien die orthogonale Matrix

$$Z_I = (\underbrace{Z_I^{(1)}}_{q} \quad \underbrace{Z_I^{(2)}}_{n-q}) \in \mathbb{R}^{n\times n},$$

die obere Dreiecksmatrix R_I, die untere Dreiecksmatrix mit Einsen in der Diagonalen L_I sowie die (positiv definite) Diagonalmatrix D_I mit

$$Z_I^{(1)T} A_I^T = R_I, \qquad A_I Z_I^{(2)} = 0, \qquad Z_I^{(2)T} Q Z_I^{(2)} = L_I D_I L_I^T.$$

Man entwickle ein effizientes, stabiles Verfahren zur Berechnung der Matrizen Z_{I^+}, R_{I^+}, L_{I^+} und D_{I^+} mit den zu Z_I, R_I, L_I bzw. D_I analogen Eigenschaften.

8. Man programmiere das Verfahren von Fletcher und teste das Programm an konvexen, quadratischen Optimierungsaufgaben mit den folgenden Daten:

   (a) Es sei $m := 4$, $m_0 := 4$ und $n := 3$, ferner sei

$$Q := \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 0 \\ 2 & 0 & 2 \end{pmatrix}, \qquad c := \begin{pmatrix} -8 \\ -6 \\ -4 \end{pmatrix}$$

   sowie

$$A := \begin{pmatrix} -1 & -1 & -2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad b := \begin{pmatrix} -3 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

   Wie bei W. Hock, K. Schittkowski (1981)[8] starte man mit der zulässigen Lösung $x := (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})^T$ (und damit $I := \emptyset$).

[8] Hock, W. and K. Schittkowski (1981) *Test Examples for Nonlinear Programming Codes.* Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin-Heidelberg-New York.

(b) Es sei $m := 7$, $m_0 := 7$ und $n := 4$, ferner sei

$$Q := \begin{pmatrix} 2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 2 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \qquad c := \begin{pmatrix} -1 \\ -3 \\ 1 \\ -1 \end{pmatrix}$$

und

$$A := \begin{pmatrix} -1 & -2 & -1 & -1 \\ -3 & -1 & -2 & 1 \\ 0 & 1 & 4 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad b := \begin{pmatrix} -5 \\ -4 \\ \frac{3}{2} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Wie bei W. Hock, K. Schittkowski (1981, S. 96) starte man mit der zulässigen Lösung $x := (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})^T$ (und damit $I := \emptyset$).

9. Gegeben sei ein lineares Gleichungssystem der Form

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Hierbei sei $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, $A \in \mathbb{R}^{m \times n}$ mit Rang $(A) = m$. Man zeige, dass man obiges lineares Gleichungssystem mit den folgenden Schritten lösen kann:

- Bestimme eine QR-Zerlegung von $A^T \in \mathbb{R}^{n \times m}$, berechne also, etwa mit dem Householder-Verfahren, eine orthogonale Matrix $Z \in \mathbb{R}^{n \times n}$ und eine (nichtsinguläre) obere Dreiecksmatrix $R \in \mathbb{R}^{m \times m}$ mit

$$ZA^T = \begin{pmatrix} R \\ 0 \end{pmatrix}.$$

Simultan berechne man

$$\begin{pmatrix} c \\ d \end{pmatrix} := Za, \qquad \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} := ZQZ^T.$$

Hierbei ist $c \in \mathbb{R}^m$, $d \in \mathbb{R}^{n-m}$, ferner ist $B_{22} \in \mathbb{R}^{(n-m)\times(n-m)}$ symmetrisch und positiv definit (Beweis?).

- Durch Vorwärtseinsetzen bestimme man $u \in \mathbb{R}^m$ aus $R^T u = b$.
- Mit Hilfe des Cholesky-Verfahrens berechne man $v \in \mathbb{R}^{n-m}$ aus $B_{22} v = d - B_{21} u$.
- Gewinne die Anteile $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ der gesuchten Lösung aus

$$x := Z^T \begin{pmatrix} u \\ v \end{pmatrix}$$

und

$$Ry = c - B_{11} u - B_{12} v$$

durch Rückwärtseinsetzen.

3.2 Das duale Verfahren von Goldfarb-Idnani

In diesem Abschnitt betrachten wir wieder das quadratische Programm

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2}\, x^T Q x \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}, \end{array} \right.$$

wobei die Matrix Q als positiv definit vorausgesetzt wird. Ziel wird es sein, das duale Verfahren von Goldfarb-Idnani (siehe D. Goldfarb, A. Idnani (1982, 1983)[9] und auch, in einer allerdings unbefriedigenden Darstellung, P. Spellucci (1993, S. 293 ff.))[10] darzustellen, welches für moderat große, nicht speziell strukturierte quadratische Programme, bei denen die Hessesche der Zielfunktion positiv definit[11] ist, sicherlich zu den besten Verfahren gehört. Ähnlich wie das duale Simplexverfahren bei linearen Programmen erzeugt auch das Verfahren von Goldfarb-Idnani „optimale“, aber unzulässige Näherungslösungen mit monoton wachsenden Zielfunktionswerten. Ein Vorteil eines dualen Verfahrens gegenüber einem primalen, in dem eine Folge zulässiger Lösungen mit fallenden Kosten berechnet wird, besteht darin, dass nicht in einer ersten Phase eine zulässige Startnäherung bestimmt werden muß. Im ersten Unterabschnitt werden wir das Verfahren in seinen Grundzügen beschreiben, im darauf folgenden Unterabschnitt gehen wir auf einige Einzelheiten einer möglichen Implementation des Verfahrens ein.

3.2.1 Beschreibung des Verfahrens

Bei einer gegebenen Indexmenge $I \subset \{1, \ldots, m\}$ definieren wir das (im Vergleich zu (P)) relaxierte quadratische Programm (P_I) durch

$$\text{(P}_I\text{)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2}\, x^T Q x \quad \text{auf} \\ M_I := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i \in I \cap \{1, \ldots, m_0\}) \\ a_i^T x = b_i & (i \in I \cap \{m_0 + 1, \ldots, m\}) \end{array} \right\}. \end{array} \right.$$

Als grundlegend wird sich die folgende Definition herausstellen.

[9] Goldfarb, D. and A. Idnani (1982) “Dual and primal-dual methods for solving strictly convex quadratic programs.” In: *Numerical Analysis, Proceedings Cocoyoc, Mexico 1982.* (ed. J. P. Hennart), Lecture Notes in Mathematics 909, Springer-Verlag, Berlin.

Goldfarb, D. and A. Idnani (1983) “A numerically stable dual method for solving strictly convex quadratic programs.” Mathematical Programming 27, 1–33.

[10] Spellucci, P. (1993) *Numerische Verfahren der nichtlinearen Optimierung.* Birkhäuser, Basel-Boston-Berlin.

[11] Eine Verallgemeinerung des Goldfarb-Idnani-Verfahrens auf den Fall, dass die Hessesche der Zielfunktion nur positiv semidefinit ist, ist kürzlich von

N. L. Boland (1997) “A dual-active-set algorithm for positive semi-definite quadratic programming.” Mathematical Programming 78, 1–27.

angegeben worden.

Definition 2.1 Ein Paar (x, I) mit $x \in \mathbb{R}^n$ und $I \subset \{1, \ldots, m\}$ heißt ein *Lösungspaar* für das quadratische Programm (P), wenn

1. I eine Indexmenge ist, für die $\{a_i\}_{i \in I} \subset \mathbb{R}^n$ linear unabhängig sind,
2. das relaxierte quadratische Programm (P_I) zulässig ist, also $M_I \neq \emptyset$ gilt,
3. $x \in M_I$ die (eindeutige) Lösung von (P_I) ist und zusätzlich $a_i^T x = b_i$ für alle $i \in I \cap \{1, \ldots, m_0\}$ gilt.

Mit $(-Q^{-1}c, \emptyset)$ kann ein Lösungspaar für das quadratische Programm (P) sofort angegeben werden. Klar ist, dass es nur endlich viele Lösungspaare gibt. Ist ferner (x, I) ein Lösungspaar mit $x \in M$, so ist x einerseits zulässig für (P), andererseits Lösung des relaxierten Problems (P_I), insgesamt also die Lösung von (P). Ist umgekehrt (P) zulässig und $x^* \in M$ die Lösung von (P), so existiert eine Indexmenge $I^* \subset \{1, \ldots, m\}$ derart, dass (x^*, I^*) ein Lösungspaar ist (siehe auch Aufgabe 1).

Ist $I := \{i_1, \ldots, i_q\} \subset \{1, \ldots, m\}$ eine Indexmenge mit $q := \#(I)$ Elementen[12], so setzen wir naheliegenderweise (wie im letzten Abschnitt)

$$A_I := \begin{pmatrix} a_{i_1}^T \\ \vdots \\ a_{i_q}^T \end{pmatrix} \in \mathbb{R}^{q \times n}, \qquad b_I := \begin{pmatrix} b_{i_1} \\ \vdots \\ b_{i_q} \end{pmatrix} \in \mathbb{R}^q.$$

Ähnliche Bezeichnungen für andere Matrizen oder Vektoren sind entsprechend zu verstehen.

Bemerkung: Ein Paar (x, I) mit $x \in \mathbb{R}^n$, $I \subset \{1, \ldots, m\}$ (und $q := \#(I)$) ist genau dann ein Lösungspaar für das quadratische Programm (P), wenn Rang $(A_I) = q$, $A_I x = b_I$ und ein $y_I \in \mathbb{R}^q$ mit $c + Qx = A_I^T y_I$ und $y_i \geq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$ existiert.

Das zu (P) duale Programm lautet

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \phi(y) := b^T y - \frac{1}{2}\,(A^T y - c)^T Q^{-1} (A^T y - c) \quad \text{auf} \\ \qquad\qquad N := \{y \in \mathbb{R}^m : y_i \geq 0 \quad (i = 1, \ldots, m_0)\}. \end{cases}$$

Sei (x, I) ein Lösungspaar und $y_I \in \mathbb{R}^q$ ein zugehöriger Vektor von Lagrange-Multiplikatoren. Ergänzt man y_I zu einem Vektor $y \in \mathbb{R}^m$, indem man $y_i := 0$ für $i \in \{1, \ldots, m\} \setminus I$ setzt, so ist $y \in N$ dual zulässig. Ferner ist

$$\begin{aligned} \phi(y) &= b^T y - \tfrac{1}{2}\,(A^T y - c)^T Q^{-1} (A^T y - c) \\ &= b_I^T y_I - \tfrac{1}{2}\,(A_I^T y_I - c)^T Q^{-1} (A_I^T y_I - c) \\ &= (A_I^T y_I)^T x - \tfrac{1}{2}\,(Qx)^T Q^{-1} (Qx) \\ &= (c + Qx)^T x - \tfrac{1}{2}\, x^T Q x \\ &= f(x). \end{aligned}$$

[12] Diese Bezeichnung werden wir beibehalten: Ist $I \subset \{1, \ldots, m\}$ eine Indexmenge, so sei grundsätzlich $q := \#(I)$ die Anzahl der Elemente von I.

Daher ist ein Lösungspaar (x, I) „optimal“ in dem Sinne, dass ein dual zulässiges y mit $f(x) = \phi(y)$ existiert. Der schwache Dualitätssatz liefert erneut: Ist zusätzlich $x \in M$, so ist x die Lösung von (P). □

Nun können wir schon einen Modellalgorithmus zum von D. Goldfarb, A. Idnani (1982, 1983) entwickelten Verfahren zur numerischen Behandlung des quadratischen Programms (P) angeben.

- Berechne das Lösungspaar $(x^0, I^0) := (-Q^{-1}c, \emptyset)$.
- Für $k = 0, 1, \ldots$:
  - Falls $x^k \in M$, dann: STOP, x^k ist die Lösung von (P).
  - Andernfalls:
    - Bestimme verletzte Restriktion $p \in \{1, \ldots, m\} \setminus I^k$.
    - Falls $M_{I^k \cup \{p\}} = \emptyset$, dann: STOP, (P) ist nicht zulässig.
    - Andernfalls:
      - Bestimme Lösungspaar (x^{k+1}, I^{k+1}) mit $I^{k+1} = \overline{I}^k \cup \{p\}$, $\overline{I}^k \subset I^k$ und $f(x^{k+1}) > f(x^k)$.

Bemerkung: Interessant für die praktische Durchführung des obigen Modellalgorithmus ist vor allem, wie bei Vorliegen eines aktuellen Lösungspaares (x, I) und einer durch x verletzten Restriktion $p \in \{1, \ldots, m\} \setminus I$ auf effiziente Weise festgestellt werden kann, ob das relaxierte Programm $(\mathrm{P}_{I \cup \{p\}})$ nicht zulässig ist bzw. $M_{I \cup \{p\}} = \emptyset$ gilt, bzw. wie andernfalls ein neues Lösungspaar (x^+, I^+) mit $I^+ = \overline{I} \cup \{p\}$, $\overline{I} \subset I$ und $f(x^+) > f(x)$ bestimmt werden kann. Auf eine naheliegende Methode zur Beantwortung der zweiten Frage gehen wir in Aufgabe 2 ein. Ist die Durchführbarkeit des Modellalgorithmus gesichert, so ist klar, dass er nach endlich vielen Schritten abbricht, und zwar entweder mit der Lösung $x^* \in M$ von (P) oder der Information, dass (P) nicht zulässig ist. Denn einerseits gibt es nur endlich viele Lösungspaare, andererseits vergrößert sich der Zielfunktionswert von Schritt zu Schritt, wodurch ausgeschlossen wird, dass man zu einem einmal berechneten Lösungspaar zurückkehrt. □

Wir stellen uns nun auf den Standpunkt, es sei ein (aktuelles) Lösungspaar (x, I) mit einem zugehörigen Lagrange-Vektor $y_I \in \mathbb{R}^q$ bekannt. Ist $x \in M$, so ist x die Lösung von (P). Andernfalls wird eine durch x verletzte Restriktion $p \in \{1, \ldots, m\} \setminus I$ bestimmt. Für diese ist also $b_p > a_p^T x$, falls $p \in \{1, \ldots, m_0\}$, bzw. $b_p \neq a_p^T x$, falls $p \in \{m_0 + 1, \ldots, m\}$. Bei der Berechnung eines neuen Lösungspaares (x^+, I^+) mit $I^+ = \overline{I} \cup \{p\}$, $\overline{I} \subset I$ und $f(x^+) > f(x)$ werden zwei Fälle unterschieden.

- $a_p \notin \operatorname{span}\{a_i : i \in I\}$.

  Dann sind auch $\{a_i\}_{i \in I \cup \{p\}}$ linear unabhängig. Wenn möglich wird $I^+ := I \cup \{p\}$ gesetzt.

- $a_p \in \operatorname{span}\{a_i : i \in I\}$.

  Dann ist a_p von $\{a_i\}_{i \in I}$ linear abhängig. Da p auf alle Fälle in I^+ aufgenommen wird, muss mindestens ein Element aus I entfernt werden.

Beide Fälle werden getrennt untersucht. Die genaue Vorgehensweise wird in den beiden folgenden Lemmata beschrieben.

Lemma 2.2 *Sei (x, I) ein Lösungspaar für das quadratische Programm* (P) *mit einem zugehörigen Lagrange-Vektor $y_I = (y_i)_{i\in I} \in \mathbb{R}^q$, $p \in \{1, \ldots, m\}\backslash I$ eine durch x verletzte Restriktion und $a_p \notin \operatorname{span}\{a_i : i \in I\}$. Dann berechnet der folgende Algorithmus ein neues Lösungspaar (x^+, I^+) mit $I^+ = \overline{I} \cup \{p\}$, $\overline{I} \subset I$ und $f(x^+) > f(x)$ sowie einen zugehörigen Lagrange-Vektor y_{I^+}.*

(0) Gegeben (x, I, y_I, f, θ) mit $f := f(x)$, $\theta := 0$.

(1) Berechne

$$r_I := \begin{cases} (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} a_p, & \text{falls} \quad I \neq \emptyset, \\ 0 & \text{sonst} \end{cases}$$

sowie

$$z := Q^{-1}(a_p - A_I^T r_I), \qquad t_1 := \frac{b_p - a_p^T x}{a_p^T z}.$$

(2) Falls $I = \emptyset$ oder $y_i - t_1 r_i \geq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$, dann: STOP, durch

$$(x^+, I^+) := (x + t_1 z, I \cup \{p\}), \qquad y_{I^+} := \begin{pmatrix} y_I - t_1 r_I \\ \theta + t_1 \end{pmatrix}$$

ist ein Lösungspaar mit zugehörigen Lagrange-Vektor und dem Zielfunktionswert

$$f^+ := f(x^+) = f + t_1 \left(\tfrac{1}{2} t_1 + \theta\right) a_p^T z > f(x) = f$$

gegeben.

(3) Berechne

$$t_2 := \left\{ \begin{array}{ll} \min\left\{\dfrac{y_i}{r_i} : i \in I \cap \{1, \ldots, m_0\},\ r_i > 0\right\} & \text{für} \quad t_1 > 0, \\ \max\left\{\dfrac{y_i}{r_i} : i \in I \cap \{1, \ldots, m_0\},\ r_i < 0\right\} & \text{für} \quad t_1 < 0 \end{array} \right\} = \frac{y_l}{r_l}.$$

Anschließend setze

$$x^- := x + t_2 z, \qquad I^- := I \setminus \{l\}, \qquad y_{I^-} := T_l(y_I - t_2 r_I)$$

sowie

$$f^- := f + t_2 \left(\tfrac{1}{2} t_2 + \theta\right) a_p^T z\,, \qquad \theta^- := \theta + t_2.$$

Hierbei entferne der Operator $T_l\colon \mathbb{R}^q \longrightarrow \mathbb{R}^{q-1}$ die Komponente mit dem Index l. Dann mache man den Update

$$(x, I, y_I, f, \theta) := (x^-, I^-, y_{I^-}, f^-, \theta^-)$$

und gehe nach (1).

Beweis: Klar ist, dass der im Lemma angegebene Algorithmus nach endlich vielen Schritten abbricht, da aus der ursprünglich gegebenen Indexmenge I nur Elemente entfernt werden und der Algorithmus spätestens dann stoppt, wenn man zur leeren Menge kommt.

Wir nehmen an, es sei das 5-Tupel (x, I, y_I, f, θ) mit $x \in \mathbb{R}^n$, $I \subset \{1, \ldots, m\}$, $y_I \in \mathbb{R}^q$, $f \in \mathbb{R}$ und $\theta \in \mathbb{R}$ gegeben. Es sei Rang $(A_I) = q$ und p der Index einer durch x verletzten Restriktion mit $a_p \notin \operatorname{span}\{a_i : i \in I\}$. Ferner gelte

$$A_I x = b_I, \qquad c + Qx = A_I^T y_I + \theta a_p, \qquad y_i \geq 0 \quad (i \in I \cap \{1, \ldots, m_0\})$$

sowie

$$f = f(x), \qquad \theta \operatorname{sign}(b_p - a_p^T x) \geq 0.$$

Beim Start ist (x, I) das gegebene Lösungspaar, y_I ein zugehöriger Lagrange-Vektor, $f = f(x)$ und $\theta = 0$. Wie in (1) angegeben, berechnet man anschließend r_I und z. Wegen $a_p \notin \operatorname{span}\{a_i : i \in I\}$ ist $z \neq 0$. Ist $I \neq \emptyset$, so ist $A_I z = 0$ und daher

$$a_p^T z = (a_p - Qz)^T z + z^T Q z = (A_I^T r_I)^T z + z^T Q z = z^T Q z > 0.$$

Ist dagegen $I = \emptyset$, so ist $Qz = a_p$ und daher ebenfalls $a_p^T z = z^T Q z > 0$. Daher ist t_1 in Schritt (1) wohldefiniert, es ist $t_1 \neq 0$ und $\operatorname{sign} t_1 = \operatorname{sign}(b_p - a_p^T x)$. Für $t \in \mathbb{R}$ sei $x(t) := x + tz$. Dann ist

$$A_I x(t) = A_I x + t \underbrace{A_I z}_{=0} = b_I$$

und

$$a_p^T x(t) = a_p^T x + t a_p^T z = b_p + a_p^T z \left(t - \frac{b_p - a_p^T x}{a_p^T z}\right).$$

Ferner ist

$$c + Qx(t) = c + Qx + tQz = A_I^T y_I + \theta a_p + t(a_p - A_I^T r_I) = A_I^T (y_I - t r_I) + (\theta + t) a_p.$$

Hieraus folgt: Ist $I = \emptyset$ oder $y_i - t_1 r_i \geq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$, so ist durch

$$(x^+, I^+) := (x + t_1 z, I \cup \{p\}), \qquad y_{I^+} := \begin{pmatrix} y_I - t_1 r_I \\ \theta + t_1 \end{pmatrix}$$

ein neues Lösungspaar mit zugehörigem Lagrange-Vektor gegeben. Als zugehörigen Zielfunktionswert berechnet man

$$\begin{aligned} f(x^+) &= f(x) + (c + Qx)^T (x^+ - x) + \tfrac{1}{2} (x^+ - x)^T Q (x^+ - x) \\ &= f + t_1 (A_I^T y_I + \theta a_p)^T z + \tfrac{1}{2} t_1^2 z^T Q z \\ &= f + \underbrace{t_1 (\theta + \tfrac{1}{2} t_1) a_p^T z}_{>0} \\ &> f. \end{aligned}$$

Wird die Abfrage, ob $y_i - t_1 r_i \geq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$ ist, verneint, so wird t_2 in Schritt (3) berechnet. Ist $t_1 > 0$, so ist $0 \leq t_2 < t_1$. Für $t_1 < 0$ (dies kann nur bei einer

verletzten Gleichungsrestriktion eintreten) ist dagegen $t_1 < t_2 \leq 0$. In jedem Fall ist $y_i - t_2 r_i \geq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$ und $y_l - t_2 r_l = 0$. Im Algorithmus wird dann in Schritt (3) das neue 5-Tupel $(x^-, I^-, y_{I^-}, f^-, \theta^-)$ berechnet. Wegen $I^- := I \setminus \{l\}$ und Rang $(A_I) = q$ ist Rang $(A_{I^-}) = q - 1$. Ferner verletzt auch $x^- := x + t_2 z$ die p-te Restriktion wegen

$$b_p - a_p^T x^- = (t_1 - t_2)\, a_p^T z \begin{cases} > 0 & \text{für} \quad t_1 > 0, \\ < 0 & \text{für} \quad t_1 < 0. \end{cases}$$

Schließlich bestätigt man leicht, dass $(x^-, I^-, y_{I^-}, f^-, \theta^-)$ der Ausgangssituation mit $f^- \geq f$ genügt. Das Lemma ist damit bewiesen. □ □

Nun untersuchen wir den zweiten Fall, dass nämlich $a_p \in \operatorname{span}\{a_i : i \in I\}$ für ein gegebenes Lösungspaar (x, I), wobei p der Index einer durch x verletzten Restriktion ist. Die Vorgehensweise wird im folgenden Lemma erklärt. Das Lemma wird aus zwei Teilen bestehen. Im ersten wird ein Test dafür angegeben, dass $(\mathrm{P}_{I \cup \{p\}})$ und damit auch (P) nicht zulässig ist. Der zweite Teil des Lemmas geht davon aus, dass dieser Test passiert wurde. Es wird ein $l \in I \cap \{1, \ldots, m_0\}$ bestimmt, $I^- := I \setminus \{l\}$ gesetzt und ein Quintupel $(x^-, I^-, y_{I^-}, f^-, \theta^-)$ berechnet, mit dem in das Verfahren aus Lemma 2.2 eingestiegen werden kann.

Lemma 2.3 *Sei (x, I) ein Lösungspaar für* (P), *$y_I \in \mathbb{R}^q$ (mit $q := \#(I)$) ein zugehöriger Lagrange-Vektor, $p \in \{1, \ldots, m\} \setminus I$ eine durch x verletzte Restriktion mit $a_p \in \operatorname{span}\{a_i : i \in I\}$. Mit*

$$r_I := (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} a_p$$

gilt:

1. *Ist $r_i \operatorname{sign}(b_p - a_p^T x) \leq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$, so ist $(\mathrm{P}_{I \cup \{p\}})$ und damit auch* (P) *nicht zulässig.*

2. *Existiert ein $i \in I \cap \{1, \ldots, m_0\}$ mit $r_i \operatorname{sign}(b_p - a_p^T x) > 0$, so bestimme man $l \in I \cap \{1, \ldots, m_0\}$ mit*

$$\min\Big\{\frac{y_i}{r_i} : i \in I \cap \{1, \ldots, m_0\},\ r_i > 0\Big\} = \frac{y_l}{r_l}, \quad \textit{falls} \quad b_p - a_p^T x > 0,$$

bzw.

$$\max\Big\{\frac{y_i}{r_i} : i \in I \cap \{1, \ldots, m_0\},\ r_i < 0\Big\} = \frac{y_l}{r_l}, \quad \textit{falls} \quad b_p - a_p^T x < 0.$$

Setzt man anschließend

$$x^- := x, \qquad I^- := I \setminus \{l\}, \qquad \theta^- := \frac{y_l}{r_l}, \qquad y_{I^-} := T_l\Big(y_I - \frac{y_l}{r_l}\, r_I\Big),$$

wobei der Operator $T_l \colon \mathbb{R}^q \longrightarrow \mathbb{R}^{q-1}$ wieder die Komponente mit dem Index l entferne, so ist $a_p \notin \operatorname{span}\{a_i : i \in I^-\}$ und

$$A_{I^-} x^- = b_{I^-}, \qquad c + Q x^- = A_{I^-}^T y_{I^-} + \theta^- a_p$$

sowie

$$(y_{I^-})_i \geq 0 \quad (i \in I^- \cap \{1, \ldots, m_0\}), \qquad \theta^- \operatorname{sign}(b_p - a_p^T x^-) \geq 0.$$

Beweis: Nach Voraussetzung ist $a_p \in \operatorname{span}\{a_i : i \in I\}$ und $\operatorname{Rang}(A_I) = q$. Daher besitzt a_p eine eindeutige Darstellung der Form $a_p = A_I^T \lambda_I$ mit $\lambda_I \in \mathbb{R}^q$. Dann ist aber

$$A_I Q^{-1} a_p = A_I Q^{-1} A_I^T \lambda_I \qquad \text{bzw.} \qquad \lambda_I = (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} a_p = r_I.$$

Durch $a_p = A_I^T r_I$ ist also die gesuchte Darstellung von a_p gefunden.

Wir nehmen nun an, es sei $r_i \operatorname{sign}(b_p - a_p^T x) \leq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$. Angenommen, es gäbe ein $z \in \mathbb{R}^n$ derart, dass $x + z$ zulässig für $(\mathrm{P}_{I \cup \{p\}})$ ist. Wegen $A_I x = b_I$ ist dann notwendig

$$a_i^T z \begin{cases} \geq 0 & \text{für} \quad i \in I \cap \{1, \ldots, m_0\}, \\ = 0 & \text{für} \quad i \in I \cap \{m_0 + 1, \ldots, m\}. \end{cases}$$

Das Erfülltsein der p-ten Restriktion durch $x + z$ impliziert

$$a_p^T z \begin{cases} \geq b_p - a_p^T x \quad (> 0) & \text{für} \quad p \in \{1, \ldots, m_0\}, \\ = b_p - a_p^T x \quad (\neq 0) & \text{für} \quad p \in \{m_0 + 1, \ldots, m\}, \end{cases}$$

so dass $a_p^T z \operatorname{sign}(b_p - a_p^T x) > 0$ ist. Andererseits ist

$$\begin{aligned} a_p^T z \operatorname{sign}(b_p - a_p^T x) &= r_I^T A_I z \operatorname{sign}(b_p - a_p^T x) \\ &= \sum_{i \in I \cap \{1, \ldots, m_0\}} \underbrace{a_i^T z}_{\geq 0} \underbrace{r_i \operatorname{sign}(b_p - a_p^T x)}_{\leq 0} \\ &\leq 0. \end{aligned}$$

Damit ist die Annahme, $(\mathrm{P}_{I \cup \{p\}})$ sei zulässig, zum Widerspruch geführt. Der erste Teil des Lemmas ist bewiesen.

Zum Nachweis des zweiten Teiles nehmen wir an, es sei $r_i \operatorname{sign}(b_p - a_p^T x) > 0$ für ein $i \in I \cap \{1, \ldots, m_0\}$ und bestimmen, wie angegeben, $l \in I \cap \{1, \ldots, m_0\}$ sowie (x^-, I^-, y_{I^-}). Wegen $r_l \neq 0$, $I^- := I \setminus \{l\}$ und der eindeutigen Darstellung von a_p durch

$$a_p = \sum_{i \in I^-} r_i a_i + r_l a_l$$

ist $a_p \notin \operatorname{span}\{a_i : i \in I^-\}$. Wegen $x^- = x$, $I^- \subset I$ und $A_I x = b_I$ ist trivialerweise $A_{I^-} x = b_{I^-}$. Schließlich ist

$$c + Q x^- = A_I^T y_I = \sum_{i \in I^-} y_i a_i + y_l a_l = \sum_{i \in I^-} \Big(y_i - \frac{y_l}{r_l} r_i \Big) a_i + \frac{y_l}{r_l} a_p = A_{I^-}^T y_{I^-} + \frac{y_l}{r_l} a_p.$$

Die restlichen Aussagen gelten nach Wahl des Index l. Das Lemma ist bewiesen. □ □

Damit ist das Verfahren von Goldfarb-Idnani im Prinzip beschrieben. Wir fassen die Schritte zusammen.

- Input: Gegeben sind die Daten des obigen quadratischen Programms (P), bei welchem $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit ist.

(0) Bestimme das unrestringierte Minimum.
Berechne Q^{-1} und setze $(x, I, f) := (-Q^{-1}c, \emptyset, -\frac{1}{2}\, c^T Q^{-1} c)$ sowie $q := 0$.

(1) Bestimme eine verletzte Restriktion, falls eine solche existiert.
Falls $x \in M$, dann: STOP, x ist die Lösung von (P). Andernfalls bestimme man eine von x verletzte Restriktion $p \in \{1, \ldots, m\} \setminus I$, z. B. die am stärksten verletzte Restriktion (was genauer erklärt werden müßte). Anschließend setze man $\sigma := \operatorname{sign}(b_p - a_p^T x)$ und $\theta := 0$.

(2) Bestimme primale und duale Richtungen.
Falls $I = \emptyset$ (bzw. $q = 0$), so setze $H_I := Q^{-1}$. Andernfalls berechne

$$N_I := (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} \in \mathbb{R}^{q \times n}, \qquad H_I := Q^{-1}(I - A_I^T N_I) \in \mathbb{R}^{n \times n}.$$

Dann berechne man $z := H_I a_p$ und, falls $I \neq \emptyset$, $r_I := N_I a_p$.

(3) Bestimme primale und duale Schrittweiten.
Setze

$$t_1 := \begin{cases} \dfrac{b_p - a_p^T x}{a_p^T z} & \text{für} \quad z \neq 0, \\ \sigma \cdot \infty & \text{für} \quad z = 0. \end{cases}$$

Setze

$$t_2 := \begin{cases} \sigma \cdot \infty & \text{falls } I = \emptyset \text{ oder } \sigma r_i \leq 0 \text{ für alle } i \in I \cap \{1, \ldots, m_0\}, \\ \dfrac{y_l}{r_l} & \text{sonst,} \end{cases}$$

wobei

$$\frac{y_l}{r_l} = \begin{cases} \min\left\{ \dfrac{y_i}{r_i} : i \in I \cap \{1, \ldots, m_0\},\ r_i > 0 \right\} & \text{für} \quad \sigma = +1, \\ \max\left\{ \dfrac{y_i}{r_i} : i \in I \cap \{1, \ldots, m_0\},\ r_i < 0 \right\} & \text{für} \quad \sigma = -1. \end{cases}$$

Anschließend berechne man

$$t := \begin{cases} \min(t_1, t_2) & \text{für} \quad \sigma = +1, \\ \max(t_1, t_2) & \text{für} \quad \sigma = -1. \end{cases}$$

(4) Test auf Unzulässigkeit.
Ist $t = \sigma \cdot \infty$, dann: STOP, (P) ist nicht zulässig.

(5) Dualer Schritt.
Falls $t_1 = \sigma \cdot \infty$, so setze

$$\theta := \theta + t, \qquad y_{I \setminus \{l\}} := T_l(y_I - t r_I), \qquad I := I \setminus \{l\}, \qquad q := q - 1,$$

und gehe nach (2).

(6) Primaler und dualer Schritt.
Setze
$$x := x + tz, \qquad f := f + t(\tfrac{1}{2}\,t + \theta)\,a_p^T z, \qquad \theta := \theta + t.$$
(a) Ist $t = t_1$, so setze
$$y_{I\cup\{p\}} := \begin{pmatrix} y_I - tr_I \\ \theta \end{pmatrix}, \qquad I := I \cup \{p\}, \qquad q := q + 1$$
und gehe nach (1).
(b) Ist $t = t_2$, so setze
$$y_{I\setminus\{l\}} := T_l(y_I - tr_I), \qquad I := I \setminus \{l\}, \qquad q := q - 1$$
und gehe nach (2).

- Output: Das Verfahren bricht (bei exakter Arithmetik) nach einer endlichen Zahl von Schritten mit der Lösung von (P) ab oder es liefert die Information, dass (P) nicht zulässig ist.

Beispiel: Wie im Anschluss an das primale Verfahren von Fletcher betrachten wir die Aufgabe

$$\text{(P)} \quad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := \begin{pmatrix} -2 \\ -6 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \text{unter der Nebenbedingung} \\ \begin{pmatrix} -1 & -1 \\ 1 & -2 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq \begin{pmatrix} -2 \\ -2 \\ -3 \end{pmatrix}. \end{array} \right.$$

Zu Beginn ist
$$(x^0, I^0) = \left(\begin{pmatrix} 10 \\ 8 \end{pmatrix}, \emptyset \right), \qquad f^0 = -34$$
das Lösungspaar zum Start mit Kosten f^0. Alle drei Ungleichungsrestriktionen sind durch x^0 verletzt, da
$$\begin{pmatrix} -1 & -1 \\ 1 & -2 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} 10 \\ 8 \end{pmatrix} - \begin{pmatrix} -2 \\ -2 \\ -3 \end{pmatrix} = \begin{pmatrix} -16 \\ -4 \\ -25 \end{pmatrix}.$$
Wir wählen $p = 1$ als die am stärksten verletzte Restriktion. Da es sich um eine verletzte Ungleichungsrestriktion handelt, wird $\sigma = 1$ gesetzt, danach $\theta = 0$. Wir berechnen
$$z = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} -3 \\ -2 \end{pmatrix}.$$
Als primale und duale Schrittweiten berechnen wir
$$t_1 = 3.2000, \qquad t_2 = +\infty, \qquad t = 3.2000.$$

Anschließend wird der primale Schritt gemacht und

$$x^1 = \begin{pmatrix} 0.4000 \\ 1.6000 \end{pmatrix}, \qquad f^1 = -8.4000, \qquad \theta = 3.2000$$

berechnet, der duale Schritt liefert

$$y^1 = (3.2000), \qquad I^1 = \{1\}.$$

Im nächsten Schritt ist nur die zweite Restriktion noch verletzt, es ist also $p = 2$, ferner ist wieder $\sigma = 1$ und $\theta = 0$. Wir berechnen

$$N_{I^1} = \begin{pmatrix} -0.6000 & -0.4000 \end{pmatrix}, \qquad H_{I^1} = \begin{pmatrix} 0.2000 & -0.2000 \\ -0.2000 & 0.2000 \end{pmatrix}$$

und und die primalen bzw. dualen Richtungen

$$z = \begin{pmatrix} 0.6000 \\ -0.6000 \end{pmatrix}, \qquad r_{I^1} = (0.2000).$$

Anschließend berechnet man die Schrittweiten

$$t_1 = 0.4444, \qquad t_2 = 16.0000, \qquad t = 0.4444.$$

Im primalen und dualen Schritt wird zunächst

$$x^2 = \begin{pmatrix} 0.6667 \\ 1.3333 \end{pmatrix}, \qquad f^2 = -8.2222, \qquad \theta = 0.4444,$$

danach

$$y^2 = \begin{pmatrix} 3.1111 \\ 0.4444 \end{pmatrix}, \qquad I^2 = \{1, 2\}$$

berechnet. Da x^2 zulässig ist, hat man die Lösung gefunden. □

3.2.2 Implementation des Verfahrens

In diesem Unterabschnitt wollen wir einige Bemerkungen zur numerischen Realisierung des Verfahrens von Goldfarb-Idnani machen. Wichtige Hinweise hierzu (und ein Fortran-Programm) findet man bei M. J. D. Powell (1983)[13], einige davon werden wir im folgenden schildern.

Gegeben sei wieder die quadratische Optimierungsaufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2}\, x^T Q x \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}. \end{array} \right.$$

[13] Powell, M. J. D. (1983) "ZQPCVX A Fortran subroutine for convex quadratic programming." Report DAMTP/1983/NA17, Department for Applied Mathematics and Theoretical Physics, University of Cambridge.

Hierbei seien $a_1, \ldots, a_m \in \mathbb{R}^n$, $b = (b_i) \in \mathbb{R}^m$, $c = (c_j) \in \mathbb{R}^n$ und $Q = (q_{ij}) \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Die Matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ besitze a_i^T als i-te Zeile, $i = 1, \ldots, m$.

In dem zu beschreibenden Algorithmus wird *nicht* getestet, ob die Matrix Q symmetrisch ist (es wird nur die obere Hälfte von Q benutzt, also nur die q_{ij} mit $i \leq j$). Die positive Definitheit von Q kann dadurch geprüft werden, dass eine Cholesky-Zerlegung von Q vorgenommen wird. Stellt sich hierbei heraus, dass Q nicht „numerisch positiv definit“ ist, so kann sukzessive ein kleines Vielfaches der Einheitsmatrix zu Q addiert werden (so dass ein in der Zielfunktion eventuell gestörtes quadratisches Programm gelöst wird), bis die Cholesky-Zerlegung für die so abgeänderte Matrix durchführbar bzw. diese positiv definit ist. Hierauf wollen wir aber nicht eingehen.

Wir werden die im letzten Unterabschnitt eingeführten Bezeichnungen benutzen. Das Verfahren von Goldfarb-Idnani hat den folgenden Input und Output:

- Input: Gegeben sind die Daten des obigen quadratischen Programms (P). Das sind also:
  - Die nichtnegative ganze Zahl m ist die Anzahl der Restriktionen.
  - Die nichtnegative ganze Zahl m_0 mit $0 \leq m_0 \leq m$ gibt die Anzahl der Ungleichungsrestriktionen an.
  - Die Anzahl $n \in \mathbb{N}$ der Variablen.
  - Der Vektor $c = (c_j) \in \mathbb{R}^n$ und die symmetrische, positiv definite Matrix $Q = (q_{ij}) \in \mathbb{R}^{n \times n}$. Die Zielfunktion im quadratischen Programm (P) ist gegeben durch $f(x) := c^T x + \frac{1}{2} x^T Q x$.
  - Die Matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ und der Vektor $b = (b_i) \in \mathbb{R}^m$. Die Menge der zulässigen Lösungen des gegebenen quadratischen Programms ist

$$M := \left\{ x = (x_j) \in \mathbb{R}^n : \begin{array}{ll} \displaystyle\sum_{j=1}^{n} a_{ij} x_j \geq b_i & (i = 1, \ldots, m_0), \\ \displaystyle\sum_{j=1}^{n} a_{ij} x_j = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$

  Schließlich wird noch eine kleine Zahl $\epsilon > 0$ eingegeben. Reelle Zahlen, die betragsmäßig kleiner als ϵ sind, werden als Null angesehen.

- Output: Ausgegeben werden
  - Eine ganze Zahl $k_{\max}$, die Informationen über den Ausgang des Verfahrens gibt. Ist $k_{\max} > 0$, so ist das Verfahren erfolgreich mit einem optimalen Lösungspaar abgebrochen, $k_{\max}$ gibt in diesem Falle die Anzahl der berechneten Lösungspaare an. Ist dagegen $k_{\max} < 0$, so hat sich das gegebene quadratische Programm (P) als nicht zulässig herausgestellt.

  Die folgenden Größen sind nur dann sinnvoll besetzt, wenn das Verfahren erfolgreich war, also $k_{\max} > 0$ ist.

- In $x = (x_j) \in \mathbb{R}^n$ steht die gefundene Lösung.
- $f_{\min}$ gibt den Wert der Zielfunktion in der gefundenen Lösung x an, es ist also $f_{\min} := c^T x + \frac{1}{2}\, x^T Q x$.
- Die nichtnegative ganze Zahl q gibt die Anzahl der Elemente der Indexmenge I an, für die (x, I) ein optimales Lösungspaar ist.
- Die Elemente der (optimalen) Indexmenge I werden in $i_{\text{act}}(1), \ldots, i_{\text{act}}(q)$ ausgegeben. Hierbei ist natürlich $1 \le i_{\text{act}}(i) \le m$ für $i = 1, \ldots, q$.
- In $y_1, \ldots, y_q$ werden die Lagrange-Multiplikatoren zum optimalen Lösungspaar (x, I) gespeichert.

Bei einem erfolgreichen Ausgang des Verfahrens ist also $x \in M$ zulässig, für $i = 1, \ldots, q$ ist $y_i \ge 0$, wenn $1 \le i_{\text{act}}(i) \le m_0$, und

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \qquad i = i_{\text{act}}(1), \ldots, i_{\text{act}}(q),$$

sowie

$$c_j + \sum_{k=1}^{n} q_{jk} x_k = \sum_{i=1}^{q} a_{i_{\text{act}}(i)j}\, y_i, \qquad j = 1, \ldots, n.$$

Die Input-Daten werden durch den Algorithmus nicht überschrieben, sie werden also nicht verändert.

Nun kommen wir zu einer genaueren Beschreibung des Verfahrens. Im ersten Schritt wird $q := 0$ gesetzt. Nun ist die Cholesky-Zerlegung der symmetrischen Matrix Q zu bilden. Gesucht ist also eine untere Dreiecksmatrix L mit positiven Diagonalelementen und $Q = LL^T$ bzw. eine obere Dreiecksmatrix U mit positiven Diagonalelementen und $Q = U^T U$. Wie die Cholesky-Zerlegung berechnet wird, ist aus der numerischen Mathematik wohlbekannt, hierauf wollen wir nicht näher eingehen.

Nachdem die Cholesky-Zerlegung $Q = U^T U$ von Q erhalten wurde, berechnen wir im nächsten Schritt $Z := U^{-1}$. Insbesondere ist dann $ZZ^T = Q^{-1}$, was sich später als wichtig herausstellen wird. Im Anschluss hieran wird das unrestringierte Minimum $x := -Q^{-1}c$ sowie der zugehörige Funktionswert $f := -\frac{1}{2}\, c^T Q^{-1} c = \frac{1}{2} c^T x$ berechnet.

Ist (x, I) ein aktuelles Lösungspaar, so muss getestet werden, ob x zulässig ist, und andernfalls die am meisten verletzte Restriktion bestimmt werden. Auch hierauf wollen wir nicht näher eingehen.

Sei $I \subset \{1, \ldots, m\}$ (mit $q := \#(I)$) nun eine nichtleere Indexmenge mit der Eigenschaft, dass die Vektoren $\{a_i\}_{i \in I} \subset \mathbb{R}^n$ linear unabhängig sind, also Rang $(A_I) = q$ gilt. Insbesondere sei also $1 \le q \le n$. Die Hauptarbeit beim Verfahren von Goldfarb-Idnani besteht in der Berechnung der Matrizen

$$N_I := (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} \in \mathbb{R}^{q \times n}, \qquad H_I := Q^{-1}(I - A_I^T N_I) \in \mathbb{R}^{n \times n}$$

bzw. der Vektoren $z := H_I a_p$ und $r_I := N_I a_p$, wobei $p \notin I$ eine durch $x \in \mathbb{R}^n$ verletzte Restriktion ist (siehe Schritt (2) in der Zusammenfassung des Verfahrens von Goldfarb-Idnani am Schluss des vorigen Unterabschnitts). Von Schritt zu Schritt verändert sich

die Indexmenge I um genau ein Element, was für eine effiziente Implementation natürlich ausgenutzt werden sollte. Zu unterscheiden sind hier die Fälle, ob zur Indexmenge I das Element p hinzugefügt, oder ein Element $l \in I$ entfernt wird. Ein direktes Updaten von N_I und H_I ist möglich, in den Aufgaben 3 und 4 im letzten Abschnitt wurden hierzu Hinweise gegeben. Insgesamt erhält man einfache Update-Formeln zur Berechnung von N_I und H_I, welche zeigen, dass man diese Matrizen von Schritt zu Schritt mit höchstens $O(n^2)$ flops berechnen kann.

Wie oft in einem entsprechenden Zusammenhang ist es aber besser, geeignete Zerlegungen von N_I und H_I von Schritt zu Schritt upzudaten. Wir schildern die Vorschläge von M. J. D. Powell (1983), die im wesentlichen denen von D. Goldfarb, A. Idnani (1983) entsprechen, welche wiederum auf Vorschlägen von P. E. Gill, W. Murray (1978) basieren.

Sei $I \subset \{1, \ldots, m\}$ wieder eine (nicht notwendig nichtleere) Indexmenge mit der Eigenschaft, dass die Vektoren $\{a_i\}_{i \in I}$ linear unabhängig sind bzw. Rang $(A_I) = q$ gilt. Es existiere eine (nichtsinguläre) Matrix $Z_I \in \mathbb{R}^{n \times n}$, so daß

$$(*) \qquad Z_I Z_I^T = Q^{-1}, \qquad Z_I^T A_I^T = \begin{pmatrix} R_I \\ 0 \end{pmatrix} \begin{matrix} \} \, q \\ \} \, n-q \end{matrix}$$

mit einer oberen Dreiecksmatrix $R_I \in \mathbb{R}^{q \times q}$, deren Diagonalelemente wegen der Rangvoraussetzung an A_I nicht verschwinden, die also nichtsingulär ist. Man beachte, dass diese Annahme für $I = \emptyset$ trivialerweise erfüllt ist, da in diesem Falle $Z_\emptyset := Z$ mit der oben berechneten oberen Dreiecksmatrix Z, für welche $ZZ^T = Q^{-1}$ gilt, gesetzt werden kann. Es wird sich herausstellen, dass in Z_I und R_I alle Informationen zur Berechnung der Matrizen N_I und H_I sowie der Vektoren $z := H_I a_p$ und $r_I := N_I a_p$ enthalten sind. Um dies einzusehen, denke man sich Z_I zerlegt in der Form

$$Z_I = (\underbrace{Z_I^{(1)}}_{q} \quad \underbrace{Z_I^{(2)}}_{n-q}),$$

d. h. in $Z_I^{(1)} \in \mathbb{R}^{n \times q}$ stehen die ersten q Spalten von Z_I, in $Z_I^{(2)} \in \mathbb{R}^{n \times (n-q)}$ die restlichen $n - q$ Spalten. Dann ist nach einfacher Rechnung

$$N_I = R_I^{-1} Z_I^{(1)T}, \qquad H_I = Z_I^{(2)} Z_I^{(2)T}.$$

Zur Berechnung von $z := H_I a_p$ und $r_I := N_I a_p$ ist es daher zweckmäßig, zunächst

$$d_I := Z_I^T a_p = \begin{pmatrix} d_I^{(1)} \\ d_I^{(2)} \end{pmatrix} \begin{matrix} \} \, q \\ \} \, n-q \end{matrix},$$

anschließend $z := Z_I^{(2)} d_I^{(2)}$ zu berechnen und r_I aus $R_I r_I = d_I^{(1)}$ durch Rückwärtseinsetzen zu erhalten.

Entscheidend ist nun, wie man Z_{I^+} und R_{I^+} mit der obigen Eigenschaft $(*)$ bestimmt, wenn I^+ dadurch aus I hervorgeht, dass zu I ein Element $p \notin I$ hinzugefügt

wird (dies geschieht nur dann, wenn $z := H_I a_p \neq 0$, also a_p von $\{a_i\}_{i\in I}$ linear unabhängig ist), oder ein Element $l \in I$ aus I entfernt wird. Diese beiden Fälle werden nun getrennt untersucht. Gemeinsam ist aber beiden Fällen, dass der Ansatz

$$Z_{I^+} := Z_I \Omega_I^T$$

mit einer orthogonalen Matrix $\Omega_I \in \mathbb{R}^{n\times n}$ gemacht wird. Wegen

$$Z_{I^+} Z_{I^+}^T = Z_I \underbrace{\Omega_I^T \Omega_I}_{=I} Z_I^T = Z_I Z_I^T = Q^{-1}$$

ist dann die erste Bedingung in $(*)$ automatisch erfüllt.

In beiden Fällen spielen Givens-Rotationen, also spezielle orthogonale Matrizen, eine besondere Rolle. Wir benutzen die Funktion "rot", die zu (α, β) ein Tripel (c, s, γ) mit $c^2 + s^2 = 1$ und

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}$$

bestimmt. Für $i < k$ bezeichnen wir eine Givens-Rotation, die nur in den Positionen (i,i), (i,k), (k,i) und (k,k) von der Einheitsmatrix abweicht und dort mit c, s, $-s$ und c mit $c^2 + s^2 = 1$ besetzt ist, mit G_{ik}.

Wir wollen uns auf den etwas einfacheren Fall beschränken, dass nämlich $I^+ := I \cup \{p\}$ mit einem $p \notin I$. Der Ansatz $Z_{I\cup\{p\}} := Z_I \Omega_I^T$ mit

$$\Omega_I := \begin{pmatrix} I_q & 0 \\ 0 & \Omega_I^{(2)} \end{pmatrix} \begin{matrix} \} \, q \\ \} \, n-q \end{matrix}$$

(I_q sei die Einheitsmatrix in $\mathbb{R}^{q\times q}$) und der orthogonalen Matrix $\Omega_I^{(2)} \in \mathbb{R}^{(n-q)\times(n-q)}$ liefert

$$Z_{I\cup\{p\}}^T A_{I\cup\{p\}}^T = \Omega_I Z_I^T \begin{pmatrix} A_I^T & a_p \end{pmatrix} = \Omega_I \begin{pmatrix} R_I & d_I^{(1)} \\ 0 & d_I^{(2)} \end{pmatrix} = \begin{pmatrix} R_I & d_I^{(1)} \\ 0 & \Omega_I^{(2)} d_I^{(2)} \end{pmatrix},$$

wobei

$$\begin{pmatrix} d_I^{(1)} \\ d_I^{(2)} \end{pmatrix} := \begin{pmatrix} Z_I^{(1)T} a_p \\ Z_I^{(2)T} a_p \end{pmatrix}, \qquad Z_I = \begin{pmatrix} Z_I^{(1)} & Z_I^{(2)} \end{pmatrix}.$$

Es kommt also darauf an, die orthogonale Matrix $\Omega_I^{(2)}$ so zu bestimmen, daß $\Omega_I^{(2)} d_I^{(2)}$ ein Vielfaches des ersten Einheitsvektors im $\mathbb{R}^{n-q}$ ist. Hierzu multipliziert man $d_I^{(2)}$ sukzessive mit $n-q-1$ Givens-Rotationen

$$G_{n-q-1,n-q}, \ldots, G_{23}, G_{12} \in \mathbb{R}^{(n-q)\times(n-q)}.$$

Die erste, nämlich $G_{n-q-1,n-q}$, annulliert die letzte Komponente von $d_I^{(2)}$, die nächste macht die vorletzte Komponente von $G_{n-q-1,n-q} d_I^{(2)}$ zu Null, bis schließlich G_{12} die zweite Komponente von $G_{23} \cdots G_{n-q-1,n-q} d_I^{(2)}$ zum Verschwinden bringt. Einmal erzeugte Nullen bleiben hierbei offenbar erhalten. Die gesuchte orthogonale Matrix $\Omega_I^{(2)}$ hat daher die Form

$$\Omega_I^{(2)} := G_{12} G_{23} \cdots G_{n-q-1,n-q}.$$

Damit ist

$$R_{I\cup\{p\}} := \left.\begin{pmatrix} R_I & d_I^{(1)} \\ 0^T & \delta_I \end{pmatrix}\right\} \; q+1$$

mit der ersten Komponente δ_I von $\Omega_I^{(2)} d_I^{(2)}$. Nun ist es keineswegs nötig, sich die Givens-Rotationen $G_{n-q-1,n-q}, \ldots, G_{12}$ zu merken oder gar $\Omega_I^{(2)}$ zu berechnen. Denn wegen

$$Z_{I\cup\{p\}} = Z_I \Omega_I^T = \begin{pmatrix} Z_I^{(1)} & Z_I^{(2)} \end{pmatrix} \begin{pmatrix} I_q & 0 \\ 0 & \Omega_I^{(2)T} \end{pmatrix} = \begin{pmatrix} Z_I^{(1)} & Z_I^{(2)} \Omega_I^{(2)T} \end{pmatrix}$$

und

$$Z_I^{(2)} \Omega_I^{(2)T} = Z_I^{(2)} G_{n-q-1,n-q}^T \cdots G_{23}^T G_{12}^T$$

genügt es, $Z_I^{(2)}$ sukzessive von rechts mit $G_{n-q-1,n-q}^T, \ldots, G_{12}^T$ zu multiplizieren. Diese Multiplikationen können sozusagen parallel zur sukzessiven Multiplikation von $d_I^{(2)}$ mit $G_{n-q-1,n-q}, \ldots, G_{12}$ erfolgen. Sobald die beiden Multiplikationen durchgeführt sind, kann man die entsprechende Givens-Rotation vergessen. In Pseudocode könnte dies folgendermaßen aussehen, wobei die Funktion "rot" benutzt wird.

- Input:
  - Eine Matrix $Z \in \mathbb{R}^{n\times n}$.
  - Eine Indexmenge $I = \{i_{\text{act}}(1), \ldots, i_{\text{act}}(q)\}$ mit $0 \le q \le n-1$.
  - Ein $p \in \{1, \ldots, m\} \setminus I$.
  - Eine obere Dreiecksmatrix $R \in \mathbb{R}^{q\times q}$.
  - Ein Vektor $d \in \mathbb{R}^n$.

  Hierbei ist

$$ZZ^T = Q^{-1}, \qquad Z^T A_I^T = \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad d = Z^T a_p.$$

- $q := q+1, \quad i_{\text{act}}(q) := p$

  Für $j = n-1, \ldots, q$:

  $(c, s, d_j) := \text{rot}(d_j, d_{j+1})$

  Für $i = 1, \ldots, n$:

  $\text{temp} := c z_{ij} + s z_{i,j+1}, \quad z_{i,j+1} := -s z_{ij} + c z_{i,j+1}, \quad z_{ij} := \text{temp}$

  Für $i = 1, \ldots, q$:

  $r_{i,q} := d_i$

- Output:
  - Eine Matrix $Z \in \mathbb{R}^{n\times n}$.

– Die Anzahl der Elemente der gegebenen Indexmenge I hat sich um eines erhöht, es ist also $q := q + 1$ und $i_{\text{act}}(q) := p$ gesetzt worden. Die neue Indexmenge I ist durch $I = \{i_{\text{act}}(1), \ldots, i_{\text{act}}(q)\}$ gegeben.
– Eine obere Dreiecksmatrix $R \in \mathbb{R}^{q\times q}$.

Nach Abschluss ist

$$ZZ^T = Q^{-1}, \qquad Z^T A_I^T = \begin{pmatrix} R \\ 0 \end{pmatrix}.$$

Damit haben wir einige Details einer möglichen Implementation des Verfahrens von Goldfarb-Idnani besprochen. Weitere Feinheiten sind bei M. J. D. Powell (1983) angegeben. Trotzdem sollte es mit den hier angegebenen Hinweisen möglich sein, ein einigermaßen effizientes Programm zu schreiben. Hierbei sollte man aber insbesondere für den Fall, dass die Matrix Q kleine Eigenwerte besitzt, in einigen Schritten die Möglichkeit einer iterativen Verbesserung eines Lösungspaares (x, I) und eines zugehörigen Lagrange-Vektors berücksichtigen. Hinweise hierzu findet man im Anschluss in Aufgabe 3.

3.2.3 Aufgaben, Ergänzungen

1. Gegeben sei das quadratische Programm

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2}\, x^T Q x \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}. \end{array} \right.$$

   Hierbei seien $a_1, \ldots, a_m \in \mathbb{R}^n \setminus \{0\}$, $b = (b_i) \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Die Matrix $A \in \mathbb{R}^{m\times n}$ besitze a_i^T als i-te Zeile, $i = 1, \ldots, m$. Sei (P) zulässig, $x^* \in M$ sei die eindeutige Lösung von (P). Man zeige, dass eine Indexmenge $I^* \subset \{1, \ldots, m\}$ existiert derart, daß (x^*, I^*) ein Lösungspaar für (P) ist.

2. Gegeben sei wieder das quadratische Programm (P) aus Aufgabe 1. Sei (x, I) ein Lösungspaar, $p \in \{1, \ldots, m\} \setminus I$ eine durch x verletzte Restriktion und $M_{I\cup\{p\}} \neq \emptyset$. Ferner sei $a_p \notin \operatorname{span}\{a_i : i \in I\}$, so dass die Vektoren $\{a_i\}_{i\in I\cup\{p\}}$ linear unabhängig sind. Sei x^+ die (eindeutige) Lösung von $(\text{P}_{I\cup\{p\}})$, $\overline{I} := \{i \in I : a_i^T x^+ = b_i\}$ und $I^+ := \overline{I} \cup \{p\}$. Man zeige, daß (x^+, I^+) ein Lösungspaar mit $f(x^+) > f(x)$ ist.

3. Gegeben sei die symmetrische, positiv definite Matrix $Q \in \mathbb{R}^{n\times n}$ und der Vektor $c \in \mathbb{R}^n$. Sei $I \subset \{1, \ldots, m\}$ (mit $q := \#(I)$) eine nichtleere Indexmenge mit der Eigenschaft, dass die Vektoren $\{a_i\}_{i\in I} \subset \mathbb{R}^n$ linear unabhängig sind bzw. Rang $(A_I) = q$ gilt. Es existiere eine Matrix $Z_I \in \mathbb{R}^{n\times n}$, so daß

$$(*) \qquad Z_I Z_I^T = Q^{-1}, \qquad Z_I^T A_I^T = \begin{pmatrix} R_I \\ 0 \end{pmatrix} \begin{matrix} \} \, q \\ \} \, n-q \end{matrix}$$

   mit einer oberen Dreiecksmatrix $R_I \in \mathbb{R}^{q\times q}$. Sei

$$Z_I = (\underbrace{Z_I^{(1)}}_{q} \quad \underbrace{Z_I^{(2)}}_{n-q}).$$

Bei gegebenem $b_I \in \mathbb{R}^q$ berechne man $\hat{x},\ x \in \mathbb{R}^n$ aus

$$\hat{x} := Z_I^{(1)} R_I^{-T} b_I, \qquad x := \begin{cases} \hat{x} - Z_I^{(2)} Z_I^{(2)T} c & \text{für} \quad q < n, \\ \hat{x} & \text{für} \quad q = n. \end{cases}$$

Man zeige, dass x eine Lösung des durch lineare Gleichungen restringierten quadratischen Programms

$$\text{Minimiere} \quad f(x) := c^T x + \tfrac{1}{2}\, x^T Q x \quad \text{unter der Nebenbedingung} \quad A_I x = b_I$$

mit zugehörigem Lagrange-Vektor $y_I := R_I^{-1} Z_I^{(1)T}(c + Q\hat{x})$ ist.

4. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Mit einer Spalten-Version des Cholesky-Verfahrens soll getestet werden, ob A positiv definit ist. Dies könnte folgendermaßen aussehen:

   - Gegeben sei die symmetrische Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$.
   - Für $k = 1, \ldots, n$:
     Berechne $\tilde{a}_{kk} := a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2$.
     Falls $\tilde{a}_{kk} \le 0$, dann:
       STOP: A nicht positiv definit.
     Andernfalls:
       Berechne $l_{kk} := (\tilde{a}_{kk})^{1/2}$.
       Für $i = k+1, \ldots, n$:
         Berechne $l_{ik} := (a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj})/l_{kk}$.

   Angenommen, das Verfahren breche im k-ten Schritt wegen $\tilde{a}_{kk} \le 0$ ab. Dann ist

$$A = \begin{pmatrix} L_1 & 0 \\ L_2 & I \end{pmatrix} \begin{pmatrix} L_1^T & L_2^T \\ 0 & \tilde{A} \end{pmatrix}$$

   mit

$$L_1 := \begin{pmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{k-1,1} & \cdots & l_{k-1,k-1} \end{pmatrix}, \qquad L_2 := \begin{pmatrix} l_{k1} & \cdots & l_{k,k-1} \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{n,k-1} \end{pmatrix}$$

   und

$$\tilde{A} := \begin{pmatrix} \tilde{a}_{kk} & \cdots & \tilde{a}_{kn} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{nk} & \cdots & \tilde{a}_{nn} \end{pmatrix} := \begin{pmatrix} a_{kk} & \cdots & a_{kn} \\ \vdots & \ddots & \vdots \\ a_{nk} & \cdots & a_{nn} \end{pmatrix} - L_2 L_2^T.$$

   Schließlich sei $x_1 \in \mathbb{R}^{k-1}$ die eindeutige Lösung von $L_1^T x_1 = -L_2^T e_1$, wobei $e_1 \in \mathbb{R}^{n-k+1}$ den ersten Einheitsvektor bezeichnet. Man zeige, dass

$$\lambda_{\min}(A) \le \frac{\tilde{a}_{kk}}{\|x_1\|_2^2 + 1}.$$

   Man muss also mindestens $-\tilde{a}_{kk}/(\|x_1\|_2^2+1)$ zu den Diagonalelementen von A addieren, um eine positiv semidefinite Matrix zu erhalten.

3.3 Quadratische Programme mit Box-Constraints

Unser Ziel in diesem Abschnitt besteht darin, einige Ideen eines (nicht einfach zu lesenden) Aufsatzes von T. F. Coleman, Y. Li (1996)[14] darzustellen. Dieses Verfahren ist in MATLAB implementiert worden. Eine gut lesbare Darstellung der Ergebnisse von Coleman-Li würde allerdings den Rahmen dieser Vorlesung sprengen. Kurz werden wir auch noch auf quadratische Programme mit Vorzeichenbeschränkungen eingehen. Gemeinsam ist den beiden Ansätzen, dass das gegebene Problem in ein nichtlineares Gleichungssystem bzw. eine unrestringierte Optimierungsaufgabe umformuliert wird, auf welche anschließend Varianten des Newton-Verfahrens angewandt werden.

3.3.1 Problemstellung, Optimalitätsbedingungen

Gegeben sei die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := c^T x + \tfrac{1}{2} x^T Q x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : l \le x \le u\}.$$

Hierbei sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und i. Allg. indefinit, $l \in \{\mathbb{R} \cap \{-\infty\}\}^n$ und $u \in \{\mathbb{R} \cap \{+\infty\}\}^n$ mit $l < u$. Gewisse Variable sind also nach unten und/oder oben beschränkt, man spricht von Box-Constraints. Bei Coleman-Li (1996) findet man man einige Hinweise auf neuere Arbeiten über quadratische Programme mit Box-Constraints. Wir wollen für (P) die notwendigen Optimalitätsbedingungen erster und zweiter Ordnung, sowie die hinreichenden Optimalitätsbedingungen zweiter Ordnung aufstellen, siehe auch Coleman-Li (1994)[15]. Die notwendige Optimalitätsbedingung erster Ordnung für eine lokale Lösung $x^* \in M$ von (P) besagt, dass

$$(*) \qquad \nabla f(x^*)_i \begin{cases} = 0, & \text{falls} \quad l_i < x_i^* < u_i, \\ \le 0, & \text{falls} \quad x_i^* = u_i, \\ \ge 0, & \text{falls} \quad x_i^* = l_i, \end{cases} \qquad i = 1, \ldots, n.$$

Dies erhält man entweder durch die Anwendung des Satzes von Kuhn-Tucker oder, einfacher, aus der notwendigen Bedingung, dass $\nabla f(x^*)^T (x - x^*) \ge 0$ für alle $x \in M$. Wir wollen uns überlegen, dass diese Aussagen äquivalent dazu sind, dass x^* einem gewissen nichtlinearen Gleichungssystem genügt. Hierzu definieren wir bei gegebenem $x \in \mathbb{R}^n$ den Vektor $v(x) \in \mathbb{R}^n$ durch:

- Ist $\nabla f(x)_i < 0$ und $u_i < \infty$, dann sei $v(x)_i := x_i - u_i$.
- Ist $\nabla f(x)_i \ge 0$ und $l_i > -\infty$, dann sei $v(x)_i := x_i - l_i$.
- Ist $\nabla f(x)_i < 0$ und $u_i = \infty$, dann sei $v(x)_i := -1$.
- Ist $\nabla f(x)_i \ge 0$ und $l_i = -\infty$, dann sei $v(x)_i := 1$

[14] T. F. Coleman, Y. Li (1996) "A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables". SIAM J. Optim. 6, 1040–1068.

[15] T. F. Coleman, Y. Li (1994) "On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds". Mathematical Programming 67, 189–224.

und anschließend die Diagonalmatrix $D(x) \in \mathbb{R}^{n\times n}$ durch

$$D(x) := \operatorname{diag}\left(|v(x)_1|^{1/2}, \ldots, |v(x)_n|^{1/2}\right).$$

Dann sind die obigen notwendigen Optimalitätsbedingungen erster Ordnung $(*)$ äquivalent zu

$$(**) \qquad D(x^*)^2 \nabla f(x^*) = 0$$

bzw.

$$|v(x^*)_i| \nabla f(x^*)_i = 0, \qquad i = 1, \ldots, n.$$

Um dies zu beweisen, sei $x^* \in M$ vorgegeben. Es gelte $(*)$ und $i \in \{1, \ldots, n\}$ sei fest vorgeben. Wir wollen zeigen, dass $|v(x^*)_i| \nabla f(x^*)_i = 0$. Dies ist für $\nabla f(x^*)_i = 0$ trivialerweise richtig, so dass wir $\nabla f(x^*)_i \neq 0$ annehmen können. Ist $\nabla f(x^*)_i < 0$, so ist $x_i^* = u_i < \infty$ und daher $v(x^*)_i = x_i^* - u_i = 0$. Ist dagegen $\nabla f(x^*)_i > 0$, so ist $x_i^* = l_i$ und daher $v(x^*)_i = x_i^* - l_i = 0$. Nun nehmen wir umgekehrt an, es gelte $(**)$ und $i \in \{1, \ldots, n\}$ sei fest vorgegeben. Ist $l_i < x_i^* < u_i$, so ist $\nabla f(x^*)_i = 0$. Denn andernfalls folgt aus $(**)$, dass $v(x^*)_i = 0$, was aber wegen $l_i < x_i^* < u_i$ nicht möglich ist. Ist $x_i^* = u_i < \infty$ und $\nabla f(x^*)_i \neq 0$, so ist wieder $v(x^*)_i = 0$. Wäre $\nabla f(x^*)_i > 0$, so müsste $l_i = -\infty$ sein (andernfalls wäre $v(x^*)_i = x_i^* - l_i = 0$, also $l_i = x_i^* = u_i$, was wegen $l < u$ nicht möglich ist), dann aber $v(x^*)_i = 1$, ein Widerspruch. Also folgt aus $x_i^* = u_i$, dass $\nabla f(x^*)_i \leq 0$. Entsprechend ist $\nabla f(x^*)_i \geq 0$, falls $x_i^* = l_i > -\infty$. Damit ist nachgewiesen, dass $(*)$ für ein $x^* \in M$ genau dann gilt, wenn x^* eine Lösung von $D(x)^2 \nabla f(x) = 0$ ist.

Nun kommen wir zu den notwendigen Optimalitätsbedingungen zweiter Ordnung. Wieder sei also x^* eine lokale Lösung von (P) (nach wie vor nutzen wir nicht aus, dass f eine quadratische Funktion, sondern setzen nur voraus, dass f zweimal stetig differenzierbar ist). Wir definieren die Indexmenge der in x^* freien Restriktionen durch

$$F^* := \{i \in \{1, \ldots, n\} : l_i < x_i^* < u_i\}.$$

Dies ist also genau das Komplement der Indexmenge aller in x^* aktiven Restriktionen. Nun sei $p \in \mathbb{R}^n$ beliebig mit $p_i = 0$ für alle $i \in \{1, \ldots, n\} \setminus F^*$. Dann ist $x^* + tp \in M$ für alle hinreichend kleinen $|t|$. Da $\phi(t) := f(x^* + tp)$ bei $t = 0$ ein lokales Minimum annimmt, ist nicht nur $\phi'(0) = 0$ bzw. $\nabla f(x^*)^T p = 0$ (das liefert weniger als wir schon wissen, nämlich $\nabla f(x^*)_i = 0$, $i \in F^*$), sondern auch $\phi''(0) = p^T \nabla^2 f(x^*) p \geq 0$. Dies bedeutet, dass

$$H_{F^*} := \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x^*)\right)_{(i,j)\in F^*\times F^*}$$

positiv semidefinit ist. Die Matrix H_{F^*} ist eine Submatrix von $\nabla^2 f(x^*)$, die durch Streichen der Zeilen und Spalten zu in x^* aktiven Indizes entsteht. Die positive Semidefinitheit von H_{F^*} und $(*)$ (bzw. $(**)$) sind also die notwendigen Optimalitätsbedingungen zweiter Ordnung.

Die hinreichenden Optimalitätsbedingungen zweiter Ordnung formulieren wir nur für nichtentartete Punkte $x \in M$. Hierbei heißt ein $x \in M$ *nichtentartet*, wenn

$$\nabla f(x)_i = 0 \Longrightarrow l_i < x_i < u_i, \qquad i = 1, \ldots, n.$$

Genügt ein nichtentartetes $x^* \in M$ den notwendigen Bedingungen erster Ordung $(*)$ (bzw. $(**)$) und ist die Submatrix H_{F^*} von $\nabla^2 f(x^*)$ positiv definit, so ist x^* eine isolierte lokale Lösung von (P). Um dies einzusehen beachten wir zunächst, dass (Nichtentartung und notwendige Bedingung erster Ordnung)

$$\nabla f(x^*)_i \begin{cases} = 0, & \text{falls} \quad l_i < x_i^* < u_i, \\ < 0, & \text{falls} \quad x_i^* = u_i, \\ > 0, & \text{falls} \quad x_i^* = l_i, \end{cases} \qquad i = 1, \ldots, n.$$

Wäre x^* keine isolierte lokale Lösung von (P), so existiert eine Folge $\{x_k\} \subset M \setminus \{x^*\}$ mit $f(x_k) \leq f(x^*)$ und $x_k \to x^*$. Es ist

$$x_k = x^* + t_k p_k \qquad \text{mit} \qquad t_k := \|x_k - x^*\|, \qquad p_k := \frac{x_k - x^*}{\|x_k - x^*\|}.$$

Da man aus $\{p_k\}$ eine konvergente Teilfolge auswählen kann, können wir o. B. d. A. annehmen, dass $p_k \to p$ mit $p \neq 0$ und (siehe Beweis der allgemeinen hinreichenden Bedingungen zweiter Ordnung)

$$\nabla f(x^*)^T p \leq 0, \qquad p_i \begin{cases} \leq 0, & \text{falls} \quad x_i^* = u_i, \\ \geq 0, & \text{falls} \quad x_i^* = l_i, \end{cases} \qquad i = 1, \ldots, n.$$

Hieraus folgt offenbar $p_i = 0$ für alle $i \in \{1, \ldots, n\} \setminus F^*$ und folglich $\nabla f(x^*)^T p = 0$ (hierbei sollte man p_i natürlich nicht mit dem i-ten Folgenglied von $\{p_k\}$ verwechseln). Da außerdem $\{x_k\} \subset M$, ist für alle k auch

$$(p_k)_i \begin{cases} \leq 0, & \text{falls} \quad x_i^* = u_i, \\ \geq 0, & \text{falls} \quad x_i^* = l_i, \end{cases} \qquad i = 1, \ldots, n.$$

Daher ist

$$\nabla f(x^*)^T p_k = \sum_{i: l_i < x_i^* < u_i} \underbrace{\nabla f(x^*)_i}_{=0} (p_k)_i + \sum_{i: x_i^* = u_i} \underbrace{\nabla f(x^*)_i}_{<0} \underbrace{(p_k)_i}_{\leq 0} + \sum_{i: x_i^* = l_i} \underbrace{\nabla f(x^*)_i}_{>0} \underbrace{(p_k)_i}_{\geq 0} \geq 0$$

für alle k. Folglich ist

$$\begin{aligned} 0 &\geq f(x_k) - f(x^*) \\ &= f(x^* + t_k p_k) - f(x^*) \\ &= \underbrace{t_k}_{>0} \underbrace{\nabla f(x^*)^T p_k}_{\geq 0} + \frac{1}{2} t_k^2 p_k^T \nabla^2 f(\tilde{x}_k) p_k \\ &\geq \frac{1}{2} t_k^2 p_k^T \nabla^2 f(\tilde{x}_k) p_k, \end{aligned}$$

wobei $\tilde{x}_k$ auf der Verbindungsstrecke zwischen x_k und x^* liegt, so dass auch $\tilde{x}_k \to x^*$. Dann ist $p_k^T \nabla^2 f(\tilde{x}_k) p_k \leq 0$, nach dem Grenzübergang $k \to \infty$ folgt $p^T \nabla^2 f(x^*) p \leq 0$. Wegen $p \neq 0$ und $p_i = 0$ für alle $i \in \{1, \ldots, n\} \setminus F^*$ ist dies ein Widerspruch zur vorausgesetzten positiven Definitheit von H_{F^*}.

3.3.2 Motivation des Verfahrens, lokale Konvergenz

Wir gehen jetzt aus von dem quadratischen Programm (P) mit Box-Constraints. Wie wir im letzten Unterabschnitt gesehen haben, genügt ein $x \in M$ genau dann den notwendigen Optimalitätsbedingungen erster Ordnung, wenn $D(x)^2 \nabla f(x) = 0$. Hierbei ist natürlich $\nabla f(x) = c + Qx$, ferner ist

$$D(x) = \text{diag}\,(|v(x)_1|^{1/2}, \ldots, |v(x)_n|^{1/2}) = \text{diag}\,(|v(x)|^{1/2}),$$

wobei naheliegende Bezeichnungen benutzt wurden und $v(x) \in \mathbb{R}^n$ im vorigen Abschnitt definiert wurde:

- Ist $\nabla f(x)_i < 0$ und $u_i < \infty$, dann sei $v(x)_i := x_i - u_i$.
- Ist $\nabla f(x)_i \geq 0$ und $l_i > -\infty$, dann sei $v(x)_i := x_i - l_i$.
- Ist $\nabla f(x)_i < 0$ und $u_i = \infty$, dann sei $v(x)_i := -1$.
- Ist $\nabla f(x)_i \geq 0$ und $l_i = -\infty$, dann sei $v(x)_i := 1$.

Zur Abkürzung definieren wir $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ durch $F(x) := D(x)^2 \nabla f(x)$. Es liegt nahe, auf das nichtlineare Gleichungssystem $F(x) = 0$ das Newton-Verfahren anzuwenden. Eine geeignete Modifikation ist nötig, da die Abbildung F zwar stetig, aber nicht überall differenzierbar ist. Dies ist in einem $x \in \mathbb{R}^n$ genau dann der Fall, wenn $v(x)_i = 0$ für ein $i \in \{1, \ldots, n\}$. Da das Verfahren von Coleman-Li, wie wir sehen werden, nur innere Punkte von M erzeugt, ist in solchen Punkten die Differenzierbarkeit kein Problem. Sei nun $x \in \text{int}\,(M)$. Wir wollen die Funktionalmatrix von F berechnen, um die Newton-Richtung ageben zu können. Zunächst berechnen wir aber die Funktionalmatrix $J^v(x)$ der Abbildung $x \mapsto |v(x)|$, wobei wir ausnutzen, dass keine Komponente von $v(x)$ (wegen $x \in \text{int}\,(M)$) verschwindet. Offenbar ist

$$J^v(x) = \text{diag}\,(J^v_{11}(x), \ldots, J^v_{nn}(x))$$

mit

$$J^v_{ii}(x) := \begin{cases} -1, & \text{falls} \quad v(x)_i = x_i - u_i, \\ 1, & \text{falls} \quad v(x)_i = x_i - l_i, \\ 0, & \text{sonst.} \end{cases}$$

Dann ist aber

$$\begin{aligned} F(x+h) - F(x) &= \text{diag}\,(|v(x+h)|)\nabla f(x+h) - \text{diag}\,(|v(x)|)\nabla f(x) \\ &= \text{diag}\,(|v(x)| + J^v(x)h)(\nabla f(x) + Qh) - \text{diag}\,(|v(x)|)\nabla f(x) \\ &\qquad + o(h) \\ &= \text{diag}\,(|v(x)|)Qh + \text{diag}\,(J^v(x)h)\nabla f(x) + o(h) \\ &= [\text{diag}\,(|v(x)|)Q + J(x)\text{diag}\,(|\nabla f(x)|)]h + o(h), \end{aligned}$$

wobei

$$J(x) := \text{diag}\,(|J^v_{11}(x)|, \ldots, |J^v_{nn}(x)|).$$

Also ist

$$F'(x) = D(x)^2 Q + J(x)\text{diag}\,(|\nabla f(x)|).$$

Daher ist die Newton-Richtung gegeben durch

$$\begin{aligned} p^N(x) &:= -[D(x)^2 Q + J(x)\text{diag}\,(|\nabla f(x)|)]^{-1} D(x)^2 \nabla f(x) \\ &= -D(x)[D(x)QD(x) + J(x)\text{diag}\,(|\nabla f(x)|)]^{-1} D(x)\nabla f(x). \end{aligned}$$

Daher liegt es nahe, zunächst

$$M(x) := D(x)QD(x) + J(x)\text{diag}\,(|\nabla f(x)|)$$

zu bilden, dann $\overline{p}^N(x)$ aus

$$M(x)\overline{p}^N(x) = -D(x)\nabla f(x)$$

zu berechnen und schließlich die Newton-Richtung aus

$$p^N(x) := D(x)\overline{p}^N(x)$$

zu erhalten. Wir wollen (bei der folgenden Behauptung wird bei Coleman-Ki (1996) auf Coleman-Li (1994) verwiesen, wir konnten dieses Ergebnis dort aber nicht finden) uns überlegen:

- In dem nichtentarteten $x^* \in M$ seien die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt. Dann gibt es eine Umgebung U^* von x^* derart, dass $M(x)$ für alle $x \in U^* \cap M$ (symmetrisch und) positiv definit ist.

Die Matrix $M(x)$ ist natürlich trivialerweise symmetrisch. Wir überlegen uns, dass $M(x^*)$ positiv definit ist. Mit F^* bezeichnen wir wieder die Indexmenge der freien Restriktionen. Wir beachten, dass $v(x^*)_i = 0$ genau dann, wenn $i \notin F^*$. Für ein beliebiges $p \in \mathbb{R}^n \setminus \{0\}$ ist

$$\begin{aligned} p^T D(x^*)QD(x^*)p &= \sum_{i,j=1}^{n} q_{ij} v(x^*)_i p_i v(x^*)_j p_j \\ &= \sum_{\substack{i,j=1 \\ i,j\in F^*}}^{n} q_{ij} v(x^*)_i p_i v(x^*)_j p_j \\ &\geq 0, \end{aligned}$$

wobei das Gleichheitszeichen wegen der positiven Definitheit von $(q_{ij})_{(i,j)\in F^*\times F^*}$ genau dann eintritt, wenn $v(x^*)_i p_i = 0$ bzw. $p_i = 0$ für alle $i \in F^*$. Angenommen, dies ist der Fall. Dann ist

$$p^T J(x^*)\text{diag}\,(|\nabla f(x^*)|)p = \sum_{i\notin F^*} p_i^2 \underbrace{|\nabla f(x^*)_i|}_{>0} > 0.$$

Insgesamt ist gezeigt, dass $M(x^*) = D(x^*)QD(x^*) + J(x^*)\text{diag}\,(|\nabla f(x^*)|)$ positiv definit. Aus Stetigkeitsgründen ist $M(\cdot)$ auch noch in einer Umgebung von x^* positiv definit.

Um die strikte Zulässigkeit zu erhalten, wird man natürlich eine Schrittweite einführen müssen. Als neue Näherung hat man also $x_+ := x + tp^N(x)$, wobei die Schrittweite $t > 0$ u. a. dafür zu sorgen hat, dass mit x auch x_+ strikt zulässig ist, also in $\operatorname{int}(M)$ liegt. Außerdem sollte $|1-t| = O(\|x-x^*\|)$ gelten, damit lokal im wesentlichen das ungedämpfte Verfahren verwandt wird. Bei Coleman-Li (1994, S. 213) wird näher auf die lokale (quadratische) Konvergenz eingegangen.

3.3.3 Vorzeichenbeschränkte quadratische Programme

Man betrachte[16] das quadratische Programm

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := c^T x + \dfrac{1}{2} x^T Q x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : x_j \geq 0 \ \ (j = 1, \ldots, n_0)\}. \end{cases}$$

Hierbei sei $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit. Probleme dieser Art treten entweder direkt, oder auch als duales Programm zu einem sogenannten Least-Distance-problem auf, wie wir im nächsten Beispiel zeigen wollen.

Beispiel: Gegeben sei das quadratische Programm

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := \dfrac{1}{2} \|x - z\|^2 \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}. \end{cases}$$

Hierbei seien $a_1, \ldots, a_m \in \mathbb{R}^n$, $b_1, \ldots, b_m \in \mathbb{R}$, $z \in \mathbb{R}^n$ gegeben. Zur Abkürzung setzen wir wieder

$$A := \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix} \in \mathbb{R}^{m \times n}, \qquad b := \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m.$$

Das Problem (P) wird in der Literatur auch ein Least-Distance-Problem genannt, da es darin besteht, einen gegebenen Punkt z auf ein Polyeder M zu projizieren, also insbesondere den kürzesten Abstand von z zu M zu berechnen. Das zu (P) duale Programm ist

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \phi(y) := b^T y - \dfrac{1}{2}(A^T y + z)^T (A^T y + z) \quad \text{auf} \\ N := \{y \in \mathbb{R}^m : y_i \geq 0 \ \ (i = 1, \ldots, m_0)\}. \end{cases}$$

[16] Wir schildern hier einige Aspekte der Arbeit

W. Li, J. Swetits (1993) "A Newton method for convex regression, data smoothing, and quadratic programming with bounded constraints". SIAM J. Optimization 3, 466–488,

wobei wir aber keineswegs genau diesen Autoren folgen. Siehe auch

W. Li (1996) "Differentiable piecewise quadratic exact penalty functions for quadratic functions with simple bound constraints". SIAM J. Optimization 6, 299–315.

Dieses Problem ist offenbar ein konvexes, vorzeichenbeschränktes quadratisches Programm, wie es eingangs dieses Unterabschnitts angegeben wurde. Die Zulässigkeit von (P) impliziert die Lösbarkeit von (D). Ist $y^* \in N$ eine Lösung von (D), so ist $x^* := A^T y^* + z$ die (eindeutige) Lösung von (P). □

Wegen der notwendigen und hinreichenden Optimalitätsbedingungen, angewandt auf das vorzeichenbeschränkte Eingangsproblem (P), ist eine Lösung $x^* \in M$ von (P) charakterisiert durch

$$(Qx^* + c)_j \begin{cases} \geq 0 & (j = 1, \ldots, n_0), \\ = 0 & (j = n_0 + 1, \ldots, n) \end{cases}$$

und

$$(x^*)^T (Qx^* + c) = 0.$$

Für einen Vektor $x = (x_j) \in \mathbb{R}^n$ sei im folgenden der Vektor $x_+ \in \mathbb{R}^n$ definiert durch

$$(x_+)_j = \begin{cases} \max(0, x_j) & (j = 1, \ldots, n_0), \\ x_j & (j = n_0 + 1, \ldots, n). \end{cases}$$

Man überzeugt sich leicht davon, dass x_+ die orthogonale Projektion von x auf

$$C := \{x \in \mathbb{R}^n : x_j \geq 0 \ \ (j = 1, \ldots, n_0)\}$$

ist. Dann kann man zunächst die Optimierungsaufgabe (P) in eine Fixpunktaufgabe überführen, wie das folgende, einfach zu beweisende Lemma aussagt.

Lemma 3.1 *Ist $x^* \in \mathbb{R}^n$ eine Lösung von*

$$(*) \qquad x = [x - \alpha(Qx + c)]_+$$

mit einem $\alpha > 0$, so ist $x^* \in M$ eine Lösung von (P). Ist umgekehrt $x^* \in M$ eine Lösung von (P) und $\alpha > 0$ beliebig, so ist x^* eine Lösung von $(*)$.

Die Idee bei W. Li, J. Swetits (1993) besteht darin, die vorzeichenbeschränkte, konvexe quadratische Optimierungsaufgabe (P) bzw. die Fixpunktaufgabe $(*)$ als eine äquivalente *unrestringierte* konvexe Optimierungsaufgabe zu schreiben. Dass dies möglich ist, und wie dies geschehen kann, ist die Aussage des folgenden Lemmas.

Lemma 3.2 *Mit obigen Bezeichnungen und $\alpha > 0$ definiere man $f_\alpha \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch*

$$f_\alpha(x) := \frac{1}{2} x^T (I - \alpha Q) x - \frac{1}{2} \|[x - \alpha(Qx + c)]_+\|^2,$$

wobei $\|\cdot\|$ die euklidische Norm auf dem $\mathbb{R}^n$ bedeutet. Dann gilt:

1. *Die Abbildung $f_\alpha \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ ist stetig partiell differenzierbar und besitzt den Gradienten*

$$\nabla f_\alpha(x) = (I - \alpha Q)\{x - [x - \alpha(Qx + c)]_+\}.$$

2. *Ist* $0 < \alpha \|Q\| < 1$, *so ist* f_α *konvex auf dem* $\mathbb{R}^n$. *Mit* $B := I - \alpha Q$ *ist genauer* $B(I-B)$ *positiv semidefinit und*

$$f_\alpha(y) - f_\alpha(x) - \nabla f_\alpha(x)^T(y-x) \geq \frac{1}{2}(y-x)^T B(I-B)(y-x) \qquad \text{für alle } x, y \in \mathbb{R}^n.$$

3. *Ist* $0 < \alpha \|Q\| < 1$, *so ist* $x^* \in \mathbb{R}^n$ *genau dann eine Lösung von (P), wenn* x^* *eine Lösung der unrestringiertem konvexen Optimierungsaufgabe*

$$\textit{Minimiere} \quad f_\alpha(x) := \frac{1}{2}x^T(I-\alpha Q)x - \frac{1}{2}\|[x - \alpha(Qx+c)]_+\|^2, \quad x \in \mathbb{R}^n$$

ist.

Beweis: Zur Abkürzung setzen wir

$$B := I - \alpha Q, \qquad d := -\alpha c,$$

so dass f_α kürzer geschrieben werden kann als

$$f_\alpha(x) = \frac{1}{2}x^T Bx - \frac{1}{2}\|(Bx+d)_+\|^2.$$

Die Abbildung $h(t) := \frac{1}{2}(t_+)^2$ ist stetig differenzierbar mit $h'(t) = t_+$, daher ist

$$\lim_{t\to 0} \frac{f_\alpha(x+tp) - f_\alpha(x)}{t} = [B(x - (Bx+d)_+)]^T p,$$

f_α stetig partiell differenzierbar und $\nabla f_\alpha(x) = B(x - (Bx+d)_+)$. Damit ist der erste Teil des Satzes bewiesen.

Für den zweiten Teil des Satzes beachte man, dass $B := I - \alpha Q$ für $0 < \alpha \|Q\| < 1$ positiv definit ist, ferner ist $\|B\| \leq 1$. Dann ist

$$\begin{aligned}
f_\alpha(y) - f_\alpha(x) - \nabla f_\alpha(x)^T(y-x) &= \frac{1}{2}y^T By - \frac{1}{2}x^T Bx \\
&\quad - \frac{1}{2}\|(By+d)_+\|^2 + \frac{1}{2}\|(Bx+d)_+\|^2 \\
&\quad - [x - (Bx+d)_+]^T B(y-x) \\
&= \frac{1}{2}(y-x)^T B(y-x) + (Bx+d)_+^T B(y-x) \\
&\quad - \frac{1}{2}\|(By+c)_+\|^2 + \frac{1}{2}\|(Bx+c)_+\|^2 \\
&\geq \frac{1}{2}(y-x)^T B(y-x) - \frac{1}{2}\|B(y-x)\|^2 \\
&= \frac{1}{2}(y-x)^T B(I-B)(y-x).
\end{aligned}$$

Nicht ganz offensichtlich ist in dieser Gleichungs-Ungleichungskette nur die einzige auftretende Ungleichung bzw.

$$(Bx+d)_+^T B(y-x) - \frac{1}{2}\|(By+d)_+\|^2 + \frac{1}{2}\|(Bx+d)_+\|^2 \geq -\frac{1}{2}\|B(y-x)\|^2.$$

Zur Abkürzung setzen wir

$$p := Bx + d, \qquad q := By + d$$

und zeigen, dass

$$f_i := (p_i)_+(q_i - p_i) - \frac{1}{2}(q_i)_+^2 + \frac{1}{2}(p_i)_+^2 + \frac{1}{2}(q_i - p_i)^2 \geq 0, \qquad i = 1, \ldots, n.$$

Durch Aufsummieren erhält man die gewünschte Ungleichung. Für $i = n_0 + 1, \ldots, n$ und alle $i \in \{1, \ldots, n_0\}$ mit $p_i \geq 0$ und $q_i \geq 0$ ist $f_i = 0$. Für $i \in \{1, \ldots, n_0\}$ mit $p_i \leq 0$ und $q_i \leq 0$ ist $f_i = \frac{1}{2}(q_i - p_i)^2 \geq 0$. Ist schließlich $i \in \{1, \ldots, n_0\}$ mit $p_i \geq 0$ und $q_i \leq 0$ bzw. $p_i \leq 0$ und $q_i \geq 0$, so ist $f_i = \frac{1}{2}q_i^2$ bzw. $f_i = \frac{1}{2}p_i^2 - q_i p_i \geq 0$. Damit haben wir schließlich nachgewiesen, dass für alle $x, y \in \mathbb{R}^n$ gilt

$$f_\alpha(y) - f_\alpha(x) - \nabla f_\alpha(x)^T(y - x) \geq \frac{1}{2}(y - x)^T B(I - B)(y - x).$$

Sind $0 < \lambda_1 \leq \cdots \leq \lambda_n \leq 1$ die Eigenwerte von B (natürlich wird weiter $0 < \alpha \|Q\| < 1$ angenommen), so sind $\lambda_i(1 - \lambda_i)$ die Eigenwerte von $B(I - B)$. Diese liegen sämtlich in $[0, 1]$, daher ist insbesondere $B(I - B)$ positiv semidefinit und folglich

$$f_\alpha(y) - f_\alpha(x) - \nabla f_\alpha(x)^T(y - x) \geq 0 \qquad \text{für alle } x, y \in \mathbb{R}^n.$$

Aus einer bekannten Charakterisierung konvexer Funktionen folgt die Konvexität von $f_\alpha \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf dem gesamten $\mathbb{R}^n$.

Ein $x^* \in \mathbb{R}^n$ ist genau dann (globales) Minimum der konvexen Funktion f_α, wenn

$$\nabla f_\alpha(x^*) = (I - \alpha Q)\{x^* - [x^* - \alpha(Qx^* + c)]_+\} = 0.$$

Für $0 < \alpha \|Q\| < 1$ ist dies wiederum äquivalent dazu, dass x^* Lösung der Fixpunktgleichung $(*)$ in Lemma 3.1 ist, was andererseits äquivalent dazu ist, dass x^* das Ausgangsproblem (P) löst. □
□

Bemerkung: Ist Q positiv definit, so ist $B(I - B)$ mit $B := I - \alpha Q$ für $0 < \alpha \|Q\| < 1$ positiv definit, und daher f_α gleichmäßig konvex auf dem gesamten $\mathbb{R}^n$. □

Nun geht es darum, wie man die unrestringierte Optimierungsaufgabe

$$(**) \qquad \text{Minimiere} \quad f_\alpha(x) := \frac{1}{2}y^T By - \frac{1}{2}\|(Bx + d)_+\|^2, \quad x \in \mathbb{R}^n$$

lösen kann, wobei wir wieder die Abkürzungen

$$B := I - \alpha Q, \qquad d := -\alpha c$$

benutzen. Es liegt nahe, auf $(**)$ das Newton-Verfahren mit exakter Schrittweite anzuwenden. Die Schwierigkeit besteht darin, dass f_α zwar einmal, aber i. Allg. nicht

zweimal stetig partiell differenzierbar ist, weil der Gradient ∇f_α stückweise linear und daher die Hessesche $\nabla^2 f_\alpha$ stückweise konstant ist. Dies soll präzisiert werden, wozu wir

$$h_\alpha(x) := \frac{1}{2}\|(Bx+d)_+\|^2$$

setzen. Bei gegebenem $x \in \mathbb{R}^n$ sei

$$J_0(x) := \{j \in \{1,\dots,n_0\} : (Bx+d)_j = 0\}$$

und

$$J(x) := \{j \in \{1,\dots,n_0\} : (Bx+d)_j > 0\} \cup \{n_0+1,\dots,n\}.$$

Die x, für die $J_0(x) \neq \emptyset$, sind offenbar kritisch insofern, als h_α bzw. f_α dort nicht zweimal differenzierbar sind. Bezeichnet man mit $b_1^T,\dots,b_n^T$ die Zeilen von B (wegen der Symmetrie von B sind dann $b_1,\dots,b_n$ die Spalten von B) und mit $d_1,\dots,d_n$ die Komponenten von d, so sei

$$H_j := \{x \in \mathbb{R}^n : b_j^T x + d_j = 0\}, \qquad j = 1,\dots,n_0,$$

mit H_j^+ bzw. H_j^- seien die zugehörigen offenen Halbräume bezeichnet, also

$$H_j^+ := \{x \in \mathbb{R}^n : b_j^T x + d_j > 0\}, \qquad H_j^- := \{x \in \mathbb{R}^n : b_j^T x + d_i < 0\}.$$

Sei $\beta = (\beta_j) \in \{-1,1\}^{n_0}$, also β ein Vektor mit n_0 Komponenten, die sämtlich gleich -1 oder 1 sind, und

$$D_\beta := \{x \in \mathbb{R}^n : \operatorname{sign}(b_j^T x + d_j) = \beta_j,\ \ j = 1,\dots,n_0\}.$$

Hierbei wird $\operatorname{sign}(0) = 0$ vereinbart. Auf diese Weise zerlegen die n_0 Hyperebenen H_j, $j = 1,\dots,n_0$, den $\mathbb{R}^n$ in 2^{n_0} Mengen, auf denen die Hessesche von h_α bzw. f_α jeweils konstant ist. Für $x \in D_\beta$ ist genauer

$$h_\alpha(x) = \frac{1}{2}\|(Bx+d)_+\|^2 = \frac{1}{2}\sum_{\substack{j=1\\ \beta_j=1}}^{n_0}(b_j^T x + c_j)^2 + \frac{1}{2}\sum_{j=n_0+1}^{n}(b_j^T x + d_j)^2$$

und daher

$$\nabla^2 h_\alpha(x) = \sum_{\substack{j=1\\ \beta_j=1}}^{n_0} b_j b_j^T + \sum_{j=n_0+1}^{n} b_j b_j^T.$$

Definiert man die Diagonalmatrix

$$\Sigma_\beta := \operatorname{diag}(\sigma_1,\dots,\sigma_n)$$

durch

$$\sigma_j := \begin{cases} 1 & \text{falls} \quad j \in \{1,\dots,n_0\},\ \beta_j = 1, \\ 0 & \text{falls} \quad j \in \{1,\dots,n_0\},\ \beta_j = -1, \\ 1 & \text{falls} \quad j \in \{n_0+1,\dots,n\}, \end{cases}$$

so ist

$$\nabla^2 h_\alpha(x) = B\Sigma_\beta B \qquad \text{für } x \in D_\beta$$

und daher

$$\nabla^2 f_\alpha(x) = B - B\Sigma_\beta B.$$

Nun formulieren wir ein geeignet modifiziertes Newton-Verfahren zur Lösung der unrestringierten Optimierungsaufgabe $(**)$ bezw. der vorzeichenbeschränkten quadratischen Optimierungsaufgabe (P). Hierbei setzen wir voraus, daß die Matrix $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv *definit* ist, ferner die Konstante $\alpha > 0$ so klein gewählt ist, dass $\alpha \|Q\| < 1$. Die vorzeichenbeschränkte Optimierungsaufgabe (P) bzw. die unrestringierte Optimierungsaufgabe $(**)$ besitzen dann jeweils die Lösung x^*. Mit

$$B := I - \alpha Q = (\ b_1 \ \cdots \ b_n\ ), \qquad d := -\alpha c = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix}$$

und

$$f_\alpha(x) := \frac{1}{2}x^T Bx - \frac{1}{2}\|(Bx+d)_+\|^2$$

hatten wir in Lemma 3.2 bzw. dessen Beweis nachgewiesen, dass f_α stetig partiell differenzierbar ist,

$$f_\alpha(y) - f_\alpha(x) - \nabla f_\alpha(x)^T(y-x) \geq \frac{1}{2}(y-x)^T B(I-B)(y-x)$$

gilt und B sowie $B(I-B)$ positiv definit sind. Die Hessesche $\nabla^2 f_\alpha$ ist stückweise konstant und positiv definit, für $x \in D_\beta$ hat sie die Form $\nabla^2 f_\alpha(x) = B(I - \Sigma_\beta B)$ mit einer Diagonalmatrix die fast mit der Einheitsmatrix übereinstimmt insofern, als das j-te Diagonalelement eine 0 ist, wenn $j \in \{1,\dots,n_0\}$ und $b_j^T x + d_j = 0$. Ferner ist $\nabla f_\alpha(x) = B[x - (Bx+d)_+]$, die Newton-Richtung also

$$p := -\nabla^2 f(x)^{-1}\nabla f_\alpha(x) = -(I - \Sigma_\beta B)^{-1}[x - (Bx+d)_+].$$

Nun formulieren wir einen Schritt des Newton-Verfahrens mit exakter Schrittweite zur Lösung der unrestringierten Optimierungsaufgabe $(**)$ (und damit des vorzeichenbeschränkten quadratischen Programms (P)).

- Input: Gegeben $x_k \in \mathbb{R}^n$.
- Falls $x_k - (Bx_k + d)_+ = 0$, dann: STOP, x_k ist die Lösung von (P).
- Bestimme $\beta_k \in \{-1,1\}^{n_0}$ mit $x_k \in \mathrm{cl}\,(D_{\beta_k})$, berechne die Diagonalmatrix Σ_{β_k} (s. o.) und anschließend die Newton-Richtung $p_k := \nabla^2 f_\alpha(x_k)^{-1}\nabla f_\alpha(x_k)$ bzw.

$$p_k := -(I - \Sigma_{\beta_k} B)^{-1}[x_k - (Bx_k + d)_+].$$

- Bestimme $t_k > 0$ mit $\nabla f_\alpha(x_k + t_k p_k)^T p_k = 0$ bzw.

$$[x_k - (Bx_k + t_k Bp_k + d)_+]^T (Bp_k) = 0.$$

- Berechne $x_{k+1} := x_k + t_k p_k$.
- Output: Neue Näherung x_{k+1} mit $f_\alpha(x_{k+1}) = \min_{t\geq 0} f_\alpha(x_k + tp_k) < f_\alpha(x_k)$ (oder die Information, dass x_k die Lösung ist).

Die Durchführbarkeit dieses Verfahrens ist völlig klar, denn f_α ist unter den getroffenen Voraussetzungen eine stetig partiell differenzierbare, gleichmäßig konvexe Funktion, ferner ist die Newton-Richtung eine Abstiegsrichtung. Im folgenden Satz (siehe auch W. Li, J. Swetits (1993, Theorem 3.2)) wird ausgesagt, dass das Verfahren nach endlich vielen Schritten mit der Lösung x^* von (P) bzw. $(**)$ abbricht.

Satz 3.3 *Unter obigen Voraussetzungen (insbesondere sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, die Konstante $\alpha > 0$ sei so klein gewählt, dass $\alpha \|Q\| < 1$) bricht das Newton-Verfahren mit exakter Schrittweite bei beliebigem Startwert $x_0 \in \mathbb{R}^n$ nach endlich vielen Schritten mit der Lösung des vorzeichenbeschränkten quadratischen Programms (P) bzw. der unrestringierten Optimierungsaufgabe $(**)$ ab.*

Beweis: Wir machen einen Widerspruchsbeweis, nehmen also an, das Verfahren würde nicht vorzeitig mit der Lösung abbrechen und eine Folge $\{x_k\}$ erzeugen. Die Niveaumenge

$$L_0 := \{x \in \mathbb{R}^n : f_\alpha(x) \leq f_\alpha(x_0)\}$$

ist kompakt (siehe Aufgabe 2), die Folge $\{x_k\}$ besitzt also einen Häufungspunkt x^*. Dieser ist Limes einer Teilfolge $\{x_k\}_{k\in K}$ mit einer unendlichen Teilmenge $K \subset \mathbb{N}$. Es gibt nur 2^{n_0} Mengen $\mathrm{cl}\,(D_\beta)$ mit $\beta \in \{-1, 1\}^{n_0}$. In wenigstens einer dieser Mengen, etwa in $\mathrm{cl}\,(D_{\beta^*})$, müssen unendlich viele der Folgenglieder liegen, also etwa $\{x_k\}_{k\in K^*}$ mit einer (unendlichen) Menge $K^* \subset K$ (und natürlich auch x^* selber). Für $k \in K^*$ ist also

$$p_k = -(I - \Sigma_{\beta^*} B)^{-1}[x_k - (Bx_k + d)_+].$$

Offenbar ist

$$\lim_{k\in K^*, k\to\infty} p_k = p^* := -(I - \Sigma_{\beta^*} B)^{-1}[x^* - (Bx^* + d)_+].$$

Angenommen, es wäre $x^* \neq (Bx^* + d)_+$ bzw. $p^* \neq 0$. Wie wir aus der unrestringierten Optimierung (die benötigten Voraussetzungen sind wegen Aufgabe 2 erfüllt) wissen, existiert eine (vom Iterationsindex k unabhängige) Konstante $\theta > 0$ mit

$$f_\alpha(x_k) - f_\alpha(x_{k+1}) \geq \theta \left(\frac{\nabla f_\alpha(x_k)^T p_k}{\|p_k\|}\right)^2 \qquad \text{für alle } k.$$

Da $\{f_\alpha(x_k)\}$ monoton fallend, nach unten beschränkt ist, konvergiert $\{f_\alpha(x_k)\}$, insbesondere gilt $\lim_{k\to\infty}[f_\alpha(x_k) - f_\alpha(x_{k+1})] = 0$. Dann ist aber auch

$$0 = \lim_{k\in K^*, k\to\infty} \frac{\nabla f_\alpha(x_k)^T p_k}{\|p_k\|} = \frac{[x^* - (Bx^* + d)_+]^T B(I - \Sigma_{\beta^*} B)^{-1}[x^* - (Bx^* + d)_+]}{\|p^*\|}.$$

Nun ist auch die Matrix $B(I - \Sigma_{\beta^*} B)^{-1}$ symmetrisch und positiv definit, woraus wir schließlich doch $x^* = (Bx^* + d)_+$ erhalten.

Bisher haben wir bewiesen: Bricht das Verfahren nicht nach endlich vielen Schritten ab, so ist die Lösung x^* von (P) der einzige Häufungspunkt der durch das Verfahren erzeugten Folge $\{x_k\}$. Dieses x^* liegt, wie unendlich viele der x_k, in einer Menge $\mathrm{cl}\,(D_{\beta^*})$. Hierbei ist $\beta^* \in \{-1,1\}^{n_0}$ und

$$D_{\beta^*} = \{x \in \mathbb{R}^n : \mathrm{sign}\,(b_j^T x + d_j) = \beta_j^*,\ \ j = 1,\ldots,n_0\}.$$

Sei nun x_k das erste Folgenglied, das in $\mathrm{cl}\,(D_{\beta^*})$ liegt. Wir wollen uns überlegen, dass dann $x_{k+1} = x^*$ bzw. $\nabla f_\alpha(x_{k+1}) = 0$, der Algorithmus also im nächsten Schritt mit der Lösung x^* stehen bleibt. Auf D_{β^*} ist f_α eine quadratische Funktion, und zwar ist

$$f_\alpha(x) = f_\alpha(x^*) + \frac{1}{2}(x-x^*)^T B(I - \Sigma_{\beta^*} B)(x-x^*) \qquad \text{für } x \in \mathrm{cl}\,(D_{\beta^*}).$$

Nun definiere man $g_\alpha\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ als die (auf dem ganzen $\mathbb{R}^n$) quadratische Funktion

$$g_\alpha(x) := f_\alpha(x^*) + \frac{1}{2}(x-x^*)^T B(I - \Sigma_{\beta^*} B)(x-x^*).$$

Dann ist $\nabla g_\alpha(x^*) = 0$, also ist x^* nicht nur Minimum von f_α sondern auch Minimum von g_α auf dem ganzen $\mathbb{R}^n$ bzw.

$$g_\alpha(x^*) = \min_{x\in\mathbb{R}^n} g_\alpha(x).$$

Nun ist einerseits

$$x_k - \nabla^2 g_\alpha(x_k)^{-1} \nabla g_\alpha(x_k) = x^*,$$

da das Newton-Verfahren, angewandt auf die gleichmäßig konvexe, quadratische Funktion g_α, das Minimum von g_α in einem Schritt findet, andererseits ist wegen $f_\alpha(x) = g_\alpha(x)$ für $x \in \mathrm{cl}\,(D_{\beta^*})$ und $x_k \in \mathrm{cl}\,(D_{\beta^*})$ offenbar

$$x_k - \nabla g_\alpha(x_k)^{-1} \nabla g_\alpha(x_k) = x_k - \nabla^2 f_\alpha(x_k)^{-1} \nabla f_\alpha(x_k) = x_k + p_k.$$

Dann ist aber

$$f_\alpha(x_{k+1}) \le f_\alpha(x_k + p_k) = f_\alpha(x^*) = \min_{x\in\mathbb{R}^n} f_\alpha(x)$$

und folglich $x_{k+1} = x^*$. Der Satz ist damit bewiesen. □ □

Nun ist geklärt, dass das oben formulierte Newton-Verfahren unter den Voraussetzungen von Satz 3.3 die Lösung in endlich vielen Schritten findet. Für die Umsetzung (wir setzen weiter voraus, dass $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit ist und $\alpha > 0$ so klein gewählt ist, daß $\alpha\,\|Q\| < 1$) sollten wir noch kurz darauf eingehen, wie man die exakte Schrittweite bestimmt. Sei also $x \in \mathbb{R}^n$ eine aktuelle Näherung, für die $x \neq (Bx+d)_+$, die also noch nicht die Lösung der gestellten Aufgabe ist. Sei $p \in \mathbb{R}^n$ die Newton-Richtung (insbesondere ist $p \neq 0$) und

$$g(t) := \nabla f_\alpha(x+tp)^T p = [x + tp - (Bx + tBp + d)_+]^T (Bp),$$

zu lösen ist die Nullstellenaufgabe $g(t) = 0$. Aus Teil 2 von Lemma 3.2 erhalten wir

$$f_\alpha(z) - f_\alpha(y) - \nabla f_\alpha(y)^T (z-y) \ge \frac{1}{2}(z-y)^T B(I-B)(z-y)$$

und (vertausche die Rollen von y und z)

$$f_\alpha(y) - f_\alpha(z) - \nabla f_\alpha(z)^T(y-z) \geq \frac{1}{2}(z-y)^T B(I-B)(z-y)$$

für alle $y, z \in \mathbb{R}^n$. Durch Additon dieser beiden Ungleichungen folgt

$$[\nabla f_\alpha(z) - \nabla f_\alpha(y)]^T(z-y) \geq (z-y)^T B(I-B)(z-y) \qquad \text{für alle } y, z \in \mathbb{R}^n.$$

Setzt man hier wiederum speziell $y := x + sp$ und $z := x + tp$ mit $s \leq t$, so folgt

$$[\nabla f_\alpha(x+tp) - \nabla f_\alpha(x+sp)]^T p \geq (t-s)p^T B(I-B)p,$$

mit der positiven, nur von p und B abhängenden Konstanten $c_0 := p^T B(I-B)p$ ist also

$$g(t) - g(s) \geq c_0(t-s) \qquad \text{für } s \leq t.$$

Insbesondere ist g auf $\mathbb{R}$ monoton wachsend (und natürlich auch stetig). Genauer ist

$$g(t) = [x + tp - (Bx + tBp + d)_+]^T(Bp) = \gamma + \beta t + v^T(b - tv)_+,$$

wobei wir zur Abkürzung

$$v := -Bp, \qquad \gamma := -x^T v, \qquad \beta := -p^T v, \qquad b := Bx + d$$

gesetzt haben. Die Funktion g ist stückweise linear, wobei höchstens n_0 "Knicke" auftreten können. Um die Bezeichnungen etwas einfacher zu gestalten, wollen wir nun annehmen, dass $n_0 = n$, dass also im Ausgangsproblem alle Variablen vorzeichenbeschränkt sind. Im folgenden ist also x_+ der "nichtnegative Anteil" des Vektors x. Zur Bestimmung einer Nullstelle der stückweise linearen, monoton wachsenden Funktion g kann man den folgenden Algorithmus benutzen. Hierbei vergessen wir, dass $g(0) < 0$, so dass wegen der Monotonie von g eine Nullstelle nur in $(0, \infty)$ liegen kann.

1. Streiche alle verschwindenden Komponenten von v und die entsprechenden Komponenten von b. Sei $J \subset \{1, \ldots, n\}$ die Menge der verbleibenden Indizes.

2. Wähle $j_0 \in J$ und setze $t := b_{j_0}/v_{j_0}$. Berechne

$$g(t) := \gamma + \beta t + \sum_{j \in J} v_j \max(0, b_j - tv_j).$$

3. Falls $g(t) = 0$, dann: STOP, t ist die gesuchte Nullstelle.

4. Für alle $i \in J$:

   (a) Falls ($g(t) > 0$, $v_i > 0$ und $b_i/v_i \geq t$) oder ($g(t) < 0$, $v_i < 0$ und $b_i/v_i \leq t$):
   Berechne $\gamma := \gamma + v_i b_i$, $\beta := \beta - v_i^2$ und setze $J := J \setminus \{i\}$.

   (b) Falls ($g(t) > 0$, $v_i < 0$ und $b_i/v_i \geq t$) oder ($g(t) < 0$, $v_i > 0$ und $b_i/v_i \leq t$):
   Setze $J := J \setminus \{i\}$.

5. Falls $J \neq \emptyset$, so gehe zu Schritt 2.

6. Ist $\beta \neq 0$, so ist $t := -\gamma/\beta$ eine Nullstelle von g, andernfalls hat g keine Nullstelle.

Zumindestens auf den ersten Blick ist keineswegs klar, dass dieser Algorithmus korrekt ist. Nehmen wir z. B. an, es sei $g(t) > 0$, $v_i > 0$ und $b_i/v_i \geq t$. Wegen der Monotonie von g liegt die zu bestimmende Nullstelle von g links von t, wegen $b_i/v_i \geq t$ auch links von b_i/v_i. Wegen $v_i > 0$ ist $b_i - sv_i \geq 0$ für alle s links von b_i/v_i. Für diese s ist daher

$$g(s) = \gamma + \beta s + \sum_{j \in J} v_j \max(0, b_j - sv_j) = \gamma + v_i b_i + (\beta - v_i^2)s + \sum_{j \in J \setminus \{i\}} v_j \max(0, b_j - sv_j).$$

Ist dagegen $g(t) < 0$, $v_i < 0$ und $b_i/v_i \leq t$, so braucht man nicht mehr links von t und erst recht nicht links von b_i/v_i nach einer Nullstelle von g zu suchen. Wegen $v_i < 0$ ist $b_i - sv_i \geq 0$ für alle s rechts von b_i/v_i, so dass dort wieder

$$g(s) = \gamma + v_i b_i + (\beta - v_i^2)s + \sum_{j \in J \setminus \{i\}} v_j \max(0, b_j - sv_j).$$

Entsprechend ist die Argumentation für 4b. Ist etwa $g(t) > 0$, $v_i < 0$ und $b_i/v_i \geq t$, so liegt die gesuchte Nullstelle von g links von t und erst recht links von b_i/v_i. Für $s \leq b_i/v_i$ ist aber

$$g(s) = \gamma + \beta s + \sum_{j \in J \setminus \{i\}} v_j \max(0, b_j - sv_j).$$

Entsprechendes gilt für den zweiten Fall in Schritt 4b. In jedem Schritt wird wenigstens ein Index, nämlich der in Schritt 2 gewählte Index j_0, aus der Indexmenge J entfernt. Daher endet der Algorithmus nach endlich vielen Schritten mit $J = \emptyset$. Dort, wo eine Nullstelle von g nur liegen kann, ist g durch $g(t) = \gamma + \beta t$ gegeben, so dass für $\beta \neq 0$ die gesuchte Nullstelle durch $t = -\gamma/\beta$ gegeben ist.

Damit dürfte die Korrektheit des angegebenen Algorithmus klar sein. Auf weitere Modifikationen wird bei W. Li, J. Swetits (1993) eingegangen.

3.3.4 Aufgaben

1. Man beweise Lemma 3.1. Bleibt die Aussage von Lemma 3.1 richtig, wenn man α durch eine positive Diagonalmatrix ersetzt?

2. Sei $f_\alpha \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ definiert wie in Lemma 3.2, also durch

$$f_\alpha(x) := \frac{1}{2} x^T (I - \alpha Q)x - \frac{1}{2} \| [x - \alpha(Qx + c)]_+ \|^2,$$

wobei $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit ist und $\alpha > 0$ so klein gewählt ist, dass $\alpha \, \|Q\| < 1$. Man zeige:

(a) Der Gradient ∇f_α ist auf dem $\mathbb{R}^n$ global lipschitzstetig, es existiert also eine Konstante $\gamma > 0$ mit

$$\|\nabla f_\alpha(x) - \nabla f_\alpha(y)\| \leq L \, \|x - y\| \qquad \text{für alle } x, y \in \mathbb{R}^n.$$

(b) Ist Q sogar positiv definit, so ist bei beliebigem $x_0 \in \mathbb{R}^n$ die Niveaumenge

$$L_0 := \{x \in \mathbb{R}^n : f_\alpha(x) \le f_\alpha(x_0)\}$$

kompakt.

3. Gegeben[17] sei das quadratische Programm

(P) Minimiere $c^T x + \frac{1}{2} x^T Q x$ unter den Nebenbedingungen $l \le Ax \le u$.

Hierbei sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, ferner seien $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^n$ und $l, u \in \mathbb{R}^m$ mit $l \le u$ gegeben. Für einen Vektor $v \in \mathbb{R}^m$ seien die Vektoren v_+ bzw. $(v)_l^u$ in naheliegenderweise als Projektion von v auf den nichtnegativen Orthanten bzw. den Quader $[l, u]$ definiert. Man zeige:

(a) Für alle $v \in \mathbb{R}^m$ ist

$$(v)_l^u = v + (l - v)_+ - (v - u)_+.$$

(b) Es ist $x \in \mathbb{R}^n$ genau dann die Lösung von (P), wenn ein $w \in \mathbb{R}^m$ mit

$$Qx + c - A^T w = 0, \qquad Ax = (Ax - \alpha w)_l^u$$

existiert.

(c) Sei $\alpha > 0$ beliebig. Dann ist $x(w) := Q^{-1}(A^T w - c)$ mit einem $w \in \mathbb{R}^m$ genau dann die Lösung von (P), wenn $Ax(w) = (Ax(w) - \alpha w)_l^u$.

(d) Sei $\alpha > 0$ beliebig. Dann ist $x(w) := Q^{-1}(A^T w - c)$ mit einem $w \in \mathbb{R}^m$ genau dann die Lösung von (P), wenn

$$\phi_\alpha(w) := AQ^{-1}(A^T w - c) - [AQ^{-1}(A^T w - b) - \alpha w]_l^u = 0.$$

(e) Sei $\alpha > 0$. Zur Abkürzung setze man

$$B_\alpha := \alpha I - AQ^{-1}A^T, \qquad a := l + AQ^{-1}c, \qquad b := -(AQ^{-1}c + u).$$

Dann ist

$$\phi_\alpha(w) = \alpha w - (a + B_\alpha w)_+ + (b - B_\alpha w)_+.$$

(f) Mit $\alpha > 0$ definiere man $\Phi_\alpha\colon \mathbb{R}^m \longrightarrow \mathbb{R}$ durch

$$\Phi_\alpha(w) := \frac{\alpha}{2} w^T B_\alpha w - \frac{1}{2}\|(a + B_\alpha w)_+\|^2 - \frac{1}{2}\|(b - B_\alpha w)_+\|^2,$$

wobei $\|\cdot\|$ die euklidische Norm auf dem $\mathbb{R}^m$ bedeutet. Dann gilt:

i. Die Abbildung Φ_α ist auf dem $\mathbb{R}^m$ stetig partiell differenzierbar und besitzt den Gradienten

$$\nabla\Phi_\alpha(w) = B_\alpha[\alpha w - (a + B_\alpha w)_+ + (b - B_\alpha w)_+] = B_\alpha \phi_\alpha(w).$$

ii. Ist $\alpha > \|AQ^{-1}A^T\|$, so ist Φ_α auf dem $\mathbb{R}^m$ konvex, genauer ist

$$\Phi_\alpha(w) - \Phi_\alpha(v) - \nabla\Phi_\alpha(v)^T(w - v) \ge \frac{1}{2}(w - v)^T B_\alpha(\alpha I - B_\alpha)(w - v) \ge 0$$

für alle $v, w \in \mathbb{R}^m$.

[17] Siehe auch

W. Li, J. Swetits (1997) "A new algorithm for solving strictly convex quadratic programs". SIAM J. Optimization 7, 595–619.

Kapitel 4

Linear restringierte Optimierungsaufgaben

In diesem Kapitel betrachten wir Optimierungsaufgaben der Form

$$\text{(P)} \quad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$

Es handelt sich hier also um die Aufgabe, eine nichtlineare, und i. Allg. nicht quadratische, Zielfunktion unter (affin) linearen Ungleichungs- und Gleichungsrestriktionen zu minimieren. I. Allg. werden wir mindestens einmalige stetige Differenzierbarkeit der Zielfunktion f voraussetzen. Wir werden wieder die Abkürzungen

$$A := \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix} \in \mathbb{R}^{m \times n}, \qquad b := \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m$$

benutzen. Wir werden zunächst auf die *Methode der aktiven Mengen*, danach auf *Verfahren der zulässigen Richtungen* eingehen.

4.1 Die Methode der aktiven Mengen

Auf der Darstellung bei P. E. Gill, W. Murray, M. H. Wright (1981, S. 155 ff.)[1] und R. Fletcher (1987, S. 259 ff.)[2] aufbauend schildern wir zunächst die Methode der aktiven Mengen. Ähnlich wie in der quadratischen Optimierung (siehe das Verfahren von Fletcher in Unterabschnitt 3.1.1) wird hierbei eine Folge von Optimierungsaufgaben mit linearen Gleichungsrestriktionen gelöst. Daher beschäftigen wir uns hiermit im nächsten Unterabschnitt.

[1] Gill, P. E., W. Murray and M. H. Wright (1981) *Practical Optimization.* Academic Press, London-New York.

[2] Fletcher, R. (1987) *Practical Methods of Optimization. Second Edition.* John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.

4.1.1 Lineare Gleichungsrestriktionen

In diesem Unterabschnitt betrachten wir Aufgaben der Form

(P) Minimiere $f(x)$ unter der Nebenbedingung $Ax = b$,

wobei $A \in \mathbb{R}^{m\times n}$ mit $\operatorname{Rang}(A) = m$. Natürlich kann man bei nichtquadratischem $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ nicht erwarten, dass man diese Aufgabe in endlich vielen Schritten lösen kann, es liegt aber nahe, dass man sie auf eine unrestringierte Optimierungsaufgabe zurückführen kann. Wir nehmen hierzu an, es sei eine nichtsinguläre Matrix

$$(\underbrace{Y}_{m} \quad \underbrace{Z}_{n-m}) \in \mathbb{R}^{n\times n}$$

mit $AY = I$ und $AZ = 0$ bekannt. Insbesondere seien die Spalten von Z eine Basis von $\operatorname{Kern}(A)$, ferner ist Yb eine Lösung des linearen Gleichungssystems $Ax = b$. In einfacher Weise kann dann das durch eine lineare Gleichung restringierte Problem auf die unrestringierte Optimierungsaufgabe

(P_x) Minimiere $\psi(u) := f(x + Zu), \quad u \in \mathbb{R}^{n-m}$

zurückgeführt werden, wobei $Ax = b$, also x den Gleichungsrestriktionen genügt (z. B. ist $x = Yb$). Hierbei hat man sich unter x eine aktuelle zulässige Näherung für eine gesuchte (kritische, lokale, globale) Lösung von (P) vorzustellen (im ersten Schritt setzt man also etwa $x^{(0)} := Yb$). Hierzu kann im Prinzip jedes Verfahren der unrestringierten Optimierung herangezogen werden, also etwa Quasi-Newton-Verfahren und hier insbesondere das BFGS-Verfahren. Der Gradient von ψ (natürlich bezüglich u) ist durch

$$\nabla\psi(u) = Z^T \nabla f(x + Zu),$$

den sogenannten *reduzierten Gradienten*, die Hessesche durch

$$\nabla^2\psi(u) = Z^T \nabla^2 f(x + Zu) Z,$$

die *reduzierte Hessesche*, gegeben. Eine notwendige Bedingung dafür, dass u^* eine Lösung von (P_x) besteht darin, dass $\nabla\psi(u^*) = 0$ und $\nabla^2\psi(u^*)$ positiv semidefinit ist. Dies wiederum ist gleichwertig damit, dass in $x^* := x + Zu^*$ die notwendigen Optimalitätsbedingungen zweiter Ordnung für das durch eine lineare Gleichung restringierte Problem (P) erfüllt sind. Denn $\nabla\psi(u^*) = 0$ ist gleichwertig mit

$$\nabla f(x^*) \in \operatorname{Kern}(Z^T) = \operatorname{Bild}(Z)^\perp = \operatorname{Kern}(A)^\perp = \operatorname{Bild}(A^T),$$

wegen $\operatorname{Kern}(A) = \operatorname{Bild}(Z)$ ist ferner $\nabla^2\psi(u^*)$ genau dann positiv semidefinit, wenn $\nabla^2 f(x^*)$ auf $\operatorname{Kern}(A)$ positiv semidefinit ist. Für das weitere ist es wichtig, auch den zu einer kritischen Lösung x^* von (P) gehörenden Lagrange-Vektor (oder zumindestens eine Näherung) zu berechnen, also einen Vektor $y^* \in \mathbb{R}^m$ mit $\nabla f(x^*) = A^T y^*$. Das ist aber einfach, wenn man die nichtsinguläre Matrix $(\ Y \quad Z\ )$ mit $AY = I$ und $AZ = 0$ bestimmt hat. Denn wir wissen schon, dass $\nabla f(x^*) \in \operatorname{Bild}(A^T)$, also $\nabla f(x^*) = A^T y^*$

mit einem gewissen, wegen der Rangvoraussetzung eindeutig bestimmten $y^* \in \mathbb{R}^m$. Eine Multiplikation mit Y^T liefert

$$Y^T \nabla f(x^*) = Y^T A^T y^* = (AY)^T y^* = y^*,$$

so dass $y^* := Y^T \nabla f(x^*)$ der gesuchte Multiplikator ist. Aus einer Näherung x_k für x^* erhält man schließlich eine Näherung für y^* durch $y_k := Y^T \nabla f(x_k)$.

Bemerkung: Einige Bemerkungen sollten noch dazu gemacht werden, wie die nichtsinguläre Matrix $\begin{pmatrix} Y & Z \end{pmatrix}$ mit $AY = I$ und $AZ = 0$ bestimmt werden kann. Darauf sind wir früher schon eingegangen, daher sind die folgenden Aussagen zum Teil Wiederholungen. Die übliche Methode besteht darin, eine QR-Zerlegung (z. B. mit Householder-Matrizen) von A^T zu bestimmen, also eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ und eine (nichtsinguläre) obere Dreiecksmatrix $R \in \mathbb{R}^{m\times m}$ mit

$$A^T = Q \begin{pmatrix} R \\ 0 \end{pmatrix} = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix} = Q_1 R$$

mit $Q_1 \in \mathbb{R}^{n\times m}$ und $Q_2 \in \mathbb{R}^{n\times(n-m)}$. Dann haben die Matrizen

$$Y := Q_1 R^{-T}, \qquad Z := Q_2$$

offenbar die geforderten Eigenschaften. Den Vektor Yb berechnet man durch Vorwärtseinsetzen und anschließende Multiplikation mit Q_1. Den Multiplikator

$$y^* = Y^T \nabla f(x^*) = R^{-1} Q_1^T \nabla f(x^*)$$

gewinnt man durch Rückwärtseinsetzen. Die angegebene Methode heißt *orthogonale Faktorisierungsmethode* (siehe R. Fletcher (1987, S. 234)).

Allgemeiner kann man zur Bestimmung passender Y und Z folgendermaßen vorgehen: Wähle eine Matrix $V \in \mathbb{R}^{n\times(n-m)}$ derart, dass $\begin{pmatrix} A^T & V \end{pmatrix}$ nichtsingulär ist. Die Inverse dieser Matrix denke man sich durch

$$\begin{pmatrix} A^T & V \end{pmatrix}^{-1} = \begin{pmatrix} Y^T \\ Z^T \end{pmatrix}$$

partitioniert. Dann ist offenbar $AY = I$ und $AZ = 0$. Speziell kann man z. B.

$$V = \begin{pmatrix} 0 \\ I \end{pmatrix}$$

(mit der $m \times m$-Einheitsmatrix I) wählen. Denkt man sich A zerlegt in $A = \begin{pmatrix} A_1 & A_2 \end{pmatrix}$ mit nichtsingulärem $A_1 \in \mathbb{R}^{m\times m}$ und $A_2 \in \mathbb{R}^{m\times(n-m)}$, benutzt man ferner die Identität

$$\begin{pmatrix} A_1^T & 0 \\ A_2^T & I \end{pmatrix}^{-1} = \begin{pmatrix} A_1^{-T} & 0 \\ -A_2^T A_1^{-T} & I \end{pmatrix} = \begin{pmatrix} Y^T \\ Z^T \end{pmatrix},$$

so erhält man Matrizen Y und Z, die den gestellten Forderungen genügen. Setzt man andereseits $V := Q_2$, wobei Q_2 wie in der orthogonalen Faktorisierungsmethode bestimmt ist, so erhält man auch Y und Z wie in dieser Methode. Dies folgt aus

$$\begin{pmatrix} A^T & V \end{pmatrix}^{-1} = \begin{pmatrix} Q_1 R & Q_2 \end{pmatrix}^{-1} = \begin{pmatrix} R^{-1} Q_1^T \\ Q_2^T \end{pmatrix}.$$

Man hat also einige Möglichkeiten, die Matrizen Y und Z zu bestimmen. □

Wie sehen nun Quasi-Newton-Verfahren zur Lösung von (P_x) bzw. (P) genauer aus? Die zulässige Näherung x für (P) stehe zur Verfügung, was der Näherung $u = 0$ von (P_x) entspricht. Ist $B \in \mathbb{R}^{(n-m)\times(n-m)}$ symmetrisch und positiv definit, sowie

$$B \approx \nabla^2\psi(0) = Z^T\nabla^2 f(x)Z,$$

also B eine Näherung für die aktuelle reduzierte Hessesche, so ist die Quasi-Newton-Richtung im u-Raum durch

$$p := -B^{-1}\nabla\psi(0) = -B^{-1}Z^T\nabla f(x)$$

gegeben. Die neue Näherung im u-Raum ist $u_+ := 0 + tp = -tB^{-1}Z^T\nabla f(x)$ mit einer gewissen Schrittweite $t > 0$, was der neuen Näherung

$$x_+ := x + Zu^+ = x + tZp = x - tZB^{-1}Z^T\nabla f(x)$$

im x-Raum entspricht. Man wird also die Schrittweite $t > 0$ wenigstens näherungsweise so bestimmen, dass f auf dem von x ausgehenden Strahl in Richtung Zp minimiert wird. Beim BFGS-Verfahren bestimmt man die neue Matrix B_+ bekanntlich durch

$$B_+ := B - \frac{(Bs)(Bs)^T}{s^TBs} + \frac{yy^T}{y^Ts},$$

wobei

$$y := \nabla\psi(u_+) - \nabla\psi(0) = Z^T[\nabla f(x_+) - \nabla f(x)]$$

und

$$s := u_+ - 0 = u_+.$$

Bekanntlich ist mit B auch B_+ (symmetrisch und) positiv definit, wenn $y^Ts > 0$. Dies ist, wie man aus der unrestringierten Optimierung weiß, keine große Einschränkung. Denn dies ist immer erfüllt, wenn $\psi(\cdot)$ gleichmäßig konvex (siehe Aufgabe 1) ist, wenn die sogenannte Powell-Schrittweite (wir kommen auf diese noch zurück) oder überhaupt eine hinreichend genaue Schrittweitenstrategie benutzt wird. Die Hauptarbeit besteht in der Lösung des linearen Gleichungssystems $Bp = -\nabla\psi(0)$. Für eine stabile Implementation ist es ratsam, aus einer Cholesky-Zerlegung (oder einer LDL^T-Zerlegung) von B eine entsprechende Zerlegung von B_+ zu berechnen. Hinweise hierzu werden in Aufgabe 2 gegeben.

4.1.2 Der allgemeine Fall

Nun gehen wir davon aus, dass wir nichtlineare Optimierungsaufgaben mit linearen Gleichungen als Nebenbedingung lösen können und betrachten das Ausgangsproblem (P). Wir geben den folgenden konzeptionellen Algorithmus an (siehe R. Fletcher (1987, S. 265)). Die Menge der in einem $x \in M$ aktiven Indizes bezeichnen wir wieder mit $I(x)$, die Gleichungsindizes $\{m_0+1, \ldots, m\}$ sind dann in $I(x)$ enthalten. Für eine Indexmenge $I \subset \{1, \ldots, m\}$ seien die Matrix A_I und der Vektor b_I wieder in gewohnter Weise definiert. Wir setzen voraus, dass $A_{I(x)}$ für jedes $x \in M$ vollen Rang besitzt.

1. Gegeben sei ein Paar (x, I) mit $x \in M$, $I = I(x)$ und Rang $(A_I) = \#(I)$. Setze $q := \#(I)$.

2. Bestimme eine Lösung $p^* \in \mathbb{R}^n$ und einen zugehörigen Lagrange-Vektor $y_I^* \in \mathbb{R}^q$ der Optimierungsaufgabe

   (P_0) Minimiere $f(x+p)$ unter der Nebenbedingung $A_I p = 0$.

   Bestimme ferner $l \in I \cap \{1, \ldots, m_0\}$ mit

   $$y_l^* = \min_{i \in I \cap \{1,\ldots,m_0\}} y_i^*.$$

3. Falls $p^* = 0$ und $y_l^* \geq 0$, dann: STOP, da x kritische Lösung von (P) ist.

4. Andernfalls:

   (a) Falls $p^* = 0$, dann setze $x^+ := x$ und $I^+ := I \setminus \{l\}$ und gehe nach 5.

   (b) Andernfalls:

      i. Bestimme die maximale Schrittweite

      $$s(x, p^*) := \min\Big\{ \frac{b_i - a_i^T x}{a_i^T p^*} : i \in \{1, \ldots, m\} \setminus I,\ a_i^T p^* < 0 \Big\},$$

      wobei $s(x, p^*) := +\infty$ gesetzt wird, wenn kein $i \notin I$ mit $a_i^T p^* < 0$ existiert.

      ii. Berechne $t^* > 0$ mit

      $$f(x + t^* p^*) \approx \min_{t \in [0, s(x,p^*)]} f(x + t p^*).$$

      iii. Setze $x^+ := x + t^* p^*$.

      iv. Falls $t^* = s(x, p^*) = (b_r - a_r^T x)/(a_r^T p^*)$, so setze $I^+ := I \cup \{r\}$, andernfalls setze $I^+ := I$. Gehe nach 5.

5. Setze $(x, I) := (x^+, I^+)$ und gehe nach 2.

Einige Bemerkungen zu diesem Algorithmus sind angebracht. Ist in Schritt 3 der Abbruchtest erfüllt, ist also $p^* = 0$ und $y_i^* \geq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$, so setze man $y_i^* := 0$ für alle $i \in \{1, \ldots, m\} \setminus I$ und erkennt anschließend, dass in x die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind, also x eine kritische Lösung ist. Ist zwar $p^* = 0$, aber $y_l^* = \min_{i \in I \cap \{1,\ldots,m_0\}} y_i^* < 0$, so ist im nächsten Schritt die Lösung p^{**} der Aufgabe, $f(x+p)$ unter der Nebenbedingung $A_{I \setminus \{l\}} p = 0$ zu minimieren, nicht der Nullvektor, da $y_l^* \neq 0$ und $\{a_i\}_{i \in I}$ linear unabhängig sind. Weiter ist

$$\nabla f(x)^T p^{**} = (A_I^T y_I^*)^T p^{**} = \underbrace{y_l^*}_{<0} a_l^T p^{**}.$$

I. allg. scheint nicht gesichert zu sein, dass $a_l^T p^{**} > 0$ bzw. p^{**} eine Abstiegsrichtung ist. Ist allerdings f gleichmäßig konvex, so existiert eine Konstante $c > 0$ mit

$$c \, \|p^{**}\|^2 \leq [\nabla f(x + p^{**}) - \nabla f(x)]^T p^{**} = -y_l^* \, a_l^T p^{**},$$

so dass in diesem Falle $a_l^T p^{**} > 0$ und p^{**} eine Abstiegsrichtung für f in x ist. Wegen $a_i^T p^{**} = 0$ für $i \in I(x) \setminus \{l\}$ und $a_l^T p^{**} > 0$ ist p^{**} auch eine zulässige Richtung in der aktuellen Näherung x. Jetzt nehmen wir an, es sei $p^* \neq 0$. Natürlich ist p^* eine in x zulässige Richtung, da ja $A_{I(x)} p^* = 0$. Damit ist die maximale Schrittweite $s(x, p^*) > 0$ wohldefiniert. Jetzt stellt sich die Frage, ob p^* auch eine Abstiegsrichtung für f in x ist, d. h. ob $\nabla f(x)^T p^* < 0$ ist. Es scheint, als wenn auch hierzu die gleichmäßige Konvexität der Zielfunktion f (wenigstens lokal) gegeben sein muss. Zunächst existiert ein Vektor $y_I^* \in \mathbb{R}^q$ mit

$$\nabla f(x + p^*) = A_I^T y_I^*,$$

woraus man mit $A_I p^* = 0$ erhält, dass $\nabla f(x + p^*)^T p^* = 0$. Mit einer positiven Konstanten c ist daher

$$c \, \|p^*\|^2 \leq [\nabla f(x + p^*) - \nabla f(x)]^T p^* = -\nabla f(x)^T p^*$$

und damit $\nabla f(x)^T p^* < 0$ bzw. p^* eine Abstiegsrichtung für f in x. I. allg. ist $s(x, p^*) < \infty$, in diesem Falle wird (mindestens[3]) eine bisher inaktive Ungleichungsrestriktion aktiv.

4.1.3 Aufgaben

1. Ist $D \subset \mathbb{R}^n$ konvex, so heißt eine Funktion $f\colon D \longrightarrow \mathbb{R}$ bekanntlich auf D *gleichmäßig konvex*, wenn eine Konstante $c > 0$ mit

   $$(1 - \lambda) f(x_1) + \lambda f(x_2) - f((1 - \lambda)x_1 + \lambda x_2) \geq \frac{c}{2} \lambda (1 - \lambda) \, \|x_1 - x_2\|^2$$

   für alle $x_1, x_2 \in D$, $\lambda \in [0, 1]$ existiert.

   Gegeben sei die Optimierungsaufgabe

   (P) $\qquad$ Minimiere $\quad f(x) \quad$ auf $\quad M := \{x \in \mathbb{R}^n : Ax = b\}.$

   Hierbei seien $A \in \mathbb{R}^{m \times n}$ mit Rang $(A) = m$ und $b \in \mathbb{R}^m$ gegeben. Wie in Unterabschnitt 4.1.1 geschildert ordne man (P) die unrestringierte Optimierungsaufgabe

   (P_x) $\qquad$ Minimiere $\quad \psi(u) := f(x + Zu), \quad u \in \mathbb{R}^{n-m},$

   zu, wobei x zulässig für (P) und die Spalten von $Z \in \mathbb{R}^{n \times (n-m)}$ (mit Rang $(Z) = n - m$) eine Basis von Kern (A) bilden. Man zeige: Ist f gleichmäßig konvex auf M, so ist ψ gleichmäßig konvex auf $\mathbb{R}^{n-m}$.

[3] Aus Komplexitätsgründen ist es sinnvoll, wenn sich die Indexmenge I in jedem Schritt um höchstens ein Element verändert. Hier könnte kritisch sein, wenn das Minimum in der Definition der maximalen Schittweite von mehr als einem Element angenommen wird.

2. Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ gegeben (bei der Anwendung in Unterabschnitt 4.1.1 ist n durch $n - m$ zu ersetzen). Es sei eine Cholesky-Zerlegung von B bekannt, also eine untere Dreiecksmatrix L mit positiven Diagonalelementen mit $B = LL^T$. Ferner sei

$$B_+ := B - \frac{(Bs)(Bs)^T}{s^T Bs} + \frac{yy^T}{y^T s}.$$

Man zeige:

(a) Ist

$$w := (y^T s)^{1/2} \frac{L^T s}{\|L^T s\|}, \qquad J_+^T := L^T + \frac{w(y - Lw)^T}{y^T s},$$

so ist $B_+ = J_+ J_+^T$.

(b) Die Matrix J_+ ist nichtsingulär und daher B_+ positiv definit.

(c) Ist $J_+^T = Q_+ R_+$ eine QR-Zerlegung von J_+^T, wobei (Q_+ orthogonal und) R_+ eine obere Dreiecksmatrix mit positiven Diagonalelementen ist, so ist $B_+ = L_+ L_+^T$ mit $L_+ := R_+^T$ eine Cholesky-Zerlegung von B_+.

(d) Die QR-Zerlegung einer durch eine Matrix vom Rang 1 gestörten oberen Dreiecksmatrix kann in $O(n^2)$ Flops berechnet werden.

4.2 Verfahren der zulässigen Richtungen

4.2.1 Einige grundlegende Begriffe

In Definition 3.2 in Abschnitt 2.3 hatten wir den Kegel $F(M; x)$ der zulässigen Richtungen an eine Menge $M \subset \mathbb{R}^n$ in einem Punkt $x \in M$ durch

$$F(M; x) := \left\{ p \in \mathbb{R}^n : \begin{array}{l} \text{Es existiert eine Folge } \{t_k\} \subset \mathbb{R}_+ \text{ mit} \\ t_k \to 0 \text{ und } x + t_k p \in M \text{ für alle } k \end{array} \right\}$$

definiert. Ein großer Vorteil linearer Restriktionen, also einem Polyeder M in obiger Darstellung als Restriktionenmenge, besteht darin, dass der Kegel der zulässigen Richtungen leicht angegeben werden kann. Bezeichnet man wieder mit

$$I(x) := \{i \in \{1, \ldots, m_0\} : a_i^T x = b_i\}$$

die Menge der in $x \in M$ aktiven Ungleichungsrestriktionen, so ist offenbar

$$F(M; x) = \{p \in \mathbb{R}^n : a_i^T p \geq 0 \ \ (i \in I(x)),\ a_i^T p = 0 \ \ (i = m_0 + 1, \ldots, m)\}.$$

Ist $p \in F(M; x)$ und $\nabla f(x)^T p < 0$, so sprechen wir von einer *zulässigen Abstiegsrichtung* in $x \in M$. Gibt es zu einem $x \in M$ keine zulässige Absiegsrichtung, ist also $\nabla f(x)^T p \geq 0$ für alle $p \in F(M; x)$, so kann man z. B. mit Hilfe des Farkas-Lemmas leicht zeigen, dass in x die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind, d. h. es existiert ein $y \in \mathbb{R}^m$ mit

$$y_i \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x) = A^T y, \qquad y^T (Ax - b) = 0.$$

Eine zulässige Lösung, in der die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind, nennen wir auch eine *kritische Lösung* von (P). Auch die Umkehrung der obigen Aussage ist richtig: Ist $x \in M$ eine kritische Lösung von (P), so gibt es in x keine zulässige Abstiegsrichtung. Ein weiterer Vorteil linearer Restriktionen besteht darin, dass man ziemlich einfach die *maximale Schrittweite* berechnen kann. Allgemein bezeichnen wir für konvexes $M \subset \mathbb{R}^n$ bei gegebenen $x \in M$, $p \in F(M; x)$ mit

$$s(x, p) := \sup\{t > 0 : x + tp \in M\}$$

die maximale Schrittweite. Hierbei ist $s(x, p) = +\infty$ möglich, wenn nämlich der gesamte, von x in Richtung p ausgehende Strahl innerhalb von M verläuft. Ist M durch ein Polyeder mit der Darstellung wie im linear restringierten Programm (P) gegeben, so ist offenbar

$$s(x, p) = \min\Big\{ \frac{a_i^T x - b_i}{(-a_i^T p)} : i \in \{1, \ldots, m_0\} \setminus I(x),\ a_i^T p < 0 \Big\}.$$

Ist $a_i^T p \geq 0$ für alle $i \in \{1, \ldots, m_0\} \setminus I(x)$, so ist natürlich $s(x, p) = +\infty$.

Ein Modellalgorithmus für ein Verfahren der zulässigen Richtungen sieht dann folgendermaßen aus:

- Gegeben $x_0 \in M$.
- Für $k = 0, 1, \ldots$:
  - Falls $F(M; x_k) \cap \{p \in \mathbb{R}^n : \nabla f(x_k)^T p_k < 0\} = \emptyset$, dann: STOP, x_k ist kritische Lösung von (P).
  - Andernfalls:
    - Wähle $p_k \in F(M; x_k)$ mit $\nabla f(x_k)^T p_k < 0$.
    - Wähle $t_k \in (0, s(x_k, p_k)]$ mit $f(x_k + t_k p_k) < f(x_k)$.
    - Setze $x_{k+1} := x_k + t_k p_k$.

4.2.2 Schrittweitenstrategien

Nun ist es einfach, die aus der unrestringierten Optimierung her bekannten Schrittweitenstrategien zu übertragen. Wir werden uns zwar im weiteren auf linear restringierte nichtlineare Optimierungsaufgaben der Form (P) konzentrieren, für die jetzt folgenden Aussagen über Schrittweitenstrategien würde es aber genügen, dass die Menge M der zulässigen Lösungen konvex und abgeschlossen ist. Wie in der unrestringierten Optimierung setzen wir jetzt voraus:

(V) (a) Mit einem gegebenen $x_0 \in M$ (gewöhnlich Startwert eines Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\} \cap M$ kompakt.

(b) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar.

(c) Der Gradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit

$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma \|x - y\| \qquad \text{für alle } x, y \in L_0.$$

Nun kommen wir zur Definition verschiedener Schrittweitenstrategien. In jedem Falle seien die aktuelle Näherung $x \in L_0$ und eine zulässige Abstiegsrichtung $p \in \mathbb{R}^n$ vorgegeben.

Die *Minimum-Schrittweite* $t_M(x,p)$ ist definiert als globales Minimum von

$$\phi(t) := f(x + tp)$$

auf $[0, s(x,p)]$. Da die Niveaumenge L_0 als kompakt vorausgesetzt wurde, existiert $t_M(x,p)$ auch dann, wenn $s(x,p) = +\infty$.

Die *Curry-Schrittweite* $t_C(x,p)$ ist die erste Nullstelle von

$$\phi'(t) = \nabla f(x + tp)^T p$$

in $(0, s(x,p)]$, falls eine solche existiert, andernfalls ist $t_C(x,p) := s(x,p)$.

Diese beiden Schrittweiten nennt man *exakte Schrittweiten*, da zu ihrer Realisierung eine eindimensionale Optimierungsaufgabe bzw. Nullstellenaufgabe exakt gelöst werden muss. Wie in der unrestringierten Optimierung ist es auch hier wichtig, die durch eine gegebene Schrittweitenstrategie erreichbare Verminderung der Zielfunktion nach unten abzuschätzen. Für die gerade eben definierten exakten Schrittweiten erhält man das folgende Ergebnis, das wir hier ohne Beweis angeben (siehe auch Aufgabe 3).

Lemma 2.1 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Dann existiert eine Konstante $\theta_C > 0$ derart, dass*

$$\begin{aligned} f(x) - f(x + t_M(x,p)p) &\geq f(x) - f(x + t_C(x,p)p) \\ &\geq \theta_C \min\Big[-s(x,p)\,\nabla f(x)^T p, \Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2\Big] \end{aligned}$$

für alle nicht kritischen $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$.

Genau wie in der unrestringierten Optimierung spielen auch bei linear restringierten nichtlinearen Optimierungsaufgaben sogenannte *inexakte* Schrittweitenstrategien eine wichtige Rolle. Die beiden wichtigsten sind die *Powell-Schrittweite* (gelegentlich auch nach P. Wolfe benannt) und die *Armijo-Schrittweite*. Diese wollen wir nun genau definieren.

Bei der *Powell-Schrittweite* sind zwei Konstanten $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ vorgegeben. Man setze $t_P(x,p) := s(x,p)$, falls

$$s(x,p) < +\infty \qquad \text{und} \qquad f(x + s(x,p)p) \leq f(x) + \alpha s(x,p)\,\nabla f(x)^T p,$$

andernfalls wähle man $t_P(x,p) \in (0, s(x,p))$ beliebig mit

$$f(x + t_P(x,p)p) \leq f(x) + \alpha t_P(x,p)\,\nabla f(x)^T p, \qquad \nabla f(x + t_P(x,p)p)^T p \geq \beta\,\nabla f(x)^T p.$$

Natürlich stellt sich die Frage, ob die Powell-Schrittweite überhaupt existiert. Hier können wir uns auf den Fall beschränken, dass $s(x,p) < +\infty$, da man das entsprechende Ergebnis sonst aus der unrestringierten Optimierung kennt. Zur Abkürzung setzen wir

$$\psi(t) := f(x) + \alpha t \nabla f(x)^T p - f(x+tp).$$

Dann ist $\psi(0) = 0$ und $\psi'(0) = -(1-\alpha)\,\nabla f(x)^T p > 0$. Angenommen, es sei $\psi(s(x,p)) < 0$. Dann existiert $t_P(x,p) \in (0, s(x,p))$ mit

$$0 < \psi(t_P(x,p)) = f(x) + \alpha\, t_P(x,p)\, \nabla f(x)^T p - f(x + t_P(x,p)p)$$

und

$$0 = \psi'(t_P(x,p)) = \alpha\, \nabla f(x)^T p - \nabla f(x+t_P(x,p)p)^T p > \beta\, \nabla f(x)^T p - \nabla f(x+t_P(x,p)p)^T p.$$

Insgesamt ist die Existenz der Powell-Schrittweite bewiesen.

Im folgenden Lemma wird eine Lemma 2.1 entsprechende Aussage für die Powell-Schrittweite formuliert. Wieder verzichten wir auf einen Beweis (siehe Aufgabe 4)..

Lemma 2.2 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Dann existiert eine Konstante $\theta_P > 0$ derart, dass*

$$f(x) - f(x + t_P(x,p)p) \geq \theta_P \, \min\Big[-s(x,p)\, \nabla f(x)^T p, \Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2\Big]$$

für alle nicht kritischen $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$.

Bemerkung: Eine geringfügige Modifikation der Powell-Schrittweite ist sinnvoll, wenn eine bestimmte Schrittweite, etwa die Schrittweite $t = 1$ ausgezeichnet ist. Das ist immer dann der Fall, wenn die Richtung eine Newton- oder Quasi-Newton-Richtung ist (was das genau ist, das werden wir später erläutern). Denn setzt man etwa $\tilde{s}(x,p) := \min(1, s(x,p))$, so kann man bei vorgegebenen Konstanten $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ die (modifizierte) Powell-Schrittweite $t_P(x,p) := \tilde{s}(x,p)$ setzen, falls

$$f(x + \tilde{s}(x,p)p) \leq f(x) + \alpha\, \tilde{s}(x,p)\, \nabla f(x)^T p,$$

andernfalls bestimme man $t_P(x,p) \in (0, \tilde{s}(x,p))$ mit

$$f(x + t_P(x,p)p) \leq f(x) + \alpha t_P(x,p)\, \nabla f(x)^T p, \qquad \nabla f(x + t_P(x,p)p)^T p \geq \beta\, \nabla f(x)^T p.$$

Natürlich existiert auch diese (modifizierte) Powell-Schrittweite, ferner gilt eine Lemma 2.2 entsprechende Aussage. □

Nun definieren wir schließlich noch die *Armijo-Schrittweite*. Während bei der Powell-Schrittweite eine bestimmte Schrittweite ausgezeichnet sein kann, aber nicht sein muss, wird bei der Armijo-Schrittweite davon ausgegangen, dass die Schrittweite $t = 1$ eine besondere Rolle spielt und nur Schrittweiten in $(0, \tilde{s}(x,p)]$ mit $\tilde{s}(x,p) := \min(1, s(x,p))$ sinnvoll sind. Wir geben eine Version für die Armijo-Schrittweite an, die sich an der Darstellung bei J. Werner (1992, 166 ff.) für unrestringierte Optimierungsaufgaben orientiert.

- Seien $\alpha \in (0, \frac{1}{2})$ und $0 < l \leq u < 1$ gegeben. Setze $\rho_0 := \tilde{s}(x, p)$.
- Für $j = 0, 1, \ldots$:

  Falls $f(x + \rho_j p) \leq f(x) + \alpha\rho_j \nabla f(x)^T p$, dann: Setze $t_A(x, p) := \rho_j$, STOP.

  Andernfalls: Wähle $\rho_{j+1} \in [l\rho_j, u\rho_j]$.

Ist z. B. $l = u =: \rho$, so ist die Armijo-Schrittweite gegeben durch $t_A(x, p) = \rho^j \tilde{s}(x, p)$, wobei j die kleinste nichtnegative ganze Zahl mit

$$f(x + \rho^j \tilde{s}(x, p)p) \leq f(x) + \alpha\rho^j \tilde{s}(x, p) \nabla f(x)^T p$$

ist. In dieser Form wird die Armijo-Schrittweite i. allg. in der Literatur angegeben, wegen der größeren Flexibilität ziehen wir aber obige Darstellung vor.

Die *Existenz* der Armijo-Schrittweite ist einfach einzusehen. Denn würde die obige Schleife zur Definition der Armijo-Schrittweite nicht vorzeitig abbrechen, so würde eine Folge $\{\rho_j\} \subset \mathbb{R}_+$ mit $\rho_j \to 0+$ und

$$\frac{f(x + \rho_j p) - f(x)}{\rho_j} > \alpha \nabla f(x)^T p, \qquad j = 0, 1, \ldots$$

existieren. Mit $j \to \infty$ erhielten wir $(1 - \alpha)\nabla f(x)^T p \geq 0$, was ein Widerspruch zu $\alpha < 1$ und $\nabla f(x)^T p < 0$ ist.

Nun kommt noch die Lemma 2.1 und 2.2 entsprechende Aussage für die Armijo-Schrittweite (siehe Aufgabe 5).

Lemma 2.3 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Dann existiert eine Konstante $\theta_A > 0$ derart, dass*

$$f(x) - f(x + t_A(x, p)p) \geq \theta_A \min\Big[-\tilde{s}(x, p) \nabla f(x)^T p, \Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2\Big]$$

für alle nicht kritischen $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$.

Die Aussagen der Lemmata 2.1, 2.2 und 2.3 reduzieren sich natürlich genau auf die aus der unrestringierten Optimierung bekannten Resultate, wenn $s(x, p) = +\infty$.

4.2.3 Richtungsstrategien

Gegeben sei wieder die linear restringierte Optimierungsaufgabe (P). Für eine gegebene aktuelle Näherung $x \in M$ und $\epsilon \geq 0$ sei

$$I_\epsilon(x) := \{i \in \{1, \ldots, m_0\} : a_i^T x - b_i \leq \epsilon\}$$

die Indexmenge der in x ϵ-aktiven Ungleichungsrestriktionen. Man beachte, dass $I_0(x)$ die Menge der in x aktiven Ungleichungsrestriktionen ist, ferner ist offenbar $I_\epsilon(x) = \{1, \ldots, m_0\}$ für alle hinreichend großen ϵ. Schließlich seien noch die folgenden Voraussetzungen (siehe der vorige Unterabschnitt) erfüllt:

(V) (a) Mit einem gegebenen $x_0 \in M$ (gewöhnlich Startwert eines Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \le f(x_0)\} \cap M$ kompakt.

(b) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar.

(c) Der Gradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit

$$\|\nabla f(x) - \nabla f(y)\| \le \gamma \|x - y\| \qquad \text{für alle } x, y \in L_0.$$

Wir stellen uns nun das folgende Problem: Gegeben sei ein nicht kritisches $x \in L_0$, etwa eine aktuelle Näherung für eine (lokale oder kritische) Lösung von (P). Gesucht ist eine in x zulässige Abstiegsrichtung, also ein $p \in F(M; x)$ mit $\nabla f(x)^T p < 0$. Das folgende Lemma gibt eine Antwort auf dieses Problem.

Lemma 2.4 *Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $\epsilon \ge 0$ und $x \in M$. Sei p die eindeutige Lösung des quadratischen Programms*

$$(\mathrm{P}_\epsilon(x)) \qquad \begin{cases} \textit{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T B p \quad \textit{unter den Nebenbedingungen} \\ a_i^T p \ge b_i - a_i^T x \quad (i \in I_\epsilon(x)), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m). \end{cases}$$

Dann gilt: Ist $p \ne 0$, so ist p eine in x zulässige Abstiegsrichtung mit $0 < p^T B p \le -\nabla f(x)^T p$, andernfalls ist x eine kritische Lösung von (P).

Beweis: Natürlich besitzt das quadratische Hilfsproblem $(\mathrm{P}_\epsilon(x))$ eine eindeutige Lösung, da es zulässig ist ($p = 0$ genügt allen Restriktionen) und die Zielfunktion gleichmäßig konvex ist. Die Lösung p ist durch die Existenz von Multiplikatoren y_i, $i \in I_\epsilon(x) \cup \{m_0 + 1, \ldots, m\}$ charakterisiert, welche den Bedingungen

$$y_i \ge 0 \quad (i \in I_\epsilon(x)), \qquad \nabla f(x) + Bp = \sum_{i \in I_\epsilon(x)} y_i a_i + \sum_{i=m_0+1}^{m} y_i a_i$$

sowie

$$y_i(a_i^T p + a_i^T x - b_i) = 0 \quad (i \in I_\epsilon(x))$$

genügen. Ist $p = 0$ und definiert man $y_i := 0$ für alle $i \in \{1, \ldots, m_0\} \setminus I_\epsilon(x)$, so ist

$$y_i \ge 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x) = \sum_{i=1}^{m} y_i a_i, \qquad y_i(a_i^T x - b_i) = 0 \quad (i = 1, \ldots, m).$$

Das wiederum bedeutet, dass in x die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind bzw. x eine kritische Lösung von (P) ist. Sei daher nun $p \ne 0$. Offensichtlich ist $p \in F(M; x)$. Ferner ist

$$\nabla f(x)^T p + p^T B p = \sum_{i \in I_\epsilon(x)} y_i a_i^T p = \sum_{i \in I_\epsilon(x)} \underbrace{y_i}_{\ge 0} \underbrace{(b_i - a_i^T x)}_{\le 0} \le 0,$$

so dass, wie behauptet, $0 < p^T Bp \leq -\nabla f(x)^T p$, insbesondere p also auch eine Abstiegsrichtung ist. □
□

Bemerkung: Die Motivation für die in Lemma 2.4 angegebene Richtungsstrategie dürfte klar sein. Ist nämlich $x \in M$ eine aktuelle Näherung und f in x zweimal differenzierbar, so ist

$$f(x+p) \approx f(x) + \nabla f(x)^T p + \tfrac{1}{2}\, p^T \nabla^2 f(x) p.$$

Daher liegt es nahe, B als eine Approximation an die Hessesche $\nabla^2 f(x)$ zu wählen. Für $B := \nabla^2 f(x)$ wird man von dem *Newton-Verfahren* zur Lösung des linear restringierten Programms (P) sprechen, wobei man sich die beschriebene Richtungsstrategie noch mit einer geeigneten Schrittweitenstrategie kombinieren muss. Ferner ist $x + p \in M$ genau dann, wenn

$$a_i^T p \geq b_i - a_i^T x \quad (i = 1, \ldots, m_0), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m).$$

Für große ϵ sind dies genau die Restriktionen des quadratischen Programms $(\mathrm{P}_\epsilon(x))$ in Lemma 2.4. Nun möchte man in dem Hilfsproblem zur Berechnung der Richtung möglichst wenige Restriktionen haben. Das andere Extrem besteht darin, $\epsilon = 0$ zu wählen. Dann ist $F(M; x)$ die Restriktionenmenge im Programm $(\mathrm{P}_0(x))$. Ein Beispiel von P. Wolfe (siehe z. B. R. Fletcher (1987, S. 276)) zeigt aber, dass man *nicht* durchgehend $\epsilon := 0$ setzen sollte, weil dann das Phänomen des sogenannten "Zigzagging" auftreten kann. □

Bemerkung: Ist in Lemma 2.4 die Matrix B nur noch positiv semidefinit (ist z. B. $B = 0$, was im wesentlichen einer Linearisierung der Zielfunktion entspricht), so braucht das Problem $(\mathrm{P}_\epsilon(x))$ nicht lösbar zu sein. Ist allerdings $\epsilon = +\infty$ und M kompakt, so ist die Lösbarkeit gesichert. Denn die resultierende Aufgabe $(\mathrm{P}_\infty(x))$ besteht darin, $\nabla f(x)^T p + \frac{1}{2} p^T Bp$ unter der Nebenbedingung $x + p \in M$ zu minimieren. Diese hat wegen der Kompaktheit von M (bzw. der nichtleeren, kompakten Menge $M - x$) eine (allerdings nicht notwendig eindeutige) Lösung. Ist p eine (globale) Lösung von $(\mathrm{P}_\infty(\mathrm{x}))$, so existiert ein $y \in \mathbb{R}^m$ mit

$$y_i \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x) + Bp = \sum_{i=1}^{m} y_i a_i$$

und

$$y_i (a_i^T p + a_i^T x - b_i) = 0 \quad (i = 1, \ldots, m_0).$$

Insbesondere ist

$$\nabla f(x)^T p + p^T Bp = \sum_{i=1}^{m_0} y_i a_i^T p = \sum_{i=1}^{m_0} y_i (b_i - a_i^T x) \leq 0.$$

Daher ist p eine zulässige Richtung mit

$$\nabla f(x)^T p \leq -p^T Bp \leq 0.$$

Ist also $\nabla f(x)^T p \neq 0$, so ist p eine zulässige Abstiegsrichtung. Ist dagegen $\nabla f(x)^T p = 0$, so ist auch $p^T Bp = 0$, folglich $Bp = 0$ und damit (Beweis?) x eine kritische Lösung von (P). Auch in dem Fall, dass B nur positiv semidefinit ist, hat man also eine Möglichkeit, eine zulässige Abstiegsrichtung zu bestimmen, oder festzustellen, dass die aktuelle Näherung eine kritische Lösung ist. □

4.2.4 Konvergenzaussagen

Nun stellt sich naheliegenderweise die Frage, ob die in Lemma 2.4 angegebene Richtungsstrategie, kombiniert mit einer der vorgestellten Schrittweitenstrategien ein konvergentes Verfahren ergibt. Die einfachste Aussage hierzu formulieren wir in dem folgenden Satz.

Satz 2.5 *Gegeben sei die linear restringierte Optimierungsaufgabe* (P)*, die Voraussetzungen* (V) (a)–(c) *seien erfüllt. Sei* $\{B_k\} \subset \mathbb{R}^{n\times n}$ *eine Folge symmetrischer Matrizen, die gleichmäßig positiv definit und beschränkt sei, d. h. es mögen positive Konstanten* μ *und* η *mit*

$$\mu \|p\|^2 \leq p^T B_k p \leq \eta \|p\|^2 \qquad \text{für alle } p \in \mathbb{R}^n,\ k = 0, 1, \ldots,$$

existieren. Mit einem Startwert $x_0 \in \mathbb{R}^n$*, mit dem (V) erfüllt ist, und einem* $\epsilon > 0$ *betrachte man das folgende Verfahren:*

- *Für* $k = 0, 1, \ldots$*:*
  - *Sei* p_k *die Lösung des quadratischen Programms*

    $$\begin{array}{ll} \text{Minimiere} & \nabla f(x_k)^T p + \frac{1}{2} p^T B_k p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \geq b_i - a_i^T x_k \quad (i \in I_\epsilon(x_k)), & a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m). \end{array}$$

  - *Falls* $p_k = 0$*, dann: STOP,* x_k *ist eine kritischee Lösung von* (P).
  - *Berechne* $t_k := t_M(x_k, p_k)$*,* $t_C(x_k, p_k)$*,* $t_P(x_k, p_k)$ *oder* $t_A(x_k, p_k)$*. Hierbei wird vorausgesetzt, dass die für die Powell- bzw. die Armijo-Schrittweite benötigten Konstanten fest vorgegeben sind.*
  - *Setze* $x_{k+1} := x_k + t_k p_k$*.*

Dann gilt: Das Verfahren ist ein durchführbares Verfahren der zulässigen Richtungen. Bricht es nicht schon nach endlich vielen Schritten mit einer stationären Lösung ab, so erzeugt es eine Folge $\{x_k\}$ *mit der Eigenschaft, dass jeder Häufungspunkt* x^* *von* $\{x_k\}$ *eine kritische Lösung von* (P) *ist. Besitzt* (P) *genau eine kritische Lösung* x^* *in der kompakten Niveaumenge* L_0*, so konvergiert die gesamte Folge* $\{x_k\}$ *gegen* x^**.*

Beweis: Wegen Lemma 2.4 ist obiges Verfahren ein durchführbares Verfahren der zulässigen Richtungen, welches bei vorzeitigem Abbruch eine stationäre Lösung von (P) gefunden hat. Wir können daher davon ausgehen, dass das Verfahren eine Folge von Näherungen $\{x_k\} \subset L_0$, eine Folge $\{p_k\}$ von in x_k zulässigen Abstiegsrichtungen p und eine Folge von Schrittweiten $\{t_k\} \subset \mathbb{R}_+$ erzeugt. Der Beweis dafür, dass jeder Häufungspunkt x^* von $\{x_k\}$ eine kritische Lösung von (P) ist, erfolgt in mehreren Schritten.

(a) Die Richtungsfolge $\{p_k\}$ ist beschränkt, d. h. es existiert eine Konstante $c_0 > 0$ mit $\|p_k\| \le c_0$ für $k = 0, 1, \ldots$.

Denn: Wegen Lemma 2.4 und der gleichmäßigen positiven Definitheit der Folge $\{B_k\}$ symmetrischer Matrizen ist

$$0 < \mu \|p_k\|^2 \le p_k^T B_k p_k \le -\nabla f(x_k)^T p_k \le C \|p_k\|$$

für $k = 0, 1, \ldots$ mit einer Konstanten $C > 0$, die etwa so groß gewählt ist, dass $\|\nabla f(x)\| \le C$ für alle $x \in L_0$, was wegen der in (V) (a) vorausgesetzten Kompaktheit der Niveaumenge L_0 sicher möglich ist. Damit ist

$$\|p_k\| \le \frac{C}{\mu} =: c_0 \qquad \text{für } k = 0, 1, \ldots,$$

die Richtungsfolge $\{p_k\}$ ist also beschränkt.

(b) Die Folge $\{s(x_k, p_k)\}$ maximaler Schrittweiten ist durch eine positive Konstante nach unten beschränkt, d. h. es existiert ein $\delta > 0$ derart, dass $s(x_k, p_k) \ge \delta$ für $k = 0, 1, \ldots$. Insbesondere ist auch die Folge $\{\tilde{s}(x_k, p_k)\}$ mit $\tilde{s}(x_k, p_k) := \min(1, s(x_k, p_k)$ nach unten durch eine positive Konstante beschränkt.

Denn: Es ist

$$s(x_k, p_k) = \min\Big\{ \frac{b_i - a_i^T x_k}{a_i^T p_k} : i \in \{1, \ldots, m_0\} \setminus I(x_k),\ a_i^T p_k < 0 \Big\}.$$

Für alle $i \in \{1, \ldots, m\} \setminus I_\epsilon(x_k)$ mit $a_i^T p_k < 0$ ist

$$\frac{b_i - a_i^T x_k}{a_i^T p_k} > -\frac{\epsilon}{a_i^T p_k} \ge \frac{\epsilon}{\|a_i\| \, \|p_k\|} \ge c_1$$

mit einer positiven Konstanten c_1, wobei die in (a) bewiesene Beschränktheit der Richtungsfolge $\{p_k\}$ eingeht. Ist dagegen $i \in I_\epsilon(x_k)$ und $a_i^T p_k < 0$, so folgt

$$\frac{b_i - a_i^T x_k}{a_i^T p_k} \ge 1.$$

Mit $\delta := \min(c_1, 1)$ ist auch (b) bewiesen.

(c) Es ist $\lim_{k\to\infty} \nabla f(x_k)^T p_k = 0$ und $\lim_{k\to\infty} p_k = 0$.

Denn: Wegen (a) existiert eine Konstante $c_0 > 0$ mit $\|p_k\| \le c_0$, wegen (b) existiert eine Konstante $\delta > 0$ mit $\tilde{s}(x_k, p_k) \ge \delta$ für $k = 0, 1, \ldots$. Aus Lemma 2.1 (Minimum- und Curry-Schrittweite),Lemma 2.2 (Powell-Schrittweite) und Lemma 2.3 (Armijo-Schrittweite) erhält man die Existenz einer von k unabhängigen positiven Konstanten θ mit

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\ge \theta \, \min\Big[-\tilde{s}(x_k, p_k) \, \nabla f(x_k)^T p_k, \Big(\frac{\nabla f(x_k)^T p_k}{\|p_k\|} \Big)^2 \Big] \\ &\ge \theta \, \min\Big[-\delta \, \nabla f(x_k)^T p_k, \frac{1}{c_0^2} \, (\nabla f(x_k)^T p_k)^2 \Big]. \end{aligned}$$

Als monoton fallende, nach unten beschränkte Folge ist $\{f(x_k)\}$ konvergent und folglich $\lim_{k\to\infty}(f(x_k)-f(x_{k+1}))=0$. Aus obiger Abschätzung folgt dann auch, wie behauptet, dass $\lim_{k\to\infty}\nabla f(x_k)^T p_k = 0$. Da

$$\mu\,\|p_k\|^2 \le p_k^T B_k p_k \le -\nabla f(x_k)^T p_k$$

mit $\mu>0$, gilt auch $\lim_{k\to\infty} p_k = 0$.

(d) Jeder Häufungspunkt x^* von $\{x_k\}$ ist eine kritische Lösung von (P).

Denn: Sei $x^*\in M$ ein Häufungspunkt von $\{x_k\}$, also Limes einer Teilfolge $\{x_k\}_{k\in K}$ mit einer nicht endlichen Teilmenge $K\subset\mathbb{N}$. Sei $p^*\in F(M;x^*)$ eine beliebige in x^* zulässige Richtung. Wir werden zeigen, dass $\nabla f(x^*)^T p^*\ge 0$ gilt. Damit wird gezeigt sein, dass es in x^* keine zulässige Abstiegsrichtung gibt, bzw. dass x^* eine kritische Lösung von (P) ist.

Nach Konstruktion ist p_k die Lösung von

$$(\mathrm{P}_k)\qquad \begin{cases} \text{Minimiere}\quad \nabla f(x_k)^T p + \frac{1}{2}\,p^T B_k p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \ge b_i - a_i^T x_k \quad (i\in I_\epsilon(x_k)), \qquad a_i^T p = 0 \quad (i=m_0+1,\dots,m). \end{cases}$$

Wir wollen uns überlegen, dass ein $s_0>0$ existiert derart, dass $s_0 p^*$ für alle hinreichend großen $k\in K$ zulässig für das quadratische Programm (P_k) ist, also

$$a_i^T(s_0p^*) \ge b_i - a_i^T x_k \quad (i\in I_\epsilon(x_k)), \qquad a_i^T(s_0p^*) = 0 \quad (i=m_0+1,\dots,m)$$

für alle hinreichend großen $k\in K$ gilt. Nach Definition der Indexmenge $I(x^*)$ der in x^* aktiven Ungleichungsrestriktionen existiert ein $\zeta>0$ mit $a_i^T x^* - b_i \ge \zeta$ für alle $i\in\{1,\dots,m_0\}\setminus I(x^*)$. Für alle hinreichend großen $k\in K$, etwa $k\ge k_0$, ist daher $a_i^T x_k - b_i \ge \frac{1}{2}\zeta$ für alle $i\in\{1,\dots,m_0\}\setminus I(x^*)$. Nun wähle man $s_0>0$ so klein, dass $\frac{1}{2}\,\zeta \ge -a_i^T(s_0p^*)$ für alle $i\in\{1,\dots,m_0\}$ mit $a_i^T p^*<0$. Um nachzuweisen, dass s_0p^* für alle $k\ge k_0$ zulässig für (P_k) ist, nehmen wir $k\in K$ und $k\ge k_0$ an und geben uns ein $i\in I_\epsilon(x_k)$ vor. Für $i\in I(x^*)$ ist $a_i^T p^*\ge 0$, da $p^*\in F(M;x^*)$, und folglich $a_i^T(s_0p^*)\ge 0\ge b_i - a_i^T x_k$. Den selben Schluss können wir machen, wenn $i\in I_\epsilon(x_k)\setminus I(x^*)$ und $a_i^T p^*\ge 0$. Daher können wir jetzt annehmen, es sei $i\in I_\epsilon(x_k)\setminus I(x^*)$ und $a_i^T p^*<0$. Nach Definition von ζ ist dann

$$a_i^T x_k - b_i \ge \tfrac{1}{2}\,\zeta \ge -a_i^T(s_0p^*)$$

sogar für alle $i\in\{1,\dots,m_0\}\setminus I(x^*)$, erst recht also für alle $i\in I_\epsilon(x_k)\setminus I(x^*)$. Für alle hinreichend großen $k\in K$ ist damit s_0p^* zulässig für (P_k). Da $p=0$ trivialerweise zulässig ist, ist aus Konvexitätsgründen sp^* für alle $s\in[0,s_0]$ und alle hinreichend großen $k\in K$ zulässig für (P_k). Da aber p_k die Lösung von (P_k) ist, ist

$$\begin{aligned} \nabla f(x_k)^T p_k + \tfrac{1}{2}\,\mu\,\|p_k\|^2 &\le \nabla f(x_k)^T p_k + \tfrac{1}{2}\,p_k^T B_k p_k \\ &\le s\,\nabla f(x_k)^T p^* + \tfrac{1}{2}\,s^2\,(p^*)^T B_k p^* \\ &\le s\,\nabla f(x_k)^T p^* + \tfrac{1}{2}\,s^2\eta\,\|p^*\|^2 \end{aligned}$$

für alle $s \in [0, s_0]$ und alle hinreichend großen $k \in K$. Mit $k \in K$ und $k \to \infty$ erhält man wegen $\nabla f(x_k)^T p_k \to 0$, $p_k \to 0$ (siehe (c)) und $x_k \to x^*$, dass $0 \leq \nabla f(x^*)^T p^* + \frac{1}{2}\,\mu s\,\|p^*\|^2$ für alle $s \in (0, s_0]$. Mit $s \to 0+$ folgt $\nabla f(x^*)^T p^* \geq 0$, womit schließlich auch (d) bewiesen ist.

(e) Besitzt (P) genau eine kritische Lösung x^* in der Niveaumenge L_0, so konvergiert die gesamte Folge $\{x_k\}$ gegen x^*.

Denn: Angenommen, $\{x_k\}$ würde nicht gegen x^* konvergieren. Dann existiert eine unendliche Teilmenge $K \subset \mathbb{N}$ und ein $\delta > 0$ mit $\|x_k - x^*\| \geq \delta$ für alle $k \in K$. Aus $\{x_k\}_{k\in K} \subset L_0$ kann eine gegen ein $x^{**} \in L_0$ konvergente Teilfolge ausgewählt werden. Dann ist auch x^{**} ein Häufungspunkt von $\{x_k\}$ und damit nach (d) eine kritische Lösung von (P). Da aber $\|x^{**} - x^*\| \geq \delta$ ergibt sich ein Widerspruch zur Voraussetzung, dass (P) genau eine kritische Lösung in L_0 besitzt. □ □

Bemerkung: Natürlich erscheint es wünschenswert zu sein, dass die Anzahl der Restriktionen des in jedem Schritt zu lösenden quadratischen Hilfsproblems möglichst klein ist. Wird $\epsilon = 0$ gewählt, so lautet das entsprechende quadratische Programm

$$\text{Minimiere} \quad \nabla f(x)^T p + \tfrac{1}{2}\, p^T B p, \quad p \in F(M; x).$$

Durch ein Beispiel von P. Wolfe kann man zeigen, dass das Verfahren aus Satz 2.5 mit $\epsilon := 0$, $B_k := I$ für alle k und der exakten Schrittweite eine Folge $\{x_k\}$ liefern kann, welche gegen einen Punkt konvergiert, welcher keine kritische Lösung von (P) ist. Die Konvergenz wird hier verhindert durch das sogenannte „Zigzagging“. Wünschenswert wäre es, dass $I(x^k)$ nach endlich vielen Schritten konstant ist, dass also nach endlich vielen Schritten die richtige Indexmenge aktiver Ungleichungsrestriktionen gefunden ist. Genau das ist in dem Beispiel nicht der Fall. □

Bemerkung: Eine zu Satz 2.5 ganz entsprechende Konvergenzaussage kann zum Verfahren von Frank-Wolfe gamacht werden. Dieses unterscheidet sich von dem obigen Vefahren nur darin, dass die Richtung p_k eine Lösung des linearen Programms

$$(\mathrm{P}_k) \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad \nabla f(x_k)^T p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \geq b_i - a_i^T x_k \quad (i = 1, \ldots, m_0), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m) \end{array} \right.$$

ist, und die Abbruchbedingung durch $\nabla f(x_k)^T p_k = 0$ gegeben ist (siehe Aufgabe 6). □

Nun wollen wir noch eine Aussage über die Konvergenzgeschwindigkeit des in Satz 2.5 angegebenen Verfahrens machen. Wir werden hierzu voraussetzen, dass das Verfahren eine Folge $\{x_k\}$ liefert, welche gegen einen Punkt x^* konvergiert, der einerseits kritisch ist, für den andererseits die Hessesche $\nabla^2 f(x^*)$ (symmetrisch und) positiv definit ist. Nicht verschwiegen werden soll, dass dies eine unangemessen starke Voraussetzung ist. Angemessen wäre die Voraussetzung, dass in x^* die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt sind.

Satz 2.6 *Die Voraussetzungen* (V) (a)–(c) *seien erfüllt. Man betrachte das Verfahren aus Satz 2.5, bei dem in jedem Schritt* $t_k = t_A(x_k, p_k)$ *die Armijo-Schrittweite und* $\{B_k\}$

eine Folge symmetrischer, gleichmäßig positiv definiter Matrizen ist. Das Verfahren liefere eine Folge $\{x_k\}$, die gegen eine kritische Lösung x^* von (P) *konvergent ist, und eine Richtungsfolge $\{p_k\}$. Die Zielfunktion f sei auf einer Umgebung von x^* zweimal stetig differenzierbar, die Hessesche $\nabla^2 f(\cdot)$ sei auf dieser Umgebung lipschitzstetig, $\nabla^2 f(x^*)$ sei positiv definit. Ferner gelte*

$$(*) \qquad \lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x_k)]p_k\|}{\|p_k\|} = 0.$$

Dann gilt:

(i) Es ist $t_k = t_A(x_k, p_k) = 1$ für alle hinreichend großen k.

(ii) Die Folge $\{x_k\}$ konvergiert superlinear gegen x^*.

(iii) Ist $B_k = \nabla^2 f(x_k)$, so konvergiert $\{x_k\}$ sogar quadratisch gegen x^*.

Beweis: Im Beweis von Satz 2.5 wurde gezeigt, dass die Richtungsfolge $\{p_k\}$ gegen den Nullvektor konvergiert. Hieraus folgt aber, dass $s(x_k, p_k) \geq 1$ für alle hinreichend großen k ist. Daher ist in (i) zu zeigen, dass

$$\frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} \geq \alpha$$

für alle hinreichend großen k gilt, wobei $\alpha \in (0, \frac{1}{2})$ vorgegeben ist. Nun existiert eine konvexe Umgebung U^* von x^*, auf der f zweimal stetig partiell differenzierbar ist, zu der es ferner ein $\hat{\mu} > 0$ mit

$$\hat{\mu}\,\|p\|^2 \leq p^T \nabla^2 f(x) p \qquad \text{für alle } x \in U^*,\ p \in \mathbb{R}^n$$

gibt. Schließlich kann U^* auch gleich noch so klein gewählt werden, dass $\nabla^2 f(\cdot)$ auf U^* lipschitzstetig mit einer Lipschitzkonstanten $L > 0$ ist. Für alle hinreichend großen k sind x_k und $x_k + p_k$ in U^* enthalten. Damit ist

$$\begin{aligned}
\frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} &= 1 + \frac{1}{2}\,\frac{p_k^T \nabla^2 f(x_k + \theta_k p_k) p_k}{\nabla f(x_k)^T p_k} \\
&\qquad (\text{mit } \theta_k \in (0,1)) \\
&= \frac{1}{2} + \frac{1}{2}\,\frac{p_k^T \nabla^2 f(x_k + \theta_k p_k) p_k + \nabla f(x_k)^T p_k}{\nabla f(x_k)^T p_k} \\
&\geq \frac{1}{2} - \frac{1}{2}\,\frac{p_k^T [\nabla^2 f(x_k + \theta_k p_k) - B_k] p_k}{p_k^T B_k p_k} \\
&\qquad (\text{wegen } p_k^T B_k p_k \leq -\nabla f(x_k)^T p_k) \\
&\geq \frac{1}{2} - \frac{1}{2\mu}\,\frac{\|[\nabla^2 f(x_k + \theta_k p_k) - B_k] p_k\|}{\|p_k\|} \\
&\qquad (\text{wegen } \mu\,\|p_k\|^2 \leq p_k^T B_k p_k)
\end{aligned}$$

$$\geq \frac{1}{2} - \frac{1}{2\mu}\Big(\underbrace{\|\nabla^2 f(x_k+\theta_k p_k) - \nabla^2 f(x_k)\|}_{\to 0} + \underbrace{\frac{\|[\nabla^2 f(x_k) - B_k]p_k\|}{\|p_k\|}}_{\to 0}\Big).$$

Wegen $\alpha \in (0, \frac{1}{2})$ ist daher

$$\frac{f(x_k+p_k)-f(x_k)}{\nabla f(x_k)^T p_k} \geq \alpha$$

und folglich $t_A(x_k, p_k) = 1$ für alle hinreichend großen k. Damit ist (i) bewiesen.

Für alle hinreichend großen k ist

$$\begin{aligned}
\hat{\mu}\,\|x_{k+1}-x^*\|^2 &\leq (x_{k+1}-x^*)^T\nabla^2 f(x_k)(x_{k+1}-x^*)\\
&= (p_k + x_k - x^*)^T\nabla^2 f(x_k)(x_{k+1}-x^*)\\
&= (\nabla^2 f(x_k)p_k)^T(x_{k+1}-x^*) + (x_k-x^*)^T\nabla^2 f(x_k)(x_{k+1}-x^*)\\
&= [(\nabla^2 f(x_k) - B_k)p_k]^T(x_{k+1}-x^*) + (B_k p_k)^T(x_{k+1}-x^*)\\
&\quad - [\nabla^2 f(x_k)(x^*-x_k)]^T(x_{k+1}-x^*).
\end{aligned}$$

Da p_k Lösung des quadratischen Hilfsproblems (P_k) ist, existieren $y_i^{(k)}$ für $i \in I_\epsilon(x_k) \cup \{m_0+1,\ldots,m\}$ mit

$$y_i^{(k)} \geq 0 \quad (i \in I_\epsilon(x_k)), \qquad \nabla f(x_k) + B_k p_k = \sum_{i\in I_\epsilon(x_k)} y_i^{(k)} a_i + \sum_{i=m_0+1}^{m} y_i^{(k)} a_i$$

sowie

$$y_i^{(k)}(a_i^T p_k + a_i^T x_k - b_i) = 0 \quad (i \in I_\epsilon(x_k)).$$

Daher ist

$$\begin{aligned}
(B_k p_k)^T(x_{k+1}-x^*) &= -\nabla f(x_k)^T(x_{k+1}-x^*) + \sum_{i\in I_\epsilon(x_k)} y_i^{(k)} a_i^T(x_{k+1}-x^*)\\
&= -\nabla f(x_k)^T(x_{k+1}-x^*) + \sum_{i\in I_\epsilon(x_k)} y_i^{(k)} a_i^T(x_k+p_k-x^*)\\
&= -\nabla f(x_k)^T p_k + \sum_{i\in I_\epsilon(x_k)} \underbrace{y_i^{(k)}(b_i - a_i^T x^*)}_{\leq 0}\\
&\leq -\nabla f(x_k)^T(x_{k+1}-x^*).
\end{aligned}$$

Damit erhalten wir

$$\begin{aligned}
\hat{\mu}\,\|x_{k+1}-x^*\|^2 &\leq [(\nabla^2 f(x_k) - B_k)p_k]^T(x_{k+1}-x^*) - \nabla f(x_k)^T(x_{k+1}-x^*)\\
&\quad - [\nabla^2 f(x_k)(x^*-x_k)]^T(x_{k+1}-x^*)\\
&\leq [(\nabla^2 f(x_k) - B_k)p_k]^T(x_{k+1}-x^*)\\
&\quad + [\nabla f(x^*) - \nabla f(x_k) - \nabla^2 f(x_k)(x^*-x_k)]^T(x_{k+1}-x^*)\\
&\quad (\text{wegen } \nabla f(x^*)^T(x_{k+1}-x^*) \geq 0)
\end{aligned}$$

und nach Anwendung der Cauchy-Schwarzschen Ungleichung

$$\hat{\mu}\,\|x_{k+1}-x^*\| \le \|[\nabla^2 f(x_k)-B_k]p_k\| + \|\nabla f(x^*)-\nabla f(x_k)-\nabla^2 f(x_k)(x^*-x_k)\|.$$

Ist daher $B_k = \nabla^2 f(x_k)$, handelt es sich bei dem Verfahren aus Satz 2.5 also um das Newton-Verfahren, so ist für alle hinreichend großen k daher

$$\begin{aligned}
\hat{\mu}\,\|x_{k+1}-x^*\| &\le \|\nabla f(x^*)-\nabla f(x_k)-\nabla^2 f(x_k)(x^*-x_k)\| \\
&= \Big\|\int_0^1 [\nabla^2 f(x_k+t(x^*-x_k))-\nabla f(x_k)](x^*-x_k)\,dt\Big\| \\
&\le \int_0^1 \|\nabla f(x_k+t(x^*-x_k))-\nabla f(x_k)\|\,dt\,\|x_k-x^*\| \\
&\le \frac{L}{2}\,\|x_k-x^*\|^2,
\end{aligned}$$

da $\nabla^2 f(\cdot)$ auf einer Umgebung von x^* lipschitzstetig (mit einer Lipschitzkonstanten $L>0$) ist. Also ist das Verfahren in diesem Fall tatsächlich quadratisch konvergent. Andernfalls wird

$$\lim_{k\to\infty} \frac{\|[\nabla^2 f(x_k)-B_k]p_k\|}{\|p_k\|} = 0$$

vorausgesetzt und man erhält wegen

$$\|p_k\| = \|(x_{k+1}-x^*)+(x^*-x_k)\| \le \|x_k-x^*\| + \|x_{k+1}-x^*\|,$$

dass

$$\hat{\mu}\,\frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|} \le \frac{\|[\nabla^2 f(x_k)-B_k]p_k\|}{\|p_k\|}\Big(1+\frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|}\Big) + \frac{L}{2}\,\|x_k-x^*\|,$$

woraus wegen $\lim_{k\to\infty}\|[\nabla^2 f(x_k)-B_k]p_k\|/\,\|p_k\| = 0$ und $\lim_{k\to\infty}\|x_k-x^*\| = 0$ die superlineare Konvergenz der Folge $\{x_k\}$ gegen x^* folgt. □ □

Bemerkung: Natürlich wird man sich fragen, wie die Folge symmetrischer, postiv definiter Matrizen $\{B_k\} \subset \mathbb{R}^{n\times n}$ gewählt werden sollte, um unter möglichst schwachen Voraussetzungen lokal superlineare Konvergenz zu sichern. Nach Konstruktion sollte B_k eine Approximation an $\nabla^2 f(x_k)$ sein. Da man meistens nicht bereit sein wird, die Hessesche der Zielfunktion zu berechnen, ist man auf Quasi-Newton-Verfahren angewiesen. Bei der BFGS-Update-Formel ist z. B.

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$

mit $s_k := y_{k+1} - y_k$ und $y_k := \nabla f(x_{k+1}) - \nabla f(x_k)$. Nicht verschwiegen werden sollte aber, dass für diese Wahl selbst bei gleichmäßig konvexer Zielfunktion nicht die Konvergenz oder gar die superlineare Konvergenz gezeigt werden konnte. Außerdem ist zu beachten, dass jeder Iterationsschritt relativ "teuer" ist und Informationen über den aktuellen Schritt offenbar nur schwer oder gar nicht effizient für den nächsten herangezogen werden können. □

4.2.5 Aufgaben

1. Gegeben sei eine linear restringierte nichtlineare Optimierungsaufgabe mit einer stetig differenzierbaren Zielfunktion. Man zeige, dass eine zulässige Lösung genau dann eine kritische Lösung ist, wenn es in ihr keine zulässige Abstiegsrichtung gibt.

2. Gegeben sei die Optimierungsaufgabe

   (P) $$\text{Minimiere} \quad f(x) \quad \text{auf} \quad M,$$

   wobei $M \subset \mathbb{R}^n$ konvex ist. Sei $x^* \in M$ und die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in x^* stetig differenzierbar. Man zeige:

   (a) Ist x^* eine lokale Lösung von (P), so ist $\nabla f(x^*)^T(x - x^*) \geq 0$ für alle $x \in M$.

   (b) Sei
   $$M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$
   Dann ist $\nabla f(x^*)^T(x - x^*) \geq 0$ für alle $x \in M$ genau dann, wenn x^* eine kritische Lösung von (P) ist, also ein $y^* \in \mathbb{R}^m$ mit
   $$y_i^* \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x^*) = A^T y^*, \qquad (y^*)^T(Ax^* - b) = 0$$
   existiert. Hierbei ist, wie stets in diesem Zusammenhang, $A \in \mathbb{R}^{m \times n}$ die Matrix, die a_i^T als i-te Zeile besitzt, ferner ist b_i die i-te Komponente von $b \in \mathbb{R}^m$.

3. Man zeige: Genügt die Zielfunktion f von (P) den Voraussetzungen (V) (a)–(c), so existiert eine Konstante $\theta_C > 0$ derart, dass
   $$\begin{aligned} f(x) - f(x + t_M(x,p)p) &\geq f(x) - f(x + t_C(x,p)p) \\ &\geq \theta_C \, \min\left[-s(x,p)\, \nabla f(x)^T p, \left(\frac{\nabla f(x)^T p}{\|p\|} \right)^2 \right] \end{aligned}$$
   für alle nicht kritischen $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$. Hierbei bedeutet $t_M = t_M(x,p)$ die Minimum-Schrittweite, $t_C = t_C(x,p)$ die Curry-Schrittweite und $s = s(x,p)$ die maximale Schrittweite in x in Richtung p, ferner $\|\cdot\|$ die euklidische Norm.

4. Man zeige: Genügt die Zielfunktion f von (P) den Voraussetzungen (V) (a)–(c), so existiert eine Konstante $\theta_P > 0$ derart, dass
   $$f(x) - f(x + t_P(x,p))p) \geq \theta_C \, \min\left[-s(x,p)\, \nabla f(x)^T p, \left(\frac{\nabla f(x)^T p}{\|p\|} \right)^2 \right]$$
   für alle nicht kritischen $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$. Hierbei bedeutet $t_P(x,p)$ die Powell-Schrittweite und $s(x,p)$ die maximale Schrittweite in x in Richtung p, ferner $\|\cdot\|$ die euklidische Norm.

5. Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Dann existiert eine Konstante $\theta_A > 0$ derart, dass
   $$f(x) - f(x + t_A(x,p)p) \geq \theta_A \, \min\left[-\tilde{s}(x,p)\, \nabla f(x)^T p, \left(\frac{\nabla f(x)^T p}{\|p\|} \right)^2 \right]$$

für alle nicht kritischen $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$. Hierbei bedeutet $t_A(x,p)$ die Armijo-Schrittweite und $\tilde{s}(x,p) := \min(s(x,p),1)$ die eventuell reduzierte maximale Schrittweite, ferner $\|\cdot\|$ die euklidische Norm.

6. Gegeben sei das linear restringierte Programm

$$\text{(P)} \quad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0+1, \ldots, m) \end{array} \right\}.$$

Die Menge der zulässigen Lösungen M sei nichtleer und kompakt, ferner seien die üblichen Voraussetzungen (V) (a)–(c) erfüllt. Man betrachte das Verfahren von Frank-Wolfe:

- Für $k = 0, 1, \ldots$:
  - Sei p_k eine Lösung des linearen Programms

$$\left\{ \begin{array}{c} \text{Minimiere} \quad \nabla f(x_k)^T p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \geq b_i - a_i^T x_k \quad (i = 1, \ldots, m_0), \qquad a_i^T p = 0 \quad (i = m_0+1, \ldots, m). \end{array} \right.$$

  - Falls $\nabla f(x_k)^T p_k = 0$, dann: STOP, x_k ist kritische Lösung von (P).
  - Berechne $t_k := t_M(x_k, p_k)$, $t_C(x_k, p_k)$, $t_P(x_k, p_k)$ oder $t_A(x_k, p_k)$.
  - Setze $x_{k+1} := x_k + t_k p_k$.

Dann gilt: Bricht das Verfahren nicht vorzeitig mit einer kritischen Lösung von (P) ab, so liefert es eine Folge $\{x_k\}$ mit der Eigenschaft, dass jeder Häufungspunkt von $\{x_k\}$ eine kritische Lösung von (P) ist.

7. Gegeben sei das linear restringierte Programm

$$\text{(P)} \quad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0+1, \ldots, m) \end{array} \right\}.$$

Sei $x \in M$ eine aktuelle Näherung, in der die Zielfunktion f von (P) stetig differenzierbar ist, und $B \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit. Hiermit betrachte man das quadratische Hilfsproblem

$$\text{(P}(x)\text{)} \qquad \left\{ \begin{array}{l} \text{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T B p \quad \text{unter den Nebenbedingungen} \\ \begin{array}{rcll} a_i^T p & \geq & b_i - a_i^T x & (i = 1, \ldots, m_0), \\ a_i^T p & = & 0 & (i = m_0+1, \ldots, m), \end{array} \qquad \|p\|_\infty \leq 1. \end{array} \right.$$

Sei p^* eine Lösung von (P(x)). Man zeige: Ist $\nabla f(x)^T p^* = 0$, so ist x eine kritische Lösung von (P), andernfalls ist p^* eine zulässige Abstiegsrichtung in x.

Hinweis: Man wende den Satz von Kuhn-Tucker auf das Hilfsproblem (P(x)) an, wobei die Restriktion $\|p\|_\infty \leq 1$ durch die beiden linearen Ungleichungsrestriktionen $-e \leq p \leq e$ (wobei e einmal wieder der Vektor ist, dessen Komponenten alle gleich 1 sind) ersetzt wird.

8. Gegeben sei das linear restringierte Programm

$$\text{(P)} \quad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0+1, \ldots, m) \end{array} \right\}.$$

Sei $x \in M$ eine aktuelle Näherung, in der die Zielfunktion f von (P) stetig differenzierbar ist, und $B \in \mathbb{R}^{n\times n}$ symmetrisch (aber nicht notwendig positiv semidefinit). Mit einem $\Delta > 0$ betrachte man das Hilfsproblem

$$(\mathrm{P}_{x,\Delta}) \quad \left\{ \begin{array}{c} \text{Minimiere} \quad \phi_x(p) := \nabla f(x)^T p + \frac{1}{2} p^T B p \quad \text{unter den Nebenbedingungen} \\ x + p \in M, \qquad \|p\| \le \Delta, \end{array} \right.$$

wobei $\|\cdot\|$ eine beliebige Norm auf dem $\mathbb{R}^n$ ist. Dann gilt: Ist $\min(\mathrm{P}_{x,\Delta}) = 0$, also $p^* := 0$ eine Lösung von $(\mathrm{P}_{x,\Delta})$, so ist $x \in M$ eine kritische Lösung von (P).

9. Gegeben sei die linear restringierte Optimierungsaufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Ax \le b\}$$

mit

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix} \in \mathbb{R}^{m\times n}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m$$

und stetig differenzierbarer Zielfunktion f. Sei $x \in M$ eine zulässige Lösung, ferner $I := I(x)$ die Indexmenge der in x aktiven Restriktionen. Die Matrix $A_I \in \mathbb{R}^{\#(I)\times n}$ sei in naheliegender Weise definiert, sie habe vollen Rang, d. h. $\{a_i\}_{i\in I}$ seien linear unabhängig. Schließlich sei

$$P := I - A_I^T (A_I A_I^T)^{-1} A_I$$

(eine Verwechslung der Einheitsmatrix I und der Indexmenge I ist extrem unwahrscheinlich). Man zeige:

(a) Ist $p := -P\nabla f(x) \neq 0$, so ist p eine zulässige Abstiegsrichtung in x.

(b) Ist $P\nabla f(x) = 0$ und $y := -(A_I A_I^T)^{-1} A_I \nabla f(x) \ge 0$, so ist x eine kritische Lösung von (P).

(c) Ist $P\nabla f(x) = 0$ und $y := -(A_I A_I^T)^{-1} A_I \nabla f(x) \not\ge 0$, ist ferner $l \in I$ ein Index mit $y_l < 0$, so setze man $I := I \setminus \{l\}$ und

$$P := I - A_I^T (A_I A_I^T)^{-1} A_I.$$

Dann ist $p := -P\nabla f(x)$ eine zulässige Abstiegsrichtung in x.

10. Sei $M \subset \mathbb{R}^n$ nichtleer, konvex und abgeschlossen (z. B. sei M ein Polyeder) und $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf einer offenen Obermenge von M stetig differenzierbar. Wir nennen $x \in M$ eine *kritische Lösung* von (P), wenn $\nabla f(x)^T (z - x) \ge 0$ für alle $z \in M$, also die notwendige Optimalitätsbedingung erster Ordnung erfüllt ist. Mit $P_M\colon \mathbb{R}^n \longrightarrow M$ sei die Projektionsabbildung auf M bezüglich der euklidischen Norm $\|\cdot\|$ bezeichnet. Sei $x \in M$ keine stationäre Lösung der Aufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(z), \quad z \in M,$$

und $x(t) := P_M(x - t\nabla f(x))$. Man zeige:

(a) Es ist $x \neq x(t)$ für alle $t > 0$.

(b) Es ist

$$\lim_{t\to 0+} \frac{f(x) - f(x(t))}{\nabla f(x)^T (x - x(t))} = 1.$$

(c) Es ist $f(x(t)) < f(x)$ für alle hinreichend kleinen $t > 0$.

Kapitel 5

Nichtlinear restringierte Optimierungsaufgaben

In diesem Kapitel werden Verfahren zur Lösung der nichtlinear restringierten Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$$

entwickelt und analysiert. Wir werden voraussetzen, dass die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sowie die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ glatt, also mindestens einmal stetig differenzierbar sind. Gelegentlich werden wir nur nichtlineare Gleichungen als Restriktionen betrachten. Dies ist zumindestens theoretisch keine Einschränkung, denn die Ungleichungsrestriktion $g_i(x) \leq 0$ ist äquivalent zu $g_i(x)+y_i^2 = 0$. Mit Hilfe von l (nichtlinear auftretenden) Schlupfvariablen können also die l Ungleichungsrestriktionen in Gleichungen überführt werden. I. allg. dürfte dies für die Praxis aber kein adäquater Zugang sein. Verfahren der zulässigen Richtungen sind zumindestens bei nichtlinearen Gleichungen als Nebenbedingungen nicht praktikabel, u. a. da die Zulässigkeit der Näherungslösungen zu bewahren den selben Schwierigkeitsgrad wie das Lösen nichtlinearer Gleichungssysteme besitzt. Auch wenn z. B. bei konvexen, quadratischen Ungleichungsrestriktionen Verfahren der zulässigen Richtungen durchaus möglich sind, werden wir auf diese in diesem Kapitel nicht mehr eingehen.

5.1 Straffunktionen

5.1.1 Differenzierbare Straffunktionen

Eine naheliegende Idee besteht darin, dass statt der restringierten Aufgabe (P) eine Folge unrestringierter Optimierungsaufgaben gelöst wird, wobei die Verletzung der gegebenen Restriktionen zunehmend härter bestraft wird. Wir wollen diese simple Idee bei durch nichtlineare Gleichungen restringierte Optimierungsaufgaben ein wenig genauer untersuchen (siehe R. Fletcher (1987, S. 277 ff.)). Gegeben sei also die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : h(x) = 0\}.$$

Mit einem $\sigma > 0$ wird dieser Aufgabe die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \tfrac{1}{2}\sigma \, \|h(x)\|^2, \quad x \in \mathbb{R}^n,$$

wobei $\|\cdot\|$ natürlich die euklidische Norm auf dem $\mathbb{R}^m$ bedeutet, zugeordnet. Die Zielfunktion $\Phi_\sigma(\cdot)$ von (P_σ) heißt eine (quadratische) *Penalty-Funktion* oder auch *Straffunktion*, da sie das Verletztsein der Nebenbedingung durch erhöhte Kosten bestraft. Genauer ist $\Phi_\sigma(x) = f(x)$ für alle $x \in M$, während für $x \notin M$ offenbar $\Phi_\sigma(x) \to +\infty$ mit $\sigma \to \infty$. Man hofft, dass man mit wachsendem σ (globale, lokale, stationäre) Lösungen von (P) durch Lösungen von (P_σ) approximieren kann. Ein ganz primitives Penalty-Verfahren könnte dann folgendermaßen aussehen:

- Wähle $\sigma_0 > 0$.
- Für $k = 0, 1, \ldots$:
  - Bestimme eine (globale, lokale, stationäre) Lösung $x(\sigma_k)$ von (P_{σ_k}).
  - Wähle $\sigma_{k+1} > \sigma_k$, z. B. $\sigma_{k+1} := 10\sigma_k$.

Beispiel: Betrachte die Aufgabe

$$(\mathrm{P}) \quad \text{Minimiere} \quad f(x) := -x_1 - x_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^2 : h(x) := 1 - x_1^2 - x_2^2 = 0\}.$$

Die Lösung x^* und den zugehörigen Lagrange-Multiplikator erhält man sehr leicht aus den notwendigen Bedingungen erster Ordnung. Ist x^* eine lokale Lösung, so ist $\nabla h(x^*) \neq 0$, die Arrow-Hurwicz-Uzawa Constraint Qualification also erfüllt. Daher existiert ein y^* mit $\nabla f(x^*) + y^* \nabla h(x^*) = 0$. Zusammen mit $h(x^*) = 0$ ergibt dies ein nichtlineares Gleichungssystem für (x^*, y^*), als Lösung der gegebenen Aufgabe (P) erhält man $x^* = (1/\sqrt{2}, 1/\sqrt{2})^T$. Mit Hilfe von

$$\nabla \Phi_\sigma(x) = \begin{pmatrix} -1 \\ -1 \end{pmatrix} - \sigma \begin{pmatrix} (1 - x_1^2 - x_2^2)x_1 \\ (1 - x_1^2 - x_2^2)x_2 \end{pmatrix}$$

erhält man aus $\nabla \Phi_\sigma(x(\sigma)) = 0$, dass $x_1(\sigma) = x_2(\sigma)$ als Lösung von $(1 - 2x^2)x = -1/\sigma$ zu bestimmen ist, was bei gegebenem $\sigma > 0$ zumindestens numerisch leicht möglich ist. Bei R. Fletcher (1987, S. 280) findet man einige numerische Ergebnisse. □

Im folgenden Satz nehmen wir an (ohne es genau vorauszusetzen, siehe auch Theorem 12.1.1 bei R. Fletcher (1987, S. 281)), die Aufgabe (P_σ) besitze für jedes $\sigma > 0$ eine globale Lösung $x(\sigma)$, ferner sei (P) zulässig.

Satz 1.1 *Sei $0 < \sigma \leq \tau$. Dann ist*

$$\Phi_\sigma(x(\sigma)) \leq \Phi_\tau(x(\tau)), \qquad \|h(x(\sigma))\|^2 \geq \|h(x(\tau))\|^2, \qquad f(x(\sigma)) \leq f(x(\tau)).$$

Ist $\{\sigma_k\}$ monoton wachsend und $\sigma_k \to \infty$, so gilt $\lim_{k\to\infty} h(x(\sigma_k)) = 0$, ferner ist jeder Häufungspunkt x^* von $\{x(\sigma_k)\}$ eine Lösung von (P).

Beweis: Sei $0 < \sigma \leq \tau$. Dann ist

$$\Phi_\sigma(x(\sigma)) \leq \Phi_\sigma(x(\tau)) \leq \Phi_\tau(x(\tau)),$$

womit die erste Behauptung bewiesen ist. Wegen

$$\Phi_\tau(x(\tau)) \leq \Phi_\tau(x(\sigma))$$

ist

$$\begin{aligned} \tfrac{1}{2}(\tau-\sigma)[\|h(x(\sigma))\|^2 - \|h(x(\tau))\|^2] &= \underbrace{\Phi_\tau(x(\sigma)) - \Phi_\tau(x(\tau))}_{\geq 0} \\ &\quad + \underbrace{\Phi_\sigma(x(\tau)) - \Phi_\sigma(x(\sigma))}_{\geq 0} \\ &\geq 0, \end{aligned}$$

woraus die zweite Behauptung folgt. Dann ist aber

$$f(x(\tau)) - f(x(\sigma)) = \underbrace{\Phi_\sigma(x(\tau)) - \Phi_\sigma(x(\sigma))}_{\geq 0} + \tfrac{1}{2}\sigma[\underbrace{\|h(x(\sigma))\|^2 - \|h(x(\tau))\|^2}_{\geq 0}] \geq 0,$$

womit auch die dritte Behauptung bewiesen ist.

Nach Definition von $x(\sigma)$ ist

$$\Phi_\sigma(x(\sigma)) \leq \inf_{x\in M} \Phi_\sigma(x) = \inf_{x\in M} f(x) = \inf(\mathrm{P}).$$

Als monoton fallende (bzw. genauer: monoton nicht wachsende), nach unten beschränkte Folge ist $\{\|h(x(\sigma_k))\|\}$ konvergent. Angenommen, es sei $c := \lim_{k\to\infty} \|h(x(\sigma_k))\| > 0$. Dann wäre

$$\begin{aligned} \inf(\mathrm{P}) &\geq \Phi_{\sigma_k}(x(\sigma_k)) \\ &= f(x(\sigma_k)) + \tfrac{1}{2}\sigma_k \|h(x(\sigma_k))\|^2 \\ &\geq f(x(\sigma_k)) + \tfrac{1}{2}\sigma_k c^2 \\ &\geq f(x(\sigma_0)) + \tfrac{1}{2}\sigma_k c^2 \\ &\to \infty, \end{aligned}$$

ein Widerspruch. Ist schließlich x^* ein Häufungspunkt der Folge $\{x(\sigma_k)\}$, so ist $h(x^*) = 0$ bzw. $x^* \in M$ wegen $\lim_{k\to\infty} h(x(\sigma_k)) = 0$. Daher ist $f(x^*) \geq \inf(\mathrm{P})$. Andererseits ist

$$f(x(\sigma_k)) \leq \Phi_{\sigma_k}(x(\sigma_k), \sigma_k) \leq \inf(\mathrm{P})$$

und folglich $f(x^*) \leq \inf(\mathrm{P})$. Insgesamt haben wir gezeigt, dass $x^* \in M$ und $f(x^*) = \inf(\mathrm{P})$ gilt bzw. x^* eine Lösung von (P) ist. Der Satz ist damit bewiesen. □ □

Im letzten Satz wurde (ohne Differenzierbarkeitsbedingungen an die Zielfunktion f oder die Restriktionsabbildung h sowie ohne Regularitätsbedingungen an den Häufungspunkt x^* der Folge $\{x(\sigma_k)\}$) vorausgesetzt, dass bei gegebenem $\sigma > 0$ eine globale Lösung $x(\sigma)$ der unrestringierten Optimierungsaufgabe (P_σ) existiert. Das ist im folgenden Satz (siehe R. Fletcher (1987, S. 282)) anders.

Satz 1.2 *Sei $\{\sigma_k\}$ eine Folge positiver Zahlen mit $\sigma_k \to \infty$, $x_k := x(\sigma_k)$ eine lokale Lösung der unrestringierten Optimierungsaufgabe*

$$(\mathrm{P}_k) \qquad \textit{Minimiere} \quad \Phi_k(x) := f(x) + \tfrac{1}{2}\sigma_k \|h(x)\|^2, \quad x \in \mathbb{R}^n,$$

und $x_k \to x^*$.

(a) Sind f und h auf einer Umgebung von x^* stetig partiell differenzierbar und Rang $(h'(x^*)) = m$, *so ist x^* eine kritische (oder auch stationäre) Lösung von (P), d. h. es ist $h(x^*) = 0$ und es existiert ein $y^* \in \mathbb{R}^m$ mit*

$$\nabla f(x^*) + h'(x^*)^T y^* = 0.$$

Mit $y_k := \sigma_k h(x_k)$ gilt $y_k \to y^*$, ferner ist $\Phi_k(x_k) \to f(x^*)$. Genauer ist

$$h(x_k) = y^*/\sigma_k + o(1/\sigma_k), \qquad \sigma_k \|h(x_k)\|^2 = \|y^*\|^2/\sigma_k + o(1/\sigma_k),$$

wobei wir $g_k = o(1/\sigma_k)$ schreiben, wenn $\sigma_k g_k \to 0$.

(b) Seien f und h auf einer Umgebung von x^* zweimal stetig partiell differenzierbar, wieder sei Rang $(h'(x^*)) = m$. *In x^* sei die hinreichende Optimalitätsbedingung zweiter Ordnung erfüllt, es existiere also ein $y^* \in \mathbb{R}^m$ mit $\nabla f(x^*) + h'(x^*)^T y^* = 0$ und der Eigenschaft, dass*

$$W^* := \nabla^2 f(x^*) + \sum_{j=1}^{m} y_j^* \nabla^2 h_j(x^*)$$

auf Kern $(h'(x^*))$ *positiv definit ist. Dann ist*

$$f(x^*) = \Phi_k(x_k) + \tfrac{1}{2}\sigma_k \|h(x_k)\|^2 + o(1/\sigma_k)$$

und

$$x_k - x^* = (T^*)^T y^*/\sigma_k + o(1/\sigma_k),$$

wobei $T^* \in \mathbb{R}^{m\times n}$ durch

$$\begin{pmatrix} W^* & h'(x^*)^T \\ h'(x^*) & 0 \end{pmatrix}^{-1} = \begin{pmatrix} H^* & (T^*)^T \\ T^* & U^* \end{pmatrix}$$

gegeben ist.

Beweis: Da x_k als lokale Lösung von (P_k) insbesondere eine kritische Lösung von (P_k) ist, ist

$$\nabla \Phi_k(x_k) = \nabla f(x_k) + \sigma_k h'(x_k)^T h(x_k) = 0.$$

Mit $y_k := \sigma_k h(x_k)$ ist daher

$$(*) \qquad \nabla f(x_k) + h'(x_k)^T y_k = 0.$$

Da Rang $(h'(x^*)) = m$ und $x_k \to x^*$ ist auch Rang $(h'(x_k)) = m$ für alle hinreichend großen k (Beweis?) und folglich

$$y_k = -[h'(x_k)h'(x_k)^T]^{-1}h'(x_k)\nabla f(x_k) \to -[h'(x^*)h'(x^*)^T]^{-1}h'(x^*)\nabla f(x^*) =: y^*.$$

Aus $(*)$ folgt mit $k \to \infty$, dass $\nabla f(x^*) + h'(x^*)^T y^* = 0$. Wegen $h(x_k) = y_k/\sigma_k$ sowie $y_k \to y^*$ und $\sigma_k \to \infty$ ist $h(x^*) = 0$. Wegen $\|y_k\|^2 = \sigma_k^2 \|h(x_k)\|^2 \to \|y^*\|^2$ sowie $\sigma_k \to \infty$ folgt $\sigma_k \|h(x_k)\|^2 \to 0$ und damit $\Phi_k(x_k) \to f(x^*)$. Damit ist der erste Teil des Satzes bewiesen.

Zum Beweis des zweiten Teils beachten wir, dass (ohne Benutzung der hinreichenden Optimalitätsbedingungen zweiter Ordnung)

$$\begin{aligned} f(x^*) &= f(x_k) - (x_k - x^*)^T \nabla f(x_k) + o(\|x_k - x^*\|) \\ &= f(x_k) + (x_k - x^*)^T h'(x_k)^T y_k + o(\|x_k - x^*\|) \end{aligned}$$

und

$$\begin{aligned} 0 &= h(x^*) \\ &= h(x_k) - h'(x_k)(x_k - x^*) + o(\|x_k - x^*\|). \end{aligned}$$

Zusammen erhält man

$$\begin{aligned} f(x^*) &= f(x_k) + h(x_k)^T y_k + o(\|x_k - x^*\|) \\ &= f(x_k) + \sigma_k \|h(x_k)\|^2 + o(\|x_k - x^*\|) \\ &= \Phi_k(x_k) + \tfrac{1}{2}\sigma_k \|h(x_k)\|^2 + o(\|x_k - x^*\|). \end{aligned}$$

Weiter ist

$$\begin{aligned} 0 &= \nabla f(x_k) + h'(x_k)^T y_k \\ &= \nabla f(x^*) + \nabla^2 f(x^*)(x_k - x^*) + o(\|x_k - x^*\|) + h'(x_k)^T (y_k - y^*) + h'(x_k)^T y^* \\ &= \underbrace{\nabla f(x^*) + h'(x^*)^T y^*}_{=0} + W^*(x_k - x^*) + h'(x^*)^T (y_k - y^*) + o(\|x_k - x^*\|). \end{aligned}$$

Wegen

$$h(x_k) = \underbrace{h(x^*)}_{=0} + h'(x^*)(x_k - x^*) + o(\|x_k - x^*\|)$$

ist

$$\begin{pmatrix} 0 \\ h(x_k) \end{pmatrix} = \begin{pmatrix} W^* & h'(x^*)^T \\ h'(x^*) & 0 \end{pmatrix} \begin{pmatrix} x_k - x^* \\ y_k - y^* \end{pmatrix} + o(\|x_k - x^*\|).$$

Die Koeffizientenmatrix in dieser Beziehung ist nichtsingulär. Denn ist

$$\begin{pmatrix} W^* & h'(x^*)^T \\ h'(x^*) & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

so liefert die zweite Gleichung, dass $u \in \text{Kern}\,(h'(x^*))$. Eine Multiplikation der ersten Gleichung von links mit u^T ergibt $u^T W^* u = 0$. Wegen der hinreichenden Optimalitätsbedingungen zweiter Ordnung ist W^* auf Kern $(h'(x^*))$ positiv definit, so dass $u = 0$

folgt. Aus der ersten Gleichung folgt damit $h'(x^*)^T v = 0$, aus der Rangvoraussetzung folgt $v = 0$ und damit insgesamt die Nichtsingularität der angegebenen Matrix. Folglich ist

$$\begin{pmatrix} x_k - x^* \\ y_k - y^* \end{pmatrix} = \begin{pmatrix} W^* & h'(x^*)^T \\ h'(x^*) & 0 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ h(x_k) \end{pmatrix} + o(\|x_k - x^*\|)$$
$$= \begin{pmatrix} H^* & (T^*)^T \\ T^* & U^* \end{pmatrix} \begin{pmatrix} 0 \\ h(x_k) \end{pmatrix} + o(\|x_k - x^*\|).$$

Insbesondere ist

$$x_k - x^* = (T^*)^T h(x_k) + o(\|x_k - x^*\|).$$

Hieraus folgt wegen $h(x_k) = y^*/\sigma_k + o(1/\sigma_k)$, dass $x_k - x^* = O(1/\sigma_k)$, insgesamt folgen wegen

$$f(x^*) = \Phi_k(x_k) + \tfrac{1}{2}\sigma_k \|h(x_k)\|^2 + o(\|x_k - x^*\|) = \Phi_k(x_k) + \tfrac{1}{2}\sigma_k \|h(x_k)\|^2 + o(1/\sigma_k)$$

und

$$x_k - x^* = (T^*)^T h(x_k) + o(\|x_k - x^*\|) = (T^*)^T y^*/\sigma_k + o(1/\sigma_k)$$

die restlichen Behauptungen. □ □

Bemerkung: Wir zitieren einige Sätze aus R. Fletcher (1987, S. 283):

- This well-developed theoretical background may make it appear that, apart from the inefficiency of sequential minimization, the method is a robust one which can be used with confidence. In fact this is not true at all and there are severe numerical difficulties which arise when the method is used in practice. These are caused by the fact that as $\sigma_k \to \infty$, it is increasingly difficult to solve the problem (P_{σ_k}).

Die Lösung der unrestringierten Optimierungsaufgabe (P_σ) zu finden, bedeutet anschaulich, in einem mit wachsendem σ immer langgestreckteren Tal den tiefsten Punkt zu finden, was schwierig ist. □

Beispiel: Gegeben sei die Optimierungsaufgabe (siehe P. Spelluci (1993, S. 401)[1]

(P) Minimiere $f(x) := (x_1+2)^2+x_2^2$ auf $M := \{x \in \mathbb{R}^2 : h(x) := x_1^2+x_2^2-1 = 0\}$.

Zunächst berechnen wir mit der Lagrangeschen Multiplikatorenregel die (eindeutige) Lösung $x^* = (x_1^*, x_2^*)^T$. Es existiert $y^* \in \mathbb{R}$ mit

$$\nabla f(x^*) + y^* \nabla h(x^*) = 2 \begin{pmatrix} x_1^* + 2 + y^* x_1^* \\ x_2^* + y^* x_2^* \end{pmatrix} = 0.$$

Aus der zweiten Gleichung erhält man, dass $x_2^* = 0$ oder $y^* = -1$. Die Annahme, es sei $y^* = -1$ liefert über die erste Gleichung zu einem Widerspruch. Also ist $x_2^* = 0$. Aus der ersten Gleichung folgt

$$x_1^* = -\frac{2}{1+y^*},$$

[1] Spellucci, P. (1993) *Numerische Verfahren der nichtlinearen Optimierung.* Birkhäuser, Basel-Boston-Berlin.

die Nebenbedingung $h(x^*) = 0$ liefert $y^* = 1$ bzw. $x_1^* = -1$ oder $y^* = -3$ bzw. $x_1^* = 1$. In $x^* = (-1, 0)^T$ nimmt die Zielfunktion auf M ihr Minimum an, in $(1, 0)$ ihr Maximum. Beim Penalty-Verfahren mit einer quadratischen Straffunktion wird der Aufgabe (P) mit $\sigma > 0$ die Schar unrestringierter Optimierungsaufgaben

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \tfrac{1}{2}\sigma \, \|h(x)\|^2, \quad x \in \mathbb{R}^2$$

gegenüber gestellt. Eine Lösung $x(\sigma)$ von (P_σ) bestimmt man aus

$$0 = \nabla \Phi_\sigma(x) = 2 \begin{pmatrix} x_1 + 2 + \sigma x_1 (x_1^2 + x_2^2 - 1) \\ x_2 + \sigma x_2 (x_1 + x_2^2 - 1) \end{pmatrix}.$$

Für eine Lösung ist notwendigerweise $x_2 = 0$ (andernfalls erhielte man einen Widerspruch zur ersten Gleichung). Also ist $x_1(\sigma)$ als Lösung von

$$\sigma x_1^3 + (1 - \sigma) x_1 + 2 = 0$$

zu bestimmen. Mit $\rho := 1/\sigma$ hat man also die kubische Gleichung

$$x_1^3 + (\rho - 1) x_1 + 2\rho = 0$$

zu lösen. Dies ist bekanntlich exakt möglich, uns interessiert aber nur eine Entwicklung einer Lösung nach ρ bzw. $1/\sigma$. Auf

$$F(x, \rho) := x^3 + (\rho - 1)x + 2\rho = 0$$

wenden wir den Satz über implizite Funktionen an. Die Gleichung $F(x, 0) = 0$ hat die drei Lösungen -1, 0 und 1. Da wir die exakte Lösung von (P) ja schon kennen, interessiert uns von diesen drei Lösungen nur die erste. Wegen

$$\frac{\partial F}{\partial x}(-1, 0) = 4 \neq 0$$

liefert der Satz über implizite Funktionen

$$x_1(\sigma) = -1 - \tfrac{3}{4}\frac{1}{\sigma} + O(1/\sigma^2).$$

Es ist nicht schwierig (nur etwas mühsam) nachzuweisen, daß $\nabla^2 \Phi_\sigma(x(\sigma)) = O(\sigma)$. □

Ist eine nichtlineare Optimierungsaufgabe mit Ungleichungsrestriktionen gegeben, etwa

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0\},$$

so ist eine naheliegende Penalty-Funktion durch

$$\Phi_\sigma(x) := f(x) + \sigma \sum_{i=1}^{l} \max(g_i(x), 0)^2$$

gegeben. Diese Straffunktion ist bei glattem g einmal stetig differenzierbar, während die zweite Ableitung Sprünge besitzt. Trotzdem können im wesentlichen die gleichen

theoretischen Aussagen wie oben bei durch Gleichungen restringierte Optimierungsaufgaben gemacht werden.

Beispiel: Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := x^2 \quad \text{auf} \quad M := \{x \in \mathbb{R} : g(x) := 1 - x \leq 0\}.$$

Offenbar ist $x^* = 1$ die (eindeutige) Lösung von (P), der zugehörige Lagrange-Multiplikator ist $y^* = 2$. Wir wollen die Lösung $x(\sigma)$ von

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := x^2 + \sigma \max(1 - x, 0)^2, \quad x \in \mathbb{R}$$

bestimmen. Es ist

$$\Phi'_\sigma(x) = \begin{cases} 2x + \sigma 2(x-1), & x < 1, \\ 2x, & x \geq 1. \end{cases}$$

Daher ist

$$x(\sigma) + \sigma(x(\sigma) - 1) = 0,$$

bzw.

$$x(\sigma) = \frac{\sigma}{1+\sigma},$$

in der Tat ist auch hier $\lim_{\sigma\to\infty} x(\sigma) = x^*$. Ferner ist $x(\sigma) - x^* = O(1/\sigma)$. □

Bemerkung: Natürlich sind auch andere als quadratische Straffunktionen denkbar. Ist z. B. eine Optimierungsaufgabe ohne Gleichungsrestrionen, also

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0\}$$

gegeben, so kann man hierzu die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \frac{1}{\sigma}\sum_{i=1}^{l} y_i \exp(\sigma g_i(x)), \quad x \in \mathbb{R}^n$$

betrachten, wobei $y > 0$. Für $x \in M$ ist $f(x) \leq \Phi_\sigma(x) \leq f(x) + \|y\|_1/\sigma$ für alle $\sigma > 0$ und damit $\Phi_\sigma(x) \to f(x)$ mit $\sigma \to \infty$, während offenbar $\Phi_\sigma(x) \to \infty$ mit $\sigma \to \infty$ für alle $x \notin M$. Solche Straffunktionen kommen u. a. bei P. Tseng, D. P. Bertsekas (1993)[2] und R. Comminetti, J. San Martin (1994)[3] vor. □

5.1.2 Nichtdifferenzierbare, exakte Straffunktionen

Gegeben sei jetzt wieder die nichtlineare Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$$

mit Gleichungen und Ungleichungen als Restriktionen. Wir nehmen an, $x^* \in M$ sei eine (globale, lokale, kritische) Lösung von (P). Ferner wird wieder angenommen, die

[2] Tseng, P. and D. P. Bertsekas (1993) "On the convergence of the exponential multiplier method for convex programming." Mathematical Programming 60, 1–19.

[3] Comminetti, R. and J. San Martin (1994) "Asymptotic analysis of the exponential penalty trajectory in linear programming." Mathematical Programming 67, 169–187.

Zielfunktion f und die Restriktionsabbildungen g, h seien glatt (d. h. alle Ableitungen, die wir hinschreiben, existieren und sind stetig). Die zu (P) gehörende (differenzierbare) quadratische Straffunktion

$$\Phi_\sigma(x) := f(x) + \sigma \left(\sum_{i=1}^{l} \max(g_i(x), 0)^2 + \tfrac{1}{2} \|h(x)\|^2 \right)$$

hat den Nachteil, dass die zugehörige unrestringierte Optimierungsaufgabe mit wachsendem σ immer schlechter konditioniert ist. Man stellt sich daher die Frage, ob man nicht dem restringierten Problem (P) *eine* unrestringierte Optimierungsaufgabe zuordnen kann mit der Eigenschaft, dass x^* eine lokale Lösung dieser (unrestringierten) Aufgabe ist. Es stellt sich heraus, dass dies in der Tat im wesentlichen möglich ist, die dabei auftretenden Straffunktionen (die dann auch *exakt* genannt werden) aber nichtdifferenzierbar sind. Die bekannteste nichtdifferenzierbare exakte Straffunktion ist die L_1 (exakte) Straffunktion, welche durch

$$\Psi_\sigma(x) := f(x) + \sigma \left(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1 \right)$$

definiert ist und zuerst von T. Pietrzyowski (1969)[4] eingeführt wurde. Hierbei ist $\sigma > 0$ ein geeigneter Parameter und $\|\cdot\|_1$ die Betragssummennorm (oder auch L_1-Norm) auf dem $\mathbb{R}^m$. Man beachte, dass Ψ_σ wieder die charakteristischen Eigenschaften einer Straffunktion hat, d. h. es ist $\Psi_\sigma(x) = f(x)$ für alle $x \in M$, während $\Psi_\sigma(x) \to +\infty$ mit $\sigma \to \infty$ für alle $x \notin M$. Offenbar ist Ψ_σ nicht im üblichen Sinne differenzierbar, so dass es sich bei der unrestringierten Aufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x), \quad x \in \mathbb{R}^n$$

um eine "nichtglatte" (nonsmooth) Optimierungsaufgabe handelt. Denkbar wäre es aber auch, die einzelnen Komponenten der Restriktionsabbildungen zu gewichten, also etwa mit der Funktion

$$\Psi_{\alpha\beta}(x) := f(x) + \sum_{i=1}^{l} \alpha_i \max(g_i(x), 0) + \sum_{ji=1}^{m} \beta_j \, |h_j(x)|$$

mit $\alpha, \beta > 0$ zu arbeiten (siehe P. Spellucci (1993, S. 457)). Dies läuft aber natürlich nur darauf hinaus, die Komponenten der Restriktionsabbildungen mit positiven Zahlen durchzumultiplizieren, trotzdem stellt sich diese Straffunktion im Zusammenhang mit der Methode der sequentiellen quadratischen Minimierung als nützlich heraus (siehe Unterabschnitt 5.1.3). Weitere nichtdifferenzierbare Straffunktionen sind denkbar, z. B.

$$\Psi_\sigma(x) := f(x) + \max(0, g_1(x), \ldots, g_l(x), |h_1(x)|, \ldots, |h_m(x)|),$$

siehe z. B. D. P. Bertsekas (1982, S. 194)[5].

[4] Pietrzykowski, T. (1969) "An exact potential method for constrained maxima." SIAM J. Numer. Anal. 6, 299–304.

[5] Bertsekas, D. P. (1982) *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press, New York.

Beispiel: Zu

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := x^2 \quad \text{auf} \quad M := \{x \in \mathbb{R} : h(x) := x - 1 = 0\}$$

gehört die unrestringierte Optimierungsaufgabe

$$\text{(P}_\sigma\text{)} \qquad \text{Minimiere} \quad \Psi_\sigma(x) := x^2 + \sigma\,|x-1|, \quad x \in \mathbb{R}.$$

Natürlich ist $x^* := 1$ die einzige zulässige Lösung und damit die Lösung von (P). In Abbildung 5.1.2 links geben wir die Abbildung Ψ_σ für $\sigma = 0.5$ an. Man erkennt, dass

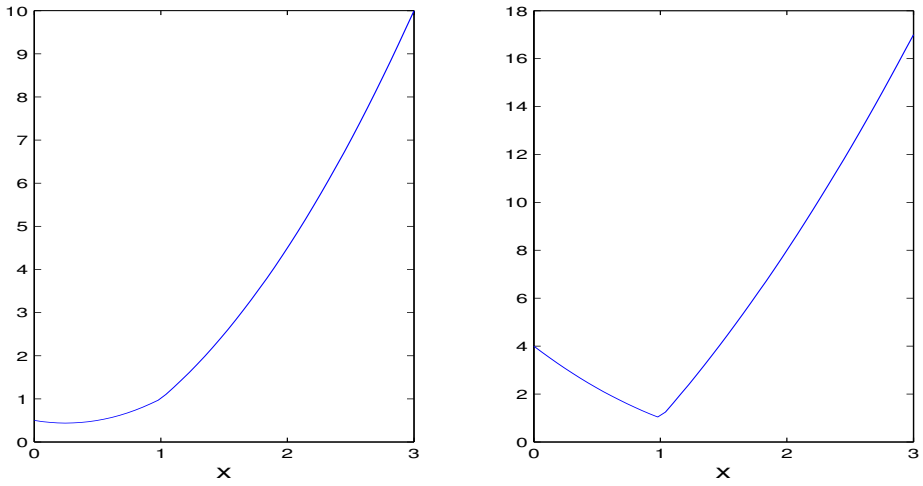

Abbildung 5.1: Die exakte L_1-Straffunktion mit $\sigma = 0.5$ und $\sigma = 4$

$x^* = 1$ keine Lösung von $(\text{P}_{0.5})$ ist. Im Gegensatz hierzu zeichnen wir in Abbildung 5.1.2 rechts die exakte L_1-Straffunktion mit $\sigma = 4$. Offensichtlich besitzt Ψ_4 in $x^* = 1$ ein Minimum. Für $x \geq 1$ ist $\Psi_\sigma(x) \geq \Psi_\sigma(1)$ für alle $\sigma > 0$. Für $x < 1$ ist dagegen $\Psi_\sigma(x) = x^2 + \sigma(1-x)$ und folglich $\Psi'_\sigma(x) = 2x - \sigma < 2 - \sigma$. Für alle $\sigma > 2$ ist daher $x^* = 1$ die Lösung der unrestringierten Optimierungsaufgabe (P_σ). □

Nun interessiert, ob allgemein (unter geeigneten Voraussetzungen) eine Aussage wie im letzten Beispiel gemacht werden kann, dass also zu einer Lösung x^* von (P) ein $\sigma^* > 0$ derart existiert, dass x^* für alle $\sigma > \sigma^*$ eine Lösung von (P_σ) ist. Hierzu benötigt man hinreichende Optimalitätsbedingungen für die nichtglatte Optimierungsaufgabe

$$\text{(P}_\sigma\text{)} \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma\left(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1\right), \quad x \in \mathbb{R}^n.$$

Kompliziert (oder positiv gewendet: interessant) wird dies dadurch, dass Ψ_σ nicht im üblichen Sinne differenzierbar ist. Es ist naheliegend, dass man schrittweise vorgeht, und zunächst *notwendige Optimalitätsbedingungen erster Ordnung* für (P_σ) aufstellt. Nach wie vor setzen wir voraus, dass f, g und h glatt sind. Zunächst wollen wir uns überlegen, dass die Richtungsableitung

$$\Psi'_\sigma(x^*; p) := \lim_{t \to 0+} \frac{\Psi_\sigma(x^* + tp) - \Psi_\sigma(x^*)}{t}$$

in $x^* \in \mathbb{R}^n$ (in dem die Daten f, g und h glatt sind) in jede Richtung p existiert.

Lemma 1.3 *Die (exakte) L_1-Straffunktion*

$$\Psi_\sigma(x) := f(x) + \sigma \Big(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1 \Big)$$

ist in x^* in jede Richtung p richtungsdifferenzierbar. Ferner ist

$$\begin{aligned} \Psi'_\sigma(x^*; p) \;\; = \;\; & \nabla f(x^*)^T p + \sigma \Big(\sum_{i \in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*)^T p \\ & + \sum_{j \in J^*} |\nabla h_j(x^*)^T p| + \sum_{j \notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*)^T p \Big). \end{aligned}$$

Hierbei ist

$$I^* := \{i \in \{1, \ldots, l\} : g_i(x^*) = 0\}, \qquad J^* := \{j \in \{1, \ldots, m\} : h_j(x^*) = 0\},$$

ferner sind τ_i, $i \in \{1, \ldots, l\} \setminus I^*$, durch

$$\tau_i := \begin{cases} 1, & \text{falls} \quad g_i(x^*) > 0, \\ 0, & \text{falls} \quad g_i(x^*) < 0, \end{cases} \qquad i \in \{1, \ldots, l\} \setminus I^*$$

definiert.

Beweis: O. B. d. A. können wir offenbar $l = m = 1$ annehmen, so dass

$$\Psi_\sigma(x) = f(x) + \sigma \, [\max(g(x), 0) + |h(x)|].$$

Wir definieren $r \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ bzw. $q \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$r(x) := \max(g(x), 0), \qquad q(x) := |h(x)|.$$

Sei $p \in \mathbb{R}^n$ vorgegeben. Ist $g(x^*) > 0$, so ist $r(x^* + tp) = g(x^* + tp)$ für alle hinreichend kleinen $t > 0$ und daher $r'(x^*; p) = \nabla g(x^*)^T p$. Ist dagegen $g(x^*) < 0$, so ist $r(x^* + tp) = 0$ für alle hinreichend kleinen $t > 0$ und damit $r'(x^*; p) = 0$. Sei daher schließlich $g(x^*) = 0$. Mit $t \to 0+$ ist dann aber

$$\frac{r(x^* + tp) - r(x^*)}{t} = \frac{\max(t \nabla g(x^*)^T p + o(t), 0)}{t} \to \max(\nabla g(x^*)^T p, 0).$$

Ähnlich einfach kann die Existenz der Richtungsableitung von q und

$$q'(x^*; p) = \begin{cases} |\nabla h(x^*)^T p|, & \text{falls} \quad h(x^*) = 0, \\ \operatorname{sign}[h(x^*)] \nabla h(x^*)^T p, & \text{falls} \quad h(x^*) \neq 0 \end{cases}$$

nachgewiesen werden. Damit ist die Aussage des Lemmas bewiesen. □ □

Wir nennen $x^* \in \mathbb{R}^n$ eine *kritische Lösung* (oder auch stationäre Lösung) von (P_σ), wenn $\Psi'_\sigma(x^*; p) \geq 0$ für alle $p \in \mathbb{R}^n$, wenn es also keine Richtung p gibt mit $\Psi'_\sigma(x^*; p) < 0$ gibt, also keine "unmittelbare" Abstiegsrichtung. Im folgenden Lemma wird eine Charakterisierung dafür angegeben, dass ein $x^* \in \mathbb{R}^n$ (in dem die Daten f, g und h stetig differenzierbar sind) eine kritische Lösung von (P_σ) ist.

Lemma 1.4 *Für $x^* \in \mathbb{R}^n$ seien $I^* \subset \{1,\dots,l\}$, $J^* \subset \{1,\dots,m\}$ sowie τ_i, $i \in \{1,\dots,l\} \setminus I^*$, wie in Lemma 1.3 definiert. Dann ist $x^* \in \mathbb{R}^n$ genau dann eine kritische Lösung von (P_σ), wenn Zahlen $\hat{u}_i$, $i \in I^*$, und $\hat{v}_j$, $j \in J^*$, existieren mit*

$$0 \le \hat{u}_i \le 1 \quad (i \in I^*), \qquad -1 \le \hat{v}_j \le 1 \quad (j \in J^*)$$

und

$$\begin{aligned} 0 \;=\; & \nabla f(x^*) + \sigma \Big(\sum_{i \in I^*} \hat{u}_i \nabla g_i(x^*) + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*) \\ & + \sum_{j \in J^*} \hat{v}_j \nabla h_j(x^*) + \sum_{j \notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*) \Big). \end{aligned}$$

Beweis: Wir nehmen zunächst an, dass es Zahlen $\hat{u}_i$, $i \in I^*$, $\hat{v}_j$, $j \in J^*$, mit den angegebenen Eigenschaften gibt. Sei $p \in \mathbb{R}^n$ beliebig. Dann ist

$$\begin{aligned} 0 \;=\; & \nabla f(x^*)^T p + \sigma \Big(\sum_{i \in I^*} \hat{u}_i \nabla g_i(x^*)^T p + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*)^T p \\ & + \sum_{j \in J^*} \hat{v}_j \nabla h_j(x^*)^T p + \sum_{i \notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*)^T p \Big) \\ \le \; & \nabla f(x^*)^T p + \sigma \Big(\sum_{i \in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*)^T p \\ & + \sum_{j \in J^*} |\nabla h_j(x^*)^T p| + \sum_{j \notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*)^T p \Big) \\ = \; & \Psi'_\sigma(x^*; p), \end{aligned}$$

also $x^* \in \mathbb{R}^n$ eine kritische Lösung von (P_σ).

Nun sei umgekehrt $x^* \in \mathbb{R}^n$ eine kritische Lösung von (P_σ). Wir machen einen Widerspruchsbeweis und nehmen an, es gäbe keine $\hat{u}_i$, $i \in I^*$, und $\hat{v}_j$, $j \in J^*$, mit den angegebenen Eigenschaften. Das bedeutet, dass das Gleichungs-Ungleichungssystem

$$\begin{cases} \displaystyle\sum_{i \in I^*} u_i \nabla g_i(x^*) + \sum_{j \in J^*} v_j \nabla h_j(x^*) = c, \\ 0 \le u_i \le 1 \quad (i \in I^*), \qquad -1 \le v_j \le 1 \quad (j \in J^*) \end{cases}$$

nicht lösbar ist, wobei

$$c := -\Big(\frac{1}{\sigma} \nabla f(x^*) + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*) + \sum_{j \notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*) \Big).$$

Es liegt nahe, das Farkas-Lemma anzuwenden. Zur Vereinfachung der Notation seien die Matrizen

$$A := (\nabla g_i(x^*))_{i \in I^*} \in \mathbb{R}^{n \times q}, \qquad B := (\nabla h_j(x^*))_{j \in J^*} \in \mathbb{R}^{n \times r}$$

definiert mit $q := \#(I^*)$, $r := \#(J^*)$, ferner sei

$$u = (u_i)_{i \in I^*} \in \mathbb{R}^q, \qquad v = (v_j)_{j \in J^*} \in \mathbb{R}^r,$$

schließlich sei e ein Vektor der Länge q bzw. r, dessen Komponenten alle gleich 1 sind. Hiermit besagt die Widerspruchsannahme, dass das Gleichungs-Ungleichungssystem

$$Au + Bv = c, \qquad 0 \le u \le e, \qquad -e \le v \le e$$

nicht lösbar ist. Etwas anders geschrieben bedeutet dies, dass

$$\begin{pmatrix} c \\ e \\ e \\ e \end{pmatrix} - \begin{pmatrix} A & B \\ I & 0 \\ 0 & I \\ 0 & -I \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \in \{0\} \times \mathbb{R}^q_{\ge 0} \times \mathbb{R}^r_{\ge 0} \times \mathbb{R}^r_{\ge 0},$$
$$(u, v) \in \mathbb{R}^q_{\ge 0} \times \mathbb{R}^r$$

nicht lösbar ist. Das verallgemeinerte Farkas-Lemma (siehe Lemma 1.8 in Abschnitt 2.1) liefert die Existenz eines $p \in \mathbb{R}^n$ und von nichtnegativen Vektoren $\alpha \in \mathbb{R}^q$ sowie $\beta, \gamma \in \mathbb{R}^r$ mit

$$-A^T p + \alpha \ge 0, \qquad -B^T p + \beta - \gamma = 0, \qquad -c^T p + e^T \alpha + e^T(\beta + \gamma) < 0$$

bzw.

$$\nabla g_i(x^*)^T p \le \alpha_i \quad (i \in I^*), \qquad \nabla h_j(x^*)^T p = \beta_j - \gamma_j \quad (j \in J^*)$$

und

$$-c^T p + \sum_{i \in I^*} \alpha_i + \sum_{j \in J^*} (\beta_j + \gamma_j) < 0.$$

Dann ist

$$\begin{aligned}
\frac{1}{\sigma}\Psi'_\sigma(x^*; p) &= \frac{1}{\sigma}\nabla f(x^*)^T p + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*)^T p + \sum_{j \notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*)^T p \\
&\quad + \sum_{i \in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{j \in J^*} |\nabla h_j(x^*)^T p| \\
&\le -c^T p + \sum_{i \in I^*} \max(\alpha_i, 0) + \sum_{j \in J^*} |\beta_j - \gamma_j| \\
&= -c^T p + \sum_{i \in I^*} \alpha_i + \sum_{j \in J^*} |\beta_j - \gamma_j| \\
&\le -c^T p + \sum_{i \in I^*} \alpha_i + \sum_{j \in J^*} (\beta_j + \gamma_j) \\
&< 0,
\end{aligned}$$

ein Widerspruch dazu, dass x^* eine kritische Lösung von (P_σ) ist. □ □

Bemerkung: Bei R. Fletcher (1987, S. 298 ff.) wird obiges Ergebnis für den Fall, dass keine Gleichungen als Restriktionen auftreten, bewiesen (allerdings nicht als Satz formuliert), wobei aber (seltsamerweise) vorausgesetzt wird, dass $\{\nabla g_i(x^*)\}_{i \in I^*}$ linear

unabhängig sind. Das Resultat selber scheint von T. F. Coleman, A. R. Conn (1980, Corollary 1)[6] zu stammen, wobei auch hier die Voraussetzung über die lineare Unabhängigkeit von $\{\nabla g_i(x^*)\}_{i\in I^*}$ zusammen mit $\{\nabla h_j(x^*)\}_{j\in J^*}$ gemacht wird. □

Unter einer kritischen Lösung des Ausgangsproblems

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \le 0,\ h(x) = 0\}$$

verstehen wir natürlich ein $x^* \in M$, zu welchem es ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$(*) \qquad u^* \ge 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad (u^*)^T g(x^*) = 0$$

gibt. Genauer sagen wir, ein Tripel (x^*, u^*, v^*) mit $x^* \in M$ und $(*)$ sei ein *Kuhn-Tucker-Tripel* zu (P). Dann gilt:

Satz 1.5 *Ist (x^*, u^*, v^*) ein Kuhn-Tucker-Tripel zu (P), so ist x^* für alle σ mit $\sigma \ge \max(\|u^*\|_\infty, \|v^*\|_\infty)$ eine kritische Lösung von (P_σ).*

Beweis: Da x^* zulässig für (P), ist mit den Bezeichnungen der letzten beiden Lemmata $J^* = \{1, \dots, m\}$, I^* die Menge der in x^* aktiven Ungleichungsrestriktionen und $\tau_i = 0$ für $i \in \{1, \dots, l\} \setminus I^*$. Daher ist wegen Lemma 1.4 die Existenz von Zahlen $\hat{u}_i$, $i \in I^*$, und $\hat{v}_j$, $j \in J^*$ mit

$$0 \le \hat{u}_i \le 1 \quad (i \in I^*), \qquad -1 \le \hat{v}_j \le 1 \quad (j \in J^*)$$

und

$$0 = \nabla f(x^*) + \sigma\Big(\sum_{i\in I^*} \hat{u}_i \nabla g_i(x^*) + \sum_{j\in J^*} \hat{v}_j \nabla h_j(x^*)\Big)$$

zu zeigen. Da aber (x^*, u^*, v^*) ein Kuhn-Tucker-Tripel ist, ist $u_i^* \ge 0$, $i \in I^*$, $u_i^* = 0$, $i \in \{1, \dots, l\} \setminus I^*$, und

$$0 = \nabla f(x^*) + \sum_{i\in I^*} u_i^* \nabla g_i(x^*) + \sum_{j\in J^*} v_j^* \nabla h_j(x^*).$$

Setzt man daher

$$\hat{u}_i := \frac{1}{\sigma} u_i^* \quad (i \in I^*), \qquad \hat{v}_j := \frac{1}{\sigma} v_j^* \quad (j \in J^*),$$

so ist x^* offenbar für alle $\sigma \ge \max(\|u^*\|_\infty, \|v^*\|_\infty)$ eine kritische Lösung von (P_σ). □□

Nun geben wir Bedingungen dafür an, dass $x^* \in \mathbb{R}^n$ eine isolierte, lokale Lösung der unrestringierten Optimierungsaufgabe (P_σ) ist. Eine sehr ähnliche Aussage findet man bei T. F. Coleman, A. R. Conn (1980, Corollary 3), siehe auch (für nichtlineare L_1-Funktionen) C. Charalambous (1979, Theorem 3)[7].

[6] Coleman, T. F. and A. R. Conn (1980) "Second-order conditions for an exact penalty function." Mathematical Programming 19, 178–185.

[7] Charalambous, C. (1979) "On the conditions for optimality of the nonlinear l_1 problem." Mathematical Programming 17, 123–135.

Lemma 1.6 *Zu $x^* \in \mathbb{R}^n$ seien die Indexmengen $I^* \subset \{1,\dots,l\}$, $J^* \subset \{1,\dots,m\}$ sowie τ_i, $i \in \{1,\dots,l\} \setminus I^*$, wie in Lemma 1.3 definiert. Es mögen $\hat{u}_i$, $i \in I^*$, und $\hat{v}_j$, $j \in J^*$, mit*

$$0 \le \hat{u}_i \le 1 \quad (i \in I^*), \qquad -1 \le \hat{v}_j \le 1 \quad (j \in J^*)$$

und

$$\begin{aligned} 0 \;=\; & \nabla f(x^*) + \sigma \Big(\sum_{i \in I^*} \hat{u}_i \nabla g_i(x^*) + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*) \\ & + \sum_{j \in J^*} \hat{v}_j \nabla h_j(x^*) + \sum_{j \notin J^*} \operatorname{sign}\left[h_j(x^*)\right] \nabla h_j(x^*) \Big) \end{aligned}$$

existieren. Hiermit definiere man die Mengen

$$A^* := \left\{ p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \begin{cases} \le 0 & \text{für } i \in I^* \text{ mit } \hat{u}_i = 0, \\ = 0 & \text{für } i \in I^* \text{ mit } \hat{u}_i \in (0,1), \\ \ge 0 & \text{für } i \in I^* \text{ mit } \hat{u}_i = 1 \end{cases} \right\}$$

und

$$B^* := \left\{ p \in \mathbb{R}^n : \nabla h_j(x^*)^T p \begin{cases} \le 0 & \text{für } j \in J^* \text{ mit } \hat{v}_j = -1, \\ = 0 & \text{für } j \in J^* \text{ mit } \hat{v}_j \in (-1,1), \\ \ge 0 & \text{für } j \in J^* \text{ mit } \hat{v}_j = 1 \end{cases} \right\}.$$

Die Matrix

$$\begin{aligned} W_\sigma^* \;:=\; & \nabla^2 f(x^*) + \sigma \Big(\sum_{i \in I^*} \hat{u}_i \nabla^2 g_i(x^*) + \sum_{i \notin I^*} \tau_i \nabla^2 g_i(x^*) \\ & + \sum_{j \in J^*} \hat{v}_j \nabla^2 h_j(x^*) + \sum_{j \notin J^*} \operatorname{sign}\left[h_j(x^*)\right] \nabla^2 h_j(x^*) \Big) \end{aligned}$$

sei positiv definit auf $A^* \cap B^*$, d. h. es sei

$$p^T W_\sigma^* p > 0 \qquad \text{für alle } p \in A^* \cap B^* \setminus \{0\}.$$

Dann ist x^* eine isolierte, lokale Lösung von (P_σ), d. h. es existiert eine Umgebung U^* von x^* mit $\Psi_\sigma(x^*) < \Psi_\sigma(x)$ für alle $x \in U^* \setminus \{x^*\}$.

Beweis: Angenommen, die Behauptung sei falsch. Dann gibt es eine gegen x^* konvergierende Folge $\{x_k\}$, $x_k \neq x^*$ für alle k, mit $\Psi_\sigma(x_k) \le \Psi_\sigma(x^*)$. Man stelle x_k dar in der Form

$$x_k = x^* + \underbrace{\|x_k - x^*\|}_{=:t_k} \underbrace{\frac{x_k - x^*}{\|x_k - x^*\|}}_{=:p_k} = x^* + t_k p_k.$$

Wegen $x_k \to x^*$ gilt $t_k \to 0$. Aus $\{p_k\}$ kann eine konvergente Teilfolge ausgewählt werden. Daher nehmen wir o. B. d. A. an, die Folge $\{p_k\}$ konvergiere schon gegen ein p,

welches wegen $\|p\| \neq 0$ vom Nullvektor verschieden ist. Wegen $x_k = x^* + t_k p + r_k$ mit $r_k := t_k(p_k - p)$ und $r_k/t_k \to 0$ kann leicht gezeigt werden, dass

$$0 \geq \frac{\Psi_\sigma(x_k) - \Psi_\sigma(x^*)}{t_k} = \frac{\Psi_\sigma(x^* + t_k p + r_k) - \Psi_\sigma(x^*)}{t_k} \to \Psi'_\sigma(x^*; p).$$

Also ist (siehe Lemma 1.3)

$$\begin{aligned}
0 &\geq \Psi'_\sigma(x^*; p) \\
&= \nabla f(x^*)^T p + \sigma\Big(\sum_{i\in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{i\notin I^*} \tau_i \nabla g_i(x^*)^T p \\
&\qquad + \sum_{j\in J^*} |\nabla h_j(x^*)^T p| + \sum_{j\notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*)^T p\Big) \\
&= \sigma\Big(\sum_{i\in I^*} [\underbrace{\max(\nabla g_i(x^*)^T p, 0) - \hat{u}_i \nabla g_i(x^*)^T p}_{\geq 0}] + \sum_{j\in J^*} [\underbrace{|\nabla h_j(x^*)^T p| - \hat{v}_j \nabla h_j(x^*)^T p}_{\geq 0}]\Big).
\end{aligned}$$

Hieraus folgt

$$\max(\nabla g_i(x^*)^T p, 0) = \hat{u}_i \nabla g_i(x^*)^T p \quad (i \in I^*)$$

und

$$|\nabla h_j(x^*)^T p| = \hat{v}_j \nabla h_j(x^*)^T p \quad (j \in J^*).$$

Aus der ersten Beziehung folgt $p \in A^*$, aus der zweiten $p \in B^*$, insgesamt also $p \in A^* \cap B^* \setminus \{0\}$. Nun besteht unser Ziel natürlich darin, $p^T W^*_\sigma p \leq 0$ nachzuweisen, womit der gewünschte Widerspruch erreicht wäre. Für alle hinreichend großen k ist

$$\begin{aligned}
0 &\geq \Psi_\sigma(x_k) - \Psi_\sigma(x^*) \\
&= f(x_k) - f(x^*) \\
&\qquad + \sigma\Big(\sum_{i=1}^{l} [\max(g_i(x_k), 0) - \max(g_i(x^*), 0)] + \sum_{j=1}^{m} [|h_j(x_k)| - |h_j(x^*)|]\Big) \\
&= f(x_k) - f(x^*) + \sigma\Big(\sum_{i\in I^*} \max(g_i(x_k), 0) + \sum_{j\in J^*} |h_j(x_k)|\Big) \\
&\qquad + \sigma\Big(\sum_{i\notin I^*} [\max(g_i(x_k), 0) - \max(g_i(x^*), 0)] + \sum_{j\notin J^*} [|h_j(x_k)| - |h_j(x^*)|]\Big) \\
&\geq f(x_k) - f(x^*) + \sigma\Big(\sum_{i\in I^*} \hat{u}_i [g_i(x_k) - \underbrace{g_i(x^*)}_{=0}] + \sum_{j\in J^*} \hat{v}_j [h_j(x_k) - \underbrace{h_j(x^*)}_{=0}]\Big) \\
&\qquad + \sigma\Big(\sum_{i\notin I^*} \tau_i [g_i(x_k) - g_i(x^*)] + \sum_{j\notin J^*} \operatorname{sign}[h_j(x^*)][h_j(x_k) - h_j(x^*)]\Big) \\
&= t_k\Big[\nabla f(x^*) + \sigma\Big(\sum_{i\in I^*} \hat{u}_i \nabla g_i(x^*) + \sum_{i\notin I^*} \tau_i \nabla g_i(x^*) \\
&\qquad + \sum_{j\in J^*} \hat{v}_j \nabla h_j(x^*) + \sum_{j\notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*)\Big)\Big]^T p_k \\
&\qquad + \tfrac{1}{2} t_k^2 p_k^T W_k p_k \\
&= \tfrac{1}{2} t_k^2 p_k^T W_k p_k
\end{aligned}$$

mit

$$W_k = \nabla^2 f(x_k^{(0)}) + \sigma\Big(\sum_{i\in I^*} \hat{u}_i \nabla^2 g_i(\hat{x}_k^{(i)}) + \sum_{i\notin I^*} \tau_i \nabla^2 g_i(x_k^{(i)}) + \sum_{j\in J^*} \hat{v}_j \nabla^2 h_j(\overline{x}_k^{(j)}) + \sum_{j\notin J^*} \operatorname{sign}\,[h_j(x^*)] \nabla^2 h_j(\tilde{x}_k^{(j)})\Big),$$

wobei $x_k^{(0)}$, $\hat{x}_k^{(i)}$ für $i \in I^*$ usw. jeweils zwischen x_k und x^* liegen, so dass z. B. $\hat{x}_k^{(i)} \to x^*$, $i \in I^*$. Daher erhält man aus $p_k^T W_k p_k \le 0$ nach dem Grenzübergang $k \to \infty$, dass $p^T W_\sigma^* p \le 0$, womit der gesuchte Widerspruch erhalten ist. □ □

Der nächste Satz sagt aus, dass ein Punkt $x^* \in M$, in dem die hinreichenden Optimalitätsbedingungen zweiter Ordnung für die restringierte Optimierungsaufgabe (P) erfüllt sind, für alle hinreichend großen σ eine isolierte, lokale Lösung von (P_σ) ist.

Satz 1.7 *In $x^* \in M$ seien die hinreichenden Optimalitätsbedingungen zweiter Ordnung für (P) erfüllt, d. h. es existieren $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ derart, dass (x^*, u^*, v^*) ein Kuhn-Tucker-Tripel ist, für welches*

$$p^T\Big[\nabla^2 f(x^*) + \sum_{i=1}^{l} u_i^* \nabla^2 g_i(x^*) + \sum_{j=1}^{m} v_j^* \nabla^2 h_j(x^*)\Big] p > 0$$
$$\text{für alle } p \in L^0(M; x^*) \setminus \{0\}.$$

Hierbei sei

$$L^0(M; x^*) := \Big\{p \in \mathbb{R}^n : \begin{array}{l} \nabla g_i(x^*)^T p = 0 \ \ (i \in I_+^*), \\ \nabla g_i(x^*)^T p \le 0 \ \ (i \in I^* \setminus I_+^*), \end{array} \ h'(x^*)p = 0\Big\},$$

wobei

$$I^* := \{i \in \{1, \dots, l\} : g_i(x^*) = 0\}$$

die Menge der in x^* aktiven Ungleichungsrestriktionen ist und

$$I_+^* := \{i \in I^* : u_i^* > 0\}.$$

Dann ist x^* für alle σ mit $\sigma > \max(\|u^*\|_\infty, \|v^*\|_\infty)$ eine isolierte, lokale Lösung von (P_σ).

Beweis: Da x^* nach Voraussetzung zulässig für (P) ist, haben wir wegen Lemma 1.6 für alle σ mit $\sigma > \max(\|u^*\|_\infty, \|v^*\|_\infty)$ die Existenz von Zahlen $\hat{u}_i$, $i \in I^*$, und $\hat{v}_j$, $j \in \{1, \dots, m\}$, mit

$$0 \le \hat{u}_i \le 1 \quad (i \in I^*), \qquad -1 \le \hat{v}_j \le 1 \quad (j \in \{1, \dots, m\})$$

und

$$0 = \nabla f(x^*) + \sigma\Big(\sum_{i\in I^*} \hat{u}_i \nabla g_i(x^*) + \sum_{j=1}^{m} \hat{v}_j \nabla h_j(x^*)\Big)$$

zu zeigen, für die die Matrix

$$W_\sigma^* := \nabla^2 f(x^*) + \sigma\Big(\sum_{i\in I^*} \hat{u}_i \nabla^2 g_i(x^*) + \sum_{j=1}^{m} \hat{v}_j \nabla^2 h_j(x^*)\Big)$$

auf $A^* \cap B^*$ positiv definit ist, wobei

$$A^* := \left\{ p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \begin{cases} \le 0 & \text{für } i \in I^* \text{ mit } \hat{u}_i = 0, \\ = 0 & \text{für } i \in I^* \text{ mit } \hat{u}_i \in (0,1), \\ \ge 0 & \text{für } i \in I^* \text{ mit } \hat{u}_i = 1 \end{cases} \right\}$$

und

$$B^* := \left\{ p \in \mathbb{R}^n : \nabla h_j(x^*)^T p \begin{cases} \le 0 & \text{für } j \in \{1,\dots,m\} \text{ mit } \hat{v}_j = -1, \\ = 0 & \text{für } j \in \{1,\dots,m\} \text{ mit } \hat{v}_j \in (-1,1), \\ \ge 0 & \text{für } j \in \{1,\dots,m\} \text{ mit } \hat{v}_j = 1 \end{cases} \right\}.$$

Nun definiere man

$$\hat{u}_i := \frac{1}{\sigma} u_i^* \quad (i \in I^*), \qquad \hat{v}_j := \frac{1}{\sigma} v_j^* \quad (j \in \{1,\dots,m\}).$$

Wegen $\sigma > \max(\|u^*\|_\infty, \|v^*\|_\infty)$ ist

$$0 \le \hat{u}_i < 1 \quad (i \in I^*), \qquad -1 < \hat{v}_j < 1 \quad (j \in \{1,\dots,m\}).$$

Daher ist $A^* \cap B^* = L^0(M; x^*)$ und die Behauptung folgt. □ □

Bemerkung: Auch der letzte Satz ist im wesentlichen bei R. Fletcher (1987, S. 300 ff.) angegeben worden, wobei allerding eine genaue Formulierung fehlt. Eine Verallgemeinerung dieses Satzes (es werden allgemeinere Straffunktionen zugelassen) findet man bei S.-P. Han, O. L. Mangasarian (1979, Theorem 4.6)[8]. □

Beispiel: Als Beispiel betrachten wir die Aufgabe

$$\text{(P)} \quad \text{Minimiere} \quad f(x) := x - \tfrac{1}{2}x^2 \quad \text{auf} \quad M := \left\{ x \in \mathbb{R} : g(x) := \begin{pmatrix} -x \\ x-1 \end{pmatrix} \le 0 \right\}.$$

Offenbar ist $x^* := 0$ die Lösung zu (P), mit $u^* := (1,0)^T$ ist (x^*, u^*) ein Kuhn-Tucker-Paar, in dem auch die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt sind (es ist $L^0(M; x^*) = \{0\}$). Wir betrachten die exakte L_1-Penalty-Funktion mit $\sigma := 1$ (wegen Satz 1.7 müssten wir eigentlich $\sigma > 1$ wählen), also die Aufgabe

$$(\mathrm{P}_1) \qquad \text{Minimiere} \quad \Psi_1(x) := x - \tfrac{1}{2}x^2 + \max(-x, 0) + \max(x-1, 0), \quad x \in \mathbb{R}.$$

Offensichtlich ist $x^* = 0$ aber keine lokale Lösung von (P_1), da $\Psi_1(x) = -\frac{1}{2}x^2$ für $x \le 0$. Sehr wohl im Einklang mit Satz 1.5 ist $x^* = 0$ aber eine kritische Lösung von

[8] Han, S.-P. and O. L. Mangasarian (1979) "Exact penalty functions in nonlinear programming." Mathematical Programming 17, 251–269.

(P_1). Denn als Richtungsableitung in x^* in Richtung p berechnet man nach Lemma 1.3 $\Psi_1'(x^*; p) = p + \max(-p, 0) \geq 0$. □

Bemerkung: Von S.-P. Han, O. L. Mangasarian (1979, Theorem 4.4) stammt eine weitere interessante Aussage über den Zusammenhang zwischen dem nichtlinear restringierten Problem (P) und der unrestringierten Aufgabe (P_σ) (mit der exakten L_1-Straffunktion Ψ_σ als Zielfunktion. Es gilt nämlich:

- *Sei $x^* \in M$ eine isolierte, lokale Lösung von (P). Die Daten von (P), also die Zielfunktion f und die Restriktionsabbildungen g, h, seien auf einer Umgebung von x^* stetig differenzierbar. Ferner sei die Arrow-Hurwicz-Uzawa Constraint Qualification erfüllt. Mit der Indexmenge $I(x^*)$ der in x^* aktiven Ungleichungsrestriktionen gelte also:*
  - *Es existiert ein $\hat{p} \in \mathbb{R}^n$ mit $\nabla g_i(x^*)^T \hat{p} < 0$, $i \in I(x^*)$, und $h'(x^*)\hat{p} = 0$.*
  - *Die Gradienten $\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ sind linear unabhängig.*

  Dann gibt es ein $\sigma^* > 0$ derart, dass x^* für alle $\sigma \geq \sigma^*$ eine lokale Lösung von (P_σ) ist.

Der Beweis hierzu bei S.-P. Han, O. L. Mangasarian (1979) ist nicht einfach und benutzt ein Resultat von T. Pietrzykowski (1970)[9] (siehe Aufgabe 6), welches wiederum nicht ganz einfach zu beweisen scheint. Einen "Beweis" obiger Aussage findet man auch bei P. Spellucci (1993, S. 469). □

Nun interessiert eine Umkehrung der letzten beiden Sätze eigentlich mehr als deren Aussage selber. Denn man stellt sich ja vor, dass man das Ausgangsproblem (P) dadurch zu lösen versucht, dass man mit einem hinreichend großen $\sigma > 0$ die unrestringierte Optimierungsaufgabe (P_σ) löst und möchte dann sicher sein, auch das Ausgangsproblem (in einem geeigneten Sinne) gelöst zu haben. Eine solche Aussage ist leider i. allg. nicht richtig, wie das folgende Beispiel zeigt, da Lösungen von (P_σ) i. allg. nicht zulässig für (P) sein werden.

Beispiel: Betrachte die (triviale) Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := 0 \quad \text{auf} \quad M := \{x \in \mathbb{R} : h(x) := x^3 + 3x^2 + 3 = 0\}.$$

Das zugehörige unrestringierte Problem ist (der Faktor σ spielt hier keine Rolle und wird deswegen weggelassen bzw. gleich 1 gesetzt)

$$(P_1) \qquad \text{Minimiere} \quad \Psi_1(x) := |x^3 + 3x^2 + 3|, \quad x \in \mathbb{R}.$$

In Abbildung 5.2 zeichnen wir Ψ_1 auf $[-1, 1]$. Man erkennt, dass $x^* = 0$ eine strikte, lokale Lösung von (P_1) ist, aber natürlich keine Lösung von (P), da $x^* = 0$ für (P) nicht zulässig ist. □

Andererseits gilt: Ist x^* eine kritische Lösung von (P_σ) für ein $\sigma > 0$ und darüber hinaus $x^* \in M$, also x^* zulässig für (P), so ist (x^*, u^*, v^*) mit einem geeigneten Paar $(u^*, v^*) \in$

[9] Pietrzykowski, T. (1970) "The potential method for conditional maxima in the locally compact metric spaces." Numer. Math. 14, 325–329.

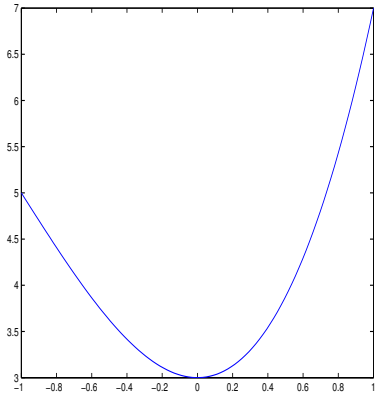

Abbildung 5.2: Die L_1-Penalty-Funktion

$\mathbb{R}^l \times \mathbb{R}^m$ ein Kuhn-Tucker-Tripel für (P) (bzw. x^* eine kritische Lösung von (P)). Der Beweis hierfür ist völlig trivial, wenn man die Charakterisierung kritischer Lösungen von (P_σ) in Lemma 1.4 und die Definition eines Kuhn-Tucker-Tripels berücksichtigt.

Die obigen Überlegungen sollen nicht suggerieren, dass es vom praktischen Standpunkt empfehlenswert ist, eine restringierte Optimierungsaufgabe mit Hilfe einer exakten Straffunktion auf eine unrestringierte Optimierungsaufgabe zurückzuführen. Wichtiger sind die exakten Straffunktionen im Zusammenhang mit der Schrittweitenbestimmung bei der Methode der sequentiellen quadratischen Optimierung. Hierauf werden wir im nächsten Unterabschnitt eingehen.

5.1.3 Die Methode der sequentiellen quadratischen Optimierung

Gegeben sei wieder die nichtlinear restringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\},$$

wobei f, g, h wie üblich als glatt vorausgesetzt werden. Der Aufgabe (P) ordnen wir die exakte L_1-Straffunktion

$$\Psi_{\alpha\beta}(x) := f(x) + \Big(\sum_{i=1}^{l} \alpha_i \max(g_i(x), 0) + \sum_{j=1}^{m} \beta_j |h_j(x)|\Big)$$

zu, bei der wir also auch (eventuell) unterschiedliche Gewichte $\alpha_i > 0$, $i = 1, \ldots, l$, und $\beta_j > 0$, $j = 1, \ldots, m$, zulassen. Wir wollen ein Verfahren schildern, das auf S.-P. Han (1976, 1977)[10] zurückgeht und von dem bei P. Spellucci (1993, S. 474) immerhin behauptet wird, dass es zu der zur Zeit am effizientesten allgemein einsetzbaren

[10] Han, S.-P. (1976) "Superlinearly convergent variable metric algorithms for general nonlinear programming problems." Mathematical Programming 11, 263–282.

Han, S.-P. (1977) "A globally convergent method for nonlinear programming." J. O. T. A. 22, 297–309.

Methode zur Lösung nichtlinearer Optimierungsaufgaben führt. Im folgenden Lemma wird angegeben, wie man zu einer aktuellen Näherung $x \in \mathbb{R}^n$ für eine (globale, lokale, kritische) Lösung eine Suchrichtung $p \in \mathbb{R}^n$ bestimmen kann, die für die exakte Straffunktion $\Psi_{\alpha\beta}$ für alle hinreichend großen α, β eine Abstiegsrichtung in x ist (wenn nicht x schon eine zulässige, kritische Lösung von (P) ist). Hierzu erinnern wir an die Darstellung der Richtungsableitung der nichtdifferenzierbaren L_1-Straffunktion, wobei es keine Schwierigkeiten machen sollte, dass diesmal nicht alle Gewichte gleich sind. Es ist

$$\begin{aligned}\Psi'_{\alpha\beta}(x;p) &= \nabla f(x)^T p + \sum_{i\in I} \alpha_i \max(\nabla g_i(x)^T p, 0) + \sum_{i\notin I} \alpha_i \tau_i \nabla g_i(x)^T p \\ &\quad + \sum_{j\in J} \beta_j |\nabla h_j(x)^T p| + \sum_{j\notin J} \beta_j \operatorname{sign}[h_j(x)] \nabla h_j(x)^T p,\end{aligned}$$

wobei

$$I := \{i \in \{1,\dots,l\} : g_i(x) = 0\}, \qquad J := \{j \in \{1,\dots,m\} : h_j(x) = 0\}$$

und τ_i, $i \in \{1,\dots,l\} \setminus I$, durch

$$\tau_i := \begin{cases} 1, & \text{falls} \quad g_i(x) > 0, \\ 0, & \text{falls} \quad g_i(x) < 0, \end{cases} \qquad i \in \{1,\dots,l\} \setminus I$$

definiert ist.

Lemma 1.8 *Gegeben sei ein Paar $(x, B) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}$, wobei B symmetrisch und positiv definit ist. Es wird vorausgesetzt, dass das quadratische Programm*

$$(\mathrm{Q}_{x,B}) \qquad \begin{cases} \textit{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T B p \quad \textit{unter den Nebenbedingungen} \\ g(x) + g'(x)p \le 0, \qquad h(x) + h'(x)p = 0 \end{cases}$$

zulässig ist. Die dann eindeutige Lösung von $(Q_{x,B})$ werde mit p bezeichnet. Dann gilt:

1. *Ist $p = 0$, so ist x eine zulässige, kritische Lösung von (P).*
2. *Ist $p \neq 0$, so ist p für alle hinreichend großen α, β eine Abstiegsrichtung für $\Psi_{\alpha\beta}$ in x.*

Beweis: Die Lösung p von $(\mathrm{Q}_{x,B})$ ist durch die Existenz von $(u, v) \in \mathbb{R}^l \times \mathbb{R}^n$ mit

$$u \ge 0, \qquad \nabla f(x) + Bp + g'(x)^T u + h'(x)^T v = 0, \qquad u^T[g(x) + g'(x)p] = 0$$

charakterisiert. Ist $p = 0$, so ist x zulässig für (P), ferner ist (x, u, v) offensichtlich ein Kuhn-Tucker-Tripel für (P) bzw. x eine kritische Lösung von (P). Daher sei jetzt $p \neq 0$. Mit den obigen Bezeichnungen ist dann

$$\Psi'_{\alpha\beta}(x;p) = \nabla f(x)^T p + \sum_{i\in I} \alpha_i \max(\nabla g_i(x)^T p, 0) + \sum_{i\notin I} \alpha_i \tau_i \nabla g_i(x)^T p$$

$$
\begin{aligned}
&\quad + \sum_{j\in J} \beta_j |\nabla h_j(x)^T p| + \sum_{j\notin J} \beta_j \text{sign}\,[h_j(x)] \nabla h_j(x)^T p \\
= &\quad \underbrace{-p^T B p - \sum_{i=1}^{l} u_i \nabla g_i(x)^T p - \sum_{j=1}^{m} v_j \nabla h_j(x)^T p}_{=\nabla f(x)^T p} \\
&\quad + \sum_{i\in I} \alpha_i \max(\nabla g_i(x)^T p, 0) + \sum_{i\notin I} \alpha_i \tau_i \nabla g_i(x)^T p \\
&\quad + \sum_{j\in J} \beta_j |\nabla h_j(x)^T p| + \sum_{j\notin J} \beta_j \text{sign}\,[h_j(x)] \nabla h_j(x)^T p \\
\leq &\quad -p^T B p + \sum_{i=1}^{l} u_i g_i(x) + \sum_{j=1}^{m} v_j h_j(x) \\
&\quad + \sum_{i\in I} \alpha_i \underbrace{\max(-g_i(x), 0)}_{=0} - \sum_{i\notin I} \alpha_i \tau_i g_i(x) \\
&\quad + \sum_{j\in J} \beta_j \underbrace{|h_j(x)|}_{=0} - \sum_{j\notin J} \beta_j \text{sign}\,[h_j(x)] h_j(x) \\
= &\quad \underbrace{-p^T B p}_{<0} - \sum_{i\notin I} (\alpha_i \tau_i - u_i) g_i(x) - \sum_{j\notin J} (\beta_j - v_j \text{sign}\,[h_j(x)]) |h_j(x)|.
\end{aligned}
$$

Für $g_i(x) < 0$ ist

$$(\alpha_i \tau_i - u_i) g_i(x) = -u_i g_i(x) \geq 0.$$

Wählt man daher $\alpha, \beta > 0$ so groß, dass

$$\alpha_i \geq u_i \quad (i = 1, \ldots, l), \qquad \beta_j \geq |v_j| \quad (j = 1, \ldots, m),$$

so ist

$$\Psi'_{\alpha\beta}(x; p) \leq -p^T B p < 0,$$

also p eine Abstiegsrichtung in x für die Straffunktion $\Psi_{\alpha\beta}$. Das Lemma ist damit bewiesen. □ □

Bemerkungen: Das obige Lemma findet man auch bei P. Spellucci (1993, S. 477). Merkwürdigerweise wird hier vorausgesetzt, dass für das quadratische Hilfsproblem ($Q_{x,B}$) die Slatersche Constraint Qualification erfüllt ist. Für nichtlineare Optimierungsaufgaben, bei denen nur Ungleichungen als Restriktionen auftreten, findet man ihn auch bei S.-P. Han (1977, Theorem 3.1). Hingewiesen sei noch darauf, dass man z. B. bei dem Verfahren von Goldfarb-Idnani auch die zu einer Lösung gehörenden Lagrange-Multiplikatoren mitgeliefert bekommt, so dass es relativ einfach ist, geeignete Vektoren α, β zu bestimmen. Schließlich soll noch auf die Frage eingegangen werden, unter welchen Voraussetzungen das quadratische Hilfsproblem ($Q_{x,B}$) notwendigerweise zulässig ist. Hier gilt:

- Ist g (komponentenweise) konvex und h affin linear, ist ferner das Programm (P) zulässig, existiert also ein $\hat{x} \in \mathbb{R}^n$ mit $g(\hat{x}) \leq 0$ und $h(\hat{x}) = 0$, so ist das

quadratische Programm $(\mathrm{Q}_{x,B})$ für jedes $x \in \mathbb{R}^n$ zulässig, d. h. für jedes $x \in \mathbb{R}^n$ existiert ein $p \in \mathbb{R}^n$ mit

$$g(x) + g'(x)p \leq 0, \qquad h(x) + h'(x)p = 0.$$

Um dies einzusehen, braucht man offenbar nur $p := \hat{x} - x$ zu setzen. □

Die folgende Aussage (treten nur Ungleichungen in den Restriktionen auf, so findet man dieses Ergebnis bei S.-P. Han (1977, Lemma 3.3)) ist für Konvergenzuntersuchungen nützlich:

Lemma 1.9 *Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ stetig differenzierbar, $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ (komponentenweise) konvex und stetig differenzierbar, $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear. Ferner wird vorausgesetzt:*

(a) Es existiert ein $\hat{x} \in \mathbb{R}^n$ mit $g(\hat{x}) < 0$, $h(\hat{x}) = 0$.

(b) Die Abbildung h ist surjektiv.

Ist dann $L \subset \mathbb{R}^n$ eine kompakte Menge und $\gamma \leq \delta$ positive Zahlen, so existiert zu einem beliebigen Paar (x, B) mit $x \in L$ und einer symmetrischen, positiv definiten Matrix $B \in \mathbb{R}^{n\times n}$ mit

$$\gamma \|z\|^2 \leq z^T B z \leq \delta \|z\|^2 \qquad \text{für alle } z \in \mathbb{R}^n$$

eine Zahl $r = r(L, \gamma, \delta)$ mit der Eigenschaft: Ist p die Lösung des quadratischen Programms

$$(\mathrm{Q}_{x,B}) \qquad \begin{cases} \text{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T B p \quad \text{unter den Nebenbedingungen} \\ g(x) + g'(x)p \leq 0, \qquad h(x) + h'(x)p = 0, \end{cases}$$

und sind (u, v) zugehörige Lagrange-Vektoren, so ist $\max(\|u\|, \|v\|) \leq r$. Ferner existiert auch eine Konstante $q = q(L, \gamma, \delta)$ mit $\|p\| \leq q$.

Beweis: Zunächst ist das Problem $(\mathrm{Q}_{x,B})$ natürlich wegen der vorigen Bemerkung zulässig, da wir ja mit (a) insbesondere die Zulässigkeit von (P) vorausgesetzt haben. Das Paar (u, v) genügt den Bedingungen

$$u \geq 0, \qquad \nabla f(x) + Bp + g'(x)^T u + h'(x)^T v = 0, \qquad u^T[g(x) + g'(x)p] = 0,$$

ferner ist natürlich

$$g(x) + g'(x)p \leq 0, \qquad h(x) + h'(x)p = 0.$$

Man definiere $\hat{p} := \hat{x} - x$. Dann ist $\hat{p}$ zulässig für das quadratische Programm $(\mathrm{Q}_{x,B})$. Ferner ist

$$\begin{aligned} \nabla f(x)^T \hat{p} + \tfrac{1}{2}\hat{p}^T B \hat{p} - \nabla f(x)^T p - \tfrac{1}{2} p^T B p &= [\nabla f(x) + Bp]^T(\hat{p} - p) \\ &\quad + \tfrac{1}{2}(\hat{p} - p)^T B(\hat{p} - p) \\ &= -u^T g'(x)(\hat{p} - p) + \tfrac{1}{2}(\hat{p} - p)^T B(\hat{p} - p) \\ &\geq -u^T g'(x)(\hat{p} - p) \\ &= -u^T[g(x) + g'(x)\hat{p}] + \underbrace{u^T[g(x) + g'(x)p]}_{=0} \\ &\geq -u^T g(\hat{x}) \\ &\geq \min_{i=1,\ldots,l}[-g_i(\hat{x})]\, \|u\|_1. \end{aligned}$$

Mit

$$\eta := \min_{i=1,\dots,l}[-g_i(\hat{x})]$$

ist also

$$\eta \,\|u\|_1 \le \nabla f(x)^T\hat{p} + \tfrac{1}{2}\,\hat{p}^T B\hat{p} - \nabla f(x)^T p - \tfrac{1}{2}p^T Bp.$$

Weiter ist

$$\nabla f(x)^T p + \tfrac{1}{2}p^T Bp \ge -\tfrac{1}{2}\nabla f(x)^T B^{-1}\nabla f(x),$$

denn die quadratische Funktion $\phi(q) := \nabla f(x)^T q + \frac{1}{2}q^T Bq$ nimmt auf dem gesamten $\mathbb{R}^n$ ihr Minimum in $q^* := -B^{-1}\nabla f(x)$ an. Daher ist

$$\begin{aligned}
\|u\|_1 &\le \frac{1}{\eta}[\nabla f(x)^T\hat{p} + \tfrac{1}{2}\hat{p}^T B\hat{p} + \tfrac{1}{2}\nabla f(x)^T B^{-1}\nabla f(x)] \\
&\le \frac{1}{\eta}[\|\nabla f(x)\|\,\|\hat{p}\| + \tfrac{1}{2}\delta\,\|\hat{p}\|^2 + \tfrac{1}{2}\|\nabla f(x)\|^2/\gamma] \\
&\le \frac{1}{\eta}[\zeta\xi + \tfrac{1}{2}\delta\xi^2 + \tfrac{1}{2}\zeta^2/\gamma],
\end{aligned}$$

wobei wir zur Abkürzung

$$\zeta := \max_{x\in L}\|\nabla f(x)\|, \qquad \xi := \max_{x\in L}\|\hat{x} - x\|$$

gesetzt haben. Zu zeigen bleibt die Beschränktheit der zu den Gleichungen gehörenden Lagrange-Vektoren. Hierzu beachten wir, dass wir die Surjektivität der Abbildung $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ vorausgesetzt haben. Da h affin linear ist, hat die (von x unabhängige) Funktionalmatrix $A := h'(x)$ vollen Rang m, insbesondere ist $\mathrm{Kern}\,(A^T) = \{0\}$. Dies wiederum impliziert, dass $AB^{-1}A^T$ nichtsingulär ist. Wegen

$$AB^{-1}\nabla f(x) + \underbrace{Ap}_{=-h(x)} + AB^{-1}g'(x)^T u + AB^{-1}A^T v = 0$$

ist

$$v = -(AB^{-1}A^T)^{-1}[AB^{-1}\nabla f(x) - h(x) + AB^{-1}g'(x)^T u],$$

woraus man abliest, dass auch $\|v\|$ durch eine Konstante abgeschätzt werden kann, die nur von (L, γ, δ) abhängt. Die Beschränktheit von $\|p\|$ unabhängig von (x, B) erkennt man aus

$$p = -B^{-1}[\nabla f(x) + g'(x)^T u + h'(x)^T v].$$

Damit ist das Lemma bewiesen. □ □

Wir lassen bei den folgenden Überlegungen möglichst den Iterationsindex k weg, denken aber daran, dass $x = x_k$ eine aktuelle Näherung ist, die Suchrichtung $p = p_k$ mit Hilfe einer symmetrischen, positiv definiten Matrix $B = B_k$ durch Lösen eines quadratischen Programms gewonnen wird und natürlich auch die Schrittweite t vom Iterationsindex abhängt. Von einer aktuellen Näherung x, für die das quadratische Programm $(\mathrm{Q}_{x,B})$ zulässig ist, geht man also in Richtung p, wobei p die Lösung von $(\mathrm{Q}_{x,B})$ ist. Als neue Näherung wird man daher $x_+ := x + tp$ mit geeigneter Schrittweite $t > 0$ bestimmen. Es stellen sich nun natürlich die folgenden Fragen:

- Wie sollte die Schrittweite $t > 0$ bestimmt werden?
- Durch welche Update-Formel sollte die neue Matrix B_+ berechnet werden?

Zunächst zur Frage nach einer geeigneten Schrittweite. Bei S.-P. Han (1977) wird eine asymptotisch exakte Schrittweite vorgeschlagen, d. h. mit vorgegebenen positiven ϵ, δ wird $t > 0$ so bestimmt, dass

$$\Psi_{\alpha\beta}(x + tp) \leq \min_{s\in[0,\delta]} \Psi_{\alpha\beta}(x + sp) + \epsilon,$$

wobei die vom Iterationsindex abhängenden Zahlen ϵ hinreichend schnell gegen Null konvergieren. Hierauf wollen wir aber nicht eingehen, sondern die Armijo-Schrittweite auf den vorliegenden Fall übertragen. Mehrere Autoren haben hierzu beigetragen, u. a. S.-P. Han (1981) im Zusammenhang mit unrestringierten Min-Max-Optimierungsaufgaben. Ausgangspunkt ist die im Beweis von Lemma 1.8 gemachte Beobachtung, dass für hinreichend große positive α, β die Richtungsableitung $\Psi'_{\alpha\beta}(x; p)$ in Richtung p, der Lösung des quadratischen Hilfsprogramms $(\mathrm{Q}_{x,B})$, der Abschätzung

$$\Psi'_{\alpha\beta}(x; p) \leq -p^T Bp$$

genügt, insbesondere also also eine Abstiegsrichtung für die L_1-Straffunktion ist. Die Armijo-Schrittweite t bestimme man durch den folgenden Algorithmus:

- Seien $\sigma \in (0, 1)$ und $0 < l \leq u < 1$ gegeben, setze $\rho_0 := 1$.
- Für $j = 0, 1, \ldots$:
  - Falls
    $$\Psi_{\alpha\beta}(x + \rho_j p) \leq \Psi_{\alpha\beta}(x) - \sigma\rho_j p^T Bp,$$
    dann: $t := \rho_j$, STOP.
  - Andernfalls: Wähle $\rho_{j+1} \in [l\rho_j, u\rho_j]$.

Z. B. kann man $l = u = \frac{1}{2}$ setzen, was bedeutet, dass man die Schrittweite halbiert, bis eine gewisse Abschätzung erfüllt ist. Es ist klar, dass die Armijo-Schrittweite existiert bzw. der obige Algorithmus nach endlich vielen Schritten abbricht. Denn wäre die zu testende Ungleichung für kein j erfüllt, so wäre $\{\rho_j\} \subset \mathbb{R}_+$ eine Nullfolge und

$$-\sigma p^T Bp \leq \lim_{j\to\infty} \frac{\Psi_{\alpha\beta}(x + \rho_j p) - \Psi_{\alpha\beta}(x)}{\rho_j} = \Psi'_{\alpha\beta}(x; p) \leq -p^T Bp,$$

was wegen $\sigma \in (0, 1)$ und $p^T Bp > 0$ ein Widerspruch ist. Der nächste Punkt, der untersucht werden muß, ist die Abschätzung[11] der Zielfunktionsminderung. Genauer sei $x_+ := x + tp$, wobei $p \neq 0$ die Lösung des quadratischen Hilfsprogramms $(\mathrm{Q}_{x,B})$ und $t > 0$ die zugehörige Armijo-Schrittweite ist. Wir setzen voraus, dass mit vom Iterationsindex unabhängigen positiven Konstanten γ, δ gilt, dass

$$\gamma \|z\|^2 \leq z^T Bz \leq \delta \|z\|^2 \qquad \text{für alle } z \in \mathbb{R}^n.$$

[11] Siehe auch P. Spellucci (1993, S. 479 ff.).

Weiter setzen wir voraus, dass die Niveaumenge

$$L_{\alpha\beta} := \{x \in \mathbb{R}^n : \Psi_{\alpha\beta}(x) \le \Psi_{\alpha\beta}(x_0)\}$$

(mit einem gewissen $x_0 \in \mathbb{R}^n$) kompakt ist. Ist dann $\hat{t} > 0$ die erste positive Nullstelle von $\Psi_{\alpha\beta}(x+tp) - \Psi_{\alpha\beta}(x)$ (die Existenz ist wegen der Kompaktheit von $L_{\alpha\beta}$ gesichert), so ist $x + sp \in L_{\alpha\beta}$ für alle $s \in [0, \hat{t}]$ und daher

$$\begin{aligned} g_i(x+tp) &= g_i(x) + t\nabla g_i(x)^T p + \int_0^t [\nabla g_i(x+sp) - \nabla g_i(x)]^T p \, ds \\ &= (1-t)g_i(x) + t\underbrace{[g_i(x) + \nabla g_i(x)^T p]}_{\le 0} + \int_0^t [\nabla g_i(x+sp) - \nabla g_i(x)]^T p \, ds \\ &\le (1-t)g_i(x) + \tfrac{1}{2}Ct^2 \|p\|^2, \end{aligned}$$

wobei C eine (o. B. d. A. von i unabhängige) Lipschitzkonstante von $\nabla g_i(\cdot)$ auf $L_{\alpha\beta}$ ist. Für alle $t \in [0, \min(1, \hat{t})]$ ist daher

$$\max(g_i(x+tp), 0) \le (1-t)\max(g_i(x), 0) + \tfrac{1}{2}Ct^2 \|p\|^2, \qquad i = 1, \ldots, l.$$

Für alle $t \in [0, \hat{t}]$ ist entsprechend

$$\begin{aligned} h_j(x+tp) &= h_j(x) + t\nabla h_j(x)^T p + \int_0^t [\nabla h_j(x+sp) - \nabla h_j(x)]^T p \, ds \\ &= (1-t)h_j(x) + t\underbrace{[h_j(x) + \nabla h_j(x)^T p]}_{=0} \int_0^t [\nabla h_j(x+sp) - \nabla h_j(x)]^T p \, ds \\ &\le (1-t)h_j(x) + \tfrac{1}{2}Ct^2 \|p\|^2, \end{aligned}$$

wobei C auch noch gemeinsame Lipschitzkonstante der Gradienten ∇h_j auf der Niveaumenge $L_{\alpha\beta}$ ist. Für alle $t \in [0, \min(1, \hat{t})]$ ist daher

$$|h_j(x+tp)| \le (1-t)|h_j(x)| + \tfrac{1}{2}Ct^2 \|p\|^2, \qquad j = 1, \ldots, m.$$

Weiter ist

$$f(x+tp) \le f(x) + t\nabla f(x)^T + \tfrac{1}{2}Ct^2 \|p\|^2$$

für alle $t \in [0, \hat{t}]$, wobei schließlich die Konstante $C > 0$ auch noch so groß gewählt ist, dass sie als Lipschitzkonstante von $\nabla f(\cdot)$ auf $L_{\alpha\beta}$ dienen kann. Für alle $t \in [0, \min(1, \hat{t})]$ ist daher schließlich

$$\begin{aligned} \Psi_{\alpha\beta}(x+tp) &\le f(x) + t\nabla f(x)^T p + (1-t)\sum_{i=1}^{l} \alpha_i \max(g_i(x), 0) \\ &\qquad + (1-t)\sum_{j=1}^{m} \beta_j |h_j(x)| + \tfrac{1}{2}C(1+l+m)t^2 \|p\|^2 \\ &= \Psi_{\alpha\beta}(x) - t\Big[\sum_{i=1}^{l} \alpha_i \max(g_i(x), 0) + \sum_{j=1}^{m} \beta_j |h_j(x)| - \nabla f(x)^T p\Big] \end{aligned}$$

$$
\begin{aligned}
&\quad + \tfrac{1}{2}C(1+l+m)t^2\,\|p\|^2 \\
&= \Psi_{\alpha\beta}(x) - t\Big[p^TBp + \sum_{i=1}^{l}[\alpha_i \max(g_i(x),0) + u_i\nabla g_i(x)^Tp] \\
&\qquad + \sum_{j=1}^{m}[\beta_j|h_j(x)| + v_j\nabla h_j(x)^Tp]\Big] + \tfrac{1}{2}Ct^2(1+l+m)\,\|p\|^2 \\
&\le \Psi_{\alpha\beta}(x) - t\Big[p^TBp + \sum_{i=1}^{l}[\alpha_i \max(g_i(x),0) - u_ig_i(x)] \\
&\qquad + \sum_{j=1}^{m}[\beta_j|h_j(x)| - v_jh_j(x)|] + \tfrac{1}{2}C(1+l+m)t^2\,\|p\|^2 \\
&\le \Psi_{\alpha\beta}(x) - tp^TBp + \tfrac{1}{2}Ct^2(1+l+m)\,\|p\|^2,
\end{aligned}
$$

wenn $\alpha, \beta > 0$ so groß gewählt sind, dass

$$\alpha_i \ge u_i \quad (i = 1,\dots,l), \qquad \beta_j \ge |v_j| \quad (j = 1,\dots,m).$$

Ist $\hat{t} \le 1$, so folgt hieraus (setze $t := \hat{t}$), dass

$$\hat{t} \ge t^* := \frac{2}{C(l+m+1)}\,\frac{p^TBp}{\|p\|^2}.$$

Für alle $t \in [0, \min(1, t^*)]$ ist daher

$$(*) \qquad \Psi_{\alpha\beta}(x+tp) \le \Psi(x) - tp^TBp + \tfrac{1}{2}C(l+m+1)t^2\,\|p\|^2.$$

Angenommen, der Test zur Bestimmung der Armijo-Schrittweite ist schon ganz am Anfang erfüllt. Dann ist $t = 1$ und daher

$$\Psi_{\alpha\beta}(x) - \Psi_{\alpha\beta}(x+tp) \ge \sigma p^TBp \ge \sigma\gamma\,\|p\|^2.$$

Nun nehmen wir an, der Test sei nicht schon am Anfang erfüllt. Mit $s := \rho_{j-1}$ gelten dann die Ungleichungen

$$\Psi_{\alpha\beta}(x+tp) \le \Psi_{\alpha\beta}(x) - \sigma tp^TBp, \qquad \Psi_{\alpha\beta}(x+sp) > \Psi_{\alpha\beta}(x) - \sigma sp^TBp.$$

Ferner ist $ls \le t$. Ist $s \le \min(1, t^*)$, so liefert $(*)$, dass

$$
\begin{aligned}
\Psi_{\alpha\beta}(x) - \sigma sp^TBp &< \Psi_{\alpha\beta}(x+sp) \\
&\le \Psi_{\alpha\beta}(x) - sp^TBp + \tfrac{1}{2}C(l+m+1)s^2\,\|p\|^2,
\end{aligned}
$$

daher

$$\frac{2(1-\sigma)}{C(l+m+1)}\,\frac{p^TBp}{\|p\|^2} \le s$$

und folglich

$$\Psi_{\alpha\beta}(x) - \Psi_{\alpha\beta}(x+tp) \ge \sigma tp^TBp \ge \sigma lsp^TBp \ge \frac{2l\sigma(1-\sigma)}{C(l+m+1)}\Big(\frac{p^TBp}{\|p\|}\Big)^2.$$

Nun sei $s > \min(1, t^*)$. Wegen $s \leq 1$ ist dann $s > t^*$. Damit ist in diesem Fall $t \geq ls > lt^*$ und daher

$$\Psi_{\alpha\beta}(x) - \Psi_{\alpha\beta}(x + tp) \geq \sigma t p^T B p \geq l\sigma t^* p^T B p \geq \frac{2l\sigma}{C(l+m+1)} \Big(\frac{p^T B p}{\|p\|}\Big)^2.$$

Wir haben daher die Existenz einer Konstanten $\theta > 0$ erhalten, die von der aktuellen Näherung x, der Matrix B (deren Eigenwerte in $[\gamma, \delta]$ liegen) und der Lösung p des quadratischen Programms $(\mathrm{Q}_{x,B})$ unabhängig ist, mit

$$\Psi_{\alpha\beta}(x) - \Psi_{\alpha\beta}(x + tp) \geq \theta \, \|p\|^2.$$

Hierbei wurde noch vorausgesetzt, dass die Niveaumenge $L_{\alpha\beta}$ kompakt ist und α, β hinreichend groß gewählt sind. Sei $\{x_k\}$ eine Folge, die durch das Verfahren erzeugt ist. Wir wollen uns überlegen, daß jeder Häufungspunkt x^* von $\{x_k\}$ (wegen $\{x_k\} \subset L_{\alpha\beta}$ existiert mindestens ein Häufungspunkt) eine (zulässige) kritische Lösung des gegebenen nichtlinearen Optimierungsproblems (P) ist. Zunächst ist $\{\Psi_{\alpha\beta}(x_k)\}$ eine monoton fallende, nach unten beschränkte Folge. Wegen $\Psi_{\alpha\beta}(x_k) - \Psi_{\alpha\beta}(x_{k+1}) \geq \theta \, \|p_k\|^2$ ist $\lim_{k\to\infty} p_k = 0$. Da weiter

$$g(x_k) + g'(x_k)p_k \leq 0, \qquad h(x_k) + h'(x_k)p_k = 0,$$

ist x^* zulässig für (P). Aus

$$u_k \geq 0, \quad \nabla f(x_k) + B_k p_k + g'(x_k)^T u_k + h'(x_k)^T v_k = 0, \quad u_k^T [g(x_k) + g'(x_k)^T u_k] = 0$$

folgt mit beschränkten Folgen $\{u_k\}$ und $\{v_k\}$ (hinreichende Bedingungen hierfür haben wir oben angegeben) die Existenz von $u^* \in \mathbb{R}^l$, $v^* \in \mathbb{R}^m$ mit

$$u^* \geq 0, \qquad \nabla f(x^*) + f'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad (u^*)^T g(x^*).$$

Zusammen mit der Zulässigkeit von x^* bedeutet dies, daß x^* eine kritische Lösung von (P) ist.

Bemerkung: Wir haben beschrieben, wie man die L_1-Straffunktion zur Schrittweitenbestimmung in einem Verfahren, bei dem die Richtungen durch Lösen eines quadratischen Programms berechnet werden, benutzen kann. Wir haben einige Punkte einer möglichen Konvergenzanalyse angesprochen (Zulässigkeit der Hilfsprobleme, Beschränktheit der Lagrange-Multiplikatoren, Verminderung der L_1-Straffunktion bei Verwendung der Armijo-Schrittweite), verzichten aber auf die genaue Formulierung eines Konvergenzsatzes. Eine Beschreibung eines praktikablen Verfahrens (allerdings ohne theoretische Konvergenzergebnisse) findet man bei M. J. D. Powell (1978)[12]. Hier werden auch Vorschläge für das Updaten der Matrix B gemacht. Im wesentlichen bedeutet dies: Ist x eine aktuelle Näherung und B die aktuelle positiv definite Matrix, p die Lösung des quadratischen Programms $(\mathrm{Q}_{x,B})$ mit einem zugehörigen Paar von

[12] Powell, M. J. D. (1978) "A fast algorithm for nonlinearly constrained optimization calculations." In *Numerical Analysis*, (G. A. Watson, ed.), Lecture Notes in Mathematics 630, Springer-Verlag, 144–157.

Lagrange-Multiplikatoren (u, v) (wie schon erwähnt, liefert z. B. das Verfahren von Goldfarb-Idnani diese mit), berechnet man dann eine geeignete Schrittweite t (etwa nach Armijo mit der L_1-Straffunktion) und setzt $x_+ := x + tp$, ist schließlich die Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ wie üblich durch

$$L(x, u, v) := f(x) + g(x)^T u + h(x)^T v$$

definiert, so liegt es nahe,

$$s := x_+ - x, \qquad y := \nabla_x L(x_+, u, v) - \nabla_x L(x, u, v)$$

zu setzen und anschließend zur Bestimmung von B_+ den BFGS-Update, also

$$B_+ := B - \frac{(Bs)(Bs)^T}{s^T Bs} + \frac{yy^T}{y^T s}$$

zu machen. Bei dieser Vorgehensweise ist aber nicht gesichert, dass mit B auch B_+ positiv definit ist. Denn bekanntlich ist B_+ positiv definit, wenn $y^T s > 0$, was etwa bei einer unrestringierten Optimierungsaufgabe und gleichmäßig konvexer Zielfunktion automatisch der Fall ist. Bei der obigen Definition von y kann aber $y^T s \leq 0$ eintreten. Daher schlägt Powell eine Modifikation vor, bei der y ersetzt wird durch den Vektor

$$z := \theta y + (1 - \theta) Bs,$$

wobei $\theta \in [0, 1]$ möglichst nahe bei 1 unter der Nebenbedingung $y^T z \geq 0.2 s^T Bs$ gewählt wird. Dies führt auf

$$\theta := \begin{cases} 1, & \text{falls} \quad y^T s \geq 0.2 s^T Bs, \\ \dfrac{0.8 s^T Bs}{s^T Bs - y^T s}, & \text{falls} \quad y^T s < 0.2 s^T Bs. \end{cases}$$

Anschließend macht man den BFGS-Update

$$B_+ := B - \frac{(Bs)(Bs)^T}{s^T Bs} + \frac{zz^T}{z^T s}$$

und ist sich durch diese Konstruktion sicher, dass mit B auch B_+ positiv definit ist. □

5.1.4 Aufgaben

1. Gegeben sei das quadratische Programm

   (P) Minimiere $f(x) := c^T x + \frac{1}{2} x^T Q x$ auf $M := \{x \in \mathbb{R}^n : h(x) := Ax - b = 0\}$

   mit symmetrischem, positiv definitem $Q \in \mathbb{R}^{n \times n}$ und $A \in \mathbb{R}^{m \times n}$ mit Rang $(A) = m$. Man bilde die quadratische Straffunktion Φ_σ und berechne das unrestringierte Minimum $x(\sigma)$ von Φ_σ. Man zeige, dass $x^* := \lim_{\sigma \to \infty} x(\sigma)$ existiert und die eindeutige Lösung von (P) ist. Ferner überlege man sich, dass auch der Lagrange-Multiplikator zu x^* eindeutig ist und durch $\lim_{\sigma \to \infty} \sigma h(x(\sigma))$ gegeben ist.

2. Gegeben sei das quadratische Programm

$$\text{(P)} \quad \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : h(x) := Ax - b = 0\}$$

mit symmetrischem, positiv definitem $Q \in \mathbb{R}^{n\times n}$ und $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$. Man betrachte die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}^*_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + (y^*)^T h(x) + \tfrac{1}{2}\sigma \|h(x)\|^2, \quad x \in \mathbb{R}^n,$$

wobei y^* der (eindeutige) Lagrange-Multiplikator zur Lösung x^* von (P) ist. Man zeige, dass x^* für jedes $\sigma \geq 0$ die eindeutige Lösung von (P^*_σ) ist.

3. Gegeben sei (siehe P. Spellucci (1993, S. 394)) die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := x_1^2 + 4x_1x_2 + 5x_2^2 - 10x_1 - 20x_2 \quad \text{auf} \\ M := \{x \in \mathbb{R}^2 : h(x) := x_1 + x_2 - 2 = 0\}. \end{cases}$$

Dieser Aufgabe ordne man die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \tfrac{1}{2}\sigma h(x)^2, \quad x \in \mathbb{R}^2$$

zu. Man bestimme die Lösung $x(\sigma)$ von (P_σ) und bestätige die Aussage von Aufgabe 1, berechne also z. B. die Lösung x^* von (P) und weise $x^* = \lim_{\sigma\to\infty} x(\sigma)$ nach. Weiter bestimme man den zu x^* gehörenden Lagrange-Multiplikator y^* und zeige, dass $\lim_{\sigma\to\infty} \sigma h(x(\sigma)) = y^*$.

4. Gegeben sei die Optimierungsaufgabe (siehe P. Spellucci (1993, S. 453))

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := (x_1 + 2)^2 + 9(x_2 + 3)^2 \quad \text{unter der Nebenbedingung} \\ g(x) := 1 - x_1 - x_2 \leq 0. \end{cases}$$

   (a) Man berechne die Lösung x^* von (P) und einen zugehörigen Lagrange-Multiplikator u^*.

   (b) Bei gegebenem $\sigma > 0$ bestimme man die Lösung $x(\sigma)$ der unrestringierten Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \frac{\sigma}{2} \max(g(x), 0)^2, \quad x \in \mathbb{R}^2$$

   und zeige, dass $\lim_{\sigma\to\infty} x(\sigma) = x^*$.

   (c) Wie erhält man durch Lösen von (P_σ) für hinreichend großes σ eine Näherung für den Lagrange-Multiplikator u^*?

5. Gegeben sei die zulässige, restringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$$

und hierzu die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma \underbrace{\left(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1 \right)}_{=:S(x)}, \quad x \in \mathbb{R}^n.$$

Existiert dann ein $\sigma^* > 0$ und ein $x^* \in \mathbb{R}^n$ derart, daß x^* für alle $\sigma \geq \sigma^*$ eine (globale) Lösung von (P_σ) ist, so ist x^* eine Lösung von (P), insbesondere also zulässig für (P).

Hinweis: Siehe S.-P. Han, O. L. Mangasarian (1979, Theorem 4.1)[13], der Beweis ist einfach.

6. Gegeben sei die restringierte Optimierungsaufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M,$$

wobei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ stetig und $M \subset \mathbb{R}^n$ abgeschlossen ist. Mit $\sigma > 0$ betrachte man hierzu die unrestringierte Aufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad P_\sigma(x) := f(x) + \sigma S(x), \quad x \in \mathbb{R}^n,$$

wobei $S\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ stetig ist mit

$$S(x) \begin{cases} = 0 & \text{für} \quad x \in M, \\ > 0 & \text{für} \quad x \notin M. \end{cases}$$

Ist dann $x^* \in M$ eine isolierte, lokale Lösung von (P), so existiert ein $\sigma^* > 0$ derart, dass es zu jedem $\sigma \geq \sigma^*$ ein Paar $(x(\sigma), \epsilon(\sigma)) \in \mathbb{R}^n \times \mathbb{R}_+$ mit

$$x(\sigma) \in B(x^*; \epsilon(\sigma)), \qquad \lim_{\sigma \to \infty} \epsilon(\sigma) = 0$$

und

$$P_\sigma(x(\sigma)) \leq P_\sigma(x) \qquad \text{für alle } x \in B(x^*; \epsilon(\sigma))$$

gilt, wobei $B(x^*; \epsilon(\sigma))$ die offene (euklidische) Kugel um x^* mit dem Radius $\epsilon(\sigma)$ bedeutet.

Hinweis: Siehe T. Pietzykowski (1970)[14]. Der Beweis dort ist überraschend verwickelt. Wer schafft einen einfacheren?

7. Gegeben sei die Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \begin{cases} \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma\left(\displaystyle\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1\right) \\ \text{auf} \quad M := \{x \in \mathbb{R}^n : l \leq x \leq u\}. \end{cases}$$

Hierbei seien $l, u \in \mathbb{R}^n$ zwei Vektoren mit $l < u$ (eine Verwechslung der unteren (lower) Schranke l mit der Anzahl l der g_i sollte vermieden werden). Man übertrage den Begriff der kritischen Lösung auf die Aufgabe (P_σ) und gebe notwendige und hinreichende Bedingungen dafür an, dass ein $x^* \in M$ kritische Lösung von (P_σ) ist.

8. Der restringierten, nichtlinearen Optimierungsaufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$$

[13] Han, S.-P. and O. L. Mangasarian (1979) "Exact penalty functions in nonlinear programming." Mathematical Programming 17, 251–269.

[14] Pietrzykowski, T. (1970) "The potential method for conditional maxima in the locally compact metric spaces." Numer. Math. 14, 325–329.

mit glatten $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ordne man die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma P(x), \quad x \in \mathbb{R}^n$$

mit

$$P(x) := \max(0, g_1(x), \ldots, g_l(x), |h_1(x)|, \ldots, |h_m(x)|)$$

zu. Man berechne die Richtungsableitung $\Psi'_\sigma(x^*; p)$ in einem Punkt $x^* \in \mathbb{R}^n$ in die Richtung $p \in \mathbb{R}^n$ und gebe notwendige und hinreichende Bedingungen dafür an, dass x^* eine kritische Lösung von (P_σ) ist, also $\Psi'_\sigma(x^*; p) \geq 0$ für alle $p \in \mathbb{R}^n$ gilt.

9. Gegeben sei die *konvexe* Optimierungsaufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\},$$

die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sei also konvex, die Restriktionabbildung $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ komponentenweise konvex und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear. Sei $x^* \in M$ eine Lösung von (P), ferner gelte die Slatersche Constraint Qualification, es existiere also $\hat{x} \in \mathbb{R}^n$ mit $g(\hat{x}) < 0$ und $h(\hat{x}) = 0$ und die Abbildung h sei surjektiv. Ist dann (u^*, v^*) eine Lösung des zu (P) dualen Programms, so ist x^* für alle $\sigma \geq \sigma^* := \max(\|u^*\|_\infty, \|v^*\|_\infty)$ eine globale Lösung von

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma\left(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1\right), \quad x \in \mathbb{R}^n.$$

Hinweis: Eine etwas allgemeinere Version der obigen Aussage findet man bei S.-P. Han, O. L. Mangasarian (1979, Theorem 4.9). Man sollte aber nicht dort nachsehen, sondern den einfachen Beweis selber finden.

5.2 Barriere- und Straffunktionen bei konvexen Optimierungsaufgaben

5.2.1 Einführung

Gegeben sei die Optimierungsaufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

Hierbei setzen wir generell voraus:

(V1) Die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildung $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sind stetig differenzierbar und (komponentenweis) konvex, die Abbildung $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ist affin linear, also (P) ein konvexes Programm. Ferner habe die (konstante) Funktionalmatrix h' den Rang m.

(V2) Die Menge M_{opt} der Lösungen von (P) ist nichtleer und kompakt, insbesondere also (P) zulässig.

Der Aufgabe (P) wird mittels eines Parameters $\sigma > 0$ eine Schar von Optimierungsaufgaben der Form

$$(\mathrm{P}_\sigma) \qquad \begin{cases} \text{Minimiere} \quad f_\sigma(x) := f(x) + \dfrac{1}{\sigma}\displaystyle\sum_{i=1}^{l} \theta(\sigma g_i(x)) \quad \text{auf} \\ \qquad M_\sigma := \{x \in \mathbb{R}^n : \sigma g(x) < \eta e,\ h(x) = 0\} \end{cases}$$

zugeordnet. Hierbei ist $\theta \in C^2(-\infty, \eta)$ mit $0 \le \eta \le \infty$ eine Funktion mit gewissen, später zu präzisierenden Eigenschaften. Zu diesen soll aber auf alle Fälle gehören, dass θ auf $(-\infty, \eta)$ konvex und monoton nicht fallend und damit die Zielfunktion f_σ von (P_σ) auf M_σ konvex ist. Generell setzen wir weiter voraus, dass für $\eta = 0$ das relative Innere

$$M^0 := \{x \in \mathbb{R}^m : g(x) < 0,\ h(x) = 0\}$$

von M nichtleer ist. Damit ist gesichert, dass (P_σ) für alle $\sigma > 0$ zulässig ist. Wir wollen Antworten auf die folgenden Fragen geben:

1. Unter welchen Voraussetzungen ist auch die Lösungsmenge $(M_\sigma)_{\mathrm{opt}}$ von (P_σ) für alle $\sigma > 0$ nichtleer und kompakt?

2. Sei $\{\sigma_k\} \subset \mathbb{R}_+$ eine Folge mit $\sigma_k \to \infty$, für alle $k \in \mathbb{N}$ sei ferner $x_k \in M_{\sigma_k}$ eine Lösung von (P_{σ_k}). Unter welchen Voraussetzungen ist $\{x_k\}$ beschränkt und gehört jeder Häufungspunkt x^* von $\{x_k\}$ zu M_{opt}?

3. Unter welchen Voraussetzungen besitzt das Problem (P_σ) für jedes $\sigma > 0$ genau eine Lösung $x_\sigma \in M_\sigma$?

4. Unter welchen Voraussetzungen existiert $x_\infty = \lim_{\sigma\to\infty} x_\sigma$ und ist eine (gewisse) Lösung von (P)?

5. Eine Lösung $x_\sigma \in M_\sigma$ von (P_σ) ist charakterisiert durch die Existenz eines $v_\sigma \in \mathbb{R}^m$ mit
$$\nabla f(x_\sigma) + g'(x_\sigma)^T u_\sigma + (h')^T v_\sigma = 0$$
mit
$$(u_\sigma)_i := \psi'(\sigma g_i(x_\sigma)), \qquad i = 1, \ldots, l.$$
Unter welchen Voraussetzungen existiert $(u_\infty, v_\infty) = \lim_{\sigma\to\infty}(u_\sigma, v_\sigma)$ und ist eine (gewisse) Lösung des zu (P) dualen Programms?

Die ersten beiden Fragen sind u. a. von A. Auslender, R. Cominetti, M. Haddou (1997)[15] behandelt worden. Die Eindeutigkeit einer Lösung von (P_σ), wenn das Ausgangsproblem ein quadratisch restringiertes quadratisches Programm ist, ist z. B. von A. V. Fiacco (1995, Theorem 4)[16] für die klassischen logarithmischen Barrieren (hier ist $\eta = 0$

[15] Auslender, A., R. Cominetti and M. Haddou (1997) "Asymptotic analysis for penalty and barrier methods in convex and linear programming." Mathematics of Operations Research 22, 43–62.

[16] Fiacco, A. V. (1995) "Objective function and logarithmic barrier function properties in convex programming: level sets, solution attainment and strict convexity." Optimization 34, 213–222.

und $\theta(t) = -\log(-t)$) bewiesen worden. Stetigkeitsaussagen für die primale Trajektorie $\{x_\sigma\}$ bzw. die duale Trajektorie $\{(u_\sigma, v_\sigma)\}$ werden ebenfalls bei A. Auslender, R. Cominetti, M. Haddou (1997) bewiesen, wobei sich diese Autoren auf lineare Programme beschränken. Vorläufer dieser Aussagen (auch nur für lineare Programme) stammen von N. Megiddo (1989)[17] (logarithmische Barrieren) und R. Cominetti, J. San Martin (1994)[18] (exponentielle Strafen, hier ist $\eta = +\infty$ und $\theta(t) = \exp(t)$). Unser Ziel ist es, diese Aussagen weitgehend auf quadratisch restringierte quadratische Programme zu übertragen. Es wird schon bald klar werden, wann wir von f_σ als einer Barriere- bzw. Straffunktion sprechen. Wichtig ist, dass die Restriktionenmenge M_σ des Hilfproblems (P_σ) jedenfalls relativ zu der affin linearen Gleichungsnebenbedingung offen ist, so dass es sich bei (P_σ) im wesentlichen um eine unrestringierte Optimierungsaufgabe handelt.

5.2.2 Existenz einer Lösung des Hilfsproblems

In diesem Unterabschnitt sollen hinreichende Bedingungen an die Funktion θ dafür angegeben werden, dass die Aufgabe (P_σ) eine nichtleere, kompakte Lösungsmenge $(M_\sigma)_{\mathrm{opt}}$ besitzt.

Satz 2.1 *Die Funktion $\theta \in C^2(-\infty, \eta)$ mit $0 \le \eta \le \infty$ sei konvex und monoton nicht fallend. Es gelte*

(A1) $\lim_{t\to\eta-} \theta(t) = +\infty$,

(A2) $\lim_{t\to\infty} \theta(-t)/t = 0$,

und, falls $\eta = \infty$,

(A3) $\lim_{t\to\infty} \theta(t)/t = +\infty$.

Dann ist die Menge $(M_\sigma)_{\mathrm{opt}}$ der Lösungen von (P_σ) nichtleer und kompakt.

Beweis: Bei festem $\sigma > 0$ wählen wir ein $\hat{x} \in M_\sigma$ und bilden die Niveaumenge

$$\hat{L}_\sigma := M_\sigma \cap \{x \in \mathbb{R}^n : f_\sigma(x) \le f_\sigma(\hat{x})\}.$$

Zunächst zeigen wir, dass $\hat{L}_\sigma$ beschränkt ist. Ist dies nicht der Fall, so existiert eine Folge $\{x_k\} \subset \hat{L}_\sigma$ mit $\|x_k\| \to \infty$ und $x_k/\|x_k\| \to p$. Wir wollen zeigen, dass dann $x^* + tp \in M_{\mathrm{opt}}$ für alle $t \ge 0$ mit einem beliebigen $x^* \in M_{\mathrm{opt}}$, was wegen $p \neq 0$ einen Widerspruch zur vorausgesetzten Beschränktheit von M_{opt} ergibt. Offensichtlich ist $p \in \mathrm{Kern}\,(h')$ und daher $h(x^* + tp) = 0$ für alle $t \ge 0$. Wir geben uns ein $t > 0$ beliebig vor. Für alle hinreichend großen k ist $t/\|x_k\| \in (0, 1]$. Wegen der Konvexität von f ist

$$(*) \qquad f\Big(\Big(1 - \frac{t}{\|x_k\|}\Big)x^* + \frac{t}{\|x_k\|}x_k\Big) \le \Big(1 - \frac{t}{\|x_k\|}\Big)f(x^*) + t\frac{f(x_k)}{\|x_k\|}.$$

[17] Megiddo, N. (1989) "Pathways to the optimal set." In *Interior Point and Related Methods* (N. Megiddo, ed.). Springer-Verlag, New York, 131–158.

[18] Comminetti, R. and J. San Martin (1994) "Asymptotic analysis of the exponential penalty trajectory in linear programming." Mathematical Programming 67, 169–187.

Eine entsprechende Ungleichung gilt für g_i, $i = 1, \ldots, l$. Hieran erkennen wir, dass es genügt, die Beziehungen

$$\limsup_{k\to\infty} \frac{f(x_k)}{\|x_k\|} \le 0, \qquad \limsup_{k\to\infty} \frac{g_i(x_k)}{\|x_k\|} \le 0 \quad (i = 1, \ldots, l)$$

nachzuweisen. Denn dann ist $f(x^* + tp) \le f(x^*)$ und $g(x^* + tp) \le g(x^*)$ und damit $x^* + tp \in M_{\text{opt}}$ für alle $t \ge 0$. Wegen $(*)$ bzw. den entsprechenden Ungleichungen für die g_i, $i = 1, \ldots, l$, existiert eine Konstante $c_0 > 0$ mit

$$\frac{f(x_k)}{\|x_k\|} \ge -c_0, \qquad \frac{g_i(x_k)}{\|x_k\|} \ge -c_0 \quad (i = 1, \ldots, l)$$

für alle $k \in \mathbb{N}$. Wegen $f_\sigma(x_k) \le f_\sigma(\hat{x})$ ist

$$\frac{f(x_k)}{\|x_k\|} \le \frac{f_\sigma(\hat{x})}{\|x_k\|} - \frac{1}{\sigma}\sum_{i=1}^{l} \frac{\theta(\sigma g_i(x_k))}{\|x_k\|} \le \frac{f_\sigma(\hat{x})}{\|x_k\|} - \frac{l\theta(-\sigma c_0 \|x_k\|)}{\sigma \|x_k\|} \to 0$$

wegen (A2), daher $\limsup_{k\to\infty} f(x_k)/\|x_k\| \le 0$. Für $\eta < \infty$ ist es einfach,

$$\limsup_{k\to\infty} \frac{g_i(x_k)}{\|x_k\|} \le 0, \qquad i = 1, \ldots, l,$$

nachzuweisen, denn dann ist ja $g_i(x_k)/\|x_k\| \le \eta/(\sigma \|x_k\|)$, $i = 1, \ldots, l$. Daher nehmen wir jetzt an, dass $\eta = \infty$. Angenommen, für ein $i \in \{1, \ldots, l\}$ sei

$$\limsup_{k\to\infty} g_i(x_k)/\|x_k\| > 0.$$

Da man notfalls zu Teilfolgen übergehen kann, existiert ein $\epsilon > 0$ mit $g_i(x_k) \ge \epsilon \|x_k\|$ für alle k. Dann ist

$$\begin{aligned}
\frac{\theta(\epsilon\sigma \|x_k\|)}{\sigma \|x_k\|} &\le \frac{\theta(\sigma g_i(x_k))}{\sigma \|x_k\|} \\
&\le \frac{f_\sigma(\hat{x})}{\|x_k\|} - \frac{f(x_k)}{\|x_k\|} - \frac{1}{\sigma}\sum_{\substack{j=1\\ j\ne i}}^{l} \frac{\theta(\sigma g_i(x_k))}{\|x_k\|} \\
&\le \underbrace{\frac{f_\sigma(\hat{x})}{\|x_k\|}}_{\to 0} + c_0 - \underbrace{\frac{(l-1)\theta(-\sigma c_0 \|x_k\|)}{\sigma \|x_k\|}}_{\to 0}.
\end{aligned}$$

Die rechte Seite bleibt beschränkt, während die linke Seite wegen (A3) gegen $+\infty$ konvergiert. Das ist natürlich ein Widerspruch. Damit ist die Beschränktheit der Niveaumenge $\hat{L}_\sigma$ bewiesen. Zum Beweis der Abgeschlossenheit von $\hat{L}_\sigma$ nehmen wir an, $\{x_k\} \subset \hat{L}_\sigma$ sei eine Folge, die gegen ein $x \in \mathbb{R}^n$ konvergiert. Mit $h(x_k) = 0$ ist natürlich auch $h(x) = 0$, ferner ist $\sigma g(x) \le \eta e$ und wegen (A1) sogar $\sigma g(x) < \eta e$. Schließlich ist f_σ auf M_σ stetig und daher $x \in \hat{L}_\sigma$. Damit folgt die Kompaktheit von $\hat{L}_\sigma$ und dann auch die der Lösungsmenge $(M_\sigma)_{\text{opt}}$ von (P_σ). □ □

Wir geben nun Beispiele von monoton nicht fallenden, konvexen Funktionen $\theta \in C^2(-\infty, \eta)$ mit $0 \leq \eta \leq \infty$ an, für die die Voraussetzungen (A1)–(A3) in Satz 2.1 erfüllt sind.

Beispiele: Wir unterscheiden die Fälle, dass $\eta = \infty$, $\eta \in (0, \infty)$ und $\eta = 0$.

- Sei $\eta = \infty$.

Die Funktion $\theta(t) := \exp(t)$ (exponentielle Straffunktion) ist auf $(-\infty, \infty)$ monoton wachsend, strikt konvex und genügt offensichtlich den Bedingungen (A1)–(A3). Weiter definiere man

$$\theta(t) := \begin{cases} t + \frac{1}{2}t^2, & \text{falls} \quad t \geq -\frac{1}{2}, \\ -\frac{1}{4}\log(-2t) - \frac{3}{8}, & \text{falls} \quad t \leq -\frac{1}{2}. \end{cases}$$

Man rechnet leicht nach, dass $\theta \in C^2(-\infty, \infty)$. Ferner ist

$$\theta'(t) = \begin{cases} 1 + t, & \text{falls} \quad t \geq -\frac{1}{2}, \\ -1/(4t), & \text{falls} \quad t \leq -\frac{1}{2}, \end{cases}$$

ist $\theta'(t) > 0$ auf $(-\infty, \infty)$, also monoton wachsend. Schließlich ist

$$\theta''(t) = \begin{cases} 1, & \text{falls} \quad t \geq -\frac{1}{2}, \\ 1/(4t^2), & \text{falls} \quad t \leq -\frac{1}{2}, \end{cases}$$

woraus man abliest, dass θ auf $(-\infty, \infty)$ strikt konvex ist. Die Bedingungen (A1)–(A3) sind offensichtlich erfüllt.

- Es ist $\eta \in (0, \infty)$.

Sei $\eta = 1$ und $\theta(t) := -\log(1-t)$ (modifizierte logarithmische Barriere). Offensichtlich ist $\theta \in C^2(-\infty, 1)$ monoton wachsend und strikt konvex, auch die Eigenschaften (A1)–(A2) sind erfüllt. Das gleiche gilt mit $\eta = 1$ offenbar für $\theta(t) := t/(1-t)$ (modifizierte hyperbolische Barrierefunktion).

- Es ist $\eta = 0$.

Hier spricht man von Innere-Punkt-Verfahren. Die bekannteste Barrierefunktion ist natürlich die klassische logarithmische Barrierefunktion $\theta(t) := -\log(-t)$, die natürlich die geforderten Eigenschaften besitzt. Das gleiche gilt für $\theta(t) := -1/t$ (inverse Barriere). □

5.2.3 Lösungsfolgen und ihre Häufungspunkte

Das Ziel in diesem Unterabschnitt ist, einen Beweis für den folgenden Satz zu liefern.

Satz 2.2 *Die Funktion $\theta \in C^2(-\infty, \eta)$ mit $0 \le \eta \le \infty$ möge den Voraussetzungen von Satz 2.1 genügen. Ist dann $\{\sigma_k\} \subset \mathbb{R}_+$ eine Folge mit $\sigma_k \to \infty$ und $x_k \in M_k$ eine Lösung von*

$$(\mathrm{P}_k) \qquad \begin{cases} \text{Minimiere} \quad f_k(x) := f(x) + \dfrac{1}{\sigma_k} \displaystyle\sum_{i=1}^{l} \theta(\sigma_k g_i(x)) \quad \text{auf} \\ \qquad M_k := \{x \in \mathbb{R}^n : \sigma_k g(x) < \eta e,\ h(x) = 0\}, \end{cases}$$

so ist die Folge $\{x_k\}$ beschränkt, ferner ist jeder Häufungspunkt von $\{x_k\}$ eine Lösung[19] *von (P). Schließlich gilt $\lim_{k\to\infty} \min(\mathrm{P}_k) = \min(\mathrm{P})$.*

Beweis: Zunächst zeigen wir die Beschränktheit von $\{x_k\}$. Angenommen, dies sei nicht der Fall. Da wir notfalls zu Teilfolgen übergehen können, ist o. B. d. A. $\|x_k\| \to \infty$ und $x_k/\|x_k\| \to p$. Wie wir uns im Beweis von Satz 2.1 überlegten, erhalten wir einen Widerspruch zur vorausgesetzten Kompaktheit von M_{opt}, wenn wir

$$\limsup_{k\to\infty} \frac{f(x_k)}{\|x_k\|} \le 0, \qquad \limsup_{k\to\infty} \frac{g_i(x_k)}{\|x_k\|} \le 0 \quad (i = 1, \ldots, l)$$

nachweisen. Wie wir dort außerdem gesehen haben, existiert eine Konstante $c_0 > 0$ mit

$$\frac{f(x_k)}{\|x_k\|} \ge -c_0, \qquad \frac{g_i(x_k)}{\|x_k\|} \ge -c_0 \quad (i = 1, \ldots, l)$$

für alle $k \in \mathbb{N}$. Man wähle ein $x_0 \in M$, wenn $\eta > 0$, bzw. ein $x_0 \in M^0$, wenn $\eta = 0$. Dann ist

$$f_k(x_k) = f(x_k) + \frac{1}{\sigma_k} \sum_{i=1}^{l} \theta(\sigma_k g_i(x_k)) \le f_k(x_0) \le f(x_0) + \frac{l}{\sigma_k} \theta(\sigma_k a)$$

mit $a := \max_{i=1,\ldots,l} g_i(x_0)$. Dann ist

$$\frac{f(x_k)}{\|x_k\|} \le \frac{f(x_0)}{\|x_k\|} + \frac{l\theta(\sigma_k a)}{\sigma_k \|x_k\|} - \frac{l\theta(-c_0 \sigma_k \|x_k\|)}{\sigma_k \|x_k\|}.$$

Da die drei Summanden auf der rechten Seite wegen (A2) gegen 0 konvergieren, ist

$$\limsup_{k\to\infty} f(x_k)/\|x_k\| \le 0.$$

Für $\eta < \infty$ folgt aus $g(x_k) < (\eta/\sigma_k)e$, dass

$$\limsup_{k\to\infty} g_i(x_k)/\|x_k\| \le 0, \qquad i = 1, \ldots, l.$$

[19] Besitzt (P) insbesondere eine eindeutige Lösung x^*, so konvergiert die Folge $\{x_k\}$ gegen x^*.

Sei daher jetzt $\eta = \infty$. Wie im entsprechenden Teil des Beweises von Satz 2.1 nehmen wir an, es existiere ein $i \in \{1, \ldots, l\}$ und ein $\epsilon > 0$ mit $g_i(x_k) \geq \epsilon \|x_k\|$ für alle k. Dann ist

$$\begin{aligned}
\frac{\theta(\epsilon\sigma_k \|x_k\|)}{\sigma_k \|x_k\|} &\leq \frac{\theta(\sigma_k g_i(x_k))}{\sigma_k \|x_k\|} \\
&\leq \frac{f(x_0)}{\|x_k\|} + \frac{l\theta(\sigma_k a)}{\sigma_k \|x_k\|} - \frac{f(x_k)}{\|x_k\|} - \sum_{\substack{j=1 \\ j\neq i}}^{l} \frac{\theta(\sigma_k g_j(x_k))}{\sigma_k \|x_k\|} \\
&\leq \frac{f(x_0)}{\|x_k\|} + \frac{l\theta(\sigma_k a)}{\sigma_k \|x_k\|} + c_0 - \frac{(l-1)\theta(-c_0\sigma_k \|x_k\|)}{\sigma_k \|x_k\|}.
\end{aligned}$$

Wieder konvergiert die linke Seite wegen (A3) gegen $+\infty$, während die rechte Seite beschränkt bleibt, erneut ein Widerspruch. Damit ist die Beschränktheit der Folge $\{x_k\}$ bewiesen.

Nun sei $x^* \in \mathbb{R}^n$ ein Häufungspunkt von $\{x_k\}$. Wegen $h(x_k) = 0$ ist natürlich auch $h(x^*) = 0$. Wegen $g(x_k) < (\eta/\sigma_k)e$ und $\sigma_k \to \infty$, ist für $\eta < \infty$ auch $g(x^*) \leq 0$ und damit $x^* \in M$. Sei daher jetzt $\eta = \infty$. Angenommen, es existiert ein $i \in \{1, \ldots, l\}$ mit $g_i(x^*) > 0$. Dann existiert ein $\epsilon > 0$ und eine gegen x^* konvergente Teilfolge $\{x_k\}_{k\in K}$ derart, dass $g_i(x_k) \geq \epsilon$ für alle $k \in K$. Mit einem $x_0 \in M$ ist dann

$$\begin{aligned}
\frac{\theta(\epsilon\sigma_k)}{\sigma_k} &\leq \frac{\theta(\sigma_k g_i(x_k))}{\sigma_k} \\
&= f_k(x_k) - f(x_k) - \sum_{\substack{j=1 \\ j\neq i}}^{l} \frac{\theta(\sigma_k g_j(x_k))}{\sigma_k} \\
&\leq f_k(x_0) - f(x_k) - \sum_{\substack{j=1 \\ j\neq i}}^{l} \frac{\theta(\sigma_k g_j(x_k))}{\sigma_k} \\
&\leq f(x_0) + \sum_{i=1}^{l} \frac{\theta(\sigma_k a)}{\sigma_k} - f(x_k) - \sum_{\substack{j=1 \\ j\neq i}}^{l} \frac{\theta(\sigma_k g_j(x_k))}{\sigma_k} \\
&\leq f(x_0) + \sum_{i=1}^{l} \frac{\theta(\sigma_k a)}{\sigma_k} - f(x_k) - \frac{(l-1)\theta(-\sigma_k b)}{\sigma_k}
\end{aligned}$$

für alle $k \in K$, wobei wir wieder $a := \max_{i=1,\ldots,l} g_i(x_0)$ gesetzt und $b > 0$ so gewählt haben, dass $g_j(x_k) \geq -b$ für alle $j \in \{1, \ldots, l\} \setminus \{i\}$ und alle $k \in K$. Die linke Seite konvergiert wegen (A3) wieder gegen $+\infty$, während die rechte beschränkt bleibt, ein Widerspruch. Damit ist insgesamt nachgewiesen, dass jeder Häufungspunkt der Folge $\{x_k\}$ zulässig für (P) ist. Um nachzuweisen, dass jeder Häufungspunkt x^* von $\{x_k\}$ zu M_{opt} gehört, wählen wir zunächst $b > 0$ mit $g_i(x_k) \geq -b$ für alle $i \in \{1, \ldots, l\}$ und alle k, was natürlich wegen der Beschränktheit von $\{x_k\}$ möglich ist. Für $\eta > 0$ wählen wir $x_0 \in M_{\text{opt}}$ beliebig, wegen der Optimalität von x_k für (P_k) ist

$$f(x_k) + \frac{l\theta(-\sigma_k b)}{\sigma_k} \leq f_k(x_k) \leq f_k(x_0) \leq \min(\mathrm{P}) + \frac{l\theta(\sigma_k a)}{\sigma_k},$$

wobei wieder $a := \max_{i=1,\dots,l} g_i(x_0)$. Mit $k \to \infty$ folgt $f(x^*) \le \min(\mathrm{P})$ und folglich $x^* \in M_{\mathrm{opt}}$. Für $\eta = 0$ sei $x_0 \in M^0$, ferner wähle man $x_0' \in M_{\mathrm{opt}}$. Für alle $\lambda \in (0,1]$ ist $x_\lambda' := \lambda x_0 + (1-\lambda)x_0' \in M^0$. Mit festem $\lambda \in (0,1]$ ist

$$f(x_k) + \frac{l\theta(-\sigma_k b)}{\sigma_k} \le f_k(x_k) \le f_k(x_\lambda') \le f(x_\lambda') + \frac{l\theta(\sigma_k a_\lambda')}{\sigma_k},$$

wobei $a_\lambda' := \max_{i=1,\dots,l} g_i(x_\lambda')$. Mit $k \to \infty$ folgt

$$f(x^*) \le f(x_\lambda') \le \lambda f(x_0) + (1-\lambda)\min(\mathrm{P}),$$

mit $\lambda \to 0+$ ist $f(x^*) \le \min(\mathrm{P})$ und folglich $x^* \in M_{\mathrm{opt}}$.

Es bleibt, $\lim_{k\to\infty} \min(\mathrm{P}_k) = \min(\mathrm{P})$ zu zeigen. Wegen $\min(\mathrm{P}_k) = f_k(x_k)$ folgt dies aber leicht aus der obigen Argumentation, wobei man wieder die Fälle $\eta > 0$ und $\eta = 0$ getrennt behandelt. Ist $\eta > 0$, so wähle man $x_0 \in M_{\mathrm{opt}}$, setze $a := \min_{i=1,\dots,l} g_i(x_0)$, bestimme eine Konstante $b > 0$ mit $g_i(x_k) \ge -b$, $i = 1,\dots,l$, für alle k und erhalte wieder

$$f(x_k) + \underbrace{\frac{l\theta(-\sigma_k b)}{\sigma_k}}_{\to 0} \le f_k(x_k) \le \min(\mathrm{P}) + \underbrace{\frac{l\theta(\sigma_k a)}{\sigma_k}}_{\to 0}.$$

Unter Benutzung der beiden schon bewiesenen Teile, dass nämlich die Folge $\{x_k\}$ beschränkt und jeder Häufungspunkt von $\{x_k\}$ eine Lösung von (P) ist, folgt

$$\min(\mathrm{P}) \le \liminf_{k\to\infty} f_k(x_k) \le \limsup_{k\to\infty} f_k(x_k) \le \min(\mathrm{P})$$

und damit die Behauptung. Ist dagegen $\eta = 0$, so seien $x_0 \in M^0$ und $x_0' \in M_{\mathrm{opt}}$ gewählt. Mit $\lambda \in (0,1]$ sei wieder $x_\lambda' := \lambda x_0 + (1-\lambda)x_0'$. Mit derselben Argumentation wie gerade eben erhält man für jedes $\lambda \in (0,1]$, dass

$$\min(\mathrm{P}) \le \liminf_{k\to\infty} f_k(x_k) \le \limsup_{k\to\infty} f_k(x_k) \le f(x_\lambda') \le \lambda f(x_0) + (1-\lambda)\min(\mathrm{P}).$$

Wieder erhält man mit $\lambda \to 0+$ die Behauptung.

Damit ist der Satz bewiesen. □ □

5.2.4 Eindeutigkeit einer Lösung des Hilfsproblems

Unter unseren Standardvoraussetzungen untersuchen wir, wann bei festem $\sigma > 0$ die Aufgabe

$$(\mathrm{P}_\sigma) \qquad \begin{cases} \text{Minimiere} \quad f_\sigma(x) := f(x) + \dfrac{1}{\sigma}\displaystyle\sum_{i=1}^{l} \theta(\sigma g_i(x)) \quad \text{auf} \\ M_\sigma := \{x \in \mathbb{R}^n : \sigma g(x) < \eta e,\ h(x) = 0\} \end{cases}$$

eindeutig lösbar ist. Wir setzen jetzt voraus, dass f und g zweimal stetig differenzierbar sind. Dann existiert $\nabla^2 f_\sigma(x)$ für jedes $x \in M_\sigma$ und es ist einfach einzusehen, dass eine Lösung von (P_σ) eindeutig ist, wenn $\nabla^2 f_\sigma(x)$ auf Kern (h') positiv definit ist. Es ist

$$\nabla f_\sigma(x) = \nabla f(x) + \sum_{i=1}^{l} \theta'(\sigma g_i(x)) \nabla g_i(x)$$

und

$$\nabla^2 f_\sigma(x) = \nabla^2 f(x) + \sum_{i=1}^{l} [\sigma\theta''(\sigma g_i(x))\nabla g_i(x)\nabla g_i(x)^T + \theta'(\sigma g_i(x))\nabla^2 g_i(x)].$$

Wir setzen jetzt noch voraus, dass θ auf $(-\infty, \eta)$ monoton wachsend und strikt konvex ist, also $\theta'(t) > 0$ und $\theta''(t) > 0$ für alle $t \in (-\infty, \eta)$ gilt. Als positive Linearkombination positiv semidefiniter Matrizen ist $\nabla^2 f_\sigma(x)$ positiv semidefinit. Sei daher $p \in \text{Kern}\,(h') \setminus \{0\}$ und $p^T \nabla^2 f_\sigma(x) p = 0$. Dann folgt

$$(*) \qquad \nabla^2 f(x)p = 0, \qquad \nabla^2 g_i(x)p = 0, \qquad \nabla g_i(x)^T p = 0 \quad (i = 1, \ldots, l).$$

Als Hauptergebnis erhalten wir:

Satz 2.3 *Sei (P) ein konvexes, quadratisch restringiertes quadratisches Programm, also*

$$f(x) := c_0^T x + \tfrac{1}{2} x^T Q_0 x, \qquad g_i(x) := \beta_i + c_i^T x + \tfrac{1}{2} x^T Q_i x \quad (i = 1, \ldots, l)$$

mit symmetrischen, positiv semidefiniten Matrizen $Q_0, Q_1, \ldots, Q_l \in \mathbb{R}^{n\times n}$. Sei $\theta \in C^2(-\infty, \eta)$ mit $0 \le \eta \le \infty$ monoton wachsend und strikt konvex auf $(-\infty, \eta)$, ferner seien die Standardvoraussetzungen an das Problem (P) erfüllt, insbesondere also die Menge M_{opt} der Lösungen von (P) nichtleer und kompakt. Dann besitzt die Aufgabe (P_σ) für jedes $\sigma > 0$ höchstens eine Lösung, wenn die Bedingungen (A1)–(A3) aus Satz 2.1 erfüllt sind also genau eine Lösung.

Beweis: Wir geben uns ein $p \in \text{Kern}\,(h')$ vor und schließen aus $(*)$, dass $p = 0$ bzw. $\nabla^2 f_\sigma(x)$ auf $\text{Kern}\,(h')$ positiv definit ist. Spezialisiert auf die vorliegende Situation bedeutet $(*)$, dass $Q_i p = 0$, $i = 0, \ldots, l$, und $c_i^T p = 0$, $i = 1, \ldots, l$. O. B. d. A. ist ferner $c_0^T p \le 0$ (ersetze notfalls p durch $-p$). Mit einem beliebigen $x^* \in M_{\text{opt}}$ ist dann $f(x^* + tp) \le f(x^*)$, $g(x^* + tp) = g(x^*)$ und $h(x^* + tp) = h(x^*)$ und damit $x^* + tp \in M_{\text{opt}}$ für alle $t \ge 0$. Da M_{opt} nach Voraussetzung beschränkt ist, ist $p = 0$ und daher $\nabla^2 f_\sigma(x)$ für jedes $x \in M_\sigma$ auf $\text{Kern}\,(h')$ positiv definit, woraus die Eindeutigkeit einer Lösung von (P_σ) folgt. □ □

5.2.5 Konvergenz der primalen Trajektorie

Wie in Satz 2.3 betrachten wir ein (konvexes) quadratisch restringiertes quadratisches Programm. Ist $\theta \in C^2(-\infty, \eta)$ mit $0 \le \eta \le \infty$ monoton wachsend und strikt konvex, sind ferner die Bedingungen (A1)–(A3) in Satz 2.1 erfüllt, so besitzt die Aufgabe (P_σ) für jedes $\sigma > 0$ genau eine Lösung $x_\sigma \in M_\sigma$. In diesem Unterabschnitt wollen wir untersuchen, unter welchen Bedingungen $x_\infty = \lim_{\sigma\to\infty} x_\sigma$ existiert und eine (gewisse) Lösung von (P) ist. Wegen Satz 2.2 ist dies jedenfalls dann der Fall, wenn (P) eindeutig lösbar ist. Ist dies nicht der Fall, so muss man aus der Lösungsmenge M_{opt} eine bestimmte aussondern, gegen die die Folge $\{x_k\}$ mit $x_k := x_{\sigma_k}$ und $\sigma_k \to \infty$ konvergiert. Hierzu definieren wir die Indexmengen

$$I := \{i \in \{1, \ldots, l\} : g_i(x) = 0 \text{ für alle } x \in M_{\text{opt}}\}, \qquad J := \{1, \ldots, l\} \setminus I.$$

Also ist I die Indexmenge derjenigen Ungleichungsrestriktionen, die für jede Lösung aktiv sind. Ein erstes einfaches Ergebnis ist

Lemma 3.1 *Gegeben sei das konvexe, quadratisch restringierte quadratische Programm (P), dessen Lösungsmenge M_{opt} nichtleer und kompakt sei. Ist dann $g(x) = 0$ für alle $x \in M_{\text{opt}}$ (bzw. $J = \emptyset$), so ist (P) eindeutig lösbar.*

Beweis: Wir nehmen an, $x^*, x^{**} \in M_{\text{opt}}$ seien zwei Lösungen von (P). Dann ist auch $\frac{1}{2}(x^* + x^{**}) \in M_{\text{opt}}$ eine Lösung von (P). Folglich ist

$$0 = \tfrac{1}{2}[\underbrace{g_i(x^*)}_{=0} + \underbrace{g_i(x^{**})}_{=0}] - \underbrace{g_i(\tfrac{1}{2}(x^* + x^{**}))}_{=0} = \tfrac{1}{8}(x^* - x^{**})^T Q_i (x^* - x^{**})$$

und daher $Q_i(x^* - x^{**}) = 0$, $i = 1, \ldots, l$. Wegen $g_i(x^*) = g_i(x^{**})$ folgt dann auch $c_i^T(x^* - x^{**}) = 0$, $i = 1, \ldots, l$. Entsprechend folgt aus

$$f(x^*) = f(x^{**}) = f(\tfrac{1}{2}(x^* + x^{**})) = \min\,(\mathrm{P}),$$

dass auch $Q_0(x^* - x^{**}) = 0$ und $c_0^T(x^* - x^{**}) = 0$. Daher ist $x^{**} + t(x^* - x^{**}) \in M_{\text{opt}}$ für alle $t \in \mathbb{R}$, woraus $x^* = x^{**}$ wegen der Beschränktheit von M_{opt} folgt. □ □

Wegen des letzten Lemmas kann im folgenden $J \neq \emptyset$ angenommen werden. Die "auszusondernde" Lösung von (P) ist eindeutige Lösung einer gewissen Optimierungsaufgabe, deren "Prototyp" wir jetzt im ersten Teil des folgenden Satzes angeben. Im zweiten Teil des Satzes wird gezeigt, dass $x_\infty = \lim_{\sigma\to\infty} x_\sigma$ existiert und eine gewisse Lösung von (P) ist.

Satz 3.2 *Gegeben sei das konvexe, quadratisch restringierte quadratische Programm (P), dessen Lösungsmenge M_{opt} nichtleer und kompakt sei. Sei $\theta \in C^2(-\infty, \eta)$ mit $0 \leq \eta \leq \infty$ monoton wachsend und strikt konvex auf $(-\infty, \eta)$, ferner seien die Bedingungen (A1)–(A3) aus Satz 2.1 erfüllt. Weiter gelte die Bedingung*

(A4) *Es existieren Abbildungen $\alpha\colon \mathbb{R}_+ \longrightarrow \mathbb{R}$ und $\beta\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ derart, dass*

$$\theta_\infty(\delta) := \lim_{\sigma\to\infty} \beta(\sigma)[\theta(\sigma\delta) - \alpha(\sigma)]$$

für jedes $\delta < 0$ existiert und θ_∞ auf $(-\infty, 0)$ eine stetige, monoton wachsende, strikt konvexe Funktion mit $\lim_{\delta\to 0-} \theta_\infty(\delta) = +\infty$ ist.

Dann gilt:

(a) Die Aufgabe

$$(\mathrm{P}_\infty) \qquad \begin{cases} \text{Minimiere} \quad f_\infty(x) := \displaystyle\sum_{j\in J} \theta_\infty(g_i(x)) \quad \text{auf} \\ \\ M^*_{\text{opt}} := \{x \in M_{\text{opt}} : g_j(x) < 0 \ \ (j \in J)\} \end{cases}$$

besitzt genau eine Lösung $x_\infty \in M^*_{\text{opt}}$. Hierbei sei

$$I := \{i \in \{1, \ldots, l\} : g_i(x) = 0 \ \text{ für alle } x \in M_{\text{opt}}\}, \qquad J := \{1, \ldots, l\} \setminus I.$$

(b) Ist $x_\sigma \in M_\sigma$ mit $\sigma > 0$ die Lösung von

$$(\mathrm{P}_\sigma) \qquad \left\{ \begin{array}{l} \textit{Minimiere} \quad f_\sigma(x) := f(x) + \dfrac{1}{\sigma}\displaystyle\sum_{i=1}^{l} \theta(\sigma g_i(x)) \quad \textit{auf} \\ \qquad M_\sigma := \{x \in \mathbb{R}^n : \sigma g(x) < \eta e,\ h(x) = 0\}, \end{array} \right.$$

so gilt $\lim_{\sigma\to\infty} x_\sigma = x_\infty$.

Beweis: Zunächst müssen wir uns überlegen, dass (P_∞) zulässig bzw. $M^*_{\mathrm{opt}} \neq \emptyset$ ist. Nach Definition der Indexmenge J existiert zu jedem $j \in J$ ein $x^{(j)} \in M_{\mathrm{opt}}$ mit $g_j(x^{(j)}) < 0$. Offenbar ist dann

$$x := \frac{1}{\#(J)} \sum_{j\in J} x^{(j)} \in M^*_{\mathrm{opt}}.$$

Zum Nachweis der Existenz einer Lösung von (P_∞) bilde man mit einem $\hat{x} \in M^*_{\mathrm{opt}}$ die Niveaumenge

$$\hat{L} := \{x \in M^*_{\mathrm{opt}} : f_\infty(x) \le f_\infty(\hat{x})\}.$$

Wegen $\hat{L} \subset M^*_{\mathrm{opt}} \subset M_{\mathrm{opt}}$ ist $\hat{L}$ beschränkt, da $\lim_{t\to 0-} \theta_\infty(t) = +\infty$, ist $\hat{L}$ offenbar auch kompakt. Da f_∞ auf $\hat{L}$ stetig ist, folgt die Existenz einer Lösung von (P_∞). Sind $x^*, x^{**} \in M^*_{\mathrm{opt}}$ zwei Lösungen von (P_∞), so ist es auch $\frac{1}{2}(x^* + x^{**})$. Daher ist

$$\begin{aligned} 0 &= \tfrac{1}{2}[f_\infty(x^*) + f_\infty(x^{**})] - f_\infty(\tfrac{1}{2}(x^* + x^{**})) \\ &= \sum_{j\in J} \{\tfrac{1}{2}[\theta_\infty(g_j(x^*)) + \theta_\infty(g_j(x^{**}))] - \theta_\infty(g_j(\tfrac{1}{2}(x^* + x^{**})))\} \\ &\ge \sum_{j\in J} \{\underbrace{\tfrac{1}{2}[\theta_\infty(g_j(x^*)) + \theta_\infty(g_j(x^{**}))] - \theta_\infty(\tfrac{1}{2}(g_j(x^*) + g_j(x^{**})))}_{\ge 0}\} \\ &\qquad \text{(da } g_j \text{ konvex und } \theta_\infty \text{ monoton nicht fallend und konvex)} \\ &\ge 0. \end{aligned}$$

Also ist

$$\theta_\infty(\tfrac{1}{2}(g_j(x^*) + g_j(x^{**}))) = \tfrac{1}{2}[\theta_\infty(g_j(x^*)) + \theta_\infty(g_j(x^{**}))], \qquad j \in J.$$

Da θ_∞ nach Voraussetzung monoton wachsend und strikt konvex auf $(-\infty, 0)$ ist, ist

$$g_j(x^*) = g_j(x^{**}) = g_j(\tfrac{1}{2}(x^* + x^{**})), \qquad j \in J.$$

Genau wie in Lemma 3.1 folgt hieraus, dass $x^{**} + t(x^* - x^{**}) \in M_{\mathrm{opt}}$ für alle $t \in \mathbb{R}$ und hieraus $x^* = x^{**}$. Damit ist der erste Teil des Satzes bewiesen.

Sei nun $\{\sigma_k\} \subset \mathbb{R}_+$ mit $\sigma_k \to \infty$, ferner $x_k = x_{\sigma_k}$ die Lösung von (P_{σ_k}). Sei x^* ein Häufungspunkt der nach Satz 2.2 beschränkten Folge. O. B. d. A. (notfalls gehe man zu Teilfolgen über) konvergiere $\{x_k\}$ gegen x^*. Ebenfalls nach Satz 2.2 ist $x^* \in M_{\mathrm{opt}}$. Wir wollen zeigen, dass x^* eine Lösung von (P_∞) ist. Ist dies gelungen, so konvergiert

offenbar die gesamte Folge $\{x_k\}$ gegen die eindeutige Lösung x_∞ von (P_∞). Um dies einzusehen, beachten wir zunächst, dass

$$c_i^T(x_\infty - x^*) = 0, \qquad Q_i(x_\infty - x^*) = 0 \quad (i \in \{0\} \cup I).$$

Definiert man daher

$$x_k^* := x_k + x_\infty - x^*,$$

so gilt

$$f(x_k^*) = f(x_k), \qquad g_i(x_k^*) = g_i(x_k) \quad (i \in I)$$

für alle k. Wegen $h(x_k^*) = h(x_k)$ sowie $x_k^* \to x_\infty$ und $\sigma_k g_j(x_k^*) \to -\infty$, $j \in J$, ist ferner x_k^* für alle hinreichend großen k zulässig für (P_{σ_k}). Für diese k ist daher

$$\sum_{j\in J} \theta(\sigma_k g_j(x_k)) \le \sum_{j\in J} \theta(\sigma_k g_j(x_k^*)).$$

Mit

$$\delta_j < g_j(x^*) \le 0, \qquad g_j(x_\infty) < \delta_j^* < 0$$

folgt aus der Monotonie von θ, dass

$$\sum_{j\in J} \theta(\sigma_k \delta_j) \le \sum_{j\in J} \theta(\sigma_k \delta_j^*)$$

für alle hinreichend großen k. Diese Aussage bleibt richtig, wenn wir in jedem Summanden auf beiden Seiten $\alpha(\sigma_k) \in \mathbb{R}$ abziehen und mit $\beta(\sigma_k) \in \mathbb{R}_+$ multiplizieren:

$$\sum_{j\in J} \beta(\sigma_k)[\theta(\sigma_k \delta_j) - \alpha(\sigma_k)] \le \sum_{j\in J} \beta(\sigma_k)[\theta(\sigma_k \delta_j^*) - \alpha(\sigma_k)].$$

Dann erhält man nach dem Grenzübergang $k \to \infty$, dass

$$\sum_{j\in J} \theta_\infty(\delta_j) \le \sum_{j\in J} \theta_\infty(\delta_j^*).$$

Da hier $\delta_j < g_j(x^*)$ beliebig ist und nach Voraussetzung $\lim_{t\to 0-} \theta_\infty(t) = +\infty$, ist $g_j(x^*) < 0$, $j \in J$, bzw. $x^* \in M_{\mathrm{opt}}^*$. Mit $\delta_j \nearrow g_j(x^*)$ und $\delta_j^* \searrow g_j(x_\infty)$, $j \in J$, folgt aus Stetigkeitsgründen

$$\sum_{j\in J} \theta_\infty(g_j(x^*)) \le \sum_{j\in J} \theta_\infty(g_j(x_\infty)),$$

d. h. x^* löst (P_∞). Damit ist auch der zweite Teil des Satzes bewiesen. □ □

Nun geben wir Beispiele an, bei denen die Bedingung (A4) in Satz 3.2 erfüllt ist.

Beispiele: Wir unterscheiden wieder die Fälle, dass $\eta = \infty$, $\eta \in (0, \infty)$ und $\eta = 0$.

- Sei $\eta = \infty$.

Für $\theta(t) := \exp(t)$ können wir die Existenz von Funktionen α und β mit den angegebenen Eigenschaften nicht zeigen. Es ist zu vermuten, dass die exponentielle Straffunktion der Voraussetzung (A4) nicht genügt (Beweis?).

Sei

$$\theta(t) := \begin{cases} t + \frac{1}{2}t^2, & \text{falls} \quad t \geq -\frac{1}{2}, \\ -\frac{1}{4}\log(-2t) - \frac{3}{8}, & \text{falls} \quad t \leq -\frac{1}{2}. \end{cases}$$

Man definiere $\alpha\colon \mathbb{R}_+ \longrightarrow \mathbb{R}$ und $\beta\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ durch

$$\alpha(\sigma) := -(\tfrac{3}{8} + \tfrac{1}{4}\log(2\sigma)), \qquad \beta(\sigma) := 4.$$

Für jedes $\delta < 0$ ist $\sigma\delta \leq -\frac{1}{2}$ für alle hinreichend großen σ (genauer für alle $\sigma \geq -1/(2\delta)$), für diese σ ist

$$\beta(\sigma)[\theta(\sigma\delta) - \alpha(\sigma)] = -\log(-\delta).$$

Da weiter $\theta_\infty\colon (-\infty, 0) \longrightarrow \mathbb{R}$, definiert durch $\theta_\infty(\delta) := -\log(-\delta)$ monoton wachsend und strikt konvex ist, ist die Bedingung (A4) erfüllt.

- Es ist $\eta \in (0, \infty)$.

Sei $\eta := 1$ und $\theta(t) := -\log(1 - t)$. Man definiere $\alpha\colon \mathbb{R}_+ \longrightarrow \mathbb{R}$ und $\beta\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ durch

$$\alpha(\sigma) := -\log\sigma, \qquad \beta(\sigma) := 1.$$

Dann ist

$$\beta(\sigma)[\theta(\sigma\delta) - \alpha(\sigma)] = -\log)1/\sigma - \delta),$$

die Bedingung (A4) ist also mit $\theta_\infty(\delta) := -\log(-\delta)$ erfüllt.

Ist dagegen $\eta := 1$ und $\theta(t) := t/(1 - t)$, so definiere man $\alpha\colon \mathbb{R}_+ \longrightarrow \mathbb{R}$ und $\beta\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ durch

$$\alpha(\sigma) := -1, \qquad \beta(\sigma) := \sigma.$$

Dann ist

$$\beta(\sigma)[\theta(\sigma\delta) - \alpha(\sigma)] = \frac{1}{1/\sigma - \delta},$$

die Bedingung (A4) ist also mit $\theta_\infty(\delta) := -1/\delta$ erfüllt.

- Es ist $\eta = 0$.

Ist $\theta(t) := -\log(-t)$, so definiere man $\alpha\colon \mathbb{R}_+ \longrightarrow \mathbb{R}$ und $\beta\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ durch

$$\alpha(\sigma) := -\log\sigma, \qquad \beta(\sigma) := 1.$$

Dann ist

$$\beta(\sigma)[\theta(\sigma\delta) - \alpha(\sigma)] = -\log(-\delta),$$

die Bedingung (A4) ist mit $\theta_\infty(\delta) := -\log(-\delta)$ erfüllt.

Ist dagegen $\eta := 0$ und $\theta(t) := -1/t$, so definiere man $\alpha\colon \mathbb{R}_+ \longrightarrow \mathbb{R}$ und $\beta\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ durch

$$\alpha(\sigma) := 0, \qquad \beta(\sigma) := \sigma.$$

Dann ist

$$\beta(\sigma)[\theta(\sigma\delta) - \alpha(\sigma)] = -1/\delta,$$

die Bedingung (A4) ist mit $\theta_\infty(\delta) := -1/\delta$ erfüllt. □

Wir haben bisher nicht die Stetigkeit der primalen Trajektorie im Falle exponentieller Strafen ($\eta := \infty$ und $\theta(t) := \exp(t)$) zeigen können. Bei A. Auslender, R. Cominetti, M. Haddou (1997, S. 56) wird hierauf in einem allgemeineren Zusammenhang eingegangen, während wir uns hier auf exponentielle Strafen konzentrieren wollen.

Im folgenden Satz wird das sogenannte *Zentrum* der Menge M_{opt} der Lösungen einer konvexen quadratisch restringierten quadratischen Optimierungsaufgabe definiert und gezeigt, dass es aus genau einem Punkt besteht.

Satz 3.3 *Gegeben sei die konvexe, quadratisch restringierte quadratische Optimierungsaufgabe (P), deren Lösungsmenge M_{opt} nichtleer und kompakt sei. Man betrachte den folgenden Algorithmus:*

- *Sei $S_0 := M_{\text{opt}}$ und $I_0 := \{i \in \{1, \ldots, l\} : g_i(x) = 0 \;\; \text{für alle } x \in S_0\}$.*
- *Für $k = 0, 1, \ldots$:*
  - *Falls $I_k = \{1, \ldots, l\}$, dann: $p := k$, STOP, da S_p aus genau einem Punkt besteht.*
  - *Andernfalls berechne*

$$\gamma_{k+1} := \min_{x \in S_k} \max_{i \notin I_k} g_i(x), \qquad S_{k+1} := \{x \in S_k : \max_{i \notin I_k} g_i(x) = \gamma_{k+1}\}$$

und anschließend

$$J_{k+1} := \{i \notin I_k : g_i(x) = \gamma_{k+1} \;\; \text{für alle } x \in S_{k+1}\}, \qquad I_{k+1} := I_k \cup J_{k+1}.$$

Dieser Algorithmus ist durchführbar und bricht nach endlich vielen Schritten mit der einpunktigen Menge S_p, dem sogenannten Zentrum von M_{opt}, ab.

Beweis: Zunächst überlegen wir uns, dass der Algorithmus durchführbar ist, dann, dass er nach endlich vielen Schritten abbricht, da (solange kein Abbruch) die Folge der in $\{1, \ldots, l\}$ enthaltenen Indexmengen $I_0, I_1, \ldots$ streng aufsteigend ist und schließlich, dass das Zentrum von M_{opt} aus genau einem Element besteht. Für den Beweis der Durchführbarkeit überlegen wir uns:

- Die Menge $S_k \subset M_{\text{opt}}$ ist nichtleer, konvex und kompakt.

Die Aussage zeigen wir durch vollständige Induktion nach k, wobei der Induktionsanfang bei $k = 0$ liegt. Die Menge $S_0 = M_{\text{opt}}$ ist nach Voraussetzung nichtleer und kompakt, als Lösungsmenge eines konvexen Programms ferner konvex. Nun nehmen wir an, die Aussage sei für k richtig. Die Funktion $\phi_k \colon S_k \longrightarrow \mathbb{R}$, definiert durch $\phi_k(x) := \max_{i \notin I_k} g_i(x)$, ist stetig. Insbesondere besitzt die Aufgabe, ϕ_k auf S_k zu minimieren, mindestens eine Lösung und die Lösungsmenge S_{k+1} ist nichtleer, kompakt und konvex. Damit ist die Durchführbarkeit des Algorithmus gezeigt.

Nun zeigen wir, dass der Algorithmus nach endlich vielen Schritten abbricht. Hierzu überlegen wir uns:

- Ist I_k eine echte Teilmenge von $\{1,\dots,l\}$, so ist

$$J_{k+1} := \{i \notin I_k : g_i(x) = \gamma_{k+1} \text{ für alle } x \in S_{k+1}\} \neq \emptyset$$

und daher $I_{k+1} := I_k \cup J_{k+1}$ eine echte Obermenge von I_k.

Wir machen einen Widerspruchsbeweis und nehmen an, es sei $J_{k+1} = \emptyset$. Dann existiert zu jedem $i \notin I_k$ ein $x^{(i)} \in S_{k+1}$ mit $g_i(x^{(i)}) < \gamma_{k+1}$. Wegen der Konvexität von S_{k+1} ist auch

$$x := \frac{1}{l - \#(I_k)} \sum_{j \notin I_k} x^{(j)} \in S_{k+1}.$$

Für ein beliebiges $i \notin I_k$ folgt aus der Konvexität von g_i, dass

$$\begin{aligned} g_i(x) &= g_i\Big(\frac{1}{l-\#(I_k)} \sum_{j\notin I_k} x^{(j)}\Big) \\ &\leq \frac{1}{l-\#(I_k)} \sum_{j\notin I_k} g_i(x^{(j)}) \\ &= \frac{1}{l-\#(I_k)} \Big(\sum_{\substack{j\notin I_k \\ j\neq i}} \underbrace{g_i(x^{(j)})}_{\leq \gamma_{k+1}} + \underbrace{g_i(x^{(i)})}_{<\gamma_{k+1}}\Big) \\ &< \gamma_{k+1}, \end{aligned}$$

folglich ist $\gamma_{k+1} = \max_{i\notin I_k} g_i(x) < \gamma_{k+1}$, ein Widerspruch. Damit ist auch gezeigt, dass der Algorithmus nach endlich vielen Schritten abbricht.

Angenommen, das Verfahren breche im p-ten Schritt ab, es sei also $I_p = \{1,\dots,l\}$. Die Menge $\{1,\dots,l\}$ kann dargestellt werden als disjunkte Vereinigung von $p+1$ Indexmengen:

$$\{1,\dots,l\} = I_0 + \bigcup_{k=1}^{p} J_k.$$

Wir definieren

$$G_i := \begin{cases} 0, & i \in I_0, \\ \gamma_1, & i \in J_1, \\ \vdots & \vdots \\ \gamma_p, & i \in J_p, \end{cases} \qquad i = 1,\dots,l.$$

Für alle $x \in S_p$ ist dann $g_i(x) = G_i$, $i = 1,\dots,l$. Hieraus wollen wir schließen, dass S_p einpunktig ist. Hierzu nehmen wir an, es seien $x^*, x^{**} \in S_p$. Wegen der Konvexität von S_p ist $\frac{1}{2}(x^* + x^{**}) \in S_p$, folglich ist

$$0 = \tfrac{1}{2}[g_i(x^*) + g_i(x^{**})] - g_i(\tfrac{1}{2}(x^* + x^{**})) = \tfrac{1}{8}(x^* - x^{**})^T Q_i (x^* - x^{**}),$$

also $Q_i(x^* - x^{**}) = 0$, $i = 1,\dots,l$. Weiter ist

$$0 = g_i(x^*) - g_i(x^{**}) = (c_i + Q_i x^{**})^T (x^* - x^{**}) + \tfrac{1}{2} \underbrace{(x^* - x^{**})^T Q_i (x^* - x^{**})}_{=0} = c_i^T (x^* - x^{**}),$$

also $c_i^T(x^* - x^{**})$, $i = 1, \ldots, l$. Ebenso erhält man aus der Konvexität von M_{opt} und $f(x^*) = f(x^{**})$, dass $Q_0(x^* - x^{**}) = 0$ und $c_0^T(x^* - x^{**}) = 0$. Dann ist aber $f(x^* + t(x^{**} - x^*)) = f(x^*)$ und $g(x^* + t(x^{**} - x^*)) = g(x^*)$ für alle $t \in \mathbb{R}$, insbesondere $x^* + t(x^{**} - x^*) \in M_{\text{opt}}$ für alle $t \in \mathbb{R}$. Aus der vorausgesetzten Kompaktheit von M_{opt} folgt $x^* = x^{**}$. Damit ist der Satz bewiesen. □ □

Nun kommen wir zu einem Satz 3.2 entsprechenden Satz.

Satz 3.4 *Gegeben sei das konvexe, quadratisch restringierte quadratische Programm (P), dessen Lösungsmenge M_{opt} nichtleer und kompakt sei. Ist $x_\sigma \in M_\sigma$ mit $\sigma > 0$ die Lösung von*

$$(\mathrm{P}_\sigma) \qquad \begin{cases} \textit{Minimiere} \quad f_\sigma(x) := f(x) + \dfrac{1}{\sigma} \displaystyle\sum_{i=1}^{l} \exp(\sigma g_i(x)) \quad \textit{auf} \\ \qquad\qquad M_\sigma := \{x \in \mathbb{R}^n : h(x) = 0\}, \end{cases}$$

so existiert $\lim_{\sigma \to \infty} x_\sigma$ und stimmt mit dem Zentrum x^∞ von M_{opt} überein.

Beweis: Sei $\{\sigma_k\} \subset \mathbb{R}_+$ eine Folge mit $\sigma_k \to \infty$, ferner $x_k = x_{\sigma_k}$ die Lösung von (P_{σ_k}). Sei x^* ein Häufungspunkt der nach Satz 2.2 beschränkten Folge. O. B. d. A. (notfalls gehe man zu Teilfolgen über) konvergiere $\{x_k\}$ gegen x^*. Ebenfalls nach Satz 2.2 ist $x^* \in M_{\text{opt}}$. Wir wollen zeigen, dass x^* das Zentrum x^∞ von M_{opt} ist, wobei wir natürlich annehmen können, dass I_0 eine echte Teilmenge von $\{1, \ldots, l\}$, ist. Mit den Bezeichnungen von Satz 3.3 zeigen wir zunächst:

- Es existiert ein $x \in S_k$ mit $g_i(x) < 0$, $i \notin I_k$, $k = 0, \ldots, p$.
- Es ist $\gamma_{k+1} < 0$, $k = 0, \ldots, p-1$.

Diese Aussage wird durch vollständige Induktion nach k bewiesen. Für alle $i \notin I_0$ existiert ein $x^{(i)} \in S_0$ mit $g_i(x^{(i)}) < 0$. Dann ist

$$x := \frac{1}{l - \#(I_0)} \sum_{i \notin I_0} x^{(i)}$$

ein Punkt mit $g_i(x) < 0$, $i \notin I_0$. Insbesondere existiert ein $x \in S_0$ mit $\max_{i \notin I_0} g_i(x) < 0$. Nach Definition ist S_1 die Lösungsmenge der Aufgabe, ϕ_0, definiert durch $\phi_0(x) := \max_{i \notin I_0} g_i(x)$ auf $S_0 = M_{\text{opt}}$ zu minimieren, folglich ist $\gamma_1 = \max_{i \notin I_0} g_i(x) < 0$ für alle $x \in S_1$. Der Induktionsschritt kann ganz entsprechend bewiesen werden.

Nun zeigen wir:

- Für alle k ist $x_k^* := x_k + (x^\infty - x^*)$ zulässig für (P_{σ_k}).

Dies ist trivial, denn mit $h(x) = Ax - b$ ist

$$h(x_k^*) = A(x_k + (x^\infty - x^*)) - b = Ax_k - b = h(x_k) = 0.$$

Damit haben wir nachgewiesen, dass x_k^* für alle k zulässig für (P_{σ_k}) ist.

Da x_k Lösung von (P_{σ_k}) und x_k^* für alle k für (P_{σ_k}) zulässig ist, ist

$$(*) \qquad f(x_k) + \frac{1}{\sigma_k} \sum_{i=1}^{l} \exp(\sigma_k g_i(x_k)) \leq f(x_k^*) + \frac{1}{\sigma_k} \sum_{i=1}^{l} \exp(\sigma_k g_i(x_k^*))$$

für alle k. Wir zeigen durch vollständige Induktion nach j, dass $x^* \in S_j$, $j = 0, \ldots, p$. Da $S_p = \{x^\infty\}$ einpunktig ist, ist dann $x^* = x^\infty$ und der Satz ist bewiesen. Wegen $x^* \in M_{\mathrm{opt}} = S_0$ ist der Induktionsanfang gelegt. Angenommen, es ist $x^* \in S_j$ mit einem $1 \leq j < p$. Dann ist

$$g_i(x^*) = g_i(x^\infty) = g_i(\tfrac{1}{2}(x^* + x^\infty)), \qquad i \in I_j.$$

Hieraus folgt $g_i(x_k^*) = g_i(x_k)$, $i \in I_j$, und alle k, ferner ist natürlich auch $f(x_k^*) = f(x_k)$ für alle k. Aus $(*)$ erhält man damit

$$\sum_{i \notin I_j} \exp(\sigma_k g_i(x_k)) \leq \sum_{i \notin I_j} \exp(\sigma_k g_i(x_k^*))$$

für alle k. Nun wähle man $\delta_i < g_i(x^*) \leq 0$, $i \notin I_j$, und $g_i(x^\infty) < \delta_i^* < 0$, $i \notin I_j$. Für alle hinreichend großen k ist dann auch $\delta_i < g_i(x_k)$ und $g_i(x_k^*) < \delta_i^*$ für alle $i \notin I_j$. Wegen der strengen Monotonie von exp und log ist

$$\frac{1}{\sigma_k} \log\Big(\sum_{i \notin I_j} \exp(\sigma_k \delta_i)\Big) < \frac{1}{\sigma_k} \log\Big(\sum_{i \notin I_j} \exp(\sigma_k \delta_i^*)\Big)$$

für alle hinreichend großen k. Hieraus folgt

$$(**) \quad \lim_{k\to\infty} \frac{1}{\sigma_k} \log\Big(\sum_{i \notin I_j} \exp(\sigma_k \delta_i)\Big) = \max_{i \notin I_j} \delta_i \leq \max_{i \notin I_j} \delta_i^* = \lim_{k\to\infty} \frac{1}{\sigma_k} \log\Big(\sum_{i \notin I_j} \exp(\sigma_k \delta_i^*)\Big).$$

Um dies einzusehen beachten wir, dass

$$\begin{aligned}
\max_{i \notin I_j} \delta_i &= \frac{1}{\sigma_k} \log\Big(\exp\Big(\sigma_k \max_{i \notin I_j} \delta_i\Big)\Big) \\
&\leq \frac{1}{\sigma_k} \log\Big(\sum_{i \notin I_j} \exp(\sigma_k \delta_i)\Big) \\
&\leq \frac{1}{\sigma_k} \log\Big((l - \#(I_j)) \exp\Big(\sigma_k \max_{i \notin I_j} \delta_i\Big)\Big) \\
&= \underbrace{\frac{\log(l - \#(I_j))}{\sigma_k}}_{\to 0} + \max_{i \notin I_j} \delta_i,
\end{aligned}$$

Mit $\delta_i \nearrow g_i(x^*)$ und $\delta_i^* \searrow g_i(x^\infty)$, $i \notin I_j$, folgt aus Stetigkeitsgründen, dass

$$\max_{i \notin I_j} g_i(x^*) \leq \max_{i \notin I_j} g_i(x^\infty).$$

Da $x^\infty \in S_{j+1}$, also eine Lösung der Aufgabe, $\phi_j(x) := \max_{i \notin I_j} g_i(x)$ auf S_j zu minimieren, ist es auch x^*. Also ist $x^* \in S_{j+1}$ und der Beweis ist vollständig. □
□

5.2.6 Konvergenz der dualen Trajektorie

Wir beschränken uns in diesem Unterabschnitt auf die Untersuchung der klassischen Methode der logarithmischen Barrieren. Wie schon in den letzten Unterabschnitten betrachten wir das konvexe, quadratisch restringierte quadratische Programm

$$(\mathrm{P}) \quad \begin{cases} \text{Minimiere} \quad f(x) := c_0^T x + \frac{1}{2} x^T Q_0 x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : g_i(x) := \beta_i + c_i^T x + \frac{1}{2} x^T Q_i x \le 0 \ (i = 1, \ldots, l),\ h(x) = 0\}, \end{cases}$$

wobei nach wie vor $Q_0, Q_1, \ldots, Q_l \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit sind, ferner $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ durch $g(x) := (g_1(x), \ldots, g_l(x))^T$ definiert und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear ist und die (konstante) Funktionalmatrix $h' \in \mathbb{R}^{m \times n}$ den vollen Rang m besitzt. Weiter wird vorausgesetzt, dass das relative Innere M^0 von M nichtleer und die Menge M_{opt} der Lösungen von (P) nichtleer und kompakt ist. Hierzu betrachteten wir das im wesentlichen unrestringierte Hilfsproblem

$$(\mathrm{P}_\sigma) \qquad \begin{cases} \text{Minimiere} \quad \Phi_\sigma(x) := f(x) - \dfrac{1}{\sigma} \displaystyle\sum_{i=1}^{l} \log(-g_i(x)) \quad \text{auf} \\ M^0 := \{x \in \mathbb{R}^n : g(x) < 0,\ h(x) = 0\} \end{cases}$$

und wissen bisher, dass (P_σ) für jedes $\sigma > 0$ genau eine Lösung $x_\sigma \in M^0$ besitzt und $x_\infty = \lim_{\sigma \to \infty} x_\sigma$ existiert und eine gewisse Lösung von (P) ist. Weiter ist die Lösung $x_\sigma \in M^0$ von (P_σ) durch die Existenz eines Vektors $v_\sigma \in \mathbb{R}^m$ mit

$$\nabla f(x_\sigma) + \sum_{i=1}^{l} \Big(-\frac{1}{\sigma g_i(x_\sigma)} \Big) \nabla g_i(x_\sigma) + (h')^T v_\sigma = 0$$

charakterisiert. Wegen der Rangvoraussetzung an h' ist v_σ eindeutig festgelegt. Wir wollen uns überlegen, dass die duale Trajektorie $\{(u_\sigma, v_\sigma) : \sigma > 0\}$, wobei $(u_\sigma)_i := -1/(\sigma g_i(x_\sigma))$, $i = 1, \ldots, l$, mit $\sigma \to \infty$ gegen eine Lösung (u^*, v^*) des zu (P) dualen Programms konvergiert. Dass dies richtig ist, ist genau die Aussage des folgenden Satzes. Zunächst erinnern wir aber an einige Ergebnisse der Dualitätstheorie bei konvexen Programmen (wobei hier die quadratische Struktur des primalen Programms (P) häufig, aber nicht immer irrelevant ist). Die Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ zu (P) ist bekanntlich durch

$$L(x, u, v) := f(x) + u^T g(x) + v^T h(x)$$

gegeben, das zugehörige duale Programm durch

$$(\mathrm{D}) \qquad \begin{cases} \text{Maximiere} \quad \phi(u, v) := \inf\limits_{x \in \mathbb{R}^n} L(x, u, v) \quad \text{auf} \\ N := \{(u, v) \in \mathbb{R}^l \times \mathbb{R}^m : u \ge 0,\ \phi(u, v) > -\infty\}. \end{cases}$$

Für das konvexe, quadratisch restringierte quadratische Programm (P) können wir aussagen, dass die Menge der dual zulässigen Lösungen sich darstellen läßt als

$$N = \{(u, v) \in \mathbb{R}^l \times \mathbb{R}^m : u \ge 0,\ \exists\ z \in \mathbb{R}^n \text{ mit } \nabla_x L(z, u, v) = 0\}.$$

Insbesondere ist (u_σ, v_σ) für jedes $\sigma > 0$ dual zulässig. Wegen Korollar 3.7 in Abschnitt 2.3 gibt es zu einer Lösung $x^* \in M_{\text{opt}}$ ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u^* \geq 0, \qquad \nabla_x L(x^*, u^*, v^*) = 0, \qquad g(x^*)^T u^* = 0.$$

Folglich ist $(u^*, v^*) \in N$ dual zulässig, wegen $\phi(u^*, v^*) = L(x^*, u^*, v^*) = f(x^*)$ und des schwachen Dualitätssatzes ist (u^*, v^*) eine Lösung von (D), insbesondere ist die Menge N_{opt} der Lösungen von (D) nichtleer.

Satz 3.5 *Gegeben sei das konvexe, quadratisch restringierte quadratische Programm (P), die Voraussetzungen (V1), (V2) seien erfüllt. Es sei $M^0 \neq \emptyset$. Ferner existiere ein strikt komplementäres optimales Paar*[20] *$(\hat{x}, (\hat{u}, \hat{v})) \in M_{\text{opt}} \times N_{\text{opt}}$, d. h. es sei $-g(\hat{x}) + \hat{u} > 0$. Zu gegebenem $\sigma > 0$ ist die Lösung $x_\sigma \in M^0$ von (P_σ) durch die Existenz eines Vektors v_σ mit*

$$\nabla f(x_\sigma) + g'(x_\sigma)^T u_\sigma + (h')^T v_\sigma = 0$$

charakterisiert, wobei $u_\sigma \in \mathbb{R}^l$ durch

$$(u_\sigma)_i := -\frac{1}{\sigma g_i(x_\sigma)}, \qquad i = 1, \ldots, l,$$

gegeben ist. Dann existiert $(u_\infty, v_\infty) := \lim_{\sigma \to \infty} (u_\sigma, v_\sigma)$ und ist eine Lösung des zu (P) dualen Programms (D).

Beweis: Wie im Beweis von Satz 3.2 definiere man die Indexmengen

$$I := \{i \in \{1, \ldots, l\} : g_i(x) = 0 \text{ für alle } x \in M_{\text{opt}}\}, \qquad J := \{1, \ldots, l\} \setminus I.$$

Weiter sei $x_\infty := \lim_{\sigma \to \infty} x_\sigma$, wobei wir schon wissen (siehe den Beweis von Satz 3.2), dass $g_j(x_\infty) < 0$, $j \in J$. Für jedes $(u, v) \in N_{\text{opt}}$ ist daher $u_j = 0$, $j \in J$. Dies erkennt man aus

$$f(x_\infty) = \phi(u, v) \leq f(x_\infty) + u^T g(x_\infty) + v^T \underbrace{h(x_\infty)}_{=0} \leq f(x_\infty).$$

Wir können im weiteren annehmen, dass $I \neq \emptyset$. Denn andernfalls ist $J = \{1, \ldots, l\}$ und daher $g(x_\infty) < 0$. Dies wiederum impliziert die Konvergenz $(u_\sigma, v_\sigma) \to (u_\infty, v_\infty) \in N_{\text{opt}}$ für $\sigma \to \infty$, wobei $u_\infty = 0$.

Im ersten Schritt zeigen wir:

- Für jedes $\sigma > 0$ ist (u_σ, v_σ) die eindeutige Lösung von

$$(\mathrm{D}_\sigma) \qquad \left\{ \begin{array}{c} \text{Maximiere} \quad \phi_\sigma(u, v) := \phi(u, v) + \dfrac{1}{\sigma} \displaystyle\sum_{i=1}^{l} \log u_i \quad \text{auf} \\ N^0 := \{(u, v) \in N : u > 0\}. \end{array} \right.$$

Ferner ist

$$(*) \qquad \sum_{i \in I} \frac{u_i}{(u_\sigma)_i} \leq l - \sum_{j \in J} \frac{g_j(x_\infty)}{g_j(x_\sigma)} \leq l \qquad \text{für alle } (u, v) \in N_{\text{opt}}.$$

[20] Dies ist bekanntlich für den Spezialfall linearer Programme keine zusätzliche Voraussetzung.

Denn: Zunächst beachten wir, dass $(u_\sigma, v_\sigma) \in N^0$. Für ein beliebiges $(u, v) \in N^0$ ist

$$\begin{aligned}
\phi_\sigma(u,v) - \phi_\sigma(u_\sigma, v_\sigma) &\leq L(x_\sigma, u, v) - L(x_\sigma, u_\sigma, v_\sigma) + \frac{1}{\sigma}\sum_{i=1}^{l}[\log u_i - \log(u_\sigma)_i] \\
&= (u - u_\sigma)^T g(x_\sigma) + \frac{1}{\sigma}\sum_{i=1}^{l}[\log u_i - \log(u_\sigma)_i] \\
&= -\frac{1}{\sigma}\sum_{i=1}^{l}[u_i - (u_\sigma)_i]\frac{1}{(u_\sigma)_i} + \frac{1}{\sigma}\sum_{i=1}^{l}[\log u_i - \log(u_\sigma)_i] \\
&\leq -\frac{1}{\sigma}\sum_{i=1}^{l}[u_i - (u_\sigma)_i]\frac{1}{(u_\sigma)_i} + \frac{1}{\sigma}\sum_{i=1}^{l}[u_i - (u_\sigma)_i]\frac{1}{(u_\sigma)_i} \\
&\qquad \text{(da der Logarithmus auf } \mathbb{R}_+ \text{ konkav)} \\
&= 0.
\end{aligned}$$

Also ist (u_σ, v_σ) eine Lösung von (D_σ). Ist $(u_\sigma^*, v_\sigma^*) \in N^0$ eine weitere Lösung von (D_σ), so folgt aus der obigen Gleichung-Ungleichungskette, dass

$$\phi(u_\sigma^*, v_\sigma^*) = L(x_\sigma, u_\sigma^*, v_\sigma^*), \qquad \log(u_\sigma^*)_i - \log(u_\sigma)_i = \frac{1}{(u_\sigma)_i}[(u_\sigma^*)_i - (u_\sigma)_i] \quad (i = 1, \ldots, l).$$

Aus dem zweiten Satz von l Gleichungen folgt $u_\sigma^* = u_\sigma$. Aus der ersten Beziehung folgt (wir nutzen $u_\sigma^* = u_\sigma$ schon aus)

$$\nabla_x L(x_\sigma, u_\sigma, v_\sigma^*) = 0 = \nabla_x L(x_\sigma, u_\sigma, v_\sigma),$$

wegen der Rangvoraussetzung an h' ist $v_\sigma^* = v_\sigma$, womit gezeigt ist, daß (u_σ, v_σ) die eindeutige Lösung von (D_σ) ist. Zum Beweis von $(*)$ geben wir uns $(u, v) \in N_{\mathrm{opt}}$ beliebig vor. Dann ist

$$(u(t), v(t)) := (u_\sigma, v_\sigma) + t[(u, v) - (u_\sigma, v_\sigma)] \in N^0 \qquad \text{für alle } t \in (0, 1]$$

und daher

$$\begin{aligned}
0 &\geq \frac{\phi_\sigma(u(t), v(t)) - \phi_\sigma(u_\sigma, v_\sigma)}{t} \\
&= \frac{\phi(u(t), v(t)) - \phi(u_\sigma, v_\sigma)}{t} + \frac{1}{\sigma}\sum_{i=1}^{l}\frac{\log[(u_\sigma)_i + t(u_i - (u_\sigma)_i)] - \log(u_\sigma)_i}{t} \\
&\geq \frac{(1-t)\phi(u_\sigma, v_\sigma) + t\phi(u, v) - \phi(u_\sigma, v_\sigma)}{t} + \frac{1}{\sigma}\sum_{i=1}^{l}\frac{u_i - (u_\sigma)_i}{(u_\sigma)_i + t(u_i - (u_\sigma)_i)} \\
&\qquad \text{(da } \phi \text{ konkav auf } N \text{ und log konkav auf } \mathbb{R}_+).
\end{aligned}$$

Mit $t \to 0+$ folgt

$$0 \geq \phi(u, v) - \phi(u_\sigma, v_\sigma) + \frac{1}{\sigma}\sum_{i=1}^{l}\frac{u_i - (u_\sigma)_i}{(u_\sigma)_i}$$

und damit

$$\begin{aligned}
\sum_{i\in I}\frac{u_i}{(u_\sigma)_i} &= \sum_{i=1}^{l}\frac{u_i}{(u_\sigma)_i}\\
&\le l+\sigma[\phi(u_\sigma,v_\sigma)-\phi(u,v)]\\
&= l+\sigma[\phi(u_\sigma,v_\sigma)-f(x_\infty)]\\
&\le l+\sigma[L(x_\infty,u_\sigma,v_\sigma)-f(x_\infty)]\\
&= l+\sigma u_\sigma^T g(x_\infty)\\
&= l-\sum_{i=1}^{l}\frac{g_i(x_\infty)}{g_i(x_\sigma)}\\
&= l-\sum_{j\in J}\frac{g_j(x_\infty)}{g_j(x_\sigma)}\\
&\le l.
\end{aligned}$$

Damit ist $(*)$ bewiesen. Im nächsten Schritt zeigen wir:

- Sei $\{\sigma_k\}\subset\mathbb{R}_+$ eine Folge mit $\sigma_k\to\infty$, zur Abkürzung setze man $(u_k,v_k):=(u_{\sigma_k},v_{\sigma_k})$. Dann gilt:
  - Die Folge $\{u_k\}$ ist beschränkt.
  - Konvergiert die Teilfolge $\{u_k\}_{k\in K}\subset\{u_k\}$ gegen u^*, so konvergiert die Teilfolge $\{v_k\}_{k\in K}\subset\{v_k\}$ gegen ein v^*. Ferner ist $(u^*,v^*)\in N_{\text{opt}}$, $u_i^*>0$, $i\in I$, und (u^*,v^*) eine Lösung von

$$(\mathrm{D}_\infty)\qquad \begin{cases} \text{Maximiere}\quad \phi_\infty(u,v):=\displaystyle\sum_{i\in I}\log u_i \quad\text{auf}\\[2ex] N_{\text{opt}}^*:=\{(u,v)\in N_{\text{opt}}: u_i>0\ \ (i\in I)\}. \end{cases}$$

  - Sind $(u^*,v^*),(u^{**},v^{**})\in N_{\text{opt}}^*$ zwei Lösungen von (D_∞), so ist $u^*=u^{**}$.

Denn: Sei $x_k:=x_{\sigma_k}$ die Lösung von (P_{σ_k}). Es ist

$$\phi(u_k,v_k)=L(x_k,u_k,v_k)=f(x_k)-\frac{l}{\sigma_k}\ge\min(\mathrm{P})-\frac{l}{\sigma}=:\delta,$$

wobei $0<\sigma\le\sigma_k$ für alle k. Hieraus folgt, dass $\{u_k\}$ beschränkt ist. Denn nach Voraussetzung existiert ein $\tilde{x}\in M^0$, daher ist

$$\delta\le\phi(u_k,v_k)\le L(\tilde{x},u_k,v_k)=f(\tilde{x})+u_k^Tg(\tilde{x})\le f(\tilde{x})-\epsilon\,\|u_k\|_1$$

mit $\epsilon:=\min_{i=1,\dots,l}(-g_i(\tilde{x}))$ und folglich $\|u_k\|_1\le[f(\tilde{x})-\delta]/\epsilon$. Aus

$$\nabla_x L(x_k,u_k,v_k)=\nabla f(x_k)+g'(x_k)^Tu_k+(h')^Tv_k=0,$$

der Konvergenz von $\{x_k\}$ und Rang $(h') = m$ folgt aus der Konvergenz der Teilfolge $\{u_k\}_{k\in K}$ gegen ein u^* auch die Konvergenz der entsprechenden Teilfolge $\{v_k\}_{k\in K}$ gegen ein v^*. Wegen $u_k > 0$, $\nabla_x L(x_k, u_k, v_k) = 0$ und $u_k^T g(x_k) = -l/\sigma_k$ folgt

$$u^* \geq 0, \qquad \nabla_x L(x_\infty, u^*, v^*) = 0, \qquad (u^*)^T g(x_\infty) = 0,$$

woraus sich wiederum $(u^*, v^*) \in N_{\text{opt}}$ ergibt. Nach Voraussetzung existiert zu (P) ein Paar $(\hat{x}, (\hat{u}, \hat{v})) \in M_{\text{opt}} \times N_{\text{opt}}$ mit $-g(\hat{x}) + \hat{u} > 0$. Nach Definition der Indexmenge I ist $g_i(\hat{x}) = 0$, $i \in I$, und daher $\hat{u}_i > 0$, $i \in I$. Aus der Beziehung $(*)$ im ersten Beweisschritt erhält man

$$\sum_{i\in I} \frac{\hat{u}_i}{(u_k)_i} \leq l \qquad \text{für alle } k$$

und hieraus $u_i^* > 0$, $i \in I$. Nun zeigen wir, dass (u^*, v^*) eine Lösung von (D_∞) ist. Eben haben wir schon bewiesen, dass $(u^*, v^*) \in N_{\text{opt}}^*$, also (u^*, v^*) zulässig für (D_∞) ist. Sei $(u, v) \in N_{\text{opt}}^*$ beliebig. Aus $(*)$ im ersten Beweisschritt erhalten wir

$$\sum_{i\in I} \frac{u_i}{(u_k)_i} \leq l - \sum_{j\in J} \frac{g_j(x_\infty)}{g_j(x_k)} \qquad \text{für alle } k.$$

Mit dem Grenzübergang $k \to \infty$, $k \in K$, erhalten wir wegen $g_j(x_k) \to g_j(x_\infty) < 0$, $j \in J$, daß

$$\sum_{i\in I} \frac{u_i}{u_i^*} \leq l - \#(J) = \#(I)$$

bzw.

$$\frac{1}{\#(I)} \sum_{i\in I} \frac{u_i}{u_i^*} \leq 1.$$

Die Ungleichung vom geometrisch-arithmetischen Mittel liefert

$$\prod_{i\in I} \frac{u_i}{u_i^*} \leq 1,$$

anschließendes Logarithmieren $\phi_\infty(u, v) \leq \phi_\infty(u^*, v^*)$. Daher ist (u^*, v^*) eine Lösung von (D_∞). Ist $(u^{**}, v^{**}) \in N_{\text{opt}}^*$ eine weitere Lösung, so folgt aus der strikten Konkavität des Logarithmus auf $\mathbb{R}_+$, daß $u_i^{**} = u_i^*$, $i \in I$. Andererseits ist $u_i^{**} = u_i^* = 0$, $i \notin I$, und daher $u^{**} = u^*$. Damit ist der zweite Beweisschritt abgeschlossen.

Nun zum Schluss des Beweises. Wir zeigen die Konvergenz der Folge $\{u_k\}$ gegen die eindeutige erste Komponente u^* einer Lösung (u^*, v^*) von (D_∞). Denn angenommen, $u_k \not\to u^*$. Dann existiert eine Teilfolge $\{u_k\}_{k\in K}$ und ein $\epsilon > 0$ mit $\|u_k - u^*\| \geq \epsilon$ für alle $k \in K$. Aus $\{u_k\}_{k\in K}$ kann eine gegen ein u^{**} konvergente Teilfolge $\{u_k\}_{k\in K_1}$ mit $K_1 \subset K$ ausgewählt werden. Wir wissen, dass auch die Folge $\{v_k\}_{k\in K_1}$ gegen ein v^{**} konvergent ist und (u^{**}, v^{**}) eine Lösung von (D_∞) ist. Also ist $u^* = u^{**}$, ein Widerspruch zu $\|u^{**} - u^*\| \geq \epsilon > 0$. Damit ist der Satz schließlich bewiesen. □ □

Bemerkung: Es ist möglich, den letzten Satz auf weitere Straf- bzw. Barrierefunktionen zu übertragen, worauf wir allerdings nicht mehr eingehen wollen. □

5.2.7 Primal-duale Verfahren bei konvexen, quadratisch restringierten quadratischen Programmen

Wie in den letzten beiden Abschnitt betrachten wir auch hier das konvexe, quadratisch restringierte quadratische Programm

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\},$$

wobei

$$f(x) := c_0^T x + \tfrac{1}{2} x^T Q_0 x, \quad g_i(x) := \beta_i + c_i^T x + \tfrac{1}{2} x^T Q_i x \quad (i = 1, \ldots, l), \quad h(x) := b - Ax$$

mit symmetrischen, positiv semidefiniten Matrizen $Q_0, Q_1, \ldots, Q_l$. Weiter setzen wir voraus, daß Rang $(h') = m$ maximal ist. Bei festem $\sigma > 0$ definiere man die Abbildung

$$F_\sigma \colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^l \longrightarrow \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^l$$

durch

$$F_\sigma(x, u, v, z) := \begin{pmatrix} \nabla f(x) + g'(x)^T u + (h')^T v \\ g(x) + z \\ h(x) \\ \sigma U z - e \end{pmatrix}.$$

Für $u \in \mathbb{R}^l$ sei hierbei $U := \operatorname{diag}(u_1, \ldots, u_l)$, entsprechende Bezeichnungen benutzen wir auch für andere Vektoren. Dann gilt:

Lemma 3.6 *Unter den Voraussetzungen und mit den Bezeichnungen von Satz 3.5 besitzt das nichtlineare Gleichungssystem*

$$F_\sigma(x, u, v, z) = 0$$

genau eine Lösung (x, u, v, z) mit $u > 0$ und $z > 0$. Diese ist durch $(x_\sigma, u_\sigma, v_\sigma, -g(x_\sigma))$ gegeben. Ferner ist die Funktionalmatrix $F'_\sigma(x, u, v, z)$ für alle $(x, u, v, z) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^l$ mit $u > 0$, $z > 0$ nichtsingulär.

Beweis: Offenbar ist (x, u, v, z) genau dann eine Nullstelle von F_σ, wenn $x \in M^0$, $u_i = -1/(\sigma g_i(x_\sigma))$, $i = 1, \ldots, l$, und $\nabla f(x) + g'(x)^T u + (h')^T v = 0$, was wiederum äquivalent dazu ist, daß x die eindeutige Lösung von (P_σ) ist. Hieraus folgt die Aussage des ersten Teiles des Satzes. Als Funktionalmatrix von F_σ berechnen wir

$$F'_\sigma(x, u, v, z) = \begin{pmatrix} Q(u) & g'(x)^T & (h')^T & 0 \\ g'(x) & 0 & 0 & I \\ h' & 0 & 0 & 0 \\ 0 & \sigma Z & 0 & \sigma U \end{pmatrix},$$

wobei wir zur Abkürzung

$$Q(u) := Q_0 + \sum_{i=1}^{l} u_i Q_i$$

gesetzt haben. Angenommen, (p^x, p^u, p^v, p^z) sei aus dem Kern von $F'_\sigma(x,u,v,z)$. Das ist gleichbedeutend mit

$$\begin{array}{rcrcrcrcl} Q(u)p^x & + & g'(x)^T p^u & + & (h')^T p^v & & & = & 0, \\ g'(x)p^x & & & & & + & p^z & = & 0, \\ h'p^x & & & & & & & = & 0, \\ & & \sigma Z p^u & & & + & \sigma U p^z & = & 0. \end{array}$$

Multipliziert man die erste Gleichung von links mit $(p^x)^T$ und nutzt man die dritte, anschließend die zweite und die vierte Gleichung aus, so erhält man

$$\begin{aligned} 0 &= (p^x)^T Q(u) p^x + (g'(x)p^x)^T p^u \\ &= (p^x)^T Q(u) p^x - (p^z)^T p^u \\ &= (p^x)^T Q(u) p^x + (p^u)^T U^{-1} Z p^u. \end{aligned}$$

Nun ist $Q(u)$ positiv semidedinit, die positive Diagonalmatrix $U^{-1}Z$ ist postiv definit und daher $Q(u)p^x = 0$ und $p^u = 0$. Aus der vierten der obigen Gleichungen folgt, daß auch $p^z = 0$, die zweite Gleichung ergibt $g'(x)p^x = 0$. Wegen $u > 0$ folgt aus $Q(u)p^x = 0$, daß $Q_i p^x = 0$, $i = 1, \ldots, l$, anschließend aus $g'(x)p^x$, daß $c_i^T p^x = 0$, $i = 1, \ldots, l$. Da M_{opt} nach Voraussetzung nichtleer und kompakt ist, ist $p^x = 0$. Aus der Rangvoraussetzung an h' und der ersten Gleichung schließt man dann noch auf $p^v = 0$, womit bewiesen ist, daß der Kern von $F'_\sigma(x,u,v,z)$ trivial bzw. $F'_\sigma(x,u,v,z)$ für $u > 0$, $z > 0$ nichtsingulär ist. Damit ist das Lemma bewiesen. □ □

Wegen des zweiten Teiles des vorigen Lemmas ist für ein Quadrupel (x,u,v,z) mit $u > 0$, $z > 0$ die Newton-Richtung als Lösung des linearen Gleichungssystems

$$\begin{pmatrix} Q(u) & g'(x)^T & (h')^T & 0 \\ g'(x) & 0 & 0 & I \\ h' & 0 & 0 & 0 \\ 0 & \sigma Z & 0 & \sigma U \end{pmatrix} \begin{pmatrix} p^x \\ p^u \\ p^v \\ p^z \end{pmatrix} = - \begin{pmatrix} \nabla f(x) + g'(x)^T u + (h')^T v \\ g(x) + z \\ h(x) \\ \sigma U z - e \end{pmatrix}$$

erklärt. Ist $x \in M^0$ und $z = -g(x)$ (dann ist automatisch $z > 0$), so lautet dieses lineare Gleichungssystem

$$\begin{pmatrix} Q(u) & g'(x)^T & (h')^T & 0 \\ g'(x) & 0 & 0 & I \\ h' & 0 & 0 & 0 \\ 0 & \sigma Z & 0 & \sigma U \end{pmatrix} \begin{pmatrix} p^x \\ p^u \\ p^v \\ p^z \end{pmatrix} = - \begin{pmatrix} \nabla f(x) + g'(x)^T u + (h')^T v \\ 0 \\ 0 \\ \sigma U z - e \end{pmatrix}.$$

Mit Hilfe der vierten Gleichung kann p^z eliminiert und durch p^u ausgedrückt werden:

$$p^z = -U^{-1} Z p^u + (1/\sigma) U^{-1} e - z.$$

Das reduzierte lineare Gleichungssystem lautet dann

$$\begin{pmatrix} Q(u) & g'(x)^T & (h')^T \\ g'(x) & -U^{-1}Z & 0 \\ h' & 0 & 0 \end{pmatrix} \begin{pmatrix} p^x \\ p^u \\ p^v \end{pmatrix} = - \begin{pmatrix} \nabla f(x) + g'(x)^T u + (h')^T v \\ (1/\sigma) U^{-1} e - z \\ 0 \end{pmatrix}.$$

Ein Modellalgorithmus für einen Schritt eines primal-duales Innere-Punkt-Verfahrens zur Lösung von (P) (und des dazu dualen Programms (D)) könnte folgendermaßen aussehen:

- Gegeben $(x, u, v) \in M^0 \times \mathbb{R}^l_{>0} \times \mathbb{R}^m$. Berechne $z := -g(x)$.
- Mit einem $\rho > 1$ berechne
$$\sigma := \frac{\rho n}{u^T z}.$$
- Berechne die Newton-Richtung durch Lösen des linearen Gleichungssystems
$$\begin{pmatrix} Q(u) & g'(x)^T & (h')^T \\ g'(x) & -U^{-1}Z & 0 \\ h' & 0 & 0 \end{pmatrix} \begin{pmatrix} p^x \\ p^u \\ p^v \end{pmatrix} = - \begin{pmatrix} \nabla f(x) + g'(x)^T u + (h')^T v \\ (1/\sigma)U^{-1}e - z \\ 0 \end{pmatrix}.$$
- Berechne die maximale Schrittweite
$$t_{\max} := \sup\{t > 0 : g(x + tp^x) < 0,\ u + tp^u > 0\}.$$
- Mit einem $\tau \in (0, 1)$ berechne man $t := \min(1, \tau t_{\max})$.
- Berechne
$$\begin{pmatrix} x_+ \\ u_+ \\ v_+ \end{pmatrix} := \begin{pmatrix} x \\ u \\ v \end{pmatrix} + t \begin{pmatrix} p^x \\ p^u \\ p^v \end{pmatrix}.$$

Für einen konkreten Algorithmus muß insbesondere erklärt werden, wie ρ und τ zu bestimmen sind. In einer späteren Bemerkung werden wir zeigen, daß die maximale Schrittweite bei konvexen, quadratisch restringierten quadratischen Problemen verhältnismäßig einfach berechnet werden kann. Die hier vorgeschlagene Wahl des Parameters σ wird später bei linearen Programmen motiviert. Leider scheint es bisher keine Konvergenzaussagen (bis auf Spezialfälle, etwa lineare Programme) für das oben angegebene primal-duale Innere-Punkt-Verfahren zu geben.

Beispiel: Von J. J. Sylvester (1857) stammt die Aufgabe, zu vorgegebenen Punkten $a_1, \ldots, a_l \in \mathbb{R}^n$ (bei Sylvester ist $n = 2$) diejenige euklidische Kugel zu finden, die unter der Nebenbedingung, daß sie die vorgegebenen Punkte $a_1, \ldots, a_l$ enthält, minimalen Radius besitzt. Hierzu formulieren wir die Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x, \delta) := \delta \quad \text{auf} \\ M := \{(x, \delta) \in \mathbb{R}^n \times \mathbb{R} : g_i(x, \delta) := \frac{1}{2}\|x - a_i\|^2 - \delta \le 0,\ i = 1, \ldots, l\}. \end{cases}$$

Hierbei bedeute $\|\cdot\|$ natürlich die euklidische Norm auf dem $\mathbb{R}^n$. Bei (P) handelt es sich offensichtlich um ein konvexes Problem, ferner ist

$$M_0 := \{(x, \delta) \in \mathbb{R}^n \times \mathbb{R} : g_i(x, \delta) := \tfrac{1}{2}\|x - a_i\|^2 - \delta < 0,\ i = 1, \ldots, l\} \neq \emptyset,$$

denn hierzu braucht man sich ja natürlich nur $x \in \mathbb{R}^n$ beliebig zu wählen (z. B. $x := (1/l)\sum_{i=1}^l a_i$) und anschließend ein hinreichend großes $\delta > 0$ zu bestimmen.

Wir wollen uns überlegen, daß (P) eindeutig lösbar ist und damit insbesondere M_{opt} einpunktig und insbesondere kompakt ist. Die Lösbarkeit erhält man offenbar sofort durch die Beobachtung, daß eine Niveaumenge zu (P) kompakt ist. Die Eindeutigkeit kann man folgendermaßen einsehen: Sind (x_1^*, δ^*) und (x_2^*, δ^*) zwei Lösungen von (P), so ist natürlich auch (x^*, δ^*) mit $x^* := \frac{1}{2}(x_1^* + x_2^*)$ eine Lösung von (P). Dann ist

$$\sqrt{2\delta^*} = \max_{i=1,\dots,l} \|\tfrac{1}{2}(x_1^* - a_i) + \tfrac{1}{2}(x_2^* - a_i)\| = \|\tfrac{1}{2}(x_1^* - a_j) + \tfrac{1}{2}(x_2^* - a_j)\|$$

mit einem $j \in \{1, \dots, l\}$. Dann ist aber

$$\begin{aligned} \sqrt{2\delta^*} &= \|\tfrac{1}{2}(x_1^* - a_j) + \tfrac{1}{2}(x_2^* - a_j)\| \\ &\le \tfrac{1}{2}\|x_1^* - a_j\| + \tfrac{1}{2}\|x_2^* - a_j\| \\ &\le \tfrac{1}{2} \max_{i=1,\dots,l} \|x_1^* - a_i\| + \tfrac{1}{2} \max_{i=1,\dots,l} \|x_2^* - a_i\| \\ &= \sqrt{2\delta^*}. \end{aligned}$$

Da die euklidische Norm strikt konvex ist, folgt hieraus $x_1^* = x_2^*$, insgesamt also die eindeutige Lösbarkeit von (P). Nun wollen wir das zu (P) duale Programm aufstellen. Die Lagrange-Funktion zu (P) ist durch

$$L((x, \delta), u) := \delta + \sum_{i=1}^{l} u_i [\tfrac{1}{2}\|x - a_i\|^2 - \delta]$$

bzw.

$$L((x, \delta), u) = \Big(1 - \sum_{i=1}^{l} u_i\Big)\delta + \tfrac{1}{2} \sum_{i=1}^{l} u_i \, \|x - a_i\|^2$$

gegeben. Hieraus liest man ab, daß

$$N := \{u \in \mathbb{R}^l : u \ge 0, \; e^T u = 1\}$$

die Menge der dual zulässigen Lösungen ist, wobei e einmal wieder den Vektor (des $\mathbb{R}^l$) bedeutet, dessen Komponenten alle gleich 1 sind. Die auf N zu maximierende Zielfunktion im dualen Programm ist

$$\phi(u) = L((z, \eta), u) = \tfrac{1}{2} \sum_{i=1}^{l} u_i \, \|x - a_i\|^2,$$

wobei $(z, \eta) \in \mathbb{R}^n \times \mathbb{R}$ aus

$$0 = \nabla_{(x,\delta)} L((z, \eta), u) = \begin{pmatrix} \sum_{i=1}^{l} u_i (z - a_i) \\ 0 \end{pmatrix}$$

zu bestimmen ist. Dies führt auf $z = \sum_{j=1}^{l} u_j a_j$, anschließend berechnet man die duale Zielfunktion durch

$$\phi(u) = \tfrac{1}{2} \sum_{i=1}^{l} u_i \Big\| \sum_{j=1}^{l} u_j a_j - a_i \Big\|^2 = \tfrac{1}{2} \sum_{i=1}^{l} u_i \, \|a_i\|^2 - \tfrac{1}{2} \Big\| \sum_{i=1}^{l} u_i a_i \Big\|^2.$$

Nun wollen wir noch das beim primal-dualen Verfahren auftretende lineare Gleichungssystem genauer betrachten. Hierbei gehen wir davon aus, daß $(x, \delta) \in M_0$ und $u \in N_0$ gegeben sind, wobei natürlich

$$N_0 := \{u \in \mathbb{R}^l : u > 0,\ e^T u = 1\}.$$

Mit der oben benutzten Notation ist

$$Q(u) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \sum_{i=1}^{l} u_i \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}.$$

Mit

$$g(x, \delta) := \begin{pmatrix} \frac{1}{2}\|x - a_1\|^2 - \delta \\ \vdots \\ \frac{1}{2}\|x - a_l\|^2 - \delta \end{pmatrix}$$

ist weiter

$$g'(x, \delta) = \begin{pmatrix} (x - a_1)^T & -1 \\ \vdots & \vdots \\ (x - a_l)^T & -1 \end{pmatrix}.$$

Für einen Newton-Schritt ist das lineare Gleichungssystem

$$\begin{pmatrix} I & 0 & x - a_1 & \cdots & x - a_l \\ 0 & 0 & -1 & \cdots & -1 \\ (x - a_1)^T & -1 & g_1(x, \delta)/u_1 & & \\ \vdots & \vdots & & \ddots & \\ (x - a_l)^T & -1 & & & g_l(x, \delta)/u_l \end{pmatrix} \begin{pmatrix} p^x \\ p^\delta \\ p_1^u \\ \vdots \\ p_l^u \end{pmatrix} = - \begin{pmatrix} x - \sum_{i=1}^{l} u_i a_i \\ 0 \\ 1/(\sigma u_1) + g_1(x, \delta) \\ \vdots \\ 1/(\sigma u_l) + g_l(x, \delta) \end{pmatrix}$$

zu lösen. □

Bemerkung: Die maximale Schrittweite kann bei konvexen, quadratisch restringierten Problemen, wie wir sie hier betrachten, noch relativ einfach berechnet werden. Denn sei $g(x) < 0$. Zur Abkürzung setze man

$$\gamma_i := g_i(x), \qquad \delta_i := \nabla g_i(x)^T p^x, \qquad \epsilon_i := (p^x)^T Q_i p^x, \qquad i = 1, \ldots, l.$$

Dann ist

$$g_i(x + tp^x) = \gamma_i + \delta_i t + \tfrac{1}{2}\epsilon_i t^2.$$

Anschließend berechne man

$$s_i := \begin{cases} +\infty, & \text{falls } \epsilon_i = 0 \text{ und } \delta_i \le 0, \\ -\dfrac{\gamma_i}{\delta_i}, & \text{falls } \epsilon_i = 0 \text{ und } \delta_i > 0, \\ -\dfrac{\delta_i}{\epsilon_i} + \sqrt{\left(\dfrac{\delta_i}{\epsilon_i}\right)^2 - 2\dfrac{\gamma_i}{\epsilon_i}}, & \text{falls } \epsilon_i > 0, \end{cases} \qquad i = 1, \ldots, l.$$

Definiert man $t^1_{\max} := \min_{i=1,\dots,l} s_i$, so ist $g(x + tp^x) < 0$ für alle $t \in [0, t^1_{\max})$. Definiert man ferner

$$t^2_{\max} := \min_{i=1,\dots,l} \Big\{ -\frac{u_i}{p^u_i} : p^u_i < 0 \Big\},$$

so ist offenbar durch

$$t_{\max} := \min(t^1_{\max}, t^2_{\max})$$

die maximale Schrittweite in x in Richtung der Newton-Richtung bestimmt. □

Beispiel: Als Spezialfall[21] von (P) betrachten wir ein lineares Programm in Normalform, also

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b\}.$$

Hierbei sei $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$. Weiter wird vorausgesetzt, daß

$$M_0 := \{x \in \mathbb{R}^n : x > 0,\ Ax = b\} \neq \emptyset, \qquad N_0 := \{y \in \mathbb{R}^m : A^T y < c\} \neq \emptyset.$$

Wir wissen (siehe Aufgabe 7 in Abschnitt 2.2), daß dann die Mengen M_{opt} der Lösungen von (P) und N_{opt} der Lösungen des zu (P) dualen linearen Programms

$$\text{(D)} \qquad \text{Maximiere} \quad b^T y \quad \text{auf} \quad N := \{y \in \mathbb{R}^m : A^T y \leq c\}$$

nichtleer und kompakt sind. Mit $f(x) := c^T x$, $g(x) := -x$ und $h(x) := b - Ax$ ordnet sich das lineare Programm (P) in Normalform der bisher betrachteten allgemeinen Problemstellung unter. Das lineare Gleichungssystem zur Bestimmung der Newton-Richtung lautet (man beachte, daß jetzt $z = x$)

$$\begin{pmatrix} 0 & I & A^T \\ I & U^{-1}X & 0 \\ A & 0 & 0 \end{pmatrix} \begin{pmatrix} p^x \\ p^u \\ p^v \end{pmatrix} = \begin{pmatrix} c - u - A^T v \\ (1/\sigma)U^{-1}e - x \\ 0 \end{pmatrix}.$$

Ist im Ausgangstripel (x, u, v) das Paar (x, v) strikt zulässig für (P) bzw. (D), also $(x, v) \in M_0 \times N_0$, und $u = c - A^T v$ (dann notwendigerweise ein positiver Vektor), so vereinfacht sich das letzte lineare Gleichungssystem zu

$$\begin{pmatrix} 0 & I & A^T \\ I & U^{-1}X & 0 \\ A & 0 & 0 \end{pmatrix} \begin{pmatrix} p^x \\ p^u \\ p^v \end{pmatrix} = \begin{pmatrix} 0 \\ (1/\sigma)U^{-1}e - x \\ 0 \end{pmatrix}.$$

Aus der ersten Gleichung erhält man $p^u = -A^T p^v$, Einsetzen in die zweite Gleichung liefert nach anschließender Multiplikation mit A, daß

$$p^v = -(AU^{-1}XA^T)^{-1}[(1/\sigma)AU^{-1}e - b].$$

[21] Primal-duale Innere-Punkt-Verfahren bei linearen Programmen werden ausführlich bei S. J. Wright (1997) behandelt.

Wichtiger aber ist, daß man p^u über die erste Gleichung eliminieren kann. Die zweite Gleichung lautet dann

$$p^x - U^{-1}XA^Tp^v = (1/\sigma)U^{-1}e - x$$

bzw. nach Multiplikation von $X^{-1}U$ unter Berücksichtigung der Tatsache, daß Diagonalmatrizen vertauschbar sind

$$X^{-1}Up^x - A^Tp^v = (1/\sigma)X^{-1}e - u.$$

Die Größe von

$$x^Tu = x^T(c - A^Tv) = c^Tx - b^Tv$$

gibt genau die Dualitätslücke an und kann als ein Maß für die Güte von (x, v) aufgefasst werden. Wir wollen untersuchen, wie sich diese Dualitätslücke verändert, wenn man einen gewissen Schritt t in die Newton-Richtung geht. Hierzu definieren wir

$$x(t) := x + tp^x, \qquad u(t) := u + tp^u, \qquad v(t) := v + tp^v.$$

Dann ist

$$\begin{aligned} x(t)^Tu(t) &= (x + tp^x)^T(u + tp^u) \\ &= (x + tp^x)^T(c - A^Tv - tA^Tp^v) \\ &= (x + tp^x)(c - A^Tv(t)) \\ &= c^Tx(t) - b^Tv(t), \end{aligned}$$

also wird auch in einem neuen Schritt durch $x(t)^Tu(t)$ die Dualitätslücke angegeben (wenn nur $t > 0$ so klein, daß $x(t) > 0$ und $u(t) > 0$, so daß $x(t) \in M_0$ und $v(t) \in N_0$). Nun rechnen wir die Dualitätslücke genauer aus:

$$\begin{aligned} x(t)^Tu(t) &= c^T(x + tp^x) - b^T(v + tp^v) \\ &= c^Tx - b^Tv + t(\underbrace{c - A^Tv}_{=u})^Tp^x - tb^Tp^v \\ &\qquad (\text{wegen } Ap^x = 0) \\ &= c^Tx - b^Tv + te^TUp^x - te^TXA^Tp^v \\ &= c^Tx - b^Tv + te^T(Up^x + Xp^u) \\ &\qquad (\text{wegen } p^u = -A^Tp^v) \\ &= c^Tx - b^Tv + te^T[(1/\sigma)e - Ux] \\ &= c^Tx - b^Tv + t\frac{n}{\sigma} - tu^Tx \\ &= \Big[1 - t\Big(1 - \frac{1}{\rho}\Big)\Big]x^Tu, \end{aligned}$$

wenn

$$\sigma = \frac{n\rho}{x^Tu},$$

wobei sinnvollerweise $\rho > 1$. In der Praxis geht man folgendermaßen vor:

- Gegeben $(x, v) \in M_0 \times N_0$ und $u \in \mathbb{R}^m$ mit $u = c - A^T v$.
- Mit einem $\rho > 1$ berechne
$$\sigma := \frac{n\rho}{x^T u}.$$
- Berechne die Lösung (p^x, p^v) des linearen Gleichungssystems
$$\begin{pmatrix} X^{-1}U & -A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p^x \\ p^v \end{pmatrix} = \begin{pmatrix} (1/\sigma)X^{-1}e - u \\ 0 \end{pmatrix},$$
anschließend $p^u := -A^T p^v$.
- Berechne die maximale Schrittweite: Zunächst berechne
$$t^x_{\max} := \min_{j=1,\dots,n} \Big\{ -\frac{x_j}{p^x_j} : p^x_j < 0 \Big\}, \qquad t^u_{\max} := \min_{i=1,\dots,m} \Big\{ -\frac{u_i}{p^u_i} : p^u_i < 0 \Big\},$$
anschließend
$$t_{\max} := \min(t^x_{\max}, t^u_{\max}).$$
- Berechne die Schrittweite: Mit einem $\tau \in (0, 1)$ setze $t := \min(1, \tau t_{\max})$.
- Mache den Update
$$x_+ := x + tp^x, \qquad v_+ := v + tp^v, \qquad u_+ = u + tp^u.$$
Offenbar ist dann $(x_+, v_+) \in M_0 \times N_0$ und $u_+ = c - A^T v_+$, das neue Tripel (x_+, u_+, v_+) genügt also den Eingangsvoraussetzungen.

Hiermit ist ein typischer Schritt für ein primal-duales Innere-Punkt-Verfahren bei einer linearen Optimierungsaufgabe in Normalform beschrieben. Um aus diesem Modellalgorithmus ein implementierbares Verfahren zu machen, müßte die Wahl der (von der Iterationsstufe abhängigen) Parameter $\rho > 1$ und $\tau \in (0, 1)$ spezifiziert werden. Versehen wir die Näherungen mit Iterationsindizes, schreiben also x_k statt x, x_{k+1} statt x_+ usw., ferner ρ_k statt ρ usw., so erhalten wir
$$\frac{x^T_{k+1} v_{k+1}}{x^T_k v_k} = 1 - t_k \Big(1 - \frac{1}{\rho_k}\Big).$$
Ziel ist es natürlich, die Dualitätslücke möglichst schnell gegen Null konvergieren zu lassen. Z. B. liegt lineare Konvergenz vor, wenn eine Konstante $\delta > 0$ mit
$$t_k \Big(1 - \frac{1}{\rho_k}\Big) \geq \delta$$
existiert, während superlineare Konvergenz für
$$\lim_{k\to\infty} t_k \Big(1 - \frac{1}{\rho_k}\Big) = 1$$
gegeben ist. Letzteres ist etwa der Fall, wenn $t_k \to 1$ und $\rho_k \to +\infty$. Mit diesen etwas vagen Andeutungen zur Konvergenz bei primal-dualen Innere-Punkt-Verfahren bei linearen Optimierungsaufgaben wollen wir es genug sei lassen. □

5.2.8 Aufgaben

1. Sei $C \subset \mathbb{R}^n$ nichtleer, konvex und abgeschlossen. Für ein $x \in C$ und ein $p \in \mathbb{R}^n$ sei $\{x + tp : t \geq 0\} \subset C$, also der gesamte von x in Richtung p ausgehende Halbstrahl in C enthalten. Man zeige, dass für ein beliebiges $z \in C$ auch der Halbstrahl $\{z + tp : t \geq 0\}$ in C enthalten ist.

2. Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex. Man zeige:

   (a) Für jedes $x \in \mathbb{R}^n$ und jedes $p \in \mathbb{R}^n$ existiert (im eigentlichen oder uneigentlichen Sinne)
   $$f_\infty(p) := \lim_{t\to\infty} \frac{f(x+tp) - f(x)}{t}$$
   und ist durch
   $$f_\infty(p) = \sup_{z\in\mathbb{R}^n} [f(z+p) - f(z)]$$
   gegeben, ist also insbesondere (wie die Notation es erwarten lässt) von x unabhängig.

   (b) Die konvexe Optimierungsaufgabe

   (P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$

   sei zulässig. Dann ist die Menge M_{opt} der Lösungen von (P) genau dann nichtleer und kompakt, wenn das System
   $$f_\infty(p) \leq 0, \qquad (g_i)_\infty(p) \leq 0 \quad (i = 1, \ldots, l), \qquad h'p = 0$$
   nur trivial lösbar ist.

3. Gegeben sei die konvexe, quadratisch restringierte quadratische Optimierungsaufgabe

   (P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$,

   wobei
   $$f(x) := c_0^T x + \tfrac{1}{2} x^T Q_0 x, \qquad g_i(x) := \beta_i + c_i^T x + \tfrac{1}{2} x^T Q_i x \quad (i = 1, \ldots, l)$$
   und
   $$h(x) := Ax - b$$
   mit symmetrischen, positiv semidefiniten Matrizen $Q_0, Q_1, \ldots, Q_l$. Weiter setzen wir voraus, dass (P) zulässig ist. Man zeige:

   (a) Die Menge M_{opt} der Lösungen von (P) ist genau dann nichtleer und kompakt, wenn das System

   (∗) $$c_i^T p \leq 0, \quad Q_i p = 0 \qquad (i = 0, \ldots, l), \quad Ap = 0$$

   nur trivial lösbar ist.

(b) Die Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ zu (P) ist natürlich durch

$$L(x,u,v) := f(x) + g(x)^T u + h(x)^T v$$

gegeben. Das zu (P) duale Programm ist bekanntlich

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \phi(u,v) := \inf\limits_{x\in\mathbb{R}^n} L(x,u,v) \quad \text{auf} \\ N := \{(u,v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \phi(u,v) > -\infty\}. \end{cases}$$

Da eine konvexe quadratische Funktion genau dann auf dem $\mathbb{R}^n$ nach unten beschränkt ist, wenn ihr Gradient eine Nullstelle besitzt, ist die Menge der dual zulässigen Lösungen durch

$$N = \{(u,v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \exists\, z \in \mathbb{R}^n \ \text{ mit } \ \nabla_x L(z,u,v) = 0\}$$

gegeben. Weiter sei

$$N^0 := \{(u,v) \in N : u > 0\}.$$

Man zeige: Die Menge M_{opt} der Lösungen von (P) ist genau dann nichtleer und kompakt, wenn

$$(**) \qquad c_i^T p = 0, \quad Q_i p = 0 \qquad (i = 0,\ldots,l), \quad Ap = 0$$

nur trivial lösbar ist und N^0 nichtleer ist.

Kapitel 6

Lösungen zu den Aufgaben

6.1 Aufgaben in Kapitel 1

1. Gegeben sei die konvexe Optimierungsaufgabe

   (P) $$\text{Minimiere} \quad f(x) \quad \text{auf} \quad M,$$

   d. h. die Menge $M \subset \mathbb{R}^n$ der zulässigen Lösungen von (P) sei konvex, die Zielfunktion $f\colon M \longrightarrow \mathbb{R}$ sei auf M konvex. Man zeige:

   (a) Die Menge M_{opt} der (globalen) Lösungen von (P) ist konvex.

   (b) Ist $f\colon M \longrightarrow \mathbb{R}$ auf M sogar *strikt konvex*, gilt also die Implikation

   $$x, y \in M, \quad x \neq y, \quad \lambda \in (0,1) \Longrightarrow f((1-\lambda)x + \lambda y) < (1-\lambda)f(x) + \lambda f(y),$$

   so besteht die Menge M_{opt} der Lösungen von (P) aus höchstens einem Punkt.

   (c) Sei (P) zulässig (d. h. $M \neq \emptyset$), M abgeschlossen und f auf M stetig. Dann gilt:

   i. Existiert ein $x_0 \in M$ derart, dass die Niveaumenge $L_0 := \{x \in M : f(x) \leq f(x_0)\}$ kompakt ist, so ist M_{opt} nichtleer und kompakt.

   ii. Ist M_{opt} nichtleer und kompakt, so ist die Niveaumenge $L_0 := \{x \in M : f(x) \leq f(x_0)\}$ für jedes $x_0 \in M$ kompakt.

Lösung: Sind $x_1^*, x_2^* \in M_{\text{opt}}$ zwei Lösungen von (P), so ist natürlich $f(x_1^*) = f(x_2^*) = \min(\text{P})$ und $x_1^*, x_2^* \in M$. Mit einem vorgegebenem $\lambda \in [0,1]$ ist wegen der Konvexität von M auch $(1-\lambda)x_1^* + \lambda x_2^* \in M$. Wegen der Konvexität von f ist

$$f((1-\lambda)x_1^* + \lambda x_2^*) \leq (1-\lambda)f(x_1^*) + \lambda f(x_2^*) = \min(\text{P}).$$

Also ist auch $(1-\lambda)x_1^* + \lambda x_2^* \in M_{\text{opt}}$.

Gäbe es bei strikt konvexem f zwei verschiedene Lösungen $x_1^*, x_2^* \in M_{\text{opt}}$, so wäre einerseits wegen des schon bewiesenen Teils der Aufgabe auch $\frac{1}{2}(x_1^* + x_2^*) \in M_{\text{opt}}$, andererseits

$$f(\tfrac{1}{2}x_1^* + \tfrac{1}{2}x_2^*) < \tfrac{1}{2}f(x_1^*) + \tfrac{1}{2}f(x_2^*) = \min(\text{P}),$$

was natürlich einen Widerspruch bedeutet.

Nun sei M nichtleer und abgeschlossen, f auf M stetig. Mit einem $x_0 \in M$ sei die zugehörige Niveaumenge L_0 kompakt. Dann nimmt f auf L_0 das Minimum an, so dass $M_{\text{opt}} \neq \emptyset$. Da M_{opt} natürlich abgeschlossen ist, ist $M_{\text{opt}} \subset L_0$ kompakt.

Der letzte Teil der Aufgabe ist der einzige nicht ganz triviale Teil. Wir nehmen an, M_{opt} sei nichtleer und kompakt, ferner wird ein beliebiges $x_0 \in M$ gewählt und hiermit die Niveaumenge L_0 definiert. Wegen der Abgeschlossenheit von M und der Stetigkeit von f auf M ist diese natürlich abgeschlossen. Wir haben zu zeigen, dass sie auch beschränkt ist. Wäre dies nicht der Fall, so gibt es eine Folge $\{x_k\} \subset L_0$ mit $\|x_k\| \to \infty$ (hierbei ist $\|\cdot\|$ eine beliebige Norm auf dem $\mathbb{R}^n$). Aus der Folge $\{p_k\}$ mit $p_k := x_k/\|x_k\|$ ist eine gegen ein $p \in \mathbb{R}^n$ mit $\|p\| = 1$ konvergente Teilfolge auswählbar. O. B. d. A. gilt schon $p_k \to p$. Wir wollen zeigen, dass mit einem beliebigen $x^* \in M_{\text{opt}}$ der gesamte Strahl $\{x^* + tp : t \geq 0\}$ in M_{opt} liegt, was einen Widerspruch zur vorausgesetzten Kompaktheit (und damit Beschränktheit) von M_{opt} bedeutet. Sei $t > 0$ beliebig vorgegeben. Für alle hinreichend großen k ist $t/\|x_k\| \in (0, 1]$ und daher

$$\left(1 - \underbrace{\frac{t}{\|x_k\|}}_{\to 0}\right) x^* + t \underbrace{\frac{x_k}{\|x_k\|}}_{\to p} \in M$$

wegen der Konvexität von M, aus der Abgeschlossenheit von M folgt $x^* + tp \in M$. Weiter ist

$$\underbrace{f\left(\left(1 - \frac{t}{\|x_k\|}\right) x^* + \frac{t}{\|x_k\|} x_k\right)}_{\to f(x^*+tp)} \leq \underbrace{\left(1 - \frac{t}{\|x_k\|}\right) f(x^*)}_{\to f(x^*)} + \underbrace{\frac{t}{\|x_k\|}}_{\to 0} \underbrace{f(x_k)}_{\leq f(x_0)}$$

und damit $f(x^* + tp) \leq f(x^*)$ wegen der Konvexität und Stetigkeit von f auf M. Damit ist $x^* + tp \in M_{\text{opt}}$ bewiesen und der gewünschte Widerspruch erreicht.

2. Sei $M \subset \mathbb{R}^n$ konvex und $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf einer offenen Obermenge von M stetig differenzierbar. Man zeige:

   (a) f ist genau dann auf M konvex, wenn

   $$\nabla f(x)^T (y - x) \leq f(y) - f(x) \qquad \text{für alle } x, y \in M.$$

   (b) Ist f auf M konvex, so ist ein $x^* \in M$ genau dann eine Lösung der konvexen Optimierungsaufgabe, f auf M zu minimieren, wenn $\nabla f(x^*)^T (x - x^*) \geq 0$ für alle $x \in M$.

Lösung: Zum Beweis des ersten Teiles nehmen wir an, f sei auf M konvex. Mit vorgegebenen $x, y \in M$ und $t \in (0, 1]$ ist dann

$$f((1-t)x + ty) \leq (1-t) f(x) + t f(y)$$

und daher

$$\frac{f(x + t(y - x)) - f(x)}{t} \leq f(y) - f(x)$$

woraus mit $t \to 0+$ die Ungleichung $\nabla f(x)^T (y - x) \leq f(y) - f(x)$ folgt. Setzt man umgekehrt die Gültigkeit dieser Ungleichung für beliebige $x, y \in M$ voraus, gibt man

sich $t \in [0,1]$ vor und definiert $z := (1-t)x + ty$ mit vorgegebenen $x, y \in M$. Wegen der Konvexität von M ist $z \in M$. Aus

$$\begin{aligned} \nabla f(z)^T(x-z) &\leq f(x) - f(z), \\ \nabla f(z)^T(y-z) &\leq f(y) - f(z) \end{aligned}$$

erhält nach Multiplikation mit $(1-t)$ bzw. t und anschließender Addition, dass

$$0 \leq (1-t)f(x) + tf(y) - f((1-t)x + ty)$$

bzw. die Konvexität von f auf M.

Sei f auf M konvex und $0 \leq \nabla f(x^*)^T(x-x^*)$ für alle $x \in M$. Wegen des ersten Teils der Aufgabe ist dann $0 \leq \nabla f(x^*)^T(x-x^*) \leq f(x) - f(x^*)$, also x^* eine Lösung der Aufgabe, f auf M zu minimieren. Ist dies umgekehrt der Fall und $x \in M$ beliebig, so ist

$$\frac{f(x^* + t(x-x^*))}{t} \geq 0$$

für alle $t \in (0,1]$, mit $t \to 0+$ folgt die Behauptung[1].

3. Sei $M \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex, $z \in \mathbb{R}^n$ vorgegeben. Dann besitzt die Aufgabe

(P) $$\text{Minimiere} \quad \|x-z\|_2 \quad \text{auf} \quad M$$

genau eine Lösung x^*. Ferner ist ein $x^* \in M$ genau dann eine Lösung von (P), wenn $(x^*-z)^T(x-x^*) \geq 0$ für alle $x \in M$.

Hinweis: Es handelt sich hier um den *Projektionssatz für konvexe Mengen*. Die Existenz einer Lösung zeige man mit Hilfe der Kompaktheit von Niveaumengen, die Eindeutigkeit durch die strikte Konvexität von $f(x) := \frac{1}{2}\|x-z\|_2^2$, schließlich führe man die Charakterisierung einer Lösung auf eine Aussage in Aufgabe 2 zurück.

Lösung: Man wähle sich $x_0 \in M$ beliebig und bilde die Niveaumenge $L_0 := \{x \in M : \|x-z\|_2 \leq \|x_0 - z\|_2\}$. Diese ist der Durchschnitt der abgeschlossenen Menge M und einer abgeschlossenen Kugel, also kompakt, woraus die Existenz einer Lösung folgt. Für $x, y \in M$ und $\lambda \in (0,1)$ gilt für die im Hinweis angegebene Abbildung f nach leichter Rechnung

$$(1-\lambda)f(x) + \lambda f(y) - f((1-\lambda)x + \lambda y) = \frac{\lambda(1-\lambda)}{2}\|x-y\|_2^2,$$

woraus unmittelbar die strikte Konvexität von f folgt. Wegen $\nabla f(x^*)^T(x-x^*) = (x^*-z)^T(x-x^*)$ folgt der letzte Teil des Projektionssatzes aus dem zweiten Teil von Aufgabe 2.

4. Man betrachte die Optimierungsaufgabe

(P) $$\text{Minimiere} \quad f(x) := \sum_{j=1}^{n} x_j \ln \frac{x_j}{p_j} \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : e^T x = 1,\ x \geq 0\}.$$

[1] Für diesen Teil der Aufgabe haben wir die Konvexität von M, nicht aber die von f ausgenutzt.

Hierbei sei $e := (1,\dots,1)^T \in \mathbb{R}^n$, die positiven reellen Zahlen $p_1,\dots,p_n$ seien vorgegeben. Ferner ist natürlich $0\ln 0$ durch 0 definiert. Man zeige, dass (P) eine eindeutige Lösung x^* besitzt. Anschließend überlege man sich, dass $x^* > 0$ bzw. x^* nur positive Komponenten besitzt. Mit Hilfe der Lagrangeschen Multiplikatorenregel berechne man x^*.

Lösung: Mit einem beliebigen $x_0 \in M$ (etwa $x_0 := \frac{1}{n}e$) bilden wir die Niveaumenge $L_0 := \{x \in M : f(x) \le f(x_0)\}$. Da f auf M stetig und M kompakt ist, ist L_0 ebenfalls kompakt, woraus die Existenz eine Lösung von (P) folgt. Die Zielfunktion f ist strikt konvex. Um dies zu zeigen, genügt es offenbar die strikte Konvexität von $h(t) := t\ln t$ auf $[0,\infty)$ zu zeigen. Wegen $h''(t) = 1/t > 0$ für $t > 0$ ist h auf $(0,\infty)$ strikt konvex. Wegen $h(\lambda t) < \lambda h(t)$ für $\lambda \in (0,1)$ und $t > 0$ ist h sogar auf $[0,\infty)$ strikt konvex. Also besitzt (P) genau eine Lösung x^*. Angenommen, es wäre $x_j^* = 0$ für wenigstens ein $j \in \{1,\dots,n\}$. Wegen $e^T x^* = 1$ existiert ein $k \in \{1,\dots,n\}$ mit $x_k^* > 0$. Für alle hinreichend kleinen $t > 0$ ist dann $x^*(t) \in M$, wobei

$$x_i^*(t) := \begin{cases} t, & \text{falls} \quad i = j, \\ x_k^* - t, & \text{falls} \quad i = k, \\ x_i^*, & \text{sonst.} \end{cases}$$

Dann ist

$$\begin{aligned} f(x^*(t)) - f(x^*) &= t\ln\frac{t}{p_j} + (x_k^* - t)\ln\frac{x_k^* - t}{p_k} - x_k^*\ln\frac{x_k^*}{p_k} \\ &= t\ln\left(t\frac{p_k}{p_j}\right) + (x_k^* - t)\ln(x_k^* - t) - x_k^*\ln x_k^*. \end{aligned}$$

Mit $t \to 0+$ ist daher

$$\frac{f(x^*(t)) - f(x^*)}{t} = \ln\left(t\frac{p_k}{p_j}\right) + \frac{(x_k^* - t)\ln(x_k^* - t) - x_k^*\ln x_k^*}{t} \to -\infty + (-\ln x_k^* - 1).$$

Folglich ist $f(x^*(t)) < f(x^*)$ für alle hinreichend kleinen $t > 0$, ein Widerspruch dazu, dass x^* eine Lösung von (P). Also ist $x^* > 0$. Eine Anwendung der Lagrangeschen Multiplikatorenregel auf die Aufgabe

$$(\mathrm{P}_0) \qquad \text{Minimiere} \quad f(x) := \sum_{j=1}^n x_j \ln\frac{x_j}{p_j} \quad \text{auf} \quad M_0 := \{x \in \mathbb{R}^n : e^T x = 1,\ x > 0\}$$

liefert die Existenz eines Multiplikators $v^* \in \mathbb{R}$ mit $\nabla f(x^*) + v^* e = 0$ bzw.

$$\ln\frac{x_j^*}{p_j} + 1 + v^* = 0, \qquad j = 1,\dots,n.$$

Mit einer Konstanten c^* ist also $x_j^*/p_j = c^*$, $j = 1,\dots,n$. Aus der Nebenbedingung $e^T x^* = 1$ erhalten wir schließlich die Lösung

$$x_j^* = p_j \Big/ \sum_{i=1}^n p_i, \qquad j = 1,\dots,n.$$

5. Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch $f(x) := c^T x + \frac{1}{2} x^T Q x$ mit symmetrischem $Q \in \mathbb{R}^{n\times n}$ definiert. Dann ist $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$ genau dann, wenn Q positiv semidefinit ist und ein $x^* \in \mathbb{R}^n$ mit $\nabla f(x^*) = 0$ existiert.

   Lösung: Sei $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$. Angenommen, Q sei nicht positiv semidefinit. Dann existiert ein $x_0 \in \mathbb{R}^n$ mit $x_0^T Q x_0 < 0$. Für $t > 0$ definiere man $x_0(t) := t x_0$. Offensichtlich gilt dann $f(x_0(t)) \to -\infty$ mit $t \to \infty$, ein Widerspruch. Also ist Q positiv semidefinit. Angenommen, das lineare Gleichungssystem $\nabla f(x) = c + Qx = 0$ sei nicht lösbar. Dann ist

   $$-c \notin \text{Bild}\,(Q) = \text{Kern}\,(Q)^{\perp},$$

   d. h. es existiert ein $x_0 \in \text{Kern}\,(Q)$ mit $c^T x_0 \neq 0$, etwa $c^T x_0 < 0$. Dann ist $f(t x_0) = t c^T x_0 \to -\infty$ mit $t \to \infty$, ein Widerspruch.

   Ist umgekehrt Q positiv semidefinit und existiert ein $x^* \in \mathbb{R}^n$ mit $\nabla f(x^*) = 0$, so ist x^* unrestringiertes Minimum der konvexen Funktion f und daher erst recht $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$.

6. Gegeben sei das zweiseitig quadratisch restringierte quadratische Programm

   $$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := c_0^T x + \frac{1}{2} x^T Q_0 x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : \alpha_i \le g_i(x) := c_i^T x + \frac{1}{2} x^T Q_i x \le \beta_i,\ i = 1, \ldots, m\}. \end{array} \right.$$

   Hierbei seien $Q_0, Q_1, \ldots, Q_m \in \mathbb{R}^{n\times n}$ symmetrisch, $\alpha_i \le \beta_i$, $i = 1, \ldots, m$. Dann gilt:

   (a) Ist (P) zulässig und existieren $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ derart, dass $Q_0 + \sum_{i=1}^m \lambda_i Q_i$ positiv definit ist, so besitzt (P) eine Lösung.

   (b) Existiert zu $x^* \in M$ ein Vektor $\lambda^* = (\lambda_i^*) \in \mathbb{R}^m$ mit

   - $\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) = 0$,
   - $\lambda_i^*(\alpha_i - g_i(x^*)) \le 0 \le \lambda_i^*(g_i(x^*) - \beta_i)$, $i = 1, \ldots, m$,
   - $Q_0 + \sum_{i=1}^m \lambda_i^* Q_i$ ist positiv semidefinit,

   so ist x^* eine globale Lösung der (i. allg. nichtkonvexen) Optimierungsaufgabe (P). Ist $Q_0 + \sum_{i=1}^m \lambda_i^* Q_i$ sogar positiv definit, so ist x^* eindeutige Lösung von (P).

   Hinweis: Sie beweisen eine Verallgemeinerung eines Teils von Theorem 2.1 bei R. J. Stern, H. Wolkowicz (1995)[2].

   Lösung: Sei (P) zulässig, ferner möge reelle Zahlen $\lambda_1, \ldots, \lambda_m$ derart existieren, dass $Q_0 + \sum_{i=1}^m \lambda_i Q_i$ positiv definit ist. Wir zeigen die Existenz einer Lösung von(P) dadurch, dass wir mit beliebigem $x_0 \in M$ die Kompaktheit der Niveaumenge $L_0 := \{x \in M : f(x) \le f(x_0)\}$ nachweisen. Zu zeigen bleibt nur die Beschränktheit, was durch Widerspruch geschieht. Angenommen, es existiert eine Folge $\{x_k\} \subset L_0$ mit $\|x_k\| \to \infty$ (hierbei sei $\|\cdot\|$ irgendeine Norm auf dem $\mathbb{R}^n$). O. B. d. A. konvergiert die Folge $\{p_k\}$ mit $p_k := x_k/\|x_k\|$ gegen ein p, welches natürlich notwendigerweise vom Nullvektor verschieden ist. Aus $\{x_k\} \subset M$ bzw.

   $$\frac{\alpha_i}{\|x_k\|^2} \le c_i^T \frac{x_k}{\|x_k\|} \frac{1}{\|x_k\|} + \frac{1}{2} \frac{x_k^T}{\|x_k\|} Q \frac{x_k}{\|x_k\|} \le \frac{\beta_i}{\|x_k\|}, \qquad i = 1, \ldots, m,$$

[2] R. J. Stern, H. Wolkowicz (1995) Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. SIAM J. Optim. 5, 286–313.

erhält man mit $k \to \infty$, dass $p^T Q_i p = 0$, $i = 1, \ldots, m$. Entprechend folgt aus $f(x_k) \leq f(x_0)$, dass $p^T Q_0 p \leq 0$.Folglich ist

$$p^T \left(Q_0 + \sum_{i=1}^{m} \lambda_i Q_i \right) p = p^T Q_0 p \leq 0,$$

ein Widerspruch.

Nun wird vorausgesetzt, dass es zu $x^* \in M$ einen Vektor $\lambda^* \in \mathbb{R}^m$ mit den angegebenen Eigenschaften gibt. Sei $x \in M$ beliebig. Dann ist

$$\begin{aligned} f(x) - f(x^*) &= \nabla f(x^*)^T (x - x^*) + \frac{1}{2}(x - x^*)^T Q_0 (x - x^*) \\ &= -\sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*)^T (x - x^*) + \frac{1}{2}(x - x^*)^T \left(Q_0 + \sum_{i=1}^{m} \lambda_i^* Q_i \right)(x - x^*) \\ &\quad - \frac{1}{2} \sum_{i=1}^{m} \lambda_i^* (x - x^*)^T Q_i (x - x^*) \\ &\geq -\sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*)^T (x - x^*) - \frac{1}{2} \sum_{i=1}^{m} \lambda_i^* (x - x^*)^T Q_i (x - x^*) \\ &= \sum_{i=1}^{m} \lambda_i^* [g_i(x^*) - g_i(x)]. \end{aligned}$$

Wir zeigen, dass jeder der Summanden nichtnegativ ist. Bei vorgegebenem i machen wir eine Fallunterscheidung. Ist $\alpha_i < g_i(x^*) < \beta_i$, so ist notwendigerweise $\lambda_i^* = 0$, der Summand verschwindet also sogar. Ist $\alpha_i = g_i(x^*) < \beta_i$, so ist $\lambda_i^* \leq 0$ und daher

$$\lambda_i^* [g_i(x^*) - g_i(x)] = \underbrace{\lambda_i^*}_{\leq 0} \underbrace{[\alpha_i - g_i(x)]}_{\leq 0} \geq 0.$$

Der Fall $\alpha_i < g_i(x^*) = \beta_i$ verläuft entsprechend. Ist schließlich $\alpha_i = g_i(x^*) = \beta_i$, so ist $\lambda_i^* [g_i(x^*) - g_i(x)] = 0$. An der obigen Gleichungs-Ungleichungskette erkennt man ferner die Gültigkeit der behaupteten Eindeutigkeitsaussage.

7. Gegeben seien $c \in \mathbb{R}^n \setminus \{0\}$, die symmetrische, positiv definite Matrix $Q \in \mathbb{R}^{n \times n}$ sowie $x_0 \in \mathbb{R}^n$. Hiermit betrachte man die Optimierungsaufgabe

(P) Minimiere $c^T x$ auf $M := \{x \in \mathbb{R}^n : (x - x_0)^T Q (x - x_0) \leq 1\}$.

Man zeige, dass (P) eine eindeutige Lösung $x^* \in M$ besitzt und bestimme diese.

Lösung: Da M kompakt ist, besitzt (P) trivialerweise eine Lösung x^*, welche offenbar nicht im Innern von M liegen kann (andernfalls wäre $c = 0$). Mit

$$h(x) := (x - x_0)^T Q (x - x_0) - 1$$

reduziert sich (P) also auf die Aufgabe, $c^T x$ unter der Nebenbedingung $h(x) = 0$ zu minimieren. Wegen $\nabla h(x^*) = 2Q(x^* - x_0) \neq 0$ kann die Lagrangesche Multiplikatorenregel angewandt werden. Diese ergibt die Existenz eines Multiplikators $v^* \in \mathbb{R}$ mit $c + 2v^* Q(x^* - x_0) = 0$. Hieraus erhält man (es ist notwendig $v^* \neq 0$)

$$x^* = x_0 - \frac{1}{2v^*} Q^{-1} c.$$

Zur Berechnung des Multiplikators ziehe man nun noch die Nebenbedingung $h(x^*) = 0$ heran. Aus

$$0 = h(x^*) = \left(\frac{1}{2v^*}\right)^2 c^T Q^{-1} c - 1$$

erhält man

$$v^*_{1,2} = \pm\frac{1}{2}\sqrt{c^T Q^{-1} c}$$

und hiermit

$$x^*_{1,2} = x_0 \mp \frac{1}{\sqrt{c^T Q^{-1} c}}\, Q^{-1} c$$

als einzige Kandidaten für eine Lösung von (P). Hieraus liest man ab, dass

$$x^* := x_0 - \frac{1}{\sqrt{c^T Q^{-1} c}}\, Q^{-1} c$$

die eindeutige Lösung von (P) ist.

8. Beim Maximalflussproblem ist ein Netzwerk $(\mathcal{N}, \mathcal{A})$ mit zwei ausgezeichneten Knoten q (Quelle) und s (Senke) gegeben, ferner nichtnegative Kapazitäten u_{ij} auf den Pfeilen $(i,j) \in \mathcal{A}$. Ein *Fluss* $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ heißt *zulässig*, wenn er den Kapazitätsbeschränkungen, also

$$0 \le x_{ij} \le u_{ij}, \qquad (i,j) \in \mathcal{A},$$

und der Flussgleichung genügt. Diese besagt, dass in jedem Knoten außer der Quelle und Senke genau so viel Fluss ankommt wie auch wieder abtransportiert wird, also

$$\sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik} = 0, \qquad k \in \mathcal{N} \setminus \{q, s\},$$

gilt. Unter diesen Bedingungen ist der Fluss $\sum_{j:(q,j)\in\mathcal{A}} x_{qj}$ zu maximieren. Ein *Schnitt* im Netzwerk eine Partition der Knotenmenge $\mathcal{N}$ (bzw. $\{1, \ldots, m\}$) in zwei (disjunkte) Mengen $\mathcal{N}_1$ und $\mathcal{N}_2$ mit $q \in \mathcal{N}_1$ und $s \in \mathcal{N}_2$. Zu einem Schnitt $(\mathcal{N}_1, \mathcal{N}_2)$ definieren wir die zugehörige *Kapazität* $C(\mathcal{N}_1, \mathcal{N}_2)$ als die Summe aller Kapazitätsschranken über Pfeilen, die in $\mathcal{N}_1$ starten und in $\mathcal{N}_2$ enden, also in der oben eingeführten Notation durch

$$C(\mathcal{N}_1, \mathcal{N}_2) := \sum_{\substack{(i,j)\in\mathcal{A} \\ i\in\mathcal{N}_1,\, j\in\mathcal{N}_2}} u_{ij}.$$

Man zeige: Ist $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ ein zulässiger Fluss und $(\mathcal{N}_1, \mathcal{N}_2)$ ein Schnitt mit zugehöriger Kapazität $C(\mathcal{N}_1, \mathcal{N}_2)$, so ist

$$\sum_{j:(q,j)\in\mathcal{A}} x_{qj} \le C(\mathcal{N}_1, \mathcal{N}_2).$$

Besteht hier sogar Gleichheit, so ist x ein maximaler Fluss (und $(\mathcal{N}_1, \mathcal{N}_2)$ ein minimaler Schnitt). Mit dieser Aussage bestimme man in dem in der folgenden Abbildung angegebenen Netzwerk einen maximalen Fluss und einen minimalen Schnitt.

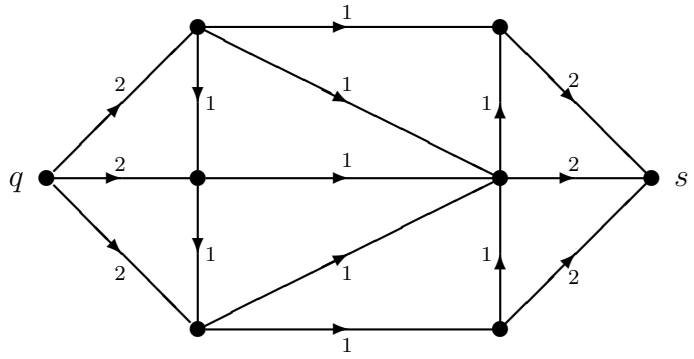

Abbildung 6.1: Maximaler Fluss, minimaler Schnitt?

Lösung: Es ist

$$\begin{aligned}
\sum_{j:(q,j)\in\mathcal{A}} x_{qj} &= \sum_{j:(q,j)\in\mathcal{A}} x_{qj} - \underbrace{\sum_{i:(i,q)\in\mathcal{A}} x_{iq}}_{=0} + \sum_{k\in\mathcal{N}_1\setminus\{q\}} \underbrace{\left(\sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik}\right)}_{=0} \\
&= \sum_{k\in\mathcal{N}_1}\left(\sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik}\right) \\
&= \sum_{k\in\mathcal{N}_1}\left(\sum_{j\in\mathcal{N}_2:(k,j)\in\mathcal{A}} \underbrace{x_{kj}}_{\le u_{kj}} - \underbrace{\sum_{i\in\mathcal{N}_2:(i,k)\in\mathcal{A}} x_{ik}}_{\ge 0}\right) \\
&\qquad + \underbrace{\sum_{k\in\mathcal{N}_1}\left(\sum_{j\in\mathcal{N}_1:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i\in\mathcal{N}_1:(i,k)\in\mathcal{A}} x_{ik}\right)}_{=0} \\
&\le \sum_{k\in\mathcal{N}_1}\sum_{j\in\mathcal{N}_2:(k,j)\in\mathcal{A}} u_{kj} \\
&= C(\mathcal{N}_1,\mathcal{N}_2).
\end{aligned}$$

Damit ist der erste Teil der Aufgabe gelöst. In Abbildung 6.2 geben wir einen Schnitt in dem gegebenen Netzwerk an. Die zu $\mathcal{N}_1$ gehörenden Knoten sind durch $\circ$, solche zu $\mathcal{N}_2$ durch $\bullet$ gekennzeichnet. Hier gibt es vier Pfeile, die Knoten aus $\mathcal{N}_1$ mit Knoten aus $\mathcal{N}_2$ verbinden, die zugehörige Kapazität ist 5.

9. Seien $a_1,\dots,a_m \in \mathbb{R}^n$ mit $\|a_i\|_2 = 1$, $i = 1,\dots,m$, und $b_1,\dots,b_m \in \mathbb{R}$ gegeben. Die Menge

$$P := \{x \in \mathbb{R}^n : a_i^T x \le b_i,\ (i = 1,\dots,m)\}$$

sei nichtleer und beschränkt. Man zeige: Ist $(x^*, r^*) \in \mathbb{R}^n \times \mathbb{R}$ eine Lösung der linearen Optimierungsaufgabe

$$\text{Maximiere} \quad r \quad \text{auf} \quad M := \{(x,r) \in \mathbb{R}^n \times \mathbb{R} : r \ge 0,\ a_i^T x + r \le b_i\ (i = 1,\dots,m)\},$$

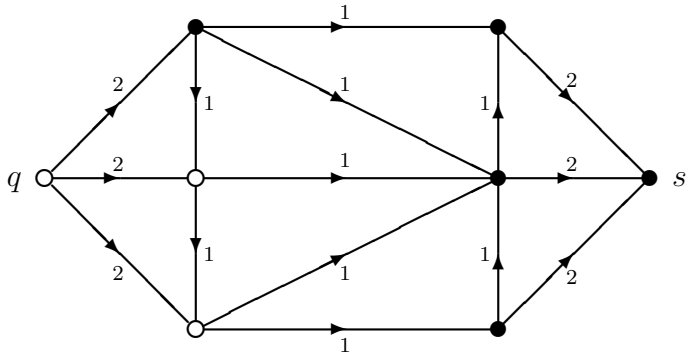

Abbildung 6.2: Ein Schnitt im Netzwerk mit Kapazität 5

so ist $B[x^*; r^*] := \{y \in \mathbb{R}^n : \|y - x^*\|_2 \le r^*\}$ die größte (euklidische) Kugel (d. h. die Kugel mit maximalem Radius), die in P enthalten ist. Also kann man die Inkugel zu einem Polytop (kompakter Polyeder) durch Lösen eines linearen Programms bestimmen.

Lösung: Wir zeigen: Ist $(x, r) \in M$, so ist die euklidische Kugel $B[x; r]$ um x mit Radius r in P enthalten. Denn sei $y \in B[x; r]$. Dann ist unter Benutzung der Cauchy-Schwarzschen Ungleichung

$$a_i^T y = a_i^T x + a_i^T (y - x) \le a_i^T x + \underbrace{\|a_i\|_2}_{=1} \underbrace{\|y - x\|_2}_{\le r} \le a_i^T x + r \le b_i, \qquad i = 1, \ldots, m,$$

also $y \in P$. Die Aussage ist bewiesen.

10. Gegeben seien m paarweise verschiedene Punkte $a_1, \ldots, a_m$ im $\mathbb{R}^n$, positive Gewichte $w_1, \ldots, w_m$ und eine nichtleere, konvexe und abgeschlossene Menge $M \subset \mathbb{R}^n$. Hiermit betrachte man das sogenannte *Fermat-Weber Problem*

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{i=1}^{m} w_i \, \|x - a_i\|_2 \quad \text{auf} \quad M,$$

wobei $\|\cdot\|_2$ natürlich die euklidische Norm auf dem $\mathbb{R}^n$ bedeutet. Man zeige:

(a) Die Optimierungsaufgabe (P) besitzt mindestens eine (globale) Lösung.

(b) Sind die gegebenen Punkte $a_1, \ldots, a_m$ nicht kollinear, liegen sie also nicht alle auf einer Geraden, so ist (P) sogar eindeutig lösbar.

Lösung: Die Existenz mindestens einer Lösung von (P) sieht man leicht ein, wenn man beachtet, dass mit einem $x_0 \in M$ die Niveaumenge $L_0 := M \cap \{x \in \mathbb{R}^n : f(x) \le f(x_0)\}$ kompakt ist. Seien nun $a_1, \ldots, a_m$ nicht kollinear und $x_1, x_2 \in M$ zwei verschiedene Lösungen von (P). Da (P) ein kovexes Programm ist, ist nach Aufgabe 1 auch $\frac{1}{2}(x_1 + x_2)$ eine Lösung von (P). Daher ist

$$\begin{aligned} 0 &= f\left(\frac{x_1 + x_2}{2}\right) - \frac{1}{2}\left[f(x_1) + f(x_2)\right] \\ &= \frac{1}{2} \sum_{i=1}^{m} w_i \, [\underbrace{\|(x_1 - a_i) + (x_2 - a_i)\|_2 - (\|x_1 - a_i\|_2 + \|x_2 - a_i\|_2)}_{\le 0}] \end{aligned}$$

und folglich

$$\|(x_1 - a_i) + (x_2 - a_i)\|_2 = \|x_1 - a_i\|_2 + \|x_2 - a_i\|_2, \qquad i = 1, \ldots, m.$$

Für $i = 1, \ldots, m$ existiert daher (Gleichheit in der Dreiecksungleichung bzw. der Cauchy-Schwarzschen Ungleichung) ein $\lambda_i > 0$ mit

$$x_1 - a_i = \lambda_i(x_2 - a_i), \qquad i = 1, \ldots, m,$$

wobei $\lambda_i \neq 1$ ist, da $x_1 \neq x_2$ angenommen wurde. Folglich ist

$$a_i - a_j = \frac{\lambda_i - \lambda_j}{(1 - \lambda_i)(1 - \lambda_j)}\,(x_1 - x_2), \qquad i, j = 1, \ldots, m.$$

Im Widerspruch zur Voraussetzung liegen also die gegebenen Punkte $a_1, \ldots, a_m$ sämtlich auf einer Geraden, die Eindeutigkeit ist bewiesen.

11. Man löse das folgende, auf S. Lhulier (1782) zurückgehende geometrische Problem: Die Längen a_1 bzw. a_2 der Grundlinien zweier Dreiecke sowie die Summe l der Längen ihrer vier Schenkel seien gegeben, wobei natürlich $l > a_1 + a_2$ vorausgesetzt sei. Unter allen Paaren von Dreiecken mit diesen Eigenschaften bestimme man dasjenige, für welches die Summe der Flächeninhalte der beiden Dreiecke maximal ist. Für $a_1 = 1$, $a_2 = 2$ und $l = 5$ berechne man numerisch die Länge der gesuchten Schenkel.

Lösung: Nach der Formel von Heron ist der Flächeninhalt Δ eines Dreiecks mit den Seitenlängen a, b, c durch

$$\Delta = [s(s-a)(s-b)(s-c)]^{1/2} \qquad \text{mit} \qquad s := \frac{1}{2}(a+b+c)$$

gegeben. Die Längen der gesuchten Schenkel seien b_1, c_1 bzw. b_2, c_2. Die optimalen Dreiecke müssen natürlich gleichschenklig sein (Beweis?). Der Flächeninhalt eines gleichschenkligen Dreiecks (a sei die Länge der Grundlinie, b die der beiden Schenkel, ist durch

$$\Delta = \frac{1}{2}a\sqrt{b^2 - \frac{a^2}{4}}$$

gegeben. Zu lösen ist also die Aufgabe,

$$\Delta(b_1, b_2) := \frac{1}{2}a_1\sqrt{b_1^2 - \frac{a_1^2}{4}} + \frac{1}{2}a_2\sqrt{b_2^2 - \frac{a_2^2}{4}}$$

unter der Nebenbedingung

$$b_1 + b_2 = \frac{l}{2} =: \hat{l}$$

(und $b_1 > 0$, $b_2 > 0$) zu maximieren. Ist (b_1^*, b_2^*) eine Lösung, so ist wegen der Lagrangeschen Multiplikatorenregel

$$\frac{a_1 b_1^*}{\sqrt{(b_1^*)^2 - a_1^2/4}} = \frac{a_2 b_2^*}{\sqrt{(b_2^*)^2 - a_2^2/4}}.$$

Einsetzen von $b_2^* = \hat{l} - b_1^*$ liefert für b_1^* die Gleichung

$$\frac{a_1 b_1^*}{\sqrt{(b_1^*)^2 - a_1^2/4}} = \frac{a_2(\hat{l} - b_1^*)}{\sqrt{(\hat{l} - b_1^*)^2 - a_2^2/4}}.$$

Daher ist b_1^* als Nullstelle von

$$f(b_1) := \frac{1}{(\hat{l} - b_1)^2} - \frac{1}{b_1^2} - 4\left[\frac{1}{a_2^2} - \frac{1}{a_1^2}\right]$$

in $(0, \hat{l})$ zu bestimmen. Da f auf $(0, \hat{l})$ monoton wachsend ist und

$$\lim_{b_1 \to 0+} f(b_1) = -\infty, \qquad \lim_{b_1 \to \hat{l}-} f(b_1) = +\infty$$

gilt, existiert b_1^* eindeutig. Für $a_1 = 1$, $a_2 = 2$ und $l = 2.5$ erhalten wir für die Schenkellänge des ersten Dreiecks $b_1 = 0.553515$, für die des zweiten $b_2 = 0.696485$.

6.2 Aufgaben in Kapitel 2

6.2.1 Aufgaben in Abschnitt 2.1

1. Sei $K \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex, ferner $P_K \colon \mathbb{R}^n \longrightarrow K \subset \mathbb{R}^n$ die zugehörige Projektionsabbildung. Man zeige:

   (a) Es ist
   $$\|P_K(x) - P_K(y)\| \leq \|x - y\| \qquad \text{für alle } x, y \in \mathbb{R}^n.$$

   (b) Ist $L \subset \mathbb{R}^n$ ein linearer Teilraum, so ist P_L eine lineare Abbildung und $x^T P_L(y) = P_L(x)^T y$ für alle $x, y \in \mathbb{R}^n$.

   (c) Ist $L := \operatorname{span}\{v_1, \ldots, v_p\}$ mit linear unabhängigen $v_1, \ldots, v_p \in \mathbb{R}^n$ und $V := (\, v_1 \ \cdots \ v_p \,)$, so ist
   $$P_L(x) = V(V^T V)^{-1} V^T x \qquad \text{für alle } x \in \mathbb{R}^n.$$

   Lösung: Eine Anwendung der notwendigen und hinreichenden Optimalitätsbedingungen des Projektionssatzes liefert die Gültigkeit von

   $$[x - P_K(x)]^T [P_K(y) - P_K(x)] \leq 0, \qquad [y - P_K(y)]^T [P_K(x) - P_K(y)] \leq 0.$$

   Eine Addition dieser beiden Ungleichungen liefert

   $$[P_K(x) - P_K(y) - (x - y)]^T [P_K(x) - P_K(y)] \leq 0$$

   bzw. mit der Cauchy-Schwarzschen Ungleichung

   $$\|P_K(x) - P_K(y)\|^2 \leq (x - y)^T [P_K(x) - P_K(y)] \leq \|x - y\| \, \|P_K(x) - P_K(y)\|,$$

   woraus die erste Behauptung folgt.

   Bei gegebenem $z \in \mathbb{R}^n$ ist $P_L(z) \in L$ charakterisiert durch $[z - P_L(z)]^T x = 0$ für alle $x \in L$. Ist daher $[z_1 - P_L(z_1)]^T x = 0$ und $[z_2 - P_L(z_2)]^T x = 0$ jeweis für alle $x \in L$, so erhält man durch Multiplikation mit α_1 und α_2 sowie anschließender Addition, dass

   $$[(\alpha_1 z_1 + \alpha_2 z_2) - \underbrace{(\alpha_1 P_L(z_1) + \alpha_2 P_L(z_2))}_{\in L}]^T x = 0 \qquad \text{für alle } x \in L$$

und hiermit

$$\alpha_1 P_L(z_1) + \alpha_2 P_L(z_2) = P_L(\alpha_1 z_1 + \alpha_2 z_2).$$

Für beliebige $x, y \in \mathbb{R}^n$ ist

$$[x - P_L(x)]^T P_L(y) = 0, \qquad [y - P_L(y)]^T P_L(x).$$

Daher ist

$$x^T P_L(y) = P_L(x)^T P_L(y) = P_L(x)^T y.$$

Die Matrix V, deren Spalten gerade die Basiselemente des linearen Teilraumes L bilden besitzt vollen Rang, daher ist $V^T V$ nichtsingulär. Zu zeigen ist

$$[z - V(V^T V)^{-1} V^T z]^T x = 0 \qquad \text{für alle } x \in L.$$

Ein beliebiges $x \in L$ besitzt die eindeutige Darstellung $x = Vy$, Einsetzen liefert sofort die Behauptung.

2. Seien $l, u \in \mathbb{R}^n$ zwei Vektoren mit $l \leq u$. Hiermit definiere man den Quader

$$Q := \{x \in \mathbb{R}^n : l \leq x \leq u\}.$$

Man zeige, dass für $x \in \mathbb{R}^n$ die Projektion $P_Q(x)$ von x auf Q durch

$$P_Q(x)_j = \begin{cases} l_j, & \text{falls} \quad x_j < l_j, \\ x_j, & \text{falls} \quad l_j \leq x_j \leq u_j, \\ u_j, & \text{falls} \quad u_j < x_j, \end{cases} \qquad j = 1, \ldots, n,$$

gegeben ist.

Lösung: Für $x \in \mathbb{R}^n$ ist $P_Q(x) \in Q$. Wegen des Projektionssatzes bleibt die charakteriserende Eigenschaft

$$(P_Q(x) - x)^T (z - P_Q(x)) \geq 0 \qquad \text{für alle } z \in Q$$

nachzuprüfen. Wir definieren die Indexmengen

$$\begin{aligned} J_- &:= \{j \in \{1, \ldots, n\} : x_j < u_j\}, \\ J_0 &:= \{j \in \{1, \ldots, n\} : l_j \leq x_j \leq u_j\}, \\ J_+ &:= \{j \in \{1, \ldots, n\} : u_j < x_j\}. \end{aligned}$$

Für ein gegebenes $z \in Q$ ist dann

$$\begin{aligned} (P_Q(x) - x)^T (z - P_Q(x)) &= \sum_{j=1}^{n} (P_Q(x)_j - x_j)(z_j - P_Q(x)_j) \\ &= \sum_{j \in J_-} \underbrace{(l_j - x_j)}_{>0} \underbrace{(z_j - l_j)}_{\geq 0} + \sum_{j \in J_+} \underbrace{(u_j - x_j)}_{<0} \underbrace{(z_j - u_j)}_{\leq 0} \\ &\geq 0. \end{aligned}$$

Damit ist die Behauptung nachgewiesen.

3. Zwei nichtleere, konvexe Mengen $A, B \subset \mathbb{R}^n$ sind genau dann stark trennbar, wenn $0 \notin \mathrm{cl}\,(B - A)$.

   Lösung: Mit A und B ist auch $B - A$ und dann auch $\mathrm{cl}\,(B - A)$ konvex. Aus dem Korollar 1.4 zum starken Trennungssatz folgt die Existenz eines $y \in \mathbb{R}^n \setminus \{0\}$ mit $0 < \inf_{x \in \mathrm{cl}\,(B-A)} y^T x$. Also existiert ein $\gamma > 0$ mit $0 < \gamma \leq y^T b - y^T a$, $a \in A$, $b \in B$, woraus $\sup_{a \in A} y^T a < \inf_{b \in B} y^T b$ und damit die starke Trennbarkeit von A und B folgt.

   Umgekehrt seien A und B stark trennbar, es existiere also ein $y \in \mathbb{R}^n \setminus \{0\}$ mit $\sup_{a \in A} y^T a < \inf_{b \in B} y^T b$. Wäre $0 \in \mathrm{cl}\,(B - A)$, so existierten Folge $\{a_k\} \subset A$ und $\{b_k\} \subset B$ mit $b_k - a_k \to 0$ und damit $y^T b_k - y^T a_k \to 0$. Andererseits ist

   $$y^T a_k \leq \sup_{a \in A} y^T a < \inf_{b \in B} y^T b \leq y^T b_k,$$

   offensichtlich ein Widerspruch.

4. Sei $C \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex mit nichtleerem Inneren $\mathrm{int}\,(C)$. Man zeige, dass es zu jedem $x^* \in C \setminus \mathrm{int}\,(C)$ ein $y \in \mathbb{R}^n \setminus \{0\}$ mit

   $$C \subset \{x \in \mathbb{R}^n : y^T x \geq y^T x^*\}$$

   gibt.

   Hinweis: Man zeige, dass mit C auch $\mathrm{int}\,(C)$ konvex ist und wende auf $\{x^*\}$ und $\mathrm{int}\,(C)$ den Trennungssatz an. Anschließend zeige man, dass $C = \mathrm{cl}\,(\mathrm{int}\,(C))$.

   Lösung: Wie im Hinweis angegeben, zeigen wir zunächst, dass mit C auch $\mathrm{int}\,(C)$ konvex ist. Seien hierzu $x_1, x_2 \in \mathrm{int}\,(C)$ sowie $\lambda \in (0, 1)$. Wir zeigen, dass auch $(1 - \lambda)x_1 + \lambda x_2 \in \mathrm{int}\,(C)$. Da $x_1 \in \mathrm{int}\,(C)$, existiert ein $\epsilon_1 > 0$ derart, dass die euklidische Kugel um x_1 mit dem Radius ϵ_1 noch ganz in C enthalten ist. Wir wollen zeigen, dass die Kugel um $(1 - \lambda)x_1 + \lambda x_2$ mit dem Radius $(1 - \lambda)\epsilon_1$ in C enthalten ist. Sei hierzu $\|(1-\lambda)x_1+\lambda x_2-x\| \leq (1-\lambda)\epsilon_1$, zu zeigen ist $x \in C$. Man definiere $z := (x-\lambda x_2)/(1-\lambda)$, was $x = (1 - \lambda)z + \lambda x_2$ impliziert. Wir zeigen, dass z in der ϵ_1-Kugel um x_1 und damit in C liegt, was wegen der Konvexität von C auch $x \in C$ impliziert. Denn es ist

   $$\|x_1 - z\| = \left\| x_1 - \frac{1}{1-\lambda}(x - \lambda x_2) \right\| = \frac{1}{1-\lambda} \underbrace{\|(1-\lambda)x_1 + \lambda x_2 - x\|}_{\leq (1-\lambda)\epsilon_1} \leq \epsilon_1.$$

   Die disjunkten, konvexen Mengen $\{x^*\}$ und $\mathrm{int}\,(C)$ lassen sich nach dem Trennungssatz 1.10 trennen, es existiert also ein $y \in \mathbb{R}^n \setminus \{0\}$ mit $y^T x^* \leq y^T x$ für alle $x \in \mathrm{int}\,(C)$ und damit auch für alle $x \in \mathrm{cl}\,(\mathrm{int}\,(C))$. Nun ist jedes $x \in C$ Limes einer Folge aus $\mathrm{int}\,(C)$, wie z. B. sofort aus dem Beweis des ersten Teiles der Aufgabe folgt: Ist $x_1 \in \mathrm{int}\,(C)$ beliebig, ferner $\{\lambda_k\} \subset (0, 1)$ eine beliebige Folge mit $\lambda_k \to 1$, so ist $x_k := (1 - \lambda_k)x_1 + \lambda_k x \in \mathrm{int}\,(C)$ und $x_k \to x$, also $x \in \mathrm{cl}\,(\mathrm{int}\,(C))$.

5. Eine nichtleere, abgeschlossene, konvexe Menge $C \subset \mathbb{R}^n$ ist der Durchschnitt aller abgeschlossenen Halbräume, die C enthalten.

   Hinweis: Man wende den starken Trennungssatz an.

   Lösung: Sei K der Durchnitt aller abgeschlossenen Halbräume, die C enthalten. Dann ist $C \subset K$. Angenommen, es existiert ein $x^* \in K \setminus C$. Wegen des Korollars 1.4 zum starken Trennungssatzes existiert ein $y \in \mathbb{R}^n \setminus \{0\}$ mit $y^T x^* < \gamma := \inf_{x \in C} y^T x$. Dann ist $H := \{x \in \mathbb{R}^n : y^T x = \gamma\}$ eine Hyperebene, die C im zugehörigen Halbraum enthält. Dieser Halbraum enthält nicht $x^* \in K$, was ein Widerspruch zur Definition von K ist.

6. Sei $C \subset \mathbb{R}^n$ ein nichtleerer, abgeschlossener, konvexer Kegel. Dann ist $(C^+)^+ = C$. Eine (dumme) Zusatzfrage: Kann Gleichheit auch gelten, wenn C nicht abgeschlossen, nicht konvex oder kein Kegel ist?

   Hinweis: Man überzeuge sich davon, dass die Inklusion $C \subset (C^+)^+$ trivial ist. Mit Hilfe des starken Trennungssatzes zeige man anschließend, dass aus $z \notin C$ auch $z \notin (C^+)^+$ folgt.

   Lösung: Wie im Hinweis angegeben, zeigen wir zunächst die Gültigkeit von $C \subset (C^+)^+$. Seien hierzu $z \in C$ und $x \in C^+$ beliebig. Dann ist $x^T z \geq 0$ und daher $x \in (C^+)^+$. Ist $z \notin C$, so lassen sich $\{z\|$ und C stark trennen. Es existiert also ein $y \in \mathbb{R}^n \setminus \{0\}$ mit $y^T z < \inf_{x \in C} y^T x$. Da C ein Kegel ist, ist $y^T x \geq 0$ für alle $x \in C$ bzw. $y \in C^+$ (Beweis?). Dann ist also $y^T z < 0 \leq y^T x$ für alle $x \in C$, also $z \notin (C^+)^+$. Da für ein beliebiges $K \subset \mathbb{R}^n$ die Menge K^+ stets ein abgeschlossener, konvexer Kegel ist, kann Gleichheit nicht gelten, wenn eine dieser Eigenschaften nicht vorhanden ist.

7. Man zeige, dass jeder endlich erzeugte Kegel sich als dualer Kegel eines polyedrischen Kegels darstellen lässt. Genauer zeige man: Ist $U \in \mathbb{R}^{n \times m}$, so ist

$$\{Uy : y \geq 0\} = \{x \in \mathbb{R}^n : U^T x \geq 0\}^+.$$

   Lösung: Mit Hilfe des Farkas-Lemmas erhält man

$$\begin{aligned} \{x \in \mathbb{R}^n : U^T x \geq 0\}^+ &= \{v \in \mathbb{R}^n : v^T x \geq 0 \text{ für alle } x \in \mathbb{R}^n \text{ mit } U^T x \geq 0\} \\ &\qquad \text{(Definition des dualen Kegels)} \\ &= \{v \in \mathbb{R}^n : U^T x \geq 0,\ v^T x < 0 \text{ unlösbar}\} \\ &= \{v = Uy : y \geq 0\}. \end{aligned}$$

   Damit ist die Behauptung bewiesen.

8. Sei $A \in \mathbb{R}^{m \times n}$. Man beweise den Alternativsatz von Gordan: Genau eine der beiden Aussagen

   (I) $\qquad Ax = 0, \quad x \geq 0, \quad x \neq 0 \qquad$ hat eine Lösung $x \in \mathbb{R}^n$

   bzw.

   (II) $\qquad A^T y > 0 \qquad$ hat eine Lösung $y \in \mathbb{R}^m$

   ist richtig.

   Lösung: (I) und (II) sind nicht gleichzeitig lösbar, wie man leicht nachweist. Angenommen, (I) sei nicht lösbar. Mit $e := (1, \ldots, 1)^T \in \mathbb{R}^n$ ist auch

$$\begin{pmatrix} A \\ e^T \end{pmatrix} x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad x \geq 0$$

   nicht lösbar. Aus dem Farkas-Lemma folgt die Lösbarkeit von

$$\begin{pmatrix} A^T & e \end{pmatrix} \begin{pmatrix} y \\ \delta \end{pmatrix} \geq 0, \qquad \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} y \\ \delta \end{pmatrix} < 0.$$

   Also ist $\delta < 0$ und folglich

$$A^T y \geq -\delta e > 0,$$

   also (II) lösbar.

9. Sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Man beweise den Alternativsatz von Gale: Genau eine der beiden Aussagen

   (I) $\qquad Ax \le b \qquad$ hat eine Lösung $x \in \mathbb{R}^n$

   bzw.

   (II) $\qquad A^T y = 0, \quad y \ge 0, \quad b^T y < 0 \qquad$ hat eine Lösung $y \in \mathbb{R}^m$

   ist richtig.

   Lösung: Es kann entweder so argumentiert werden wie im Anschluss an das Farkas-Lemma oder indem man im verallgemeinerten Farkas-Lemma 1.8 als K den nichtnegativen Orthanten im $\mathbb{R}^m$ und als C den gesamten $\mathbb{R}^n$ nimmt.

10. Von A. Dax (1997)[3] stammt ein "elementarer" Beweis des Farkas-Lemmas. Wir wollen die Quintessenz dieses Arguments wiedergeben. Gegeben seien also wieder $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und hiermit die Systeme

    (I) $\qquad Ax = b, \qquad x \ge 0$

    und

    (II) $\qquad A^T y \ge 0, \qquad b^T y < 0.$

    Man zeige der Reihe nach:

    (a) Die Optimierungsaufgabe

    (P) $\qquad$ Minimiere $\quad f(x) := \dfrac{1}{2}\|Ax - b\|_2^2, \quad x \ge 0$

    besitzt eine Lösung x^*.

    (b) Ist (I) nicht lösbar bzw. $y^* := Ax^* - b \ne 0$, so ist y^* eine Lösung von (II).

    Lösung: Die Lösbarkeit des vorzeichenbeschränkten linearen Ausgleichsproblems (P) hatten wir uns schon in Kapitel 1 überlegt, wobei wir dort die Abgeschlossenheit endlich erzeugter konvexer Kegel benutzt hatten, was inzwischen durch Lemma 1.5 bewiesen wurde. Wir nehmen an, (I) sei nicht lösbar und daher $y^* := Ax^* - b \ne 0$. Als Lösung der konvexen Optimierungsaufgabe (P) ist $x^* \ge 0$ charakterisiert (siehe Aufgabe 2 in Kapitel 1) durch

    $(*)$ $\qquad 0 \le \nabla f(x^*)^T(x - x^*) = (A^T y^*)^T(x - x^*) \qquad$ für alle $x \ge 0.$

    Hieraus folgt

    $$(A^T y^*)_i \begin{cases} = 0, & \text{falls} \quad x_i^* > 0, \\ \ge 0, & \text{falls} \quad x_i^* = 0. \end{cases}$$

    Daher ist $A^T y^* \ge 0$. Setzt man $x := 0$ in $(*)$, so erhält man

    $$0 \le (A^T y^*)^T(-x^*) = -(y^*)^T A x^* = -\|y^*\|_2^2 - b^T y^*,$$

    so dass $b^T y^* \le -\|y^*\|_2^2 < 0$ und daher y^* eine Lösung von (II) ist.

[3] A. Dax (1997) An elementary proof of Farkas' Lemma. SIAM Rev. 39, 503–507.

11. Man beweise den folgenden Satz von Fan-Glicksburg-Hoffman (siehe O. L. Mangasarian (1969, S. 63) und R. T. Rockafellar (1972, S. 186 ff.)):

    Sei $C \subset \mathbb{R}^n$ nichtleer und konvex, die Abbildung $g\colon C \longrightarrow \mathbb{R}^l$ (komponentenweise) konvex, die Abbildung $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear. Besitzt dann

    (I) $$x \in C, \quad g(x) < 0, \quad h(x) = 0$$

    keine Lösung, so besitzt

    (II) $$(u, v) \in \mathbb{R}^l \times \mathbb{R}^m \setminus \{(0,0)\}, \quad u \geq 0, \quad \inf_{x \in C}[u^T g(x) + v^T h(x)] \geq 0$$

    eine Lösung.

    Hinweis: Besitzt (I) keine Lösung, so ist

    $$(0,0) \notin \{(g(x) + z, h(x)) \in \mathbb{R}^l \times \mathbb{R}^m : x \in C,\ z > 0\}.$$

    Man überzeuge sich davon, dass die rechtsstehende Menge konvex ist und wende den Trennungsatz für konvexe Mengen an.

    Lösung: Zur Abkürzung setzen wir

    $$K := \{(g(x) + z, h(x)) \in \mathbb{R}^l \times \mathbb{R}^m : x \in C,\ z > 0\}.$$

    Die Konvexität von K ist leicht einzusehen, wir übergehen den einfachen Beweis. Die Unlösbarkeit von (I) besagt gerade, dass $(0,0) \notin K$. Wegen des Trennungssatzes 1.10 existiert $(u, v) \in \mathbb{R}^l \times \mathbb{R}^m \setminus \{(0,0)\}$ mit

    $$0 \leq u^T(g(x) + z) + v^T h(x) \qquad \text{für alle } x \in C,\ z > 0.$$

    Hält man hier x fest, so folgt, dass $u^T z$ für alle $z > 0$ durch eine Konstante nach unten beschränkt ist, was $u \geq 0$ impliziert. Offensichtlich folgt hieraus die Behauptung.

12. Man beweise die folgende Variante zum Satz von Fan-Glicksburg-Hoffman (siehe O. L. Mangasarian (1969, S. 65)):

    Sei $C \subset \mathbb{R}^n$ nichtleer und konvex, die Abbildung $g\colon C \longrightarrow \mathbb{R}^l$ (komponentenweise) konvex. Dann ist genau eine der Aussagen

    (I) $$\text{Es existiert } x \in C \text{ mit } g(x) < 0$$

    bzw.

    (II) $$\text{Es existiert } u \in \mathbb{R}^l \setminus \{0\} \text{ mit } u \geq 0 \text{ und } \inf_{x \in C} u^T g(x) \geq 0$$

    richtig.

    Lösung: Angenommen, (I) und (II) seien gleichzeitig durch x bzw. u lösbar. Dann ist

    $$0 > u^T g(x) \inf_{z \in C} u^T g(z) \geq 0,$$

    ein Widerspruch. Nun nehmen wir an, (I) sei nicht lösbar. Dann ist $0 \notin K := \{g(x) + z : x \in C,\ z > 0\}$. Eine Anwendung des Trennungssatzes liefert ein $u \in \mathbb{R}^l \setminus \{0\}$ mit $0 \leq u^T(g(x) + z)$ für alle $x \in C$, $z > 0$. Hieraus folgt offenbar wieder die Behauptung.

13. Man beweise: Ist $A \subset \mathbb{R}^n$ nichtleer und konvex, so ist $\mathrm{ri}\,(A) \neq \emptyset$ (siehe z. B. J.-B. Hirriart-Urruty, C. Lemaréchal (1993, S. 103) oder auch R. T. Rockafellar (1972, Theorem 6.2)).

Lösung: Das relative Innere von A wurde definiert als

$$\mathrm{ri}\,(A) := \{x \in A : \text{Es existiert } \epsilon > 0 \text{ mit } B[x;\epsilon] \cap \mathrm{aff}\,(A) \subset A\}.$$

Hierbei bezeichnet $\mathrm{aff}\,(A)$ die affine Hülle von A und $B[x;\epsilon]$ die (z. B. euklidische) Kugel um x mit dem Radius ϵ. O. B. d. A. sei $0 \in A$ (andernfalls verschiebe man A), ferner sei 0 nicht der einzige Punkt von A (andernfalls ist $A = \mathrm{ri}(A) = \{0\}$). Dann ist $\mathrm{aff}\,(A) = \mathrm{span}\,(A)$ ein (nichttrivialer) linearer Teilraum, er sei aufgespannt von den linear unabhängigen $\{a_1, \ldots, a_m\} \subset A$, $1 \leq m \leq n$. Wegen der Konvexität von A (und $0 \in A$) ist

$$a := \frac{1}{m+1} \sum_{i=1}^{m} a_i \in A.$$

Wir wollen zeigen, dass a im relativen Inneren von A liegt. Hierzu setzen wir $\eta := 1/[m(m+1)]$ und zeigen

$$|\alpha_i| \leq \eta \quad (i = 1, \ldots, m) \Longrightarrow a + \sum_{i=1}^{m} \alpha_i a_i \in A.$$

Denn ist $|\alpha_i| \leq \eta$, $i = 1, \ldots, m$, so ist

$$a + \sum_{i=1}^{m} \alpha_i a_i = \sum_{i=1}^{m} \left(\frac{1}{m+1} + \alpha_i \right) a_i \in A,$$

da $0 \in A$, $1/(m+1) + \alpha_i \geq 0$, $i = 1, \ldots, m$, und

$$\sum_{i=1}^{m} \left(\frac{1}{m+1} + \alpha_i \right) \leq \frac{m}{m+1} + \frac{1}{m+1} = 1.$$

Die Behauptung folgt dann wegen der Äquivalenz von Normen auf dem endlichdimensionalen Raum $\mathrm{span}\,(A)$.

6.2.2 Aufgaben in Abschnitt 2.2

1. Gegeben sei das lineare Programm

(P) $\qquad$ Minimiere $\quad c^T x \quad$ auf $\quad M := \{x : x \geq 0,\ b - Ax \leq 0\}$.

Man stelle das zu (P) duale lineare Programm auf.

Lösung: Die Lagrange-Funktion ist

$$L(x, y) := c^T x + y^T (b - Ax) = b^T y + (c - A^T y)^T x,$$

für ein $y \geq 0$ ist der Wert der dualen Zielfunktion daher

$$\phi(y) := \inf_{x \geq 0} L(x, y) = \begin{cases} b^T y, & \text{falls} \quad c - A^T y \geq 0, \\ -\infty, & \text{sonst.} \end{cases}$$

Das zu (P) duale lineare Programm ist daher

(D) $\qquad$ Maximiere $\quad b^T y \quad$ auf $\quad N := \{y : y \geq 0,\ c - A^T y \geq 0\}$.

2. Gegeben sei das lineare Programm

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Gx \leq h,\ Ax = b\},$$

wobei l Ungleichungen und m Gleichungen auftreten. Man stelle das zu (P) duale lineare Programm auf.

Lösung: Die zu (P) gehörige Lagrange-Funktion ist

$$L(x,u,v) := c^T x + u^T(Gx-h) + v^T(Ax-b) = -h^T u - b^T v + (c - G^T u - A^T v)^T x.$$

Für ein Paar $(u,v) \in \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ ist daher

$$\phi(u,v) = \inf_{x \in \mathbb{R}^n} L(x,u,v) = \begin{cases} -h^T u - b^T v, & \text{falls} \quad c - G^T u - A^T v = 0, \\ -\infty, & \text{sonst.} \end{cases}$$

Das zu (P) duale Programm ist daher

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad -(h^T u + b^T v) \quad \text{auf} \\ N := \{(u,v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ G^T u + A^T v = c\}. \end{cases}$$

3. Seien $a_1, \ldots, a_l \in \mathbb{R}^n$ gegeben. Es sei die kleinste euklidische Kugel zu bestimmen, die $a_1, \ldots, a_l$ enthält. Man formuliere diese Aufgabe als eine Optimierungsaufgabe (P), bei der eine lineare Zielfunktion unter konvexen quadratischen Ungleichungsrestriktionen zu minimieren ist und stelle das zugehörige duale Programm (D) auf. Weiter zeige man, dass beide Probleme lösbar sind und $\max(\text{D}) = \min(\text{P})$ gilt.

Lösung: Zum Teil haben wir diese Aufgabe schon in Kapitel 1 behandelt. Das primale Problem kann geschrieben werden als

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(\delta,x) := \delta \quad \text{auf} \\ M := \{(\delta,x) \in \mathbb{R} \times \mathbb{R}^n : \frac{1}{2}\|x - a_i\|_2^2 \leq \delta,\ i = 1, \ldots, l\}. \end{cases}$$

Dieses Problem ist lösbar. Wenn man mit Kanonen auf Spatzen schießen will, so kann man den Existenzsatz für konvexe quadratisch restringierte quadratische Programme anwenden. Aber natürlich führt auch ein einfaches Kompaktheitsargument zum Ziel. Ist (δ^*, x^*) eine Lösung von (P), so ist x^* der Mittelpunkt und $r^* := \sqrt{2\delta^*}$ der Radius der gesuchten euklidischen Kugel. Auch das zu (P) duale Programm haben wir in Kapitel 1 schon berechnet, es ist

$$\text{(D)} \qquad \begin{cases} \text{Minimiere} \quad \phi(u) := \frac{1}{2}\sum_{i=1}^{l} u_i \|a_i\|_2^2 - \frac{1}{2}\left\|\sum_{i=1}^{m} u_i a_i\right\|_2^2 \quad \text{auf} \\ N := \{u \in \mathbb{R}^m : u \geq 0,\ e^T u = 1\}. \end{cases}$$

Da N kompakt und die Zielfunktion ϕ stetig ist, ist (D) trivialerweise lösbar. Etwa wegen des Dualitätssatzes 2.9 ist $\min(\text{P}) = \max(\text{D})$.

4. Gegeben sei das konvexe Programm

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \in C,\ g(x) \leq 0\}.$$

Hierbei wird vorausgesetzt:

(V) $C \subset \mathbb{R}^n$ ist nichtleer und konvex, $f\colon C \longrightarrow \mathbb{R}$ und $g\colon C \longrightarrow \mathbb{R}^l$ sind (komponentenweise) konvex.

Ferner sei die Slatersche Constraint Qualification erfüllt, d. h. es existiere ein $\hat{x} \in C$ mit $g(\hat{x}) < 0$. Man zeige: Ist (P) zulässig und $\inf(\mathrm{P}) > -\infty$, so ist die Menge N_{opt} der Lösungen des zu (P) dualen Programms

$$\text{(D)} \qquad \text{Maximiere} \quad \phi(u) := \inf_{x \in C} L(x,u) \quad \text{auf} \quad N := \{u \in \mathbb{R}^l : u \geq 0,\ \phi(u) > -\infty\}$$

nichtleer und kompakt. Hierbei ist $L(x,u) := f(x) + u^T g(x)$ die zu (P) gehörende Lagrange-Funktion.

Lösung: Wegen des starken Dualitätssatzes 2.3 ist $N_{\mathrm{opt}} \neq \emptyset$. Zunächst zeigen wir die Abgeschlossenheit von N_{opt}. Sei hierzu $\{u_k\} \subset N_{\mathrm{opt}}$ eine Folge mit $u_k \to u$. Natürlich ist $u \geq 0$. Mit einem beliebigen $z \in C$ ist ferner

$$\max(\mathrm{D}) = \phi(u_k) = \inf_{x \in C} L(x,u_k) \leq L(z,u_k) \to L(z,u).$$

Daher ist $\max(\mathrm{D}) \leq \phi(u)$, woraus $u \in N_{\mathrm{opt}}$ und damit die Abgeschlossenheit von N_{opt} folgt. Nun zeigen wir, dass N_{opt} auch beschränkt ist. Sei hierzu $u \in N_{\mathrm{opt}}$ beliebig. Dann ist

$$\max(\mathrm{D}) = \phi(u) = \inf_{x \in C} L(x,u) \leq f(\hat{x}) + u^T g(\hat{x}).$$

Wegen $g(\hat{x}) < 0$ existiert ein $\epsilon > 0$ mit $g(\hat{x}) \leq -\epsilon e$, wobei e wieder einmal der Vektor ist, dessen Komponenten alle gleich 1 sind. Daher ist

$$0 \leq u^T e = \|u\|_1 \leq \frac{f(\hat{x}) - \max(\mathrm{D})}{\epsilon},$$

also N_{opt} beschränkt.

5. Gegeben sei die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := c^T x + \tfrac{1}{2} x^T Q x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : \tfrac{1}{2}\|x\|_2^2 \leq \tfrac{1}{2}\Delta^2\},$$

wobei $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit ist und $\Delta > 0$. Man stelle das zu (P) duale Programm (D) auf und zeige, dass (P) und (D) lösbar sind und $\max(\mathrm{D}) = \min(\mathrm{P})$ gilt.

Lösung: Die zu (P) gehörende Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}$ ist durch

$$L(x,u) := c^T x + \tfrac{1}{2} x^T Q x + \tfrac{1}{2} u(\|x\|_2^2 - \Delta^2)$$

gegeben. Das duale Problem ist gegeben durch

$$\text{(D)} \qquad \text{Maximiere} \quad \phi(u) := \inf_{x \in \mathbb{R}^n} L(x,u) \quad \text{auf} \quad N := \{u \in \mathbb{R} : u \geq 0,\ \phi(u) > -\infty\}.$$

Das Problem (P) ist lösbar, da M kompakt. Die Slatersche Constraint Qualification ist erfüllt (setze $\hat{x} := 0$), wegen des starken Dualitätssatzes 2.3 folgt die Lösbarkeit von (D) und $\min(\mathrm{P}) = \max(\mathrm{D})$.

6. Unter der Voraussetzung

(V) $C \subset \mathbb{R}^n$ ist nichtleer und konvex, $f\colon C \longrightarrow \mathbb{R}$ und $g\colon C \longrightarrow \mathbb{R}^l$ sind (komponentenweise) konvex, $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ist affin linear

betrachte man das konvexe Programm

(P) Minimiere $f(x)$ $M := \{x \in \mathbb{R}^n : x \in C,\ g(x) \leq 0,\ h(x) = 0\}$.

Ein Tripel $(x^*, u^*, v^*) \in C \times \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ nennen wir einen *Sattelpunkt* der Lagrange-Funktion $L(x, u, v) := f(x) + u^T g(x) + v^T h(x)$, wenn

$$\begin{cases} L(x^*, u, v) \leq L(x^*, u^*, v^*) \leq L(x, u^*, v^*) \\ \quad \text{für alle} \quad (x, u, v) \in C \times \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m. \end{cases}$$

Man zeige:

(a) Ist $x^* \in M$ eine Lösung von (P) und ist die Slatersche Constraint Qualification aus dem starken Dualitätssatz 2.3 erfüllt, so existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ derart, dass (x^*, u^*, v^*) ein Sattelpunkt von L ist.

(b) Ist $(x^*, u^*, v^*) \in C \times \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ ein Sattelpunkt von L, so ist x^* eine Lösung von (P).

Lösung: Sei $x^* \in M$ eine Lösung von (P) und die Slatersche Constraint Qualification erfüllt. Wegen des starken Dualitätssatzes 2.3 existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit $u^* \geq 0$ und

$$(*) \qquad \min(\mathrm{P}) = f(x^*) = \inf_{x \in C} L(x, u^*, v^*),$$

wobei L natürlich die zu (P) gehörende Lagrange-Funktion bezeichnet. Hieraus folgt

$$f(x^*) \leq L(x^*, u^*, v^*) = f(x^*) + \underbrace{(u^*)^T g(x^*)}_{\leq 0} + \underbrace{(v^*)^T h(x^*)}_{=0} \leq f(x^*),$$

so dass $L(x^*, u^*, v^*) \leq L(x, u^*, v^*)$ für alle $x \in C$ folgt. Für beliebige $(u, v) \in \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ ist andererseits

$$L(x^*, u, v) = f(x^*) + \underbrace{u^T g(x^*)}_{\leq 0} + \underbrace{v^T h(x^*)}_{=0} \leq f(x^*) = L(x^*, u^*, v^*),$$

insgesamt ist gezeigt, dass (x^*, u^*, v^*) ein Sattelpunkt der Lagrange-Funktion ist.

Sei nun umgekehrt (x^*, u^*, v^*) ein Sattelpunkt von L. Wir zeigen zunächst, dass x^* zulässig für (P) ist. Angenommen, es wäre $-(g(x^*), h(x^*)) \notin \mathbb{R}^l_{\geq 0} \times \{0\}$. Der starke Trennungssatz liefert nach einfacher Argumentation die Existenz eines Paares $(\hat{u}, \hat{v}) \in \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ mit

$$\hat{u}^T g(x^*)) + \hat{v}^T h(x^*) > 0.$$

Dann ist aber $(u^* + \hat{u}, v^* + \hat{v}) \in \mathbb{R}^l_{\geq 0} \times \mathbb{R}^m$ und

$$L(x^*, u^* + \hat{u}, v^* + \hat{v}) > L(x^*, u^*, v^*),$$

ein Widerspruch dazu, dass (x^*, u^*, v^*) ein Sattelpunkt von L ist. Wieder ist $f(x^*) = L(x^*, u^*, v^*)$, für ein beliebiges $x \in M$ ist daher

$$f(x^*) = L(x^*, u^*, v^*) \leq L(x, u^*, v^*) = f(x) + \underbrace{(u^*)^T g(x)}_{\leq 0} + \underbrace{(v^*)^T h(x)}_{=0} \leq f(x),$$

also $x^* \in M$ eine Lösung von (P).

7. Seien $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$. Hiermit betrachte man die zueinander dualen linearen Programme

   (P) $\qquad$ Minimiere $\quad c^T x \quad$ auf $\quad M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b\}$

   und

   (D) $\qquad$ Maximiere $\quad b^T y \quad$ auf $\quad N := \{y \in \mathbb{R}^m : A^T y \leq c\}.$

   Es werde vorausgesetzt, daß

   $$M_0 := \{x \in \mathbb{R}^n : x > 0,\ Ax = b\} \neq \emptyset, \qquad N_0 := \{y \in \mathbb{R}^m : A^T y < c\} \neq \emptyset$$

   und Rang $(A) = m$. Man zeige, dass dann die Mengen M_{opt} und N_{opt} der optimalen Lösungen von (P) bzw. (D) nichtleer und kompakt sind.

   Lösung: Da insbesondere vorausgesetzt wird, dass (P) und (D) zulässig sind, sind (P) und (D) jeweils lösbar bzw. $M_{\text{opt}} \neq \emptyset$ und $N_{\text{opt}} \neq \emptyset$.

   Es ist

   $$M_{\text{opt}} = \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b,\ c^T x = \min(\text{P})\}.$$

   Angenommen, M_{opt} wäre nicht beschränkt. Dann existiert eine Folge $\{x_k\} \subset M_{\text{opt}}$ mit $\|x_k\| \to \infty$. O. B. d. A. konvergiert die Folge $\{p_k\}$, wobei $p_k := x_k/\|x_k\|$, gegen ein p. Dieser Vektor p ist vom Nullvektor verschieden und genügt

   $$p \geq 0, \qquad Ap = 0, \qquad c^T p = 0.$$

   Mit einem $y \in N_0$ ist dann

   $$0 < (c - A^T y)^T p = c^T p - y^T A p = 0,$$

   ein Widerspruch. Also ist M_{opt} beschränkt und dann auch, da die Abgeschlossenheit trivial ist, kompakt.

   Es ist

   $$N_{\text{opt}} = \{y \in \mathbb{R}^m : A^T y \leq c,\ b^T y = \max(\text{D})$$

   Wäre N_{opt} nicht beschränkt, so existierte entsprechend der obigen Argumentation ein $q \neq 0$ mit

   $$A^T q \leq 0, \qquad b^T q = 0.$$

   Mit einem $x \in M_0$ wäre

   $$0 = b^T q = (Ax)^T q = x^T A^T q.$$

   Da hier $x > 0$ und $A^T q \leq 0$ folgt $A^T q = 0$. Wegen Rang $(A) = m$ ist $q = 0$, ein Widerspruch. Insgesamt ist die Aufgabe gelöst.

8. Gegeben sei ein Vektor $x = (x_j) \in \mathbb{R}^n$ und $r \in \{1, \ldots, n\}$. Sei $p = \{p_1, \ldots, p_n\}$ eine Permutation von $\{1, \ldots, n\}$ mit $x_{p_1} \geq \cdots \geq x_{p_n}$. Man zeige, dass

   $$\sum_{j=1}^{r} x_{p_j} = \max\{x^T z : 0 \leq z \leq e,\ e^T z = r\},$$

   wobei e der Vektor im $\mathbb{R}^n$ ist, dessen Komponenten alle gleich 1 sind.

Lösung: Sei $z^* \in \mathbb{R}^n$ definiert durch

$$z^*_{p_j} := \begin{cases} 1, & j = 1, \ldots, r, \\ 0, & j = r+1, \ldots, n. \end{cases}$$

Dann ist $0 \le z^* \le e$, $e^T z^* = r$ und daher

$$\sum_{j=1}^{r} x_{p_j} = x^T z^* \le \max\{x^T z : 0 \le z \le e,\ e^T z = r\}.$$

Andererseits sei ein $z \in \mathbb{R}^n$ mit $0 \le z \le e$ und $e^T z = r$ vorgegeben. Dann ist

$$\begin{aligned} x^T z &= \sum_{j=1}^{n} x_j z_j \\ &= \sum_{j=1}^{n} x_{p_j} z_{p_j} \\ &= \sum_{j=1}^{r} x_{p_j} z_{p_j} + \sum_{j=r+1}^{n} x_{p_j} z_{p_j} \\ &= \sum_{j=1}^{r} x_{p_j} + \sum_{j=1}^{r} \underbrace{x_{p_j}(z_{p_j} - 1)}_{\le x_{p_r}(z_{p_j}-1)} + \sum_{j=r+1}^{n} \underbrace{x_{p_j} z_{p_j}}_{\le x_{p_r} z_{p_j}} \\ &\le \sum_{j=1}^{r} x_{p_j} + x_{p_r} \underbrace{\left(\sum_{j=1}^{n} z_{p_j} - r \right)}_{=0} \\ &= \sum_{j=1}^{r} x_{p_j}. \end{aligned}$$

Insgesamt ist die Aufgabe gelöst.

9. Gegeben sei das lineare Programm

   (P) Minimiere $c^T x$ auf $M := \{x \in \mathbb{R}^n : x \ge 0,\ Ax = b\}$.

   Hierbei seien $A \in \mathbb{R}^{m \times n}$ mit $m < n$ und Rang $(A) = m$ sowie $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ gegeben.

   (a) Man zeige, dass eine Matrix $B \in \mathbb{R}^{(n-m)\times n}$ mit Rang $(B) = n - m$ und $AB^T = 0$ existiert. Sind B_1 und B_2 zwei Matrizen mit diesen beiden Eigenschaften, so existiert eine nichtsinguläre Matrix $T \in \mathbb{R}^{(n-m)\times(n-m)}$ mit $B_1 = TB_2$.

   (b) Sei $B \in \mathbb{R}^{(n-m)\times n}$ wie in (a) gegeben, ferner sei $d := A^T(AA^T)^{-1}b$. Hiermit betrachte man das (von der Wahl von B unabhängige) lineare Programm

   (D) Minimiere $d^T y$ auf $N := \{y \in \mathbb{R}^n : y \ge 0,\ By = Bc\}$.

   Man begründe, weshalb (D) mit einigem Recht als zu (P) duales Programm bezeichnet werden kann, und beweise insbesondere einen schwachen und einen starken Dualitätssatz:

(i) Sind $x \in M$ und $y \in N$, so ist $x^T y \geq 0$.

(ii) Sind (P) und (D) zulässig, so besitzen beide Programme Lösungen $x^* \in M$ bzw. $y^* \in N$ und es ist $(x^*)^T y^* = 0$.

Lösung: Wegen Rang $(A) = m$ ist $\dim \operatorname{Kern}(A) = n - m$. Sei $\{b_1, \ldots, b_{n-m}\}$ eine Basis von Kern (A) und $B := (\ b_1 \ \cdots \ b_{n-m}\ )^T$. Dann ist $B \in \mathbb{R}^{(n-m)\times n}$ eine Matrix mit Rang $(B) = n - m$, ferner ist offensichtlich $AB^T = 0$, da $B^T z \in \operatorname{Kern}(A)$ für jedes $z \in \mathbb{R}^{n-m}$. Die zweite Aussage in (a) ist trivial. Das „übliche“ zu (P) duale lineare Programm lautet

$$\text{Maximiere} \quad b^T u \quad \text{unter der Nebenbedingung} \quad A^T u \leq c$$

bzw. nach Einführung einer Schlupfvariablen

$$(\mathrm{D}') \qquad \text{Maximiere} \quad b^T u \quad \text{auf} \quad N' := \{(u, y) \in \mathbb{R}^m \times \mathbb{R}^n : y \geq 0,\ A^T u + y = c\}.$$

Für $(u, y) \in N'$ ist $u = (AA^T)^{-1}Ac - (AA^T)^{-1}Ay$ und daher $b^T u = c^T d - d^T y$. Schließlich ist $(u, y) \in N'$ genau dann, wenn $y \in N$ und $u = (AA^T)^{-1}Ac - (AA^T)^{-1}Ay$. Daher entsteht (D) aus (D'), indem man die Variable u eliminiert, so dass als Variable in (D) einzig die (nichtnegative) Schlupfvariable übrig bleibt. Die Aussagen (i) und (i) sind dann einfach zu beweisen. Der Vorteil der hier gewählten Formulierung des dualen Programms gegenüber der üblichen besteht natürlich darin, dass hier das duale Programm, genau wie das primale Programm, in Normalform vorliegt.

10. Gegeben seien symmetrische, positiv semidefinite Matrizen $A_1, \ldots, A_m \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n \setminus \{0\}$ und $v > 0$. Es wird vorausgesetzt, dass die Matrix $A(y) := \sum_{i=1}^m y_i A_i$ für jedes $y > 0$ positiv definit ist. Man betrachte die beiden Probleme

$$(\mathrm{P}_1) \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad \dfrac{1}{2} c^T x \quad \text{auf} \\ P := \left\{ (x, y) \in \mathbb{R}^n \times \mathbb{R}^m : \displaystyle\sum_{i=1}^m y_i A_i x = c,\ e^T y = v,\ y \geq 0 \right\} \end{array} \right.$$

und

$$(\mathrm{P}_2) \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad \delta \quad \text{auf} \\ M := \left\{ (z, \delta) \in \mathbb{R}^n \times \mathbb{R} : \dfrac{v}{2} z^T A_i z - c^T z - \delta \leq 0,\ i = 1, \ldots, m \right\}. \end{array} \right.$$

Man beachte, dass (P_2) eine konvexe, quadratisch restringierte Optimierungsaufgabe mit einer linearen Zielfunktion ist. Man zeige[4]:

(a) Die beiden Optimierungsaufgaben (P_1) und (P_2) sind zulässig.

(b) Sei $(x, y) \in P$ zulässig für (P_1) und $(z, \delta) \in M$ zulässig für (P_2). Dann ist $\delta \geq -\frac{1}{2} c^T x$. Hieraus schließe man, dass (P_2) lösbar ist und $\inf(\mathrm{P}_1) \geq -\min(\mathrm{P}_2)$ gilt.

[4] Ähnliche Aussagen werden bei

A. Ben-Tal, M. P. Bendsøe (1993) A new method for optimal truss topology design. SIAM J. Optim. 3, 322–358

gemacht.

(c) Man zeige, dass die Slatersche Constraint Qualification für das Programm (P_2) erfüllt ist. Hieraus schließe man, dass das zu (P_2) duale Programm (D_2) lösbar ist und keine Dualitätslücke auftritt, also $\min(\mathrm{P}_2) = \max(\mathrm{D}_2)$ gilt.

(d) Sei u^* eine Lösung des zu (P_2) dualen Programms. Man setze $y^* := vu^*$ und zeige die Existenz eines $x^* \in \mathbb{R}^n$ mit der Eigenschaft, dass (x^*, y^*) eine Lösung von (P_1) ist.

Lösung: Man setze $y := (v/n)e$. Dann ist $y > 0$ und $e^T y = v$. Nach Voraussetzung ist die Matrix $A(y) := \sum_{i=1}^m y_i A_i$ positiv definit, insbesondere nichtsingulär. Setzt man $x := A(y)^{-1}c$, so ist $(x, y) \in P$ zulässig für (P_1). Um die Zulässigkeit von (P_2) nachzuweisen, wähle man $z \in \mathbb{R}^n$ beliebig. Ist dann

$$\delta \geq \max_{i=1,\ldots,m} \left\{ \frac{v}{2} z^T A_i x - c^T x \right\},$$

so ist $(z, \delta) \in M$ zulässig für (P_2).

Sei $(x, y) \in P$ zulässig für (P_1) und (z, δ) zulässig für (P_2). Dann ist

$$\delta \geq \frac{v}{2} z^T A_i z - c^T z, \qquad i = 1, \ldots, m.$$

Multipliziert man diese Ungleichungen mit $y_i \geq 0$, addiert sie und berücksichtigt, dass $e^T y = v > 0$, so erhält man

$$\begin{aligned}
\delta + \frac{1}{2} c^T x &\geq \frac{1}{2} z^T \left(\sum_{i=1}^m y_i A_i \right) z - c^T z + \frac{1}{2} c^T x \\
&= \frac{1}{2} z^T \left(\sum_{i=1}^m y_i A_i \right) z - z^T \left(\sum_{i=1}^m y_i A_i \right) x + \frac{v}{2} x^T \left(\sum_{i=1}^m y_i A_i \right) x \\
&= \frac{1}{2} (z - x)^T \left(\sum_{i=1}^m y_i A_i \right) (z - x) \\
&\geq 0,
\end{aligned}$$

also, wie behauptet, $\delta \geq -\frac{1}{2} c^T x$. Insbesondere ist $\inf(\mathrm{P}_2) > -\infty$. Aus dem Existenzsatz 2.7 für konvexe, quadratisch restringierte quadratische Programme folgt die Existenz einer Lösung (z^*, δ^*) von (P_2), weiter ist $\inf(\mathrm{P}_1) \geq -\min(\mathrm{P}_2)$.

Ist $z \in \mathbb{R}^n$ beliebig und

$$\delta > \max_{i=1,\ldots,m} \left\{ \frac{v}{2} z^T A_i x - c^T x \right\},$$

so sind durch (z, δ) alle Ungleichungsrestriktionen strikt erfüllt, d. h. es gilt die Slatersche Constraint Qualification. Der starke Dualitätssatz 2.3 zeigt, dass das zu (P_2) duale Programm ebenfalls lösbar ist und keine Dualitätslücke auftritt.

Die Lagrange-Funktion zu (P_2) ist

$$L(z, \delta, u) := \delta + \sum_{i=1}^m u_i \left(\frac{v}{2} z^T A_i z - c^T z - \delta \right),$$

mit

$$\phi(u) := \inf_{(z,\delta) \in \mathbb{R}^n \times \mathbb{R}} L(z, \delta, u)$$

ist das duale Problem, wie üblich, gerade

$$(\mathrm{D}_2) \qquad \text{Maximiere} \quad \phi(u) \quad \text{auf} \quad N := \{u \in \mathbb{R}^m : u \geq 0,\ \phi(u) > -\infty\}.$$

Für ein gegebenes $u \geq 0$ ist offenbar $\phi(u) > -\infty$ genau dann, wenn $e^T u = 1$ und ein $x \in \mathbb{R}^n$ mit

$$\sum_{i=1}^{m} u_i v A_i x = c$$

existiert, in diesem Falle ist

$$\phi(u) = \sum_{i=1}^{m} u_i \left(\frac{v}{2} x^T A_i x - c^T x \right) = \frac{v}{2} \sum_{i=1}^{m} u_i x^T A_i x - c^T x = -\frac{1}{2} c^T x.$$

Definiert man daher $y := vu$, so ist das Paar (x, y) zulässig für (P_2). Sei nun $u^* \in N$ eine Lösung von (D_2), $x^* \in \mathbb{R}^n$ ein zugehöriger Vektor mit $\sum_{i=1}^m u_i^* v A_i x^* = c$. Wir wollen zeigen, dass (x^*, y^*) mit $y^* := vu^*$ eine Lösung von (P_1) ist. Wegen der schon bewiesenen Teile (b) und (c) der Aufgabe ist

$$\inf(\mathrm{P}_1) \geq -\min(\mathrm{P}_2) = -\phi(u^*) = \frac{1}{2} c^T x^* \geq \inf(\mathrm{P}_1).$$

Also ist $(x^*, y^*) \in P$ eine Lösung von (P_1).

6.2.3 Aufgaben in Abschnitt 2.3

1. Man zeige, dass $x^* := (1, 1, 2)^T$ die Lösung von

$$(\mathrm{P}) \qquad \left\{ \begin{array}{lrrrrr} \text{Minimiere} & -5x_2 + \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) & & & & \text{unter den Nebenbedingungen} \\ & -4x_1 & -3x_2 & & \geq & -8 \\ & 2x_1 & + x_2 & & \geq & 2 \\ & & -2x_2 & +x_3 & \geq & 0 \\ & x_1 & -2x_2 & +x_3 & = & 1 \end{array} \right.$$

ist.

Lösung: Es handelt sich hier um eine konvexe Optimierungsaufgabe, die notwendigen Optimalitätsbedingungen sind daher auch hinreichend. Offensichtlich ist x^* zulässig, wobei die erste und die zweite Ungleichungsrestriktion inaktiv sind. Zu bestimmen sind daher $u_3^* \geq 0$, $v^* \in \mathbb{R}$ mit

$$\begin{pmatrix} 1 \\ -4 \\ 2 \end{pmatrix} + u_3^* \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix} + v^* \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Mit $u_3^* = 1$, $v^* = -1$ sind alle diese Bedingungen erfüllt. Da die Zielfunktion strikt konvex ist, ist x^* die einzige Lösung.

2. Für die Aufgabe

$$(\mathrm{P}) \qquad \left\{ \begin{array}{l} \text{Minimiere} \quad f(x) := x_1^2 + 4x_2^2 + 16x_3^2 \quad \text{unter der Nebenbedingung} \\ \qquad\qquad h(x) := x_1 x_2 x_3 - 1 = 0 \end{array} \right.$$

bestimme man alle Punkte, in denen die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind und prüfe anschließend mit Optimalitätsbedingungen zweiter Ordnung, ob dies lokale Lösungen sind.

Lösung: Die notwendigen Optimalitätsbedingungen sind in x^* erfüllt, wenn ein $v^* \in \mathbb{R}$ mit

$$\begin{pmatrix} 2x_1^* \\ 8x_2^* \\ 32x_3^* \end{pmatrix} + v^* \begin{pmatrix} x_2^* x_3^* \\ x_1^* x_3^* \\ x_1^* x_2^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

$$x_1^* x_2^* x_3^* - 1 = 0$$

existiert. Sicher ist, dass die Komponenten von x^* nicht verschwinden, weil andernfalls die Nebenbedingung nicht erfüllt wäre. Aus den ersten Gleichungen folgt

$$\frac{2x_1^*}{x_2^* x_3^*} = \frac{8x_2^*}{x_1^* x_3^*} = \frac{32x_3^*}{x_1^* x_2^*}.$$

Aus diesen Gleichungen folgt unschwer

$$x_1^* = \pm 2x_2^*, \qquad x_2^* = \pm 2x_3^*.$$

Dann ist

$$1 = x_1^* x_2^* x_3^* = \pm 8(x_3^*)^3$$

und folglich

$$x_3^* = \pm\frac{1}{2}, \qquad x_2^* = \pm 1, \qquad x_1^* = \pm 2.$$

Von diesen 8 möglichen Lösungen bleiben nur 4 übrig, denn wegen der Nebenbedingung ist die Anzahl negativer Komonenten von x^* gerade. Diese möglichen Lösungen sind

$$x^{(1)} := \begin{pmatrix} 2 \\ 1 \\ \frac{1}{2} \end{pmatrix}, \qquad x^{(2)} := \begin{pmatrix} -2 \\ -1 \\ \frac{1}{2} \end{pmatrix}, \qquad x^{(3)} := \begin{pmatrix} 2 \\ -1 \\ -\frac{1}{2} \end{pmatrix}, \qquad x^{(4)} := \begin{pmatrix} -2 \\ 1 \\ -\frac{1}{2} \end{pmatrix}.$$

Diese vier Vektoren genügen jeweils den notwendigen Opyimalitätsbedingung (und der Restriktion), der zugehörige Multiplikator ist jeweils $v^* = -8$. Es ist $f(x^{(i)}) = 12$, $i = 1, 2, 3, 4$. Nun prüfen wir mit den hinreichenden Optimalitätsbedingungen nach, welche der angegebenen potentiellen Lösungen den hinreichenden Bedingungen zweiter Ordnung genügt. Es muss jeweils nachgeprüft werden, ob $\nabla^2 f(x^*) + v^* \nabla^2 h(x^*)$ auf Kern $(h'(x^*))$ positiv definit ist. Es ist

$$\nabla^2 f(x^*) + v^* \nabla^2 h(x^*) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 32 \end{pmatrix} - 8 \begin{pmatrix} 0 & x_3^* & x_2^* \\ x_3^* & 0 & x_1^* \\ x_2^* & x_1^* & 0 \end{pmatrix}.$$

Wir gehen die vier Fälle der Reihe nach durch.

(a) Für $x^* = x^{(1)}$ ist zu prüfen, ob

$$\begin{pmatrix} \frac{1}{2} \\ 1 \\ 2 \end{pmatrix}^T \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = 0, \quad p \neq 0 \Longrightarrow p^T \begin{pmatrix} 2 & -4 & -8 \\ -4 & 8 & -16 \\ -8 & -16 & 32 \end{pmatrix} p > 0.$$

Wegen

$$\operatorname{span}\left\{\begin{pmatrix} \frac{1}{2} \\ 1 \\ 2 \end{pmatrix}\right\}^{\perp} = \operatorname{span}\left\{\begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}, \begin{pmatrix} -4 \\ 0 \\ 1 \end{pmatrix}\right\}$$

ist nachzuprüfen, ob die Matrix

$$\begin{pmatrix} 0 & -2 & 1 \\ -4 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & -4 & -8 \\ -4 & 8 & -16 \\ -8 & -16 & 32 \end{pmatrix} \begin{pmatrix} 0 & -4 \\ -2 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 128 & 64 \\ 64 & 128 \end{pmatrix}$$

positiv definit ist. Dies ist der Fall, daher ist bei $x^{(1)}$ ein lokales Minimum der Aufgabe, die Zielfunktion f unter der angegebenen Gleichungsrestriktion zu minimieren.

(b) Für $x^* = x^{(2)}$ ist nachzuprüfen, ob

$$\begin{pmatrix} -\frac{1}{2} \\ -1 \\ 2 \end{pmatrix}^T \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = 0, \quad p \neq 0 \Longrightarrow p^T \begin{pmatrix} 2 & -4 & 8 \\ -4 & 8 & 16 \\ 8 & 16 & 32 \end{pmatrix} p > 0.$$

Wegen

$$\operatorname{span}\left\{\begin{pmatrix} -\frac{1}{2} \\ -1 \\ 2 \end{pmatrix}\right\}^{\perp} = \operatorname{span}\left\{\begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}\right\}$$

ist nachzuprüfen, ob die Matrix

$$\begin{pmatrix} 0 & 2 & 1 \\ 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & -4 & 8 \\ -4 & 8 & 16 \\ 8 & 16 & 32 \end{pmatrix} \begin{pmatrix} 0 & 4 \\ 2 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 128 & 64 \\ 64 & 128 \end{pmatrix}$$

positiv definit ist. Dies ist der Fall, daher ist bei $x^{(1)}$ ein lokales Minimum der Aufgabe, die Zielfunktion f unter der angegebenen Gleichungsrestriktion zu minimieren.

(c) Die beiden restlichen Fälle können entsprechend behandelt werden, auch hier liegt jeweils eine lokale Lösung vor.

3. Gegeben sei die Optimierungsaufgabe

(P) Minimiere $f(x)$ unter der Nebenbedingung $x \geq 0$.

Sei $x^* \geq 0$ eine lokale Lösung von (P) und die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in x^* stetig differenzierbar. Man stelle die notwendigen Optimalitätsbedingungen erster Ordnung auf.

Lösung: Ist x^* eine lokale Lösung so existiert ein $u^* \in \mathbb{R}^n$ mit

$$u^* \geq 0, \qquad \nabla f(x^*) - u^* = 0, \qquad (u^*)^T x^* = 0.$$

Dies bedeutet also, dass $\nabla f(x^*) \geq 0$ und

$$x_j^* > 0 \Longrightarrow \frac{\partial f}{\partial x_j}(x^*) = 0.$$

Hierfür muss man natürlich nicht den Satz von Kuhn-Tucker anwenden, die Aussage folgt auch aus $\nabla f(x^*)^T (x - x^*) \geq 0$ für alle $x \geq 0$.

4. Ganz ohne Constraint Qualification kann man immer noch den *Satz von F. John* beweisen:

   Sei x^* eine lokale Lösung von

   $$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

   Hierbei seien die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ auf einer Umgebung von x^* stetig differenzierbar. Dann existiert ein von Null verschiedenes Tripel $(u_0^*, u^*, v^*) \in \mathbb{R} \times \mathbb{R}^l \times \mathbb{R}^m$ mit

   $$(u_0^*, u^*) \geq (0,0), \qquad u_0^* \nabla f(x^*) + g'(x^*)^T u^* + h'(x^*)^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

   Ist die Arrow-Hurwicz-Uzawa Constraint Qualification erfüllt, so ist hier notwendigerweise $u_0^* > 0$.

   Lösung: Die Arrow-Hurwicz-Uzawa Constraint Qualification besagt, dass ein $\hat{p} \in \mathbb{R}^n$ mit $\nabla g_i(x^*)^T \hat{p} < 0$, $i \in I(x^*)$, und $h'(x^*)\hat{p} = 0$ existiert und Rang $(h'(x^*)) = m$ ist. Wir können davon ausgehen, dass die Arrow-Hurwicz-Uzawa Constraint Qualification nicht erfüllt ist, da andernfalls die Aussage des Satzes wegen Satz 3.5 mit $u_0^* := 1$ richtig ist. Ist Rang $(h'(x^*)) < m$, so lassen sich die Gradienten $\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ nichttrivial zu Null kombinieren und die Aussage ist mit $(u_0^*, u^*) := (0,0)$ richtig. Daher bleibt der Fall zu betrachten, dass Rang $(h'(x^*)) = m$, aber das System

   $$\nabla g_i(x^*)^T p < 0 \ \ (i \in I(x^*)), \qquad h'(x^*)p = 0$$

   nicht lösbar ist. Mit dem Farkas-Lemma erhält man die Existenz von $u_i^* \geq 0$, $i \in I(x^*)$, die nicht alle verschwinden, sowie von $v_i^* \in \mathbb{R}$, $i = 1, \ldots, m$, mit

   $$\sum_{i \in I(x^*)} u_i^* \nabla g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla h_i(x^*) = 0.$$

   Diese Anwendung des Farkas-Lemmas verläuft im Prinzip folgendermaßen: Angenommen, das System

   $$Ap < 0, \qquad Bp = 0$$

   sei nicht lösbar. Dann hat auch

   $$-Ap + \delta e \geq 0, \qquad -Bp = 0, \qquad \delta = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} p \\ \delta \end{pmatrix} < 0$$

   keine Lösung. Das verallgemeinerte Farkas-Lemma zeigt die Existenz von (u, v) mit

   $$\begin{pmatrix} -A^T & -B^T \\ e^T & 0^T \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad u \geq 0.$$

   Also existiert ein Paar (u, v) mit $u \geq 0$, $u \neq 0$ und $A^T u + B^T v = 0$. Damit ist der Beweis des Satzes von F. John vollständig, wenn wir uns noch den letzten Zusatz überlegen. Sei also (u_0^*, u^*, v^*) ein Tripel, das den Bedingungen des Satzes von F. John genügt, ferner sei die Arrow-Hurwicz-Uzawa Constraint Qualification erfüllt. Angenommen, es

wäre $u_0^* = 0$. Die zu inaktiven Ungleichungsrestriktionen gehörenden Multiplikatoren u_i^* verschwinden wegen der Gleichgewichtsbedingung. Aus

$$\sum_{i\in I(x^*)} \underbrace{u_i^*}_{\geq 0} \underbrace{\nabla g_i(x^*)^T \hat{p}}_{<0} + \sum_{i=1}^{m} v_i^* \underbrace{\nabla h_i(x^*)^T \hat{p}}_{=0} = 0$$

folgt zunächst $u_i^* = 0$, $i \in I(x^*)$, danach wegen Rang $(h'(x^*)) = m$ auch $v^* = 0$, ein Widerspruch zu $(u_0^*, u^*, v^*) \neq (0, 0, 0)$.

5. Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}.$$

Hierbei seien $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ auf dem $\mathbb{R}^n$ konvex und stetig differenzierbar, $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear. Wie üblich sei die Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ zu (P) durch

$$L(x, u, v) := f(x) + u^T g(x) + v^T h(x)$$

definiert. Das zu (P) sogenannte *Wolfe-duale* Programm (siehe P. Wolfe (1961)[5]) ist dann durch

$$\text{(D)} \qquad \left\{ \begin{array}{c} \text{Maximiere} \quad L(z, u, v) \quad \text{auf} \\ N := \{(z, u, v) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \nabla_x L(z, u, v) = 0\} \end{array} \right.$$

gegeben. Man zeige:

(a) Ist $x \in M$ und $(z, u, v) \in N$, so ist $L(z, u, v) \leq f(x)$. Zwischen (P) und (D) gilt also ein schwacher Dualitätssatz.

(b) Die (schwache) Slatersche Constraint Qualification sei erfüllt, d. h. es existiere ein $\hat{x} \in M$ mit $g_i(\hat{x}) < 0$ für alle i, für die g_i nicht affin linear ist. Ist dann $x^* \in M$ eine Lösung von (P), so existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ derart, dass $(x^*, u^*, v^*) \in N$ und $f(x^*) = L(x^*, u^*, v^*)$. Ferner ist (u^*, v^*) eine Lösung des zu (P) Lagrange-dualen Programms.

Lösung: Sei $x \in M$ und $(z, u, v) \in N$. Dann ist $L(\cdot, u, v)$ konvex und daher

$$0 = \nabla_x L(z, u, v)^T (x - z) \leq L(x, u, v) - L(z, u, v) \leq f(x) - L(z, u, v),$$

womit der erste Teil schon bewiesen ist.

Sei $x^* \in M$ eine Lösung von (P) und die schwache Slatersche Constraint Qualification erfüllt. Wegen Korollar 3.7 existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u^* \geq 0, \qquad \nabla_x L(x^*, u^*, v^*) = 0, \qquad g(x^*)^T u^* = 0.$$

Offensichtlich ist $(x^*, u^*, v^*) \in N$ und

$$L(x^*, u^*, v^*) = f(x^*) + \underbrace{(u^*)^T g(x^*)}_{=0} + (v^*)^T \underbrace{h(x^*)}_{=0} = f(x^*).$$

[5] Wolfe (1961) "A duality theorem for nonlinear programming." Quarterly of Applied Mathematics 19, 239–244.

Wegen der Konvexität von $L(\cdot, u^*, v^*)$ folgt aus $\nabla_x L(x^*, u^*, v^*) = 0$, dass

$$f(x^*) = L(x^*, u^*, v^*) = \inf_{x \in \mathbb{R}^n} L(x, u^*, v^*),$$

was zeigt, dass (u^*, v^*) eine Lösung des zu (P) Lagrange-dualen Programms ist.

6. Gegeben sei die Optimierungsaufgabe

(P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : h(x) = 0\}$.

Sei $x^* \in M$ eine lokale Lösung von (P) und $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sowie $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ auf einer Umgebung von x^* zweimal stetig differenzierbar. In x^* seien die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt, d. h. es existiere ein $v^* \in \mathbb{R}^m$ mit $\nabla f(x^*) + h'(x^*)^T v^* = 0$ und der Eigenschaft, dass

$$W^* := \nabla^2 f(x^*) + \sum_{i=1}^m v_i^* \nabla^2 h_i(x^*)$$

auf Kern $(h'(x^*))$ positiv definit ist. Schließlich sei Rang $(h'(x^*)) = m$. Man zeige, dass es ein $\sigma_0 > 0$ gibt derart, dass x^* für jedes $\sigma > \sigma_0$ eine isolierte, lokale Lösung der unrestringierten Optimierungsaufgabe

(P_σ) Minimiere $\Phi_\sigma(x) := f(x) + (v^*)^T h(x) + \frac{1}{2}\sigma \|h(x)\|^2, \quad x \in \mathbb{R}^n,$

ist. Hierbei sei $\|\cdot\|$ die euklidische Norm.

Hinweis: Man zeige, dass $\nabla \Phi_\sigma(x^*) = 0$ für alle $\sigma > 0$ und $\nabla^2 \Phi_\sigma(x^*)$ für alle hinreichend großen $\sigma > 0$ positiv definit ist.

Lösung: Es ist

$$\nabla \Phi_\sigma(x^*) = \underbrace{\nabla f(x^*) + h'(x^*)^T v^*}_{=0} + \sigma h'(x^*)^T \underbrace{h(x^*)}_{=0} = 0.$$

Daher haben wir uns noch zu überlegen, dass

$$\nabla^2 \Phi_\sigma(x^* = W^* + \sigma \left[h'(x^*)^T h'(x^*) + \sum_{i=1}^m \underbrace{h_i(x^*)}_{=0} \nabla^2 h_i(x^*) \right] = W^* + \sigma h'(x^*)^T h'(x^*)$$

für alle hinreichend großen $\sigma > 0$ positiv definit ist. Zur Abkürzung setzen wir $B := h'(x^*)$. Sei $p = u + B^T v$ mit $u \in \text{Kern}\,(B)$ ein beliebiges Element des $\mathbb{R}^n$. Dann ist

$$\begin{aligned} p^T \nabla^2 \Phi_\sigma(x^*) p &= (u + B^T v)^T [W^* + \sigma B^T B](u + B^T v) \\ &= u^T W^* u + 2u^T W^* B^T v + v^T B W^* B^T v + \sigma v^T (BB^T)(BB^T) v \\ &\geq \lambda_0 \|u\|^2 - 2\|W^* B^T\| \, \|u\| \, \|v\| - \|B W^* B^T\| \, \|v\|^2 + \sigma \mu_0 \|v\|^2 \end{aligned}$$

mit

$$\lambda_0 := \min_{u \in \text{Kern}\,(B) \setminus \{0\}} \frac{u^T W^* u}{u^T u}, \qquad \mu_0 := \min_{v \in \mathbb{R}^m \setminus \{0\}} \frac{v^T (BB^T)(BB^T) v}{v^T v}.$$

Hier ist $\lambda_0 > 0$, da W^* auf Kern (B) positiv definit ist. Ferner ist $\mu_0 > 0$, denn wegen Rang $(B) = m$ ist BB^T und damit auch $(BB^T)(BB^T)$ positiv definit. Setzt man zur Abkürzung $\gamma := \|W^* B^T\|$ und $\delta := \|B W^* B^T\|$, so ist also

$$p^T \nabla^2 \Phi_\sigma(x^*) p \geq \lambda_0 \|u\|^2 - 2\gamma \|u\| \, \|v\| + (\sigma \mu_0 - \delta) \|v\|^2.$$

O. B. d. A. ist $v \neq 0$ (andernfalls ist $p^T \nabla^2 \Phi_\sigma(x^*) p \geq \lambda_0 \|u\|^2 > 0$ für $u \neq 0$). Für $\sigma > [\delta + \gamma^2/\lambda_0]/\mu_0$ bzw. $\sigma\mu_0 - \delta > \gamma^2/\lambda_0$ ist

$$\begin{aligned} p^T \nabla^2 \Phi_\sigma(x^*) p &> \lambda_0 \|u\|^2 - 2\gamma \|u\| \|v\| + \frac{\gamma^2}{\lambda_0} \|v\|^2 \\ &= \frac{(\lambda_0 \|u\| - \gamma \|v\|)^2}{\lambda_0} \\ &\geq 0. \end{aligned}$$

Damit ist die Aufgabe gelöst.

7. Sei (x^*, v^*) ein Kuhn-Tucker-Paar zu der Optimierungsaufgabe

(P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : h(x) = 0\}$,

also (x^*, v^*) eine Nullstelle der durch

$$T(x, v) := \begin{pmatrix} \nabla f(x) + h'(x)^T v \\ h(x) \end{pmatrix}$$

definierten Abbildung $T\colon \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n \times \mathbb{R}^m$. Man berechne die Funktionalmatrix von T in (x^*, v^*) und untersuche, unter welchen Voraussetzungen diese nichtsingulär ist. Hierbei sind natürlich $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ als zweimal stetig differenzierbar auf einer Umgebung von x^* vorausgesetzt.

Lösung: Offenbar ist

$$T'(x^*, v^*) = \begin{pmatrix} \nabla^2 f(x^*) + \sum_{i=1}^m v_i^* \nabla^2 h_i(x^*) & h'(x^*)^T \\ h'(x^*) & 0 \end{pmatrix}.$$

Mit

$$A := \nabla^2 f(x^*) + \sum_{i=1}^m v_i^* \nabla^2 h_i(x^*), \qquad B := h'(x^*)$$

ist also

$$T'(x^*, v^*) = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}.$$

Ist Rang $(B) = m$ und ist die symmetrische Matrix A auf Kern (B) positiv definit, so ist $T'(x^*, v^*)$ nichtsingulär. Denn in diesem Falle folgt aus

$$T'(x^*, v^*) \begin{pmatrix} p \\ q \end{pmatrix} = 0 \qquad \text{bzw.} \qquad \begin{aligned} Ap + B^T q &= 0, \\ Bp &= 0 \end{aligned}$$

durch Multiplikation der ersten Gleichung mit p^T von links unter Berücksichtigung der zweiten Gleichung, dass $p^T A p = 0$. Da $p \in$ Kern (A) und A auf Kern (B) positiv definit ist, ist $p = 0$. Aus $B^T q = 0$ und Rang $(B) = m$ folgt auch $q = 0$.

8. Als Hoffman-Theorem (siehe A. J. Hoffman (1952)[6]) wollen wir die folgende Aussage verstehen (auch wenn sie nicht ganz mit der Originalversion übereinstimmt). Hierbei

[6] Hoffman, A. J., "On approximate solutions of systems of linear inequalities." J. Res. Natl. Bur. Standards, 49 (1952), pp. 263–265.

benutzen wir die folgende Bezeichnung: Für einen Vektor $y \in \mathbb{R}^l$ sei y_+ die Projektion von y auf den nichtnegativen Orthanten, also $(y_+)_i = \max(y_i, 0)$.

Sei

$$P := \{x \in \mathbb{R}^n : Ax \le b,\ Cx = d\} \neq \emptyset.$$

Hierbei seien $A \in \mathbb{R}^{l\times n}$, $b \in \mathbb{R}^l$, $C \in \mathbb{R}^{m\times n}$, $d \in \mathbb{R}^m$. Dann existiert eine Konstante $c_0 = c_0(A, C) > 0$ derart, daß

$$\operatorname{dist}(z, P) := \inf_{x\in P} \|z - x\| \le c_0 \left\| \begin{pmatrix} (Az-b)_+ \\ Cz - d \end{pmatrix} \right\| \qquad \text{für alle } z \in \mathbb{R}^n.$$

Hierbei sei $\|\cdot\|$ jeweils die euklidische Norm auf dem entsprechenden Raum.

Lösung: Für eine Indexmenge $I \subset \{1, \dots, l\}$ seien $A_I \in \mathbb{R}^{\#(I)\times n}$ und $b_I \in \mathbb{R}^{\#(I)}$ in naheliegender Weise definiert. Wir beweisen zunächst die folgende Hilfsaussage:

- Sei $I \subset \{1, \dots, l\}$, $N_I := \{x \in \mathbb{R}^n : A_I x \ge 0,\ Cx = 0\}$. Mit N_I^+ werde der zu N_I duale Kegel bezeichnet. Dann existiert eine Konstante $d_I > 0$ mit

$$d_I \|y\| \le \left\| \begin{pmatrix} (A_I y)_+ \\ Cy \end{pmatrix} \right\| \qquad \text{für alle } y \in N_I^+.$$

Um dies einzusehen, können wir zunächst annehmen, dass $N_I^+ \neq \{0\}$ (bzw. N_I echter Kegel im $\mathbb{R}^n$), da andernfalls die Aussage trivial ist. Man definiere

$$d_I := \min_{y\in N_I^+,\, \|y\|=1} \left\| \begin{pmatrix} (A_I y)_+ \\ Cy \end{pmatrix} \right\|.$$

Es ist $d_I > 0$, denn andernfalls existiert ein $y \neq 0$ mit $-y \in N_I$ und $y \in N_I^+$, was $-\|y\|^2 \ge 0$ implizieren und damit den Widerspruch $y = 0$ ergeben würde. Die angegebene Konstante d_I tut offenbar das verlangte.

Bei gegebenem $z \in \mathbb{R}^n$ betrachte man die quadratische Optimierungsaufgabe

$$\text{Minimiere} \quad \tfrac{1}{2}\|x - z\|^2, \quad x \in P.$$

Die eindeutige Lösung $x(z) \in P$ ist die Projektion von z auf P und nach Kuhn-Tucker charakterisiert durch die Existenz von Vektoren $u(z) \in \mathbb{R}^l$ und $v(z) \in \mathbb{R}^m$ mit

$$u(z) \ge 0, \qquad x(z) - z + A^T u(z) + C^T v(z) = 0, \qquad u(z)^T (Ax(z) - b) = 0.$$

Mit $I(z) \subset \{1, \dots, l\}$ werde die Indexmenge der in $x(z)$ aktiven Ungleichungsrestriktionen bezeichnet. Es ist also

$$u_{I(z)} \ge 0, \qquad x(z) - z + A_{I(z)}^T u_{I(z)} + C^T v(z) = 0.$$

Um die obige Hilfsaussage benutzen zu können, überlegen wir uns, dass $z - x(z) \in N_{I(z)}^+$. Denn für ein beliebiges $x \in N_{I(z)}$ (also $A_{I(z)} x \ge 0$ und $Cx = 0$) ist

$$x^T(z - x(z)) = x^T [A_{I(z)}^T u_{I(z)} + C^T v(z)] = \underbrace{(A_{I(z)} x)}_{\ge 0}{}^T \underbrace{u_{I(z)}}_{\ge 0} + \underbrace{(Cx)}_{=0}{}^T v(z) \ge 0.$$

Mit obiger Hilfsaussage ist daher

$$\begin{aligned}\left\|\begin{pmatrix}(Az-b)_+\\ Cz-d\end{pmatrix}\right\| &\geq \left\|\begin{pmatrix}(A_{I(z)}z-b_{I(z)})_+\\ Cz-d\end{pmatrix}\right\| \\ &= \left\|\begin{pmatrix}(A_{I(z)}(z-x(z)))_+\\ C(z-x(z))\end{pmatrix}\right\| \\ &\geq d_{I(z)}\,\|z-x(z)\| \\ &= d_{I(z)}\,\mathrm{dist}(z,P) \\ &\geq \delta\,\mathrm{dist}(z,P),\end{aligned}$$

wobei

$$\delta := \min_{I\subset\{1,\ldots,m\}} d_I.$$

Mit $c_0 := 1/\delta$ ist das Hoffman-Theorem bewiesen.

9. Mit Hilfe des Hoffman-Theorems zeige man: Ist $A \in \mathbb{R}^{l\times n}$, so existiert eine Konstante $c_0 = c_0(A) > 0$ derart, dass es zu jedem $b \in \mathrm{Bild}\,(A)$ ein $x^* \in \mathbb{R}^n$ mit $Ax^* = b$ und $\|x^*\| \leq c_0\,\|b\|$ gibt.

   Lösung: Mit vorgegebenem $b \in \mathrm{Bild}\,(A)$ sei $P_b := \{x \in \mathbb{R}^n : Ax = b\}$. Wegen des Hoffman-Theorems existiert eine Konstante $c_0 = c_0(A)$ (die also nicht von b abhängt) mit

   $$\mathrm{dist}(z,P_b) \leq c_0\,\|Az-b\| \qquad \text{für alle } z \in \mathbb{R}^n \text{ und alle } b \in \mathrm{Bild}\,(A).$$

   Setzt man hier nun $z := 0$, so erhält man offenbar die Behauptung.

10. Mit Hilfe des Hoffman-Theorems zeige man: Gegeben sei das lineare Programm

    $$\text{(P)} \qquad \text{Minimiere} \quad f(x) := c^T x, \quad x \in M.$$

    Hierbei sei $c \in \mathbb{R}^n$, $M \subset \mathbb{R}^n$ ein nichtleerer Polyeder und $\inf\,(\mathrm{P}) > -\infty$, daher die Menge M_{opt} der Lösungen von (P) nichtleer. Dann existiert eine Konstante $c_0 > 0$ derart, dass

    $$\mathrm{dist}(x,M_{\mathrm{opt}}) \leq c_0[f(x) - \min\,(\mathrm{P})] \qquad \text{für alle } x \in M.$$

    Hinweis: Man beachte, dass $M_{\mathrm{opt}} = M \cap \{x^* \in \mathbb{R}^n : c^T x^* - \min\,(\mathrm{P}) = 0\}$.

    Lösung: Wir nehmen an, der Polyeder M habe die Darstellung $M = \{x \in \mathbb{R}^n : Ax \leq b\}$ mit $A \in \mathbb{R}^{l\times n}$, $b \in \mathbb{R}^l$. Da (P) lösbar, ist $M_{\mathrm{opt}} = \{x \in \mathbb{R}^n : Ax \leq b,\ c^T x = \min\,(\mathrm{P})\}$ nichtleer. Das Hoffman-Theorem liefert die Existenz einer Konstanten $c_0 = c_0(A,c)$ mit

    $$\mathrm{dist}(x,M_{\mathrm{opt}}) \leq c_0 \left\|\begin{pmatrix}(Ax-b)_+\\ c^T x - \min\,(\mathrm{P})\end{pmatrix}\right\| \qquad \text{für alle } x \in \mathbb{R}^n.$$

    Insbesondere ist

    $$\mathrm{dist}(x,M_{\mathrm{opt}}) \leq c_0 \left\|\begin{pmatrix}0\\ c^T x - \min\,(\mathrm{P})\end{pmatrix}\right\| = c_0[c^T x - \min\,(\mathrm{P})] \qquad \text{für alle } x \in M.$$

    Damit ist die Aufgabe gelöst.

11. Gegeben sei das quadratische Programm

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := c^T x + \tfrac{1}{2} x^T Q x, \quad x \in M,$$

wobei $c \in \mathbb{R}^n$, $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit, $M \subset \mathbb{R}^n$ ein nichtleerer Polyeder und $\inf(\text{P}) > -\infty$. Die dann nichtleere Menge der Lösungen von (P) werde mit M_{opt} bezeichnet. Man zeige die Existenz einer Konstanten $c > 0$ mit

$$\operatorname{dist}(x, M_{\text{opt}}) \le c \left[f(x) - \min(\text{P}) + \sqrt{f(x) - \min(\text{P})} \right] \qquad \text{für alle } x \in M.$$

Hinweis: Der Polyeder M habe die Darstellung $M = \{x \in \mathbb{R}^n : Ax \le b\}$, wobei $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$. Eine Lösung $x_0^* \in M$ von (P) ist charakterisiert durch die Existenz eines Vektors $u_0^* \in \mathbb{R}^m$ mit

$$u_0^* \ge 0, \qquad c + Qx_0^* + A^T u_0^* = 0, \qquad (u_0^*)^T (b - Ax_0^*) = 0.$$

Man zeige, dass die Menge M_{opt} der Lösungen von (P) sich darstellen lässt als

$$M_{\text{opt}} = \{x^* \in \mathbb{R}^n : (b - Ax^*)^T u_0^* = 0,\ Qx^* = Qx_0^*,\ Ax^* \le b\}$$

und wende das Hoffman-Theorem an. (Ähnliche Ergebnisse findet man bei W. Li (1995)[7].)

Lösung: Sei $x_0^* \in M_{\text{opt}}$ eine spezielle Lösung von (P) und u_0^* ein zugehöriger Lagrange-Multiplikator. Ist $x \in M$ beliebig, so ist

$$\begin{aligned} f(x) &= f(x_0^*) + (c + Qx_0^*)^T (x - x_0^*) + \frac{1}{2}(x - x_0^*)^T Q (x - x_0^*) \\ &= f(x_0^*) - (A^T u_0^*)^T (x - x_0^*) + \frac{1}{2}(x - x_0^*)^T Q (x - x_0^*) \\ &= f(x_0^*) + \underbrace{(u_0^*)^T (b - Ax)}_{\ge 0} + \underbrace{(u_0^*)^T (Ax_0^* - b)}_{=0} + \frac{1}{2} \underbrace{(x - x_0^*)^T Q (x - x_0^*)}_{\ge 0} \\ &\ge f(x_0^*). \end{aligned}$$

Ist daher $x^* \in M_{\text{opt}}$ eine weitere Lösung, so ist

$$(u_0^*)^T (b - Ax^*) = 0, \qquad Qx^* = Qx_0^*.$$

Hieraus erkennt man, dass die Menge der Lösungen M_{opt} von (P) gegeben ist durch

$$M_{\text{opt}} = \{x^* \in \mathbb{R}^n : (b - Ax^*)^T u_0^* = 0,\ Qx^* = Qx_0^*,\ Ax^* \le b\}.$$

Wegen des Hoffman-Lemmas existiert eine Konstante $c_0 > 0$ mit

$$\begin{aligned} \operatorname{dist}(x, M_{\text{opt}}) &\le c_0 \left[|(b - Ax)^T u_0^*|^2 + \|Q(x - x_0^*)\|^2 \right]^{1/2} \\ &\le c_0 \left[|(b - Ax)^T u_0^*| + \|Q(x - x_0^*)\| \right] \qquad \text{für alle } x \in M. \end{aligned}$$

Für $x \in M$ ist offenbar

$$0 \le (b - Ax)^T u_0^* \le f(x) - f(x_0^*) = f(x) - \min(\text{P}),$$

[7] Li, W. (1995) "Error bounds for piecewise convex quadratic programs and applications." SIAM J. Control and Optimization 33, 1510– 1529.

wie man an obiger Entwicklung unschwer erkennt. Weiter benutzen wir, dass eine Konstante $\theta > 0$ mit

$$\theta \, \|Qy\|^2 \le y^T Qy \qquad \text{für alle } y \in \mathbb{R}^n$$

existiert. Für $x \in M$ ist daher

$$\|Q(x - x_0^*)\| \le \frac{1}{\sqrt{\theta}} [(x - x_0^*)^T Q(x - x_0^*)]^{1/2} \le \sqrt{\frac{2}{\theta}} \sqrt{f(x) - \min(\mathrm{P})}.$$

Damit erhalten wir mit einer hinreichend großen Konstanten c, dass

$$\begin{aligned} \operatorname{dist}(x, M_{\mathrm{opt}}) &\le c_0 \left[f(x) - \min(\mathrm{P}) + \sqrt{2/\theta} \sqrt{f(x) - \min(\mathrm{P})} \right] \\ &\le c \left[f(x) - \min(\mathrm{P}) + \sqrt{f(x) - \min(\mathrm{P})} \right] \qquad \text{für alle } x \in M. \end{aligned}$$

Ist $x \in M$ beliebig, so ist

$$\begin{aligned} f(x) &= f(x_0^*) + (c + Qx_0^*)^T (x - x_0^*) + \frac{1}{2}(x - x_0^*)^T Q(x - x_0^*) \\ &= f(x_0^*) - (A^T y_0^*)^T (x - x_0^*) + \frac{1}{2}(x - x_0^*)^T Q(x - x_0^*) \\ &= f(x_0^*) + \underbrace{(y_0^*)^T (b - Ax)}_{\ge 0} + \underbrace{(y_0^*)^T (Ax_0^* - b)}_{=0} + \frac{1}{2} \underbrace{(x - x_0^*)^T Q(x - x_0^*)}_{\ge 0} \\ &\ge f(x_0^*). \end{aligned}$$

Ist daher $x^* \in M^*$ eine weitere Lösung, so ist

$$(y_0^*)^T (b - Ax^*) = 0, \qquad Qx^* = Qx_0^*.$$

Hieraus erkennt man, dass die Menge der Lösungen M^* von (P) gegeben ist durch

$$M^* = \{x^* \in \mathbb{R}^n : (b - Ax^*)^T y_0^* = 0,\ Qx^* = Qx_0^*,\ Ax^* \le b\}.$$

Wegen des Hoffman-Lemmas existiert eine Konstante $c_0 > 0$ mit

$$\begin{aligned} \operatorname{dist}(x, M^*) &\le c_0 \left[|(b - Ax)^T y_0^*|^2 + \|Q(x - x_0^*)\|^2 \right]^{1/2} \\ &\le c_0 \left[|(b - Ax)^T y_0^*| + \|Q(x - x_0^*)\| \right] \qquad \text{für alle } x \in M. \end{aligned}$$

Für $x \in M$ ist offenbar

$$0 \le (b - Ax)^T y_0^* \le f(x) - f(x_0^*) = f(x) - \min(\mathrm{P}),$$

wie man an obiger Entwicklung unschwer erkennt. Weiter benutzen wir, dass eine Konstante $\theta > 0$ mit

$$\theta \, \|Qy\|^2 \le y^T Qy \qquad \text{für alle } y \in \mathbb{R}^n$$

existiert. Für $x \in M$ ist daher

$$\|Q(x - x_0^*)\| \le \frac{1}{\sqrt{\theta}} [(x - x_0^*)^T Q(x - x_0^*)]^{1/2} \le \sqrt{\frac{2}{\theta}} \sqrt{f(x) - \min(\mathrm{P})}.$$

Damit erhalten wir mit einer hinreichend großen Konstanten c, dass

$$\begin{aligned} \operatorname{dist}(x, M^*) &\le c_0 \left[f(x) - \min(\mathrm{P}) + \sqrt{2/\theta} \sqrt{f(x) - \min(\mathrm{P})} \right] \\ &\le c \left[f(x) - \min(\mathrm{P}) + \sqrt{f(x) - \min(\mathrm{P})} \right] \qquad \text{für alle } x \in M. \end{aligned}$$

12. Es sollen 400 m^3 Kies von einem Ort zu einem anderen transportiert werden. Dies geschehe in einer (nach oben!) offenen Box der Länge x_1, der Breite x_2 und der Höhe x_3 (jeweils in Metern gemessen). Der Boden und die beiden Längsseiten müssen aus einem Material hergestellt werden, das zwar nichts kostet, von dem aber nur 4 m^2 zur Verfügung steht. Das Material für die beiden Querseiten kostet 200 Euro pro m^2. Ein Transport der Box kostet 1 Euro. Wie hat man die Box zu konstruieren?

    Man stelle also die zugehörige Optimierungsaufgabe auf, bestimme die zulässigen Punkte, in denen die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind und überprüfe diese mit Hilfe der hinreichenden Optimalitätsbedingungen zweiter Ordnung auf Optimalität.

    Lösung: Damit in die Box überhaupt etwas hinein getan werden kann, hat man die Nebenbedingungen $x_1, x_2, x_3 > 0$. Dadurch, dass von dem Material für den Boden und die beiden Längsseiten nur 4 m^2 zur Verfügung stehen, hat man noch die Restriktion $x_1x_2 + 2x_1x_3 \leq 4$. Die Herstellung einer Box der Länge x_1, der Breite x_2 und der Höhe x_3 (in Metern) kostet $400x_2x_3$ (in Euro), als Transportkosten hat man ferner $400/(x_1x_2x_3)$ (in Euro). Insgesamt erhält man die Optimierungsaufgabe

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad f(x) := \dfrac{1}{x_1x_2x_3} + x_2x_3 \quad \text{unter den Nebenbedingungen} \\ \qquad\qquad g(x) := x_1x_2 + 2x_1x_3 \leq 4, \qquad x_1, x_2, x_3 > 0. \end{cases}$$

    Wir wollen den Satz von Kuhn-Tucker anwenden und nehmen hierzu an, x^* sei eine lokale Lösung von (P). Die Arrow-Hurwicz-Uzawa Constraint Qualification ist erfüllt (nimm als $\hat{p}$ einen beliebigen Vektor mit negativen Komponenten). Daher existiert ein $u^* \in \mathbb{R}$ mit

$$u^* \geq 0, \quad \begin{pmatrix} -1/((x_1^*)^2x_2^*x_3^*) \\ -1/(x_1^*(x_2^*)^2x_3^*) + x_3^* \\ -1/(x_1^*x_2^*(x_3^*)^2) + x_2^* \end{pmatrix} + u^* \begin{pmatrix} x_2^* + 2x_3^* \\ x_1^* \\ 2x_1^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

    und der Gleichgewichtsbedingung

$$u^*(x_1^*x_2^* + 2x_1^*x_3^* - 4) = 0.$$

    Wie man durch Inspektion erkennt, ist notwendigerweise $u^* > 0$ und damit die Ungleichungsrestriktion aktiv. Mit $A^* := 1/(x_1^*x_2^*x_3^*)$ erhalten wir

$$\frac{1}{u^*} = \frac{(x_2^* + 2x_3^*)x_1^*}{A^*} = \frac{x_1^*x_2^*}{A^* - x_2^*x_3^*} = \frac{2x_1^*x_3^*}{A^* - x_2^*x_3^*}$$

    und

$$x_1^*x_2^* + 2x_1^*x_3^* = 4.$$

    Durch Inspektion folgt hieraus zunächst $x_2^* = 2x_3^*$, aus der letzten Gleichung erhält man $x_1^* = 1/x_3^*$. Also ist $A^* = 1/(2x_3^*)$, die erste Gleichung (die beiden anderen sind schon benutzt worden)

$$\frac{(x_2^* + 2x_3^*)x_1^*}{A^*} = \frac{x_1^*x_2^*}{A^* - x_2^*x_3^*}$$

    besagt nun, dass

$$8x_3^* = \frac{2}{1/(2x_3^*) - 2(x_3^*)^2},$$

was auf $x_3^* = \frac{1}{2}$ führt. Der einzige zulässige Punkt, in dem die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind, ist daher

$$x^* = (2, 1, \tfrac{1}{2})^T,$$

der zugehörige Multiplikator ist

$$u^* = \tfrac{1}{4}.$$

"Sicherheitshalber" überprüfen wir diesen Lösungskandidaten mit Hilfe der hinreichenden Optimalitätbedingungen zweiter Ordnung. Hiernach ist nachzuprüfen, ob die Implikation (man beachte, dass die Ungleichungsrestriktion aktiv und der zugehörige Lagrange-Multiplikator positiv ist)

$$\nabla g(x^*)^T p = 0, \quad p \neq 0 \Longrightarrow p^T[\nabla^2 f(x^*) + u^* \nabla^2 g(x^*)]p > 0$$

gilt. Einsetzen liefert die gleichwertige Aussage

$$\begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}^T p = 0, \quad p \neq 0 \Longrightarrow p^T \begin{pmatrix} \frac{1}{2} & \frac{3}{4} & \frac{3}{2} \\ \frac{3}{4} & 2 & 3 \\ \frac{3}{2} & 3 & 8 \end{pmatrix} p > 0.$$

Nun ist die rechts stehende Matrix selber schon positiv definit, daher gilt die Implikation erst recht. Damit ist nachgewiesen, dass x^* eine lokale Lösung von (P) ist.

13. Man bestimme die Lösung von

$$\text{(P)} \qquad \text{Maximiere} \quad f(x) := \prod_{j=1}^{n} x_j \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0,\ e^T x = 1\},$$

wobei e einmal wieder den Vektor im $\mathbb{R}^n$ bezeichnet, dessen Komponenten sämtlich gleich 1 sind. Hiermit beweise man die Ungleichung vom geometrisch-arithmetischen Mittel, dass also für alle $x \in \mathbb{R}^n$ mit $x \geq 0$ gilt

$$\left(\prod_{j=1}^{n} x_j\right)^{1/n} \leq \frac{1}{n} \sum_{j=1}^{n} x_j.$$

Hierbei tritt Gleichheit genau dann ein, wenn $x = \alpha e$ mit $\alpha \geq 0$.

Lösung: Die Existenz einer Lösung x^* von (P) ist klar, da M kompakt und f auf M stetig ist. Natürlich ist notwendigerweise eine Lösung von (P) auch eine Lösung von

$$\text{Maximiere} \quad f(x) := \prod_{j=1}^{n} x_j \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x > 0,\ e^T x = 1\}.$$

Eine Anwendung des Satzes von Kuhn-Tucker liefert die Existenz von $v^* \in \mathbb{R}^n$ mit

$$-\prod_{\substack{j=1 \\ j \neq i}}^{n} x_j^* + v^* = 0, \qquad i = 1, \ldots, n.$$

Hieraus folgt, dass $x_1^* = \cdots = x_n^*$, wegen $e^T x^* = 1$ ist $x^* = (1/n)e$ die Lösung von (P) und der Wert ist $\max(\text{P}) = (1/n)^n$. Zum Nachweis der Ungleichung vom geometrisch-arithmetischen Mittel sei ein Vektor $x \in \mathbb{R}^n$ mit $x \geq 0$ vorgegeben. O. B. d. A. ist $x > 0$,

da die Ungleichung andernfalls trivial ist. Setzt man $z := x/e^T x$, so ist $z \in M$ und daher

$$\prod_{j=1}^{n} z_j = \frac{\prod_{j=1}^{n} x_j}{(\sum_{j=1}^{n} x_j)^n} \leq \frac{1}{n^n},$$

was genau auf die Ungleichung vom geometrisch-arithmetischen Mittel führt. Angenommen, in der Ungleichung vom geometrisch-arithmetischen Mittel trete für ein $x \in \mathbb{R}^n$ mit $x \geq 0$ Gleichheit auf. O. B. d. A. sei $x > 0$. Dann ist $z := x/e^T x$ die Lösung von (P) und daher $x = (e^T x/n)e$, was zu zeigen war.

14. Bei gegebenem $\alpha \in (0,1)$ und $r := \sqrt{n/(n-1)}$ betrachte man die Optimierungsaufgabe

$$\text{(P)} \quad \text{Minimiere} \quad f(x) := \prod_{j=1}^{n} x_j \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : e^T x = n,\ \|x - e\|_2 \leq \alpha r\}.$$

Hierbei sei e wieder der Vektor des $\mathbb{R}^n$, dessen Komponenten alle gleich 1 sind. Man zeige:

(a) (P) besitzt eine Lösung x^* und es ist notwendig $x^* > 0$ und $\|x^* - e\|_2 = \alpha r$.

(b) Eine Lösung x^* von (P) besitzt genau zwei verschiedene Komponenten. Bis auf die Reihenfolge der Komponenten kommt als Lösungskandidat also nur ein Vektor $x^{(m)}$ in Frage, dessen erste m Komponenten übereinstimmen und kleiner sind als die restlichen (ebenfalls gleichen) $(n-m)$ Komponenten. Man zeige, dass $x^* = x^{(1)}$ bis auf die Reihenfolge der Komponenten die Lösung von (P) ist.

(c) Es ist

$$\prod_{j=1}^{n} x_j \geq (1-\alpha)\left(1 + \frac{\alpha}{n-1}\right)^{n-1} \qquad \text{für alle } x \in M.$$

Hinweis: Diese Aufgabe spielt im Zusammenhang mit der Konvergenzanalyse des Karmarkar-Verfahrens eine Rolle, siehe z. B. J. Werner (1992, S. 135 ff.).

Lösung: Die Existenz einer Lösung x^* von (P) ist wegen der Kompaktheit von M und der Stetigkeit von f trivial. Es ist $x^* > 0$, da sogar jedes Element aus M im positiven Orthanten liegt. Denn sei $i \in \{1, \ldots, n\}$ fest. Wegen $e^T x = n$ ist dann

$$\begin{aligned}
(x_i - 1)^2 &= \left(-\sum_{\substack{j=1 \\ j \neq i}}^{n} (x_j - 1)\right)^2 \\
&\leq (n-1) \sum_{\substack{j=1 \\ j \neq i}}^{n} (x_j - 1)^2 \\
&\leq (n-1)[\alpha^2 r^2 - (x_i - 1)^2] \\
&< (n-1)\left[\frac{n}{n-1} - (x_i - 1)^2\right]
\end{aligned}$$

und folglich $(x_i - 1)^2 < 1$ und daher $x_i > 0$. Nun wollen wir zeigen, dass die Ungleichungsrestriktion $\|x - e\| \leq \alpha r$ für eine Lösung x^* aktiv ist. Wir nehmen an, es sei

$x^* \in M$ eine Lösung von (P) mit $\|x^* - e\| < \alpha r$ und führen dies zu einem Widerspruch. Da x^* eine Lösung von (P) ist, ist

$$f(x^*) \le (1-\alpha)\left(1 + \frac{\alpha}{n-1}\right)^{n-1} < 1.$$

Hierbei haben wir benutzt, dass der Vektor, dessen erste Komponente gleich $1-\alpha$, die restlichen gleich $1 + \alpha/(n-1)$ sind, für (P) zulässig ist. Definiert man $x^*(t) := x^* + t(e - x^*)$, so ist $x^*(t) \in M$ für alle hinreichend kleinen $|t|$ und

$$\frac{d}{dt} f(x^*(t))_{t=0} = \sum_{i=1}^{n} \left(\prod_{\substack{j=1 \\ j \ne i}}^{n} x_j^* \right) (1 - x_i^*) = f(x^*) \left(\sum_{i=1}^{n} \frac{1}{x_i^*} - n \right) > 0,$$

ein Widerspruch zur Optimalität von x^*. Die letzte Ungleichung folgt hierbei mit Hilfe der Ungleichung vom geometrisch-arithmetischen Mittel aus

$$1 < \left(\frac{1}{f(x^*)} \right)^{1/n} = \left(\prod_{i=1}^{n} \frac{1}{x_i^*} \right)^{1/n} \le \frac{1}{n} \sum_{i=1} \frac{1}{x_i^*}.$$

Damit ist der erste Teil der Aufgabe gelöst.

Nun nehmen wir an, x^* sei eine Lösung von (P). Da die Ungleichungsrestriktion in einer Lösung notwendigerweise aktiv ist, ist x^* auch Lösung von

$$\text{Minimiere} \quad f(x) := \prod_{j=1}^{n} x_j \quad \text{unter der Nebenbedingung} \quad h(x) = 0,$$

wobei

$$h(x) := \begin{pmatrix} e^T x - n \\ \|x\|_2^2 - (n + \alpha^2 r^2) \end{pmatrix}.$$

Die Constraint Qualification zur Anwendung der Lagrangeschen Multiplikatorenregel ist wegen

$$\text{Rang } h'(x^*) = \text{Rang} \begin{pmatrix} e^T \\ 2(x^*)^T \end{pmatrix} = 2$$

erfüllt, denn x^* kann kein Vielfaches von e sein. Daher existieren reelle Zahlen v_1^* und v_2^* mit

$$\prod_{\substack{j=1 \\ j \ne i}}^{n} x_j^* + v_1^* + v_2^* 2 x_i^* = 0, \qquad i = 1, \ldots, n.$$

Folglich ist

$$f(x^*) + v_1^* x_i^* + 2 v_2^* (x_i^*)^2 = 0, \qquad i = 1, \ldots, n.$$

Also genügen die Komponenten x_i^*, $i = 1, \ldots, n$, einer Lösung x^* ein und derselben quadratischen Gleichung, d. h. die Komponenten von x^* nehmen höchstens zwei und wegen $x^* \ne e$ genau zwei verschiedene Werte an. Da es auf die Reihenfolge der Komponenten nicht ankommt, kann angenommen werden, dass die ersten m Komponenten

und die übrigen $n-m$ Komponenten jeweils gleich sind, wobei $m \in \{1,\dots,n-1\}$. Für ein solches m mache man für den zugehörigen Lösungskandidaten den Ansatz

$$x_j^{(m)} = \begin{cases} 1-\alpha u^{(m)}, & j=1,\dots,m, \\ 1+\alpha v^{(m)}, & j=m+1,\dots,n. \end{cases}$$

Der Lösungskandidat $x^{(m)}$ muss den Nebenbedingungen genügen. Aus

$$n = e^T x^{(m)} = m(1-\alpha u^{(m)}) + (n-m)(1+\alpha v^{(m)})$$

folgt

$$u^{(m)} = \frac{n-m}{m} v^{(m)}.$$

Dies benutzend erhält man aus der zweiten Nebenbedingung

$$\begin{aligned} n + \alpha^2 \frac{n}{n-1} &= \|x^{(m)}\|_2^2 \\ &= m(1-\alpha u^{(m)})^2 + (n-m)(1+\alpha v^{(m)})^2 \\ &= n + \alpha^2 \frac{(n-m)n}{m} (v^{(m)})^2. \end{aligned}$$

Damit erhält man, dass die Lösungskandidaten $x^{(m)}$, also zulässige Lösungen mit genau zwei verschiedenen Komponenten für $m = 1,\dots,n-1$ genau (d. h. bis auf die Reihenfolge der Komponenten) durch

$$x_j^{(m)} = \begin{cases} 1-\alpha\sqrt{\dfrac{n-m}{(n-1)m}}, & j=1,\dots,m, \\ 1+\alpha\sqrt{\dfrac{m}{(n-1)(n-m)}}, & j=m+1,\dots,n \end{cases}$$

gegeben sind. Nun ist

$$\begin{aligned} f(x^{(m)}) &= \left(1-\alpha\sqrt{\frac{n-m}{(n-1)m}}\right)^m \left(1+\alpha\sqrt{\frac{m}{(n-1)(n-m)}}\right)^{n-m} \\ &\geq f(x^{(1)}) \\ &= (1-\alpha)\left(1+\frac{\alpha}{n-1}\right)^{n-1}. \end{aligned}$$

Um die hier auftretende Ungleichung einzusehen, überlegen wir uns, dass

$$f(x^{(1)}) \leq \cdots \leq f(x^{(m)}) \leq f(x^{(m+1)}) \leq \cdots \leq f(x^{(n-1)}).$$

Hierzu definieren wir

$$g(m) := \log f(x^{(m)}) = m \log y(m) + (n-m) \log z(m)$$

mit

$$y(m) := 1 - \frac{\alpha}{\sqrt{n-1}}\sqrt{\frac{n-m}{m}}, \qquad z(m) := 1 + \frac{\alpha}{\sqrt{n-1}}\sqrt{\frac{m}{n-m}}$$

und zeigen die Monotonie von $g(\cdot)$ auf $[1, n-1]$. Nach leichter Rechnung stellt man fest, dass

$$\begin{aligned} g'(m) &= \log y(m) - \log z(m) + \frac{m}{y(m)} y'(m) + \frac{n-m}{z(m)} z'(m) \\ &= \log y(m) - \log z(m) + \frac{1}{2}\left(\frac{1}{y(m)} + \frac{1}{z(m)}\right) \frac{\alpha}{\sqrt{n-1}} \frac{n}{\sqrt{m(n-m)}} \\ &= \log y(m) - \log z(m) + \frac{1}{2}\left(\frac{1}{y(m)} + \frac{1}{z(m)}\right)[z(m) - y(m)] \\ &\geq 0. \end{aligned}$$

Hierbei haben wir ausgenutzt, dass

$$\log z(m) - \log y(m) \leq \frac{1}{2}\left(\frac{1}{y(m)} + \frac{1}{z(m)}\right)[z(m) - y(m)]$$

wegen $0 < y(m) \leq z(m)$, was wiederum leicht aus

$$\log t \leq \frac{1}{2}\left(t - \frac{1}{t}\right) \qquad \text{für alle } t \geq 1$$

folgt. Damit ist schließlich die Aufgabe vollständig gelöst.

6.3 Aufgaben in Kapitel 3

6.3.1 Aufgaben in Abschnitt 3.1

1. Sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und $A \in \mathbb{R}^{m\times n}$ eine Matrix mit Rang $(A) = m$. Man zeige, dass dann die Matrix

$$K := \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix}$$

nichtsingulär ist. Ferner zeige man, dass mit

$$N := (AQ^{-1}A^T)^{-1}AQ^{-1}, \qquad H := Q^{-1}(I - A^T N)$$

die Inverse K^{-1} gegeben ist durch

$$K^{-1} = \begin{pmatrix} H & N^T \\ N & -NQN^T \end{pmatrix}.$$

Hinweis: Diese Aussage findet man schon bei R. Fletcher (1971).

Lösung: Angenommen, es ist

$$K \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

bzw.

$$Qu + A^T v = 0, \qquad Au = 0.$$

Multipliziert man die erste Gleichung von links mit u^T und berücksichtigt man die zweite, so folgt $u^TQu = 0$ und damit $u = 0$. Aus $A^Tv = 0$ folgt wegen Rang $(A) = m$, dass auch $v = 0$ und damit die Nichtsingularität von K. Weiter ist

$$\begin{pmatrix} H & N^T \\ N & -NQN^T \end{pmatrix} \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} = \begin{pmatrix} HQ + N^TA & HA^T \\ NQ - NQN^TA & NA^T \end{pmatrix}.$$

Nun berücksichtige man, dass offensichtlich $NA^T = I$ und daher $HA^T = 0$. Außerdem ist

$$HQ + N^TA = (I - Q^{-1}A^T(AQ^{-1}A^T)^{-1}A) + Q^{-1}A^T(AQ^{-1}A^T)^{-1}A = I$$

und

$$NQ - NQN^TA = (AQ^{-1}A^T)^{-1}A - (AQ^{-1}A^T)^{-1}AQ^{-1}A^T(AQ^{-1}A^T)^{-1}A = 0.$$

Damit ist die Aufgabe gelöst.

2. Sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch, die Matrix $A \in \mathbb{R}^{m\times n}$ habe vollen Zeilenrang, d. h. es sei Rang $(A) = m$. Hiermit definiere man die Matrix

$$K := \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}$$

und zeige:

   (a) Ist Q auf Kern (A) positiv definit. ist also $p^TQp > 0$ für alle $p \in \mathbb{R}^n \setminus \{0\}$ mit $Ap = 0$, so ist K nichtsingulär.

   (b) Ist Q positiv semidefinit und K nichtsingulär, so ist Q auf Kern (A) positiv definit.

Lösung: Der erste Teil der Aufgabe ist praktisch zu Beginn der vorigen Aufgabe gelöst worden. Genau wie dort zeigt man, dass der Kern der Matrix K nur aus dem Nullvektor besteht. Sei daher umgekehrt Q (auf dem $\mathbb{R}^n$) positiv semidefinit und K nichtsingulär. Sei $p \in$ Kern (A) und $p^TQp = 0$. Dann ist auch $Qp = 0$ (Beweis?) und folglich

$$K \begin{pmatrix} p \\ 0 \end{pmatrix} = 0.$$

Da K nach Voraussetzung nichtsingulär ist, ist $p = 0$. Daher ist Q auf Kern (A) positiv definit.

3. Sei $I \subset \{1, \ldots, m\}$ mit $q := \#(I)$ eine (nichtleere) Indexmenge, $r \in \{1, \ldots, m\} \setminus I$ und $\{a_i\}_{i\in I\cup\{r\}}$ linear unabhängig. Die Matrizen

$$N_I := (A_IQ^{-1}A_I^T)^{-1}A_IQ^{-1} \in \mathbb{R}^{q\times n}, \qquad H_I := Q^{-1}(I - A_I^TN_I) \in \mathbb{R}^{n\times n}$$

und die Vektoren

$$z := H_Ia_r \in \mathbb{R}^n, \qquad r_I := N_Ia_r \in \mathbb{R}^q$$

seien bekannt. Man zeige, daß

$$N_{I\cup\{r\}} = \begin{pmatrix} N_I - \dfrac{r_Iz^T}{a_r^Tz} \\ \dfrac{z^T}{a_r^Tz} \end{pmatrix}, \qquad H_{I\cup\{r\}} = H_I - \frac{zz^T}{a_r^Tz}.$$

Lösung: Wir überlegen uns, wie man die Inverse der symmetrischen Matrix

$$\begin{pmatrix} A & a \\ a^T & \alpha \end{pmatrix}$$

berechnen kann, wenn man A^{-1} kennt. Aus

$$\begin{pmatrix} A & a \\ a^T & \alpha \end{pmatrix} \begin{pmatrix} B & b \\ b^T & \beta \end{pmatrix} = \begin{pmatrix} AB + ab^T & Ab + \beta a \\ a^T B + \alpha b^T & a^T b + \alpha\beta \end{pmatrix}$$

erhält man sukzessive

$$B = A^{-1} - (A^{-1}a)b^T, \qquad b = -\beta A^{-1}a, \qquad \beta = \frac{1}{\alpha - a^T A^{-1} a}.$$

Mit

$$A := A_I Q^{-1} A_I^T, \qquad a := A_I Q^{-1} a_r, \qquad \alpha := a_r^T Q^{-1} a_r$$

wird

$$\frac{1}{\beta} = a_r^T Q^{-1} a_r - a_r^T Q^{-1} A_I^T (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} a_r = a_r^T Q^{-1}[I - A_I^T N_I] a_r = a_r^T z$$

und

$$b = -\frac{1}{a_r^T z}[(A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} a_r = -\frac{1}{a_r^T z} N_I a_r = -\frac{r_I}{a_r^T z}$$

sowie

$$B = (A_I Q^{-1} A_I)^T)^{-1} + \frac{r_I r_I^T}{a_r^T z}.$$

Daher ist

$$\begin{aligned}
N_{I\cup\{r\}} &= (A_{I\cup\{r\}} Q^{-1} A_{I\cup\{r\}}^T)^{-1} A_{I\cup\{r\}} Q^{-1} \\
&= \left[\begin{pmatrix} A_I \\ a_r^T \end{pmatrix} Q^{-1} \begin{pmatrix} A_I^T & a_r \end{pmatrix} \right]^{-1} \begin{pmatrix} A_I \\ a_r^T \end{pmatrix} Q^{-1} \\
&= \begin{pmatrix} A_I Q^{-1} A_I^T & A_I Q^{-1} a_r \\ (A_I Q^{-1} a_r)^T & a_r^T Q^{-1} a_r \end{pmatrix}^{-1} \begin{pmatrix} A_I Q^{-1} \\ a_r^T Q^{-1} \end{pmatrix} \\
&= \begin{pmatrix} (A_I Q^{-1} A_I^T)^{-1} + \dfrac{r_I r_I^T}{a_r^T z} & -\dfrac{r_I}{a_r^T z} \\ -\dfrac{r_I^T}{a_r^T z} & \dfrac{1}{a_r^T z} \end{pmatrix} \begin{pmatrix} A_I Q^{-1} \\ a_r^T Q^{-1} \end{pmatrix} \\
&= \begin{pmatrix} N_I - \dfrac{r_I z^T}{a_r^T z} \\ \dfrac{z^T}{a_r^T z} \end{pmatrix},
\end{aligned}$$

was für die erste Update-Formel zu zeigen war. Weiter ist

$$H_{I\cup\{r\}} = Q^{-1}(I - A_{I\cup\{r\}}^T N_{I\cup\{r\}})$$

$$\begin{aligned}
&= Q^{-1}\left[I - \begin{pmatrix} A_I^T & a_r \end{pmatrix}\begin{pmatrix} N_I - \dfrac{r_I z^T}{a_r^T z} \\ \dfrac{z^T}{a_r^T z} \end{pmatrix}\right] \\
&= Q^{-1}\left[I - A_I^T N_I + \frac{(A_I^T r_I - a_r)z^T}{a_r^T z}\right] \\
&= H_I - Q^{-1}(I - A_I^T N_I)a_r \frac{z^T}{a_r^T z} \\
&= H_I - \frac{zz^T}{a_r^T z}.
\end{aligned}$$

Das war zu zeigen.

4. Sei $I \subset \{1, \ldots, m\}$ (wieder sei $q := \#(I)$) eine nichtleere Indexmenge mit der Eigenschaft, dass die Vektoren $\{a_i\}_{i\in I} \subset \mathbb{R}^n$ linear unabhängig sind. Insbesondere sei also $1 \le q \le n$ und Rang $(A_I) = q$. Bekannt seien die Matrizen

$$N_I := (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} \in \mathbb{R}^{q\times n}, \qquad H_I := Q^{-1}(I - A_I^T N_I) \in \mathbb{R}^{n\times n}.$$

Ferner sei $l \in I$ vorgegeben. Man überlege sich, wie man auf effiziente Weise die analog definierten Matrizen $N_{I\setminus\{l\}}$ und $H_{I\setminus\{l\}}$ berechnen kann.

Lösung: Sei l das k-te Element von I. Definiert man die Matrix

$$T_k := \begin{pmatrix} e_1^T \\ \vdots \\ e_{k-1}^T \\ e_{k+1}^T \\ \vdots \\ e_q^T \end{pmatrix} \in \mathbb{R}^{(q-1)\times q},$$

wobei e_i den i-ten Einheitsvektor im $\mathbb{R}^q$ bedeutet, so ist $A_{I\setminus\{l\}} = T_k A_I$ und

$$T_k^T T_k = I - e_k e_k^T \in \mathbb{R}^{q\times q}, \qquad T_k T_k^T = I_{q-1} \in \mathbb{R}^{(q-1)\times(q-1)}.$$

Ist $B \in \mathbb{R}^{q\times q}$ symmetrisch und positiv definit, so rechnet man leicht nach, dass

$$(T_k B T_k^T)^{-1} = T_k\left[B^{-1} - \frac{(B^{-1}e_k)(B^{-1}e_k)^T}{e_k^T B^{-1} e_k}\right]T_k^T.$$

Folglich ist

$$(A_{I\setminus\{l\}} Q^{-1} A_{I\setminus\{l\}}^T)^{-1} = T_k\left[(A_I Q^{-1} A_I^T)^{-1} - \frac{[(A_I Q^{-1} A_I^T)^{-1} e_k][(A_I Q^{-1} A_I^T)^{-1} e_k]^T}{e_k^T (A_I Q^{-1} A_I^T)^{-1} e_k}\right]T_k^T.$$

Berücksichtigt man noch, dass $(A_I Q^{-1} A_I^T)^{-1} = N_I Q N_I^T$, so erhält man

$$N_{I\setminus\{l\}} = T_k N_I\left[I - \frac{(Q N_I^T e_k)(N_I^T e_k)^T}{(N_I^T e_k)^T Q (N_I^T e_k)}\right], \qquad H_{I\setminus\{l\}} = H_I + \frac{(N_I^T e_k)(N_I^T e_k)^T}{(N_I^T e_k)^T Q (N_I^T e_k)}.$$

5. Gegeben sei das (übliche) quadratische Programm (P) mit der symmetrischen, positiv definiten Matrix $Q \in \mathbb{R}^{n\times n}$. Das Paar (x, I) genüge den Bedingungen in Schritt (0) des Fletcher-Verfahrens. Sei (p, y_I) die eindeutige Lösung des linearen Gleichungssystems in Schritt (1). Es sei $x+p \in M$ und $y_l < 0$ für ein $l \in I \cap \{1, \ldots, m_0\}$. Wie im Verfahren von Fletcher setze man $x^+ := x + p$ und $I^+ := I \setminus \{l\}$. Ist dann p^+ die Lösung von

$$\text{Minimiere} \quad (c + Qx^+)^T z + \tfrac{1}{2}\, z^T Q z \quad \text{unter der Nebenbedingung} \quad A_{I^+} z = 0,$$

so ist $a_l^T p^+ > 0$ und $(c + Qx^+)^T p^+ = -(p^+)^T Q p^+ < 0$, insbesondere also $p^+ \neq 0$.

Lösung: Bezeichnet man mit $y_I = (y_i)_{i\in I}$ und $y^+_{I^+} = (y_i^+)_{i\in I^+}$ die Lagrange-Vektoren zu p bzw. p^+, so ist

$$c + Qx^+ = A_I^T y_I, \qquad c + Qx^+ + Qp^+ = A_{I^+}^T y^+_{I^+}.$$

Durch Subtraktion der ersten Gleichung von der zweiten erhält man

$$Qp^+ = A_{I^+}^T y^+_{I^+} - A_I^T y_I = \underbrace{-y_l}_{>0} a_l + \sum_{\substack{i\in I\\ i\neq l}} (y_i^+ - y_i) a_i.$$

Wegen der linearen Unabhängigkeit von $\{a_i\}_{i\in I}$ ist $Qp^+ \neq 0$ und daher

$$\underbrace{(p^+)^T Q p^+}_{>0} = \underbrace{-y_l}_{>0} a_l^T p^+ + \sum_{\substack{i\in I\\ i\neq l}} (y_i^+ - y_i) \underbrace{a_i^T p^+}_{=0}.$$

Folglich ist $a_l^T p^+ > 0$, aus $c+Qx^+ +Qp^+ = A^T_{I^+} y^+_{I^+}$ erhält man nach Multiplikation von links mit $(p^+)^T$ unter Berücksichtigung von $A_{I^+} p^+ = 0$ auch die letzte Behauptung.

6. Sei $I \subset \{1, \ldots, m\}$ (mit $q := \#(I)$) und $r \in \{1, \ldots, m\} \setminus I$. Die wie üblich definierten Matrizen A_I und $A_{I\cup\{r\}}$ mögen maximalen Zeilenrang q bzw. $q+1$ besitzen, die Matrix $Q \in \mathbb{R}^{n\times n}$ sei symmetrisch und auf Kern (A_I) positiv definit. Bekannt seien die orthogonale Matrix

$$Z_I = (\ \underbrace{Z_I^{(1)}}_{q}\ \ \underbrace{Z_I^{(2)}}_{n-q}\ ) \in \mathbb{R}^{n\times n},$$

die obere Dreiecksmatrix R_I, die untere Dreiecksmatrix mit Einsen in der Diagonalen L_I sowie die (positiv definite) Diagonalmatrix D_I mit

$$Z_I^{(1)T} A_I^T = R_I, \qquad A_I Z_I^{(2)} = 0, \qquad Z_I^{(2)T} Q Z_I^{(2)} = L_I D_I L_I^T.$$

Man setze $I^+ := I \cup \{r\}$ und entwickle ein effizientes, stabiles Verfahren zur Berechnung der Matrizen Z_{I^+}, R_{I^+}, L_{I^+} und D_{I^+} mit den zu Z_I, R_I, L_I bzw. D_I analogen Eigenschaften.

Lösung: Für orthogonales $Z_I = (\ Z_I^{(1)}\ \ Z_I^{(2)}\ )$ sind die beiden Bedingungen

$$Z_I^{(1)T} A_I^T = R_I, \qquad A_I Z_I^{(2)} = 0$$

äquivalent zu

$$A_I^T = Z_I \begin{pmatrix} R_I \\ 0 \end{pmatrix}.$$

Dies ist klar, wenn man bedenkt, dass letztere Aussage äquivalent zu

$$\begin{pmatrix} (Z_I^{(1)})^T A_I^T \\ (Z_I^{(2)})^T A_I^T \end{pmatrix} = Z_I^T A_I^T = \begin{pmatrix} R_I \\ 0 \end{pmatrix}$$

ist. Mit einer beliebigen orthogonalen Matrix $P \in \mathbb{R}^{(n-q)\times(n-q)}$ ist

$$\begin{aligned}
A_{I^+}^T &= \begin{pmatrix} A_I^T & a_r \end{pmatrix} \\
&= \begin{pmatrix} Z_I \begin{pmatrix} R_I \\ 0 \end{pmatrix} & a_r \end{pmatrix} \\
&= Z_I \begin{pmatrix} \begin{pmatrix} R_I \\ 0 \end{pmatrix} & Z_I^T a_r \end{pmatrix} \\
&= \begin{pmatrix} Z_I^{(1)} & Z_I^{(2)} \end{pmatrix} \begin{pmatrix} R_I & (Z_I^{(1)})^T a_r \\ 0 & (Z_I^{(2)})^T a_r \end{pmatrix} \\
&= \begin{pmatrix} Z_I^{(1)} & Z_I^{(2)} \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & P^T \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & P \end{pmatrix} \begin{pmatrix} R_I & (Z_I^{(1)})^T a_r \\ 0 & (Z_I^{(2)})^T a_r \end{pmatrix} \\
&= \begin{pmatrix} Z_I^{(1)} & Z_I^{(2)} P^T \end{pmatrix} \begin{pmatrix} R_I & (Z_I^{(1)})^T a_r \\ 0 & P(Z_I^{(2)})^T a_r \end{pmatrix}.
\end{aligned}$$

Um zu erreichen, dass

$$A_{I^+}^T = Z_{I^+} \begin{pmatrix} R_{I^+} \\ 0 \end{pmatrix},$$

wird man die orthogonale Matrix P so bestimmen, dass $P(Z_I^{(2)})^T a_r = \delta_I e_1$ ein Vielfaches des ersten Einheitsvektors ist. Dann ist

$$Z_{I^+}^{(1)} = \begin{pmatrix} Z_I^{(1)} & Z_I^{(2)} P^T e_1 \end{pmatrix}, \qquad R_{I^+} = \begin{pmatrix} R_I & (Z_I^{(1)})^T a_r \\ 0 & P(Z_I^{(2)})^T a_r \end{pmatrix}.$$

Eine Orthogonalbasis von $\operatorname{Kern}(A_{I^+})$ ist durch die Spalten von

$$Z_I^{(2)} P^T \begin{pmatrix} e_2 & \cdots & e_{n-q} \end{pmatrix}$$

gegeben, wobei es hierbei auf die Reihenfolge nicht ankommt. Daher kann

$$Z_{I^+}^{(2)} = Z_I^{(2)} P^T \begin{pmatrix} e_{\pi(2)} & \cdots & e_{\pi(n-q)} \end{pmatrix}$$

mit einer noch beliebigen Permutation $(\pi(2), \ldots, \pi(n-q))$ von $(2, \ldots, n-q)$ gewählt werden. Dann ist

$$\begin{aligned}
(Z_{I^+}^{(2)})^T Q Z_{I^+}^{(2)} &= \begin{pmatrix} e_{\pi(2)}^T \\ \vdots \\ e_{\pi(n-q)}^T \end{pmatrix} P(Z_I^{(2)})^T Q Z_I^{(2)} P^T \begin{pmatrix} e_{\pi(2)} & \cdots & e_{\pi(n-q)} \end{pmatrix} \\
&= \begin{pmatrix} e_{\pi(2)}^T \\ \vdots \\ e_{\pi(n-q)}^T \end{pmatrix} P L_I D_I L_I^T P^T \begin{pmatrix} e_{\pi(2)} & \cdots & e_{\pi(n-q)} \end{pmatrix} \\
&= H_{I^+}^T H_{I^+}
\end{aligned}$$

mit

$$H_{I^+} := D_I^{1/2} L_I^T P^T \begin{pmatrix} e_{\pi(2)} & \cdots & e_{\pi(n-q)} \end{pmatrix}.$$

Nun wäre es schön, wenn wir durch eine geeignete Wahl der orthogonalen Matrix P und der Permutation $(\pi(2), \ldots, \pi(n-q))$ erreichen könnten, dass H_{I^+} eine obere Hessenberg-Matrix ist. Denn dann kann man als Produkt von $n-q+1$ Givens-Rotationen eine orthogonale Matrix P_{I^+} mit

$$P_{I^+} H_{I^+} = \begin{pmatrix} N_{I^+} \\ 0 \end{pmatrix}$$

bestimmen, wobei N_{I^+} eine obere Dreiecksmatrix ist. Mit

$$(Z_{I^+}^{(2)})^T Q Z_{I^+}^{(2)} = H_{I^+}^T H_{I^+} = (P_{I^+} H_{I^+})^T P_{I^+} H_{I^+} = N_{I^+}^T N_{I^+} = L_{I^+} D_{I^+} L_{I^+}^T,$$

wobei L_{I^+} eine untere Dreiecksmatrix mit Einsen in der Diagonalen und D_{I^+} eine (positiv definite) Diagonalmatrix. Wie erreicht man also, dass H_{I^+} eine obere Hessenberg-Matrix ist? Hierzu setze man

$$\tilde{I} := \begin{pmatrix} e_{n-q} & \cdots & e_1 \end{pmatrix}$$

und bestimme der Reihe nach die Givens-Rotationen $G_{n-q-1,n-q}, \ldots, G_{12}$ so, dass der Vektor $\tilde{I}(Z_I^{(2)})^T a_r$ (dieser entsteht aus $(Z_I^{(2)})^T a_r$ dadurch, dass die Komponenten von hinten nach vorne liest) in ein Vielfaches des ersten Einheitsvektors überführt wird. Also ist P durch

$$P = G_{12} \cdots G_{n-q-1,n-q} \tilde{I}$$

gegeben. Weiter wählen wir $(\pi(2), \ldots, \pi(n-q)) := (n-q, \ldots, 2)$. Dann ist in der Tat

$$H_{I^+} = D_I^{1/2} L_I^T \tilde{I} G_{n-q-1,n-q}^T \cdots G_{12}^T \begin{pmatrix} e_{n-q} & \cdots & e_2 \end{pmatrix}$$

eine obere Dreiecksmatrix, wie man unschwer nachweist.

7. Sei $I \subset \{1, \ldots, m\}$ eine Indexmenge mit q Elementen, $l \in I$ und $I^+ := I \backslash \{l\}$. Die Matrix A_I habe maximalen Zeilenrang, es sei also Rang $(A_I) = q$, ferner sei die symmetrische Matrix $Q \in \mathbb{R}^{n \times n}$ auf Kern (A_I) positiv definit. Bekannt seien die orthogonale Matrix

$$Z_I = \begin{pmatrix} \underbrace{Z_I^{(1)}}_{q} & \underbrace{Z_I^{(2)}}_{n-q} \end{pmatrix} \in \mathbb{R}^{n \times n},$$

die obere Dreiecksmatrix R_I, die untere Dreiecksmatrix mit Einsen in der Diagonalen L_I sowie die (positiv definite) Diagonalmatrix D_I mit

$$Z_I^{(1)T} A_I^T = R_I, \qquad A_I Z_I^{(2)} = 0, \qquad Z_I^{(2)T} Q Z_I^{(2)} = L_I D_I L_I^T.$$

Man entwickle ein effizientes, stabiles Verfahren zur Berechnung der Matrizen Z_{I^+}, R_{I^+}, L_{I^+} und D_{I^+} mit den zu Z_I, R_I, L_I bzw. D_I analogen Eigenschaften.

Lösung: Sei l das k-te Element in I und daher $A_{I^+} = T_k A_I$, wobei wieder

$$T_k := \begin{pmatrix} e_1^T \\ \vdots \\ e_{k-1}^T \\ e_{k+1}^T \\ \vdots \\ e_q^T \end{pmatrix} \in \mathbb{R}^{(q-1) \times q}$$

wie in der Lösung zu Aufgabe 4. Mit einer beliebigen orthogonalen Matrix $P \in \mathbb{R}^{q\times q}$ ist dann

$$\begin{aligned} A_{I^+}^T &= A_I^T T_k^T \\ &= Z_I \begin{pmatrix} R_I \\ 0 \end{pmatrix} T_k^T \\ &= \begin{pmatrix} Z_I^{(1)} & Z_I^{(2)} \end{pmatrix} \begin{pmatrix} P^T & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} P & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} R_I T_k^T \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} Z_I^{(1)} P^T & Z_I^{(2)} \end{pmatrix} \begin{pmatrix} P R_I T_k^T \\ 0 \end{pmatrix}. \end{aligned}$$

Daher sollte die orthogonale Matrix P so bestimmt werden, dass $R_{I^+} = P R_I T_k^T$ eine obere Dreiecksmatrix ist. Als Matrix $Z_{I^+}^{(1)}$ nehmen wir die Matrix $Z_I^{(1)} P^T$, bei der aber die letzte (also q-te) Spalte noch weggelassen ist. Dann ist $Z_{I^+}^{(1)} = Z_I^{(1)} P^T T_q^T$, ferner können wir

$$Z_{I^+}^{(2)} := \begin{pmatrix} Z_I^{(2)} & z_I^{(2)} \end{pmatrix} \qquad \text{mit} \qquad z_I^{(2)} := Z_I^{(1)} P^T e_q$$

setzen um zu erreichen, dass $Z_{I^+} = \begin{pmatrix} Z_{I^+}^{(1)} & Z_I^{(2)} \end{pmatrix}$ orthogonal ist und

$$A_{I^+}^T = Z_{I^+} \begin{pmatrix} R_{I^+} \\ 0 \end{pmatrix}$$

gilt. Nun ist

$$\begin{aligned} (Z_{I^+}^{(2)})^T Q Z_{I^+}^{(2)} &= \begin{pmatrix} (Z_I^{(2)})^T \\ (z_I^{(2)})^T \end{pmatrix} Q \begin{pmatrix} Z_I^{(2)} & z_I^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} (Z_I^{(2)})^T Q Z_I^{(2)} & (Z_I^{(2)})^T Q z_I^{(2)} \\ (z_I^{(2)})^T Q Z_I^{(2)} & (z_I^{(2)})^T Q z_I^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} L_I D_I L_I^T & (Z_I^{(2)})^T Q z_I^{(2)} \\ (z_I^{(2)})^T Q Z_I^{(2)} & (z_I^{(2)})^T Q z_I^{(2)} \end{pmatrix}. \end{aligned}$$

Macht man daher den Ansatz

$$L_{I^+} = \begin{pmatrix} L_I & 0 \\ l_I^T & 1 \end{pmatrix}, \qquad D_{I^+} = \begin{pmatrix} D_I & 0 \\ 0 & \delta_I \end{pmatrix},$$

so ist

$$(Z_{I^+}^{(2)})^T Q Z_{I^+}^{(2)} = L_{I^+} D_{I^+} L_{I^+}^T$$

genau dann, wenn

$$L_I D_I l_I = (Z_I^{(2)})^T Q z_I^{(2)}, \qquad l_I^T D_I l_I + \delta_I = (z_I^{(2)})^T Q z_I^{(2)}.$$

Aus der ersten Gleichung erhält man l_I, anschließend aus der zweiten δ_I.

8. Man programmiere das Verfahren von Fletcher und teste das Programm an konvexen, quadratischen Optimierungsaufgaben mit den folgenden Daten:

(a) Es sei $m := 4$, $m_0 := 4$ und $n := 3$, ferner sei

$$Q := \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 0 \\ 2 & 0 & 2 \end{pmatrix}, \qquad c := \begin{pmatrix} -8 \\ -6 \\ -4 \end{pmatrix}$$

sowie

$$A := \begin{pmatrix} -1 & -1 & -2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad b := \begin{pmatrix} -3 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Wie bei W. Hock, K. Schittkowski (1981)[8] starte man mit der zulässigen Lösung $x := (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})^T$ (und damit $I := \emptyset$).

(b) Es sei $m := 7$, $m_0 := 7$ und $n := 4$, ferner sei

$$Q := \begin{pmatrix} 2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 2 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \qquad c := \begin{pmatrix} -1 \\ -3 \\ 1 \\ -1 \end{pmatrix}$$

und

$$A := \begin{pmatrix} -1 & -2 & -1 & -1 \\ -3 & -1 & -2 & 1 \\ 0 & 1 & 4 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad b := \begin{pmatrix} -5 \\ -4 \\ \frac{3}{2} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Wie bei W. Hock, K. Schittkowski (1981, S. 96) starte man mit der zulässigen Lösung $x := (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})^T$ (und damit $I := \emptyset$).

Lösung: In (a) erhalten wir die folgenden Werte: Zunächst ist $x^{(0)} = (0.5, 0.5, 0.5)^T$ und $I^{(0)} = \emptyset$, dann $x^{(1)} = (0.75, 0.75, 0.75)^T$ und $I^{(1)} = \{1\}$ und schließlich im nächsten Schritt die Lösung

$$x^* = \begin{pmatrix} 1.3333 \\ 0.7778 \\ 0.4444 \end{pmatrix}, \qquad I^* = \{1\}, \qquad y^* = \begin{pmatrix} 0.2222 \\ 0.0000 \\ 0.0000 \\ 0.0000 \end{pmatrix}.$$

In (b) erhalten wir die folgenden Werte: Zunächst ist $x^{(0)} = (0.5, 0.5, 0.5, 0.5)^T$ und $I^{(0)} = \emptyset$, dann erhalten wir die folgenden Ergebnisse:

$$x^{(1)} = \begin{pmatrix} 0.3696 \\ 0.7174 \\ 0.1957 \\ 0.8043 \end{pmatrix}, \qquad I^{(1)} = \{3\},$$

[8] Hock, W. and K. Schittkowski (1981) *Test Examples for Nonlinear Programming Codes.* Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin-Heidelberg-New York.

$$x^{(2)} = \begin{pmatrix} 0.3404 \\ 1.5000 \\ 0.0000 \\ 1.0000 \end{pmatrix}, \qquad I^{(2)} = \{3, 6\},$$

$$x^{(3)} = \begin{pmatrix} 0.5000 \\ 1.5000 \\ 0.0000 \\ 1.0000 \end{pmatrix}, \qquad I^{(3)} = \{6\},$$

$$x^{(4)} = \begin{pmatrix} 0.5000 \\ 1.7500 \\ 0.0000 \\ 1.0000 \end{pmatrix}, \qquad I^{(4)} = \{6, 1\}$$

und danach schließlich die Lösung x^* mit zugehörigem Multiplikator y^*:

$$x^* = \begin{pmatrix} 0.2727 \\ 2.0909 \\ 0.0000 \\ 0.5455 \end{pmatrix}, \qquad y^* = \begin{pmatrix} 0.4545 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 1.7272 \\ 0.0000 \end{pmatrix}.$$

9. Gegeben sei ein lineares Gleichungssystem der Form

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Hierbei sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$. Man zeige, dass man obiges lineares Gleichungssystem mit den folgenden Schritten lösen kann:

- Bestimme eine QR-Zerlegung von $A^T \in \mathbb{R}^{n\times m}$, berechne also, etwa mit dem Householder-Verfahren, eine orthogonale Matrix $Z \in \mathbb{R}^{n\times n}$ und eine (nichtsinguläre) obere Dreiecksmatrix $R \in \mathbb{R}^{m\times m}$ mit

$$ZA^T = \begin{pmatrix} R \\ 0 \end{pmatrix}.$$

  Simultan berechne man

$$\begin{pmatrix} c \\ d \end{pmatrix} := Za, \qquad \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} := ZQZ^T.$$

  Hierbei ist $c \in \mathbb{R}^m$, $d \in \mathbb{R}^{n-m}$, ferner ist $B_{22} \in \mathbb{R}^{(n-m)\times(n-m)}$ symmetrisch und positiv definit (Beweis?).

- Durch Vorwärtseinsetzen bestimme man $u \in \mathbb{R}^m$ aus $R^T u = b$.
- Mit Hilfe des Cholesky-Verfahrens berechne man $v \in \mathbb{R}^{n-m}$ aus $B_{22}v = d - B_{21}u$.

- Gewinne die Anteile $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ der gesuchten Lösung aus

$$x := Z^T \begin{pmatrix} u \\ v \end{pmatrix}$$

und

$$Ry = c - B_{11}u - B_{12}v$$

durch Rückwärtseinsetzen.

Lösung: Da Q positiv definit und Z orthogonal, ist ZQZ^T ebenfalls positiv definit und daher insbesondere die Diagonalblöcke B_{11} und B_{22} positiv definit. Fasst man alle Schritte zusammen, so ist mit der QR-Zerlegung

$$ZA^T = \begin{pmatrix} R \\ 0 \end{pmatrix}$$

von A^T und den Partitionierungen

$$Za = \begin{pmatrix} c \\ d \end{pmatrix}, \qquad ZQZ^T = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

offenbar

$$x = Z^T \begin{pmatrix} R^{-T}b \\ B_{22}^{-1}(d - B_{21}R^{-T}b) \end{pmatrix}, \qquad y = R^{-1}[c - B_{11}R^{-T}b - B_{12}B_{22}^{-1}(d - B_{21}R^{-T}b)].$$

Dann ist

$$Ax = AZ^T \begin{pmatrix} R^{-T}b \\ B_{22}^{-1}(d - B_{21}R^{-T}b) \end{pmatrix} = \begin{pmatrix} R^T & 0 \end{pmatrix} \begin{pmatrix} R^{-T}b \\ B_{22}^{-1}(d - B_{21}R^{-T}b) \end{pmatrix} = b.$$

Nachzuweisen bleibt also nur die Gültigkeit der ersten Gleichung $Qx + A^T y = b$, welche äquivalent zu

$$Z(Qx + A^T y) = Za = \begin{pmatrix} c \\ d \end{pmatrix}$$

ist. Nun ist

$$\begin{aligned}
Z(Qx + A^T y) &= ZQZ^T \begin{pmatrix} R^{-T}b \\ B_{22}^{-1}(d - B_{21}R^{-T}b) \end{pmatrix} \\
&\quad + ZA^T R^{-1}[c - B_{11}R^{-T}b - B_{12}B_{22}^{-1}(d - B_{21}R^{-T}b)] \\
&= \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} R^{-T}b \\ B_{22}^{-1}(d - B_{21}R^{-T}b) \end{pmatrix} \\
&\quad + \begin{pmatrix} R \\ 0 \end{pmatrix} R^{-1}[c - B_{11}R^{-T}b - B_{12}B_{22}^{-1}(d - B_{21}R^{-T}b)] \\
&= \begin{pmatrix} B_{11}R^{-T}b + B_{12}B_{22}^{-1}(d - B_{21}R^{-T}b) \\ B_{21}R^{-T}b + d - B_{21}R^{-T}b \end{pmatrix} \\
&\quad + \begin{pmatrix} c - B_{11}R^{-T}b - B_{12}B_{22}^{-1}(d - B_{21}R^{-T}b) \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} c \\ d \end{pmatrix} \\
&= Za.
\end{aligned}$$

Damit ist die Behauptung bewiesen.

6.3.2 Aufgaben in Abschnitt 3.2

1. Gegeben sei das quadratische Programm

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2}\, x^T Q x \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}. \end{array} \right.$$

Hierbei seien $a_1, \ldots, a_m \in \mathbb{R}^n \setminus \{0\}$, $b = (b_i) \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Die Matrix $A \in \mathbb{R}^{m \times n}$ besitze a_i^T als i-te Zeile, $i = 1, \ldots, m$. Sei (P) zulässig, $x^* \in M$ sei die eindeutige Lösung von (P). Man zeige, dass eine Indexmenge $I^* \subset \{1, \ldots, m\}$ existiert derart, daß (x^*, I^*) ein Lösungspaar für (P) ist.

Lösung: Sei $I_0 := \{i \in \{1, \ldots, m\} : a_i^T x^* = b_i\}$ die Menge der in x^* aktiven Restriktionen. Man betrachte die Menge $\mathcal{I}$ der Indexmengen $I \subset I_0$ mit

$$c + Qx^* \in \left\{ \sum_{i \in I} y_i a_i : y_i \geq 0 \quad (i \in I \cap \{1, \ldots, m_0\}) \right\}.$$

Diese Menge ist nichtleer, da $I_0 \in \mathcal{I}$. Sei $I^* \in \mathcal{I}$ eine Menge mit einer minimalen Anzahl von Elementen. Dann ist (x^*, I^*) ein Lösungspaar für (P). Hierzu müssen wir zeigen:

(a) Die Vektoren $\{a_i\}_{i \in I^*}$ sind linear unabhängig.
Wären $\{a_i\}_{i \in I^*}$ linear abhängig, so existierten β_i, $i \in I^*$, die nicht alle verschwinden, mit $\sum_{i \in I^*} \beta_i a_i = 0$. Nach Voraussetzung existieren y_i^*, $i \in I^*$, mit $c + Qx^* = \sum_{i \in I^*} y_i^* a_i$ und der Eigenschaft, dass $y_i^* \geq 0$ für alle $i \in I^* \cap \{1, \ldots, m_0\}$. Wegen der Minimalität von I^* ist $y_i^* \neq 0$, $i \in I^*$. Offenbar existiert ein $t \in \mathbb{R}$ derart, dass $y_i^* + t\beta_i = 0$ für wenigstens ein $i \in I^*$ und $y_i^* + t\beta_i \geq 0$ für alle $i \in I^* \cap \{1, \ldots, m_0\}$. Dies ist aber ein Widerspruch zur Minimalität von I^*.

(b) Es ist $A_{I^*} x^* = b_{I^*}$.
Dies ist richtig, da $A_{I_0} x^* = b_{I_0}$ und $I^* \subset I_0$.

(c) Es existiert ein y_{I^*} mit $c + Qx^* = A_{I^*}^T y_{I^*}$ und $y_i \geq 0$ für alle $i \in I^* \cap \{1, \ldots, m_0\}$.
Dies ist wegen $I^* \in \mathcal{I}$ und der Definition von $\mathcal{I}$ richtig.

Damit ist die Aufgabe gelöst.

2. Gegeben sei wieder das quadratische Programm (P) aus Aufgabe 1. Sei (x, I) ein Lösungspaar, $p \in \{1, \ldots, m\} \setminus I$ eine durch x verletzte Restriktion und $M_{I \cup \{p\}} \neq \emptyset$. Ferner sei $a_p \notin \operatorname{span}\{a_i : i \in I\}$, so dass die Vektoren $\{a_i\}_{i \in I \cup \{p\}}$ linear unabhängig sind. Sei x^+ die (eindeutige) Lösung von $(\mathrm{P}_{I \cup \{p\}})$, $\overline{I} := \{i \in I : a_i^T x^+ = b_i\}$ und $I^+ := \overline{I} \cup \{p\}$. Man zeige, daß (x^+, I^+) ein Lösungspaar mit $f(x^+) > f(x)$ ist.

Lösung: Klar ist, dass die Vektoren $\{a_i\}_{i \in I^+}$ linear unabhängig sind. Zum Nachweis von $A_{I^+} x^+ = b_{I^+}$ bleibt $a_p^T x^+ = b_p$ zu zeigen. Da x^+ die Lösung von $(\mathrm{P}_{I \cup \{p\}})$ ist, ist dies klar, wenn p der Index einer Gleichungsrestriktion bzw. $p \in \{m_0 + 1, \ldots, m\}$ ist. Daher kann $p \in \{1, \ldots, m_0\}$ angenommen werden. Aus den notwendigen Optimalitätsbedingungen folgt, dass zu der Lösung x^+ von $(\mathrm{P}_{I \cup \{p\}})$ ein Vektor $y^+ \in \mathbb{R}^{q+1}$ (hier ist wieder $q := \#(I)$ die Anzahl der Elemente von I) mit

$$y_i^+ \geq 0 \quad (i \in I \cap \{1, \ldots, m_0\}), \qquad y_p^+ \geq 0$$

sowie

$$c + Qx^+ = A_{I\cup\{p\}}^T y^+, \qquad (A_{I\cup\{p\}}x^+ - b_{I\cup\{p\}})^T y^+ = 0$$

existiert. Wäre nun $a_p^T x^+ > b_p$, so folgt $y_p^+ = 0$. Aus den hinreichenden Optimalitätsbedingungen erhält man, dass x^+ auch die Lösung von (P_I) ist. Wegen der Eindeutigkeit einer Lösung von (P_I) ist $x^+ = x$, was einen Widerspruch dazu ergibt, dass x die p-te Restriktion verletzt, x^+ ihr aber genügt. Daher ist (x^+, I^+) ein Lösungspaar, wenn auch noch bewiesen ist, dass x^+ die Lösung von (P_{I^+}) ist. Hierzu beachten wir, dass zu der Lösung x^+ von $(\mathrm{P}_{I\cup\{p\}})$ ein Vektor $y^+ \in \mathbb{R}^{q+1}$ mit den oben angegebenen Eigenschaften existiert. Nach Definition von $\overline{I}$ folgt aus der Gleichgewichtsbedingung, dass $y_i^+ = 0$ für $i \in I \setminus \overline{I}$ und daher

$$c + Qx^+ = \sum_{i\in I\cup\{p\}} y_i^+ a_i = \sum_{i\in\overline{I}\cup\{p\}} y_i^+ a_i = \sum_{i\in I^+} y_i^+ a_i = A_{I^+}^T y^+.$$

Aus den hinreichenden Optimalitätsbedingungen folgt, dass x^+ die Lösung von (P_{I^+}) ist. Insgesamt ist (x^+, I^+) ein Lösungspaar für (P). Wegen $x^+ \neq x$ ist schließlich mit einem zu (x, I) gehörenden Lagrange-Vektor $y_I \in \mathbb{R}^q$:

$$\begin{aligned}
f(x^+) &= f(x) + (c + Qx)^T(x^+ - x) + \frac{1}{2}(x^+ - x)^T Q(x^+ - x) \\
&> f(x) + (c + Qx)^T(x^+ - x) \\
&= f(x) + (A_I^T y_I)^T(x^+ - x) \\
&= f(x) + (A_I x^+ - b_I)^T y_I \\
&= f(x) + \sum_{i\in I\cap\{1,\ldots,m_0\}} \underbrace{y_i}_{\geq 0}\underbrace{(a_i^T x^+ - b_i)}_{\geq 0} \\
&\geq f(x),
\end{aligned}$$

womit auch $f(x^+) > f(x)$ bewiesen ist.

3. Sei $I \subset \{1, \ldots, m\}$, $q := \#(I)$. Sei $A \in \mathbb{R}^{m\times n}$ mit Rang $(A_I) = q$ und $Q \in \mathbb{R}^{n\times n}$ symmetrisch, positiv definit. Bekannt sei eine Matrix $Z_I \in \mathbb{R}^{n\times n}$ derart, dass

$$Z_I Z_I^T = Q^{-1}, \qquad Z_I^T A_I^T = \begin{pmatrix} R_I \\ 0 \end{pmatrix}$$

mit einer (nichtsingulären) oberen Dreiecksmatrix $R_I \in \mathbb{R}^{q\times q}$. Sei $l \in I$. Wie bestimmt man eine Matrix $Z_{I\setminus\{l\}} \in \mathbb{R}^{n\times n}$ derart, dass

$$Z_{I\setminus\{l\}} Z_{I\setminus\{l\}}^T = Q^{-1}, \qquad Z_{I\setminus\{l\}}^T A_{I\setminus\{l\}}^T = \begin{pmatrix} R_{I\setminus\{l\}} \\ 0 \end{pmatrix}$$

mit einer (nichtsingulären) oberen Dreiecksmatrix $R_{I\setminus\{l\}} \in \mathbb{R}^{(q-1)\times(q-1)}$ gilt?

Lösung: Sei l das k-te Element von I mit $1 \leq k \leq$. Es liegt nahe die Matrix

$$T_k := \begin{pmatrix} e_1^T \\ \vdots \\ e_{k-1}^T \\ e_{k+1}^T \\ \vdots \\ e_q^T \end{pmatrix} \in \mathbb{R}^{(q-1)\times q}$$

zu definieren, wobei e_i den i-ten Einheitsvektor im $\mathbb{R}^q$ bedeutet. Die Matrix $A_{I\setminus\{l\}}$, die man durch Entfernen der k-ten Zeile aus A_I erhält, ist dann durch $A_{I\setminus\{l\}} = T_k A_I$ gegeben. Mit dem Ansatz $Z_{I\setminus\{l\}} = Z_I \Omega_I^T$ erhalten wir

$$Z_{I\setminus\{l\}}^T A_{I\setminus\{l\}}^T = \Omega_I Z_I^T A_I^T T_k^T = \Omega_I \begin{pmatrix} R_I \\ 0 \end{pmatrix} T_k^T = \Omega_I \begin{pmatrix} R_I T_k^T \\ 0 \end{pmatrix}.$$

Hierbei ist $R_I T_k^T \in \mathbb{R}^{q\times(q-1)}$ die Matrix, die aus der oberen Dreiecksmatrix $R_I \in \mathbb{R}^{q\times q}$ dadurch entsteht, dass die k-te Spalte gestrichen wird. Daher ist

$$\begin{matrix} q \\ n-q \end{matrix} \left\{ \begin{pmatrix} R_I T_k^T \\ 0 \end{pmatrix} \right. = \begin{pmatrix} R_I^{(11)} & R_I^{(12)} \\ 0 & R_I^{(22)} \\ \underbrace{0}_{k-1} & \underbrace{0}_{q-k} \end{pmatrix} \begin{matrix} \} \ k-1 \\ \} \ q-k+1 \\ \} \ n-q \end{matrix}$$

mit der oberen Dreiecksmatrix $R_I^{(11)} \in \mathbb{R}^{(k-1)\times(k-1)}$ und der oberen Hessenberg-Matrix $R_I^{(22)} \in \mathbb{R}^{(q-k+1)\times(q-k)}$. Hier erinnern wir daran, dass eine obere Hessenberg-Matrix eine Matrix ist, bei der alle Elemente unterhalb der unteren Nebendiagonalen verschwinden. Für die orthogonale Matrix $\Omega_I \in \mathbb{R}^{n\times n}$ liegt daher der Ansatz

$$\Omega_I = \begin{pmatrix} I_{k-1} & 0 & 0 \\ 0 & \Omega_I^{(2)} & 0 \\ \underbrace{0}_{k-1} & \underbrace{0}_{q-k+1} & \underbrace{I_{n-q}}_{n-q} \end{pmatrix} \begin{matrix} \} \ k-1 \\ \} \ q-k+1 \\ \} \ n-q \end{matrix}$$

mit einer orthogonalen Matrix $\Omega_I^{(2)} \in \mathbb{R}^{(q-k+1)\times(q-k+1)}$ nahe. Wegen

$$\Omega_I \begin{pmatrix} R_I T_k^T \\ 0 \end{pmatrix} = \begin{pmatrix} I_{k-1} & 0 & 0 \\ 0 & \Omega_I^{(2)} & 0 \\ 0 & 0 & I_{n-q} \end{pmatrix} \begin{pmatrix} R_I^{(11)} & R_I^{(12)} \\ 0 & R_I^{(22)} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} R_I^{(11)} & R_I^{(12)} \\ 0 & \Omega_I^{(2)} R_I^{(22)} \\ 0 & 0 \end{pmatrix}$$

kommt es darauf an, die orthogonale Matrix $\Omega_I^{(2)}$ so zu bestimmen, dass

$$\Omega_I^{(2)} R_I^{(22)} = \begin{pmatrix} \hat{R}_I^{(22)} \\ \underbrace{0^T}_{q-k} \end{pmatrix} \begin{matrix} \} \ q-k \\ \} \ 1 \end{matrix}$$

eine obere Dreiecksmatrix ist. Dies erreicht man, indem man $R_I^{(22)}$ sukzessive mit $q-k$ Givens-Rotationen $G_{12}, \ldots, G_{q-k,q-k+1}$ von links multipliziert, d. h. die orthogonale Matrix $\Omega_I^{(2)}$ hat die Form

$$\Omega_I^{(2)} = G_{q-k,q-k+1} \cdots G_{12}.$$

Die neue obere Dreiecksmatrix $R_{I\setminus\{l\}}$ ist also gegeben durch

$$R_{I\setminus\{l\}} = \begin{pmatrix} R_I^{(11)} & R_I^{(12)} \\ \underbrace{0}_{k-1} & \underbrace{\hat{R}_I^{(22)}}_{q-k} \end{pmatrix} \begin{matrix} \} \ k-1 \\ \} \ q-k \end{matrix}$$

Nun gilt es, die neue Matrix $Z_{I\setminus\{l\}} = Z_I \Omega_I^T$ zu berechnen. Wir erinnern daran, dass wir uns Z_I zerlegt denken in $Z_I = (\ Z_I^{(1)} \quad Z_I^{(2)}\ )$. Es liegt nahe, die Matrix $Z_I^{(1)}$, in der die ersten q Spalten von Z_I stehen, weiter zu zerlegen:

$$Z_I^{(1)} = (\ \underbrace{Z_{k-1}^{(1)}}_{k-1} \quad \underbrace{Z_{q-k+1}^{(1)}}_{q-k+1}\ ).$$

Hiermit wird

$$\begin{aligned} Z_{I\setminus\{l\}} &= (\ Z_{k-1}^{(1)} \quad Z_{q-k+1}^{(1)} \quad Z_I^{(2)}\ ) \begin{pmatrix} I_{k-1} & 0 & 0 \\ 0 & \Omega_I^{(2)T} & 0 \\ 0 & 0 & I_{n-q} \end{pmatrix} \\ &= (\ Z_{k-1}^{(1)} \quad Z_{q-k+1}^{(1)} \Omega_I^{(2)T} \quad Z_I^{(2)}\ ). \end{aligned}$$

Gegenüber Z_I verändern sich in $Z_{I\setminus\{l\}}$ also nur die Spalten mit dem Index $k, \ldots, q$. Wegen

$$Z_{q-k+1}^{(1)} \Omega_I^{(2)T} = Z_{q-k+1}^{(1)} G_{12}^T \ldots G_{q-k,q-k+1}$$

kann diese Berechnung parallel zu der von $\hat{R}_I^{(22)}$ erfolgen, so dass es wie im ersten Fall nicht nötig ist, sich die Givens-Rotationen $G_{12}, \ldots, G_{q-k,q-k+1}$ zu merken.

Damit ist das Updaten von Z_I und R_I vollständig beschrieben. Auch der Vektor $d_I := Z_I^T a_p$ kann gleichzeitig upgedatet werden. Mit

$$d_I = \begin{pmatrix} d_{k-1}^{(1)} \\ d_{q-k+1}^{(1)} \\ d_I^{(2)} \end{pmatrix} \begin{matrix} \} \ k-1 \\ \} \ q-k+1 \\ \} \ n-q \end{matrix}$$

erhält man in diesem Falle

$$d_{I\setminus\{l\}} = \Omega_I d_I = \begin{pmatrix} I_{k-1} & 0 & 0 \\ 0 & \Omega_I^{(2)} & 0 \\ 0 & 0 & I_{n-q} \end{pmatrix} \begin{pmatrix} d_{k-1}^{(1)} \\ d_{q-k+1}^{(1)} \\ d_I^{(2)} \end{pmatrix} \begin{pmatrix} d_{k-1}^{(1)} \\ \Omega_I^{(2)} d_{q-k+1}^{(1)} \\ d_I^{(2)} \end{pmatrix}.$$

In Pseudocode unter Verwendung der Funktion "rot" sieht dies etwa folgendermaßen aus:

- Input:
  - Eine Matrix $Z \in \mathbb{R}^{n\times n}$,
  - Eine Indexmenge $I = \{i_{\text{act}}(1), \ldots, i_{\text{act}}(q)\}$ mit $1 \le q \le n$,
  - Ein $k \in \mathbb{N}$ mit $1 \le k \le q$.
  - Eine obere Dreiecksmatrix $R \in \mathbb{R}^{q\times q}$,
  - Ein Vektor $d \in \mathbb{R}^n$.

  Hierbei ist

$$ZZ^T = Q^{-1}, \qquad Z^T A_I^T = \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad d = Z^T a_p.$$

- $q := q - 1$
  Für $j = k, \ldots, q$:
  $i_{\text{act}}(j) := i_{\text{act}}(j+1)$
  Für $i = 1, \ldots, j+1$:
  $r_{ij} := r_{ij+1}$
  Für $j = k, \ldots, q$:
  $(c, s, r_{jj}) := \text{rot}(r_{jj}, r_{j+1,j})$
  $\text{temp} := cd_j + sd_{j+1}, \quad d_{j+1} := -sd_j + cd_{j+1}, \quad d_j := \text{temp}$
  Für $i = j+1, \ldots, q$:
  $\text{temp} := cr_{ji} + sr_{j+1,i}, \quad r_{j+1,i} := -sr_{ji} + cr_{j+1,i}, \quad r_{ji} := \text{temp}$
  Für $i = 1, \ldots, n$:
  $\text{temp} := cz_{ij} + sz_{i,j+1}, \quad z_{i,j+1} := -sz_{ij} + cz_{i,j+1}, \quad z_{ij} := \text{temp}$
- Output:
  - Eine Matrix $Z \in \mathbb{R}^{n \times n}$,
  - Aus der alten Indexmenge I ist das k-te Element gestrichen und daher $q := q - 1$ und $i_{\text{act}}(j) := i_{\text{act}}(j+1)$, $j = k, \ldots, q$ gesetzt worden,
  - Eine obere Dreiecksmatrix $R \in \mathbb{R}^{q \times q}$,
  - Ein Vektor $d \in \mathbb{R}^n$.

  Nach Abschluss ist

$$ZZ^T = Q^{-1}, \qquad Z^T A_I^T = \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad d = Z^T a_p.$$

4. Gegeben sei die symmetrische, positiv definite Matrix $Q \in \mathbb{R}^{n \times n}$ und der Vektor $c \in \mathbb{R}^n$. Sei $I \subset \{1, \ldots, m\}$ (mit $q := \#(I)$) eine nichtleere Indexmenge mit der Eigenschaft, dass die Vektoren $\{a_i\}_{i \in I} \subset \mathbb{R}^n$ linear unabhängig sind bzw. Rang $(A_I) = q$ gilt. Es existiere eine Matrix $Z_I \in \mathbb{R}^{n \times n}$, so daß

$$(*) \qquad Z_I Z_I^T = Q^{-1}, \qquad Z_I^T A_I^T = \begin{pmatrix} R_I \\ 0 \end{pmatrix} \begin{matrix} \} \, q \\ \} \, n-q \end{matrix}$$

mit einer oberen Dreiecksmatrix $R_I \in \mathbb{R}^{q \times q}$. Sei

$$Z_I = (\underbrace{Z_I^{(1)}}_{q} \quad \underbrace{Z_I^{(2)}}_{n-q}).$$

Bei gegebenem $b_I \in \mathbb{R}^q$ berechne man $\hat{x}$, $x \in \mathbb{R}^n$ aus

$$\hat{x} := Z_I^{(1)} R_I^{-T} b_I, \qquad x := \begin{cases} \hat{x} - Z_I^{(2)} Z_I^{(2)T} c & \text{für} \quad q < n, \\ \hat{x} & \text{für} \quad q = n. \end{cases}$$

Man zeige, dass x eine Lösung des durch lineare Gleichungen restringierten quadratischen Programms

$$\text{Minimiere} \quad f(x) := c^T x + \tfrac{1}{2} x^T Q x \quad \text{unter der Nebenbedingung} \quad A_I x = b_I$$

mit zugehörigem Lagrange-Vektor $y_I := R_I^{-1} Z_I^{(1)T}(c + Q\hat{x})$ ist.

Lösung: Wie üblich sei

$$N_I := (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1}, \qquad H_I := Q^{-1}(I - A_I^T N_I).$$

Dann ist $N_I = R_I^{-1} Z_I^{(1)T}$ und $H_I = Z_I^{(2)} Z_I^{(2)T}$ und folglich

$$\hat{x} = N_I^T b_I, \qquad x = \hat{x} - H_I c, \qquad y_I = N_I(c + Q\hat{x}).$$

Hiermit erhält man

$$A_I x = A_I \hat{x} = A_I N_I^T b_I = b_I.$$

Schließlich ist

$$\begin{aligned} c + Qx - A_I^T y_I &= c + Q\hat{x} - (I - A_I^T N_I)c - A_I^T N_I(c + Q\hat{x}) \\ &= \underbrace{(I - A_I^T N_I) Q N_I^T}_{=0} b_I \\ &= 0. \end{aligned}$$

Aus den hinreichenden Optimalitätsbedingungen folgt die Behauptung.

Die Aussage dieser Aufgabe ist Grundlage für eine iterative Verbesserung eines Lösungspaares, die man in den Algorithmus von Goldfarb-Idnani insbesondere dann einbauen sollte, wenn die Matrix Q kleine Eigenwerte besitzt. Ist nämlich $I \subset \{1, \ldots, m\}$, $q := \#(I)$, eine Indexmenge mit $1 \le q \le n$ und Rang $(A_I) = q$, ist $(\tilde{x}, \tilde{y}_I)$ näherungsweise eine Lösung von

$$A_I x = b_I, \qquad c + Qx = A_I^T y_I,$$

sind ferner $Z_I \in R^{n \times n}$ und $R_I \in \mathbb{R}^{q \times q}$ Matrizen, die $(*)$ genügen, so berechne man zunächst die Defekte $b_I - A_I \tilde{x}$ und $c + Q\tilde{x} - A_I^T \tilde{y}_I$ und anschließend

$$\begin{aligned} \hat{x} &:= Z_I^{(1)} R_I^{-T}(b_I - A_I \tilde{x}), \\ x &:= \tilde{x} + \begin{cases} \hat{x} - Z_I^{(2)} Z_I^{(2)T}(c + Q\tilde{x} - A_I^T \tilde{y}_I + Q\hat{x}) & \text{für } q < n, \\ \hat{x} & \text{für } q = n, \end{cases} \\ y_I &:= \tilde{y}_I + R_I^{-1} Z_I^{(1)T}(c + Q\tilde{x} - A_I^T \tilde{y}_I + Q\tilde{x}). \end{aligned}$$

5. Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch. Mit einer Spalten-Version des Cholesky-Verfahrens soll getestet werden, ob A positiv definit ist. Dies könnte folgendermaßen aussehen:

   - Gegeben sei die symmetrische Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$.
   - Für $k = 1, \ldots, n$:
     Berechne $\tilde{a}_{kk} := a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2$.
     Falls $\tilde{a}_{kk} \le 0$, dann:
     STOP: A nicht positiv definit.
     Andernfalls:
     Berechne $l_{kk} := (\tilde{a}_{kk})^{1/2}$.
     Für $i = k + 1, \ldots, n$:
     Berechne $l_{ik} := (a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj})/l_{kk}$.

Angenommen, das Verfahren breche im k-ten Schritt wegen $\tilde{a}_{kk} \leq 0$ ab. Dann ist

$$A = \begin{pmatrix} L_1 & 0 \\ L_2 & I \end{pmatrix} \begin{pmatrix} L_1^T & L_2^T \\ 0 & \tilde{A} \end{pmatrix}$$

mit

$$L_1 := \begin{pmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{k-1,1} & \cdots & l_{k-1,k-1} \end{pmatrix}, \qquad L_2 := \begin{pmatrix} l_{k1} & \cdots & l_{k,k-1} \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{n,k-1} \end{pmatrix}$$

und

$$\tilde{A} := \begin{pmatrix} \tilde{a}_{kk} & \cdots & \tilde{a}_{kn} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{nk} & \cdots & \tilde{a}_{nn} \end{pmatrix} := \begin{pmatrix} a_{kk} & \cdots & a_{kn} \\ \vdots & \ddots & \vdots \\ a_{nk} & \cdots & a_{nn} \end{pmatrix} - L_2 L_2^T.$$

Schließlich sei $x_1 \in \mathbb{R}^{k-1}$ die eindeutige Lösung von $L_1^T x_1 = -L_2^T e_1$, wobei $e_1 \in \mathbb{R}^{n-k+1}$ den ersten Einheitsvektor bezeichnet. Man zeige, dass

$$\lambda_{\min}(A) \leq \frac{\tilde{a}_{kk}}{\|x_1\|_2^2 + 1}.$$

Man muss also mindestens $-\tilde{a}_{kk}/(\|x_1\|_2^2+1)$ zu den Diagonalelementen von A addieren, um eine positiv semidefinite Matrix zu erhalten.

Lösung: Man definiere $x \in \mathbb{R}^n$ durch

$$x := \begin{pmatrix} x_1 \\ e_1 \end{pmatrix}.$$

Dann ist

$$x^T A x = \begin{pmatrix} x_1 \\ e_1 \end{pmatrix}^T \begin{pmatrix} L_1 & 0 \\ L_2 & I \end{pmatrix} \begin{pmatrix} L_1^T & L_2^T \\ 0 & \tilde{A} \end{pmatrix} \begin{pmatrix} x_1 \\ e_1 \end{pmatrix} = e_1^T \tilde{A} e_1 = \tilde{a}_{kk}$$

und daher

$$\lambda_{\min}(A) = \min_{z \neq 0} \frac{z^T A z}{z^T z} \leq \frac{x^T A x}{x^T x} = \frac{\tilde{a}_{kk}}{\|x_1\|_2^2 + 1}.$$

Damit ist die Aussage bewiesen.

6.3.3 Aufgaben in Abschnitt 3.3

1. Man beweise Lemma 3.1: Gegeben sei das vorzeichenbeschränkte quadratische Programm

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := c^T x + \dfrac{1}{2} x^T Q x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : x_j \geq 0 \ \ (j = 1, \ldots, n_0)\}. \end{cases}$$

Hierbei sei $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit. Ist $x^* \in \mathbb{R}^n$ eine Lösung von

$$(*) \qquad x = [x - \alpha(Qx + c)]_+$$

mit einem $\alpha > 0$, so ist $x^* \in M$ eine Lösung von (P). Ist umgekehrt $x^* \in M$ eine Lösung von (P) und $\alpha > 0$ beliebig, so ist x^* eine Lösung von $(*)$. Bleibt diese Aussage richtig, wenn man α in $(*)$ durch eine positive Diagonalmatrix ersetzt?

Lösung: Sei $x^* = [x^* - \alpha(Qx^* + c)]_+$. Dann ist $x_j^* \geq 0$, $j = 1, \ldots, n_0$, also x^* zulässig für (P). Es ist

$$x_j^* = \begin{cases} \max(0, x_j^* - \alpha(Qx^* + c)_j), & j = 1, \ldots, n_0, \\ x_j^* - \alpha(Qx^* + c)_j), \end{cases}$$

woraus sofort

$$(Qx^* + c)_j \begin{cases} \geq 0 & (j = 1, \ldots, n_0), \\ = 0 & (j = n_0 + 1, \ldots, n), \end{cases} \qquad (x^*)^T(Qx^* + c) = 0$$

folgt. Dies sind die notwendigen und hinreichenden Bedingungen dafür, dass x^* (P) löst. Ist umgekehrt x^* eine Lösung von (P), so genügt x^* offenbar auch $(*)$. Offensichtlich kann, da komponentenweise argumentiert wird, α durch eine positive Diagonalmatrix ersetzt werden.

2. Sei $f_\alpha \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ definiert wie in Lemma 3.2, also durch

$$f_\alpha(x) := \frac{1}{2}x^T(I - \alpha Q)x - \frac{1}{2}\|[x - \alpha(Qx + c)]_+\|^2,$$

wobei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit ist und $\alpha > 0$ so klein gewählt ist, dass $\alpha \|Q\| < 1$. Man zeige:

   (a) Der Gradient ∇f_α ist auf dem $\mathbb{R}^n$ global lipschitzstetig, es existiert also eine Konstante $\gamma > 0$ mit

$$\|\nabla f_\alpha(x) - \nabla f_\alpha(y)\| \leq L \|x - y\| \qquad \text{für alle } x, y \in \mathbb{R}^n.$$

   (b) Ist Q sogar positiv definit, so ist bei beliebigem $x_0 \in \mathbb{R}^n$ die Niveaumenge

$$L_0 := \{x \in \mathbb{R}^n : f_\alpha(x) \leq f_\alpha(x_0)\}$$

   kompakt.

Lösung: Als Gradienten von f_α haben wir

$$\nabla f_\alpha(x) = (I - \alpha Q)\{x - [x - \alpha(Qx + c)]_+\}$$

in Lemma 3.2 erhalten. Zur Abkürzung setzen wir $B := I - \alpha Q$, $d := -\alpha c$, so dass

$$\nabla f_\alpha(x) = B[x - (Bx + d)_+].$$

Für beliebige $x, y \in \mathbb{R}^n$ ist ($\|\cdot\|$ bezeichne die euklidische Norm)

$$\begin{aligned} \|\nabla f_\alpha(x) - \nabla f_\alpha(y)\| &= \|B(x - y) - B[(Bx + d)_+ - (By + d)_+]\| \\ &\leq \|B\|(1 + \|B\|)\|x - y\|. \end{aligned}$$

Hierbei haben wir ausgenutzt, dass die Projektionsabbildung $z \mapsto z_+$ nicht expandierend ist. Also ist $\nabla f(\cdot)$ global lipschitzstetig auf dem $\mathbb{R}^n$. Offenbar haben wir in diesem Teil nicht ausgenutzt, dass $\alpha\|Q\| < 1$.

In Lemma 3.2 wurde gezeigt, dass für positiv semidefinites Q und $0 < \alpha\|Q\| < 1$ gilt, dass

$$f_\alpha(y) - f_\alpha(x) - \nabla f_\alpha(x)^T(y-x) \geq \frac{1}{2}(y-x)^T B(I-B)(y-x) \qquad \text{für alle } x, y \in \mathbb{R}^n,$$

wobei $B := I - \alpha Q$. Ist Q zusätzlich positiv definit, so ist $B(I-B)$ positiv definit. Mit der positiven Konstanten $c := \lambda_{\min}(B(I-B))$ ist also

$$f_\alpha(y) - f_\alpha(x) - \nabla f_\alpha(x)^T(y-x) \geq \frac{c}{2}\|y-x\|^2 \qquad \text{für alle } x, y \in \mathbb{R}^n,$$

insbesondere ist f_α auf dem $\mathbb{R}^n$ gleichmäßig konvex. Hieraus folgt aber natürlich die Kompaktheit von Niveaumengen

$$L_0 := \{x \in \mathbb{R}^n : f_\alpha(x) \leq f_\alpha(x_0)\}.$$

Denn ist $x \in L_0$, so ist

$$\begin{aligned} 0 &\geq f_\alpha(x) - f_\alpha(x_0) \\ &\geq \nabla f_\alpha(x_0)^T(x-x_0) + \frac{c}{2}\|x-x_0\|^2 \\ &\geq -\|\nabla f_\alpha(x_0)\|\,\|x-x_0\| + \frac{c}{2}\|x-x_0\|^2 \end{aligned}$$

und folglich

$$\|x - x_0\| \leq \frac{2\,\|\nabla f(x_0)\|}{c}.$$

Damit ist die Beschränktheit von L_0 bewiesen.

3. Gegeben[9] sei das quadratische Programm

(P) Minimiere $c^T x + \frac{1}{2}x^T Q x$ unter den Nebenbedingungen $l \leq Ax \leq u$.

Hierbei sei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, ferner seien $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^n$ und $l, u \in \mathbb{R}^m$ mit $l \leq u$ gegeben. Für einen Vektor $v \in \mathbb{R}^m$ seien die Vektoren v_+ bzw. $(v)_l^u$ in naheliegenderweise als Projektion von v auf den nichtnegativen Orthanten bzw. den Quader $[l, u]$ definiert. Man zeige:

(a) Für alle $v \in \mathbb{R}^m$ ist
$$(v)_l^u = v + (l-v)_+ - (v-u)_+.$$

(b) Es ist $x \in \mathbb{R}^n$ genau dann die Lösung von (P), wenn ein $w \in \mathbb{R}^m$ mit
$$Qx + c - A^T w = 0, \qquad Ax = (Ax - \alpha w)_l^u$$
existiert.

(c) Sei $\alpha > 0$ beliebig. Dann ist $x(w) := Q^{-1}(A^T w - c)$ mit einem $w \in \mathbb{R}^m$ genau dann die Lösung von (P), wenn $Ax(w) = (Ax(w) - \alpha w)_l^u$.

[9] Siehe auch

W. Li, J. Swetits (1997) "A new algorithm for solving strictly convex quadratic programs". SIAM J. Optimization 7, 595–619.

(d) Sei $\alpha > 0$ beliebig. Dann ist $x(w) := Q^{-1}(A^T w - c)$ mit einem $w \in \mathbb{R}^m$ genau dann die Lösung von (P), wenn

$$\phi_\alpha(w) := AQ^{-1}(A^T w - c) - [AQ^{-1}(A^T w - b) - \alpha w]_l^u = 0.$$

(e) Sei $\alpha > 0$. Zur Abkürzung setze man

$$B_\alpha := \alpha I - AQ^{-1}A^T, \qquad a := l + AQ^{-1}c, \qquad b := -(AQ^{-1}c + u).$$

Dann ist

$$\phi_\alpha(w) = \alpha w - (a + B_\alpha w)_+ + (b - B_\alpha w)_+.$$

(f) Mit $\alpha > 0$ definiere man $\Phi_\alpha\colon \mathbb{R}^m \longrightarrow \mathbb{R}$ durch

$$\Phi_\alpha(w) := \frac{\alpha}{2} w^T B_\alpha w - \frac{1}{2}\|(a + B_\alpha w)_+\|^2 - \frac{1}{2}\|(b - B_\alpha w)_+\|^2,$$

wobei $\|\cdot\|$ die euklidische Norm auf dem $\mathbb{R}^m$ bedeutet. Dann gilt:

i. Die Abbildung Φ_α ist auf dem $\mathbb{R}^m$ stetig partiell differenzierbar und besitzt den Gradienten

$$\nabla\Phi_\alpha(w) = B_\alpha[\alpha w - (a + B_\alpha w)_+ + (b - B_\alpha w)_+] = B_\alpha \phi_\alpha(w).$$

ii. Ist $\alpha > \|AQ^{-1}A^T\|$, so ist Φ_α auf dem $\mathbb{R}^m$ konvex, genauer ist

$$\Phi_\alpha(w) - \Phi_\alpha(v) - \nabla\Phi_\alpha(v)^T(w - v) \geq \frac{1}{2}(w - v)^T B_\alpha(\alpha I - B_\alpha)(w - v) \geq 0$$
$$\text{für alle } v, w \in \mathbb{R}^m.$$

Lösung: Die erste Aussage ist mehr oder weniger trivial. Ist $l_i < v_i < u_i$, so ist

$$[v + (l - v)_+ - (v - u)_+]_i = v_i = (v_l^u)_i,$$

für $l_i = v_i$ ist

$$[v + (l - v)_+ - (v - u)_+]_i = l_i = (v_l^u)_i,$$

entsprechendes gilt für $v_i = u_i$.

Sei

$$I := \{i \in \{1, \ldots, m\} : l_i = u_i\}, \qquad J := \{1, \ldots, m\} \setminus I.$$

Die Restriktionen zur Indexmenge I sind Gleichungen, bei ihnen kann man also keine Aussagen über das Vorzeichen zugehöriger Lagrange-Multiplikatoren machen. Der Satz von Kuhn-Tucker liefert, dass ein für (P) zulässiges $x \in \mathbb{R}^n$ genau dann die Lösung von (P) ist, wenn ein Tripel

$$(\lambda_I, \mu_J, \nu_J) \in \mathbb{R}^{\#(I)} \times \mathbb{R}^{\#(J)} \times \mathbb{R}^{\#(J)}$$

mit

(a) $\mu_J \geq 0$, $\nu_J \geq 0$,

(b) $Qx + c - A_I^T\lambda_I - A_J^T\mu_J + A_J^T\nu_J = 0$,

(c) $(l_J - A_J x)^T\mu_J = 0$, $(A_J x - u_J)^T\nu_J = 0$

existiert. Die hierbei benutzten Bezeichnungen sollten für sich sprechen. Die Gleichgewichtsbedingung (c) kann auch komponentenweise durch

(c') $(l - Ax)_i \mu_i = 0$, $(Ax - u)_i \nu_i = 0$ für alle $i \in J$

ausgedrückt werden. Hieran erkennt man, dass $\mu_i = 0$ oder $\nu_i = 0$ für alle $i \in J$.

Nun sei x eine Lösung von (P), ferner $\alpha > 0$ beliebig und

$$(\lambda_I, \mu_J, \nu_J) \in \mathbb{R}^{\#(I)} \times \mathbb{R}^{\#(J)} \times \mathbb{R}^{\#(J)}$$

ein nach dem Satz von Kuhn-Tucker existierendes Tripel mit obigen Eigenschaften (a)–(c). Wir definieren $w \in \mathbb{R}^m$ durch

$$w_i := \begin{cases} \lambda_i, & \text{falls} \quad i \in I, \\ \mu_i, & \text{falls} \quad i \in J,\ l_i = (Ax)_i, \\ 0, & \text{falls} \quad i \in J,\ l_i < (Ax)_i < u_i, \\ -\nu_i, & \text{falls} \quad i \in J,\ (Ax)_i = u_i. \end{cases}$$

Offenbar ist dann $Qx + c - A^T w = 0$. Zu zeigen bleibt $Ax = (Ax - \alpha w)_l^u$, was komponentenweise sehr einfach geschieht.

Die umgekehrte Aussage ist auch sehr einfach zu zeigen. Wir nehmen also an, es sei $(x, w) \in \mathbb{R}^n \times \mathbb{R}^m$ ein Paar mit

$$Qx + c - A^T w = 0, \qquad Ax = (Ax - \alpha w)_l^u.$$

Aus der zweiten Beziehung erhält man sofort, dass x für (P) zulässig ist. Aus ihr folgt ferner, dass für $i \in J$ gilt:

$$w_i \begin{cases} \geq 0, & \text{falls} \quad l_i = (Ax)_i, \\ = 0, & \text{falls} \quad l_i < (Ax)_i < u_i, \\ \leq 0, & \text{falls} \quad (Ax)_i = u_i. \end{cases}$$

Definieren wir daher das Tripel

$$(\lambda_I, \mu_J, \nu_J) \in \mathbb{R}^{\#(I)} \times \mathbb{R}^{\#(J)} \times \mathbb{R}^{\#(J)}$$

durch

$$\lambda_i := w_i \quad (i \in I)$$

und

$$\mu_i := \begin{cases} w_i, & \text{falls} \quad i \in J,\ l_i = (Ax)_i, \\ 0, & \text{falls} \quad i \in J,\ l_i < (Ax)_i, \end{cases} \qquad \nu_i := \begin{cases} -w_i, & \text{falls} \quad i \in J,\ (Ax)_i = u_i, \\ 0, & \text{falls} \quad i \in J,\ (Ax)_i < u_i, \end{cases}$$

so sind die Kuhn-Tucker Bedingungen erfüllt und daher x die Lösung von (P).

Der dritte und vierte Teil der Aufgabe sind jeweils eine direkte Folgerung aus dem zweiten.

Zum Beweis von (e) beachten wir unter Benutzung von (a), dass

$$\begin{aligned}
\phi_\alpha(w) &:= AQ^{-1}(A^Tw-c) - [AQ^{-1}(A^Tw-c)-\alpha w]_l^u \\
&= AQ^{-1}(A^Tw-c) - [AQ^{-1}(A^Tw-c)-\alpha w] \\
&\qquad - [l - AQ^{-1}(A^Tw-c)+\alpha w]_+ + [AQ^{-1}(A^Tw-c)-\alpha w - u]_+ \\
&= \alpha w - [l + AQ^{-1}c + (\alpha I - AQ^{-1}A^T)w]_+ \\
&\qquad + [-AQ^{-1}c - u - (\alpha I - AQ^{-1}A^T)w]_+ \\
&= \alpha w - (a + B_\alpha w)_+ + (b - B_\alpha w)_+.
\end{aligned}$$

Nun kommen wir zum Beweis von (f). Der erste Teil ist offensichtlich, wenn man benutzt, dass die Abbildung $h(t) := \frac{1}{2}(t_+)^2$ stetig differenzierbaar mit $h'(t) = t_+$ ist. Für den zweiten Teil des Satzes beachten wir, dass alle Eigenwerte von

$$B_\alpha := \alpha I - AQ^{-1}A^T$$

für $\alpha > \|AQ^{-1}A^T\|$ in $(0, \alpha]$ liegen, insbesondere also B_α positiv definit und $B_\alpha(\alpha I - B_\alpha)$ positiv semidefinit ist. Für beliebige $v, w \in \mathbb{R}^m$ ist dann

$$\begin{aligned}
\Phi_\alpha(w) - \Phi_\alpha(v) - \nabla\Phi_\alpha(v)^T(w-v) &= \frac{\alpha}{2}w^TB_\alpha w - \frac{\alpha}{2}v^TB_\alpha v \\
&\quad - \frac{1}{2}\|(a+B_\alpha w)_+\|^2 + \frac{1}{2}\|(a+B_\alpha v)_+\|^2 \\
&\quad - \frac{1}{2}\|(b-B_\alpha w)_+\|^2 + \frac{1}{2}\|(b-B_\alpha v)_+\|^2 \\
&\quad - [B_\alpha\phi_\alpha(v)]^T(w-v) \\
&= \alpha(B_\alpha v)^T(w-v) + \frac{\alpha}{2}(w-v)^TB_\alpha(w-v) \\
&\quad - \frac{1}{2}\|(a+B_\alpha w)_+\|^2 + \frac{1}{2}\|(a+B_\alpha v)_+\|^2 \\
&\quad - \frac{1}{2}\|(b-B_\alpha w)_+\|^2 + \frac{1}{2}\|(b-B_\alpha v)_+\|^2 \\
&\quad - [B_\alpha\phi_\alpha(v)]^T(w-v) \\
&= \frac{\alpha}{2}(w-v)^TB_\alpha(w-v) \\
&\quad - \frac{1}{2}\|(a+B_\alpha w)_+\|^2 + \frac{1}{2}\|(a+B_\alpha v)_+\|^2 \\
&\quad - \frac{1}{2}\|(b-B_\alpha w)_+\|^2 + \frac{1}{2}\|(b-B_\alpha v)_+\|^2 \\
&\quad + [(a+B_\alpha v)_+ - (b-B_\alpha v)_+]^TB_\alpha(w-v).
\end{aligned}$$

Zur Abkürzung setzen wir

$$p := a + B_\alpha v, \quad q := b - B_\alpha v, \quad r := a + B_\alpha w, \quad s := b - B_\alpha w.$$

Dann ist

$$q + p = s + r = l - u \leq 0$$

und

$$r - p = -(s - q) = B_\alpha(w - v).$$

Wir wollen zeigen, daß

$$f_i := -\frac{1}{2}(r_+)_i^2 + \frac{1}{2}(p_+)_i^2 - \frac{1}{2}(s_+)_i^2 + \frac{1}{2}(q_+)_i^2 + [(p_+)_i - (q_+)_i](r_i - p_i) \geq -\frac{1}{2}(r_i - p_i)^2$$

für $i = 1, \ldots, m$. Durch Aufsummieren folgt dann die Behauptung. Sei $i \in \{1, \ldots, m\}$ fest vorgegeben. Wir machen eine Fallunterscheidung, wobei wir naheliegende Bezeichnungen benutzen: + bzw. − bedeutet, dass die entsprechende Zahl positiv bzw. negativ ist. Den Fall, dass die entsprechende Zahl verschwindet, erhält man aus einer Stetigkeitüberlegung.

- Es ist $(p_i, q_i, r_i, s_i) = (+, +, \pm, \pm)$ oder $(p_i, q_i, r_i, s_i) = (\pm, \pm, +, +)$.
  Diese Fälle sind nicht möglich, da $p_i + q_i \leq 0$ und $r_i + s_i \leq 0$.
- Es ist $(p_i, q_i, r_i, s_i) = (+, -, +, -)$.
  Dann ist
  $$f_i = -\frac{1}{2}r_i^2 + \frac{1}{2}p_i^2 + p_i(r_i - p_i) = -\frac{1}{2}(r_i - p_i)^2.$$
- Es ist $(p_i, q_i, r_i, s_i) = (+, -, -, +)$.
  Dann ist
  $$f_i = \frac{1}{2}p_i^2 + p_i(r_i - p_i) = -\frac{1}{2}(r_i - p_i)^2 + \frac{1}{2}r_i^2 \geq -\frac{1}{2}(r_i - p_i)^2.$$
- Es ist $(p_i, q_i, r_i, s_i) = (+, -, -, -)$.
  Dann ist
  $$f_i = \frac{1}{2}p_i^2 + p_i(r_i - p_i) = -\frac{1}{2}(r_i - p_i)^2 + \frac{1}{2}r_i^2 \geq -\frac{1}{2}(r_i - p_i)^2.$$
- Es ist $(p_i, q_i, r_i, s_i) = (-, +, +, -)$.
  Dann ist
  $$f_i = -\frac{1}{2}r_i^2 + \frac{1}{2}q_i^2 - q_i(r_i - p_i) = -\frac{1}{2}(r_i - p_i)^2 + \frac{1}{2}(p_i + q_i)^2 - \underbrace{(p_i + q_i)r_i}_{\leq 0} \geq -\frac{1}{2}(r_i - p_i)^2.$$
- Es ist $(p_i, q_i, r_i, s_i) = (-, +, -, +)$.
  Dann ist
  $$f_i = -\frac{1}{2}s_i^2 + \frac{1}{2}q_i^2 - q_i(r_i - p_i) = -\frac{1}{2}s_i^2 + \frac{1}{2}q_i^2 + q_i(s_i - q_i) = -\frac{1}{2}(r_i - p_i)^2.$$
- Es ist $(p_i, q_i, r_i, s_i) = (-, +, -, -)$.
  Dann ist
  $$f_i = \frac{1}{2}q_i^2 - q_i(r_i - p_i) = \frac{1}{2}q_i^2 + q_i(s_i - q_i) = -\frac{1}{2}(s_i - q_i)^2 + \frac{1}{2}s_i^2 \geq -\frac{1}{2}(r_i - p_i)^2.$$
- Es ist $(p_i, q_i, r_i, s_i) = (-, -, +, -)$.
  Dann ist
  $$f_i = -\frac{1}{2}r_i^2 = -\frac{1}{2}(r_i - p_i)^2 + \frac{1}{2}p_i^2 - \underbrace{p_i r_i}_{\leq 0} \geq -\frac{1}{2}(r_i - p_i)^2.$$

- Es ist $(p_i, q_i, r_i, s_i) = (-, -, -, +)$.
  Dann ist
$$f_i = -\frac{1}{2}s_i^2 = -\frac{1}{2}(s_i - q_i)^2 = \frac{1}{2}q_i^2 - \underbrace{s_i q_i}_{\leq 0} \geq -\frac{1}{2}(r_i - p_i)^2.$$

- Es ist $(p_i, q_i, r_i, s_i) = (-, -, -, -)$.
  Dann ist
$$f_i = 0 \geq -\frac{1}{2}(r_i - p_i)^2.$$

Das sind alle $2^4 = 16$ möglichen Fälle und die Behauptung ist bewiesen.

6.4 Aufgaben in Kapitel 4

6.4.1 Aufgaben in Abschnitt 4.1

1. Ist $D \subset \mathbb{R}^n$ konvex, so heißt eine Funktion $f: D \longrightarrow \mathbb{R}$ bekanntlich auf D *gleichmäßig konvex*, wenn eine Konstante $c > 0$ mit
$$(1-\lambda)f(x_1) + \lambda f(x_2) - f((1-\lambda)x_1 + \lambda x_2) \geq \frac{c}{2}\lambda(1-\lambda)\,\|x_1 - x_2\|^2$$
für alle $x_1, x_2 \in D$, $\lambda \in [0,1]$ existiert.

   Gegeben sei die Optimierungsaufgabe

   (P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : Ax = b\}$.

   Hierbei seien $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$ und $b \in \mathbb{R}^m$ gegeben. Wie in Unterabschnitt 4.1.1 geschildert ordne man (P) die unrestringierte Optimierungsaufgabe

   (P_x) Minimiere $\psi(u) := f(x + Zu)$, $u \in \mathbb{R}^{n-m}$,

   zu, wobei x zulässig für (P) und die Spalten von $Z \in \mathbb{R}^{n\times(n-m)}$ (mit Rang $(Z) = n-m$) eine Basis von Kern (A) bilden. Man zeige: Ist f gleichmäßig konvex auf M, so ist ψ gleichmäßig konvex auf $\mathbb{R}^{n-m}$.

   Lösung: Seien $u_1, u_2 \in \mathbb{R}^{n-m}$ und $\lambda \in [0,1]$. Dann ist
$$\begin{aligned}
& (1-\lambda)\psi(u_1) + \lambda\psi(u_2) - \psi((1-\lambda)u_1 + \lambda u_2) \\
= \;& (1-\lambda)f(x+Zu_1) + \lambda f(x+Zu_2) - f((1-\lambda)(x+Zu_1) + \lambda(x+Zu_2)) \\
\geq \;& \frac{c}{2}\lambda(1-\lambda)\,\|Z(u_1-u_2)\|^2 \\
\geq \;& \frac{cd}{2}\lambda(1-\lambda)\,\|u_1-u_2\|^2,
\end{aligned}$$
   wobei die wegen Rang $(Z) = n - m$ positive Zahl d durch
$$d := \min_{u\neq 0} \frac{\|Zu\|}{\|u\|}$$
   definiert ist. Damit ist gezeigt, dass ψ auf $\mathbb{R}^{n-m}$ gleichmäßig konvex ist.

2. Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ gegeben (bei der Anwendung in Unterabschnitt 4.1.1 ist n durch $n - m$ zu ersetzen). Es sei eine Cholesky-Zerlegung von B bekannt, also eine untere Dreiecksmatrix L mit positiven Diagonalelementen mit $B = LL^T$. Ferner sei

$$B_+ := B - \frac{(Bs)(Bs)^T}{s^T Bs} + \frac{yy^T}{y^T s}.$$

Man zeige:

(a) Ist

$$w := (y^T s)^{1/2} \frac{L^T s}{\|L^T s\|}, \qquad J_+^T := L^T + \frac{w(y - Lw)^T}{y^T s},$$

so ist $B_+ = J_+ J_+^T$.

(b) Die Matrix J_+ ist nichtsingulär und daher B_+ positiv definit.

(c) Ist $J_+^T = Q_+ R_+$ eine QR-Zerlegung von J_+^T, wobei (Q_+ orthogonal und) R_+ eine obere Dreiecksmatrix mit positiven Diagonalelementen ist, so ist $B_+ = L_+ L_+^T$ mit $L_+ := R_+^T$ eine Cholesky-Zerlegung von B_+.

(d) Die QR-Zerlegung einer durch eine Matrix vom Rang 1 gestörten oberen Dreiecksmatrix kann in $O(n^2)$ Flops berechnet werden.

Lösung: Den ersten Teil der Aufgabe löst man durch Nachrechnen, wobei wir benutzen, dass

$$Lw = \left(\frac{y^T s}{s^T Bs}\right)^{1/2} Bs.$$

Es ist nämlich

$$\begin{aligned}
J_+ J_+^T &= \left(L + \frac{(y - Lw)w^T}{y^T s}\right)\left(L^T + \frac{w(y - Lw)^T}{y^T s}\right) \\
&= LL^T + \frac{Lw(y - Lw)^T}{y^T s} + \frac{(y - Lw)(Lw)^T}{y^T s} + \left(\frac{\|w\|}{y^T s}\right)^2 (y - Lw)(y - Lw)^T \\
&= LL^T + \frac{Lw(y - Lw)^T}{y^T s} + \frac{(y - Lw)(Lw)^T}{y^T s} + \frac{(y - Lw)(y - Lw)^T}{y^T s} \\
&= LL^T - \frac{(Lw)(Lw)^T}{y^T s} + \frac{yy^T}{y^T s} \\
&= B - \frac{(Bs)(Bs)^T}{s^T Bs} + \frac{yy^T}{y^T s} \\
&= B_+.
\end{aligned}$$

Damit ist die erste Aussage bewiesen. Wegen

$$\sigma := 1 + \frac{w^T(L^{-1}y - w)}{y^T s} = \frac{w^T L^{-1} y}{y^T s} = \left(\frac{y^T s}{s^T Bs}\right)^{1/2} \neq 0$$

ist J_+ nach der Sherman-Morrison-Formel nichtsingulär. Ist $J_+^T = Q_+ R_+$ eine QR-Zerlegung von J_+^T und $L_+ := R_+^T$, so ist

$$B_+ = J_+ J_+^T = R_+^T \underbrace{Q_+^T Q_+}_{=I} R_+ = L_+ L_+^T,$$

womit auch der einfache dritte Teil bewiesen ist. Für den letzten Teil der Aufgabe nehmen wir an, es sei $A_+ = R + uv^T$ eine Störung vom Rang 1 der oberen Dreiecksmatrix R. Sei $m := \max\{i \in \{1,\ldots,n\} : u_i \neq 0\}$. Zunächst führt man den Vektor u durch sukzessive Multiplikation mit $m-1$ geeigneten Givensrotationen $G_{m-1,m},\ldots,G_{12}$, welche der Reihe nach die Komponenten mit den Indizes $m,\ldots,2$ annullieren, in ein Vielfaches u_1e_1 des ersten Einheitsvektors über. Die parallel hierzu durchgeführte Multiplikation der oberen Dreiecksmatrix R mit den Givensrotationen $G_{m-1,m},\ldots,G_{12}$ transformiert diese in eine obere Hessenberg-Matrix, die wir mit H bezeichnen. Nach Abschluss dieses ersten Schrittes ist $G_{12}\cdots G_{m-1,m}A_+ = H + u_1e_1v^T$ mit einer oberen Hessenberg-Matrix (deren Subdiagonalelemente in den Spalten $m,\ldots,n-1$ verschwinden. In einem Zwischenschritt berechnet man $H := H + u_1e_1v^T$, wodurch nur die erste Zeile verändert wird. Durch Multiplikation mit weiteren $m-1$ Givens-Rotationen $G_{12},\ldots,G_{m-1,m}$ annulliert man schließlich in einem letzten Schritt die störenden Subdiagonalelemente in den Spalten $1,\ldots,m-1$. Hierbei hat man darauf zu achten, dass die erzeugten Diagonalelemente positiv sind. Die berechnete obere Dreiecksmatrix R_+ erhält man offenbar in $O(n^2)$ flops.

6.4.2 Aufgaben in Abschnitt 4.2

1. Gegeben sei eine linear restringierte nichtlineare Optimierungsaufgabe mit einer stetig differenzierbaren Zielfunktion. Man zeige, dass eine zulässige Lösung genau dann eine kritische Lösung ist, wenn es in ihr keine zulässige Abstiegsrichtung gibt.

   Lösung: Gegeben sei die Optimierungsaufgabe

   $$\text{(P)}\quad \text{Minimiere}\quad f(x)\quad \text{auf}\quad M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1,\ldots,m_0) \\ a_i^T x = b_i & (i = m_0+1,\ldots,m) \end{array} \right\},$$

   wobei die Zielfunktion f stetig differenzierbar ist. Die Matrix $A \in \mathbb{R}^{m\times n}$ mit den Zeilen a_i^T und der Vektor $b \in \mathbb{R}^m$ mit den Komponenten b_i seien wie gewohnt definiert. Ein $x \in M$ ist eine kritische Lösung von (P), wenn ein $y \in \mathbb{R}^m$ mit

   $$y_i \geq 0 \quad (i = 1,\ldots,m_0), \qquad \nabla f(x) = A^T y, \qquad y^T(Ax - b) = 0$$

   existiert. Ferner ist $p \in \mathbb{R}^n$ eine in $x \in M$ zulässige Abstiegsrichtung, wenn $\nabla f(x)^T p < 0$ und

   $$a_i^T p \geq 0 \quad (i \in I(x)), \qquad a_i^T p = 0 \quad (i = m_0+1,\ldots,m),$$

   wobei $I(x)$ die Indexmenge der in x aktiven Ungleichungsrestriktionen bedeutet.

   Sei $x \in M$ eine kritische Lösung. Gäbe es eine in x zulässige Abstiegsrichtung p, so wäre

   $$0 > \nabla f(x)^T p = (A^T y)^T p = y^T Ap = \sum_{i\in I(x)} y_i a_i^T p \geq 0,$$

   ein Widerspruch.

   In $x \in M$ gebe es keine zulässige Abstiegsrichtung. Dann ist das System

   $$\begin{pmatrix} A_{I(x)} \\ A_= \end{pmatrix} p \in \mathbb{R}^q_{\geq 0} \times \{0\}, \qquad \nabla f(x)^T p < 0$$

nicht lösbar. Hierbei ist $q := \#(I(x))$, ferner sei $A_= \in \mathbb{R}^{(m-m_0)\times n}$ die Untermatrix von A, die zu den Gleichungsrestriktionen gehört. Das verallgemeinerte Farkas-Lemma liefert die Existenz eines Paares $(y_{I(x)}, y_=) \in \mathbb{R}^q \times \mathbb{R}^{m-m_0}$ mit

$$y_{I(x)} \geq 0, \qquad \nabla f(x) = A_{I(x)}^T y_{I(x)} + A_=^T y_=,$$

d. h. $x \in M$ ist eine kritische Lösung von (P).

2. Gegeben sei die Optimierungsaufgabe

(P) $$\text{Minimiere} \quad f(x) \quad \text{auf} \quad M,$$

wobei $M \subset \mathbb{R}^n$ konvex ist. Sei $x^* \in M$ und die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in x^* stetig differenzierbar. Man zeige:

(a) Ist x^* eine lokale Lösung von (P), so ist $\nabla f(x^*)^T(x - x^*) \geq 0$ für alle $x \in M$.

(b) Sei

$$M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$

Dann ist $\nabla f(x^*)^T(x - x^*) \geq 0$ für alle $x \in M$ genau dann, wenn x^* eine kritische Lösung von (P) ist, also ein $y^* \in \mathbb{R}^m$ mit

$$y_i^* \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x^*) = A^T y^*, \qquad (y^*)^T(Ax^* - b) = 0$$

existiert. Hierbei ist, wie stets in diesem Zusammenhang, $A \in \mathbb{R}^{m\times n}$ die Matrix, die a_i^T als i-te Zeile besitzt, ferner ist b_i die i-te Komponente von $b \in \mathbb{R}^m$.

Lösung: Die erste Aussage der Aufgabe, dass nämlich $\nabla f(x^*)^T(x - x^*) \geq 0$ für alle $x \in M$ eine notwendige Bedingung dafür ist, dass $x^* \in M$ eine lokale Lösung von (P) ist, ist schon lange bekannt, darauf wird nicht noch einmal eingegangen. Sei daher jetzt M der angegebene Polyeder.

Der zweite Teil der Aufgabe folgt sofort, wenn man beachtet, dass der Kegel $F(M; x^*)$ der in x^* zulässigen Richtungen durch (hier benötigt man nur die Konvexität von M)

$$F(M; x^*) = \{\lambda(x - x^*) : \lambda \geq 0,\ x \in M\}$$

gegeben ist, und die Aussage der vorigen Aufgabe benutzt.

3. Man zeige: Genügt die Zielfunktion f von (P) den Voraussetzungen (V) (a)–(c), so existiert eine Konstante $\theta_C > 0$ derart, dass

$$\begin{aligned} f(x) - f(x + t_M(x,p)p) &\geq f(x) - f(x + t_C(x,p)p) \\ &\geq \theta_C \min\left[-s(x,p)\, \nabla f(x)^T p, \left(\frac{\nabla f(x)^T p}{\|p\|} \right)^2 \right] \end{aligned}$$

für alle nicht kritischen $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$. Hierbei bedeutet $t_M = t_M(x,p)$ die Minimum-Schrittweite, $t_C = t_C(x,p)$ die Curry-Schrittweite und $s = s(x,p)$ die maximale Schrittweite in x in Richtung p, ferner $\|\cdot\|$ die euklidische Norm.

Lösung: Zu zeigen ist natürlich nur die zweite Ungleichung, da die erste nach Definition der Minimum-Schrittweite trivial ist. Zur Abkürzung sei

$$\psi(t) := \tfrac{1}{2}\,\phi'(0) - \phi'(t) = \tfrac{1}{2}\,\nabla f(x)^T p - \nabla f(x+tp)^T p.$$

Ferner sei $\tilde{t}(x,p)$ die erste Nullstelle von $\psi(\cdot)$ in $(0, t_C(x,p)]$, falls eine solche existiert, andernfalls sei $\tilde{t}(x,p) := t_C(x,p)$. Offenbar ist $\psi(t) > 0$ bzw. $-\nabla f(x+tp)^T p > -\frac{1}{2}\,\nabla f(x)^T p$ für alle $t \in (0, \tilde{t}(x,p))$. Dann erhält man

$$\begin{aligned} f(x) - f(x + t_C(x,p)p) &\geq f(x) - f(x + \tilde{t}(x,p)p) \\ &= -\tilde{t}(x,p)\,\nabla f(x + \theta\tilde{t}(x,p)p)^T p \qquad \text{mit } \theta \in (0,1) \\ &\geq -\tfrac{1}{2}\,\tilde{t}(x,p)\,\nabla f(x)^T p. \end{aligned}$$

Nun machen wir eine Fallunterscheidung. Ist nämlich $\tilde{t}(x,p)$ die erste Nullstelle von $\psi(\cdot)$ in $(0, t_C(x,p)]$, so ist

$$\tfrac{1}{2}\,\nabla f(x)^T p = \left[\nabla f(x) - \nabla f(x + \tilde{t}(x,p)p)\right]^T p \geq -\tilde{t}(x,p)\,\gamma\,\|p\|^2,$$

woraus

$$\tilde{t}(x,p) \geq -\frac{1}{2\gamma}\left(\frac{\nabla f(x)^T p}{\|p\|^2}\right)$$

und damit

$$f(x) - f(x + t_C(x,p)p) \geq \frac{1}{4\gamma}\left(\frac{\nabla f(x)^T p}{\|p\|}\right)^2$$

folgt. Ist dagegen $\psi(t) > 0$ für alle $t \in (0, t_C(x,p)]$ und damit $\tilde{t}(x,p) = t_C(x,p)$, so ist $t_C(x,p) = s(x,p)$ (andernfalls wäre $\psi(t_C(x,p)) = \frac{1}{2}\,\nabla f(x)^T p < 0$) und damit

$$f(x) - f(x + t_C(x,p)p) \geq -\tfrac{1}{2}\,s(x,p)\,\nabla f(x)^T p.$$

Insgesamt ist die Aussage bewiesen.

4. Man zeige: Genügt die Zielfunktion f von (P) den Voraussetzungen (V) (a)–(c), so existiert eine Konstante $\theta_P > 0$ derart, dass

$$f(x) - f(x + t_P(x,p))p) \geq \theta_C\,\min\left[-s(x,p)\,\nabla f(x)^T p, \left(\frac{\nabla f(x)^T p}{\|p\|}\right)^2\right]$$

für alle nicht kritischen $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$. Hierbei bedeutet $t_P(x,p)$ die Powell-Schrittweite und $s(x,p)$ die maximale Schrittweite in x in Richtung p, ferner $\|\cdot\|$ die euklidische Norm.

Lösung: Bei der Powell-Schrittweite sind $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ vorgegeben. Man setzt $t_P(x,p) := s(x,p)$, falls

$$s(x,p) < \infty, \qquad f(x + s(x,p)p) \leq f(x) + \alpha s(x,p)\,\nabla f(x)^T p.$$

In diesem Falle ist also

$$f(x) - f(x + t_P(x,p)p) \geq -\alpha s(x,p)\nabla f(x)^T p.$$

Andernfalls wähle man $t_P(x,p) \in (0, s(x,p))$ beliebig mit

$$f(x + t_P(x,p)p) \leq f(x) + \alpha t_P(x,p)\,\nabla f(x)^T p, \qquad \nabla f(x + t_P(x,p)p)^T p \geq \beta\,\nabla f(x)^T p.$$

Zur Abkürzung setzen wir

$$\psi(t) := f(x) + \alpha t \nabla f(x)^T p - f(x + tp).$$

Wir machen eine Fallunterscheidung. Ist nämlich $t_P(x,p) \le t_C(x,p)$, so folgt aus

$$-(1-\beta)\,\nabla f(x)^T p \le [\nabla f(\underbrace{x + t_P(x,p)p}_{\in L_0}) - \nabla f(x)]^T p \le t_P(x,p)\,\gamma\,\|p\|^2,$$

daß

$$f(x) - f(x + t_P(x,p)p) \ge -\alpha\, t_P(x,p)\,\nabla f(x)^T p \ge \left(\frac{\alpha(1-\beta)}{\gamma}\right)\left(\frac{\nabla f(x)^T p}{\|p\|}\right)^2.$$

Ist dagegegen $t_C(x,p) < t_P(x,p)$, insbesondere also $t_C(x,p) < s(x,p)$, so ist

$$-\nabla f(x)^T p = [\nabla f(x) - \nabla f(x + t_C(x,p)p)]^T p \le t_C(x,p)\,\gamma\,\|p\|^2$$

und daher

$$f(x) - f(x + t_P(x,p)p) \ge \left(\frac{\alpha}{\gamma}\right)\left(\frac{\nabla f(x)^T p}{\|p\|}\right)^2.$$

Insgesamt ist die Behauptung bewiesen.

5. Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Dann existiert eine Konstante $\theta_A > 0$ derart, daß

$$f(x) - f(x + t_A(x,p)p) \ge \theta_A \,\min\left[-\tilde{s}(x,p)\,\nabla f(x)^T p, \left(\frac{\nabla f(x)^T p}{\|p\|}\right)^2\right]$$

für alle nicht kritischen $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$. Hierbei bedeutet $t_A(x,p)$ die Armijo-Schrittweite und $\tilde{s}(x,p) := \min(s(x,p), 1)$ die eventuell reduzierte maximale Schrittweite, ferner $\|\cdot\|$ die euklidische Norm.

Lösung: Ist der Test

$$(*) \qquad f(x + \rho_j p) \le f(x) + \alpha\rho_j\,\nabla f(x)^T p$$

schon für $j = 0$ erfüllt, so ist $t_A(x,p) = \rho_0 = \tilde{s}(x,p)$ und daher

$$f(x) - f(x + t_A(x,p)p) \ge -\alpha\tilde{s}(x,p)\,\nabla f(x)^T p.$$

Andernfalls (d. h. der Test $(*)$ ist erst für ein $j > 0$ erfüllt) gelten die beiden Ungleichungen

$$f(x + \rho_j p) \le f(x) + \alpha\rho_j\,\nabla f(x)^T p, \qquad f(x + \rho_{j-1} p) > f(x) + \alpha\rho_{j-1}\,\nabla f(x)^T p.$$

Wir machen eine Fallunterscheidung. Für $\rho_{j-1} \le t_C(x,p)$ ist

$$f(x) + \alpha\rho_{j-1}\,\nabla f(x)^T p < f(x + \rho_{j-1}p) \le f(x) + \rho_{j-1}\nabla f(x)^T p + \rho_{j-1}^2\,\frac{\gamma}{2}\,\|p\|^2,$$

daher

$$\rho_j \ge l\rho_{j-1} \ge \frac{2l(\alpha-1)}{\gamma}\,\frac{\nabla f(x)^T p}{\|p\|^2}$$

und folglich

$$f(x) - f(x + t_A(x,p)p) \geq -\alpha\rho_j \nabla f(x)^T p \geq \frac{2\alpha(1-\alpha)l}{\gamma} \left(\frac{\nabla f(x)^T p}{\|p\|} \right)^2.$$

Ist dagegen $s(x,p) \geq \rho_{j-1} > t_C(x,p)$, so ist

$$\rho_j \geq l\rho_{j-1} > l\, t_C(x,p) \geq -\left(\frac{l}{\gamma}\right) \frac{\nabla f(x)^T p}{\|p\|^2}$$

und folglich

$$f(x) - f(x + t_A(x,p)p) \geq -\alpha\rho_j \nabla f(x)^T p \geq \frac{\alpha l}{\gamma} \left(\frac{\nabla f(x)^T p}{\|p\|} \right)^2.$$

Hierbei haben wir ausgenutzt, dass

$$t_C(x,p) \geq -\frac{1}{\gamma}\left(\frac{\nabla f(x)^T p}{\|p\|} \right),$$

falls $s(x,p) < \infty$. Dies wiederum folgt sofort aus

$$-\nabla f(x)^T p = [\nabla f(x + t_C(x,p)p) - \nabla f(x)]^T p \leq \gamma t_C(x,p)\, \|p\|^2.$$

Insgesamt ist die Behauptung bewiesen.

6. Gegeben sei das linear restringierte Programm

$$\text{(P)} \quad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$

Die Menge der zulässigen Lösungen M sei nichtleer und kompakt, ferner seien die üblichen Voraussetzungen (V) (a)–(c) erfüllt. Man betrachte das Verfahren von Frank-Wolfe:

- Für $k = 0, 1, \ldots$:
  - Sei p_k eine Lösung des linearen Programms

$$\left\{ \begin{array}{c} \text{Minimiere} \quad \nabla f(x_k)^T p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \geq b_i - a_i^T x_k \quad (i = 1, \ldots, m_0), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m). \end{array} \right.$$

  - Falls $\nabla f(x_k)^T p_k = 0$, dann: STOP, x_k ist kritische Lösung von (P).
  - Berechne $t_k := t_M(x_k, p_k)$, $t_C(x_k, p_k)$, $t_P(x_k, p_k)$ oder $t_A(x_k, p_k)$.
  - Setze $x_{k+1} := x_k + t_k p_k$.

Dann gilt: Bricht das Verfahren nicht vorzeitig mit einer kritischen Lösung von (P) ab, so liefert es eine Folge $\{x_k\}$ mit der Eigenschaft, dass jeder Häufungspunkt von $\{x_k\}$ eine kritische Lösung von (P) ist.

Lösung: Wir müssen uns zunächst überlegen, dass das Frank-Wolfe-Verfahren ein durchführbares Verfahren der zulässigen Richtungen ist. Sei hierzu $x_k \in M$ eine aktuelle Näherung. Das lineare Programm

$$\left\{ \begin{array}{c} \text{Minimiere} \quad \nabla f(x_k)^T p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \geq b_i - a_i^T x_k \quad (i = 1, \ldots, m_0), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m) \end{array} \right.$$

besitzt eine Lösung, denn die zugehörige Menge der zulässigen Lösungen ist $M - x_k$, also nichtleer und wegen der vorausgesetzten Kompaktheit von M auch kompakt. Da $p = 0$ zulässig ist, ist $\nabla f(x_k)^T p_k \leq 0$. Eine Lösung p_k ist charakterisiert durch die Existenz eines Vektors $y \in \mathbb{R}^m$ mit

$$y_i \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x_k) = \sum_{i=1}^{m} y_i a_i$$

und

$$y_i(a_i^T p_k + a_i^T x_k - b_i) = 0 \quad (i = 1, \ldots, m_0).$$

Ist nun $\nabla f(x_k)^T p_k = 0$, so ist

$$0 = \nabla f(x_k)^T p_k = \sum_{i=1}^{m} y_i a_i^T p_k = \sum_{i=1}^{m_0} y_i a_i^T p_k + \sum_{i=m_0+1}^{m} y_i \underbrace{a_i^T p_k}_{=0} = \sum_{i=1}^{m_0} \underbrace{y_i(b_i - a_i^T x_k)}_{\leq 0}$$

und folglich

$$y_i \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x_k) = \sum_{i=1}^{m} y_i a_i, \qquad y_i(a_i^T x_k - b_i) = 0 \quad (i = 1, \ldots, m_0),$$

also x_k eine kritische Lösung von (P). Das STOP-Kriterium besteht also zu Recht, wir können im weiteren annehmen, dass $\nabla f(x_k)^T p_k < 0$ für alle k. Die Folge $\{p_k\}$ ist beschränkt, da M kompakt und $x_k + p_k \in M$. Es ist $s(x_k, p_k) \geq 1$, da $x_k + p_k \in M$ und M konvex ist. Aus den Lemmata 2.1, 2.2, 2.3 bzw. den Aufgaben 3, 4, 5 erhält man die Existenz einer Konstanten $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \, \min\left[-\nabla f(x_k)^T p_k, \left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|}\right)^2\right].$$

Wegen $\lim_{k\to\infty}[f(x_k) - f(x_{k+1})] = 0$ und der Beschränktheit der Folge $\{p_k\}$ ist

$$\lim_{k\to\infty} \nabla f(x_k)^T p_k = 0.$$

Nun sei x^* ein Häufungspunkt von $\{x_k\}$, also Limes einer Teilfolge $\{x_k\}_{k\in K}$. Wir zeigen, dass $\nabla f(x^*)^T p^* \geq 0$ für jede in x^* zulässige Richtung $p^* \in F(M; x^*)$. Wegen der Aussage in Aufgabe 1 ist x^* dann eine kritische Lösung von (P). Als in x^* zulässige Richtung ist $a_i^T p^* \geq 0$, $i \in I(x^*)$, (hier bedeutet $I(x^*)$ natürlich die Indexmenge der in x^* aktiven Ungleichungsrestriktionen), und $a_i^T p^* = 0$, $i = m_0 + 1, \ldots, m$. Wir werden uns wie im Beweis zu Satz 2.5 überlegen, dass ein hinreichend kleines $s_0 > 0$ existiert, für welches $s_0 p^*$ für alle hinreichend großen $k \in K$ zulässig für (P_k) ist, dass also

$$a_i^T(s_0 p^*) \geq b_i - a_i^T x_k \quad (i = 1, \ldots, m_0), \qquad a_i^T(s_0 p^*) = 0 \quad (i = m_0 + 1, \ldots, m)$$

für alle hinreichend großen $k \in K$ gilt. Nach Definition der Indexmenge $I(x^*)$ der in x^* aktiven Ungleichungsrestriktionen existiert ein $\zeta > 0$ mit $a_i^T x^* - b_i \geq \zeta$ für alle $i \in \{1, \ldots, m_0\} \setminus I(x^*)$. Für alle hinreichend großen $k \in K$, etwa $k \geq k_0$, ist daher $a_i^T x_k - b_i \geq \frac{1}{2}\zeta$ für alle $i \in \{1, \ldots, m_0\} \setminus I(x^*)$. Nun wähle man $s_0 > 0$ so klein, dass $\frac{1}{2}\zeta \geq -a_i^T(s_0 p^*)$ für alle $i \in \{1, \ldots, m_0\}$ mit $a_i^T p^* < 0$. Um nachzuweisen, dass $s_0 p^*$ für alle $k \geq k_0$ zulässig für (P_k) ist, nehmen wir $k \in K$ und $k \geq k_0$ an und

geben uns ein $i \in \{1,\dots,m_0\}$ vor. Für $i \in I(x^*)$ ist $a_i^T p^* \geq 0$, da $p^* \in F(M;x^*)$, und folglich $a_i^T(s_0 p^*) \geq 0 \geq b_i - a_i^T x_k$. Den selben Schluss können wir machen, wenn $i \in \{1,\dots,m_0\} \setminus I(x^*)$ und $a_i^T p^* \geq 0$. Daher können wir jetzt annehmen, es sei $i \in \{1,\dots,m_0\} \setminus I(x^*)$ und $a_i^T p^* < 0$. Nach Definition von ζ ist dann

$$a_i^T x_k - b_i \geq \tfrac{1}{2}\,\zeta \geq -a_i^T(s_0 p^*).$$

Für alle hinreichend großen $k \in K$ ist damit $s_0 p^*$ zulässig für (P_k). Dann ist aber $\nabla f(x_k)^T p_k \leq s_0\,\nabla f(x_k)^T p^*$ für alle hinreichend großen $k \in K$. Mit $k \in K$, $k \to \infty$, folgt $0 \leq \nabla f(x^*)^T p^*$. Damit ist die Aufgabe gelöst.

7. Gegeben sei das linear restringierte Programm

$$\text{(P)}\quad \text{Minimiere}\quad f(x)\quad \text{auf}\quad M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1,\dots,m_0), \\ a_i^T x = b_i & (i = m_0+1,\dots,m) \end{array} \right\}.$$

Sei $x \in M$ eine aktuelle Näherung, in der die Zielfunktion f von (P) stetig differenzierbar ist, und $B \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Hiermit betrachte man das quadratische Hilfsproblem

$$(\mathrm{P}(x)) \qquad \left\{ \begin{array}{l} \text{Minimiere}\quad \nabla f(x)^T p + \frac{1}{2} p^T B p \quad \text{unter den Nebenbedingungen} \\ \begin{array}{rcll} a_i^T p & \geq & b_i - a_i^T x & (i = 1,\dots,m_0), \\ a_i^T p & = & 0 & (i = m_0+1,\dots,m), \end{array} \qquad \|p\|_\infty \leq 1. \end{array} \right.$$

Sei p^* eine Lösung von $(\mathrm{P}(x))$. Man zeige: Ist $\nabla f(x)^T p^* = 0$, so ist x eine kritische Lösung von (P), andernfalls ist p^* eine zulässige Abstiegsrichtung in x.

Hinweis: Man wende den Satz von Kuhn-Tucker auf das Hilfsproblem $(\mathrm{P}(x))$ an, wobei die Restriktion $\|p\|_\infty \leq 1$ durch die beiden linearen Ungleichungsrestriktionen $-e \leq p \leq e$ (wobei e einmal wieder der Vektor ist, dessen Komponenten alle gleich 1 sind) ersetzt wird.

Lösung: Eine Lösung p^* von $(\mathrm{P}(x))$ (natürlich existiert eine solche, da die Menge der zulässigen Lösungen nichtleer und kompakt ist) ist charakterisiert durch die Existenz von Vektoren $y \in \mathbb{R}^m$ und $u, v \in \mathbb{R}^n$ mit

$$y_i \geq 0 \quad (i = 1,\dots,m_0), \qquad u, v \geq 0, \qquad \nabla f(x) + Bp^* = \sum_{i=1}^{m} y_i a_i - u + v$$

und

$$y_i(a_i^T p^* + a_i^T x - b_i) = 0 \quad (i = 1,\dots,m_0), \qquad u^T(p^* - e) = 0, \qquad v^T(p^* + e) = 0.$$

Da $p = 0$ zulässig für $(\mathrm{P}(x))$, ist $\nabla f(x)^T p^* + \frac{1}{2}(p^*)^T B p^* \leq 0$, also

$$\nabla f(x)^T p^* \leq -\frac{1}{2}(p^*)^T B p^* \leq 0.$$

Ist daher $\nabla f(x)^T p^* = 0$, so ist auch $Bp^* = 0$ und folglich

$$0 = \nabla f(x)^T p^* = \sum_{i=1}^{m_0} \underbrace{y_i(b_i - a_i^T x)}_{\leq 0} - u^T e - v^T e.$$

Hieraus folgt $u = v = 0$ und $y_i(b_i - a_i^T x) = 0$, $i = 1, \ldots, m_0$. Insbesondere existiert ein $y \in \mathbb{R}^m$ mit

$$y_i \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x) = \sum_{i=1}^{m} y_i a_i, \qquad y_i(b_i - a_i^T x) = 0 \quad (i = 1, \ldots, m_0),$$

d. h. $x \in M$ ist eine kritische Lösung von (P).

8. Gegeben sei das linear restringierte Programm

$$\text{(P)} \quad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$

Sei $x \in M$ eine aktuelle Näherung, in der die Zielfunktion f von (P) stetig differenzierbar ist, und $B \in \mathbb{R}^{n \times n}$ symmetrisch (aber nicht notwendig positiv semidefinit). Mit einem $\Delta > 0$ betrachte man das Hilfsproblem

$$(\mathrm{P}_{x,\Delta}) \quad \left\{ \begin{array}{c} \text{Minimiere} \quad \phi_x(p) := \nabla f(x)^T p + \frac{1}{2} p^T B p \quad \text{unter den Nebenbedingungen} \\ x + p \in M, \qquad \|p\| \leq \Delta, \end{array} \right.$$

wobei $\|\cdot\|$ eine beliebige Norm auf dem $\mathbb{R}^n$ ist. Dann gilt: Ist $\min(\mathrm{P}_{x,\Delta}) = 0$, also $p^* := 0$ eine Lösung von $(\mathrm{P}_{x,\Delta})$, so ist $x \in M$ eine kritische Lösung von (P).

Lösung: Der Beweis ist sehr einfach und unterscheidet sich kaum von entsprechenden früheren. Da nämlich $p^* := 0$ im Innern der Δ-Kugel um den Nullpunkt liegt, ist diese Restriktion für die notwendigen Optimalitätsbedingungen irrelevant. Daher existiert ein $y \in \mathbb{R}^m$ mit

$$y_i \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x) = \sum_{i=1}^{m} y_i a_i, \qquad y_i(b_i - a_i^T x) = 0 \quad (i = 1, \ldots, m_0),$$

d. h. x ist eine kritische Lösung von (P).

9. Gegeben sei die linear restringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Ax \leq b\}$$

mit

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix} \in \mathbb{R}^{m \times n}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m$$

und stetig differenzierbarer Zielfunktion f. Sei $x \in M$ eine zulässige Lösung, ferner $I := I(x)$ die Indexmenge der in x aktiven Restriktionen. Die Matrix $A_I \in \mathbb{R}^{\#(I) \times n}$ sei in naheliegender Weise definiert, sie habe vollen Rang, d. h. $\{a_i\}_{i \in I}$ seien linear unabhängig. Schließlich sei

$$P := I - A_I^T (A_I A_I^T)^{-1} A_I$$

(eine Verwechslung der Einheitsmatrix I und der Indexmenge I ist extrem unwahrscheinlich). Man zeige:

(a) Ist $p := -P\nabla f(x) \neq 0$, so ist p eine zulässige Abstiegsrichtung in x.

(b) Ist $P\nabla f(x) = 0$ und $y := -(A_I A_I^T)^{-1} A_I \nabla f(x) \geq 0$, so ist x eine kritische Lösung von (P).

(c) Ist $P\nabla f(x) = 0$ und $y := -(A_I A_I^T)^{-1} A_I \nabla f(x) \not\geq 0$, ist ferner $l \in I$ ein Index mit $y_l < 0$, so setze man $I := I \setminus \{l\}$ und

$$P := I - A_I^T (A_I A_I^T)^{-1} A_I.$$

Dann ist $p := -P\nabla f(x)$ eine zulässige Abstiegsrichtung in x.

Lösung: Offenbar ist P die Projektionsmatrix, die den $\mathbb{R}^n$ auf Kern (A_I) projiziert. Als solche ist P symmetrisch und positiv semidefinit, ferner ist trivialerweise $p := -P\nabla f(x)$ eine in x zulässige Richtung. Ist $p \neq 0$, so ist $\nabla f(x)^T p < 0$, also p eine in x zulässige Abstiegsrichtung. Im folgenden nehmen wir an, es sei $P\nabla f(x) = 0$. Es wird $y := -(A_I A_I^T)^{-1} A_I \nabla f(x)$ gesetzt. Dann ist $\nabla f(x) + A_I^T y = 0$ und daher x eine kritische Lösung von (P), wenn $y \geq 0$. Ist dies nicht der Fall, so wähle man einen Index $l \in I$ mit $y_l < 0$. Wir setzen $\hat{I} := I \setminus \{l\}$, $\hat{P} := I - A_{\hat{I}}^T (A_{\hat{I}} A_{\hat{I}}^T)^{-1} A_{\hat{I}}$ und $\hat{p} := -\hat{P}\nabla f(x)$. Wir wollen zeigen, dass $\hat{p}$ eine zulässige Abstiegsrichtung ist. Zunächst beweisen wir, dass $\hat{p} \neq 0$. Angenommen, es wäre $\hat{p} = 0$. Mit

$$y := -(A_I A_I^T)^{-1} A_I \nabla f(x), \qquad \hat{y} := -(A_{\hat{I}} A_{\hat{I}}^T)^{-1} A_{\hat{I}} \nabla f(x)$$

wäre

$$\nabla f(x) = -A_I^T y = -A_{\hat{I}}^T \hat{y}$$

bzw.

$$(*) \qquad -\nabla f(x) = y_l a_l + \sum_{i \in I \setminus \{l\}} y_i a_i = A_I^T y = A_{\hat{I}}^T \hat{y} = \sum_{i \in I \setminus \{l\}} \hat{y}_i a_i,$$

was wegen $y_l \neq 0$ ein Widerspruch zur linearen Unabhängigkeit von $\{a_i\}_{i \in I}$ bedeutet. Wegen

$$\nabla f(x)^T \hat{p} = -\nabla f(x)^T \hat{P} \nabla f(x) < 0$$

ist $\hat{p}$ eine Abstiegsrichtung. Es ist $a_i^T \hat{p} = 0$ für alle $i \in \hat{I}$ und daher $\hat{p}$ eine in x zulässige Richtung, wenn auch noch $a_l^T \hat{p} \leq 0$ nachgewiesen werden kann. Aus

$$-\nabla f(x) = y_l a_l + \sum_{i \in I \setminus \{l\}} y_i a_i$$

erhält man unter Berücksichtigung von $a_i^T \hat{p} = 0$, $i \in I \setminus \{l\}$, $y_l < 0$ und $\nabla f(x)^T \hat{p} < 0$, dass

$$a_l^T \hat{p} = -\frac{\nabla f(x)^T \hat{p}}{y_l} < 0,$$

womit die Behauptungen sämtlich bewiesen sind.

10. Sei $M \subset \mathbb{R}^n$ nichtleer, konvex und abgeschlossen (z. B. sei M ein Polyeder) und $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf einer offenen Obermenge von M stetig differenzierbar. Wir nennen $x \in M$ eine *kritische Lösung* von (P), wenn $\nabla f(x)^T (z - x) \geq 0$ für alle $z \in M$, also die notwendige

Optimalitätsbedingung erster Ordnung erfüllt ist. Mit $P_M\colon \mathbb{R}^n \longrightarrow M$ sei die Projektionsabbildung auf M bezüglich der euklidischen Norm $\|\cdot\|$ bezeichnet. Sei $x \in M$ keine stationäre Lösung der Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(z), \quad z \in M,$$

und $x(t) := P_M(x - t\nabla f(x))$. Man zeige:

(a) Es ist $x \neq x(t)$ für alle $t > 0$.

(b) Es ist
$$\lim_{t\to 0+} \frac{f(x) - f(x(t))}{\nabla f(x)^T(x - x(t))} = 1.$$

(c) Es ist $f(x(t)) < f(x)$ für alle hinreichend kleinen $t > 0$.

Lösung: Wegen der Charakterisierung der Projektionsabbildung P_M ist

$$[x - t\nabla f(x) - P_M(x - t\nabla f(x))]^T(z - P_M(x - t\nabla f(x)) \leq 0 \qquad \text{für alle } z \in M.$$

Angenommen, für ein $t > 0$ sei $x = P_M(x - t\nabla f(x))$. Dann ist

$$-t\nabla f(x)^T(z - x) \leq 0 \qquad \text{für alle } z \in M$$

und folglich, im Widerspruch zur Voraussetzung, x eine kritische Lösung von (P).
Wegen der Charakterisierung der Projektion $x(t) = P_M(x - t\nabla f(x))$ ist insbesondere

$$0 \leq [x(t) - (x - t\nabla f(x))]^T(x - x(t)) \qquad \text{für alle } t \geq 0$$

und daher

$$\nabla f(x)^T(x - x(t)) \geq \frac{1}{t}\|x - x(t)\|^2 > 0 \qquad \text{für alle } t > 0.$$

Weiter ist

$$\begin{aligned}
\left|\frac{f(x) - f(x(t))}{\nabla f(x)^T(x - x(t))} - 1\right| &= \frac{|f(x) - f(x(t)) - \nabla f(x)^T(x - x(t))|}{\nabla f(x)^T(x - x(t))} \\
&= \frac{o(\|x - x(t)\|)}{\nabla f(x)^T(x - x(t))} \\
&= \frac{o(t)}{\nabla f(x)^T(x - x(t))} \\
&= \frac{t}{\nabla f(x)^T(x - x(t))}\,\frac{o(t)}{t},
\end{aligned}$$

wobei wir ausgenutzt haben, dass $\|x - x(t)\| = O(t)$ wegen

$$\|x - x(t)\| = \|P_M(x) - P_M(x - t\nabla f(x))\| \leq t\,\|\nabla f(x)\| \qquad \text{für alle } t > 0.$$

Die Behauptung folgt, wenn wir zeigen können, dass

$$\liminf_{t\to 0+} \frac{\nabla f(x)^T(x - x(t))}{t} > 0.$$

Angenommen, dies sei nicht der Fall. Dann existiert eine Nullfolge $\{t_k\} \subset \mathbb{R}_+$ mit

$$\lim_{k\to\infty} \frac{\nabla f(x)^T(x - x(t_k))}{t_k} = 0.$$

Für ein beliebiges $z \in M$ wäre dann

$$\begin{aligned}
\nabla f(x)^T(z-x) &= -\nabla f(x)^T(x - x(t_k)) + \nabla f(x)^T(z - x(t_k)) \\
&\geq -\|\nabla f(x)\| \, \|x - x(t_k)\| + \frac{1}{t_k}\,(x(t_k) - x)^T(x(t_k) - z) \\
&\geq -t_k \, \|\nabla f(x)\|^2 - \frac{\|x - x(t_k)\|}{t_k} \, \|z - x(t_k)\| \\
&\geq -t_k \, \|\nabla f(x)\|^2 - \|z - x(t_k)\| \left(\frac{\nabla f(x)^T(x - x(t_k))}{t_k} \right)^{1/2}
\end{aligned}$$

für alle k, woraus mit $k \to \infty$ folgt, dass $\nabla f(x)^T(z-x) \geq 0$ bzw. $x \in M$ eine kritische Lösung von (P) ist, ein Widerspruch zur Voraussetzung.

Die letzte Behauptung folgt offenbar sofort aus der eben bewiesenen, da etwa

$$f(x) - f(x(t)) \geq \frac{1}{2}\nabla f(x)^T(x - x(t)) \geq \frac{1}{2t}\,\|x - x(t)\|^2$$

für alle hinreichend kleinen $t > 0$.

6.5 Aufgaben in Kapitel 5

6.5.1 Aufgaben in Abschnitt 5.1

1. Gegeben sei das quadratische Programm

   (P) Minimiere $f(x) := c^T x + \frac{1}{2} x^T Q x$ auf $M := \{x \in \mathbb{R}^n : h(x) := Ax - b = 0\}$

   mit symmetrischem, positiv definitem $Q \in \mathbb{R}^{n\times n}$ und $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$. Man bilde die quadratische Straffunktion Φ_σ und berechne das unrestringierte Minimum $x(\sigma)$ von Φ_σ. Man zeige, dass $x^* := \lim_{\sigma\to\infty} x(\sigma)$ existiert und die eindeutige Lösung von (P) ist. Ferner überlege man sich, dass auch der Lagrange-Multiplikator zu x^* eindeutig ist und durch $\lim_{\sigma\to\infty} \sigma h(x(\sigma))$ gegeben ist.

   Lösung: Die quadratische Straffunktion zu (P) ist durch

   $$\Phi_\sigma(x) := c^T x + \frac{1}{2} x^T Q x + \frac{\sigma}{2}\|Ax - b\|^2$$

   gegeben. Wegen

   $$\nabla\Phi_\sigma(x) = c + Qx + \sigma A^T(Ax - b) = (Q + \sigma A^T A)x + c - \sigma A^T b$$

   ist bei

   $$x(\sigma) := (Q + \sigma A^T A)^{-1}(\sigma A^T b - c)$$

das unrestringierte Minimum von Φ_σ. Die Lösung x^* von (P) und den zugehörigen Lagrange-Multiplikator y^* berechnet man als Lösung von

$$\nabla f(x) + h'(x)^T y = 0, \qquad h(x) = 0$$

bzw. des linearen Gleichungssystems

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}.$$

Folglich ist

$$x^* = -Q^{-1}c + Q^{-1}A^T(AQ^{-1}A^T)^{-1}(AQ^{-1}c+b), \qquad y^* = -(AQ^{-1}A^T)^{-1}(AQ^{-1}c+b).$$

Bei der Berechnung von $\lim_{\sigma\to\infty} x(\sigma)$ benutzen wir eine Singulärwertzerlegung von $Q^{-1/2}A^T$, also eine Darstellung der Form

$$Q^{-1/2}A^T = U \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} V^T,$$

wobei $U \in \mathbb{R}^{n\times n}$ und $V \in \mathbb{R}^{m\times m}$ orthogonal sind und $\hat\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_m)$ eine Diagonalmatrix mit den positiven Singulärwerten von $Q^{-1/2}A^T$ auf der Diagonalen. Dann ist

$$\begin{aligned}
x(\sigma) &= (Q + \sigma A^T A)^{-1}(\sigma A^T b - c) \\
&= Q^{-1/2}(I + \sigma Q^{-1/2}A^T A Q^{-1/2})^{-1}(\sigma Q^{-1/2}A^T b - Q^{-1/2}c) \\
&= Q^{-1/2}\left[I + \sigma U \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} \underbrace{V^T V}_{=I} \begin{pmatrix} \hat\Sigma & 0 \end{pmatrix} U^T\right]^{-1} \left[\sigma U \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} V^T b - Q^{-1/2}c\right] \\
&= Q^{-1/2}\left[U \begin{pmatrix} I + \sigma\hat\Sigma^2 & 0 \\ 0 & I \end{pmatrix} U^T\right]^{-1} \left[\sigma U \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} V^T b - Q^{-1/2}c\right] \\
&= Q^{-1/2}U \begin{pmatrix} (I + \sigma\hat\Sigma^2)^{-1} & 0 \\ 0 & I \end{pmatrix} \sigma \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} V^T b \\
&\qquad - Q^{-1/2}U \begin{pmatrix} (I + \sigma\hat\Sigma^2)^{-1} & 0 \\ 0 & I \end{pmatrix} U^T Q^{-1/2}c \\
&\to Q^{-1/2}U \begin{pmatrix} \hat\Sigma^{-1} \\ 0 \end{pmatrix} V^T b - Q^{-1/2}U \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} U^T Q^{-1/2}c \qquad \text{mit } \sigma \to \infty.
\end{aligned}$$

Andererseits ist

$$\begin{aligned}
x^* &= Q^{-1}A^T(AQ^{-1}A^T)^{-1}(AQ^{-1}c + b) - Q^{-1}c \\
&= Q^{-1/2}Q^{-1/2}A^T(AQ^{-1/2}Q^{-1/2}A^T)^{-1}(AQ^{-1/2}Q^{-1/2}c + b) - Q^{-1/2}Q^{-1/2}c \\
&= Q^{-1/2}U \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} U^T Q^{-1/2}c + Q^{-1/2}U \begin{pmatrix} \hat\Sigma^{-1} \\ 0 \end{pmatrix} V^T b - Q^{-1/2}Q^{-1/2}c \\
&= Q^{-1/2}U \begin{pmatrix} \hat\Sigma^{-1} \\ 0 \end{pmatrix} V^T b - Q^{-1/2}U \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} U^T Q^{-1/2}c.
\end{aligned}$$

Damit ist $\lim_{\sigma\to\infty} x(\sigma) = x^*$ nachgewiesen. Weiter ist

$$\begin{aligned}
\sigma[Ax(\sigma) - b] &= \sigma[A(Q + \sigma A^TA)^{-1}(\sigma A^Tb - c) - b] \\
&= \sigma\Bigg[AQ^{-1/2}U \begin{pmatrix} \sigma\hat{\Sigma}(I + \sigma\hat{\Sigma}^2)^{-1} \\ 0 \end{pmatrix} V^Tb - b \\
&\qquad - AQ^{-1/2}U \begin{pmatrix} (I + \sigma\hat{\Sigma}^2)^{-1} & 0 \\ 0 & I \end{pmatrix} U^TQ^{-1/2}c\Bigg] \\
&= -\sigma V(I + \sigma\hat{\Sigma}^2)^{-1}V^Tb - \sigma V(\ \hat{\Sigma}(I + \sigma\hat{\Sigma}^2)^{-1} \quad 0\ )U^TQ^{-1/2}c \\
&\to -V\hat{\Sigma}^{-2}V^Tb - V(\ \hat{\Sigma}^{-1} \quad 0\ )U^TQ^{-1/2}c \quad \text{mit } \sigma\to\infty.
\end{aligned}$$

Andererseits ist

$$\begin{aligned}
y^* &= -(AQ^{-1}A^T)^{-1}(AQ^{-1}c + b) \\
&= -(AQ^{-1/2}Q^{-1/2}A^T)^{-1}(AQ^{-1/2}Q^{-1/2}c + b) \\
&= -\Bigg[V(\ \hat{\Sigma} \quad 0\ )U^TU \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T\Bigg]^{-1} \Big[V(\ \hat{\Sigma} \quad 0\ )U^TQ^{-1/2}c + b\Big] \\
&= -V(\ \hat{\Sigma}^{-1} \quad 0\ )U^TQ^{-1/2}c - V\hat{\Sigma}^{-2}V^Tb.
\end{aligned}$$

Damit ist auch $\lim_{\sigma\to\infty} \sigma h(x(\sigma)) = y^*$ nachgewiesen.

2. Gegeben sei das quadratische Programm

$$\text{(P)} \quad \text{Minimiere} \quad f(x) := c^Tx + \frac{1}{2}x^TQx \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : h(x) := Ax - b = 0\}$$

mit symmetrischem, positiv definitem $Q \in \mathbb{R}^{n\times n}$ und $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$. Man betrachte die unrestringierte Optimierungsaufgabe

$$(\text{P}^*_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + (y^*)^Th(x) + \tfrac{1}{2}\sigma\|h(x)\|^2, \quad x \in \mathbb{R}^n,$$

wobei y^* der (eindeutige) Lagrange-Multiplikator zur Lösung x^* von (P) ist. Man zeige, dass x^* für jedes $\sigma \geq 0$ die eindeutige Lösung von (P^*_σ) ist.

Lösung: Es ist

$$\nabla\Psi_\sigma(x) = c + Qx + A^Ty^* + \sigma A^T(Ax - b) = 0.$$

Da $\nabla^2\Psi_\sigma(x) = Q + \sigma A^TA$ positiv definit ist, ist Ψ_σ für jedes $\sigma \geq 0$ strikt konvex. Wegen $\nabla\Psi_\sigma(x^*) = 0$ ist x^* eindeutiges Minimum von (P^*_σ).

3. Gegeben sei (siehe P. Spellucci (1993, S. 394)) die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := x_1^2 + 4x_1x_2 + 5x_2^2 - 10x_1 - 20x_2 \quad \text{auf} \\ \qquad M := \{x \in \mathbb{R}^2 : h(x) := x_1 + x_2 - 2 = 0\}. \end{cases}$$

Dieser Aufgabe ordne man die unrestringierte Optimierungsaufgabe

$$(\text{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \tfrac{1}{2}\sigma h(x)^2, \quad x \in \mathbb{R}^2$$

zu. Man bestimme die Lösung $x(\sigma)$ von (P_σ) und bestätige die Aussage von Aufgabe 1, berechne also z. B. die Lösung x^* von (P) und weise $x^* = \lim_{\sigma\to\infty} x(\sigma)$ nach.

Weiter bestimme man den zu x^* gehörenden Lagrange-Multiplikator y^* und zeige, dass $\lim_{\sigma\to\infty} \sigma h(x(\sigma)) = y^*$.

Lösung: In Matrix-Schreibweise lautet die gegebene Optimierungsaufgabe:

$$\left\{ \begin{array}{l} \text{Minimiere} \quad f(x) := \begin{pmatrix} -10 \\ -20 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 2 & 4 \\ 4 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{auf} \\ \qquad\qquad M := \left\{ x \in \mathbb{R}^2 : \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 2 = 0 \right\}. \end{array} \right.$$

Als Lösung von (P_σ) berechnet man

$$\begin{aligned} x(\sigma) &= \left[\begin{pmatrix} 2 & 4 \\ 4 & 10 \end{pmatrix} + \sigma \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} \right]^{-1} \left[\sigma \begin{pmatrix} 1 \\ 1 \end{pmatrix} 2 + \begin{pmatrix} 10 \\ 20 \end{pmatrix} \right] \\ &= \begin{pmatrix} 2+\sigma & 4+\sigma \\ 4+\sigma & 10+\sigma \end{pmatrix}^{-1} \begin{pmatrix} 10+2\sigma \\ 20+2\sigma \end{pmatrix} \\ &= \frac{1}{4(1+\sigma)} \begin{pmatrix} 10+\sigma & -(4+\sigma) \\ -(4+\sigma) & 2+\sigma \end{pmatrix} \begin{pmatrix} 10+2\sigma \\ 20+2\sigma \end{pmatrix} \\ &= \frac{1}{2(1+\sigma)} \begin{pmatrix} 10+\sigma \\ 3\sigma \end{pmatrix} \\ &\to \begin{pmatrix} \frac{1}{2} \\ \frac{3}{2} \end{pmatrix}. \end{aligned}$$

Da

$$x^* = \begin{pmatrix} \frac{1}{2} \\ \frac{3}{2} \end{pmatrix}$$

ist dies eine erste Bestätigung des theoretischen Ergebnisses. Als Lagrange-Multiplikator berechnet man sehr einfach $y^* = 3$. Weiter ist

$$\sigma h(x(\sigma)) = \frac{\sigma}{2(1+\sigma)}[10+\sigma+3\sigma] - 2 = \frac{3\sigma}{1+\sigma} \to 3 = y^*.$$

Damit ist in diesem Spezialfall das theoretische Ergebnis von Aufgabe 1 bestätigt.

4. Gegeben sei die Optimierungsaufgabe (siehe P. Spellucci (1993, S. 453))

$$(\mathrm{P}) \quad \left\{ \begin{array}{l} \text{Minimiere} \quad f(x) := (x_1+2)^2 + 9(x_2+3)^2 \quad \text{unter der Nebenbedingung} \\ \qquad\qquad g(x) := 1 - x_1 - x_2 \le 0. \end{array} \right.$$

   (a) Man berechne die Lösung x^* von (P) und einen zugehörigen Lagrange-Multiplikator u^*.

   (b) Bei gegebenem $\sigma > 0$ bestimme man die Lösung $x(\sigma)$ der unrestringierten Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \frac{\sigma}{2} \max(g(x), 0)^2, \quad x \in \mathbb{R}^2$$

   und zeige, dass $\lim_{\sigma\to\infty} x(\sigma) = x^*$.

(c) Wie erhält man durch Lösen von (P_σ) für hinreichend großes σ eine Näherung für den Lagrange-Multiplikator u^*?

Lösung: Aus dem Satz von Kuhn-Tucker erhält man sehr leicht die Lösung x^* und den zugehörigen Lagrange-Multiplikator u^* durch

$$x^* = \frac{1}{5}\begin{pmatrix} -12 \\ 17 \end{pmatrix}, \qquad u^* = \frac{54}{5}.$$

Zur Lösung der unrestringierten Optimierungsaufgabe (P_σ) beachten wir, dass

$$\nabla\Phi_\sigma(x) = \nabla f(x) + \sigma \max(g(x), 0)\nabla g(x).$$

Es ist leicht zu sehen, dass es keine kritische Lösung x von (P_σ) mit $g(x) < 0$ gibt. Daher bestimmen wir $x(\sigma)$ als Lösung von

$$\begin{pmatrix} 2(x_1+2) \\ 18(x_2+3) \end{pmatrix} + \sigma(1 - x_1 - x_2)\begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Man erhält

$$x(\sigma) = \frac{1}{9+5\sigma}\begin{pmatrix} -18+17\sigma \\ -27-12\sigma \end{pmatrix}.$$

Offensichtlich ist $\lim_{\sigma\to\infty} x(\sigma) = x^*$. Es ist

$$g(x(\sigma)) = \frac{54}{9+5\sigma}, \qquad \sigma g(x(\sigma)) \to \frac{54}{5} = u^*.$$

Damit ist die Aufgabe gelöst.

5. Gegeben sei die zulässige, restringierte Optimierungsaufgabe

(P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : g(x) \le 0,\ h(x) = 0\}$

und hierzu die unrestringierte Optimierungsaufgabe

$$(P_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma\Bigg(\underbrace{\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1}_{=:S(x)}\Bigg), \quad x \in \mathbb{R}^n.$$

Existiert dann ein $\sigma^* > 0$ und ein $x^* \in \mathbb{R}^n$ derart, daß x^* für alle $\sigma \ge \sigma^*$ eine (globale) Lösung von (P_σ) ist, so ist x^* eine Lösung von (P), insbesondere also zulässig für (P).

Hinweis: Siehe S.-P. Han, O. L. Mangasarian (1979, Theorem 4.1)[10], der Beweis ist einfach.

Lösung: Wir zeigen zunächst, dass x^* zulässig für die restringierte Optimierungsaufgabe (P) ist. Angenommen, dies wäre nicht der Fall. Dann wäre

$$S(x^*) = \sum_{i=1}^{l} \max(g_i(x^*), 0) + \|h(x^*)\|_1 > 0.$$

[10] Han, S.-P. and O. L. Mangasarian (1979) "Exact penalty functions in nonlinear programming." Mathematical Programming 17, 251–269.

Sei $x \in M$ ein beliebiger, für (P) zulässiger Punkt und

$$\sigma > \max\left(\frac{f(x) - f(x^*)}{S(x^*)}, \sigma^*\right).$$

Dann ist

$$f(x) = \Psi_\sigma(x) \geq \Psi_\sigma(x^*) = f(x^*) + \sigma \underbrace{S(x^*)}_{>0} > f(x),$$

was ein Widerspruch ist. Um zu zeigen, dass x^* eineLösung von (P) ist, geben wir uns ein beliebiges $x \in M$ vor, ferner sei $\sigma \geq \sigma^*$. Da x^* eine Lösung von (P_σ) ist, ist dann

$$f(x^*) = \Psi_\sigma(x^*) \leq \Psi_\sigma(x) = f(x),$$

also x^* eine Lösung von (P).

6. Gegeben sei die restringierte Optimierungsaufgabe

(P) $$\text{Minimiere} \quad f(x) \quad \text{auf} \quad M,$$

wobei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ stetig und $M \subset \mathbb{R}^n$ abgeschlossen ist. Mit $\sigma > 0$ betrachte man hierzu die unrestringierte Aufgabe

(P_σ) $$\text{Minimiere} \quad P_\sigma(x) := f(x) + \sigma S(x), \quad x \in \mathbb{R}^n,$$

wobei $S\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ stetig ist mit

$$S(x) \begin{cases} = 0 & \text{für} \quad x \in M, \\ > 0 & \text{für} \quad x \notin M. \end{cases}$$

Ist dann $x^* \in M$ eine isolierte, lokale Lösung von (P), so existiert ein $\sigma^* > 0$ derart, dass es zu jedem $\sigma \geq \sigma^*$ ein Paar $(x(\sigma), \epsilon(\sigma)) \in \mathbb{R}^n \times \mathbb{R}_+$ mit

$$x(\sigma) \in B(x^*; \epsilon(\sigma)), \qquad \lim_{\sigma \to \infty} \epsilon(\sigma) = 0$$

und

$$P_\sigma(x(\sigma)) \leq P_\sigma(x) \qquad \text{für alle } x \in B(x^*; \epsilon(\sigma))$$

gilt, wobei $B(x^*; \epsilon(\sigma))$ die offene (euklidische) Kugel um x^* mit dem Radius $\epsilon(\sigma)$ bedeutet.

Hinweis: Siehe T. Pietzykowski (1970)[11]. Der Beweis dort ist überraschend verwickelt. Wer schafft einen einfacheren?

Lösung: Bisher ist diese Aufgabe nicht befriedigend gelöst worden.

7. Gegeben sei die Optimierungsaufgabe

(P_σ) $$\begin{cases} \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma\left(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1\right) \\ \text{auf} \quad M := \{x \in \mathbb{R}^n : l \leq x \leq u\}. \end{cases}$$

[11] Pietrzykowski, T. (1970) "The potential method for conditional maxima in the locally compact metric spaces." Numer. Math. 14, 325–329.

Hierbei seien $l, u \in \mathbb{R}^n$ zwei Vektoren mit $l < u$ (eine Verwechslung der unteren (lower) Schranke l mit der Anzahl l der g_i sollte vermieden werden). Man übertrage den Begriff der kritischen Lösung auf die Aufgabe (P_σ) und gebe notwendige und hinreichende Bedingungen dafür an, dass ein $x^* \in M$ kritische Lösung von (P_σ) ist.

Lösung: Man nennt $x^* \in M$ eine kritische Lösung von (P_σ), wenn $\Psi'_\sigma(x^*; p) \geq 0$ für alle $p \in F(M; x^*)$. Hierbei ist der Kegel der zulässigen Richtungen $F(M; x^*)$ in diesem Falle (also für Box-Constraints) durch

$$F(M; x^*) = \left\{ p \in \mathbb{R}^n : p_j \left\{ \begin{array}{ll} \geq 0, & \text{falls} \quad x_j^* = l_j, \\ \leq 0, & \text{falls} \quad x_j^* = u_j. \end{array} \right. \right\}$$

gegeben. In Lemma 1.3 haben wir die Richtungsableitung $\Psi'_\sigma(x^*; p)$ berechnet. Es ist

$$\begin{aligned} \Psi'_\sigma(x^*; p) \;\; = \;\; & \nabla f(x^*)^T p + \sigma \Bigg(\sum_{i \in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*)^T p \\ & + \sum_{j \in J^*} |\nabla h_j(x^*)^T p| + \sum_{j \notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*)^T p \Bigg). \end{aligned}$$

Hierbei ist

$$I^* := \{ i \in \{1, \ldots, l\} : g_i(x^*) = 0 \}, \qquad J^* := \{ j \in \{1, \ldots, m\} : h_j(x^*) = 0 \},$$

ferner sind τ_i, $i \in \{1, \ldots, l\} \setminus I^*$, durch

$$\tau_i := \left\{ \begin{array}{ll} 1, & \text{falls} \quad g_i(x^*) > 0, \\ 0, & \text{falls} \quad g_i(x^*) < 0, \end{array} \right. \qquad i \in \{1, \ldots, l\} \setminus I^*$$

definiert. Wir wollen zeigen:

- $x^* \in M$ ist genau dann eine kritische Lösung von (P_σ), wenn Zahlen $\hat{u}_i$, $i \in I^*$, und $\hat{v}_j$, $j \in J^*$, sowie $\lambda^*, \mu^* \in \mathbb{R}^n$ existieren mit

  $$0 \leq \hat{u}_i \leq 1 \quad (i \in I^*), \qquad -1 \leq \hat{v}_j \leq 1 \quad (j \in J^*),$$

  und

  $$\lambda^*, \mu^* \geq 0, \qquad (\lambda^*)^T (x^* - l) = (\mu^*)^T (u - x^*) = 0$$

  sowie

  $$\begin{aligned} 0 \;\; = \;\; & \nabla f(x^*) + \sigma \Bigg(\sum_{i \in I^*} \hat{u}_i \nabla g_i(x^*) + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*) \\ & + \sum_{j \in J^*} \hat{v}_j \nabla h_j(x^*) + \sum_{j \notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*) \Bigg) - \lambda^* + \mu^*. \end{aligned}$$

Zum Nachweis dieser Behauptung definieren wir zur Abkürzung

$$J_l^* = \{ j \in \{1, \ldots, n\} : x_j^* = l_j \}, \qquad J_u^* := \{ j \in \{1, \ldots, n\} : x_j^* = u_j \}.$$

Zunächst nehmen wir an, es würde Zahlen $\hat{u}_i$, $i \in I^*$, und $\hat{v}_j$, $j \in J^*$, sowie $\lambda^*, \mu^* \in \mathbb{R}^n$ mit den angegebenen Eigenschaften existieren. Wegen der Gleichgewichtsbedingungen

$$(\lambda^*)^T (x^* - l) = (\mu^*)^T (u - x^*) = 0$$

ist $\lambda_j^* = 0$ für $j \notin J_l^*$ und entsprechend $\mu_j^* = 0$ für $j \notin J_u^*$. Für ein beliebiges $p \in F(M; x^*)$ ist dann

$$\begin{aligned}
\Psi'_\sigma(x^*; p) &= \nabla f(x^*)^T p + \sigma\Bigg(\sum_{i\in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{i\notin I^*} \tau_i \nabla g_i(x^*)^T p \\
&\qquad + \sum_{j\in J^*} |\nabla h_j(x^*)^T p| + \sum_{j\notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*)^T p\Bigg) \\
&= \sigma\Bigg(\sum_{i\in I^*} [\underbrace{\max(\nabla g_i(x^*)^T p, 0) - \hat{u}_i \nabla g_i(x^*)^T p}_{\geq 0}] \\
&\qquad + \sum_{j\in J^*} [\underbrace{|\nabla h_j(x^*)^T p| - \hat{v}_j \nabla h_j(x^*)^T p}_{\geq 0}]\Bigg) + \underbrace{(\lambda^*)^T p}_{\geq 0} - \underbrace{(\mu^*)^T p}_{\leq 0} \\
&\geq 0,
\end{aligned}$$

also $x^* \in M$ eine kritische Lösung von (P_σ). Umgekehrt nehmen wir nun an, $x^* \in M$ sei eine kritische Lösung von (P_σ). Wir machen einen Widerspruchsbeweis und nehmen an, es gäbe keine $\hat{u}_i$, $i \in I^*$, $\hat{v}_j$, $j \in J^*$ sowie λ_j^*, $j \in J_l^*$, μ_j^*, $j \in J_u^*$, mit den angebenen Eigenschaften. Dann wäre das Gleichungs-Ungleichungssystem

$$\begin{cases}
\sigma\Big(\displaystyle\sum_{i\in I^*} u_i \nabla g_i(x^*) + \sum_{j\in J^*} v_j \nabla h_j(x^*)\Big) - \sum_{j\in J_l^*} \lambda_j e_j + \sum_{j\in J_u^*} \mu_j e_j = c, \\
\qquad 0 \leq u_i \leq 1 \quad (i \in I^*), \qquad -1 \leq v_j \leq 1 \quad (j \in J^*), \\
\qquad \lambda_j^* \geq 0 \quad (j \in J_l^*), \qquad \mu_j \geq 0 \quad (j \in J_u^*)
\end{cases}$$

nicht lösbar. Hierbei haben wir

$$c := -\nabla f(x^*) - \sigma\Big(\sum_{i\notin I^*} \tau_i \nabla g_i(x^*) + \sum_{j\notin J^*} \operatorname{sign}(h_j(x^*)) \nabla h_j(x^*)\Big)$$

gesetzt. Zur Vereinfachung der Notation definieren wir die Matrizen

$$A := (\nabla g_i(x^*))_{i\in I^*}, \quad B := (\nabla h_j(x^*))_{j\in J^*}, \quad C := (e_j)_{j\in J_l^*}, \quad D := (e_j)_{j\in J_u^*},$$

ferner die Vektoren

$$u = (u_i)_{i\in I^*}, \quad v = (v_j)_{j\in J^*}, \quad \lambda = (\lambda_j)_{j\in J_l^*}, \quad \mu = (\mu_j)_{j\in J_u^*}.$$

Ferner sei e ein Vektor geeigneter Länge, dessen Komponenten alle gleich 1 sind. Die Widerspruchsannahme besagt dann, dass das Gleichungs-Ungleichungssystem

$$\sigma(Au + Bv) - C\lambda + D\mu = c, \quad 0 \leq u \leq e, \quad -e \leq v \leq e, \quad \lambda \geq 0, \quad \mu \geq 0$$

nicht lösbar ist. Etwas anders geschrieben bedeutet dies, dass

$$\begin{pmatrix} c \\ e \\ e \\ e \end{pmatrix} - \begin{pmatrix} \sigma A & \sigma B & -C & D \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & -I & 0 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ \lambda \\ \mu \end{pmatrix} \in \{0\} \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$$

$$(u, v, \lambda, \mu) \in \mathbb{R}_{\geq 0} \times \mathbb{R} \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$$

nicht lösbar ist. Das verallgemeinerte Farkas-Lemma liefert die Existenz eines 4-Tupels

$$(q, \alpha, \beta, \gamma) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$$

mit

$$\begin{pmatrix} \sigma A^T & I & 0 & 0 \\ \sigma B^T & 0 & I & -I \\ -C^T & 0 & 0 & 0 \\ D^T & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} q \\ \alpha \\ \beta \\ \gamma \end{pmatrix} \in \mathbb{R}_{\geq 0} \times \{0\} \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$$

und

$$\begin{pmatrix} c \\ e \\ e \\ e \end{pmatrix}^T \begin{pmatrix} q \\ \alpha \\ \beta \\ \gamma \end{pmatrix} < 0.$$

Mit $p := -q$ bedeutet dies, dass

$$-\sigma A^T p + \alpha \geq 0, \qquad -\sigma B^T p + \beta - \gamma = 0, \qquad C^T p \geq 0, \qquad D^T p \leq 0$$

und

$$-c^T p + e^T \alpha + e^T (\beta + \gamma) < 0.$$

Komponentenweise bedeutet dies, dass

$$\sigma \nabla g_i(x^*)^T p \leq \alpha_i \quad (i \in I^*), \qquad \sigma \nabla h_j(x^*)^T p = \beta_j - \gamma_j \quad (j \in J^*)$$

und

$$p_j \geq 0 \quad (j \in J_l^*), \qquad p_j \leq 0 \quad (j \in J_u^*)$$

sowie

$$-c^T p + \sum_{i \in I^*} \alpha_i + \sum_{j \in J^*} (\beta_j + \gamma_j) < 0.$$

Hieran erkennen wir, dass $p \in F(M; x^*)$ eine zulässige Richtung ist. Weiter ist

$$\begin{aligned} \Psi'_\sigma(x^*; p) &= \nabla f(x^*)^T p + \sigma \Bigg(\sum_{i \in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{i \notin I^*} \tau_i \nabla g_i(x^*)^T p \\ &\qquad + \sum_{j \in J^*} |\nabla h_j(x^*)^T p| + \sum_{j \notin J^*} \operatorname{sign}[h_j(x^*)] \nabla h_j(x^*)^T p \Bigg) \\ &= -c^T p + \sigma \Bigg(\sum_{i \in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{j \in J^*} |\nabla h_j(x^*)^T p| \Bigg) \\ &\leq -c^T p + \sum_{i \in I^*} \underbrace{\max(\alpha_i, 0)}_{=\alpha_i} + \sum_{j \in J^*} \underbrace{|\beta_j - \gamma_j|}_{\leq \beta_j + \gamma_j} \\ &\leq -c^T p + \sum_{i \in I^*} \alpha_i + \sum_{j \in J^*} (\beta_j + \gamma_j) \\ &< 0. \end{aligned}$$

Also existiert in x^* eine zulässige Abstiegsrichtung, ein Widerspruch dazu, dass x^* eine kritische Lösung ist.

8. Der restringierten, nichtlinearen Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \le 0,\ h(x) = 0\}$$

mit glatten $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ordne man die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma P(x), \quad x \in \mathbb{R}^n$$

mit

$$P(x) := \max(0, g_1(x), \dots, g_l(x), |h_1(x)|, \dots, |h_m(x)|)$$

zu. Man berechne die Richtungsableitung $\Psi'_\sigma(x^*; p)$ in einem Punkt $x^* \in \mathbb{R}^n$ in die Richtung $p \in \mathbb{R}^n$ und gebe notwendige und hinreichende Bedingungen dafür an, dass x^* eine kritische Lösung von (P_σ) ist, also $\Psi'_\sigma(x^*; p) \ge 0$ für alle $p \in \mathbb{R}^n$ gilt.

Lösung: Wir definieren die Indexmengen

$$I^* := \{i \in \{1, \dots, l\} : g_i(x^*) = P(x^*)\}, \qquad J^* := \{j \in \{1, \dots, m\} : |h_j(x^*)| = P(x^*)\}.$$

Die Richtungsableitung von P in x^* in Richtung p ist gegeben durch

$$P'(x^*; p) = \begin{cases} \max\Big(0, \max\limits_{i \in I^*} \nabla g_i(x^*)^T p, \max\limits_{j=1,\dots,m} |\nabla h_j(x^*)^T p|\Big), & P(x^*) = 0, \\ \max\Big(\max\limits_{i \in I^*} \nabla g_i(x^*)^T p, \max\limits_{j \in J^*} \operatorname{sign}(h_j(x^*)) \nabla h_j(x^*)^T p\Big), & P(x^*) > 0. \end{cases}$$

Folglich ist

$$\Psi'_\sigma(x^*; p) = \nabla f(x^*)^T p + \sigma P'(x^*; p).$$

Wir wollen zeigen:

- $x^* \notin M$ ist genau dann eine kritische Lösung von (P_σ), wenn $u_i^* \ge 0$, $i \in I^*$, und $v_j^* \ge 0$, $j \in J^*$, existieren mit

$$0 = \nabla f(x^*) + \sigma\Big(\sum_{i \in I^*} u_i^* \nabla g_i(x^*) + \sum_{j \in J^*} v_j^* \operatorname{sign}(h_j(x^*)) \nabla h_j(x^*)\Big)$$

  und

$$\sum_{i \in I^*} u_i^* + \sum_{j \in J^*} v_j^* = 1.$$

Denn: Zunächst nehmen wir an, dass u_i^* und v_j^* mit den angegebenen Eigenschaften existieren. Dann ist

$$\begin{aligned} \Psi'_\sigma(x^*; p) &= \nabla f(x^*)^T p + \sigma \max\Big(\max_{i \in I^*} \nabla g_i(x^*)^T p, \max_{j \in J^*} \operatorname{sign}(h_j(x^*)) \nabla h_j(x^*)^T p\Big) \\ &= \sigma \max\Big(\max_{i \in I^*} \nabla g_i(x^*)^T p, \max_{j \in J^*} \operatorname{sign}(h_j(x^*)) \nabla h_j(x^*)^T p\Big) \\ &\qquad - \sigma\Big(\sum_{i \in I^*} u_i^* \nabla g_i(x^*)^T p + \sum_{j \in J^*} v_j^* \operatorname{sign}(h_j(x^*)) \nabla h_j(x^*)^T p\Big) \\ &\ge \sigma \max\Big(\max_{i \in I^*} \nabla g_i(x^*)^T p, \max_{j \in J^*} \operatorname{sign}(h_j(x^*)) \nabla h_j(x^*)^T p\Big) \end{aligned}$$

$$
\begin{aligned}
& - \sigma\left(\max_{i\in I^*} \nabla g_i(x^*)^T p \sum_{i\in I^*} u_i^* + \max_{j\in J^*} \operatorname{sign}\left(h_j(x^*)\right) \nabla h_j(x^*)^T p \sum_{j\in J^*} v_j^*\right) \\
\geq\ & \sigma \max\left(\max_{i\in I^*} \nabla g_i(x^*)^T p, \max_{j\in J^*} \operatorname{sign}\left(h_j(x^*)\right) \nabla h_j(x^*)^T p\right) \\
& - \sigma \max\left(\max_{i\in I^*} \nabla g_i(x^*)^T p, \max_{j\in J^*} \operatorname{sign}\left(h_j(x^*)\right) \nabla h_j(x^*)^T p\right) \times \\
& \times \underbrace{\left(\sum_{i\in I^*} u_i^* + \sum_{j\in J^*} v_j^*\right)}_{=1} \\
=\ & 0.
\end{aligned}
$$

Daher ist x^* eine kritische Lösung von (P_σ). Umgekehrt nehmen wir an, $x^* \notin M$ sei eine kritische Lösung von (P_σ). Angenommen, die Behauptung sei falsch und es würde keine $(u_i^*)_{i\in I^*}$ und $(v_j)_{j\in J^*}$ mit den angegebenen Eigenschaften geben. Mit den Matrizen

$$
A := (\nabla g_i(x^*))_{i\in I^*}, \qquad B := (\operatorname{sign}\left(h_j(x^*)\right) \nabla h_j(x^*))_{j\in J^*}
$$

hätte das System

$$
\begin{pmatrix} \sigma A & \sigma B \\ e^T & e^T \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -\nabla f(x^*) \\ 1 \end{pmatrix}, \qquad \begin{pmatrix} u \\ v \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix}
$$

keine Lösung. Das Farkas-Lemma zeigt die Existenz eines Paares (q, δ) mit

$$
\begin{pmatrix} \sigma A^T & e \\ \sigma B^T & e \end{pmatrix} \begin{pmatrix} q \\ \delta \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} -\nabla f(x^*) \\ 1 \end{pmatrix}^T \begin{pmatrix} q \\ \delta \end{pmatrix} < 0.
$$

Mit $p := -q$ impliziert dies, dass

$$
\sigma \max\left(\max_{i\in I^*} \nabla g_i(x^*)^T p, \max_{j\in J^*} \operatorname{sign}\left(h_j(x^*)\right) \nabla h_j(x^*)^T p\right) \leq \delta < -\nabla f(x^*)^T p.
$$

Folglich ist $\Psi'_\sigma(x^*; p) < 0$, ein Widerspruch dazu, dass x^* eine kritische Lösung von (P_σ) ist.

9. Gegeben sei die *konvexe* Optimierungsaufgabe

$$
(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\},
$$

die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sei also konvex, die Restriktionabbildung $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$ komponentenweise konvex und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear. Sei $x^* \in M$ eine Lösung von (P), ferner gelte die Slatersche Constraint Qualification, es existiere also $\hat{x} \in \mathbb{R}^n$ mit $g(\hat{x}) < 0$ und $h(\hat{x}) = 0$ und die Abbildung h sei surjektiv. Ist dann (u^*, v^*) eine Lösung des zu (P) dualen Programms, so ist x^* für alle $\sigma \geq \sigma^* := \max(\|u^*\|_\infty, \|v^*\|_\infty)$ eine globale Lösung von

$$
(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma\left(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1\right), \quad x \in \mathbb{R}^n.
$$

Hinweis: Eine etwas allgemeinere Version der obigen Aussage findet man bei S.-P. Han, O. L. Mangasarian (1979, Theorem 4.9). Man sollte aber nicht dort nachsehen, sondern den einfachen Beweis selber finden.

Lösung: Die zu (P) gehörende Lagrange-Funktion ist

$$L(x,u,v) := f(x) + u^T g(x) + v^T h(x),$$

das zu (P) duale Programm ist

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \phi(u,v) := \inf\limits_{x\in\mathbb{R}^n} L(x,u,v) \quad \text{auf} \\ N := \{(u,v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \phi(u,v) > -\infty\}. \end{cases}$$

Da die Slatersche Constraint Qualification vorausgesetzt ist, besitzt (D) eine Lösung (u^*, v^*) und es tritt keine Dualitätslücke ein. Mit der Lösung x^* von (P) ist also

$$f(x^*) = \phi(u^*, v^*) \leq L(x, u^*, v^*) \qquad \text{für alle } x \in \mathbb{R}^n.$$

Für ein beliebiges $x \in \mathbb{R}^n$ und $\sigma \geq \sigma^* := \max(\|u^*\|_\infty, \|v^*\|_\infty)$ ist daher

$$\begin{aligned} \Psi_\sigma(x^*) &= f(x^*) \\ &\leq L(x, u^*, v^*) \\ &= f(x) + \sum_{i=1}^{l} \underbrace{u_i^*}_{\geq 0} g_i(x) + \sum_{j=1}^{n} v_j^* h_j(x) \\ &\leq f(x) + \sum_{i=1} u_i^* \max(g_i(x), 0) + \sum_{j=1}^{m} |v_j^*|\, |h_j(x)| \\ &\leq f(x) + \|u^*\|_\infty \sum_{i=1}^{l} \max(g_i(x), 0) + \|v^*\|_\infty \|h(x)\|_1 \\ &\leq f(x) + \sigma\left(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1\right) \\ &= \Psi_\sigma(x). \end{aligned}$$

Damit ist die Behauptung bewiesen.

6.5.2 Aufgaben in Abschnitt 5.2

1. Sei $C \subset \mathbb{R}^n$ nichtleer, konvex und abgeschlossen. Für ein $x \in C$ und ein $p \in \mathbb{R}^n$ sei $\{x + tp : t \geq 0\} \subset C$, also der gesamte von x in Richtung p ausgehende Halbstrahl in C enthalten. Man zeige, dass für ein beliebiges $z \in C$ auch der Halbstrahl $\{z + tp : t \geq 0\}$ in C enthalten ist.

   Lösung: Ein merkwürdig komplizierter Beweis ist bei R. T. Rockafellar (1970, Theorem 8.3) zu finden. Einen einfacheren Beweis geben wir jetzt an. Seien $z \in C$ und $s > 0$ gegeben. Angenommen, es ist $z + sp \notin C$. Der starke Trennungssatz liefert die Existenz eines $y \in \mathbb{R}^n$ mit

   $$y^T(z + sp) < \gamma := \inf_{u\in C} y^T u.$$

Insbesondere ist

$$y^T(z+sp) < \gamma \le y^T(x+tp) \qquad \text{für alle } t \ge 0.$$

Hieraus folgt $y^T p \ge 0$. Wegen $z \in C$ ist andererseits

$$y^T z + s y^T p < \gamma \le y^T z$$

und daher $y^T p < 0$, ein Widerspruch.

2. Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex. Man zeige:

   (a) Für jedes $x \in \mathbb{R}^n$ und jedes $p \in \mathbb{R}^n$ existiert (im eigentlichen oder uneigentlichen Sinne)

   $$f_\infty(p) := \lim_{t\to\infty} \frac{f(x+tp)-f(x)}{t}$$

   und ist durch

   $$f_\infty(p) = \sup_{z\in\mathbb{R}^n} [f(z+p)-f(z)]$$

   gegeben, ist also insbesondere (wie die Notation es erwarten lässt) von x unabhängig.

   (b) Die konvexe Optimierungsaufgabe

   (P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : g(x) \le 0,\ h(x) = 0\}$

   sei zulässig. Dann ist die Menge M_{opt} der Lösungen von (P) genau dann nichtleer und kompakt, wenn das System

   $$f_\infty(p) \le 0, \qquad (g_i)_\infty(p) \le 0 \quad (i = 1,\dots,l), \qquad h'p = 0$$

   nur trivial lösbar ist.

Lösung: Um in (a) die Existenz von $\lim_{t\to\infty}[f(x+tp)-f(x)]/t$ nachzuweisen, beachte man, dass die durch

$$h(t) := \frac{f(x+tp)-f(x)}{t}$$

definierte Funktion $h\colon \mathbb{R}_+ \longrightarrow \mathbb{R}$ auf $\mathbb{R}_+$ monoton nicht fallend ist. Denn ist $0 < s \le t$, so ist

$$x + sp = \left(1 - \frac{s}{t}\right)x + \frac{s}{t}(x+tp).$$

Aus der Konvexität von f folgt

$$f(x+sp) \le \left(1-\frac{s}{t}\right)f(x) + \frac{s}{t}f(x+tp),$$

danach durch Umordnen $h(s) \le h(t)$. Damit ist die Existenz von $f_\infty(p)$ gezeigt. In (a) bleibt zu zeigen, dass

$$f_\infty(p) = \sup_{z\in\mathbb{R}^n} [f(z+p)-f(z)].$$

Hierzu zeigen wir zunächst, dass

$$\frac{f(x+tp)-f(x)}{t} \le \sup_{z\in\mathbb{R}^n} [f(z+p)-f(z)]$$

für alle $t > 0$, $x \in \mathbb{R}^n$. Denn: Seien $t > 0$ und $x \in \mathbb{R}^n$ fest. Für $t \in (k-1, k]$ mit $k \in \mathbb{N}$ ist

$$\begin{aligned} \frac{f(x+tp)-f(x)}{t} &\leq \frac{f(x+kp)-f(x)}{k} \\ &\leq f(x+(k-1)p+p)-f(x+(k-1)p) \\ &\leq \sup_{z\in\mathbb{R}^n}[f(z+p)-f(z)], \end{aligned}$$

womit (nach $t \to \infty$) die behauptete Ungleichung bewiesen ist. Zu zeigen bleibt die umgekehrte Ungleichung. Wieder sei $x \in \mathbb{R}^n$ fest vorgegeben. Wir nehmen an, es sei $f_\infty(p) < \infty$ (andernfalls zeigt die erste Ungleichung schon die Richtigkeit der Behauptung). Wir definieren den sogenannten *Epigraphen* von f durch

$$\operatorname{epi}(f) := \{(x, \mu) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq \mu\}.$$

Wegen der Konvexität von f ist $\operatorname{epi}(f)$ eine konvexe Menge, da auf dem $\mathbb{R}^n$ konvexe Funktionen dort auch stetig sind, ist $\operatorname{epi}(f)$ auch abgeschlossen. Ferner ist

$$(x, f(x)) + t(p, f_\infty(p)) \in \operatorname{epi}(f) \qquad \text{für alle } t \geq 0.$$

Da $(z, f(z)) \in \operatorname{epi}(f)$ für alle $z \in \mathbb{R}^n$, liefert eine Anwendung der Aussage von Aufgabe 1, dass

$$\begin{aligned} (z, f(z)) + t(p, f_\infty(p)) &= (z+tp, f(z)+tf_\infty(p)) \\ &\in \operatorname{epi}(f) \qquad \text{für alle } z \in \mathbb{R}^n \text{ und alle } t \geq 0 \end{aligned}$$

bzw.

$$f(z+tp) \leq f(z) + tf_\infty(p) \qquad \text{für alle } z \in \mathbb{R}^n \text{ und alle } t \geq 0.$$

Insbesondere ist

$$f(z+p) - f(z) \leq f_\infty(p) \qquad \text{für alle } z \in \mathbb{R}^n$$

und daher

$$\sup_{z\in\mathbb{R}^n}[f(z+p)-f(z)] \leq f_\infty(p).$$

Damit ist der Beweis von (a) vollständig.

Nun zum Beweis von (b). Zunächst nehmen wir an, es existiere ein $p \in \mathbb{R}^n \setminus \{0\}$ mit

$$(*) \qquad f_\infty(p) \leq 0, \qquad (g_i)_\infty(p) \leq 0 \quad (i = 1, \ldots, l), \qquad h'p = 0.$$

Mit einem beliebigen $x^* \in M_{\text{opt}}$ zeigen wir, dass $x^* + tp \in M_{\text{opt}}$ für alle $t \geq 0$ bzw. M_{opt} nicht kompakt ist. Denn für alle $t > 0$ ist

$$\frac{f(x^*+tp)-f(x^*)}{t} \leq f_\infty(p) \leq 0$$

und daher $f(x^*+tp) \leq f(x^*)$, aus $g_\infty(p) \leq 0$ und $h'(p) = 0$ folgt entsprechend $x^*+tp \in M$ und damit insgesamt $x^* + tp \in M_{\text{opt}}$ für alle $t \geq 0$. Damit ist gezeigt: Ist M_{opt} nichtleer und kompakt, so ist das System $(*)$ nur trivial lösbar. Nun nehmen wir an, das System $(*)$ sei nur trivial lösbar und zeigen mit einem $x_0 \in M$, dass die Niveaumenge

$$L_0 := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0,\ f(x) \leq f(x_0)\}$$

(nichtleer und) kompakt ist, woraus natürlich auch folgt, dass M_{opt} nichtleer und kompakt ist. Angenommen, dies sei nicht der Fall. Dann existiert eine Folge $\{x_k\} \subset L_0$ mit $\|x_k\| \to \infty$. O. B. d. A. können wir annehmen, dass $x_k/\|x_k\| \to p$, wobei natürlich $p \neq 0$. Wir wollen zeigen, dass p eine (nichttriviale) Lösung von $(*)$ ist. Nun ist

$$0 = h(x_k) = h'x_k + h(0)$$

und daher $h'p = 0$ nach Division mit $\|x_k\|$ und dem Grenzübergang $k \to \infty$. Sei $t > 0$ vorgegeben. Für alle hinreichend großen k ist $t/\|x_k\| \in (0, 1]$ und daher

$$\begin{aligned} g\left(\left(1 - \frac{t}{\|x_k\|}\right)x_0 + \frac{t}{\|x_k\|}x_k\right) &\leq \left(1 - \frac{t}{\|x_k\|}\right)g(x_0) + \frac{t}{\|x_k\|}\underbrace{g(x_k)}_{\leq 0} \\ &\leq \left(1 - \frac{t}{\|x_k\|}\right)g(x_0). \end{aligned}$$

Mit dem Grenzübergang $k \to \infty$ folgt

$$g(x_0 + tp) - g(x_0) \leq 0 \qquad \text{für alle } t > 0,$$

nach Division mit t und dem Grenzübergang $t \to +\infty$ folgt $g_\infty(p) \leq 0$. Praktisch genau so folgt, dass auch $f_\infty(p) \leq 0$ und damit p eine nichttriviale Lösung von $(*)$ ist.

3. Gegeben sei die konvexe, quadratisch restringierte quadratische Optimierungsaufgabe

(P) Minimiere $f(x)$ auf $M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ h(x) = 0\}$,

wobei

$$f(x) := c_0^T x + \tfrac{1}{2}x^T Q_0 x, \qquad g_i(x) := \beta_i + c_i^T x + \tfrac{1}{2}x^T Q_i x \quad (i = 1, \ldots, l)$$

und

$$h(x) := Ax - b$$

mit symmetrischen, positiv semidefiniten Matrizen $Q_0, Q_1, \ldots, Q_l$. Weiter setzen wir voraus, dass (P) zulässig ist. Man zeige:

(a) Die Menge M_{opt} der Lösungen von (P) ist genau dann nichtleer und kompakt, wenn das System

$$(*) \qquad c_i^T p \leq 0, \quad Q_i p = 0 \qquad (i = 0, \ldots, l), \quad Ap = 0$$

nur trivial lösbar ist.

(b) Die Lagrange-Funktion $L\colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ zu (P) ist natürlich durch

$$L(x, u, v) := f(x) + g(x)^T u + h(x)^T v$$

gegeben. Das zu (P) duale Programm ist bekanntlich

$$\text{(D)} \qquad \left\{ \begin{array}{c} \text{Maximiere} \quad \phi(u, v) := \inf\limits_{x \in \mathbb{R}^n} L(x, u, v) \quad \text{auf} \\ N := \{(u, v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \phi(u, v) > -\infty\}. \end{array} \right.$$

Da eine konvexe quadratische Funktion genau dann auf dem $\mathbb{R}^n$ nach unten beschränkt ist, wenn ihr Gradient eine Nullstelle besitzt, ist die Menge der dual zulässigen Lösungen durch

$$N = \{(u,v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ \exists\, z \in \mathbb{R}^n \text{ mit } \nabla_x L(z,u,v) = 0\}$$

gegeben. Weiter sei

$$N^0 := \{(u,v) \in N : u > 0\}.$$

Man zeige: Die Menge M_{opt} der Lösungen von (P) ist genau dann nichtleer und kompakt, wenn

$$(**) \qquad c_i^T p = 0, \quad Q_i p = 0 \qquad (i = 0, \ldots, l), \quad Ap = 0$$

nur trivial lösbar ist und N^0 nichtleer ist.

Lösung: Es ist

$$\begin{aligned} f_\infty(p) &= \lim_{t\to\infty} \frac{f(x+tp) - f(x)}{t} \\ &= (c_0 + Q_0 x)^T p + \tfrac{1}{2} \lim_{t\to\infty} t p^T Q_0 p \\ &= \begin{cases} c_0^T p, & \text{falls} \quad Q_0 p = 0, \\ +\infty, & \text{falls} \quad Q_0 p \neq 0. \end{cases} \end{aligned}$$

Aus der vorigen Aufgabe folgt dann sofort (a), da $g_\infty(p)$ entsprechend berechnet werden kann.

Zum Beweis von (b) nehmen wir zunächst an, M_{opt} sei nichtleer und kompakt und daher das Gleichungs-Ungleichungssystem $(*)$ nur trivial lösbar. Trivialerweise ist dann das Gleichungssystem $(**)$ nur trivial lösbar, so daß noch $N^0 \neq \emptyset$ zu zeigen ist. Da $(*)$ nur trivial lösbar ist, ist

$$c_i^T p \leq 0, \quad Q_i p = 0 \qquad (i = 0, \ldots, l), \quad Ap = 0, \quad \left(\sum_{i=0}^{l} c_i\right)^T p < 0$$

nicht lösbar. Eine Anwendung des (verallgemeinerten) Farkas-Lemmas liefert die Existenz von $\delta_i \geq 0$, $z_i \in \mathbb{R}^n$, $i = 0, \ldots, l$ und $w \in \mathbb{R}^m$ mit

$$\sum_{i=0}^{l} (1+\delta_i) c_i + \sum_{i=0}^{l} Q_i z_i - A^T w = 0.$$

Wegen $1 + \delta_i > 0$, $i = 0, \ldots, l$, kann man hieraus auf $N^0 \neq \emptyset$ schließen. Denn man definiere

$$\hat{z}_0 := \frac{1}{1+\delta_0} z_0, \qquad \hat{z}_i := \frac{1}{(1+\delta_i)(1+\delta_0)} z_i \quad (i = 1, \ldots, l)$$

sowie

$$u_i := \frac{1+\delta_i}{1+\delta_0} \quad (i = 1, \ldots, l), \qquad v := -\frac{1}{1+\delta_0} w.$$

Dann ist $u = (u_i) > 0$ und

$$c_0 + Q_0 \hat{z}_0 + \sum_{i=1}^{l} u_i (c_i + Q_i \hat{z}_i) + A^T v = 0.$$

Wir wollen uns nun überlegen, dass das lineare Gleichungssystem

$$\left(Q_0 + \sum_{i=1}^{l} u_i Q_i\right) z = Q_0 \hat{z}_0 + \sum_{i=1}^{l} u_i Q_i \hat{z}_i$$

eine Lösung z besitzt. Ist dies der Fall, so ist $(u, v) \in N^0$, also $N^0 \neq \emptyset$. Andernfalls existiert (z. B. starker Trennungssatz) ein

$$y \in \text{Kern}\left(Q_0 + \sum_{i=1}^{l} u_i Q_i\right), \qquad y^T \left(Q_0 \hat{z}_0 + \sum_{i=1}^{l} u_i Q_i \hat{z}_i\right) \neq 0.$$

Hieraus erhält man den gewünschten Widerspruch,

Umgekehrt nehmen wir an, $(**)$ sei nur trivial lösbar und $N^0 \neq \emptyset$. Angenommen, $p \in \mathbb{R}^n$ genüge dem System

$$(*) \qquad c_i^T p \leq 0, \quad Q_i p = 0 \qquad (i = 0, \ldots, l), \quad Ap = 0.$$

Sei $(u, v) \in N_0$, insbesondere ist $u > 0$ und es existiert ein $z \in \mathbb{R}^n$ mit

$$c_0 + Q_0 z + \sum_{i=1}^{l} u_i (c_i + Q_i z) - A^T v = 0.$$

Eine Multiplikation mit p liefert

$$0 = \underbrace{c_0^T p}_{\leq 0} + \sum_{i=1}^{l} \underbrace{u_i}_{>0} \underbrace{c_i^T p}_{\leq 0}$$

und damit $c_i^T p = 0$, $i = 0, \ldots, l$. Also ist p eine Lösung von $(**)$ und folglich $p = 0$. Dies zeigt, dass das System $(*)$ nur trivial lösbar und damit M_{opt} wegen (a) nichtleer und kompakt ist.