Vorlesung über Optimierung

Jochen Werner

Wintersemester 2007/2008
Department Mathematik der Universität Hamburg

Inhaltsverzeichnis

Kapitel 1

Einführung

Bei dem studio der Mathematik kann wohl nichts stärkeren Trost bei Unverständlichkeiten gewähren, als daß es sehr viel *schwerer ist eines andern Meditata zu verstehen, als selbst zu meditieren.*

Georg Christoph Lichtenberg

Man muss Hindernisse wegnehmen, Begriffe aufklären, Beispiele geben, alle Teilhaber zu interessieren suchen, das ist freilich beschwerlicher als befehlen, indessen die einzige Art (...) zum Zwecke zu gelangen.

Johann Wolfgang von Goethe

Eine *Optimierungsaufgabe* ist durch zwei Daten gegeben, nämlich durch die *Menge der zulässigen* (engl.: feasible) *Lösungen* M und die *Zielfunktion* (engl.: objective function) $f\colon M \longrightarrow \mathbb{R}$. 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 *Kostenfunktion* genannt. Die durch M und f gegebene Aufgabe schreiben wir in der Form

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

oder auch

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

Wir werden uns auf den Fall beschränken, dass M eine Teilmenge des $\mathbb{R}^n$ ist[1]. Diese Teilmenge ist typischerweise als Lösungsmenge endlich vieler Ungleichungen und Gleichungen gegeben. Die Optimierungsaufgaben, die wir betrachten werden, haben daher i. Allg. die folgende Form:

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

[1]Nicht eingehen werden wir auf den Fall, dass M eine Teilmenge eines unendlichdimensionalen linearen normierten Raumes ist. Man spricht in diesem Fall von einer *infiniten Optimimierungsaufgabe*. Ebenso werden wir nicht auf die wichtige Klasse *diskreter* (oder auch *kombinatorischer, ganzzahliger*) Optimierungsaufgaben eingehen, bei denen die Menge M zulässiger Lösungen eine endliche Menge ist.

Hier sind die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ gegeben, die $\leq$–Beziehung zwischen Vektoren ist komponentenweise zu verstehen[2].

Wir nennen $x^* \in M$ eine (globale) *Lösung* von (P), wenn $f(x^*) \leq f(x)$ für alle $x \in M$. Dagegen nennt man $x^* \in M$ naheliegenderweise 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.

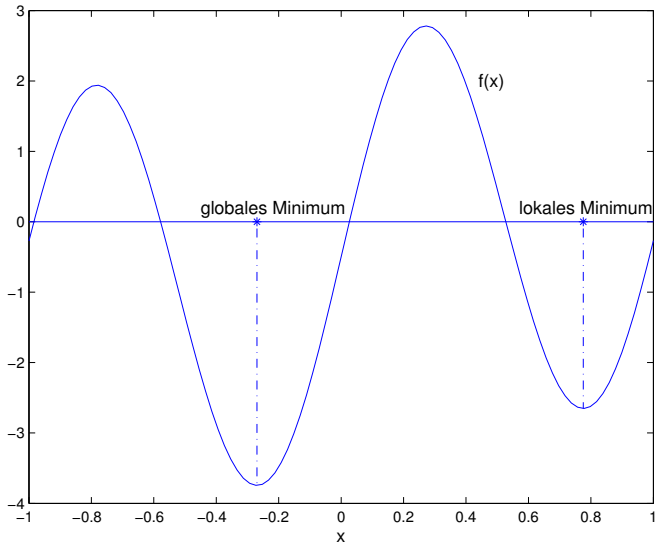


Abbildung 1.1: Globales und lokales Minimum einer Funktion $f(\cdot)$ auf $M = [-1, 1]$

In der Optimierung werden Aussagen über globale oder wenigstens lokale Lösungen einer Optimierungsaufgabe gemacht, ferner werden Verfahren motiviert und analysiert, mit deren Hilfe wenigstens eine näherungsweise Berechnung einer lokalen (oder gar globalen) Lösung möglich ist.

Man unterscheidet zwischen folgenden Optimierungsaufgaben:

- Ist $M = \mathbb{R}^n$, so spricht man von einer *unrestringierten Optimierungsaufgabe*.

  Zur Lösung unrestringierter Optimierungsaufgaben stellt die Optimization-Toolbox von MATLAB die Funktion `fminunc` zur Verfügung. Im einfachsten Fall kann man durch den Aufruf `x=fminunc(fun,x0)`, ausgehend von dem Startwert `x0` ein lokales Minimum `x` der Funktion `fun` finden.

- Ist die Zielfunktion f linear und sind die Restriktionsabbildungen g und h affin linear, so spricht man von einer *linearen Optimierungsaufgabe*.

[2]Notfalls kann eine Ungleichung mit -1 multipliziert werden, so dass wir davon ausgehen können, dass alle Ungleichungsrestriktionen *gleichgerichtet* sind. Außerdem kann das *Maximieren* einer Zielfunktion auf das Minimieren der mit -1 multiplizierten Zielfunktion zurückgeführt werden.

Z. B. ist $f(x) := c^T x$ mit $c \in \mathbb{R}^n$, $g(x) := Ax - b$ und $h(x) := A_0 x - b_0$ mit $A \in \mathbb{R}^{l \times n}$, $b \in \mathbb{R}^l$ sowie $A_0 \in \mathbb{R}^{m \times n}$, $b_0 \in \mathbb{R}^m$. Die entsprechende lineare Optimierungsaufgabe hat also die Form

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Ax \leq b,\ A_0 x = b_0\}.$$

In der Optimization-Toolbox von MATLAB findet man zur Lösung linearer Optimierungsaufgaben die Funktion `linprog`. Der Aufruf lautet im einfachsten Fall `x=linprog(c,A,b,A0,b0)`.

- Ist die Zielfunktion f quadratisch und sind die Restriktionsabbildungen g und h affin linear, so spricht man von einer *quadratischen Optimierungsaufgabe*.

  Z. B. ist $f(x) := c^T x + \frac{1}{2} x^T Q x$ mit $c \in \mathbb{R}^n$ und symmetrischer Matrix $Q \in \mathbb{R}^{n \times n}$, $g(x) := Ax - b$ und $h(x) := A_0 x - b_0$ wie eben. Die zugehörige quadratische Optimierungsaufgabe hat also die Form

  $$\text{(P)} \qquad \text{Minimiere} \quad c^T x + \tfrac{1}{2} x^T Q x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Ax \leq b,\ A_0 x = b_0\}.$$

  Zur Lösung quadratischer Optimierungsaufgaben ist in der Optimization-Toolbox von MATLAB die Funktion `quadprog` enthalten. Der Aufruf lautet im einfachsten Fall `x=quadprog(Q,c,A,b,A0,b0)`.

- Ist die Zielfunktion f nichtlinear (und nicht quadratisch), sind die Restriktionsabbildungen g und h aber affin linear, so spricht man von einer *linear restringierten nichtlinearen Optimierungsaufgabe*.

  Mit g, h wie oben hat eine linear restringierte nichtlineare Optimierungsaufgabe also die Form

  $$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Ax \leq b,\ A_0 x = b_0\}.$$

  Eine linear restringierte nichtlineare Optimierungsaufgabe kann man mit der Funktion `fmincon` der Optimization-Toolbox von MATLAB lösen. Im einfachsten Fall lautet der Aufruf `x=fmincon(fun,x0,A,b,A0,b0)`, wobei `fun` die zu minimierende Funktion und `x0` ein Startwert ist.

- Ist die Zielfunktion f oder eine der Restriktionsabbildungen g, h *nicht* affin linear, so spricht man von einer *nichtlinearen Optimierungsaufgabe*. Wir werden nur den Fall betrachten, dass f, g und h *glatt*, also wenigstens stetig differenzierbar sind.

  Zur Lösung nichtlinearer Optimierungsaufgaben ist in der Optimization-Toolbox von MATLAB die Funktion `fmincon` enthalten.

1.1 Beispiele

Dadurch, dass wir erläuterten, was eine "allgemeine" Optimierungsaufgabe ist, haben wir gegen einen Rat verstoßen, den R. P. Boas (1981) 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[3] with retractile[4] 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. An diesen mangelt es nicht, denn eigentlich versucht man immer, etwas möglichst gut zu machen, wobei i. Allg. gewisse Restriktionen zu beachten sind.

Beispiel: Eine der ältesten Optimierungsaufgaben, siehe M. CANTOR (1880, S. 228), in der Geschichte der Mathematik findet sich in Euklids 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 der Mittelpunkt der Strecke $\overline{BC}$. In Abbildung 1.2 wird dies verdeutlicht. Denn ist E beliebig auf $\overline{BC}$ und

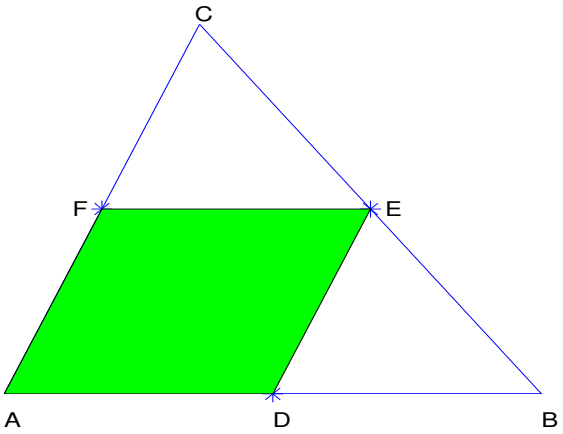


Abbildung 1.2: Die Lösung des ältesten Optimierungsproblems

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

so ist

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

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

[3] fleischfressendes Säugetier
[4] einziehbar

Es gibt zahlreiche weitere alte geometrische Optimierungsaufgaben. Genannt seien nur das Problem der Dido und Herons Problem. □

Beispiel: Das folgende Problem ist 1629 von Fermat[5] formuliert worden:

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

Bei einem Dreieck, bei dem die Winkel kleiner als 120° sind, ist der gesuchte Punkt (der sogenannte Fermat-Torricelli[6]-Punkt) der Punkt im Inneren des Dreiecks, von dem aus die drei Seiten unter einem 120°-Winkel erscheinen. In Abbildung 1.3 geben wir eine

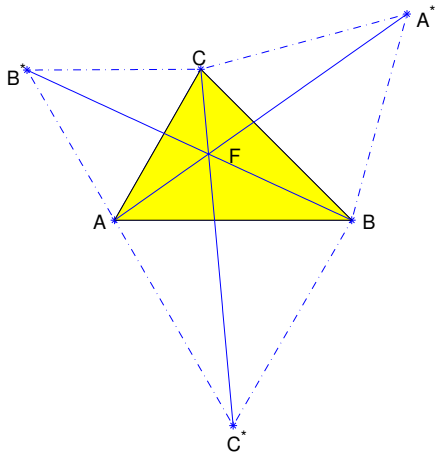


Abbildung 1.3: Konstruktion des Fermat-Torricelli-Punktes

mögliche Konstruktion des Fermat-Torricelli-Punktes an: Über den Seiten des gegebenen Dreiecks $\triangle ABC$ konstruiere man drei gleichseitige Dreiecke und gewinne dadurch Punkte A^*, B^* und C^*. Verbindet man diese Punkte mit den gegenüber liegenden Ecken des Dreiecks, also mit A, B und C, so schneiden sich diese drei Verbindungsstrecken in einem Punkt F, dem Fermat-Torricelli-Punkt.

Die Verallgemeinerung auf m Punkte im $\mathbb{R}^n$ heißt das Fermat-Weber[7]-Problem:

[5] Pierre de Fermat (1607/08–1665) war französischer Mathematiker und Jurist, bekannt natürlich vor allem durch seinen "letzten Satz".

[6] Evangelista Torricelli (1608–1647) zählt zu den bedeutendsten Physikern und Mathematikern der Barockzeit.

[7] Alfred Weber (1868–1958) war Nationalökonom, Soziologe und Kulturphilosoph und begründete die volkswirtschaftliche Standorttheorie. Er lehrte 1904–1907 in Prag, danach in Heidelberg. In Prag promovierte (mit der schlechtesten, zum Bestehen noch ausreichenden Note) 1906 Franz Kafka bei ihm. Als überzeugter Gegner des Nationalsozialismus wurde Alfred Weber bei der Bundespräsidentenwahl 1954 ohne seine Zustimmung von der KPD für das Amt des Bundespräsidenten vorgeschlagen. Alfred Weber ist Bruder (des noch berühmteren) Max Weber (1864–1920), Jurist, Soziologe und Nationalökonom.

- 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

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

  wobei $\|\cdot\|_2$ die *euklidische Norm* auf dem $\mathbb{R}^n$ bedeutet.

I. Allg. gibt es keine "geschlossene" Lösung zum Fermat-Weber-Problem, man ist auf Iterationsverfahren angewiesen. Anders ist dies bei der sehr ähnlich aussehenden Aufgabe

$$\text{Minimiere} \quad f(x) := \sum_{i=1}^{m} w_i \, \|x - a_i\|_2^2, \quad x \in \mathbb{R}^n,$$

bei der man leicht nachweisen kann, dass der (gewichtete) Schwerpunkt

$$x^* = \Big(1 \Big/ \sum_{i=1}^{m} w_i\Big) \sum_{i=1}^{m} w_i a_i$$

die Lösung ist.

Die ökonomische Interpretation (man spricht in den Wirtschaftswissenschaften auch von dem "Standortproblem" oder "location problem") 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 von den Zulieferern zum Lager und vom Lager zu den Filialen minimal wird. □

Beispiel: Die Konzentration $z(t)$ eines Stoffes in einem chemischen Prozess zur Zeit t gehorche dem Gesetz

$$z(t) = x_1 e^{x_2 t}$$

mit noch unbekannten Parametern x_1, x_2. Zur Bestimmung dieser Parameter liefern Messungen zu Zeiten t_i die Messwerte z_i, $i = 1, \ldots, 10$, die gegeben sind durch

t_i	0.9	1.5	13.8	19.8	24.1	28.2	35.2	60.3	74.6	81.3
z_i	455.2	428.6	124.1	67.3	43.2	28.1	13.1	−0.4	−1.3	−1.5

Die Parameter sind nach der Methode der kleinsten Quadrate als Lösung von

$$\text{Minimiere} \quad f(x) := \sum_{i=1}^{10} (x_1 e^{x_2 t_i} - z_i)^2, \quad x \in \mathbb{R}^2$$

zu bestimmen. Dies nennt man auch ein nichtlineares *Curve Fitting Problem*. In der Optimization-Toolbox von MATLAB hat man hierzu die Funktion `lsqcurvefit`[8], für den allgemeineren Fall eines nichtlinearen Ausgleichsproblems die Funktion `lsqnonlin`. Wir wollen demonstrieren, wie man mit Hilfe der Funktion `lsqcurvefit` obiges Problem lösen kann. Zunächst schreiben wir ein function file `Fit.m` mit dem Inhalt

[8] Das angegebene Beispiel findet man in der MATLAB-Hilfe zu dieser Funktion.

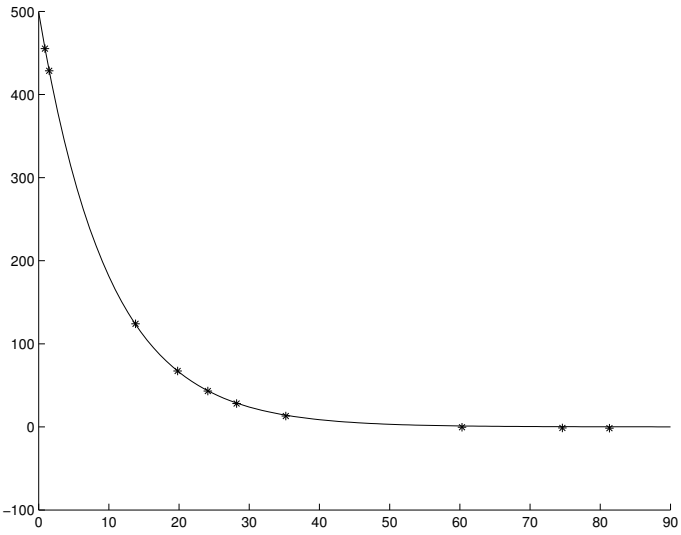


Abbildung 1.4: Daten für ein Curve Fitting Problem

```
function F=Fit(x,t);
F=x(1)*exp(x(2)*t);
```

Im folgenden Aufruf wird ein Startwert `x0` vorgegeben.

```
t=[0.9;1.5;13.8;19.8;24.1;28.2;35.2;60.3;74.6;81.3];
z=[455.2;428.6;124.1;67.3;43.2;28.1;13.1;-0.4;-1.3;-1.5];
x0=[100;-1];
x=lsqcurvefit('Fit',x0,t,z)
```

Als Lösung erhält man

$$x = \begin{pmatrix} 498.8309 \\ -0.1013 \end{pmatrix}.$$

In Abb. 1.4 haben wir den gefundenen Fit eingetragen. □

Beispiel: Sei (siehe C. Geiger, C. Kanzow (1999, S. 4)) $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ gegeben durch

$$f(x) := -x_1^2 x_2 + \frac{1}{4}(2x_1^2 - x_2^2) - \frac{1}{2}(2 - x_1^2 - x_2^2)^2.$$

Die Frage ist: Wo hat f (lokale) Minima, wo (lokale) Maxima, wie sieht f aus? Die Benutzung von MATLAB wird eine ganz wichtige Rolle in der Vorlesung spielen und hier ist eine gute Gelegenheit, dieses Programmsystem anzuwenden. In Abbildung 1.5 links geben wir einen Flächenplot von $(x, f(x))$ mit $x \in [-2, 2] \times [-2, 2]$ wieder. Rechts findet man zugehörige Höhenlinien. Den linken Plot haben wir durch

```
x_1=-2:0.2:2;x_2=x_1;
[X_1,X_2]=meshgrid(x_1,x_2);
F=-(X_1.^2).*X_2+0.25*(2*X_1.^2-X_2.^2)-0.5*(2-X_1.^2-X_2.^2).^2;
surf(X_1,X_2,F);colorbar
```

erhalten, den rechten durch anschließendes

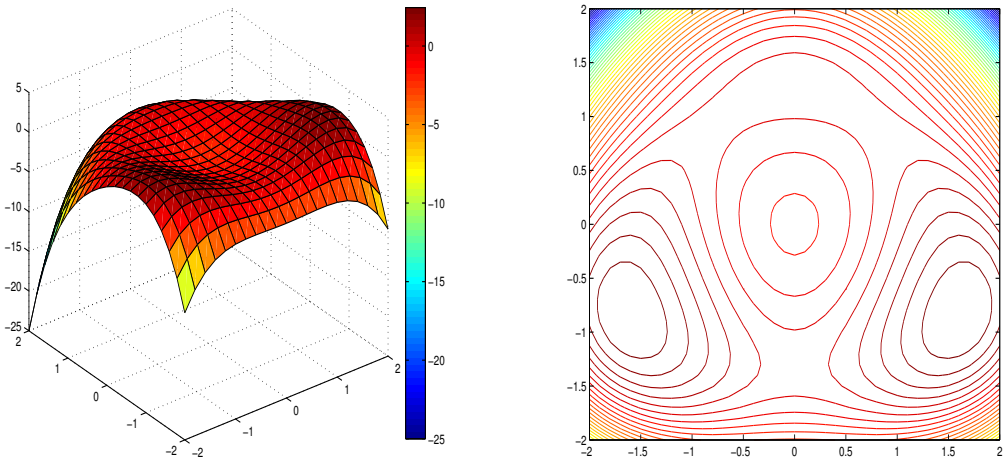


Abbildung 1.5: Flächenplot, Höhenlinienplot

```
contour(X_1,X_2,F,50);
```

Der Höhenlinienplot suggeriert, dass f mindestens drei lokale Extrema besitzt.

Als Extrema kommen nur Punkte in Frage, in denen der Gradient ∇f von f verschwindet, sogenannte *stationäre Punkte* bzw. *stationäre Lösungen*. Es ist

$$\nabla f(x) = \begin{pmatrix} -2x_1x_2 + x_1 + 2x_1(2 - x_1^2 - x_2^2) \\ -x_1^2 - \frac{1}{2}x_2 + 2x_2(2 - x_1^2 - x_2^2) \end{pmatrix}.$$

Wir bestimmen zunächst alle stationären Punkte von f und entscheiden dann, welche davon lokale Minima bzw. Maxima von f sind. Offenbar sind $(0, 0)$ und $(0, \pm\sqrt{3/2})$ stationäre Punkte. Zur Bestimmung stationärer Punkte, bei denen die erste Komponente nicht verschwindet, hat man das nichtlineare Gleichungssystem

$$\begin{aligned} -2x_2 + 1 + 2(2 - x_1^2 - x_2^2) &= 0, \\ -x_1^2 - \frac{1}{2}x_2 + 2x_2(2 - x_1^2 - x_2^2) &= 0 \end{aligned}$$

zu lösen. Mit dem Computeralgebrasystems Maple erhalten wir die weiteren stationären Lösungen $(\pm 1/\sqrt{2}, 1)$ und $(\pm\sqrt{95}/6, -5/6)$. Bekanntlich ist ein stationärer Punkt, in dem die Hessesche[9] $\nabla^2 f$ von f positiv (negativ) definit ist, ein lokales Minimum (Maximum) von f. Als Hessesche von f berechnet man

$$\nabla^2 f(x) = \begin{pmatrix} 5 - x_2 - 6x_1^2 - 2x_2^2 & -2x_1(1 + 2x_2) \\ -2x_1(1 + 2x_2) & \frac{7}{2} - x_2 - 2x_1^2 - 6x_2^2 \end{pmatrix}.$$

In der folgenden Tabelle 1.1 geben wir das Resultat unserer Berechnungen an, wobei wir massiv Maple eingesetzt haben: Hierbei nennen wir einen stationären Punkt einen

[9] Ludwig Otto Hesse (1811–1874) war ein deutscher Mathematiker. Er führte u. a. die Hesse-Matrix und deren Determinante sowie die Hessesche Normalform der Ebene ein.

Stationärer Punkt x^*	Eigenwerte von $\nabla^2 f(x^*)$	Typ
$(0,0)$	$5, \frac{7}{2}$	Lokales Minimum
$(0, \sqrt{\frac{3}{2}})$	$2 - \frac{1}{2}\sqrt{6}, -\frac{11}{2} - \frac{1}{2}\sqrt{6}$	Sattelpunkt
$(0, -\sqrt{\frac{3}{2}})$	$-\frac{11}{2} + \frac{1}{2}\sqrt{6}, 2 + \frac{1}{2}\sqrt{6}$	Sattelpunkt
$(\sqrt{\frac{1}{2}}, 1)$	$-\frac{11}{4} + \frac{1}{4}\sqrt{337}, -\frac{11}{4} - \frac{1}{4}\sqrt{337}$	Sattelpunkt
$(-\sqrt{\frac{1}{2}}, 1)$	$-\frac{11}{4} + \frac{1}{4}\sqrt{337}, -\frac{11}{4} - \frac{1}{4}\sqrt{337}$	Sattelpunkt
$(\frac{1}{6}\sqrt{95}, -\frac{5}{6})$	$-\frac{33}{4} + \frac{1}{36}\sqrt{18849}, -\frac{33}{4} - \frac{1}{36}\sqrt{18849}$	Lokales Maximum
$(-\frac{1}{6}\sqrt{95}, -\frac{5}{6})$	$-\frac{33}{4} + \frac{1}{36}\sqrt{18849}, -\frac{33}{4} - \frac{1}{36}\sqrt{18849}$	Lokales Maximum

Tabelle 1.1: Klassifikation stationärer Punkte

Sattelpunkt, wenn die Hessesche in diesem Punkt indefinit ist. Offenbar existiert kein globales Minimum von f. □

Beispiel: Auf lineare Optimierungsaufgaben wird i. Allg. im Rahmen einer Vorlesung Numerische Mathematik I eingegangen. Als Beispiele werden gerne das Produktionsplanungsproblem oder das Diätproblem angegeben. Hierauf wollen wir nicht eingehen, sondern eine andere Klasse von linearen Optimierungsaufgaben vorstellen, nämlich *Netzwerkflussprobleme*.

Ein Produkt (Öl oder Orangen oder ...) wird an gewissen Orten in einer bestimmten Menge angeboten und an anderen Orten verlangt. Schließlich gibt es Orte, die nichts anbieten und nichts verlangen, an denen das Produkt 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 Obergrenzen für die zu transportierende Menge auf dem Weg an. Gesucht ist ein kostenminimaler, zulässiger (die Kapazitätsbeschränkungen sowie die Angebote bzw. Bedürfnisse respektierender) Transportplan. Wir werden gleich für diese Aufgabenstellung ein mathematisches Modell angeben.

Gegeben sei ein *gerichteter Graph* $(\mathcal{N}, \mathcal{A})$. Hier steht $\mathcal{N}$ für die (endliche) Menge der *Knoten* (**N**odes) und $\mathcal{A}$ für die Menge der *Pfeile* (**A**rcs), also *geordneten* Paaren von Knoten. Mit jedem Knoten $k \in \mathcal{N}$ ist eine Mengenangabe b_k des 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 *Umladeknoten*.

Zu jedem Pfeil $(i,j) \in \mathcal{A}$ gehören die Kosten c_{ij} für den Fluss einer Mengeneinheit auf ihm. Mit x_{ij} wird der Fluss auf diesem Pfeil bezeichnet, die *Kapazität* des Pfeils wird durch $u_{ij} > 0$ angegeben. Gesucht wird ein Fluss im gerichteten Graphen, 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 lässt sich daher wie folgt formulieren:

$$\begin{cases} \text{Minimiere} \quad \displaystyle\sum_{(i,j)\in\mathcal{A}} c_{ij}x_{ij} \\ \text{unter den Nebenbedingungen} \\ \displaystyle\sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik} = b_k \qquad (k \in \mathcal{N}), \qquad 0 \le x_{ij} \le u_{ij} \qquad ((i,j) \in \mathcal{A}). \end{cases}$$

Diese Aufgabe wollen wir nun in Matrix-Vektorschreibweise formulieren. Dies kann folgendermaßen geschehen. Der Fluss $x = (x_{ij})$ hat so viele Komponenten wie es Pfeile gibt, ihre Anzahl sei $n := |\mathcal{A}|$. Es liegt also nahe, $\mathcal{A}$ durchzunummerieren. Es sei etwa $\mathcal{A} = \{l_1, \ldots, l_n\}$. Dann kann $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ als Vektor $x = (x_1, \ldots, x_n)^T$ mit $x_p = x_{l_p}$, $p = 1, \ldots, n$, geschrieben werden, entsprechendes gilt für die Kosten $c = (c_{ij})$ und Kapazitäten $u = (u_{ij})$. Ist ferner $m := |\mathcal{N}|$ 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 die *Knoten-Pfeil-Inzidenzmatrix* $A = (a_{kp}) \in \mathbb{R}^{m\times n}$ durch

$$a_{kp} := \begin{cases} +1, & \text{falls: Der Knoten } k \text{ ist Startknoten für den } p\text{-ten Pfeil } l_p, \\ -1, & \text{falls: Der Knoten } k \text{ ist Endknoten für den } p\text{-ten Pfeil } l_p, \\ 0 & \text{sonst,} \end{cases}$$

so erkennt man, dass obiges Netzwerkflussproblem, das sogenannte (kapazitierte) Minimale-Kosten-Fluss-Problem, in der Form

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

geschrieben werden kann. □

Beispiel: Es sollen 400 m^3 Kies von einem Ort zu einem anderen transportiert werden[10]. Dies geschehe in einer (nach oben!) offenen Box der Länge x_1, der Breite x_2 und der Höhe x_3. Der Boden (x_1x_2 m^2) und die beiden Seiten ($2x_1x_3$ m^2) müssen aus einem Material hergestellt sein, das zwar nichts kostet, von dem aber nur 4 m^2 zur Verfügung steht. Das Material für die beiden Enden ($2x_2x_3$ m^2) kostet 200 € pro m^2. Ein Transport der Box kostet 1 €. Wie hat man die Box zu konstruieren?

[10]Ein ähnliches Beispiel findet man bei C. Geiger, C. Kanzow (2002, S. 6).

Die Kosten zum Bau der Box sind $400x_2x_3$ €. Die Anzahl der Transporte ist $400/(x_1x_2x_3)$ (dies ist nicht notwendig eine natürliche Zahl, was uns aber nicht stört), so dass die Gesamtkosten zum Bau der Box und des Transportes der Kiesmenge durch

$$f(x) := 400\Big(x_2x_3 + \frac{1}{x_1x_2x_3}\Big)$$

gegeben ist. Wegen der Kapazitätsschranken für das Material des Bodens und der beiden Seiten hat man die Restriktion

$$x_1x_2 + 2x_1x_3 \leq 4.$$

Berücksichtigt man noch, dass die Variablen positiv sein müssen, so haben wir insgesamt die Optimierungsaufgabe (wir lassen den Faktor 400 weg)

$$\begin{cases} \text{Minimiere} \quad f(x) := 1/(x_1x_2x_3) + x_2x_3 \quad \text{unter den Nebenbedingungen} \\ \qquad\qquad x_1x_2 + 2x_1x_3 \leq 4, \qquad x_1, x_2, x_3 > 0. \end{cases}$$

Dies ist eine nichtlineare Optimierungsaufgabe. □

1.2 Problemstellungen

Gegeben sei die Optimierungsaufgabe

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

Hierbei sei die Menge M der zulässigen Lösungen[11] eine Teilmenge des $\mathbb{R}^n$ und die Zielfunktion $f : M \longrightarrow \mathbb{R}$ (wenigstens) auf M stetig, i. Allg. sogar glatt, d. h. ein oder zweimal stetig differenzierbar auf einer offenen Obermenge von M. Womit beschäftigt man sich in der Optimierung?

- Existenz einer (globalen) Lösung.

Mit einem $x_0 \in M$ sei die Niveaumenge (engl.: level set)

$$L_0 := \{x \in M : f(x) \leq f(x_0)\}$$

kompakt. Da eine stetige Funktion auf einer kompakten Menge ihre Extrema annimmt, besitzt (P) in diesem Falle eine globale Lösung.

Beispiel: Sei $M \subset \mathbb{R}^n$ nichtleer und abgeschlossen, weiter $f(x) := c^Tx + \frac{1}{2}x^TQx$ mit einer symmetrischen und *positiv definiten* Matrix $Q \in \mathbb{R}^n$. Wir wollen uns überlegen, dass die Niveaumenge

$$L_0 := \{x \in M : f(x) \leq f(x_0)\} = M \cap \{x \in \mathbb{R}^n : c^Tx + \tfrac{1}{2}x^TQx \leq c^Tx_0 + \tfrac{1}{2}x_0^TQx_0\}$$

[11] I. Allg. wird M als Lösungsmenge endlich vieler Gleichungen und Ungleichungen gegeben sein, also in der Form

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

wobei $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$.

kompakt ist. Hierzu genügt es nachzuweisen, dass die rechtsstehende Menge beschränkt und damit (da sie offensichtlich abgeschlossen ist) kompakt ist. Dies erkennt man aus der für $x \in L_0$ gültigen Ungleichungskette

$$\begin{aligned} 0 &\geq c^T x + \tfrac{1}{2} x^T Q x - (c^T x_0 + \tfrac{1}{2} x_0^T Q x_0) \\ &= (c + Q x_0)^T (x - x_0) + \tfrac{1}{2} (x - x_0)^T Q (x - x_0) \\ &\geq (c + Q x_0)^T (x - x_0) + \tfrac{1}{2} \lambda_{\min}(Q) \|x - x_0\|_2^2 \\ &\geq -\|c + Q x_0\|_2 \, \|x - x_0\|_2 + \tfrac{1}{2} \lambda_{\min}(Q) \|x - x_0\|_2^2, \end{aligned}$$

denn hieraus liest man ab, dass die Niveaumenge L_0 in einer (euklidischen) Kugel um x_0 mit dem Radius $2\,\|c + Q x_0\|_2 / \lambda_{\min}(Q)$ enthalten ist. Hierbei bedeutet $\lambda_{\min}(Q)$ den kleinsten (wegen der vorausgesetzten positiven Definitheit von Q positiven) Eigenwert von Q.

In Abbildung 1.6 geben wir für den Spezialfall

$$c := \begin{pmatrix} 5 \\ -3 \end{pmatrix}, \qquad Q := \begin{pmatrix} 8 & -1 \\ -1 & 1 \end{pmatrix}, \qquad x_0 := \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

und $M := \mathbb{R}^2_{\geq 0}$ die zugehörige Niveaumenge an. Das unrestringierte Minimum wurde

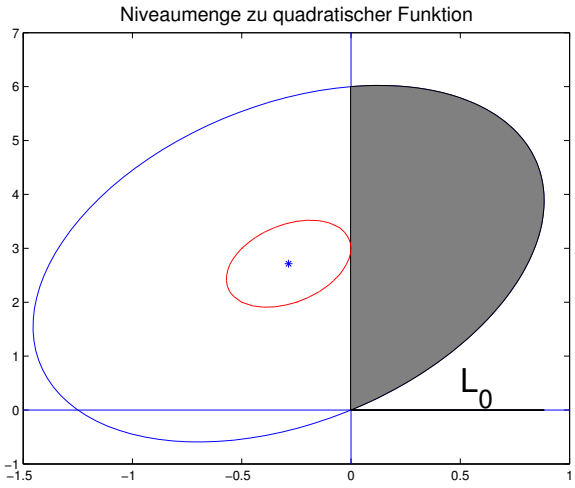


Abbildung 1.6: Niveaumenge

noch durch ein $*$ gekennzeichnet. Die Lösung des restringierten Problems ist $x^* := (0, 3)^T$. □

Mit

$$\inf(\mathrm{P}) := \inf_{x \in M} f(x)$$

bezeichnet man den *Wert* oder auch *Optimalwert* der Optimierungsaufgabe (P). Wenn (P) eine (globale) Lösung besitzt, so schreiben wir $\min(\mathrm{P})$ statt $\inf(\mathrm{P})$. Ist $\inf(\mathrm{P}) = -\infty$, so ist die Zielfunktion auf der Menge der zulässigen Lösungen nicht nach unten beschränkt und die Optimierungsaufgabe (P) besitzt dann natürlich keine Lösung.

Wir werden sehen, dass für gewisse Optimierungsaufgaben, z. B. lineare und (konvexe) quadratische Optimierungsaufgaben, aus der Zulässigkeit von (P) (also $M \neq \emptyset$) und $\inf(\mathrm{P}) > -\infty$ die Existenz eine Lösung von (P) folgt. Dies ist eine Aussage, die i. Allg. natürlich falsch ist[12].

- Eindeutigkeit einer Lösung.

Man nennt (P) eine *konvexe Optimierungsaufgabe*, wenn sowohl die Menge M der zulässigen Lösungen eine konvexe Menge[13] als auch die Zielfunktion f eine auf M konvexe Funktion[14] 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 $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, so ist M konvex, wenn die Komponenten $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, l$, konvex (auf dem $\mathbb{R}^n$) sind, und h eine affin lineare Abbildung ist, also durch $h(x) := A_0 x - b_0$ mit einer Matrix $A_0 \in \mathbb{R}^{m \times n}$ und einem Vektor $b_0 \in \mathbb{R}^m$ gegeben ist.

Offensichtlich ist die Menge M_{opt} der Lösungen einer konvexen Optimierungsaufgabe eine konvexe Menge[15]. Ist die Zielfunktion einer konvexen Optimierungsaufgabe sogar strikt konvex[16], so besitzt die entsprechende Optimierungsaufgabe höchstens eine Lösung, d. h. M_{opt} besteht aus höchstens einem Punkt.

Beispiel: Mit $c \in \mathbb{R}^n$ und der symmetrischen Matrix $Q \in \mathbb{R}^{n \times n}$ sei die Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ definiert durch

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

---

[12] Betrachte die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := e^{-x} \quad \text{auf} \quad M := \{x \in \mathbb{R} : x \geq 0\}.$$

[13] Eine Menge $M \subset \mathbb{R}^n$ heißt *konvex*, wenn mit zwei Punkten aus M auch die gesamte Verbindungsstrecke zu M gehört, wenn also

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

[14] Ist $M \subset \mathbb{R}^n$ konvex und $f : M \longrightarrow \mathbb{R}$, so heißt f auf M *konvex*, wenn

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

[15] Denn seien $x, y \in M_{\mathrm{opt}}$ und $\lambda \in [0,1]$. Dann ist $(1-\lambda)x + \lambda y \in M$ wegen der Konvexität von M und

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda) f(x) + \lambda f(y) = (1-\lambda)\min(\mathrm{P}) + \lambda \min(\mathrm{P}) = \min(\mathrm{P})$$

und daher $(1-\lambda)x + \lambda y \in M_{\mathrm{opt}}$.

[16] Ist $M \subset \mathbb{R}^n$ konvex und $f : M \longrightarrow \mathbb{R}$, so heißt f auf M *strikt konvex*, wenn

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

Für beliebige $x, y \in \mathbb{R}^n$ und $\lambda \in \mathbb{R}$ weist man leicht nach, dass

$$(1-\lambda)f(x) + \lambda f(y) - f((1-\lambda)x + \lambda y) = \tfrac{1}{2}\lambda(1-\lambda)(x-y)^T Q(x-y).$$

Hieraus erkennt man sofort: Ist Q positiv semidefinit, so ist f konvex auf dem $\mathbb{R}^n$. Ist Q sogar positiv definit, so ist f strikt konvex auf dem $\mathbb{R}^n$. □

- 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^* von f partiell differenzierbar, existieren also die partiellen Ableitungen $(\partial f/\partial x_j)(x^*)$, $j = 1, \ldots, n$, von f in x^*, so verschwindet notwendig der *Gradient* $\nabla f(x^*)$ von f in x^*, d. h. es ist

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

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

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

positiv semidefinit ist. Aber auch für restringierte Optimierungsaufgaben sind zumindestens notwendige Optimalitätsbedingungen erster Ordnung schon aus der Analysis bekannt, siehe *Lagrangesche Multiplikatorenregel*. Unsere Aufgabe wird darin bestehen, diese Aussagen auf allgemeine restringierte Optimierungsaufgaben zu übertragen.

- 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.

Ist z. B. die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ 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 x^* bekanntlich 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 $x \in U^* \setminus \{x^*\}$. Diese Aussage werden wir später auf restringierte Optimierungsaufgaben übertragen.

- Wie berechnet oder (bescheidener) approximiert man eine (lokale, globale) Lösung einer Optimierungsaufgabe?

Eine Lösung in endlich vielen Schritten (selbst bei exakter Arithmetik) zu berechnen, ist nur in wenigen Fällen möglich (z. B. eventuell durch das Simplexverfahren bei linearen Optimierungsaufgaben oder durch verwandte Verfahren bei gewissen quadratischen Optimierungsaufgaben), i. Allg. ist man auf Näherungsverfahren angewiesen. Ein Verfahren erzeugt also nach gewissen Vorschriften eine Folge $\{x_k\}$, von der man hofft, dass sie eine (lokale oder stationäre[17]) Lösung approximiert. Bei unrestringierten oder linear restringierten Optimierungsaufgaben macht es Sinn, eine zulässige Folge $\{x_k\} \subset M$ mit fallenden Kosten, also $f(x_{k+1}) < f(x_k)$, $k = 0, 1, \ldots$, zu konstruieren. Bei nichtlinear restringierten Optimierungsaufgaben, insbesondere solchen mit nichtlinearen Gleichungsrestriktionen, ist die exakte Berechnung einer zulässigen Lösung praktisch nicht möglich. Hier wird man also auf die Zulässigkeit der Näherungsfolge verzichten müssen. Dagegen versucht man von Iteration zu Iteration eine Kombination aus Zielfunktionswert und einem Maß für die Unzulässigkeit einer Iterierten, eine sogenannte Straffunktion, zu verkleinern. Ist

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

so sind mit einem $\sigma > 0$ gängige Straffunktionen z. B.

$$\Phi_\sigma(x) := f(x) + \frac{\sigma}{2}\Big(\sum_{i=1}^{l} \max(g_i(x), 0)^2 + \|h(x)\|_2^2\Big)$$

(differenzierbare quadratische Straffunktion) und

$$\Psi_\sigma(x) := f(x) + \sigma\Big(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1\Big)$$

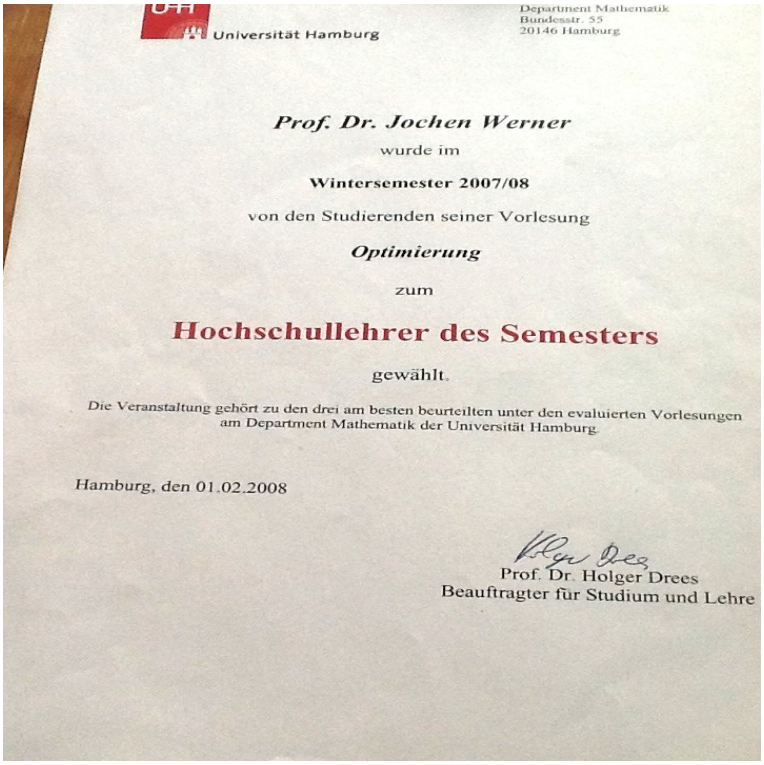
(nichtdifferenzierbare L_1-Straffunktion).

- Diverse weitere Problemstellungen sind denkbar. Eine der wichtigsten (auf die wir aber nicht eingehen werden) dürfte sein: Wie wirken sich Störungen in den Daten einer Optimierungsaufgabe, also der Zielfunktion und/oder der Menge der zulässigen Lösungen, auf die Lösungsmenge und den Wert der Optimierungsaufgabe aus?

Bemerkung: Dies ist das Skript einer Vorlesung, die ich im Wintersemester 2007/2008 in Hamburg gehalten habe. Schon einige Jahre im Ruhestand, hatte ich eine Gastprofessur am (damaligen) Department Mathematik der Universität übernommen. Da die Entscheidung, diese Aufgabe zu übernehmen, schon etwa ein halbes Jahr vor dem Beginn der Gastprofessur feststand und ich keine anderen Verpflichtungen hatte, konnte ich mich mit sehr viel Zeit auf die Vorlesung Optimierung vorbereiten. Hierdurch entstand schon vor Beginn des Wintersemesters 2007/2008 das vorliegende Skript, das ich jetzt (Dezember 2012) noch geringfügig überarbeitete und insbesondere Aufgaben

[17] Unter einer *stationären Lösung* einer Optimierungsaufgabe verstehen wir einen zulässigen Punkt, in dem die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind.

sowie ihre Lösungen hinzufügte. Während meiner Gastprofessur konnte ich mich vollständig auf das Halten der Vorlesung und der zugehörigen Übungen konzentrieren. Ansonsten konnten meine Frau und ich uns auf Hamburg mit allen dort möglichen kulturellen Aktivitäten einlassen. Dadurch hatten wir eine unvergessliche Zeit in der Stadt, in der ich aufgewachsen und zur Schule gegangen bin, in der ich studiert und meine erste eigene Vorlesung gehalten habe. Das i-Tüpfelchen war dann die Wahl zu einem der Hochschullehrer des Semesters, siehe Abbildung 1.7. □


Universität Hamburg

Department Mathematik
Bundesstr. 55
20146 Hamburg

Prof. Dr. Jochen Werner

wurde im

Wintersemester 2007/08

von den Studierenden seiner Vorlesung

Optimierung

zum

Hochschullehrer des Semesters

gewählt.

Die Veranstaltung gehört zu den drei am besten beurteilten unter den evaluierten Vorlesungen am Department Mathematik der Universität Hamburg.

Hamburg, den 01.02.2008

Prof. Dr. Holger Drees
Beauftragter für Studium und Lehre

Abbildung 1.7: Hochschullehrer des Semesters

1.3 Literatur

Wir begnügen uns hier mit der Angabe von fünf deutschsprachigen Lehrbüchern und werden bei Bedarf weitere Literatur nennen, die im Literaturverzeichnis am Ende zusammengefasst ist.

C. Geiger, C. Kanzow (1999) *Numerische Verfahren zur Lösung unrestringierte Optimierungsaufgaben*. Springer, Berlin-Heidelberg.

C. Geiger, C. Kanzow (2002) *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, Berlin-Heidelberg.

W. Alt (2002) *Nichtlineare Optimierung. Eine Einführung in Theorie, Verfahren und Anwendungen*. Vieweg, Braunschweig-Wiesbaden.

P. Spelluci (1993) *Numerische Verfahren der nichtlinearen Optimierung*. Birkhäuser, Basel-Boston-Berlin.

F. Jarre, J. Stoer (2004) *Optimierung*. Springer, Berlin-Heidelberg.

Speziell für unrestringierte Optimierungsaufgaben sei auf Kapitel 7 bei

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

hingewiesen. Noch ausführlicher und aktueller ist das Skript einer Vorlesung über Unrestringierte Optimierungsaufgaben aus dem SS 2002, das man unter `http://www.num.math.uni-goettingen.de/werner/uncopt.pdf` finden kann.

Bemerkt sei noch, dass wir sämtliche biographischen Angaben der freien Enzyklopädie Wikipedia zu verdanken haben.

1.4 Aufgaben

1. Gegeben sei die durch
$$f(x) := 2x_1^2 + x_1x_2^2 + x_2^2$$
definierte Funktion $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$. Man bestimme alle stationären Punkte von f und entscheide, welche davon lokale oder gar globale Extrema von f sind.

2. Gegeben sei die durch
$$f(x) := 2x_1^3 - 3x_1^2 - 6x_1x_2(x_1 - x_2 - 1)$$
definierte Funktion $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$.

   (a) Man bestimme alle stationären Punkte von f und entscheide, welche lokale Minima bzw. Maxima sind.

   (b) Über dem Quadrat $[-2, 2]\times[-2, 2]$ mache man mit MATLAB einen Flächen- und einen Höhenlinienplot von f.

3. 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. Wie hat man r und h bei vorgegebenen positiven V, c_1 und c_2 zu bestimmen?

4. Man löse Tartaglia's Problem: Eine Strecke der Länge 8 ist so in zwei Teile zu zerlegen, dass das Produkt aus dem Produkt und der Differenz der beiden Strecken maximal ist.

5. Sei $M \subset \mathbb{R}^n$ konvex und $f : M \longrightarrow \mathbb{R}$ auf M konvex. Man zeige, dass eine lokale Lösung der konvexen Optimierungsaufgabe

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

   sogar eine globale Lösung von (P) ist.

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

   (a) 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.

   (b) 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.

7. Sei $A \in \mathbb{R}^{m\times n}$ eine Matrix mit Rang $(A) = n$, die n Spalten von A seien also linear unabhängig. Ferner sei $b \in \mathbb{R}^m$ gegeben. Man begründe, weshalb dann das vorzeichenbeschränkte lineare Ausgleichsproblem

   $$\text{Minimiere} \quad \|Ax - b\|_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0\}$$

   eindeutig lösbar ist.

Kapitel 2

Unrestringierte Optimierungsaufgaben

Gegeben sei die unrestringierte Optimierungsaufgabe

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

Hierbei werden wir die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ als *glatt* voraussetzen. Wir unterscheiden zwischen *Schrittweiten-Verfahren* und *Trust-Region-Verfahren*. Sei $x \in \mathbb{R}^n$ eine aktuelle Näherung, es sei $\nabla f(x) \neq 0$, also x keine stationäre Lösung von (P). Wir wollen angeben, wie (im Prinzip) die beiden Verfahren eine (hoffentlich verbesserte) neue Näherung x_+ bestimmen.

Schrittweiten-Verfahren

Zunächst bestimme oder berechne man eine *Abstiegsrichtung* für f in x, also ein $p \in \mathbb{R}^n$ mit $\nabla f(x)^T p < 0$. Wegen

$$\nabla f(x)^T p = \lim_{t \to 0+} \frac{f(x+tp) - f(x)}{t}$$

ist dann $f(x+tp) < f(x)$ für alle hinreichend kleinen $t > 0$. Nun bestimme man eine gewisse *Schrittweite* $t > 0$, die einen bestimmten Mindestabstieg garantiert, für die also z. B. mit einem von (x, p) unabhängigen $\alpha \in (0, 1)$, typischerweise $\alpha = 0.0001$, die Bedingung

$$f(x) - f(x+tp) \geq -\alpha t \nabla f(x)^T p$$

erfüllt ist, und setze als neue Näherung

$$x_+ := x + tp.$$

Trust-Region-Verfahren

Neben der aktuellen Näherung x sei ein "einfaches Modell" $f_x : \mathbb{R}^n \longrightarrow \mathbb{R}$ mit $f_x(0) = f(x)$ für die i. Allg. komplizierte Abbildung $q \mapsto f(x+q)$ und ein $\Delta > 0$ gegeben. Man berechne eine globale Lösung p der Aufgabe

$$\text{Minimiere} \quad f_x(q), \quad \|q\| \leq \Delta$$

wobei $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$ ist. In Abhängigkeit von

$$r := \frac{f(x) - f(x+p)}{f(x) - f_x(p)}$$

akzeptiere man $x_+ := x + p$ als neue Näherung oder setze $x_+ := x$ und mache anschließend einen Update Δ_+ für den Trust-Region-Radius.

2.1 Schrittweitenverfahren

Ein Schrittweiten-Verfahrens ist durch eine Spezifikation einer Richtungs- und einer Schrittweitenstrategie gegeben. Wir werden zunächst einige Schrittweitenstrategien kennenlernen, die hierdurch erreichbare Verminderung im Zielfunktionswert nach unten abschätzen und erst zum Schluss auf mögliche Richtungsstrategien eingehen.

2.1.1 Schrittweitenstrategien: Wolfe- und Armijo-Schrittweite

Die folgenden Voraussetzungen seien in diesem Umterabschnitt erfüllt.

(V) (a) Mit einem gegebenen $x_0 \in \mathbb{R}^n$ (gewöhnlich Startwert eines Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ 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

$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma \|x - y\| \qquad \text{für alle } x, y \in L_0.$$

Hierbei bedeute $\|\cdot\|$ die euklidische Norm[1] auf dem $\mathbb{R}^n$.

Beispiel: Eine beliebte Testfunktion für Verfahren der unrestringierten Optimierung ist die durch

$$f(x) := 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

definierte Funktion $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, die sogenannte *Rosenbrock-Funktion*. Sie wird auch manchmal, siehe die Demos in der Optimization Toolbox von MATLAB, *banana function* genannt, da ihr Minimum in einem langgestreckten Tal liegt. Diese Funktion genügt den Voraussetzungen (V) (a)–(c). Hierbei ist die Abgeschlossenheit einer Niveaumenge wegen der Stetigkeit von f trivial, die Beschränktheit sieht man mit bloßem Auge. Daher ist eine zu f gehörende Niveaumenge kompakt. Da f beliebig oft auf dem ganzen $\mathbb{R}^n$ differenzierbar ist, ist auch (V) (b) erfüllt. Sei B_0 eine konvexe, kompakte Obermenge der (kompakten) Niveaumenge L_0, etwa eine geeignete Kugel. Wegen

$$\nabla f(x) - \nabla f(y) = \int_0^1 \nabla^2 f(y + s(x - y))(x - y)\, ds$$

folgt die Lipschitzstetigkeit von $\nabla f(\cdot)$ auf der Niveaumenge L_0 mit der Lipschitzkonstanten

$$\gamma := \max_{z \in B_0} \|\nabla^2 f(z)\|.$$

Damit erfüllt die Rosenbrock-Funktion die Bedingungen (V) (a)–(c). □

In diesem Unterabschnitt sei $x \in L_0$ gegeben, es sei x keine stationäre Lösung von (P), also $\nabla f(x) \neq 0$, und $p \in \mathbb{R}^n$ eine *Abstiegsrichtung* für f in x bzw. $\nabla f(x)^T p < 0$.

[1] Wegen der Äquivalenz der Normen auf dem $\mathbb{R}^n$ ist $\nabla f(\cdot)$ mit *einer* Norm bezüglich *jeder* Norm lipschitzstetig, wobei sich natürlich die Lipschitzkonstante verändern kann.

Z. B. ist $p := -\nabla f(x)$ (diese Richtungswahl führt auf das schon 1847 von Cauchy[2] angegebene *Gradientenverfahren*, manchmal auch *Verfahren des steilsten Abstiegs* genannt) oder allgemeiner $p := -B\nabla f(x)$ mit einer symmetrischen, positiv definiten Matrix $B \in \mathbb{R}^{n\times n}$ eine Abstiegsrichtung. Ziel dieses Unterabschnittes ist es, Strategien zur Berechnung einer Schrittweite $t > 0$ anzugeben, für welche die Verminderung $f(x) - f(x + tp)$ der Zielfunktion einerseits positiv und andererseits so groß ist, dass (einfache) Konvergenzaussagen für das entstehende, von einer speziellen Richtungsstrategie weitgehend unabhängige Verfahren gemacht werden können.

Bemerkung: Eine naheliegende Schrittweitenstrategie besteht darin, als Schrittweite $t^* > 0$ eine globale oder die erste stationäre Lösung der eindimensionalen Minimierungsaufgabe

$$\text{Minimiere} \quad f(x + tp), \quad t \in [0, \infty),$$

zu wählen. In diesem Fall wird $t^* > 0$ also so bestimmt, dass

$$f(x + t^*p) = \min_{t\geq 0} f(x + tp)$$

bzw.

$$\nabla f(x + t^*p)^T p = 0 \quad \text{und} \quad \nabla f(x + tp)^T p < 0 \quad \text{für alle } t \in [0, t^*).$$

Unter der Voraussetzung (V) ist die Existenz dieser Schrittweite gesichert. Man spricht von einer *exakten Schrittweite*, da eine eindimensionale Minimierungsaufgabe bzw. eine eindimensionale Nullstellenaufgabe zur Bestimmung der Schrittweite exakt zu lösen ist. Es ist klar, dass nur in Ausnahmefällen[3] die exakte Schrittweite (in endlich vielen Schritten) berechnet werden kann, i. Allg. muss man sich mit einer Näherung begnügen. Es kann gezeigt werden, siehe J. Werner (1992b, S. 164), dass

$$-\frac{\nabla f(x)^T p}{\gamma \|p\|^2} \leq t^*, \qquad f(x) - f(x + t^*p) \geq \frac{1}{2\gamma}\Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2.$$

Hier und im folgenden bedeute $\|\cdot\|$ die euklidische Norm, wenn nicht explizit etwas anderes gesagt wird. □

[2] Augustin Louis Cauchy (1789–1857) war ein französischer Mathematiker, insbesondere ein Pionier der Analysis.

[3] Ist z. B.

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

mit einer symmetrischen, positiv definiten Matrix $Q \in \mathbb{R}^{n\times n}$ und $p \in \mathbb{R}^n$ eine Abstiegsrichtung in x bzw. $\nabla f(x)^T p = (c + Qx)^T p < 0$, so ist

$$\frac{d}{dt} f(x + tp) = \nabla f(x + tp)^T p = (c + Qx)^T p + tp^T Qp,$$

so dass die exakte Schrittweite durch

$$t^* := -\frac{(c + Qx)Tp}{p^T Qp}$$

gegeben ist.

Da eine exakte Schrittweite i. Allg. nicht in endlich vielen Schritten realisiert werden kann, sind zunehmend *inexakte* Schrittweiten in Theorie und Praxis untersucht und angewandt worden. Bei der sogenannten *Wolfe-Schrittweite*[4] wird bei vorgegebenen $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ ein $t > 0$ so bestimmt, dass

$$\text{(a)} \qquad f(x + tp) \leq f(x) + \alpha t \nabla f(x)^T p$$

und

$$\text{(b)} \qquad \nabla f(x + tp)^T p \geq \beta \nabla f(x)^T p$$

gelten. Die Bedingung (a) ist für alle hinreichend kleinen $t > 0$ erfüllt, da

$$\lim_{t \to 0+} \frac{f(x + tp) - f(x)}{t} = \nabla f(x)^T p < \alpha \nabla f(x)^T p.$$

Die Forderung (b) sichert, dass nicht zu kleine Schrittweiten gewählt werden, da sie für alle hinreichend kleinen $t > 0$ *nicht* erfüllt ist.

Beispiel: Als Beispiel (siehe auch S. 52) betrachten wir die durch

$$f(x) := (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

definierte Funktion $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ (siehe z. B. W. Alt (2002, S. 12)). Sei $x := (-4, -4)^T$ und $p := (8, 48/7)^T$ (diese etwas seltsamen Daten haben wir von P. Spelluci (1993, S. 105) übernommen). In Abbildung 2.1 links haben wir $\phi(t) := f(x + tp)$ über dem Intervall $[0, 1]$ aufgetragen. In diesem Intervall hat $\phi(\cdot)$ offenbar zwei lokale Minima, das erste etwa bei 0.09. Rechts tragen wir zum einen $\phi(\cdot)$ über dem Intervall $[0, 0.2]$ auf, zum anderen $\psi(t) := f(x) + \alpha t \nabla f(x)^T p$, wobei wir $\alpha := 0.0001$ gewählt haben. Wir erkennen, dass die Bedingung (a) etwa für alle t aus $[0, 0.12]$ erfüllt ist. Dann haben wir auch noch $\chi(t) := [\nabla f(x + tp)^T p - \beta \nabla f(x)^T p]/\|\nabla f(x)\|$ mit $\beta := 0.9$ eingetragen und erkennen, dass die Bedingung (b) für $t > 0.02$ erfüllt ist. Also sind alle t aus dem Intervall $[0.02, 0.12]$ zulässige Wolfe-Schrittweiten. □

Im folgenden Satz (siehe auch C. Geiger, C. Kanzow (1999, S. 38)) wird gezeigt, dass eine Wolfe-Schrittweite existiert, ferner wird die hierdurch erreichbare Verminderung der Zielfunktion nach unten abgeschätzt.

Satz 1.1 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x \in L_0$ keine stationäre Lösung von (P) und p eine Abstiegsrichtung für f in x. Seien $\alpha \in (0, \frac{1}{2})$, $\beta \in (\alpha, 1)$ gegeben und*

$$T_W(x, p) := \left\{ t > 0 : \begin{array}{rcl} f(x + tp) & \leq & f(x) + \alpha t \nabla f(x)^T p, \\ \nabla f(x + tp)^T p & \geq & \beta \nabla f(x)^T p \end{array} \right\}$$

die Menge der Wolfe-Schrittweiten in x in Richtung p. Dann gilt:

[4] Bei J. Werner (1992b, S. 165) und W. Alt (2002, S. 93) heißt diese Schrittweite Powell-Schrittweite, bei C. Geiger, C. Kanzow (1999, S. 37) wird von der Wolfe-Powell-Schrittweite gesprochen.

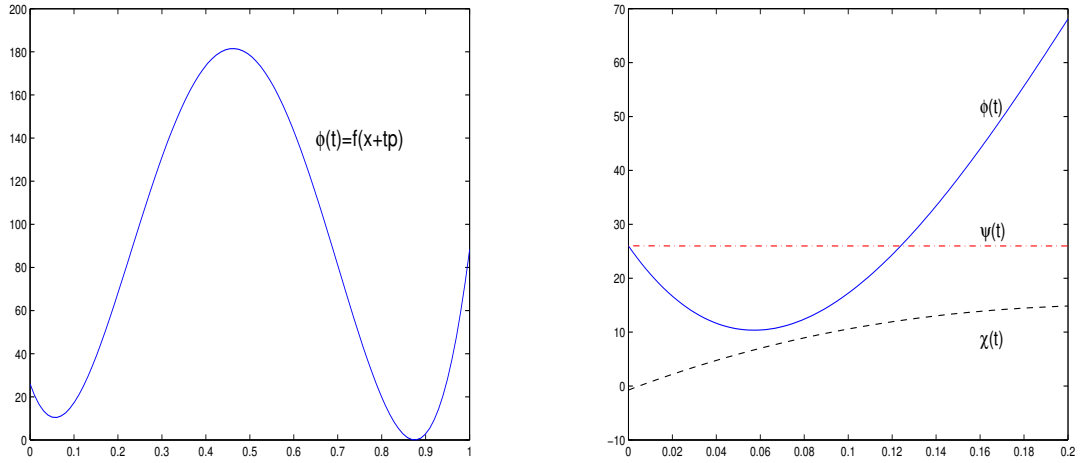


Abbildung 2.1: Exakte Schrittweite, Wolfe-Schrittweite

1. *Es ist* $T_W(x,p) \neq \emptyset$.
2. *Es existiert eine Konstante* $\theta > 0$, *die nur von* α, β *und* γ *(der Lipschitzkonstanten von* $\nabla f(\cdot)$ *auf* L_0*) abhängt, nicht aber von* x *oder* p, *mit*

$$(*) \qquad f(x) - f(x+tp) \geq \theta\Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2 \qquad \text{für alle } t \in T_W(x,p).$$

Beweis: Zur Abkürzung definieren wir

$$\Phi(t) := f(x) + \alpha t \nabla f(x)^T p - f(x+tp).$$

Dann ist $\Phi(0) = 0$, $\Phi'(0) = -(1-\alpha)\nabla f(x)^T p > 0$ und folglich $\Phi(t) > 0$ für alle hinreichend kleinen $t > 0$. Die Annahme, es sei $\Phi(\cdot)$ positiv auf $(0,\infty)$, würde $x+tp \in L_0$ für alle $t > 0$ implizieren, was ein Widerspruch zur vorausgesetzten Kompaktheit von L_0 ist. Daher existiert eine erste positive Nullstelle $\hat{t} > 0$ von $\Phi(\cdot)$. Alle Punkte aus $[0,\hat{t}]$ genügen der ersten Bedingung für eine Wolfe-Schrittweite. Wegen des Satzes von Rolle existiert ein $t \in (0,\hat{t})$ mit $\Phi'(t) = 0$. Wegen $\alpha < \beta$ und

$$0 = \Phi'(t) = \alpha \nabla f(x)^T p - \nabla f(x+tp)^T p \geq \beta \nabla f(x)^T p - \nabla f(x+tp)^T p$$

genügt t auch der zweiten Bedingung für eine Wolfe-Schrittweite. Also ist $t \in T_W(x,p)$, der erste Teil des Satzes ist bewiesen.

Sei $t \in T_W(x,p)$ gegeben. Dann ist $f(x+tp) \leq f(x)$ (direkte Folgerung aus der ersten Bedingung) und daher insbesondere $x + tp \in L_0$. Wegen der zweiten Bedingung sowie der Lipschitzstetigkeit des Gradienten auf der Niveaumenge L_0 ist

$$-(1-\beta)\nabla f(x)^T p \leq [\nabla f(x+tp) - \nabla f(x)]^T p \leq \gamma t \|p\|^2$$

und daher

$$f(x) - f(x+tp) \geq -\alpha t \nabla f(x)^T p \geq \frac{\alpha(1-\beta)}{\gamma}\Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2.$$

Die Behauptung ist mit $\theta := \alpha(1-\beta)/\gamma$ bewiesen. □

Bemerkung: Genauer wollen wir auf die Frage eingehen, wie man in endlich vielen Schritten ein $t \in T_W(x,p)$ bestimmen kann. Wir orientieren uns an J. E. Dennis, R. B. Schnabel (1983, S. 328), siehe auch C. Geiger, C. Kanzow (1999, S. 45 ff.) und P. Spellucci (1993, S. 102 ff.). Die Aufgabenstellung ist die folgende:

- Gegeben: Konstanten $\alpha \in (0, \frac{1}{2})$, $\beta \in (\alpha, 1)$ (z. B. $\alpha := 0.0001$ und $\beta := 0.9$) sowie $x \in L_0$ und $p \in \mathbb{R}^n$ mit $\nabla f(x)^T p < 0$.
- Gesucht: Schrittweite $t > 0$ mit

  (a) $$f(x+tp) \le f(x) + \alpha t \nabla f(x)^T p$$

  und

  (b) $$\nabla f(x+tp)^T p \ge \beta \nabla f(x)^T p.$$

Der Algorithmus wird aus zwei Phasen bestehen. In der ersten Phase wird ein Intervall $[t_{\min}, t_{\max}] \subset (0,\infty)$ bestimmt mit:

- Es ist $t_{\min} < t_{\max}$, weiter ist

  $$f(x+t_{\min}p) \le f(x) + \alpha t_{\min} \nabla f(x)^T p, \qquad f(x+t_{\max}p) > f(x) + \alpha t_{\max} \nabla f(x)^T p$$

  und

  $$\nabla f(x+t_{\min}p)^T p < \beta \nabla f(x)^T p.$$

  Also erfüllt $t_{\min}$ die Bedingung (a), aber nicht (b), während $t_{\max} > t_{\min}$ der Bedingung (a) nicht genügt.

Dann ist es nämlich nicht schwierig zu zeigen:

- Es existiert ein $t \in [t_{\min}, t_{\max}]$, für welches (a) und (b) erfüllt sind.

Denn: Man definiere

$$\Phi(t) := f(x) + \alpha t \nabla f(x)^T p - f(x+tp).$$

Nach Konstruktion ist $\Phi(t_{\min}) \ge 0 > \Phi(t_{\max})$ und

$$\Phi'(t_{\min}) = \alpha \nabla f(x)^T p - \nabla f(x+t_{\min}p)^T p > -(\underbrace{\beta - \alpha}_{>0}) \underbrace{\nabla f(x)^T p}_{<0} > 0.$$

Ist $t^* > t_{\min}$ die erste Nullstelle von $\Phi(\cdot)$ in $(t_{\min}, t_{\max})$, so ist (a) für alle $t \in [t_{\min}, t^*]$ erfüllt. In $(t_{\min}, t^*)$ existiert ein t^{**} mit $\Phi'(t^{**}) = 0$ bzw. $\nabla f(x+t^{**}p)^T p = \alpha \nabla f(x)^T p$. Offensichtlich erfüllt t^{**} auch (b).

Jetzt geben wir den Algorithmus zur Bestimmung des Intervalles $[t_{\min}, t_{\max}]$ an. Am Anfang wird getestet, ob die Schrittweite $t = 1$ den Bedingungen (a) und (b) genügt. Dies ist insbesondere dann vernünftig, wenn $p \approx -\nabla^2 f(x)^{-1} \nabla f(x)$, die Abstiegsrichtung also nahe bei der sogenannten *Newton-Richtung* liegt. Denn dann wird man hoffen, in der Nähe einer lokalen Lösung x^* von einem gedämpften Verfahren (die neue Näherung ist $x_+ = x + tp$ mit einer geeigneten Schrittweite $t > 0$) zu einem ungedämpften Verfahren (hier ist $x_+ = x + p$, die Schrittweite also $t = 1$) übergehen zu können.

(0) Setze $t := 1$.

(1) Gelten (a) und (b), dann: STOP, Ziel erreicht, t ist gesuchte Schrittweite.

(2) Gilt (a) (und nicht (b)), dann:
Setze $t_{\min} := t$.
Solange ((a) gilt): Setze $t := 2t$.
Setze $t_{\max} := t$.

(3) Gilt (a) nicht, dann:
Setze $t_{\max} := t$.
Solange ((a) gilt nicht) oder ((b) gilt): Setze $t := \frac{1}{2}t$.
Setze $t_{\min} := t$.

Aus den Schleifen in (2) und (3) kommt man ganz offensichtlich nach endlich vielen Durchläufen heraus. Denn einerseits gilt (a) für alle hinreichend großen t nicht (andernfalls hätte man einen Widerspruch zur Kompaktheit der Niveaumenge) andererseits gilt (a) für alle hinreichend kleinen t, während (b) für alle hinreichend kleinen t nicht gilt.

Nun kommt es darauf an, in endlich vielen Schritten einen Punkt $t \in [t_{\min}, t_{\max}]$ zu bestimmen, für welchen (a) und (b) erfüllt sind. Die primitivste Methode besteht darin, ein Bisektionsverfahren anzuwenden:

(i) Setze $t := \frac{1}{2}(t_{\min} + t_{\max})$.

(ii) Falls (a) (durch t) nicht erfüllt: $t_{\max} := t$, gehe nach (i).
Andernfalls (d. h. (a) durch t erfüllt):

> Falls (b) (durch t) erfüllt, dann: STOP, Ziel erreicht.
> Andernfalls (d. h. (b) durch t nicht erfüllt): $t_{\min} := t$, gehe nach (i).

Wir müssen uns jetzt überlegen, dass dieses Intervallhalbierungsverfahren nach endlich vielen Schritten abbricht. Angenommen, dies wäre nicht der Fall. Dann existieren Folgen $\{t^k_{\min}\}$ und $\{t^k_{\max}\}$ mit $t^k_{\min} \nearrow t$ und $t^k_{\max} \searrow t$ sowie

$$f(x + t^k_{\min} p) \le f(x) + \alpha t^k_{\min} \nabla f(x)^T p, \qquad f(x + t^k_{\max} p) > f(x) + \alpha t^k_{\max} \nabla f(x)^T p$$

und

$$\nabla f(x + t^k_{\min} p)^T p < \beta \nabla f(x)^T p.$$

Wegen des Mittelwertsatzes existiert $\theta_k \in (0, 1)$ mit

$$\begin{aligned} \alpha \nabla f(x)^T p &\le \frac{f(x + t^k_{\max} p) - f(x + t^k_{\min} p)}{t^k_{\max} - t^k_{\min}} \\ &= \nabla f(x + t^k_{\min} p + \underbrace{\theta_k (t^k_{\max} - t^k_{\min}) p}_{\to 0})^T p \\ &\to \nabla f(x + tp)^T p. \end{aligned}$$

Also ist $\alpha\nabla f(x)^T p \leq \nabla f(x+tp)^T p$. Andererseits folgt aus $\nabla f(x+t_{\min}^k p)^T p < \beta\nabla f(x)^T p$ mit $k \to \infty$, dass $\nabla f(x+tp)^T p \leq \beta\nabla f(x)^T p$. Wegen $\nabla f(x)^T p < 0$ und $\alpha < \beta$ hat man einen Widerspruch erreicht. Das obige Verfahren bricht also nach endlich vielen Schritten ab.

Sei $\phi(t) := f(x+tp)$. Im obigen Intervallhalbierungsverfahren hat man für das linke Intervallende $t_{\min}$ die Werte $\phi(t_{\min})$ und $\phi'(t_{\min}) < 0$, für das rechte Intervallende $t_{\max}$ den Wert $\phi(t_{\max})$ zur Verfügung. Statt des Mittelpunktes des Intervalls $[t_{\min}, t_{\max}]$ kann man die Minimalstelle der durch diese drei Werte bestimmten Parabel berechnen und diese, wenn sie hinreichend im Innern von $[t_{\min}, t_{\max}]$ liegt, als neuen Testwert t akzeptieren. Die gesuchte Parabel ist gegeben durch

$$q(s) = \phi(t_{\min}) + \phi'(t_{\min})(s - t_{\min}) + a_2(s - t_{\min})^2,$$

wobei

$$a_2 := \frac{\phi(t_{\max}) - (\phi(t_{\min}) + \phi'(t_{\min})(t_{\max} - t_{\min}))}{(t_{\max} - t_{\min})^2}.$$

Nun berücksichtige man, dass

$$\begin{aligned}
\phi(t_{\max}) &> f(x) + \alpha t_{\max}\nabla f(x)^T p \\
&= f(x) + \alpha t_{\min}\nabla f(x)^T p + \alpha(t_{\max} - t_{\min})\nabla f(x)^T p \\
&\geq \phi(t_{\min}) + \alpha(t_{\max} - t_{\min})\nabla f(x)^T p \\
&> \phi(t_{\min}) + \beta(t_{\max} - t_{\min})\nabla f(x)^T p \\
&> \phi(t_{\min}) + \phi'(t_{\min})(t_{\max} - t_{\min}).
\end{aligned}$$

Also ist $a_2 > 0$ und $q(\cdot)$ besitzt ein eindeutiges Minimum bei

$$t^* = t_{\min} - \frac{\phi'(t_{\min})}{2a_2} = t_{\min} + \Delta t$$

mit

$$\Delta t := -\frac{\phi'(t_{\min})(t_{\max} - t_{\min})^2}{2[\phi(t_{\max}) - (\phi(t_{\min}) + \phi'(t_{\min})(t_{\max} - t_{\min}))]} > 0.$$

Als neuen Testwert kann man dann z. B.

$$t := \begin{cases} t^*, & t^* \in [t_{\min} + \tau(t_{\max} - t_{\min}), t_{\max} - \tau(t_{\max} - t_{\min})], \\ \frac{1}{2}(t_{\min} + t_{\max}), & t^* \notin [t_{\min} + \tau(t_{\max} - t_{\min}), t_{\max} - \tau(t_{\max} - t_{\min})] \end{cases}$$

setzen. Hierbei ist $\tau \in (0, \frac{1}{2})$ gegeben, etwa $\tau := 0.1$. Damit ist eine mögliche Realisierung der Wolfe-Schrittweite beschrieben. □

Wie schon früher erwähnt spielt die Schrittweite $t = 1$ oft eine besondere Rolle. Bei der *Armijo-Schrittweite* testet man zunächst, ob bei einem vorgegebenen $\alpha \in (0, \frac{1}{2})$ die Ungleichung

$$f(x + tp) \leq f(x) + \alpha t\nabla f(x)^T p,$$

also die Bedingung für hinreichenden Abstieg, für $t := 1$ erfüllt ist. Ist dies der Fall, so wird $t = 1$ als Schrittweite akzeptiert. Andernfalls wird t "kontrolliert verkleinert" und

die Bedingung für hinreichenden Abstieg mit der neuen Schrittweite erneut getestet. Sobald diese erfüllt ist (wir wissen, dass sie für alle hinreichend kleinen $t > 0$ erfüllt ist) wird die entsprechende Schrittweite akzeptiert. In einer einfachen[5] Version, siehe auch C. Geiger, C. Kanzow (1999, S. 35), ist mit einem vorgegebenen $\rho \in (0,1)$ die Armijo-Schrittweite t definiert als $t := \rho^j$, wobei j die kleinste nichtnegative ganze Zahl mit

$$f(x + \rho^j p) \leq f(x) + \alpha \rho^j \nabla f(x)^T p$$

ist.

Im folgenden Satz wird die durch die Armijo-Schrittweite erreichte Verminderung der Zielfunktion nach unten abgeschätzt.

Satz 1.2 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x \in L_0$ keine stationäre Lösung von (P) und p eine Abstiegsrichtung für f in x. Seien $\alpha \in (0, \frac{1}{2})$, $\rho \in (0,1)$ und hiermit die Armijo-Schrittweite $t = \rho^j$ gegeben. Dann existiert eine Konstante $\theta > 0$, die nur von α, γ sowie ρ, nicht aber von x oder p abhängt, mit*

$$(**) \qquad f(x) - f(x + tp) \geq \theta \min\Big[-\nabla f(x)^T p, \Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2\Big].$$

Beweis: Zunächst machen wir eine einfache Hilfsüberlegung. Sei $\hat{t}$ die erste positive Nullstelle der durch $\psi(s) := f(x) - f(x + sp)$ definierten Abbildung $\psi : [0, \infty) \longrightarrow \mathbb{R}$. Wegen Voraussetzung (V) existiert $\hat{t}$ (wegen $\psi(0) = 0$ und $\psi'(0) > 0$ wäre andernfalls $\psi(s) > 0$ und insbesondere $x + sp \in L_0$ für alle positiven s, ein Widerspruch zur vorausgesetzten Kompaktheit von L_0) und es ist $x + sp \in L_0$ für alle $s \in [0, \hat{t}]$. Für $s \in [0, \hat{t}]$ erhält man

$$\begin{aligned} f(x + sp) &= f(x) + s\nabla f(x)^T p + \int_0^s [\nabla f(x + \sigma p) - \nabla f(x)]^T p \, d\sigma \\ &\leq f(x) + s\nabla f(x)^T p + \int_0^s \gamma \|p\|^2 \sigma \, d\sigma \\ &= f(x) + s\nabla f(x)^T p + s^2 \frac{\gamma}{2} \|p\|^2, \end{aligned}$$

wobei wir Voraussetzung (V) (c) und die Cauchy-Schwarzsche[6] Ungleichung benutzten. Setzt man hier $s := \hat{t}$, so erhält man eine Abschätzung für $\hat{t}$ nach unten, nämlich

$$-\frac{2\nabla f(x)^T p}{\gamma \|p\|^2} \leq \hat{t}.$$

Nun zum eigentlichen Beweis für die Aussage über die Armijo-Schrittweite $t = \rho^j$. Ist $j = 0$, wird also die Schrittweite $t = 1$ akzeptiert, so ist

$$f(x) - f(x + tp) \geq -\alpha \nabla f(x)^T p.$$

[5] Eine etwas allgemeinere Version findet man z. B. bei J. Werner (1992b, S. 166 ff.).

[6] Hermann Amandus Schwarz (1843–1921) war ein deutscher Mathematiker.

Ist dagegen $j > 0$, so gelten mit $s := \rho^{j-1}$ zwei Ungleichungen, nämlich:

$$f(x+tp) \le f(x) + \alpha t \nabla f(x)^T p, \qquad f(x+sp) > f(x) + \alpha s \nabla f(x)^T p.$$

Ferner ist $\rho s = t$. Mit obigem $\hat{t}$ machen wir eine Fallunterscheidung. Für $s \le \hat{t}$ ist

$$f(x) + \alpha s \nabla f(x)^T p < f(x+sp) \le f(x) + s \nabla f(x)^T p + s^2 \frac{\gamma}{2} \|p\|^2,$$

daher

$$-\frac{2\rho(1-\alpha)}{\gamma} \frac{\nabla f(x)^T p}{\|p\|^2} \le \rho s = t$$

und folglich

$$f(x) - f(x+tp) \ge -\alpha t \nabla f(x)^T p \ge \frac{2\alpha\rho(1-\alpha)}{\gamma} \Big(\frac{\nabla f(x)^T p}{\|p\|} \Big)^2.$$

Ist dagegen $s > \hat{t}$, so ist wegen der oben bewiesenen Abschätzung von $\hat{t}$ nach unten

$$-\frac{2\rho \nabla f(x)^T p}{\gamma \|p\|^2} \le \rho \hat{t} < \rho s = t$$

und daher

$$f(x) - f(x+tp) \ge -\alpha t \nabla f(x)^T p \ge \frac{2\alpha\rho}{\gamma} \Big(\frac{\nabla f(x)^T p}{\|p\|} \Big)^2.$$

Mit

$$\theta := \alpha \min(1, 2\rho(1-\alpha)/\gamma)$$

ist die Aussage des Satzes bewiesen. □

Bemerkungen: Die Voraussetzung $\alpha \in (0, \frac{1}{2})$ bei der Definition der Wolfe- und der Armijo-Schrittweite könnte durch $\alpha \in (0,1)$ ersetzt werden und die Sätze 1.1 und 1.2 würden immer noch gelten. Bei dem Nachweis der superlinearen Konvergenz von Newton- und Quasi-Newton-Verfahren wird klar werden, weshalb $\alpha \in (0, \frac{1}{2})$ vorausgesetzt wird.

Schrittweiten t, für die eine Aussage wie bei der exakten Schrittweite oder der Wolfe-Schrittweite gemacht werden kann, für die also unter den Voraussetzungen (V) (a)–(c) eine von x und p unabhängige Konstante $\theta > 0$ mit

$$(*) \qquad f(x) - f(x+tp) \ge \theta \Big(\frac{\nabla f(x)^T p}{\|p\|} \Big)^2$$

existiert, wurden von W. Warth, J. Werner (1977) *effizient* genannt. Entsprechend werden Schrittweiten t, wie z. B. die Armijo-Schrittweite, zu denen es unter den Voraussetzungen (V) (a)–(c) eine von x und p unabhängige Konstante $\theta > 0$ mit

$$(**) \qquad f(x) - f(x+tp) \ge \theta \min\Big[-\nabla f(x)^T p, \Big(\frac{\nabla f(x)^T p}{\|p\|} \Big)^2 \Big]$$

gibt, von P. Kosmol (1989, S. 92) *semi-effizient* genannt. Die Beziehungen $(*)$ und $(**)$ stellen sich als fundamental bei der Konvergenzanalyse heraus. Etwas vereinfacht gesagt: Hat man für eine Schrittweitenstrategie $(*)$ bzw. $(**)$ bewiesen, so kann man für die Konvergenzanalyse vergessen, wodurch die Schrittweitenstrategie spezifiziert ist, alleine die Richtungsstrategie spielt danach noch eine Rolle. □

2.1.2 Konvergenz des Modellalgorithmus

Unter dem Modellalgorithmus zur (approximativen) Lösung der unrestringierten Optimierungsaufgabe (P) mit glatter Zielfunktion f verstehen wir ein Verfahren der folgenden Form:

- Sei $x_0 \in \mathbb{R}^n$ gegeben.
- Für $k = 0, 1, \ldots$:
  - Test auf Abbruch: Falls $\nabla f(x_k) = 0$, dann: STOP, da x_k stationäre Lösung von (P) ist.
  - Wahl einer (Abstiegs-) Richtung: Bestimme $p_k \in \mathbb{R}^n$ mit $\nabla f(x_k)^T p_k < 0$.
  - Wahl einer Schrittweite: Bestimme $t_k > 0$ mit $f(x_k + t_k p_k) < f(x_k)$.
  - Bestimme neue Näherung: Setze $x_{k+1} := x_k + t_k p_k$.

Aus dem Modellalgorithmus wird ein konkreter Algorithmus, wenn wir die Richtungs- und die Schrittweitenstrategie spezifizieren.

Der folgende Satz gibt unter verhältnismäßig schwachen Voraussetzungen an die Zielfunktion f sowie an die benutzten Abstiegsrichtungen ein, wie man nicht anders erwarten kann, schwaches Konvergenzergebnis an[7].

Satz 1.3 *Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (V) (a)–(c). Als Schrittweite im Modellalgorithmus verwende man $t_k := t^*(x_k, p_k)$ (exakte Schrittweite), $t_k := t_W(x_k, p_k)$ (Wolfe-Schrittweite) oder $t_k := t_A(x_k, p_k)$ (Armijo-Schrittweite). Zur Abkürzung setze man $g_k := \nabla f(x_k)$. Ferner wird vorausgesetzt:*

1. *Es existiert eine Konstante $\sigma > 0$ mit*

$$-\frac{g_k^T p_k}{\|g_k\| \, \|p_k\|} \geq \sigma, \qquad k = 0, 1, \ldots .$$

2. *Es existiert eine Konstante $\tau > 0$ mit*

$$\|p_k\| \geq \tau \|g_k\|, \qquad k = 0, 1, \ldots .$$

Dann gilt: Jeder Häufungspunkt der durch den Modellalgorithmus mit Abstiegsrichtungen p_k erzeugten Folge $\{x_k\}$ ist eine stationäre Lösung von (P). Besitzt (P) genau eine stationäre Lösung x^* in der Niveaumenge L_0, so konvergiert die gesamte Folge $\{x_k\}$ gegen x^*.

Beweis: Wegen einer Bemerkung zu der exakten Schrittweite und der Sätze 1.1, 1.2 existiert eine Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \min\Big[-g_k^T p_k, \Big(\frac{g_k^T p_k}{\|p_k\|}\Big)^2\Big], \qquad k = 0, 1, \ldots .$$

[7] Siehe auch Satz 4.6 bei C. Geiger, C. Kanzow (1999, S. 27).

Wegen der Voraussetzungen 1. und 2. ist daher

$$f(x_k) - f(x_{k+1}) \geq \theta\sigma \min(\tau, \sigma) \|g_k\|^2, \qquad k = 0, 1, \ldots.$$

Da $\{f(x_k)\}$ eine monoton fallende, nach unten beschränkte Folge ist, konvergiert damit $\{g_k\}$ gegen den Nullvektor.

Ist $x^* \in L_0$ ein Häufungspunkt von $\{x_k\}$, so ist x^* Limes einer konvergenten Teilfolge $\{x_{k(j)}\} \subset \{x_k\}$. Da $\{g_{k(j)}\}$ einerseits gegen $\nabla f(x^*)$ und andererseits gegen 0 konvergiert, ist $\nabla f(x^*) = 0$, also x^* eine stationäre Lösung von (P).

Nun besitze (P) genau eine stationäre Lösung x^* in L_0. Angenommen, $\{x_k\}$ konvergiert nicht gegen x^*. Dann existiert ein $\epsilon > 0$ und eine Teilfolge $\{x_{k(j)}\} \subset \{x_k\}$ mit $\|x_{k(j)} - x^*\| \geq \epsilon$, $j = 1, 2, \ldots$. Da L_0 kompakt ist, kann aus $\{x_{k(j)}\}$ eine gegen ein $\hat{x} \in L_0$ konvergente Teilfolge ausgewählt werden. Als Häufungspunkt der Folge $\{x_k\}$ ist $\hat{x}$ wegen des gerade eben bewiesenen ersten Teils eine stationäre Lösung von (P). Wegen $\|\hat{x} - x^*\| \geq \epsilon$ ist $\hat{x} \neq x^*$. Dies ist ein Widerspruch dazu, dass x^* die einzige stationäre Lösung von (P) ist. □

Bemerkungen: Auf die zweite Voraussetzung in Satz 1.3, also die Existenz einer Konstanten $\tau > 0$ mit

$$\|p_k\| \geq \tau \|g_k\|, \qquad k = 0, 1, \ldots,$$

(hier und im folgenden wird die Abkürzung $g_k := \nabla f(x_k)$ benutzt), kann man offenbar verzichten, wenn nur die exakte Schrittweite oder die Wolfe-Schrittweite (oder eine andere effiziente Schrittweite) verwendet wird.

Eine Folge von Abstiegsrichtungen $\{p_k\}$ wird *gradientenähnlich* genannt, wenn die erste Voraussetzung in Satz 1.3 erfüllt ist, wenn es also eine Konstante $\sigma > 0$ mit

$$-\frac{g_k^T p_k}{\|g_k\| \, \|p_k\|} \geq \sigma, \qquad k = 0, 1, \ldots,$$

gibt. Diese Voraussetzung besagt, dass der Winkel zwischen $-g_k$ und p_k gleichmäßig kleiner als der rechte Winkel sein muss, daher wird sie auch *Winkelbedingung* genannt. Für $p_k := -g_k$, also dem Gradientenverfahren, kann $\sigma = 1$ gewählt werden. □

Beispiel: Man betrachte den Modellalgorithmus mit einer Richtungsfolge $\{p_k\}$, wobei $p_k = -B_k g_k$, $k = 0, 1, \ldots$, mit einer symmetrischen und positiv definiten Matrix $B_k \in \mathbb{R}^{n \times n}$. Mit $\lambda_{\min}(B_k)$ bezeichnen wir den (positiven) kleinsten Eigenwert, mit $\lambda_{\max}(B_k)$ den größten Eigenwert von B_k. Bekanntlich ist $\|B_k\| = \lambda_{\max}(B_k)$ und $\lambda_{\min}(B_k) = 1/\|B_k^{-1}\|$, wobei $\|\cdot\|$ die der euklidischen Norm zugeordnete Matrixnorm ist. Dann ist

$$-\frac{g_k^T p_k}{\|g_k\| \, \|p_k\|} = \frac{g_k^T B_k g_k}{\|g_k\| \, \|B_k g_k\|} \geq \frac{\lambda_{\min}(B_k)}{\lambda_{\max}(B_k)} = \frac{1}{\kappa(B_k)},$$

wobei $\kappa(\cdot)$ die Kondition einer Matrix (bezüglich der euklidischen Norm bzw. der Spektralnorm) bedeutet. Die erste Voraussetzung in Satz 1.3 ist daher erfüllt, wenn $\{\kappa(B_k)\}$ beschränkt ist. Wegen

$$\|p_k\| = \|B_k g_k\| \geq \frac{1}{\|B_k^{-1}\|} \|g_k\| = \lambda_{\min}(B_k) \, \|g_k\|$$

ist die zweite Voraussetzung in Satz 1.3 erfüllt, wenn $\{\|B_k^{-1}\|\}$ beschränkt ist. Beide Voraussetzungen gelten, wenn $\{B_k\}$ eine Folge symmetrischer, gleichmäßig positiv definiter und beschränkter Matrizen ist, wenn es also Konstanten $0 < c \leq d$ mit

$$c\,\|z\|^2 \leq z^T B_k z \leq d\,\|z\|^2 \qquad \text{für alle } z \in \mathbb{R}^n,\ k = 0, 1, \ldots$$

gibt. Dies wiederum ist gleichbedeutend mit der Beschränktheit der Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$. Insbesondere ist dies natürlich für $B_k = I$, also das Gradientenverfahren, der Fall. □

Es wird noch ein *globaler Konvergenzsatz* für den Modellalgorithmus folgen. Globale Konvergenz kann man nur unter einschneidenden Voraussetzungen an die Zielfunktion erwarten. Der entscheidende Begriff ist hier der der *Konvexität* bzw. der *gleichmäßigen Konvexität* einer reellwertigen Funktion.

Definition 1.4 Eine Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ heißt *gleichmäßig konvex* auf der konvexen Menge $D \subset \mathbb{R}^n$, falls eine Konstante $c > 0$ existiert mit

$$(*) \qquad (1-t)f(x) + tf(y) - f((1-t)x + ty) \geq \frac{c}{2}t(1-t)\|x-y\|^2$$
$$\text{für alle } x, y \in D,\ t \in [0,1].$$

(Dagegen heißt f bekanntlich *konvex* auf D, wenn $(*)$ mit $c = 0$ gilt.)

Bemerkung: Die Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ ist gleichmäßig konvex auf der konvexen Menge $D \subset \mathbb{R}^n$ mit der Konstanten $c > 0$, wenn sogar $f(\cdot) - (c/2)\|\cdot\|^2$ konvex auf D ist. Ist z. B. $f(x) := c^T x + \frac{1}{2}x^T Q x$ mit einer symmetrischen, positiv definiten Matrix $Q \in \mathbb{R}^{n\times n}$, so ist $f(\cdot)$ auf dem $\mathbb{R}^n$ gleichmäßig konvex mit der Konstanten $c := \lambda_{\min}(Q)$, dem kleinsten (positiven) Eigenwert von Q. Dagegen ist die Funktion $f(x) := x^4$ auf $[-1, 1]$ zwar konvex, aber nicht gleichmäßig konvex. In Abbildung 2.2 links haben wir $f(\cdot)$ auf $[-1, 1]$ dargestellt. Dagegen ist in Abbildung 2.2 rechts die Funktion $f(x) - (c/2)x^2$ über dem Intervall $[-0.05, 0.05]$ für $c = 0.1$ und $c = 0.01$ aufgetragen. □

Im folgenden Satz wird die Konvexität und die gleichmäßige Konvexität einer glatten Funktion f durch ihre ersten Ableitungen charakterisiert.

Satz 1.5 *Sei $D \subset \mathbb{R}^n$ konvex und $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ auf einer offenen Obermenge von D stetig differenzierbar. Dann gilt:*

1. *f ist genau dann auf D konvex, wenn*

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

2. *f ist genau dann auf D gleichmäßig konvex (mit einer Konstanten $c > 0$), wenn*

$$\frac{c}{2}\|y-x\|^2 + \nabla f(x)^T(y-x) \leq f(y) - f(x) \qquad \textit{für alle } x, y \in D.$$

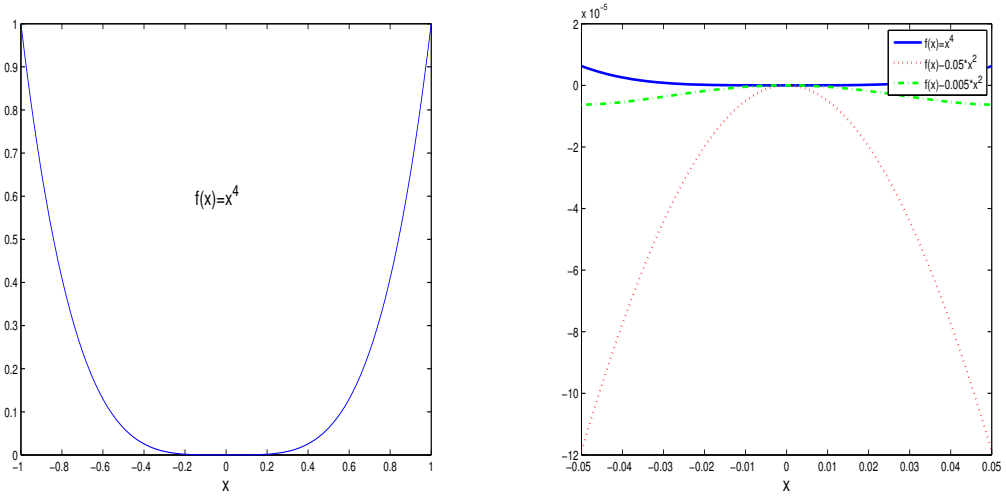


Abbildung 2.2: Eine konvexe, aber nicht gleichmäßig konvexe Funktion

Beweis: Für alle $x, y \in D$ und $t \in [0,1]$ sei

$$(1-t)f(x) + tf(y) \geq f((1-t)x + ty) + \frac{c}{2}t(1-t)\|y-x\|^2$$

mit einer Konstanten $c \geq 0$. Es sei also f konvex ($c = 0$) bzw. gleichmäßig konvex ($c > 0$). Dann ist

$$f(y) - f(x) \geq \frac{f(x + t(y-x)) - f(x)}{t} + \frac{c}{2}(1-t)\|y-x\|^2 \qquad \text{für alle } t \in (0,1].$$

Mit $t \to 0+$ folgt

$$f(y) - f(x) \geq \nabla f(x)^T(y-x) + \frac{c}{2}\|y-x\|^2.$$

Damit ist eine Richtung (nämlich "$\Longrightarrow$") bewiesen. Für die andere Richtung "$\Longleftarrow$" nehmen wir an, mit einer Konstanten $c \geq 0$ sei

$$\frac{c}{2}\|y-x\|^2 + \nabla f(x)^T(y-x) \leq f(y) - f(x) \qquad \text{für alle } x, y \in D.$$

Seien $x, y \in D$ und $t \in [0,1]$ vorgegeben. Dann ist $z := (1-t)x + ty \in D$ wegen der Konvexität von D und daher nach Voraussetzung

$$\begin{aligned} f(x) - f(z) &\geq \nabla f(z)^T(x-z) + \frac{c}{2}\|x-z\|^2, \\ f(y) - f(z) &\geq \nabla f(z)^T(y-z) + \frac{c}{2}\|y-z\|^2. \end{aligned}$$

Eine Multiplikation dieser Ungleichungen mit $(1-t)$ bzw. t und anschließende Addition ergibt

$$\begin{aligned} (1-t)f(x) + tf(y) - f((1-t)x + ty) &\geq \frac{c}{2}[(1-t)\|x-z\|^2 + t\|y-z\|^2] \\ &= \frac{c}{2}t(1-t)\|x-y\|^2. \end{aligned}$$

Also ist f konvex ($c = 0$) bzw. gleichmäßig konvex ($c > 0$). □

Bemerkung: In Abbildung 2.3 veranschaulichen wir uns die erste Aussage in Satz 1.5. Hierbei betrachten wir links ein eindimensionales, rechts ein zweidimensionales

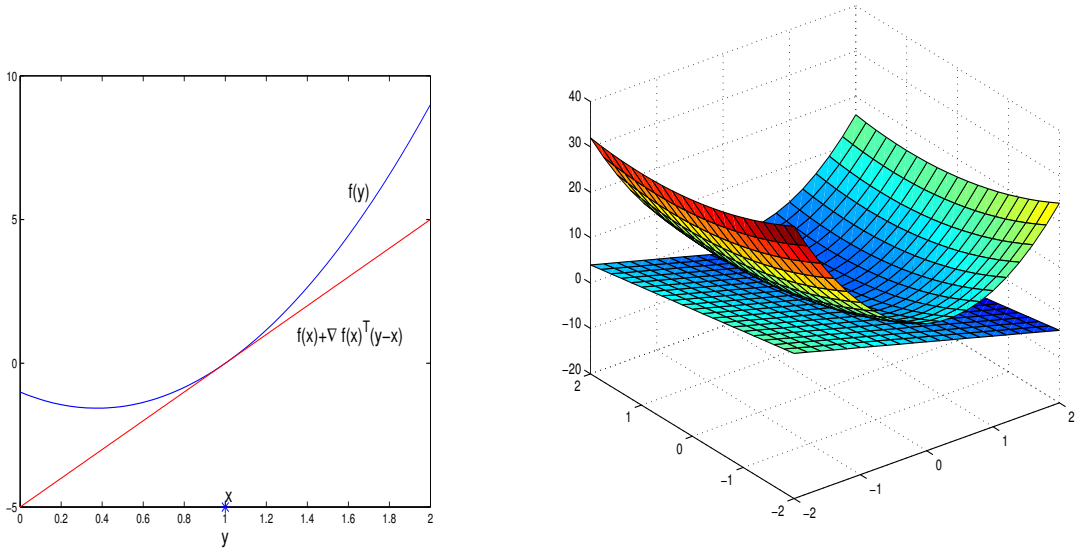


Abbildung 2.3: Charakteristische Eigenschaft konvexer Funktionen

Beispiel. □

Bevor wir den globalen Konvergenzsatz für den Modellalgorithmus bei gleichmäßig konvexer Zielfunktion formulieren und beweisen, fassen wir einige Hilfsmittel in dem folgenden Lemma zusammen.

Lemma 1.6 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die folgenden Konvexitäts- und Glattheitsvoraussetzungen an die Zielfunktion f seien erfüllt:*

(K) *(a)* *Mit einem gegebenen $x_0 \in \mathbb{R}^n$ (Startwert eines Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ konvex.*

(b) *Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar und auf L_0 gleichmäßig konvex, d. h. es existiert eine Konstante $c > 0$ mit*

$$\frac{c}{2}\|y - x\|^2 + \nabla f(x)^T(y - x) \leq f(y) - f(x) \qquad \text{für alle } x, y \in L_0.$$

.

(c) *Der Gradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit*

$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma \|x - y\| \qquad \text{für alle } x, y \in L_0.$$

Dann ist die Niveaumenge L_0 kompakt, (P) besitzt daher eine globale Lösung x^*, diese liegt in L_0 und ist die einzige stationäre Lösung von (P) in L_0. Ferner gilt die Fehlerabschätzung

$$\frac{c}{2}\|x - x^*\|^2 \le f(x) - f(x^*) \le \frac{1}{2c}\|\nabla f(x)\|^2 \qquad \text{für alle } x \in L_0.$$

Beweis: Die Niveaumenge L_0 ist abgeschlossen. Für alle $x \in L_0$ ist wegen der gleichmäßigen Konvexität von f ferner

$$\frac{c}{2}\|x - x_0\|^2 + \nabla f(x_0)^T(x - x_0) \le f(x) - f(x_0) \le 0$$

und daher mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$L_0 \subset \Big\{x \in \mathbb{R}^n : \|x - x_0\| \le \frac{2}{c}\|\nabla f(x_0)\|\Big\}.$$

Insgesamt ist L_0 kompakt, die auf L_0 stetige Funktion f nimmt auf L_0 ihr (globales) Minimum an. Da eine globale Lösung von (P) nicht außerhalb von L_0 liegen kann, ist die Existenz einer globalen Lösung $x^* \in L_0$ bewiesen. Natürlich ist $\nabla f(x^*) = 0$, also x^* auch eine stationäre Lösung von (P). Wir zeigen nun noch die behaupteten Abschätzungen, aus denen insbesondere die Eindeutigkeit einer stationären Lösung von (P) in L_0 folgt.

Die erste Ungleichung folgt direkt aus der vorausgesetzten gleichmäßigen Konvexität, indem man $y = x$ und $x = x^*$ setzt. Bei festem $x \in L_0$ ist

$$-\frac{1}{2c}\|\nabla f(x)\|^2 \le \frac{c}{2}\|x^* - x\|^2 + \nabla f(x)^T(x^* - x) \le f(x^*) - f(x).$$

Dies erkennt man daran, dass die Aufgabe

$$\text{Minimiere} \quad f_x(p) := \frac{c}{2}\|p\|^2 + \nabla f(x)^T p, \quad p \in \mathbb{R}^n,$$

die eindeutige Lösung $p^* := -(1/c)\nabla f(x)$ besitzt. Insgesamt ist der Satz damit bewiesen. □

Es folgt der globale Konvergenzsatz[8] für den Modellalgorithmus.

Satz 1.7 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzungen (K) (a)–(c) aus Lemma 1.6 seien erfüllt. Zur Lösung von (P) betrache man den Modellalgorithmus mit Abstiegsrichtungen p_k und Schrittweiten $t_k := t^*(x_k, p_k)$ (exakte Schrittweite), $t_k := t_W(x_k, p_k)$ (Wolfe-Schrittweite) oder $t_k := t_A(x_k, p_k)$ (Armijo-Schrittweite). Zur Abkürzung sei $g_k := \nabla f(x_k)$ gesetzt. Schließlich sei*

$$\delta_k := \begin{cases} \min\Big[-\dfrac{g_k^T p_k}{\|g_k\|^2}, \Big(\dfrac{g_k^T p_k}{\|g_k\|\,\|p_k\|}\Big)^2\Big], & \text{falls} \quad t_k = t_A(x_k, p_k), \\ \Big(\dfrac{g_k^T p_k}{\|g_k\|\,\|p_k\|}\Big)^2, & \text{falls} \quad t_k = t^*(x_k, p_k), t_k = t_W(x_k, p_k). \end{cases}$$

Dann gilt:

[8] Siehe auch Satz 4.7 bei C. Geiger, C. Kanzow (1999, S. 29).

1. *Ist*
$$\sum_{j=0}^{\infty} \delta_j = \infty,$$
so konvergiert die durch den Modellalgorithmus erzeugte Folge $\{x_k\}$ *gegen die eindeutige (globale) Lösung* x^* *von (P).*

2. *Existiert ein* $\delta > 0$ *mit*
$$\delta \leq \frac{1}{k+1} \sum_{j=0}^{k} \delta_j, \qquad k = 0, 1, \ldots,$$
so konvergiert die Folge $\{x_k\}$ *R-linear gegen* x^**, d. h. es existieren Konstanten* $C > 0$ *und* $q \in (0, 1)$ *mit* $\|x_k - x^*\| \leq Cq^k$*,* $k = 0, 1, \ldots$.

Beweis: Wegen der Effizienz bzw. Semi-Effizienz der gewählten Schrittweite sowie der Definition der δ_k existiert eine von k unabhängige Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \delta_k \|g_k\|^2 \geq 2c\theta\delta_k [f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots,$$

wobei auch noch die Fehlerabschätzung aus Lemma 1.6 benutzt wurde. Daher ist

$$\begin{aligned} 0 \leq f(x_{k+1}) - f(x^*) &\leq (1 - 2c\theta\delta_k)[f(x_k) - f(x^*)] \\ &\leq \prod_{j=0}^{k} (1 - 2c\theta\delta_j)[f(x_0) - f(x^*)] \\ &\leq \exp\Big(-2c\theta \sum_{j=0}^{k} \delta_j\Big)[f(x_0) - f(x^*)]. \end{aligned}$$

Wegen $\sum_{j=0}^{\infty} \delta_j = \infty$ konvergiert $\{f(x_k)\}$ gegen $f(x^*)$. Wiederum wegen der Fehlerabschätzung in Lemma 1.6 folgt die Konvergenz von $\{x_k\}$ gegen x^*.

Existiert ein $\delta > 0$ mit $\delta(k+1) \leq \sum_{j=0}^{k} \delta_j$, $k = 0, 1, \ldots$, so ist

$$f(x_k) - f(x^*) \leq \exp\Big(-2c\theta \sum_{j=0}^{k-1} \delta_j\Big)[f(x_0) - f(x^*)] \leq \exp(-2c\theta\delta k)[f(x_0) - f(x^*)].$$

Mit Hilfe von $\|x_k - x^*\| \leq \{2[f(x_k) - f(x^*)]/c\}^{1/2}$ (siehe Lemma 1.6) folgt daher

$$\|x_k - x^*\| \leq \Big\{\frac{2[f(x_0) - f(x^*)]}{c}\Big\}^{1/2} \exp(-c\theta\delta)^k, \qquad k = 0, 1, \ldots.$$

Der Satz ist damit bewiesen. □

Bemerkung: Wird im Modellalgorithmus unter den Voraussetzungen von Satz 1.7 stets die exakte oder die Wolfe-Schrittweite (oder eine andere effiziente Schrittweite) gewählt, so ist

$$\delta_k = \Big(\frac{g_k^T p_k}{\|g_k\| \, \|p_k\|}\Big)^2, \qquad k = 0, 1, \ldots.$$

Die Bedingung $\sum_{j=0}^{\infty} \delta_j = \infty$ besagt, dass der Winkel zwischen $-g_k$ und p_k sich zwar einem rechten Winkel annähern, dies aber nicht zu schnell geschehen darf. Sie ist natürlich erfüllt, wenn die Winkelbedingung aus Satz 1.3 gilt. □

Beispiel: Wir wollen das Gradientenverfahren mit exakter Schrittweite auf die Minimierung der durch

$$f(x) := 2x_1^2 + x_2^2 - 4x_1 - 2x_2$$

definierten Funktion $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ anwenden. Das eindeutige Minimum liegt in $x^* := (1,1)^T$. In Abbildung 2.4 links geben wir mit dem Startwert $x_0 := (5,-5)^T$ die ersten Iterierten an. Auch in der Nähe der Lösung beobachtet man den typischen Zick-Zack-

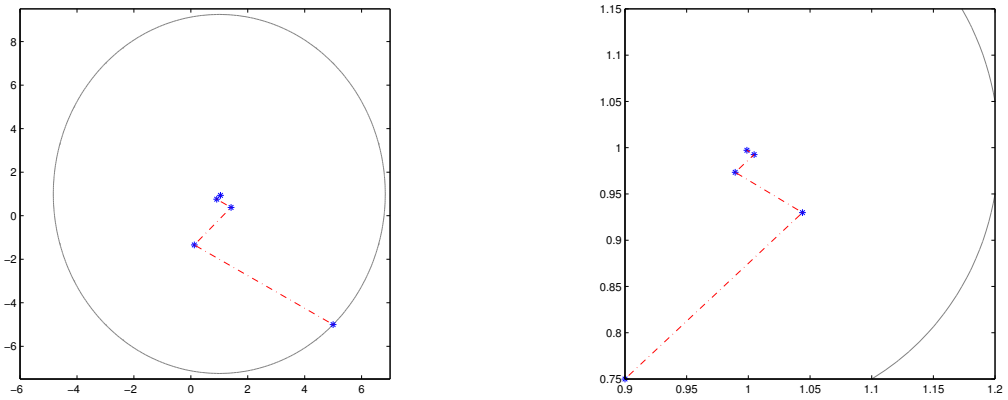


Abbildung 2.4: Das Gradientenverfahren mit exakter Schrittweite

Weg. In Abbildung 2.4 rechts sind wir z. B. mit $(0.9, 0.75)^T$ gestartet. □

2.1.3 Newton-, Quasi-Newton- und BFGS-Verfahren

Das BFGS-Verfahren (hier steht BFGS für **B**royden, **Fletcher**, **G**oldfarb, **S**hanno, die dieses Verfahren unabhängig voneinander 1970 entdeckten) gilt für glatte, nicht zu hochdimensionale, unrestringierte Optimierungsaufgaben, bei denen neben dem Zielfunktionswert auch der Gradient zur Verfügung steht, als das beste Minimierungsverfahren. Es gehört zur Klasse der *Quasi-Newton*[9]*-Verfahren* und dürfte das praktisch und theoretisch am besten erforschte Verfahren dieser Klasse sein. Das ist der Grund, weshalb wir uns auf das BFGS-Verfahren konzentrieren werden. Ein Quasi-Newton-Verfahren ist natürlich verwandt mit dem Newton-Verfahren, versucht aber die Nachteile dieses Verfahrens zu vermeiden und die Vorteile zu bewahren. Daher müssen wir zunächst ganz kurz auf das Newton-Verfahren zu sprechen kommen, wobei wir (hoffentlich) auf einige wenige Vorkenntnisse aus einer Vorlesung über Numerische Mathematik zurückgreifen können.

Gegeben sei die unrestringierte Optimierungsaufgabe (P), wobei die Zielfunktion f zweimal stetig differenzierbar ist. Ein Schritt des (ungedämpften) Newton-Verfahrens

[9] Sir Isaac Newon (1789–1857) war ein englischer Physiker, Mathematiker, Astronom, Alchemist, Philosoph und Verwaltungsbeamter (Wikipedia).

zur Bestimmung eines stationären Punktes von f bzw. einer Lösung des (i. Allg. nichtlinearen) Gleichungssystems $\nabla f(x) = 0$ liefert, ausgehend von einer aktuellen Näherung x, die neue Näherung

$$x_+ := x - \nabla^2 f(x)^{-1} \nabla f(x).$$

Ein aus der Numerischen Mathematik bekannter lokaler Konvergenzsatz, siehe z. B. J. WERNER (1992a, S. 102) für das Newton-Verfahren impliziert: Ist $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ nichtsingulär (z.B. $\nabla^2 f(x^*)$ positiv definit, also insbesondere x^* eine lokale Lösung von (P)), so ist das Newton-Verfahren lokal superlinear konvergent. Das heißt: Wählt man ein Startelement x_0 aus einer hinreichend kleinen Kugel um x^*, so ist das Newton-Verfahren

$$x_{k+1} := x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k), \qquad k = 0, 1, \ldots,$$

durchführbar (d. h. die Matrizen $\nabla^2 f(x_k)$ sind nichtsingulär) und liefert eine gegen x^* konvergente Folge $\{x_k\}$, die *superlinear* konvergiert, für die also

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

Ist die Hessesche $\nabla^2 f(\cdot)$ auf einer Kugel um x^* in x^* sogar lipschitzstetig, existieren also positive η und L mit

$$\|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leq L \|x - x^*\| \qquad \text{falls} \quad \|x - x^*\| \leq \eta,$$

so ist das Newton-Verfahren sogar *quadratisch konvergent*, d. h. es existiert eine Konstante $C > 0$ mit

$$\|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2, \qquad k = 0, 1, \ldots.$$

Man beachte, dass das ungedämpfte Newton-Verfahren versucht, das Gleichungssystem $\nabla f(x) = 0$ zu lösen bzw. einen stationären Punkt von f zu finden. Hierbei sind lokale Minima bzw. Maxima und Sattelpunkte völlig gleichberechtigt. Dem ungedämpften Newton-Verfahren ist sozusagen noch nicht gesagt worden, dass es eine Minimierungsaufgabe lösen soll.

Ein einfaches MATLAB function M-file `Unged_Newton.m` zum ungedämpften Newton-Verfahren könnte folgendermaßen aussehen:

```
function [x,iter]=Unged_Newton(fun,x_0,max_iter,tol);
%****************************************************************
%Input-Parameter:
%          fun        Zu minimierende Funktion
%                     [f,g,H]=fun(x) gibt Funktionswert,
%                     Gradient und Hessesche in x
%          x_0        Startvektor
%          max_iter   Maximale Iterationszahl
%          tol        Verfahren stoppt, wenn Norm des
%                     Gradienten  <=tol
%Output-Parameter:
```

```
%          x          Naeherungsloesung (wenn erfolgreich)
%          iter       Zahl der Iterationen
%*************************************************************
x=x_0; [f,g,H]=feval(fun,x);iter=0;
while (norm(g)>tol)&(iter<max_iter)
     p=-H\g; x=x+p;
     [f,g,H]=feval(fun,x);
     iter=iter+1;
end;
```

Beispiel: Wir kommen auf ein Beispiel aus Kapitel 1 zurück. Es sei $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ definiert durch

$$f(x) := -x_1^2 x_2 + \frac{1}{4}(2x_1^2 - x_2^2) - \frac{1}{2}(2 - x_1^2 - x_2^2)^2.$$

Wir schreiben ein MATLAB function M-file `Geiger.m` (zu Ehren von Carl Geiger, da die Funktion im Buch von Geiger-Kanzow vorkommt), in dem der Funktionswert, der Gradient und die Hessesche von f berechnet werden. Das könnte folgendermaßen aussehen:

```
function [f,g,H]=Geiger(x);
c=2-x(1)^2-x(2)^2;
f=-x(1)^2*x(2)+0.25*(2*x(1)^2-x(2)^2)-0.5*c^2;
if nargout>1
    g=[-2*x(1)*x(2)+x(1)+2*x(1)*c;-x(1)^2-0.5*x(2)+2*x(2)*c];
end;
if nargout>2
    H=[5-x(2)-6*x(1)^2-2*x(2)^2,-2*x(1)*(1+2*x(2));
       -2*x(1)*(1+2*x(2)),3.5-x(2)-2*x(1)^2-6*x(2)^2];
end;
```

Ein Aufruf

```
[x,iter]=Unged_Newton('Geiger',[5;4],100,1e-8);
```

ergibt (nach `format long`)

$$x = \begin{pmatrix} 0.70710678441217 \\ 0.99999999614887 \end{pmatrix}, \qquad \text{iter} = 18.$$

Hier wird also der Sattelpunkt $(1/\sqrt{2}, 1)$ approximiert. Mit dem Startwert $(2, -1)$ wird z. B. das (lokale) Maximum in $(\frac{1}{6}\sqrt{95}, -\frac{5}{6})$ approximiert, entsprechendes gilt z.B. auch für den Startwert $(0.3, -0.7)$. Experimente zeigen, dass man in einer kleinen Umgebung von $(0, 0)$ starten muss, um Konvergenz des ungedämpften Newton-Verfahrens gegen dieses (lokale) Minimum zu erreichen. □

Ist x eine aktuelle Näherung mit $\nabla f(x) \neq 0$, also x kein stationärer Punkt von f, und ist ferner $\nabla^2 f(x)$ positiv definit, so ist $p := -\nabla^2 f(x)^{-1} \nabla f(x)$ eine Abstiegsrichtung. Denn mit $\nabla^2 f(x)$ ist auch $\nabla^2 f(x)^{-1}$ positiv definit und daher

$$\nabla f(x)^T p = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0.$$

Es liegt in diesem Falle nahe, das Newton-Verfahren durch die Einführung einer Schrittweite $t > 0$ (exakte, Wolfe- oder Armijo-Schrittweite) zu globalisieren bzw. zu dem gedämpften Newton-Verfahren

$$x_+ := x - t\nabla^2 f(x)^{-1}\nabla f(x)$$

überzugehen.

Beispiel: Wir kommen auf das vorige Beispiel zurück und nehmen als Startwert für ein gedämpftes Newton-Verfahren $x_0 := (0.3, -0.7)^T$ und hoffen auf Konvergenz gegen das lokale Minimum $x^* = (0, 0)^T$. Dämpfen wir mit der Wolfe-Schrittweite, so wird nur im ersten Schritt eine kleinere Schrittweite als 1 genommen, danach geht das Verfahren in das ungedämpfte Verfahren über und liefert sehr schnell sehr gute Näherungen für das einzige lokale Minimum von f. Da die Funktion f kein globales Minimum besitzt, die Niveaumenge nicht kompakt ist und die Bedingung (a) zur Bestimmung der Wolfe-Schrittweite für gewisse Startwerte für beliebig große Schrittweiten erfüllt ist, ist die Grundlage unserer Implementation nicht erfüllt und man erhält für viele Startwerte (z.B. $x_0 = (0.5, 0.4)^T$) Schwierigkeiten. □

Durch die Einführung von Schrittweiten, also den Übergang zum gedämpften Verfahren

$$x_{k+1} := x_k - t_k\nabla^2 f(x_k)^{-1}\nabla f(x_k)$$

kann man versuchen, zu einem global konvergenten Verfahren zu kommen. Unter geeigneten Konvexitätsvoraussetzungen, die u. a. sichern, dass $\nabla^2 f(x_k)$ positiv definit und damit die Newton-Richtung $p_k := -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$ eine Abstiegsrichtung ist, wird man die Konvergenz des gedämpften Newton-Verfahrens erwarten. Ferner wird man die Schrittweitenstrategie so gestalten wollen, dass nach endlich vielen Schritten, wenn die durch das gedämpfte Newton-Verfahren erzeugten Näherungen erst einmal hinreichend nahe bei einer Lösung liegen, automatisch vom gedämpften zum ungedämpften Newton-Verfahren übergegangen wird. Diese Erwartungen werden durch den folgenden globalen Konvergenzsatz für das Newton-Verfahren bestätigt.

Satz 1.8 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Über die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ wird vorausgesetzt:*

(a) Mit einem $x_0 \in \mathbb{R}^n$ ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ konvex.

(b) f ist auf einer offenen Obermenge von L_0 zweimal stetig differenzierbar und es existieren positive Konstanten $c \leq \gamma$ mit

$$(*) \qquad c\,\|p\|^2 \leq p^T\nabla^2 f(x)p \leq \gamma\,\|p\|^2 \qquad \text{für alle } x \in L_0,\ p \in \mathbb{R}^n.$$

Zur Bestimmung der unter diesen Voraussetzungen eindeutig existierenden (globalen) Lösung x^* von (P) betrachte man das gedämpfte Newton-Verfahren

$$x_{k+1} := x_k + t_k p_k \qquad \text{mit} \quad p_k := -\nabla^2 f(x_k)^{-1}\nabla f(x_k),$$

wobei t_k in jedem Schritt die Wolfe- oder die Armijo-Schrittweite sei. Dann gilt: Bricht das Verfahren nicht vorzeitig mit der Lösung x^* von (P) ab, so erzeugt es eine gegen x^* konvergente Folge $\{x_k\}$. Ferner ist $t_k = 1$ für alle hinreichend größen k, nach endlich vielen Schritten geht das gedämpfte Newton-Verfahren also in das ungedämpfte über.

Beweis: Wir wollen Satz 1.7 anwenden und müssen hierzu zunächst zeigen, dass die Voraussetzungen (K) (a)–(c) aus Lemma 1.6 erfüllt sind. Es sind nur (K) (b)–(c), also die gleichmäßige Konvexität von f und die Lipschitzstetigkeit des Gradienten ∇f auf der Niveaumenge L_0, nachzuweisen. Aus $c\,\|p\|^2 \le p^T\nabla^2 f(x)p$ für alle $x \in L_0$ und alle $p \in \mathbb{R}^n$ folgt die gleichmäßige Konvexität der Zielfunktion f auf der nach Voraussetzung konvexen Niveaumenge L_0, also (K) (b). Denn seien $x, y \in L_0$. Mit einem $t_0 \in (0,1)$ ist (Taylor!)

$$f(y) - f(x) - \nabla f(x)^T(y-x) = \frac{1}{2}(y-x)^T\nabla^2 f(\underbrace{x + t_0(y-x)}_{\in L_0})(y-x) \ge \frac{c}{2}\,\|y-x\|^2,$$

nach Satz 1.5 ist f gleichmäßig konvex mit der Konstanten $c > 0$. Wegen $p^T\nabla^2 f(x)p \le \gamma\,\|p\|^2$ für alle $x \in L_0$ und alle $p \in \mathbb{R}^n$ ist $\|\nabla^2 f(x)\| \le \gamma$. Hieraus folgt die Lipschitzstetigkeit von $\nabla f(\cdot)$ auf L_0 mit der Lipschitzkonstanten $\gamma > 0$ (bezüglich der euklidischen Norm), also (K) (c). Denn für beliebige $x, y \in L_0$ ist

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\| &= \Big\| \int_0^1 \nabla^2 f(\underbrace{x + s(y-x)}_{\in L_0})(x-y)\,ds \Big\| \\
&\le \int_0^1 \underbrace{\|\nabla^2 f(x + s(y-x))\|}_{\le \gamma}\,ds\,\|x-y\| \\
&\le \gamma\,\|x-y\|.
\end{aligned}$$

Zum Nachweis der Konvergenz der Folge $\{x_k\}$ wollen wir Satz 1.7 anwenden. Die Bedingung $(*)$ in Voraussetzung (b) impliziert

$$\delta_k := \min\Big[-\frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|^2}, \Big(\frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|\,\|p_k\|}\Big)^2\Big] \ge \min\Big[\frac{1}{\gamma}, \frac{c}{\gamma}\Big] =: \delta$$

wie man unschwer nachweist. Insbesondere ist

$$\delta \le \frac{1}{k+1}\sum_{j=0}^{k} \delta_j, \qquad k = 0, 1, \ldots.$$

Wegen Satz 1.7 konvergiert die Folge $\{x_k\}$ gegen die eindeutige (globale) Lösung x^* von (P). Um $t_k = 1$ für alle hinreichend großen k nachzuweisen, müssen wir (Wolfe-Schrittweite)

$$f(x_k + p_k) \le f(x_k) + \alpha\nabla f(x_k)^T p_k, \qquad \nabla f(x_k + p_k)^T p_k \ge \beta\nabla f(x_k)^T p_k$$

bzw. (Armijo-Schrittweite)

$$f(x_k + p_k) \le f(x_k) + \alpha\nabla f(x_k)^T p_k$$

für alle hinreichend großen k nachweisen. Es genügt, die Wolfe-Schrittweite zu betrachten und

$$\lim_{k\to\infty} \frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2} > \alpha, \qquad \lim_{k\to\infty} \frac{\nabla f(x_k + p_k)^T p_k}{\nabla f(x_k)^T p_k} = 0 < \beta$$

nachzuweisen. Hier wird also benutzt, dass $\alpha \in (0, \frac{1}{2})$.

Wegen $\|p_k\| \leq \|\nabla f(x_k)\|/c$ konvergiert die Folge $\{p_k\}$ der Newton-Richtungen gegen den Nullvektor. Da außerdem o. B. d. A. x^* im Innern der Niveaumenge L_0 liegt (andernfalls wäre der Startwert x_0 die Lösung und das Verfahren im ersten Schritt abgebrochen) und $\{x_k\}$ gegen x^* konvergiert, liegt die gesamte Verbindungsstrecke zwischen x_k und $x_k + p_k$ für alle hinreichend großen k in L_0. Mit einem $\theta_k \in (0, 1)$ ist daher für diese k wegen des Mittelwertsatzes

$$\begin{aligned}\frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} &= \frac{\nabla f(x_k)^T p_k + \frac{1}{2} p_k^T \nabla^2 f(x_k + \theta_k p_k) p_k}{\nabla f(x_k)^T p_k} \\ &= 1 - \frac{p_k^T \nabla^2 f(x_k + \theta_k p_k) p_k}{2 p_k^T \nabla^2 f(x_k) p_k} \\ &= \frac{1}{2} - \frac{p_k^T [\nabla^2 f(x_k + \theta_k p_k) - \nabla^2 f(x_k)] p_k}{2 p_k^T \nabla^2 f(x_k) p_k}.\end{aligned}$$

Wegen $x_k + \theta_k p_k \to x^*$ und $x_k \to x^*$ ist

$$\frac{|p_k^T [\nabla^2 f(x_k + \theta_k p_k) - \nabla^2 f(x_k)] p_k|}{p_k^T \nabla^2 f(x_k) p_k} \leq \frac{1}{c} \|\nabla^2 f(x_k + \theta_k p_k) - \nabla^2 f(x_k)\| \to 0,$$

womit

$$\lim_{k \to \infty} \frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2} > \alpha$$

und damit

$$f(x_k + p_k) \leq f(x_k) + \alpha \nabla f(x_k)^T p_k$$

für alle hinreichend großen k bewiesen ist. Zum Nachweis der zweiten Beziehung beachte man, dass wiederum wegen des Mittelwertsatzes ein $\eta_k \in (0, 1)$ mit

$$\nabla f(x_k + p_k)^T p_k = \nabla f(x_k)^T p_k + p_k^T \nabla^2 f(x_k + \eta_k p_k) p_k$$

existiert. Folglich ist

$$\begin{aligned}\left| \frac{\nabla f(x_k + p_k)^T p_k}{\nabla f(x_k)^T p_k} \right| &= \frac{|p_k^T [\nabla^2 f(x_k + \eta_k p_k) - \nabla^2 f(x_k)] p_k}{p_k^T \nabla^2 f(x_k) p_k} \\ &\leq \frac{1}{c} \|\nabla^2 f(x_k + \eta_k p_k) - \nabla^2 f(x_k)\| \\ &\to 0.\end{aligned}$$

Insgesamt ist der Satz damit bewiesen. □

Die *Quasi-Newton-Verfahren* versuchen, die Nachteile des Newton-Verfahrens (Berechnung zweiter Ableitungen, kostspieliges Lösen linearer Gleichungssysteme, Richtungen sind nicht notwendig Abstiegsrichtungen) zu vermeiden, ohne die Vorteile (globale Konvergenz durch Einführung von Schrittweiten und automatischer Übergang zum ungedämpften Verfahren bei gleichmäßig konvexer Zielfunktion, lokal superlineare Konvergenz des ungedämpften Verfahrens) aufzugeben. Ein Schritt eines Quasi-Newton-Verfahrens sieht folgendermaßen aus:

- Gegeben $x \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $H \in \mathbb{R}^{n\times n}$. Ferner sei $g := \nabla f(x) \neq 0$ (andernfalls STOP).
- – Berechne Abstiegsrichtung $p := -H^{-1}g$.
  – Berechne Schrittweite $t = t(x, p)$, etwa die exakte Schrittweite, die Wolfe- oder die Armijo-Schrittweite.
  – Berechne neue Näherung $x_+ := x + tp$ und $g_+ := \nabla f(x_+)$.
  – Berechne symmetrische, positiv definite Matrix $H_+ \in \mathbb{R}^{n\times n}$ durch eine sogenannte *Update-Formel*. In die Berechnung von H_+ gehen i. Allg. H sowie $s := x_+ - x$ und $y := g_+ - g$ ein.

Ein Quasi-Newton-Verfahren[10] ist also (neben der Wahl der Schrittweitenstrategie) durch die Update-Formel festgelegt. Wie sollte die neue symmetrische und positiv definite Matrix H_+ berechnet werden? Hierauf gibt es keine eindeutige Antwort. Neben der Symmetrie und positiven Definitheit sollte H_+ der sogenannten *Quasi-Newton-Gleichung*

$$H_+ s = y$$

(gelegentlich auch *Sekantengleichung* genannt) genügen. Als Motivation für die Quasi-Newton-Gleichung geben wir an, dass für hinreichend glattes, etwa zweimal stetig differenzierbares f, die Beziehung

$$\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(y - x))\, dt(y - x)$$

gilt. Es liegt daher nahe, $y = H_+ s$ zu fordern, um H_+ "in die Nähe der Hesseschen zu zwingen".

Eine *notwendige* Bedingung für die Existenz einer symmetrischen, positiv definiten Matrix mit $H_+ s = y$ ist offenbar $y^T s > 0$. Diese Bedingung ist z. B. unter den Bedingungen (K) (a)–(c) aus Lemma 1.6 erfüllt, wenn also insbesondere die Zielfunktion f auf der konvexen Niveaumenge L_0 gleichmäßig konvex (mit einer Konstanten $c > 0$) ist. Denn dann ist

$$y^T s = [\nabla f(x_+) - \nabla f(x)]^T (x_+ - x) \geq c \|x_+ - x\|^2 = c \|s\|^2 > 0.$$

Aber auch ohne die gleichmäßige Konvexität von f ist i. Allg. $y^T s > 0$. Denn ist z. B. $t > 0$ eine Wolfe-Schrittweite, so ist

$$y^T s = t[\nabla f(x_+) - \nabla f(x)]^T p \geq t(\beta - 1)\nabla f(x)^T p > 0$$

mit vorgegebenem $\beta \in (0, 1)$. Ist ferner t die exakte Schrittweite, so ist $\nabla f(x_+)^T p = 0$ und daher $y^T s = -t\nabla f(x)^T p > 0$.

[10] In der Literatur variieren die Bezeichnungen für die bei einem Quasi-Newton-Verfahren auftretenden Matrizen. Sehr häufig (u.a. auch bei J. Werner (1992b)) wird B statt H benutzt. Wir gleichen uns hier C. Geiger, C. Kanzow (1999) an, denn es erscheint mnemotechnisch günstiger zu sein, die auftretenden Abstiegsrichtungen $p = -H^{-1}g$ zu nennen (statt $p = -B^{-1}g$), da H an die **H**essesche erinnert.

Es gibt viele Quasi-Newton-Verfahren, wir wollen uns auf das BFGS-Verfahren beschränken. Die Update-Formel für das BFGS-Verfahren lautet (sie hat es verdient, eingerahmt zu sein):

$$\boxed{H_+ := H - \frac{(Hs)(Hs)^T}{s^T H s} + \frac{yy^T}{y^T s}.}$$

Sie fällt hier vom Himmel, aber es gibt auch kaum einfach erklärbare Gründe, weshalb dies die beste Wahl zu sein scheint. Das folgende Lemma sagt etwas über die Störung einer nichtsingulären Matrix vom Rang 1 aus und ist mit dem Namen Sherman-Morrison verknüpft.

Lemma 1.9 *Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär und $u, v \in \mathbb{R}^n$. Dann gilt:*

1. *Die Matrix*[11] *$A + uv^T$ ist genau dann nichtsingulär, wenn $1 + v^T A^{-1} u \neq 0$.*
2. *Ist $1 + v^T A^{-1} u \neq 0$, so ist*
$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}.$$
3. *Es ist $\det(A + uv^T) = (1 + v^T A^{-1} u)\det(A)$.*

Beweis: Offenbar kann angenommen werden, dass u und v vom Nullvektor verschieden sind. Es ist $(A + uv^T)A^{-1}u = (1 + v^T A^{-1}u)u$ und damit $A + uv^T$ singulär, wenn $1 + v^T A^{-1} u = 0$. Ist dagegen $1 + v^T A^{-1} \neq 0$, so zeigt einfaches Nachrechnen

$$(A + uv^T)\Big(A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}\Big) = I$$

und damit die Gültigkeit der behaupteten Darstellung von $(A + uv^T)^{-1}$. Damit sind die ersten beiden Teile des Lemmas bewiesen.

Für den Beweis des dritten Teiles des Lemmas beachten wir, dass

$$\det(A + uv^T) = \det[A(I + A^{-1}uv^T)] = \det(I + A^{-1}uv^T)\det(A).$$

Zu zeigen bleibt, dass $\det(I + A^{-1}uv^T) = (1 + v^T A^{-1}u$. Hierzu berücksichtigen wir, dass die Determinante einer Matrix das Produkt ihrer Eigenwerte ist. Die Matrix $I + A^{-1}uv^T$ hat 1 als mindestens $(n-1)$-fachen Eigenwert mit $n-1$ linear unabhängigen Eigenvektoren aus dem $(n-1)$-dimensionalen linearem Raum bzw. der Hyperebene $\{x \in \mathbb{R}^n : v^T x = 0\}$. Ist $v^T A^{-1} u \neq 0$, so ist $1 + v^T A^{-1} u$ der verbleibende Eigenwert mit zugehörigem Eigenvektor $A^{-1}u$. Ist dagegen $v^T A^{-1} u = 0$, so ist 1 ein n-facher Eigenwert von $I + A^{-1}uv^T$. □

In dem folgenden Satz wird unter der Voraussetzung $y^T s > 0$ u. a. gezeigt, dass mit H auch der BFGS-Update H_+ symmetrisch und positiv definit ist.

[11] Beachte: uv^T ist eine $n \times n$-Matrix mit $(uv^T)_{ij} = u_i v_j$. Dagegen ist $u^T v = \sum_{j=1}^n u_j v_j$ ein Skalar.

Satz 1.10 *Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ sowie eine symmetrische, positiv definite Matrix $H \in \mathbb{R}^{n\times n}$ gegeben. Durch*

$$H_+ := H - \frac{(Hs)(Hs)^T}{s^T H s} + \frac{yy^T}{y^T s}$$

sei die Update-Matrix des BFGS-Verfahrens definiert. Dann gilt:

1. *Es ist $H_+ s = y$, d. h. H_+ genügt der Quasi-Newton-Gleichung.*
2. *Die Matrix H_+ ist symmetrisch und positiv definit.*
3. *Es ist*
$$\det(H_+) = \frac{y^T s}{s^T H s} \det(H).$$
4. *Es ist*
$$H_+^{-1} = H^{-1} + \Big(1 + \frac{y^T H^{-1} y}{y^T s}\Big)\frac{ss^T}{y^T s} - \frac{s(H^{-1}y)^T + (H^{-1}y)s^T}{y^T s}.$$

Beweis: Die Gültigkeit der Quasi-Newton-Gleichung $H_+ s = y$ ist offensichtlich. Klar ist ferner, dass mit H auch H_+ symmetrisch ist. Da $H \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, besitzt H eine Cholesky[12]-Zerlegung $H = LL^T$ mit einer unteren Dreiecksmatrix L, deren Diagonalelemente positiv sind. Wir zeigen, dass $H_+ = J_+ J_+^T$ mit einer nichtsingulären Matrix J_+, woraus die zweite Behauptung folgt. Hierzu definiere man

$$w := \Big(\frac{y^T s}{s^T H s}\Big)^{1/2} L^T s, \qquad J_+ := L + \frac{(y - Lw)w^T}{w^T w}.$$

Da

$$\sigma := 1 + \frac{w^T L^{-1}(y - Lw)}{w^T w} = \frac{w^T L^{-1} y}{w^T w} = \Big(\frac{y^T s}{s^T H s}\Big)^{1/2} \neq 0,$$

ist J_+ nach Lemma 1.9 nichtsingulär und

$$J_+^{-1} = L^{-1} - \frac{(L^{-1}y - w)w^T L^{-1}}{\sigma w^T w}, \qquad \det(J_+) = \sigma \det(L).$$

Ferner bestätigt man nach leichter Rechnung, dass $J_+ J_+^T = H_+$, womit der zweite Teil des Satzes bewiesen ist.

Es ist

$$\det(H_+) = \det(J_+)^2 = \sigma^2 \det(L)^2 = \frac{y^T s}{s^T H s} \det(H),$$

womit auch der dritte Teil bewiesen ist. Den letzten Teil kann man z. B. durch Nachrechnen beweisen. □

Wir kommen nun zum Beweis der globalen Konvergenz des BFGS-Verfahrens bei gleichmäßig konvexer Zielfunktion. Zur Vorbereitung formulieren und beweisen wir

[12] André-Louis Cholesky (1875–1918) war ein französischer Mathematiker. Cholesky studierte an der École Polytechnique und trat dann in die Armee ein, wo er Vermessungsoffizier wurde. Er starb bei Kämpfen in Nordfrankreich gegen Ende des Ersten Weltkriegs. Posthum wurde seine Methode zur numerischen Lösung von Normalgleichungen bei der Anwendung der Methode der kleinsten Quadrate veröffentlicht. Die dort benutzte Matrixzerlegung wurde ihm zu Ehren Cholesky-Zerlegung benannt.

Lemma 1.11 *Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}^{n\times n}_+ \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen. Man definiere $\psi : \mathcal{S}^{n\times n}_+ \longrightarrow \mathbb{R}$ durch*

$$\psi(A) := \operatorname{tr}(A) - \log\det(A),$$

*wobei $\operatorname{tr}(A)$ die Spur (**tr**ace) der Matrix A bedeutet. Dann gilt:*

1. *Es ist $\psi(A) \geq n$ für alle $A \in \mathcal{S}^{n\times n}_+$. Es gilt Gleichheit genau dann, wenn $A = I$.*
2. *Ist $\{A_k\} \subset \mathcal{S}^{n\times n}_+$, so ist $\{\psi(A_k)\} \subset \mathbb{R}_+$ genau dann (nach oben) beschränkt, wenn $\{\|A_k\|\}$ und $\{\|A_k^{-1}\|\}$ beschränkt sind.*

Beweis: Sind $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ die Eigenwerte der symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$, so ist

$$\psi(A) = \sum_{i=1}^{n} \underbrace{[\lambda_i(A) - \log\lambda_i(A)]}_{\geq 1} \geq n.$$

Gilt hier Gleichheit, so ist $\lambda_i(A) = 1$, $i = 1, \ldots, n$, und folglich $A = I$. Für dieses Argument spielt die Funktion $t - \log t$ auf $\mathbb{R}_+$ eine Rolle, wir veranschaulichen sie uns in Abbildung 2.5 Offenbar ist nämlich $1 \leq t - \log t$ für alle $t > 0$ und es gilt Gleichheit

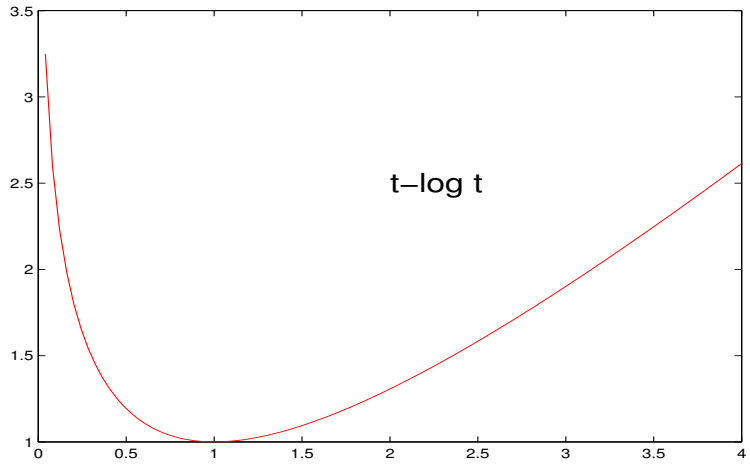


Abbildung 2.5: Die Funktion $t - \log t$

genau dann, wenn $t = 1$. Ist $\{\psi(A_k)\}$ für eine Folge $\{A_k\} \subset \mathcal{S}^{n\times n}_+$ beschränkt, so existieren positive Konstanten c und C mit $c \leq \lambda_{\min}(A_k)$ und $\lambda_{\max}(A_k) \leq C$ für alle k, woraus auch die zweite Behauptung folgt. Damit ist das Lemma bewiesen. □

Es folgt der schon angekündigte globale Konvergenzsatz für das BFGS-Verfahren.

Satz 1.12 *Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzungen (K) (a)–(c) aus Lemma 1.6 seien erfüllt. Man betrachte das durch die exakte, Wolfe- oder Armijo-Schrittweite gedämpfte BFGS-Verfahren:*

- *Gegeben der Startwert $x_0 \in \mathbb{R}^n$, sei $g_0 := \nabla f(x_0)$. Ferner sei eine symmetrische, positiv definite Matrix $H_0 \in \mathbb{R}^{n\times n}$ gegeben.*
- *Für $k = 0, 1, \ldots$:*
  - *Falls $g_k = 0$, dann: STOP, x_k ist Lösung von (P).*
  - *Berechne $p_k := -H_k^{-1} g_k$.*
  - *Sei $t_k > 0$ die exakte Schrittweite, Wolfe-Schrittweite oder Armijo-Schrittweite in x_k in Richtung p_k.*
  - *Sei $x_{k+1} := x_k + t_k p_k$ und berechne $g_{k+1} := \nabla f(x_{k+1})$.*
  - *Mit $s_k := x_{k+1} - x_k$ und $y_k := g_{k+1} - g_k$ sei*

$$H_{k+1} := H_k - \frac{(H_k s_k)(H_k s_k)^T}{s_k^T H_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

Dann gilt: Das Verfahren bricht nach endlich vielen Schritten mit der Lösung x^* von (P) ab oder es liefert eine Folge $\{x_k\}$, die R-linear gegen x^* konvergiert, d. h. es existieren Konstanten $C > 0$ und $q \in (0, 1)$ mit $\|x_k - x^*\| \le C q^k$ für alle k.

Beweis: Seien c und γ die in (K) (b)–(c) auftretenden positiven Konstanten. Die Durchführbarkeit des Verfahrens ist gesichert. Denn ist $g_k \neq 0$ und H_k symmetrisch und positiv definit, so ist $p_k = -H_k^{-1} g_k$ eine Abstiegsrichtung und daher $s_k \neq 0$. Wegen der gleichmäßigen Konvexität der Zielfunktion f ist folglich $y_k^T s_k \ge c \|s_k\|^2 > 0$ und damit auch H_{k+1} positiv definit.

Es wird angenommen, das Verfahren breche nicht schon nach endlich vielen Schritten mit der Lösung ab. Wir wollen Satz 1.7 anwenden und zeigen hierzu die Existenz einer Konstanten $\delta > 0$ mit

$$\delta \le \frac{1}{k+1} \sum_{j=0}^{k} \delta_j, \qquad k = 0, 1, \ldots,$$

wobei

$$\delta_j := \min\Big[-\frac{g_j^T p_j}{\|g_j\|^2}, \Big(\frac{g_j^T p_j}{\|g_j\| \, \|p_j\|} \Big)^2 \Big].$$

Hierzu benutzen wir die durch $\psi(A) := \operatorname{tr}(A) - \log\det(A)$ auf der Menge $\mathcal{S}_+^{n\times n}$ der symmetrischen, positiv definiten $n \times n$-Matrizen in Lemma 1.11 definierte Abbildung. Wegen $\det(H_{k+1}) = (y_k^T s_k / s_k^T H_k s_k) \det(H_k)$ (siehe den dritten Teil von Satz 1.10) erhalten wir

$$\begin{aligned}
\psi(H_{k+1}) &= \psi(H_k) - \frac{\|H_k s_k\|^2}{s_k^T H_k s_k} + \frac{\|y_k\|^2}{y_k^T s_k} - \log \frac{y_k^T s_k}{s_k^T H_k s_k} \\
&= \psi(H_k) + \log\Big(\frac{s_k^T H_k s_k}{\|s_k\| \, \|H_k s_k\|} \Big)^2 + \Big[1 - \frac{\|H_k s_k\|^2}{s_k^T H_k s_k} + \log \frac{\|H_k s_k\|^2}{s_k^T H_k s_k} \Big] \\
&\qquad + \Big[\frac{\|y_k\|^2}{y_k^T s_k} - 1 - \log \frac{y_k^T s_k}{\|s_k\|^2} \Big].
\end{aligned}$$

Man überzeugt sich leicht davon, dass der letzte Term beschränkt ist. Denn es ist

$$\frac{\|y_k\|^2}{y_k^T s_k} \le \frac{\gamma^2 \|s_k\|^2}{y_k^T s_k} \le \frac{\gamma^2}{c}$$

und daher

$$\Big[\frac{\|y_k\|^2}{y_k^T s_k} - 1 - \log \frac{y_k^T s_k}{\|s_k\|^2}\Big] \le \frac{\gamma^2}{c} - 1 - \log c =: \hat{C}.$$

Mit $C := \psi(H_0) + \hat{C}$ ist daher

$$\begin{aligned} n &\le \psi(H_{k+1}) \\ &\le \psi(H_k) + \log\Big(\frac{s_k^T H_k s_k}{\|s_k\| \, \|H_k s_k\|}\Big)^2 + \Big[1 - \frac{\|H_k s_k\|^2}{s_k^T H_k s_k} + \log \frac{\|H_k s_k\|^2}{s_k^T H_k s_k}\Big] + \hat{C} \\ &\le \psi(H_0) + \sum_{j=0}^{k} \Big\{ \log\Big(\frac{s_j^T H_j s_j}{\|s_j\| \, \|H_j s_j\|}\Big)^2 + \Big[1 - \frac{\|H_j s_j\|^2}{s_j^T H_j s_j} + \log \frac{\|H_j s_j\|^2}{s_j^T H_j s_j}\Big] \Big\} + \hat{C}(k+1) \\ &\le C(k+1) + \sum_{j=0}^{k} \Big\{ \underbrace{\log\Big(\frac{s_j^T H_j s_j}{\|s_j\| \, \|H_j s_j\|}\Big)^2}_{\le 0} + \underbrace{\Big[1 - \frac{\|H_j s_j\|^2}{s_j^T H_j s_j} + \log \frac{\|H_j s_j\|^2}{s_j^T H_j s_j}\Big]}_{\le 0} \Big\}. \end{aligned}$$

Folglich ist

$$\sum_{j=0}^{k} \Big\{ \underbrace{\log\Big(\frac{\|s_j\| \, \|H_j s_j\|}{s_j^T H_j s_j}\Big)^2}_{\ge 0} + \underbrace{\Big[\frac{\|H_j s_j\|^2}{s_j^T H_j s_j} - 1 - \log \frac{\|H_j s_j\|^2}{s_j^T H_j s_j}\Big]}_{\ge 0} \Big\} \le C(k+1)$$

für alle k. Nun überlegen wir uns:

- Sind $\alpha_0, \dots, \alpha_k \ge 0$ und $a > 0$ eine Zahl mit $\sum_{j=0}^{k} \alpha_j \le a(k+1)$, so gibt es eine Indexmenge $J_k \subset \{0, \dots, k\}$, die mindestens $\frac{1}{2}(k+1)$ Elemente enthält, und für die $\alpha_j \le 2a$ für alle $j \in J_k$.

Denn: Man definiere $I_k := \{i \in \{0, \dots, k\} : \alpha_i > 2a\}$. Dann ist

$$(k+1)a \ge \sum_{j=0}^{k} \alpha_j \ge \sum_{i \in I_k} \alpha_i > \#(I_k) 2a,$$

so dass I_k weniger als $\frac{1}{2}(k+1)$ Elemente enthält. Dann ist $J_k := \{0, \dots, k\} \setminus I_k$ die gesuchte Indexmenge.

Eine Anwendung dieser Zwischenbehauptung liefert für jedes k die Existenz einer Indexmenge $J_k \subset \{0, \dots, k\}$ mit mindestens $\frac{1}{2}(k+1)$ Elementen und

$$\log\Big(\frac{\|s_j\| \, \|H_j s_j\|}{s_j^T H_j s_j}\Big)^2 + \Big[\frac{\|H_j s_j\|^2}{s_j^T H_j s_j} - 1 - \log \frac{\|H_j s_j\|^2}{s_j^T H_j s_j}\Big] \le 2C \qquad \text{für alle } j \in J_k.$$

Insbesondere erhält man hieraus die Existenz einer von k unabhängigen Konstanten $C_1 > 0$ mit

$$\Big(\frac{\|s_j\|\,\|H_j s_j\|}{s_j^T H_j s_j}\Big)^2 + \frac{\|H_j s_j\|^2}{s_j^T H_j s_j} \le C_1 \qquad \text{für alle } j \in J_k.$$

Nun ist $s_j = t_j p_j = -t_j H_j^{-1} g_j$ und daher

$$\Big(\frac{\|s_j\|\,\|H_j s_j\|}{s_j^T H_j s_j}\Big)^2 = \Big(\frac{\|g_j\|\,\|p_j\|}{g_j^T p_j}\Big)^2, \qquad \frac{\|H_j s_j\|^2}{s_j^T H_j s_j} = -\frac{\|g_j\|^2}{g_j^T p_j}.$$

Folglich ist

$$\delta_j := \min\Big[-\frac{g_j^T p_j}{\|g_j\|^2}, \Big(\frac{g_j^T p_j}{\|g_j\|\,\|p_j\|}\Big)^2\Big] \ge \hat\delta := \frac{1}{C_1} \qquad \text{für alle } j \in J_k.$$

Für alle k ist daher

$$\sum_{j=0}^{k} \delta_j \ge \sum_{j\in J_k} \delta_j \ge \#(J_k)\hat\delta \ge \frac{\hat\delta}{2}(k+1),$$

so dass die Behauptung des Satzes aus dem allgemeinen Konvergenzsatz 1.7 folgt. □

Bemerkung: Es könnten eine Reihe weiterer bemerkenswerter Konvergenzaussagen zum BFGS-Verfahren gemacht werden, insbesondere zur lokalen superlinearen Konvergenz. Aus Zeitgründen verzichten wir darauf und verweisen auf die Vorlesung "Unrestringierte Optimierungsaufgaben" von Jochen Werner aus dem SS 2002 in Göttingen (`http://www.num.math.uni-goettingen/werner/uncopt.pdf`). □

Nun wollen wir einige Bemerkungen zur Implementation des BFGS-Verfahrens machen. Sieht man einmal von der Berechnung der Schrittweite ab, so besteht die Hauptarbeit bei der Durchführung des BFGS-Verfahrens darin, die Richtung $p := -H^{-1}g$ zu berechnen. Hierbei ist natürlich $g = \nabla f(x)$ der Gradient der Zielfunktion in der aktuellen Näherung x und $H \in \mathbb{R}^{n\times n}$ die aktuelle Update-Matrix, von der wir voraussetzen, dass sie symmetrisch und positiv definit ist. Alternativ ist die Matrix $B = H^{-1}$ bekannt. Wir wollen zwei Möglichkeiten zur Implementation besprechen[13]. Bekannt sind stets $s := x_+ - x$ und $y := g_+ - g$, wobei hier möglichst die Wolfe-Schrittweite benutzt wurde, weil diese im Gegensatz zur Armijo-Schrittweite wenigstens theoretisch die wichtige Beziehung $y^T s > 0$ sichert.

Zuerst geben wir die einfachste Methode an. Hier ist die symmetrische, positiv definite Matrix $B = H^{-1} \in \mathbb{R}^{n\times n}$ bekannt, es wird die neue Matrix $B_+ = H_+^{-1}$ durch

$$B_+ := B + \Big(1 + \frac{y^T B y}{y^T s}\Big)\frac{ss^T}{y^T s} - \frac{s(By)^T + (By)s^T}{y^T s}$$

berechnet (siehe letzter Teil von Satz 1.10). Anschließend kann die neue Richtung durch $p_+ := -B_+ g_+$ erhalten werden. Offenbar ist $O(n^2)$ die Anzahl der benötigten arithmetischen Operationen. Der Nachteil dieser Methode besteht darin, dass wir keine Kontrolle

[13] Siehe z. B. J. E. Dennis, R. B. Schnabel (1983, S. 208 ff.), J. Werner (1992b, S. 201 ff.), C. Geiger, C. Kanzow (1999, S. 179 ff.).

darüber haben, ob die neue Matrix wieder positiv definit ist. I. Allg. ist es daher vorzuziehen, eine Cholesky-Faktorisierung von H "upzudaten". Sei also eine Darstellung $H = LL^T$ mit einer unteren Dreiecksmatrix L mit positiven Diagonalelementen bekannt und H_+ das BFGS-Update von H. Gesucht ist eine obere Dreiecksmatrix L_+ mit positiven Diagonalelementen und $H_+ = L_+L_+^T$. Bei dieser Vorgehensweise wird es nicht nötig sein, die Matrizen H und H_+ zu speichern, sondern nur die entsprechenden Cholesky-Faktoren L und L_+.

- Input: Untere Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ mit positiven Diagonalelementen (Cholesky-Faktor von H) sowie $y, s \in \mathbb{R}^n$ mit $y^T s > 0$.
- Berechne
$$u := \sqrt{y^T s}\frac{L^T s}{\|L^T s\|}, \qquad J_+^T := L^T + \frac{u(y - Lu)^T}{y^T s}.$$
Dann ist $H_+ = J_+J_+^T$ (siehe den Beweis zu Satz 1.10).
- Berechne eine QR-Zerlegung $J_+^T = Q_+R_+$, wobei die obere Dreiecksmatrix R_+ positive Diagonalelemente besitzt. Mit $L_+ := R_+^T$ ist dann
$$H_+ = J_+J_+^T = R_+^T Q_+^T Q_+ R_+ = L_+L_+^T$$
die gesuchte Cholesky-Zerlegung von H_+.
- Output: Untere Dreiecksmatrix L_+ mit positiven Diagonalelementen (Cholesky-Faktor von H_+).

Hierbei bleibt noch offen, wie man die QR-Zerlegung einer Matrix $A_+ = R + uv^T$ effizient berechnen kann, wenn R eine obere Dreiecksmatrix mit positiven Diagonalelementen ist. Bei der uns interessierenden Anwendung ist $R = L^T$ und

$$u := \sqrt{y^T s}\frac{L^T s}{\|L^T s\|}, \qquad v := \frac{y - Lu}{y^T s}.$$

Das Ziel erreicht man, indem man A_+ sukzessive von links mit höchstens $2(n-1)$ Givens-Rotationen multipliziert. Sei $m := \max\{i \in \{1,\dots,n\} : u_i \neq 0\}$. Zunächst führt man den Vektor u durch sukzessive Multiplikation mit $m-1$ geeigneten Givens-Rotationen $G_{m-1,m}, \dots, G_{12}$, welche der Reihe nach die Komponenten mit den Indizes $m, \dots, 2$ annullieren, in ein Vielfaches $u_1 e_1$ des ersten Einheitsvektors über. Die parallel hierzu durchgeführte Multiplikation der oberen Dreiecksmatrix R mit den Givens-Rotationen $G_{m-1,m}, \dots, G_{12}$ transformiert diese in eine obere Hessenberg-Matrix, die wir wieder mit R bezeichnen. Wir veranschaulichen uns diesen ersten Schritt im Falle $n = 4$ und $m = 3$, wobei festbleibende Elemente mit $\bullet$, sich verändernde mit $*$ bezeichnet werden.

$$\left(\begin{array}{cccc|c} \bullet & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \end{array}\right) \xrightarrow{G_{23}} \left(\begin{array}{cccc|c} \bullet & \bullet & \bullet & \bullet & \bullet \\ & * & * & * & * \\ & * & * & * & \\ & & & \bullet & \end{array}\right) \xrightarrow{G_{12}} \left(\begin{array}{cccc|c} * & * & * & * & * \\ * & * & * & * & \\ & \bullet & \bullet & \bullet & \\ & & & \bullet & \end{array}\right)$$

Nach Abschluss dieses ersten Schrittes ist $G_{12}\cdots G_{m-1,m}A_+ = R + u_1 e_1 v^T$ mit einer Hessenberg-Matrix R (deren Subdiagonalelemente in den Spalten $m,\ldots,n-1$ verschwinden). In einem Zwischenschritt berechnet man $R := R + u_1 e_1 v^T$, wodurch nur die erste Zeile verändert wird. Durch Multiplikation mit weiteren Givens-Rotationen $G_{12},\ldots,G_{m-1,m}$ annulliert man schließlich im letzten Schritt die störenden Subdiagonalelemente in den Spalten $1,\ldots,m-1$. Hierbei hat man darauf zu achten, dass die erzeugten Diagonalelemente positiv sind. Auch diesen Schritt wollen wir uns für $n=4$, $m=3$ veranschaulichen.

$$\begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet \\ & & & \bullet \end{pmatrix} \xrightarrow{G_{12}} \begin{pmatrix} * & * & * & * \\ & * & * & * \\ & \bullet & \bullet & \bullet \\ & & & \bullet \end{pmatrix} \xrightarrow{G_{23}} \begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ & * & * & * \\ & & * & * \\ & & & \bullet \end{pmatrix}$$

Damit ist ausführlich beschrieben, wie man aus einer Cholesky-Zerlegung der positiv definiten Matrix H eine der BFGS-Update-Matrix H_+ berechnen kann. Die hierzu benötigte Anzahl arithmetischer Operationen ist offenbar im wesentlichen proportional zu n^2. Wir geben nun eine MATLAB-Funktion an, in der genau das eben beschriebene umgesetzt wird:

```
function L_plus=CholBFGS(L,y,s);
%*******************************************************
%Input-Parameter
%       L    Cholesky-Faktor von H
%       y,s    Es sei y^Ts>0
%Output-Parameter
%       L_plus Cholesky-Faktor des BFGS-Update H_plus
%*******************************************************
b=y'*s; v=L'*s; u=sqrt(b)*v/norm(v); v=(1/b)*(y-L*u);
R=L';n=length(y);
m=max(find(u));
for i=m:-1:2
    [G,u(i-1:i)]=planerot(u(i-1:i));
    R(i-1:i,i-1:n)=G*R(i-1:i,i-1:n);
end;
R(1,:)=R(1,:)+u(1)*v';
for i=1:m-1
    [G,R(i:i+1,i)]=planerot(R(i:i+1,i));
    R(i:i+1,i+1:n)=G*R(i:i+1,i+1:n);
end;
L_plus=R';
```

Hierbei haben wir die MATLAB-Funktion `planerot` benutzt. Nach `help planerot` erhält man u. a. die Information:

```
PLANEROT Givens plane rotation.
    [G,Y] = PLANEROT(X), where X is a 2-component column vector,
    returns a 2-by-2 orthogonal matrix G so that Y = G*X has Y(2) = 0.
```

Dies ist keine sogenannte built-in function, sondern sie ist selbst in MATLAB geschrieben. Man kann sie sich mittels `type planerot` oder `edit planerot` ansehen (im

Gegensatz zur built-in function `norm`). Sieht man einmal vom Kommentar ab, den wir oben schon angegeben haben, so sieht diese Funktion folgendermaßen aus:

```
function [G,x] = planerot(x)
if x(2) ~= 0
   r = norm(x);
   G = [x'; -x(2) x(1)]/r;
   x = [r; 0];
else
   G = eye(2);
end
```

Wenn die zweite Komponente x_2 des Vektors $x \in \mathbb{R}^2$ also von Null verschieden ist, so wird die orthogonale Matrix $G \in \mathbb{R}^{2\times 2}$ durch

$$G := \frac{1}{\|x\|} \begin{pmatrix} x_1 & x_2 \\ -x_2 & x_1 \end{pmatrix}$$

definiert ($\|\cdot\|$ ist nach wie vor die euklidische Norm), so dass in diesem Falle

$$Gx = \frac{1}{\|x\|} \begin{pmatrix} x_1 & x_2 \\ -x_2 & x_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \|x\| \\ 0 \end{pmatrix}.$$

Ferner wird (bei einem Aufruf `[G,y]=planerot(x)`) $y := (\|x\|, 0)^T$ gesetzt. Andernfalls ist G gleich der 2×2-Identität (und $y = x$ bzw. x wird nicht verändert).

Man macht sich leicht klar, dass in obiger Funktion `CholBFGS` die Diagonalelemente im Output `L_plus` positiv sind.

Einige wenige Worte wollen wir nun noch zur Wahl der Startmatrix H_0 verlieren. Wenn man gar nichts besseres weiß, so setzt man $H_0 = I$ bzw. $L_0 = I$, so dass am Anfang ein Gradientenschritt durchgeführt wird. Bei J. E. Dennis, R. B. Schnabel (1983, S. 209) wird auch noch die Wahl $H_0 := |f(x_0)|I$ bzw. $L_0 = \sqrt{|f(x_0)|}I$ angegeben. Hiermit ist es nun möglich, ein einfaches function-file `BFGS.m` zu schreiben, wobei wir davon ausgehen, dass die Wolfe-Schrittweite in einer Funktion `Wolfe` realisiert ist. Dies könnte z. B. folgendermaßen aussehen:

```
function [x_min,iter]=BFGS(fun,x_0,max_iter,tol);
%************************************************************
%Input-Parameter:
%            fun        Zu minimierende Funktion. Ein Aufruf
%                       [f,g]=fun(x) liefert Funktionswert
%                       und Gradienten in x
%            x_0        Startvektor
%            max_iter   Maximale Zahl der Iterationen
%            tol        Wenn norm(gradient)<=tol: exit
%Output-Parameter:
%            x_min      (Lokale) Loesung
%            iter       Zahl der Iterationen
%*******************************************************
x=x_0;
[f,g]=feval(fun,x);iter=0;
```

```
L=sqrt(abs(f))*eye(length(x_0));
while (norm(g)>tol)&(iter<max_iter)
     p=-(L'\(L\g)); t=Wolfe(fun,x,p); x_plus=x+t*p;
     [f_plus,g_plus]=feval(fun,x_plus);
     s=x_plus-x; y=g_plus-g;
     L=CholBFGS(L,y,s); g=g_plus; x=x_plus;
     iter=iter+1;
end;
if (norm(g)<=tol)
    x_min=x;
end;
```

Beispiel: Wir betrachten[14] die durch

$$f(x) := (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

definierte Funktion $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, welche vier globale Minima mit dem Funktionswert 0, vier Sattelpunkte und ein lokales Maximum bei $(-0.270845, -0.923039)^T$ besitzt. In Abbildung 2.6 findet man einen Höhenlinienplot von f über $[-0.5, 0.5] \times [-0.5, 0.5]$.

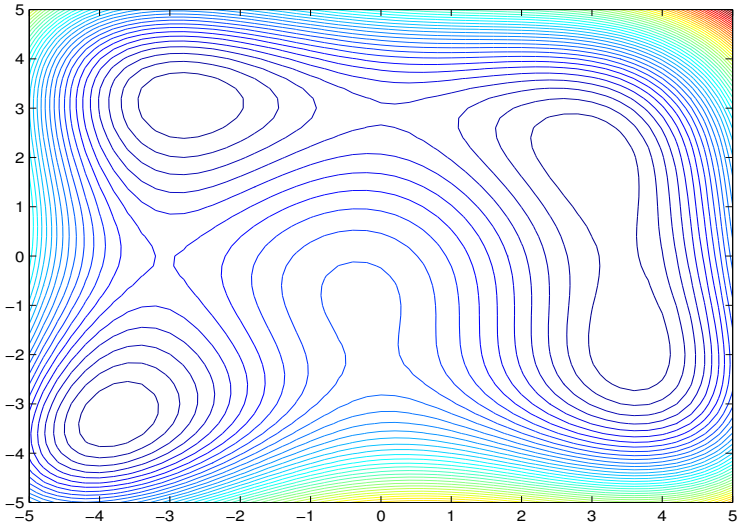


Abbildung 2.6: Höhenlinienplot der Himmelblau-Funktion

Wir schreiben ein function M-file `Himmelblau.m`[15]:

```
function [f,g,H]=Himmelblau(x);
a=x(1)^2+x(2)-11;b=x(1)+x(2)^2-7;
f=a^2+b^2;
if nargout>1
```

[14] Siehe W. Alt (2002, S. 12) und P. Spelluci (1993, S. 91 ff.).

[15] Denn die angegebene Funktion kommt zum ersten Mal bei D. M. Himmelblau (1972, S. 199) vor.

```
    g=[4*x(1)*a+2*b;2*a+4*x(2)*b];
end;
if nargout>2
    H=[4*a+8*x(1)^2+2,4*(x(1)+x(2));4*(x(1)+x(2)),4*b+8*x(2)^2+2];
end;
```

Wir wollen nur einige wenige numerische Experimente machen. Mit dem Aufruf

```
[x,iter]=Unged_Newton('Himmelblau',[3;0],100,1e-8)
```

erhalten wir

$$x = \begin{pmatrix} 3.385154183610126 \\ 0.073851879838867 \end{pmatrix}, \qquad \text{iter} = 4.$$

Hier ist x Sattelpunkt von f. Macht man dagegen den Aufruf

```
[x,iter]=BFGS('Himmelblau',[3;0],100,1e-8)
```

so ist

$$x = \begin{pmatrix} 3.584428340327278 \\ -1.848126526985198 \end{pmatrix}, \qquad \text{iter} = 10.$$

Hier ist x eines der globalen Minima von f. Nehmen wir noch einen anderen Startwert. Nach

```
[x,iter]=Unged_Newton('Himmelblau',[0;0],100,1e-8)
```

erhalten wir

$$x = \begin{pmatrix} -0.270844590678330 \\ -0.923038556403508 \end{pmatrix}, \qquad \text{iter} = 4,$$

das ist gerade das lokale Maximum von f. Mit

```
[x,iter]=BFGS('Himmelblau',[0;0],100,1e-8)
```

haben wir dagegen

$$x = \begin{pmatrix} 2.999999999999658 \\ 2.000000000001913 \end{pmatrix}, \qquad \text{iter} = 11.$$

Dies ist auch eines der globalen Minima (genauer: In $(3,2)^T$ ist ein globales Minimum von f). Der Aufruf

```
x=fminunc('Himmelblau',[0;0])
```

mit der Funktion `fminunc` aus der Optimization-Toolbox von MATLAB ergibt

$$x = \begin{pmatrix} 2.999999930595906 \\ 2.000000008784562 \end{pmatrix}.$$

Mit dem gedämpften Newton-Verfahren haben wir jeweils keinen Erfolg, da die Hessesche im Startwert indefinit bzw. negativ definit ist. □

2.1.4 Verfahren der konjugierten Gradienten

Quasi-Newton-Verfahren haben den Nachteil, dass sie Speicherplatz für eine Approximation $H \in \mathbb{R}^{n\times n}$ der Hesseschen $\nabla^2 f(x)$ der Zielfunktion f in der aktuellen Näherung x benötigen. Dies kann für großes n ein Problem werden. Weniger Speicherplatz benötigen die jetzt zu besprechenden Verfahren der konjugierten Gradienten. Wie von J. Nocedal, S. J. Wright (1999, S. 101) zu Beginn ihres Kapitels über *Conjugate Gradient Methods* ausgeführt wird, besteht ein zweifaches Interesse an Verfahren der konjugierten Gradienten. Einmal gehören diese Verfahren zu den nützlichsten Techniken zur Lösung großer linearer Gleichungssysteme mit einer symmetrischen, positiv definiten Koeffizientenmatrix bzw. der unrestringierten Minimierung einer quadratischen, gleichmäßig konvexen Zielfunktion. Zum anderen können die Verfahren zur Lösung allgemeiner unrestringierter Optimierungsaufgaben adaptiert werden. Wesentlich ist hierbei, dass keine Matrix gespeichert werden muss und die Verfahren (es gibt etliche Varianten) schneller als das Gradientenverfahren bzw. das Verfahren des steilsten Abstiegs sind. Hingewiesen sei aber darauf, dass die Wahl einer richtigen Schrittweitenstrategie nicht ganz einfach ist.

Ausgangspunkt für Verfahren der konjugierten Gradienten bei unrestringierten Optimierungsaufgaben ist das auf Hestenes-Stiefel (1952) zurückgehende CG-Verfahren (Conjugate Gradient Verfahren) für lineare Gleichungssysteme mit symmetrischer, positiv definiter Koeffizientenmatrix. Dieses Verfahren wollen wir nur sehr kurz schildern, da seine Analyse eher in eine Vorlesung über Numerische Lineare Algebra gehört, um danach auf Verfahren der konjugierten Gradienten für unrestringierte Optimierungsverfahren mit nicht notwendig quadratischer, gleichmäßig konvexer Zielfunktion einzugehen.

Man betrachte die Aufgabe

$$\text{Minimiere} \quad f(x) := \tfrac{1}{2}x^T Ax - b^T x, \quad x \in \mathbb{R}^n,$$

wobei $A \in \mathbb{R}^{n\times n}$ als symmetrisch und positiv definit vorausgesetzt wird und $b \in \mathbb{R}^n$. Diese Aufgabe besitzt genau eine Lösung x^*, nämlich die Lösung des linearen Gleichungssystems $Ax = b$. Zur Lösung betrachte man das folgende Verfahren, wobei nach wie vor $\|\cdot\|$ die euklidische Norm bezeichne.

- Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := Ax_0 - b$ und setze $p_0 := -g_0$.
- Für $k = 0, 1, \ldots$:
  - Falls $g_k = 0$, dann: $m := k$, STOP. x_m ist die gesuchte Lösung des linearen Gleichungssystems $Ax = b$.
  - Andernfalls:
    - Berechne

$$t_k := -\frac{g_k^T p_k}{p_k^T A p_k}, \qquad x_{k+1} := x_k + t_k p_k, \qquad g_{k+1} := g_k + t_k A p_k.$$

- Berechne

$$\beta_k := \frac{\|g_{k+1}\|^2}{\|g_k\|^2}, \qquad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

Über dieses Verfahren kann u. a. ausgesagt werden, dass es nach $m \leq n$ Schritten mit der gesuchten Lösung abbricht, die Gradienten paarweise aufeinander senkrecht stehen (d. h. $g_i^T g_k = 0$ für $0 \leq i < k \leq m$) und die erzeugten Richtungen A-konjugiert sind (d. h. $p_i^T A p_k = 0$ für $0 \leq i < k < m$, ferner sind $p_0, \ldots, p_{m-1}$ vom Nullvektor verschieden). Man beachte, dass die Schrittweite t_k die *exakte* Schrittweite ist. Durch eine sogenannte *Präkonditionierung* kann man zu einem äquivalenten Problem übergehen, bei welchem die Koeffizientenmatrix eine kleinere Kondition besitzt. In dieser Form wird das CG-Verfahren bei hochdimensionalen, schwach besetzten linearen Gleichungssystemen mit symmetrischer, positiv definiter Koeffizienmatrix angewandt. Hierauf wollen wir nicht eingehen, siehe z. B. das Skript "Unrestringierte Optimierungsaufgaben" (2002, S. 103 ff.) von J. Werner.

Von R. Fletcher, C. M. Reeves (1964) stammt eine erste Verallgemeinerung des Verfahrens der konjugierten Gradienten auf unrestringierte Optimierungsaufgaben

$$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n,$$

mit nicht notwendig quadratischer Zielfunktion f. Wir beschränken uns im wesentlichen auf die Beschreibung dieses Verfahrens, Varianten werden in den Aufgaben angesprochen. Wir erinnern an die Voraussetzungen (V) (a)–(c) aus Unterabschnitt 2.1.1 und die (gleichmäßigen) Konvexitätsvoraussetzungen (K) (a)–(c) aus Lemma 1.6. Im folgenden Satz wird das Fletcher-Reeves-Verfahren mit exakter Schrittweitenstrategie angegeben und unter den Voraussetzungen (V) (a)–(c) bzw. (K) (a)–(c) eine globale Konvergenzaussage bewiesen.

Satz 1.13 *Gegeben sei die Optimierungsaufgabe (P). Die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ genüge den Voraussetzungen (V) (a)–(c). Das Verfahren der konjugierten Gradienten von Fletcher-Reeves ist durch den folgenden Algorithmus gegeben:*

- *Gegeben $x_0 \in \mathbb{R}^n$, berechne $g_0 := \nabla f(x_0)$ und setze $p_0 := -g_0$.*
- *Für $k = 0, 1, \ldots$:*
  - *Falls $g_k = 0$, dann: STOP. x_k ist stationäre Lösung von (P).*
  - *Andernfalls:*
    - *Berechne die exakte Schrittweite $t_k := t^*(x_k, p_k)$.*
    - *Berechne*

$$x_{k+1} := x_k + t_k p_k, \qquad g_{k+1} := \nabla f(x_{k+1})$$

sowie

$$\beta_k := \frac{\|g_{k+1}\|^2}{\|g_k\|^2}, \qquad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

Dann gilt: Bricht das Verfahren nicht nach endlich vielen Schritten mit einer stationären Lösung von (P) ab, so liefert es eine Folge $\{x_k\}$ mit $\liminf_{k\to\infty} \|g_k\| = 0$, wenigstens ein Häufungspunkt von $\{x_k\}$ ist also eine stationäre Lösung von (P). Sind sogar die Konvexitätsvoraussetzungen (K) (a)–(c) erfüllt, so konvergiert die gesamte Folge $\{x_k\}$ gegen die dann eindeutige Lösung x^* von (P).

Beweis: Da im Verfahren stets die exakte Schrittweite gewählt wird, ist $g_k^T p_k = -\|g_k\|^2 < 0$ für $g_k \neq 0$, also p_k eine Abstiegsrichtung in x_k. Das Verfahren breche nicht vorzeitig mit einer stationären Lösung ab. Im Widerspruch zur Behauptung nehmen wir an, es sei $\liminf_{k\to\infty} \|g_k\| > 0$. Dann existiert ein $\epsilon > 0$ mit $\|g_k\| \geq \epsilon$ für alle k. Es gibt (siehe Bemerkung zu Beginn von Unterabschnitt 2.1.1) eine von k unabhängige Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta\Big(\frac{g_k^T p_k}{\|p_k\|}\Big)^2 = \theta\,\frac{\|g_k\|^4}{\|p_k\|^2} = \frac{\theta}{\alpha_k} \qquad \text{mit} \quad \alpha_k := \frac{\|p_k\|^2}{\|g_k\|^4}.$$

Für $k \geq 1$ ist

$$\alpha_k = \frac{\|p_k\|^2}{\|g_k\|^4} = \frac{\|g_k\|^2 + \beta_{k-1}^2\,\|p_{k-1}\|^2}{\|g_k\|^4} = \frac{1}{\|g_k\|^2} + \alpha_{k-1}.$$

Durch Zurückspulen erhält man

$$\alpha_k = \sum_{j=1}^{k} \frac{1}{\|g_j\|^2} + \alpha_0 = \sum_{j=0}^{k} \frac{1}{\|g_j\|^2} \leq \frac{k+1}{\epsilon^2}, \qquad k = 0, 1, \ldots,$$

und hieraus

$$(*) \qquad f(x_k) - f(x_{k+1}) \geq \frac{\theta\epsilon^2}{k+1}, \qquad k = 0, 1, \ldots.$$

Die harmonische Reihe ist bekanntlich divergent. Daher folgt aus $(*)$, dass $\{f(x_k)\}$ nicht nach unten beschränkt ist, was einen Widerspruch zur vorausgesetzten Kompaktheit der Niveaumenge L_0 darstellt.

Nun seien sogar die Voraussetzungen (K) (a)–(c) erfüllt. Ein $\epsilon > 0$ sei vorgegeben. Wegen der unter den schwächeren Voraussetzungen (V) (a)–(c) bewiesenen Aussage existiert ein $k_0 \in \mathbb{N}$ mit $\|g_{k_0}\| \leq c\epsilon$. Eine Anwendung von Lemma 1.6 liefert für alle $k \geq k_0$ die Ungleichungskette

$$\frac{c}{2}\|x_k - x^*\|^2 \leq f(x_k) - f(x^*) \leq f(x_{k_0}) - f(x^*) \leq \frac{1}{2c}\|g_{k_0}\|^2 \leq \frac{c\epsilon^2}{2}.$$

Daher ist $\|x_k - x^*\| \leq \epsilon$ für alle $k \geq k_0$, so dass auch der zweite Teil des Satzes bewiesen ist. □

Bemerkungen: Spezialisiert man das Fletcher-Reeves-Verfahren auf eine quadratische Zielfunktion, so erhält man genau das oben angegebene CG-Verfahren von Hestenes-Stiefel. Insbesondere bricht das Verfahren in diesem Fall nach $m \leq n$ Schritten ab.

Es gibt einige Variantem zum Fletcher-Reeves-Verfahren. Diese unterscheiden sich im wesentlichen in der Definition des Skalars β_k, reduzieren sich für eine quadratische Zielfunktion aber stets auf das CG-Verfahren. Beim Polak-Ribière-Verfahren wird z. B.

$$\beta_k := \frac{g_{k+1}^T(g_{k+1} - g_k)}{\|g_k\|^2}$$

gesetzt, siehe z. B. C. Geiger, C. Kanzow (1999, S. 231 ff.).

I. Allg. macht man in einem Verfahren der konjugierten Gradienten alle n Schritte einen sogenannten *restart*, indem man wieder mit der negativen Gradientenrichtung in der aktuellen Näherung beginnt. □

2.1.5 Aufgaben

1. Eine Funktion $\phi\colon [a,b] \longrightarrow \mathbb{R}$ heißt *unimodal*, wenn es genau ein $t^* \in (a,b)$ gibt mit $\phi(t^*) = \min_{t\in[a,b]} \phi(t)$, und wenn ϕ auf $[a,t^*]$ monoton fallend und auf $[t^*,b]$ monoton wachsend ist. Zur Lokalisierung des Minimums t^* der auf $[a,b]$ unimodularen Funktion ϕ betrachte man die *Methode vom goldenen Schnitt*:

   - Sei $\epsilon > 0$ (gewünschte Genauigkeit) gegeben, setze $F := (\sqrt{5}-1)/2$.
   - Berechne $\begin{cases} s := a + (1-F)(b-a), & \phi_s := \phi(s), \\ t := a + F(b-a), & \phi_t := \phi(t). \end{cases}$
   - Solange $b - a > \epsilon$:
     - Falls $\phi_s > \phi_t$, dann:

       $$a := s, \quad s := t, \quad t := a + F(b-a), \quad \phi_s := \phi_t, \quad \phi_t := \phi(t)$$

     - Andernfalls:

       $$b := t, \quad t := s, \quad s := a + (1-F)(b-a), \quad \phi_t := \phi_s, \quad \phi_s := \phi(s).$$

   - $t^* \approx (a+b)/2$.

   Man beweise, dass dieser Algorithmus nach endlich vielen Schritten mit einem Intervall $[a,b]$ abbricht, das t^* enthält.

2. Man gebe eine MATLAB-Implementation der Methode des goldenen Schnittes an und erprobe sie an den Funktionen[16]

   (a) $\phi(t) := -t/(t^2+c)$ mit $c := 2$,

   (b) $\phi(t) := (t+c)^5 - 2(t+c)^4$ mit $c := 0.004$.

   Hierbei veranschauliche man sich die Funktionen durch einen Plot.

3. Sei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ definiert durch $f(x) := c^T x + \frac{1}{2}x^T Q x$ mit $c \in \mathbb{R}^n$ und symmetrischer, positiv definiter Matrix $Q \in \mathbb{R}^{n\times n}$. Sei $x \in \mathbb{R}^n$ kein stationärer Punkt von f und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x.

[16] Siehe C. Geiger, C. Kanzow (1999, S. 52).

(a) Man berechne die exakte Schrittweite t^* und eine (von x und p unabhängige, möglichst große) Konstante $\theta > 0$ mit

$$f(x) - f(x + t^* p) \geq \theta \Big(\frac{\nabla f(x)^T p}{\|p\|} \Big)^2 .$$

(b) Zu $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ berechne man die Menge der Wolfe-Schrittweiten und zeige, dass diese ein nichtleeres Intervall ist.

4. Sei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ wie in Aufgabe 3.

(a) Zur Minimierung von f auf dem $\mathbb{R}^n$ betrachte man das Gradientenverfahren mit exakter Schrittweite. Ausgehend von einem Startwert $x_0 \in \mathbb{R}^n$ erzeuge dieses die Folge $\{x_k\}$. Man begründe, weshalb die Folge $\{x_k\}$ gegen das eindeutige Minimum x^* von f konvergiert.

(b) Man zeige, dass mit $g_k := \nabla f(x_k)$ die Beziehung

$$\frac{f(x_k) - f(x^*)}{f(x_k) - f(x_{k+1})} = \frac{(g_k^T Q g_k)(g_k^T Q^{-1} g_k)}{\|g_k\|^4}$$

gilt (hierbei sei $\|\cdot\|$ weiterhin die euklidische Norm). Hieraus schließe man, dass

$$f(x_{k+1}) - f(x^*) \leq \Big(1 - \frac{1}{\kappa(Q)} \Big) [f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots .$$

Hierbei ist $\kappa(Q) := \|Q\| \, \|Q^{-1}\| = \lambda_{\max}(Q)/\lambda_{\min}(Q)$ die *Kondition* von Q bezüglich der Spektralnorm.

(c) Man informiere sich (etwa bei C. Geiger, C. Kanzow (1999, S. 70 ff.)), wie die letzte Abschätzung mit Hilfe der Ungleichung von Kantorowitsch verbessert werden kann.

5. Die Voraussetzungen von Satz 1.3 (erster Konvergenzsatz für den Modellalgorithmus) seien erfüllt. Die Zielfunktion f besitze in der (kompakten) Niveaumenge L_0 nur endlich viele stationäre Punkte. Der Modellalgorithmus erzeuge eine Folge $\{x_k\} \subset L_0$ mit $\lim_{k\to\infty}(x_{k+1} - x_k) = 0$. Man zeige, dass dann die gesamte Folge $\{x_k\}$ gegen einen der stationären Punkte von f konvergiert.

Hinweis: Siehe J. M. Ortega, W. C. Rheinboldt (1970, S. 476). Man überlege sich, dass auch die Menge der Häufungspunkte von $\{x_k\}$ endlich ist. Wenn es genau einen Häufungspunkt gibt, so ist die Behauptung richtig. Die Annahme, dass es mehr als einen Häufungspunkt gibt, führe man zum Widerspruch.

6. Seien $y, s \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $H \in \mathbb{R}^{n\times n}$ gegeben. Es sei $(y - Hs)^T s \neq 0$. Man bestimme $\gamma \in \mathbb{R}$ und $u \in \mathbb{R}^n$ so, dass die Matrix $H_+ = H + \gamma u u^T$ der Quasi-Newton-Gleichung $H_+ s = y$ genügt. Unter welchen Voraussetzungen an H, y und s ist die so bestimmte Matrix H_+ positiv definit?

Hinweis: Zu der symmetrischen, positiv definiten Matrix $H \in \mathbb{R}^{n\times n}$ gibt es eine symmetrische, positiv definite Matrix $H^{1/2}$ mit $H^{1/2}H^{1/2} = H$ (siehe z. B. C. Geiger, C. Kanzow (1999, S. 331)). Man benutze, dass die Matrix H_+ genau dann positiv definit ist, wenn $H^{-1/2}H_+H^{-1/2}$ es ist, wobei $H^{-1/2} := (H^{1/2})^{-1}$.

7. Man gebe eine MATLAB-Implementation des CG-Verfahrens von Hestenes-Stiefel für lineare Gleichungssysteme mit symmetrischer, positiv definiter Koeffizientenmatrix an und löse hiermit die Aufgabe, die Funktion

$$f(x) := x_1^2 + 0.3x_1x_2 + 0.975x_2^2 + 0.01x_1x_3 + x_3^2 + 3x_1 - 4x_2 + x_3$$

auf dem $\mathbb{R}^3$ zu minimieren[17].

8. Gegeben sei die quadratische Zielfunktion $f(x) := \frac{1}{2}x^TAx - b^Tx$ mit einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$. Seien $p_0, \ldots, p_{n-1} \in \mathbb{R}^n$ konjugiert bezüglich A, d. h. $p_0, \ldots, p_{n-1}$ sind vom Nullvektor verschieden und es ist $p_i^TAp_j = 0$, $0 \le i < j \le n-1$. Man betrachte das folgende Verfahren zur Minimierung von $f(x)$ auf dem $\mathbb{R}^n$:

   - Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := Ax_0 - b$.
   - Für $k = 0, 1, \ldots$:
     - Falls $g_k = 0$, dann: $m := k$, f nimmt in x_m das Minimum an. STOP.
     - Andernfalls berechne

$$t_k := -\frac{g_k^Tp_k}{p_k^TAp_k}, \qquad x_{k+1} := x_k + t_kp_k, \qquad g_{k+1} := g_k + t_kAp_k.$$

   Durch vollständige Induktion nach k zeige man: Sind $g_0, \ldots, g_k \neq 0$, ist das Verfahren im k-ten Schritt also noch nicht abgebrochen, so ist x_{k+1} die Lösung der Aufgabe

$$(\mathrm{P}_k) \qquad \text{Minimiere} \quad f(x), \quad x \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}.$$

   Wegen $x_0 + \operatorname{span}\{p_0, \ldots, p_{n-1}\} = \mathbb{R}^n$ bricht das Verfahren also nach $m \le n$ Schritten mit dem Minimum von f ab.

   Hinweis: Nach Konstruktion ist klar, dass $x_{k+1} \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$. Man zeige, dass $g_{k+1}^Tp_i = 0$, $i = 0, \ldots, k$, und überlege sich, dass dies die Behauptung impliziert.

9. Das Verfahren der konjugierten Gradienten von Polak-Ribière unterscheidet sich von dem Fletcher-Reeves-Verfahren nur darin, dass

$$\beta_k := \frac{g_{k+1}^T(g_{k+1} - g_k)}{\|g_k\|^2}$$

[17] Siehe P. Spellucci (1993, S. 164).

(statt $\beta_k := \|g_{k+1}\|^2/\|g_k\|^2$) gesetzt wird. Man betrachte das dann definierte Polak-Ribière-Verfahren mit exakter Schrittweitenstrategie zur Lösung von

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

Man zeige: Sind die (gleichmäßigen) Konvexitätsvoraussetzungen (K) (a)–(c) aus Lemma 1.6 erfüllt, so liefert das Verfahren, wenn es nicht vorzeitig mit der Lösung x^* von (P) abbricht, eine Folge $\{x_k\}$, die R-linear gegen x^* konvergiert.

Hinweis: Man setze

$$\delta_k := \Big(\frac{g_k^T p_k}{\|g_k\|\,\|p_k\|}\Big)^2 = \frac{\|g_k\|^2}{\|p_k\|^2},$$

zeige $\|p_{k+1}\| \le (1+\gamma/c)\|g_{k+1}\|$, $k = 0, 1, \ldots$, folgere hieraus auf die Existenz einer Konstanten $\delta > 0$ mit $\delta_k \ge \delta$, $k = 0, 1, \ldots$, und wende Satz 1.7 an (siehe auch J. Werner (1992b, S. 234)).

2.2 Trust-Region-Verfahren

Wesentlich kürzer als im letzten Abschnitt über Schrittweitenverfahren gehen wir in diesem Abschnitt auf Trust-Region-Verfahren ein. Als Lehrbuch, welches bisher noch nicht genannt wurde, sei hier vor allem auf A. R. Conn, N. I. M. Gould, Ph. L. Toint (2000) hingewiesen. In diesem Abschnitt verstehen wir unter $\|\cdot\|$ eine beliebige Norm auf dem $\mathbb{R}^n$ bzw. die zugeordnete Matrixnorm. Die euklidische Norm bzw. die Spektralnorm als zugeordnete Matrixnorm werden wir mit $\|\cdot\|_2$ bezeichnen. Entsprechendes gilt für andere spezielle Normen. Wir werden uns auf Optimierungsaufgaben mit einer *glatten* Zielfunktion beschränken. Trust-Region-Verfahren bei diskreten nichtlinearen Approximationsaufgaben werden z. B. bei J. Werner (1992b, S. 249 ff.) behandelt.

2.2.1 Ein Modellalgorithmus

Gegeben sei wieder die unrestringierte Optimierungsaufgabe

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

Die Idee bei den Trust-Region besteht darin, die Zielfunktion f lokal auf einer Kugel (bezüglich einer geeigneten Norm) um eine aktuelle Näherung durch ein einfacheres "Modell" zu ersetzen, etwa einer linearen oder quadratischen Approximation der Zielfunktion. Dann bestimmt man ein Minimum (oder auch nur eine Approximation an das Minimum) des Modells bzw. der vereinfachten Zielfunktion auf der Kugel. Wird eine Verminderung des Zielfunktionswertes entweder nicht erreicht, oder ist diese enttäuschend gering, so hat man dem Modell auf einer zu großen Kugel um die aktuelle Näherung "vertraut", diese wird daher verkleinert und auf dieser kleineren Kugel erneut ein Minimum der Modellfunktion bestimmt. Andernfalls wird dieses Minimum als neue aktuelle Näherung akzeptiert und der Radius der Kugel vergrößert, wenn ein verschärfter Test auf hinreichende Verminderung erfolgreich bestanden wird. Das ist, sehr lax ausgedrückt, die Idee der Trust-Region-Verfahren.

Nun soll diese Idee etwas genauer gefasst werden. Bei gegebener aktueller Näherung $x \in \mathbb{R}^n$ sei ein "einfaches Modell" $f_x : \mathbb{R}^n \longrightarrow \mathbb{R}$ für die i. Allg. komplizierte Funktion $p \mapsto f(x+p)$ gegeben. Eine Minimalforderung an die Modellfunktion f_x wird $f_x(0) = f(x)$ sein. Ist z. B. f in x zweimal stetig differenzierbar, so liegt es nahe,

$$f_x(p) := f(x) + \nabla f(x)^T p + \tfrac{1}{2} p^T \nabla^2 f(x) p$$

zu setzen, wobei $\nabla^2 f(x)$ durch eine symmetrische (nicht notwendig positiv definite) Matrix $H \in \mathbb{R}^{n\times n}$ ersetzt sein kann. Ist dagegen $f(x) := \frac{1}{2}\|F(x)\|_2^2$ mit einer stetig differenzierbaren Abbildung $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, so könnte die Modellfunktion f_x durch

$$\begin{aligned} f_x(p) &:= \tfrac{1}{2}\|F(x) + F'(x)p\|_2^2 \\ &= \tfrac{1}{2}\|F(x)\|_2^2 + (F'(x)^T F(x))^T p + \tfrac{1}{2} p^T F'(x)^T F'(x) p \end{aligned}$$

gegeben sein.

Ein Schritt eines Modellalgorithmus für Trust-Region-Verfahren könnte dann folgendermaßen aussehen:

- Unabhängig vom aktuellen Iterationsschritt seien Konstanten $0 < \rho_1 < \rho_2 < 1$, $\sigma_1 \in (0,1)$ und $\sigma_2 > 1$ gegeben. Z. B. sei $\rho_1 := 0.01$, $\rho_2 := 0.9$, $\sigma_1 := 0.5$ und $\sigma_2 := 2$.

- Gegeben sei ein aktuelles Paar (x, Δ), wobei x eine Näherung für eine (stationäre, lokale, globale) Lösung von (P) ist und $\Delta > 0$ der Radius einer Kugel um 0 bzw. x ist, die bezüglich einer Norm $\|\cdot\|$ zu verstehen ist. Diese Norm ist gewöhnlich die euklidische Norm oder die Maximumnorm[18].

- Bestimme eine globale Lösung $p^* \in \mathbb{R}^n$ der Aufgabe

  $$(\mathrm{P}_x) \qquad \text{Minimiere} \quad f_x(p), \quad \|p\| \le \Delta.$$

  Natürlich sollte dieses Trust-Region-Hilfsproblem "einfach" lösbar sein.

- Falls $f(x) = f_x(p^*)$ (dies ist genau dann der Fall, wenn 0 eine Lösung von (P_x) ist), dann: STOP.

  Ist die Modellfunktion richtig gewählt, so wird x in diesem Falle wenigstens eine stationäre Lösung von (P) sein.

- Andernfalls berechne

  $$r := \frac{f(x) - f(x+p^*)}{f(x) - f_x(p^*)}.$$

  Falls $r \ge \rho_1$, dann setze $x_+ := x + p^*$ als neue Näherung und bezeichne den Iterationsschritt als *erfolgreich*. Andernfalls setze $x_+ := x$.

  - Falls $r < \rho_1$, dann wähle $\Delta_+ \in (0, \sigma_1\Delta]$, z. B. $\Delta_+ := \sigma_1\Delta$.

[18] Auf den Fall, dass die Norm noch vom aktuellen Iterationschritt abhängig ist, wollen wir nicht eingehen.

- Falls $r \in [\rho_1, \rho_2)$, dann wähle $\Delta_+ \in [\sigma_1\Delta, \Delta]$, z. B. $\Delta_+ := \Delta$.
- Falls $r \geq \rho_2$, dann wähle $\Delta_+ \in [\Delta, \sigma_2\Delta]$, z. B. $\Delta_+ := \sigma_2\Delta$.

Einige Bemerkungen zu den Tests im letzten Schritt sind angebracht. In ihm wird angenommen, dass $f(x) \neq f_x(p^*)$ (andernfalls wäre ein Ausstieg im vorherigen Schritt erfolgt). Dann ist aber $f(x) > f_x(p^*)$, da von der Modellfunktion $f_x(0) = f(x)$ angenommen wurde. Entscheidend für den Test ist die Größe r, der Quotient aus der tatächlichen und der durch das Modell vorhergesagten Verminderung des Zielfunktionswertes. Je näher r bei 1 liegt, desto besser "stimmt" das Modell.

Ist $r < \rho_1$, so hat sich keine Verminderung eingestellt oder diese ist, verglichen mit der vorhergesagten, zu gering. Das wird darauf zurückgeführt, dass dem Modell auf einer zu großen Kugel vertraut wurde. Diese wird daher entsprechend verkleinert und mit derselben aktuellen Näherung ein erneuter Versuch unternommen.

Ist sogar der verschärfte Test $r \geq \rho_2$ erfolgreich, so stimmen die tatsächliche und die vorhergesagte Verminderung hinreichend gut überein, so dass im nächsten Schritt dem Modell auf einer i. Allg. größeren Kugel vertraut wird. Für $r \in [\rho_1, \rho_2)$ ist man mit der neuen Näherung zufrieden, vergrößert den Bereich aber nicht. In beiden Fällen ist man aber erfolgreich und berechnet eine neue aktuelle Näherung.

2.2.2 Das Trust-Region-Hilfsproblem

Die wesentliche Arbeit im Modellalgorithmus für Trust-Region-Verfahren erfolgt bei der Lösung des Hilfsproblems, das (in unserem Falle: quadratische) Modell auf einer Kugel zu minimieren. Daher beschäftigen wir uns in diesem Unterabschnitt mit der numerischen Lösung des Problems

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \tfrac{1}{2} p^T H p, \quad \|p\|_2 \leq \Delta,$$

wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^n$, die symmetrische Matrix $H \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben sind.

Ganz entscheidend für die numerische Lösung von (P) ist, dass man eine globale Lösung von (P) charakterisieren kann, also notwendige und hinreichende Bedingungen dafür angeben kann, dass ein $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ eine globale Lösung von (P) ist. Da wir *nicht* voraussetzen, dass H positiv semidefinit ist, ist dies ein bemerkenswertes Ergebnis.

Satz 2.1 *Genau dann ist ein $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ eine globale Lösung von*

$$\text{(P)} \qquad \textit{Minimiere} \quad \phi(p) := f + g^T p + \tfrac{1}{2} p^T H p, \quad \|p\|_2 \leq \Delta,$$

wenn ein $\lambda^* \geq 0$ mit

(a) $(H + \lambda^* I)p^* = -g$,

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

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

existiert. Darüberhinaus ist p^* eindeutige globale Lösung von (P), wenn $H + \lambda^* I$ sogar positiv definit ist. Weiter gilt:

1. *Es ist $\phi(p^*) = f$ genau dann, wenn $g = 0$ und H positiv semidefinit ist.*
2. *Es ist*
$$f - \phi(p^*) \geq \frac{1}{2}\|g\|_2 \min\Big(\Delta, \frac{\|g\|_2}{\|H\|_2}\Big).$$

Beweis: Wir zeigen nur[19] die einfache Richtung, dass nämlich die Existenz eines $\lambda^* \geq 0$ mit (a)–(c) eine hinreichende Optimalitätsbedingung ist. Seien also $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ und ein $\lambda^* \geq 0$ mit (a)–(c) gegeben. Für ein beliebiges $p \in \mathbb{R}^n$ mit $\|p\|_2 \leq \Delta$ ist

$$\begin{aligned}
\phi(p) - \phi(p^*) &= (\underbrace{g + Hp^*}_{-\lambda^* p^*})^T(p - p^*) + \frac{1}{2}(p - p^*)^T H(p - p^*) \\
&= -\lambda^*(p^*)^T(p - p^*) + \frac{1}{2}\underbrace{(p - p^*)^T(H + \lambda^* I)(p - p^*)}_{\geq 0} - \frac{\lambda^*}{2}\|p - p^*\|_2^2 \\
&\geq -\lambda^*(p^*)^T(p - p^*) - \frac{\lambda^*}{2}\|p - p^*\|_2^2 \\
&= \frac{\lambda^*}{2}(\|p^*\|_2^2 - \|p\|_2^2) \\
&= \frac{\lambda^*}{2}(\Delta^2 - \|p\|_2^2) \\
&\geq 0,
\end{aligned}$$

und daher p^* eine globale Lösung von (P). Ist $H + \lambda^* I$ sogar positiv definit, so entnimmt man der obigen Gleichungs-Ungleichungskette, dass $p = p^*$ aus $\phi(p) = \phi(p^*)$ folgt, und das bedeutet die Eindeutigkeit der globalen Lösung.

Den Beweis dafür, dass die Existenz eines $\lambda^* \geq 0$ auch eine notwendige Bedingung für die Optimalität eines p^* ist, werden wir nicht bringen, sondern verweisen lediglich auf J. Werner (1992b, S. 239) und C. Geiger, C. Kanzow (1999, S. 259).

Ist $\phi(p^*) = f$, so ist auch $p^{**} := 0$ eine Lösung. Der hierzu nach dem ersten Teil des Satzes existierende Parameter λ^{**} verschwindet (siehe (b)) und aus (a) und (c) folgt sofort, dass $g = 0$ und H positiv semidefinit ist. Die Umkehrung erhält man aus der Gleichungs-Ungleichungskette

$$f \leq f + \underbrace{g^T p^*}_{=0} + \frac{1}{2}\underbrace{(p^*)^T H p^*}_{\geq 0} = \phi(p^*) \leq \phi(0) = f.$$

Für den Beweis des letzten Teils des Satzes können wir o. B. d. A. $g \neq 0$ annehmen. Für ein beliebiges p mit $\|p\|_2 \leq \Delta$ ist wegen der Optimalität von p^* offenbar

$$(*) \qquad f - \phi(p^*) \geq f - \phi(p) = -g^T p - \frac{1}{2}p^T H p \geq -g^T p - \frac{1}{2}\|p\|_2^2\,\|H\|_2.$$

Ist $\Delta\|H\|_2 \leq \|g\|_2$, so ist wegen $(*)$ (setze $p := -(\Delta/\|g\|_2)g$)

$$f - \phi(p^*) \geq \Delta\|g\|_2 - \frac{1}{2}\Delta^2\|H\|_2 \geq \frac{1}{2}\Delta\|g\|_2.$$

[19] Auch F. Jarre, J. Stoer (2004, S. 157) machen es sich einfach.

Ist dagegen $\Delta\|H\|_2 > \|g\|_2$, so setze man $p := -(1/\|H\|_2)g$ und erhalte aus $(*)$

$$f - \phi(p^*) \geq \frac{\|g\|_2^2}{\|H\|_2} - \frac{\|g\|_2^2}{2\|H\|_2} = \frac{\|g\|_2^2}{2\|H\|_2},$$

insgesamt ist die behauptete Abschätzung bewiesen. □

Wir kommen zur numerischen Berechnung einer oder der Lösung p^* von (P) und führen diese Berechnung auf die Lösung einer nichtlinearen Gleichung in einer Unbekannten zurück.

Seien $\lambda_1 \leq \cdots \leq \lambda_n$ die Eigenwerte der symmetrischen Matrix $H \in \mathbb{R}^{n\times n}$. Wir suchen nach einem $\lambda^* \geq \max(0, -\lambda_1)$ (da nur dann $\lambda^* \geq 0$ und $H + \lambda^* I$ positiv semidefinit ist) und einem $p^* \in \mathbb{R}^n$ mit $(H+\lambda^* I)p^* = -g$ und $\lambda^*(\|p^*\|_2 - \Delta) = 0$. Hierbei nehmen wir an (dies ist eine echte Einschränkung!), wir könnten uns bei der Suche von λ^* auf das Intervall $(-\lambda_1, \infty)$ beschränken. Dies impliziert, dass $H + \lambda^* I$ positiv definit und folglich die Lösung von (P) eindeutig ist. Die Funktion $p : (-\lambda_1, \infty) \longrightarrow \mathbb{R}$ sei durch

$$p(\lambda) := -(H + \lambda I)^{-1} g$$

definiert. Zur Berechnung von λ^* bietet sich die folgende Vorgehensweise an:

- Berechne die Cholesky-Zerlegung von H, um zu erkennen, ob H positiv definit ist[20]. Ist dies der Fall, so berechne $p(0) = -H^{-1}g$ durch Vorwärts- und Rückwärtseinsetzen. Ist $\|p(0)\|_2 \leq \Delta$, so ist $p^* := p(0)$ (mit $\lambda^* = 0$) die eindeutige globale Lösung von (P).

- Bestimme iterativ eine positive Lösung $\lambda^* \in (-\lambda_1, \infty)$ von $\|p(\lambda)\|_2 = \Delta$. Es liegt nahe, das Newton-Verfahren anzuwenden. Wegen der Singularität von $\|p(\cdot)\|_2 - \Delta$ in $-\lambda_1$ wenden wir das Newton-Verfahren aber nicht auf die Gleichung $\|p(\lambda)\|_2 - \Delta = 0$, sondern auf
$$\chi(\lambda) := \frac{1}{\|p(\lambda)\|_2} - \frac{1}{\Delta} = 0$$
an.

Nun ist es nützlich, Eigenschaften der eben definierten Funktion $\chi(\cdot)$ zu sammeln.

Lemma 2.2 *Sei $g \in \mathbb{R}^n \setminus \{0\}$ und $H \in \mathbb{R}^{n\times n}$ symmetrisch mit kleinstem Eigenwert λ_1, ferner sei $\Delta > 0$. Auf $(-\lambda_1, \infty)$ definiere man*

$$p(\lambda) := -(H + \lambda I)^{-1} g, \qquad \chi(\lambda) := \frac{1}{\|p(\lambda)\|_2} - \frac{1}{\Delta}.$$

Dann ist $\chi(\cdot)$ auf $(-\lambda_1, \infty)$ beliebig oft differenzierbar. Es ist

$$\chi'(\lambda) = -\frac{p(\lambda)^T p'(\lambda)}{\|p(\lambda)\|_2^3}, \qquad p'(\lambda) = -(H + \lambda I)^{-1} p(\lambda).$$

Weiter ist $\chi(\cdot)$ streng monoton wachsend und konkav auf $(-\lambda_1, \infty)$.

[20] Eine symmetrische Matrix ist genau dann positiv definit, wenn sie eine Cholesky-Zerlegung besitzt.

Beweis: Sei $U = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}$ eine orthogonale Matrix, die H auf Diagonalgestalt transformiert: $U^T H U = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Dann ist

$$\frac{1}{\|p(\lambda)\|_2} = \Big(\sum_{i=1}^{n} \frac{(u_i^T g)^2}{(\lambda_i + \lambda)^2}\Big)^{-1/2},$$

woraus man die Aussage über die Differenzierbarkeit von $\chi(\cdot)$ auf $(-\lambda_1, \infty)$ abliest. Differentiation von

$$\chi(\lambda) = [p(\lambda)^T p(\lambda)]^{-1/2} - \frac{1}{\Delta}$$

liefert

$$\chi'(\lambda) = -[p(\lambda)^T p(\lambda)]^{-3/2} p(\lambda)^T p'(\lambda) = -\frac{p(\lambda)^T p'(\lambda)}{\|p(\lambda)\|_2^3}.$$

Eine erneute Differentiation führt auf

$$\chi''(\lambda) = \frac{3(p(\lambda)^T p'(\lambda))^2}{\|p(\lambda)\|_2^5} - \frac{p(\lambda)p''(\lambda) + \|p'(\lambda)\|_2^2}{\|p(\lambda)\|_2^3}.$$

Durch zweimaliges Differenzieren der Gleichung $(H + \lambda I)p(\lambda) = -g$ erhält man

$$p(\lambda) + (H + \lambda I)p'(\lambda) = 0, \qquad 2p'(\lambda) + (H + \lambda I)p''(\lambda) = 0.$$

Daher ist

$$p'(\lambda) = -(H + \lambda I)^{-1} p(\lambda), \qquad p''(\lambda) = -2(H + \lambda I)^{-1} p'(\lambda).$$

Einsetzen ergibt

$$\chi'(\lambda) = \frac{p(\lambda)^T (H + \lambda I)^{-1} p(\lambda)}{\|p(\lambda)\|_2^3} > 0,$$

also ist $\chi(\cdot)$ monoton wachsend, und

$$p(\lambda)^T p''(\lambda) = [-(H + \lambda I)p'(\lambda)]^T [-2(H + \lambda I)p'(\lambda)] = 2\|p'(\lambda)\|_2^2,$$

folglich wegen der Cauchy-Schwarzschen Ungleichung

$$\chi''(\lambda) = \frac{3[(p(\lambda)^T p'(\lambda))^2 - \|p(\lambda)\|_2^2 \, \|p'(\lambda)\|_2^2]}{\|p(\lambda)\|_2^5} \leq 0.$$

Daher ist $\chi(\cdot)$ auf $(-\lambda_1, \infty)$ konkav, das Lemma ist bewiesen. □

Ein Newton-Schritt, angewandt auf die Gleichung $\chi(\lambda) = 0$, kann folgendermaßen realisiert werden:

- Input: $\Delta > 0$ und aktuelle Näherung $\lambda > -\lambda_1$.
- Berechne Cholesky-Faktorisierung $B + \lambda I = LL^T$.
- Berechne $p = p(\lambda)$ durch Vorwärts- und Rückwärtseinsetzen aus $LL^T p = -g$.

- Berechne w durch Vorwärtseinsetzen aus $Lw = p$. Wie man leicht nachrechnet ist dann
$$\chi'(\lambda) = \frac{\|w\|_2^2}{\|p\|_2^3}.$$
- Output: Die neue Newton-Iterierte
$$\lambda_+ := \lambda - \frac{\chi(\lambda)}{\chi'(\lambda)} = \lambda + \Big(\frac{\|p\|_2 - \Delta}{\Delta}\Big)\Big(\frac{\|p\|_2}{\|w\|_2}\Big)^2.$$

Bemerkung: Die Monotonie und Konkavität von $\chi(\cdot)$ hat erfreuliche Konsequenzen. Startet man das Newton-Verfahren nämlich mit einem Startwert aus $(-\lambda_1, \lambda^*)$, wobei λ^* die eindeutige Lösung von $\chi(\lambda) = 0$ ist, so erhält man eine Folge, die monoton wachsend von links (mit quadratischer Konvergenzgeschwindigkeit) gegen λ^* konvergiert. Ist der Startwert dagegen rechts von λ^*, so liegt die nächste Newton-Iterierte λ_+ links von λ^*, wobei allerdings nicht notwendig $-\lambda_1 < \lambda_+$. Man mache sich den geschilderten Sachverhalt durch eine Zeichnung klar! □

2.2.3 Globale Konvergenz

In diesem Unterabschnitt werden wir einen globalen Konvergenzsatz für ein Trust-Region-Verfahren mit quadratischem Modell formulieren, leider aber auf einen Beweis verzichten müssen. Einen Beweis findet man bei J. Werner (1992b, S. 241 ff.) oder auch auf S. 151 ff. des Skriptes einer Vorlesung über Unrestringierte Optimierungsaufgaben von J. Werner.

Satz 2.3 *Gegeben sei die unrestringierte Optimierungsaufgabe*

(P) $$\textit{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n.$$

Die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \le f(x_0)\}$ sei kompakt, wobei x_0 der Startvektor für das gleich anzugebende Verfahren ist. Die Zielfunktion f sei auf einer offenen Obermenge von L_0 stetig differenzierbar und der Gradient $\nabla f(\cdot)$ auf L_0 lipschitzstetig. Ferner sei $\{H_k\} \subset \mathbb{R}^{n\times n}$ eine beschränkte Folge symmetrischer Matrizen. Man betrachte das folgende Verfahren.

- *Gegeben seien Konstanten $0 < \rho_1 < \rho_2 < 1$, $\sigma_1 \in (0,1)$ und $\sigma_2 > 1$.*
- *Seien $x_0 \in \mathbb{R}^n$ und $\Delta_0 > 0$ gegeben. Berechne $g_0 := \nabla f(x_0)$.*
- *Für $k = 0, 1, \ldots$:*
  - *Berechne eine globale Lösung*[21] *p_k der Aufgabe*

$$(\mathrm{P}_k) \qquad \textit{Minimiere} \quad f_k(p) := f(x_k) + g_k^T p + \frac{1}{2} p^T H_k p, \quad \|p\|_2 \le \Delta_k.$$

[21] Im Beweis wird nur ausgenutzt, dass eine Abschätzung wie im letzten Teil von Satz 2.1 gilt.

- *Falls $f(x_k) = f_k(p_k)$, dann: STOP. Es ist $g_k = 0$ und daher x_k eine stationäre Lösung von (P).*
- *Berechne*
$$r_k := \frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)}.$$
- *Falls $r_k \geq \rho_1$, dann setze $x_{k+1} := x_k + p_k$ und berechne $g_{k+1} := \nabla f(x_{k+1})$. In diesem Falle nennen wir den Iterationsschritt k erfolgreich.*
- *Andernfalls setze $x_{k+1} := x_k$ sowie $g_{k+1} := g_k$.*
- *Update des Trust-Region-Radius:*
  - *Falls $r_k < \rho_1$, dann wähle $\Delta_{k+1} \in (0, \sigma_1 \Delta_k]$.*
  - *Falls $r_k \in [\rho_1, \rho_2)$, dann wähle $\Delta_{k+1} \in [\sigma_1 \Delta_k, \Delta_k]$.*
  - *Falls $r_k \geq \rho_2$, dann wähle $\Delta_{k+1} \in [\Delta_k, \sigma_2 \Delta_k]$.*

Das Verfahren breche nicht vorzeitig ab. Dann liefert es eine Folge $\{x_k\}$ mit $\lim_{k\to\infty} g_k = 0$, insbesondere ist jeder Häufungspunkt der Folge $\{x_k\}$ eine stationäre Lösung von (P).

Sind die Voraussetzungen von Satz 2.3 erfüllt, ist f auf einer offenenen Obermenge von L_0 sogar zweimal stetig differenzierbar und wird $H_0 := \nabla^2 f(x_0)$ sowie

$$H_{k+1} := \begin{cases} \nabla^2 f(x_{k+1}), & \text{Iterationsschritt } k \text{ erfolgreich,} \\ H_k, & \text{Iterationsschritt } k \text{ nicht erfolgreich,} \end{cases}$$

gesetzt (in diesem Falle spricht man naheliegenderweise vom Trust-Region-Newton-Verfahren), so kann man zeigen (siehe J. Werner (1992b, S. 241 ff.)):

- Die Folge $\{x_k\}$ besitzt mindestens einen Häufungspunkt x^*, in dem die notwendigen Optimalitätsbedingungen zweiter Ordnung erfüllt sind, für den also $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv semidefinit ist.
- Ist x^* ein Häufungspunkt der Folge $\{x_k\}$ und ist $\nabla^2 f(x^*)$ positiv definit, so konvergiert die Folge $\{x_k\}$ gegen x^*. Ferner sind nach endlich vielen Schritten alle Iterationsschritte erfolgreich, es ist $\|p_k\|_2 < \Delta_k$ für alle hinreichend großen k und $\inf_{k=0,1,\ldots} \Delta_k > 0$. Ist darüberhinaus $\nabla^2 f(\cdot)$ auf einer Kugel um x^* in x^* lipschitzstetig, so konvergiert die Folge $\{x_k\}$ von mindestens zweiter Ordnung gegen x^*.

2.2.4 Nichtlineare Ausgleichsprobleme

Einige wenige Aussagen wollen wir noch zur Anwendung des Trust-Region-Verfahrens auf ein nichtlineares Ausgleichsproblem machen[22]. Gegeben sei also die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \tfrac{1}{2}\|F(x)\|_2^2, \quad x \in \mathbb{R}^n,$$

[22] Wenn hier nicht alles verstanden wird, weil entsprechende Vorkenntnisse aus der numerischen linearen Algebra fehlen (z. B. Kenntnisse über Pseudoinverse und Singulärwertzerlegung), so ist dies nicht schlimm.

wobei $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ mit $m \geq n$ glatt sei. Die wesentliche Arbeit bei der Anwendung des Trust-Region-Verfahrens besteht in der Lösung des hierbei auftretenden in einer aktuellen Lösung x linearisierten Hilfsproblems:

$$(\mathrm{P}_{x,\Delta}) \qquad \text{Minimiere} \quad f_x(p) := \tfrac{1}{2}\|F(x) + F'(x)p\|_2^2, \quad \|p\|_2 \leq \Delta.$$

Hier ist offenbar

$$f_x(p) = \tfrac{1}{2}\|F(x)\|_2^2 + [F'(x)^T F(x)]^T p + \tfrac{1}{2} p^T F'(x)^T F'(x) p.$$

Da $H := F'(x)^T F'(x)$ positiv semidefinit ist, braucht nicht zwischen lokaler und globaler Lösung von $(\mathrm{P}_{x,\Delta})$ unterschieden zu werden. Aus Satz 2.1 erhalten wir damit:

- *Genau dann ist ein $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ eine Lösung von $(P_{x,\Delta})$, wenn ein $\lambda^* \geq 0$ mit*

  $$[F'(x)^T F'(x) + \lambda^* I]p^* = -F'(x)^T F(x), \qquad \lambda^*(\Delta - \|p^*\|_2) = 0$$

  existiert.

Es ist nützlich, etwas über die *Singulärwertzerlegung* einer Matrix zu wissen. Es gilt die folgende Aussage:

- Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ gegeben. Dann existieren orthogonale Matrizen $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ und eine Diagonalmatrix $\hat{\Sigma} := \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n)$ mit $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ derart, dass die sogenannte *Singulärwertzerlegung*

  $$A = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T$$

  von A gilt. Die Anzahl r der positiven sogenannten *singulären Werte* σ_i stimmt mit dem Rang von A überein. Ferner ist die sogenannte *Pseudoinverse* $A^+ \in \mathbb{R}^{n\times m}$ definiert durch

  $$A^+ := V(\ \hat{\Sigma}^+ \quad 0\ )U^T,$$

  wobei die Diagonalmatrix $\hat{\Sigma}^+ \in \mathbb{R}^{n\times n}$ durch

  $$\hat{\Sigma}^+ := \mathrm{diag}\,(1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0)$$

  definiert ist. Natürlich ist hier r der Rang von A bzw. die Anzahl positiver singulärer Werte. Bei vorgegebenem $b \in \mathbb{R}^m$ ist $A^+ b$ unter allen Lösungen des linearen Ausgleichsproblems, $\|Ax - b\|_2$ auf dem $\mathbb{R}^n$ zu minimieren, die eindeutige Lösung mit minimaler euklidischer Norm.

In MATLAB kann die Singulärwertzerlegung einer Matrix sehr einfach mit der Funktion `svd` berechnet werden.

Wir definieren die sogenannte *Levenberg-Marquardt-Trajektorie* $p : [0, \infty) \longrightarrow \mathbb{R}^n$ durch

$$p(\lambda) := \begin{cases} -F'(x)^+ F(x), & \lambda = 0, \\ -[F'(x)^T F'(x) + \lambda I]^{-1} F'(x)^T F(x), & \lambda > 0. \end{cases}$$

Hierbei bedeutet $F'(x)^+$ die Pseudoinverse von $F'(x)$. In MATLAB gibt es die Funktion `pinv`, mit der die Pseudoinverse berechnet werden kann. Mit Hilfe einer Singulärwertzerlegung von $F'(x)$ kann man $p(0)$ sowie $p(\lambda)$ für $\lambda > 0$ berechnen und zeigen, dass $p(0) = \lim_{\lambda \to 0+} p(\lambda)$. Denn seien orthogonale Matrizen

$$U = \begin{pmatrix} u_1 & \cdots & u_m \end{pmatrix} \in \mathbb{R}^{m \times m}, \qquad V = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} \in \mathbb{R}^{n \times n}$$

und eine "Diagonalmatrix"

$$\Sigma = \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} \in \mathbb{R}^{m \times n}, \qquad \hat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$$

mit

$$F'(x) = U \Sigma V^T$$

mit $\sigma_1 \geq \cdots \geq \sigma_n$ (die singulären Werte sind nichtnegativ) bekannt. Ferner sei $r := \operatorname{Rang}(F'(x))$ (der Rang stimmt mit der Anzahl positiver Singulärwerte überein[23]). Dann ist

$$\begin{aligned} p(\lambda) &= -(F'(x)^T F(x) + \lambda I)^{-1} F'(x)^T F(x) \\ &= -(V \Sigma^T \underbrace{U^T U}_{=I} \Sigma V^T + \lambda I)^{-1} V \Sigma^T U^T F(x) \\ &= -V(\Sigma^T \Sigma + \lambda I)^{-1} \Sigma^T U^T F(x) \\ &= -V(\hat{\Sigma}^2 + \lambda I)^{-1} \begin{pmatrix} \hat{\Sigma} & 0 \end{pmatrix} U^T F(x) \\ &= -\sum_{j=1}^{r} \frac{\sigma_j z_j}{\sigma_j^2 + \lambda} v_j, \end{aligned}$$

wobei $z := U^T F(x) = (u_j^T F(x))$. Daher ist

$$\lim_{\lambda \to 0+} p(\lambda) = -\sum_{j=1}^{r} \frac{z_j}{\sigma_j} v_j = -F'(x)^+ F(x) = p(0).$$

Ferner ist

$$\psi(\lambda) := \|p(\lambda)\|_2 = \Big(\sum_{j=1}^{r} \frac{\sigma_j^2 z_j^2}{(\sigma_j^2 + \lambda)^2} \Big)^{1/2}, \qquad \psi'(\lambda) = -\frac{1}{\psi(\lambda)} \sum_{j=1}^{r} \frac{\sigma_j^2 z_j^2}{(\sigma_j^2 + \lambda)^3}.$$

Offenbar gibt es zwei Möglichkeiten:

(a) Es ist $\lambda^* = 0$ und $\|p(0)\|_2 \leq \Delta$. Dann ist $p^* := p(0)$ eine Lösung von $(\mathrm{P}_{x,\Delta})$, und zwar eine mit minimaler euklidischer Norm.

(b) Es ist $\lambda^* > 0$ und $\|p(\lambda^*)\|_2 = \Delta$. Dann ist $p^* := p(\lambda^*)$ die eindeutige Lösung von $(\mathrm{P}_{x,\Delta})$.

[23] In MATLAB wird der Rang einer Matrix auf genau diese Weise berechnet.

Im zweiten Fall (hier ist $\psi(0) > \Delta$) wird das Newton-Verfahren wieder auf

$$\chi(\lambda) := \frac{1}{\psi(\lambda)} - \frac{1}{\Delta} = 0$$

angewandt. Auf $(0, \infty)$ ist $\chi(\cdot)$ streng monoton wachsend und konkav (siehe Lemma 2.2, es ist aber auch ein Beweis mit Hilfe obiger Darstellung von $\psi(\cdot)$ möglich), weiter ist $\psi(\cdot)$ streng monoton fallend und konvex auf $[0, \infty)$ (Beweis?). Untere und obere Schranken für die eindeutige Lösung $\lambda^* \in (0, \infty)$ von $\psi(\lambda) = \Delta$ kann man leicht bestimmen. Wegen

$$\Delta = \psi(\lambda^*) = \Big(\sum_{j=1}^{r} \frac{\sigma_j^2 z_j^2}{(\sigma_j^2 + \lambda^*)^2}\Big)^{1/2} \le \frac{1}{\lambda^*}\Big(\sum_{j=1}^{r} \sigma_j^2 z_j^2\Big)^{1/2}$$

ist

$$u_0 := \frac{1}{\Delta}\Big(\sum_{j=1}^{r} \sigma_j^2 z_j^2\Big)^{1/2}$$

eine obere Schranke für λ^*. Wegen der Konvexität von $\psi(\cdot)$ auf $[0, \infty)$ ist

$$\Delta = \psi(\lambda^*) \ge \psi(0) + \psi'(0)\lambda^*,$$

so dass

$$l_0 := -\frac{\psi(0) - \Delta}{\psi'(0)}$$

eine (positive) untere Schranke für λ^* ist. Als Startwert λ_0 für das auf $\chi(\lambda) = 0$ angewandte Newton-Verfahren

$$\lambda_{k+1} := \lambda_k + \Big(1 - \frac{\psi(\lambda_k)}{\Delta}\Big)\frac{\psi(\lambda_k)}{\psi'(\lambda_k)}$$

nimmt man dann z. B. das geometrische Mittel der unteren und der oberen Schranke für λ^*, also $\lambda_0 := \sqrt{l_0 u_0}$.

Wir geben eine einfache MATLAB-Funktion zur Lösung des Trust-Region-Hilfsproblems an:

```
function [p,lambda]=Trust_Sub(F,J,Delta);
%************************************************************
%Diese Funktion loest das Trust-Region-Subproblem
%   Minimiere f_x(p):=||F+J p||_2, ||p||_2<=Delta
%************************************************************
%Input-Parameter:
%           F,J
%           Delta       Aktueller Radius
%Output-Parameter:
%           p           (naeherungsweise) Loesung
%           lambda      zugehoeriger Multiplikator
%************************************************************
sigma=0.1;%(beende, falls | ||p(lambda)||-Delta|<=sigma*Delta)
[U,S,V]=svd(J,0);s=diag(S);toll=max(size(J))*max(s)*eps;
```

```
r=sum(s > toll);%r=rang(J)
z=U(:,1:r)'*F;s=s(1:r);
p_0=-V(:,1:r)*diag(ones(r,1)./s)*z;psi_0=norm(p_0);
if  psi_0<=Delta
    p=p_0;lambda=0;
else
    dpsi_0=-sum((z./s.^2).^2)/psi_0;
    l=-(psi_0-Delta)/dpsi_0;u=norm(s.*z)/Delta;
    lambda=sqrt(l*u);
    psi=norm((s.*z)./(s.^2+lambda));
    while abs(psi-Delta)>sigma*Delta
        dpsi=-sum((s.*z).^2./((s.^2+lambda).^3))/psi;
        lambda=lambda+(1-psi/Delta)*(psi/dpsi);
        psi=norm((s.*z)./(s.^2+lambda));
    end;
    p=-V(:,1:r)*diag(s./(s.^2+lambda))*z;
end;
```

Hierauf aufbauend schreiben wir eine Funktion, welche mit Hilfe des Trust-Region-Verfahrens ein nichtlineares Ausgleichsproblem löst.

```
function [x,iter]=Trust_Least(Fun,x_0,Delta_0,max_iter,tol);
%************************************************************
%Diese Funktion loest das nichtlineare Ausgleichsproblem
%      Minimiere f(x):=||F(x)||_2,  x in R^n
%mit dem  Trust-Region-Verfahren
%************************************************************
%Input-Parameter:
%      Fun           [F(x),F'(x)]=Fun(x)
%      x_0           Startwert
%      Delta_0       Anfangsradius
%      max_iter      maximale Zahl der Iterationen
%      tol           Toleranz
%Output parameter:
%      x             (naeherungsweise) Loesung
%      iter          Zahl der  Iterations
%************************************************************
rho_1=0.01;rho_2=0.9;sigma_1=0.5;sigma_2=2;
x=x_0;Delta=Delta_0;[F,J]=feval(Fun,x);iter=0;
[p,lambda]=Trust_Sub(F,J,Delta);
f=norm(F);f_x=norm(F+J*p);
while (f-f_x>tol)&(iter<max_iter)
     iter=iter+1;
     x_plus=x+p;
     [F_plus,J_plus]=feval(Fun,x_plus);
     f_plus=norm(F_plus);
     r=(f-f_plus)/(f-f_x);
     if (r<rho_1)
        Delta=sigma_1*Delta;
     else
        if (r>=rho_2)
            Delta=sigma_2*Delta;
        end;
     end;
```

```
    if (r>=rho_1)
        x=x_plus;F=F_plus;J=J_plus;
    end;
    [p,lambda]=Trust_Sub(F,J,Delta);
    f=norm(F);f_x=norm(F+J*p);
end;
x=x;%Ende Trust_Least
```

Beispiel: Wir kommen auf ein Beispiel aus der Einführung zurück. Dort betrachteten wir die Aufgabe

$$\text{Minimiere} \quad f(x) := \sum_{i=1}^{10} (x_1 e^{x_2 t_i} - z_i)^2, \quad x \in \mathbb{R}^2,$$

wobei

t_i	0.9	1.5	13.8	19.8	24.1	28.2	35.2	60.3	74.6	81.3
z_i	455.2	428.6	124.1	67.3	43.2	28.1	13.1	-0.4	-1.3	-1.5

und lösten diese Aufgabe mit Hilfe der MATLAB-Funktion `lsqcurvefit`. Es ist also

$$F(x) := \begin{pmatrix} x_1 e^{x_2 t_1} - z_1 \\ \vdots \\ x_1 e^{x_2 t_{10}} - z_{10} \end{pmatrix}, \qquad F'(x) = \begin{pmatrix} e^{x_2 t_1} & t_1 x_1 e^{x_2 t_1} \\ \vdots & \vdots \\ e^{x_2 t_{10}} & t_{10} x_1 e^{x_2 t_{10}} \end{pmatrix}.$$

Wir schreiben ein file `Myfun.m` mit dem Inhalt

```
function [F,J]=Myfun(x);
t=[0.9;1.5;13.8;19.8;24.1;28.2;35.2;60.3;74.6;81.3];
z=[455.2;428.6;124.1;67.3;43.2;28.1;13.1;-0.4;-1.3;-1.5];
F=x(1)*exp(x(2)*t)-z;
if nargout>1
    J=[exp(x(2)*t),x(1)*(t.*exp(x(2)*t))];
end;
```

Der Aufruf

```
[x,iter]=Trust_Least('Myfun',[100;-1],0.5,100,1e-10);
```

liefert

$$x = \begin{pmatrix} 498.8309 \\ -0.1013 \end{pmatrix}, \qquad \text{iter} = 18,$$

ferner ist

$$\|F(x)\|_2 = 3.082999658188347.$$

Wir vergleichen dieses Ergebnis mit dem, welches wir nach Anwendung von `lsqnonlin` aus der Optimization-Toolbox von MATLAB erhalten. Nach dem Aufruf

```
x=lsqnonlin('Myfun',[100;-1]);
```

erhalten wir bei `format short` dasselbe Ergebnis wie oben, diesmal ist allerdings

$$\|F(x)\|_2 = 3.082999658188315,$$

eine Winzigkeit besser als obiges Ergebnis. Die Funktion `lsqcurvefit` liefert (ohne Setzen anderer Optionen) schlechtere Ergebnisse. □

2.2.5 Aufgaben

1. Gegeben sei die Aufgabe

   (P) $$\text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T H p, \quad \|p\|_2 \leq \Delta,$$

   wobei
   $$g := \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad H := \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \Delta := \frac{1}{2}.$$
   Man definiere $p : [0, \infty) \longrightarrow \mathbb{R}$ durch $p(\lambda) := -(H + \lambda I)^{-1} g$. Auf dem Intervall $[0, 2]$ plotte man $\|p(\cdot)\|_2$ und $1/\|p(\cdot)\|_2$. Anschließend berechne man die Lösung von (P) numerisch, indem man z.B. auf
   $$\chi(\lambda) := \frac{1}{\|p(\lambda)\|_2} - \frac{1}{\Delta} = 0$$
   das Newton-Verfahren mit Startwert $\lambda_0 := 0$ anwendet. Ist $\chi(\lambda^*) = 0$, so ist $p^* := p(\lambda^*)$ die Lösung von (P).

2. Sei $f \in \mathbb{R}$, $g \in \mathbb{R}^n \setminus \{0\}$ und $\Delta > 0$. Man gebe eine Lösung von

   (P) $$\text{Minimiere} \quad \phi(p) := f + g^T p, \quad \|p\|_\infty \leq \Delta$$

   an und begründe dies.

3. Man betrachte die unrestringierte Optimierungsaufgabe

   (P) $$\text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad p \in \mathbb{R}^n,$$

   wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^n$ und $B \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix ist. Man zeige:

   (a) (P) besitzt genau dann eine Lösung, wenn B positiv semidefinit und $g \in \text{Bild}\,(B)$ ist.

   (b) (P) besitzt genau dann eine eindeutige Lösung, wenn B positiv definit ist.

4. Sei
   $$g := \begin{pmatrix} -2 \\ -20 \end{pmatrix}, \qquad B := \begin{pmatrix} 42 & 0 \\ 0 & 20 \end{pmatrix}.$$
   Für $\Delta := \frac{1}{2}$, 1, 2 berechne man eine Lösung von

   (P) $$\text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_\infty \leq \Delta.$$

5. Sei $B \in \mathbb{R}^{n \times n}$ symmetrisch mit kleinstem Eigenwert λ_1 und $g \in \mathbb{R}^n \setminus \{0\}$. Man zeige, dass die durch
   $$\psi(\lambda) := \|(B + \lambda I)^{-1} g\|_2$$
   definierte Funktion $\psi\colon (-\lambda_1, \infty) \longrightarrow \mathbb{R}$ auf $(-\lambda_1, \infty)$ monoton fallend und konvex ist.

6. **Programmieraufgabe:** Sei $H \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $g \in \mathbb{R}^n$ und $\Delta > 0$. Man schreibe eine MATLAB-Funktion `TrustStep` zur Berechnung der Lösung des Trust-Region-Hilfsproblems

   $$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T H p, \quad \|p\|_2 \le \Delta.$$

   Anschließend teste man die Funktion für den Spezialfall, dass

   $$H := \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad g := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^n$$

   mit $n := 10$ und $\Delta := 1$.

7. Gegeben sei das Trust-Region Hilfsproblem

   $$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T H p, \quad \|p\|_2 \le \Delta,$$

   wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^{n\times n} \setminus \{0\}$, die symmetrische Matrix $H \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben sind. Der sogenannte *Cauchy-Punkt* p^C ist definiert durch

   $$p^C := -\tau \frac{\Delta}{\|g\|_2} g,$$

   wobei

   $$\tau := \begin{cases} 1, & \text{falls} \quad g^T H g \le 0, \\ \min(\|g\|_2^3/(\Delta g^T H g), 1), & \text{sonst.} \end{cases}$$

   Man zeige, dass p^C eine Lösung der Aufgabe

   $$\text{Minimiere} \quad \phi(p), \quad \|p\|_2 \le \Delta,\ p \in \operatorname{span}\{g\}$$

   ist.

   Hinweis: Man stelle $p \in \operatorname{span}\{g\}$ mit $\|p\|_2 \le \Delta$ dar als $p = -\sigma(\Delta/\|g\|_2)g$ mit $|\sigma| \le 1$.

8. Gegeben sei das Trust-Region Hilfsproblem

   $$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T H p, \quad \|p\|_2 \le \Delta,$$

   wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^{n\times n} \setminus \{0\}$, die symmetrische Matrix $H \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben sind. Sei p^C der in Aufgabe 7 definierte Cauchy-Punkt. Man zeige, dass

   $$f - \phi(p^C) \ge \frac{1}{2}\|g\|_2 \min\Big(\Delta, \frac{\|g\|_2}{\|H\|_2}\Big),$$

   der Cauchy-Punkt p^C also derselben Abschätzung wie eine globale Lösung p^* von (P) genügt, siehe Satz 2.1.

   Hinweis: Man unterscheide zwischen den Fällen $g^T H g \le 0$ und $g^T H g > 0$. Für $g^T H g > 0$ betrachte man die beiden Fälle $\|g\|_2^3/g^T H g \le 1$ und $\|g\|_2^3/g^T H g > 1$.

Kapitel 3

Theoretische Grundlagen restringierter Optimierungsaufgaben

Im letzten Kapitel über unrestringierte Optimierungsaufgaben spielten notwendige und hinreichende Optimalitätsbedingungen eine wichtige Rolle. Dies gilt erst recht für restringierte Optimierungsaufgaben, wobei die entsprechenden Ergebnisse aber nicht so einfach zu erhalten sind. Wir betrachten die Optimierungsaufgabe

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

wobei die Menge $M \subset \mathbb{R}^n$ i. Allg. als Lösungsmenge eines Gleichungs-Ungleichungssystems spezifiziert und die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ hinreichend glatt ist. Bei der Herleitung notwendiger Optimalitätbedingungen, und das sind die wichtigsten Ergebnisse in diesem Kapitel, spielen zwei Mengen eine ganz entscheidende Rolle. Da sie so wichtig sind, geben wir sie jetzt schon an, bevor wir mit dem eigentlichen Stoff dieses Kapitels beginnen.

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

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

  der *Kegel*[1] *der zulässigen Richtungen* an M in x^*. Dagegen 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* bzw. der *Kegel der tangentialen Richtungen* an M in x^*.

Ist z. B. $M := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ die Oberfläche der euklidischen Einheitskugel, so ist $F(M; x^*)$ offenbar trivial, also $F(M; x^*) = \{0\}$, aber (Beweis?)

$$T(M; x^*) = \{p \in \mathbb{R}^n : (x^*)^T p = 0\}.$$

[1] Unter einem *Kegel* versteht man eine Menge, die mit einem Punkt auch jedes nichtnegative Vielfache enthält. Durch Wikipedia erfährt man: Die Redewendung *Kind und Kegel* bedeutet eigentlich alle ehelichen und unehelichen Kinder. Heute steht der Begriff für die gesamte Verwandtschaft oder auch teilweise für Kinder, Haustiere und Gepäck. Wenn jemand mit Kind und Kegel reist, so ist der Ausdruck scherzhaft zu verstehen und derjenige hat die gesamte Familie dabei.

Ganz allgemein ist $F(M;x^*) \subset T(M;x^*)$. Man kann zeigen, dass der Tangentialkegel $T(M;x^*)$ abgeschlossen ist[2] und daher den Abschluss cl $F(M;x^*)$ des Kegels der zulässigen Richtungen enthält. Die Wichtigkeit dieser beiden Mengen für notwendige Optimalitätsbedingungen erkennt man an der folgenden Beobachtung:

- *Sei $x^* \in M$ eine lokale Lösung der Optimierungsaufgabe (P). Die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ sei in x^* stetig differenzierbar. Dann ist $\nabla f(x^*)^T p \geq 0$ für alle $p \in T(M;x^*)$.*

Denn: Ist $x^* \in M$ eine lokale Lösung von (P), so existiert eine Umgebung U^* von x^* mit $f(x^*) \leq f(x)$ für alle $x \in U^* \cap M$. 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 Behauptung.

Ein Ausschlachten der in • genannten Aussage, wobei $T(M;x^*)$ häufig durch eine nichttriviale, möglichst große Teilmenge ersetzt wird, liefert mit Hilfe sogenannter Trennungssätze oder Alternativsätze, auf die wir gleich im Anschluss eingehen werden, die gewünschten notwendigen Optimalitätsbedingungen.

3.1 Trennung konvexer Mengen

3.1.1 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 3.1 wird dies mit einem $\gamma > 0$ veranschaulicht. Außerdem haben

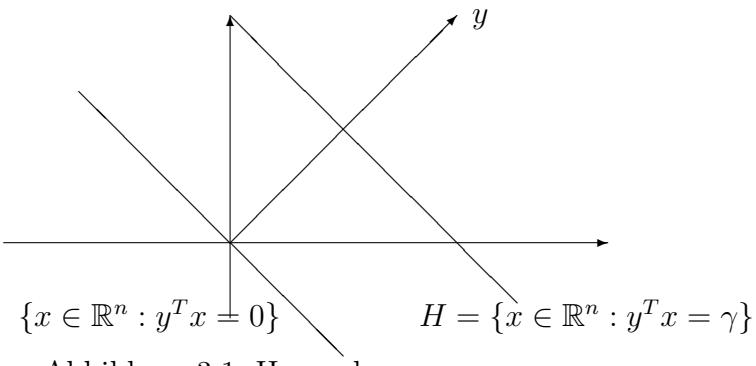


Abbildung 3.1: Hyperebene

wir noch eine parallele Hyperebene durch den Nullpunkt eingezeichnet.

[2]Siehe z. B. C. Geiger, C. Kanzow (2002, S. 42).

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 *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: Nun wollen wir die anschauliche Bedeutung dieser beiden Definitionen klären. Seien also A und B (stark) trennbare Mengen, es existiere also $y \in \mathbb{R}^n \setminus \{0\}$ mit $\sup_{a \in A} y^T a \leq \inf_{b \in B} y^T b$ bzw. $\sup_{a \in A} y^T a < \inf_{b \in B} y^T b$. Man definiere

$$\gamma := \tfrac{1}{2}[\sup_{a \in A} y^T a + \inf_{b \in B} y^T b], \qquad H := \{x \in \mathbb{R}^n : y^T x = \gamma\}.$$

Die Hyperebene H induziert zwei (abgeschlossene) Halbräume, nämlich

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

Dann gelten die folgenden beiden Aussagen, die wir uns in Abbildung 3.2 veranschaulichen, links für die erste Aussage, rechts für die zweite.

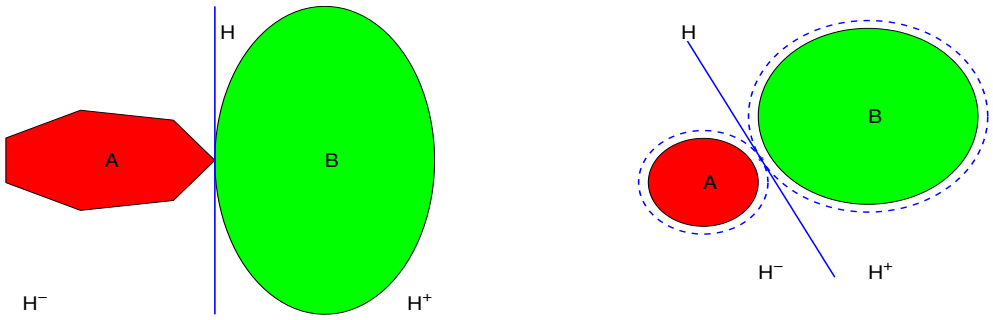


Abbildung 3.2: Trennbare und stark trennbare Mengen

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

(b) 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 (b) nicht ganz offensichtlich. Man definiere

$$\epsilon := \frac{1}{2\,\|y\|}[\inf_{b\in B} y^T b - \sup_{a\in A} y^T a] > 0.$$

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*, bei dessen Beweis wir uns ganz kurz halten können. Es sei daran erinnert, dass $\|\cdot\|$ die *euklidische* Norm bedeutet.

Satz 1.2 *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 $x \in M$.

Beweis: Man betrachte das äquivalente Problem, $g(x) := \frac{1}{2}\|x - z\|^2$ auf M zu minimieren. Die Existenz einer Lösung x^* folgt aus einem Kompaktheitsargument, die Eindeutigkeit aus der strikten Konvexität von $g(\cdot)$. Notwendig gilt $\nabla g(x^*)^T(x - x^*) = (x^* - z)^T(x - x^*) \ge 0$ für alle $x \in M$, da $M - \{x^*\} \subset F(M; x^*)$ (siehe Anfang dieses Kapitels). Wegen der Konvexität von $g(\cdot)$ ist (siehe Satz 1.5 in Unterabschnitt 2.1.2) $0 \le \nabla g(x^*)^T(x - x^*) \le g(x) - g(x^*)$ auch hinreichend für die Optimalität von x^*. □

Bemerkung: Die die Projektion x^* von z auf M charakterisierende Bedingung ist äquivalent zu $\angle(z - x^*, x - x^*) \in [\pi/2, 3\pi/2]$ für alle $x \in M$. In Abbildung 3.3 wird diese Aussage veranschaulicht. □

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

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

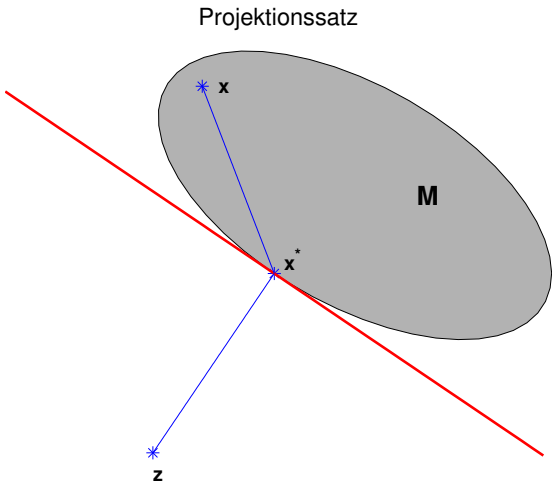


Abbildung 3.3: Veranschaulichung des Projektionssatzes

und daher

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

Also sind A und B stark trennbar. □

Das folgende Korollar ist eine 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 $y^T z < \inf_{x\in K} y^T x$.*

3.1.2 Farkas-Lemma, Trennungssatz

Ziel in diesem Unterabschnitt ist es, das Farkas[3]-Lemma (1902) und als Folgerung hieraus einen Trennungssatz für konvexe Mengen zu beweisen. Wir werden den Beweis des Farkas-Lemmas so führen, dass wir zunächst die Abgeschlossenheit sogenannter *endlich erzeugter Kegel*, das ist die Menge aller nichtnegativen Linearkombinationen endlich vieler Punkte, nachweisen und anschließend den starken Trennungssatz anwenden. In Abbildung 3.4 links veranschaulichen wir uns einen von fünf Punkten $a_1, \ldots, a_5$ des $\mathbb{R}^3$ erzeugten Kegel. Daneben bilden wir einen Kegel ab, der offensichtlich nicht endlich erzeugt ist.

Lemma 1.5 *Sei $A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \in \mathbb{R}^{m\times n}$. 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

[3]Gyula Farkas (auch Julius Farkas) (1847–1930) war ein ungarischer Physiker und Mathematiker.

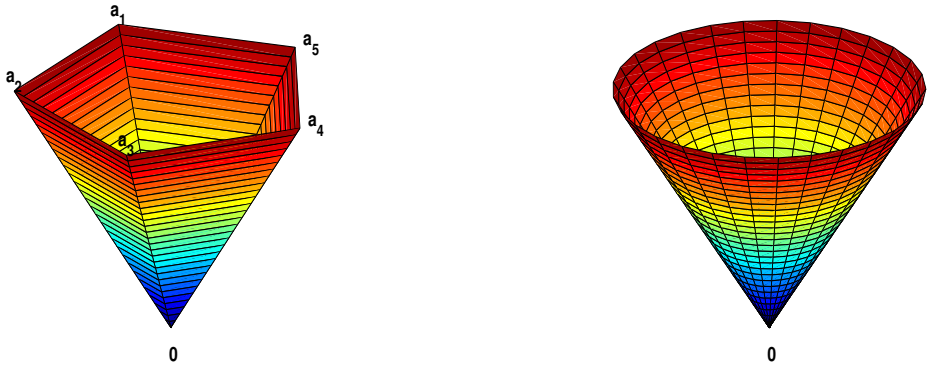


Abbildung 3.4: Kegel im $\mathbb{R}^3$

richtig. Weiter sei K ein von n Elementen $a_1, \ldots, a_n \in \mathbb{R}^m$ erzeugter konvexer Kegel, also

$$K = \Big\{ \sum_{j=1}^{n} x_j a_j : x_j \geq 0 \;\; (j = 1, \ldots, n) \Big\} =: \text{cone}\{a_1, \ldots, a_n\}.$$

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

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

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

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

Mit

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

[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 $y = Ax \in K$.

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

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

genau dann keine Lösung $x \in \mathbb{R}^n$, wenn das System

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

eine Lösung $y \in \mathbb{R}^m$ 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 $b^T y < \inf_{x \geq 0} y^T A x$. Hieraus folgt, dass y eine Lösung von (II) ist. Denn einerseits ist $b^T y < y^T A 0 = 0$, andererseits $(A^T y)^T x > b^T y$ für alle $x \geq 0$ und damit $A^T y \geq 0$. □

Bemerkung: Die anschauliche Bedeutung des Farkas-Lemmas ist sehr einfach und wird auch schon durch den Beweis deutlich. Ist (I) nicht lösbar, so bedeutet dies, dass b nicht in dem von den Spalten $a_1, \ldots, a_n$ von A erzeugten Kegel liegt. Die Lösbarkeit von (II) bedeutet die Existenz einer Hyperebene durch den Nullpunkt mit der Eigenschaft, dass b in einem zugehörigen offenenen Halbraum und $a_1, \ldots, a_n$ in dem gegenüberliegenden abgeschlossenen Halbraum liegen. In Abbildung 3.5 wird dies dargestellt. □

Bemerkung: 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

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

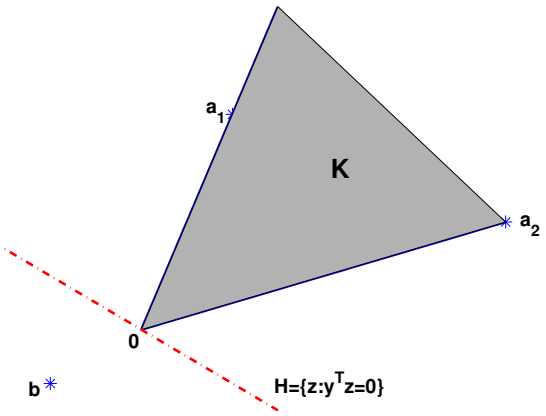


Abbildung 3.5: Veranschaulichung des Farkas-Lemmas

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* z 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 folgt der Trennungssatz für konvexe Mengen. Den Beweis haben wir O. L. Mangasarian (1969, S. 47 ff.) entnommen.

Satz 1.7 *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 1Durchschnitt 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. □

3.1.3 Aufgaben

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
   $$\|P_K(x) - P_K(y)\| \leq \|x - y\| \qquad \text{für alle } x, y \in \mathbb{R}^n.$$
   (Hierbei bedeutet $\|\cdot\|$ natürlich die euklidische Norm auf dem $\mathbb{R}^n$.) Die Projektionsabbildung ist also *nicht expandierend* auf dem $\mathbb{R}^n$.

   (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
   $$P_L(x) = V(V^T V)^{-1} V^T x \qquad \text{für alle } x \in \mathbb{R}^n.$$

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. 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 $C = \operatorname{cl}(\operatorname{int}(C))$.

4. Man zeige, dass zwei nichtleere, konvexe Mengen $A, B \subset \mathbb{R}^n$ genau dann stark trennbar sind, wenn $0 \notin \operatorname{cl}(B - A)$. Für eine Menge $C \subset \mathbb{R}^n$ bedeutet $\operatorname{cl}(C) \subset \mathbb{R}^n$ hierbei den *Abschluss* der Menge C, es ist also
   $$\operatorname{cl}(C) := \{x \in \mathbb{R}^n : \text{Es existiert eine Folge } \{x_k\} \subset C \text{ mit } x = \lim_{k\to\infty} x_k\}.$$

   Hinweis: Man überlege sich zunächst, dass mit konvexen Mengen A, B auch $B - A$ und der Abschluss $\operatorname{cl}(B - A)$ konvex sind.

5. 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.

6. Man beweise den folgenden Satz von Fan-Glicksburg-Hoffman (siehe z. B. O. L. Mangasarian (1969, S. 63)):

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

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

   keine Lösung, so besitzt

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

   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.

3.2 Notwendige und hinreichende Optimalitätsbedingungen

3.2.1 Notwendige Optimalitätsbedingungen

Wir betrachten eine Optimierungsaufgabe der Form

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

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

Der folgende Satz gibt notwendige Optimalitätsbedingungen erster Ordnung bei einer *linear restringierten* Optimierungsaufgabe an.

Satz 2.1 *Gegeben sei die linear restringierte Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad \textit{auf} \quad M := \{x \in \mathbb{R}^n : Ax \le b,\ A_0 x = b_0\}.$$

Hierbei seien $A \in \mathbb{R}^{l\times n}$, $A_0 \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^l$ und $b_0 \in \mathbb{R}^m$ gegeben. Ist $x^* \in M$ eine lokale Lösung von (P) und ist die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ in x^* stetig differenzierbar, so existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$, sogenannte Lagrange-Multiplikatoren, mit

$$u^* \ge 0, \qquad \nabla f(x^*) + A^T u^* + A_0^T v^* = 0, \qquad (b - Ax^*)^T u^* = 0.$$

Beweis: Wir definieren die Menge der in x^* *aktiven* Ungleichungsrestriktionen durch

$$I^* := \{i \in \{1, \dots, l\} : (Ax^*)_i = (b)_i\}.$$

Der Kegel $F(M; x^*)$ der zulässigen Richtungen an M in x^* kann leicht angegeben werden, es ist nämlich

$$F(M; x^*) = \{p \in \mathbb{R}^n : (Ap)_i \le 0,\ i \in I^*,\ A_0 p = 0\}.$$

Es bezeichne $A_{I^*} \in \mathbb{R}^{|I^*|\times n}$ die Untermatrix von A, die nur zu I^* gehörende Zeilen enthält. Da $x^* \in M$ eine lokale Lösung von (P) ist, ist (siehe grundlegende Beobachtung zu Beginn des Kapitels) $\nabla f(x^*)^T p \ge 0$ für alle $p \in F(M; x^*)$ bzw. das System

$$\nabla f(x^*)^T p < 0, \qquad A_{I^*} p \le 0, \qquad A_0 p = 0$$

nicht lösbar. Dies ist äquivalent dazu, dass

$$\begin{pmatrix} -A_{I^*} \\ -A_0 \\ A_0 \end{pmatrix} p \ge 0, \qquad \nabla f(x^*)^T p < 0$$

nicht lösbar ist. Das Farkas-Lemma liefert die Existenz einer Lösung $(u_{I^*}, v_+, v_-) \in \mathbb{R}^{|I^*|} \times \mathbb{R}^m \times \mathbb{R}^m$ von

$$-A_{I^*}^T u_{I^*} - A_0^T (v_+ - v_-) = \begin{pmatrix} -A_{I^*} \\ -A_0 \\ A_0 \end{pmatrix}^T \begin{pmatrix} u_{I^*} \\ v_+ \\ v_- \end{pmatrix} = \nabla f(x^*), \qquad \begin{pmatrix} u_{I^*} \\ v_+ \\ v_- \end{pmatrix} \ge 0.$$

Definiert man $u^* = (u_i^*) \in \mathbb{R}^l$ durch

$$u_i^* := \begin{cases} u_i, & \text{falls } i \in I^*, \\ 0, & \text{sonst,} \end{cases} \qquad i = 1, \dots, l,$$

und $v^* \in \mathbb{R}^m$ durch $v^* := v_+ - v_-$, so hat man in $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ ein gesuchtes Paar gefunden. □

Beispiel: Wir untersuchen das Problem, siehe D. G. Luenberger (1969, S. 203–205), eine "optimale" Gewinnwette (also keine Kombinationswette wie eine Zweier- oder Dreierwette) über insgesamt $K \in$ in einem Pferderennen mit n Pferden abzuschließen,

wobei $K > 0$. Für $j = 1, \ldots, n$ seien die Wahrscheinlichkeit $p_j \in (0,1)$ dafür, dass das j-te Pferd gewinnt, und der Betrag $s_j > 0$, den alle anderen Wetter zusammen auf das j-te Pferd setzen, bekannt. Ferner sei $c \in (0,1)$ eine Konstante, die besagt, dass der Rennveranstalter und der Staat das $(1-c)$-Fache des Gesamtwettbetrages einbehalten. Der Rest wird im Verhältnis zu dem Betrag verteilt, der auf das siegreiche Wette gesetzt wurde. Gefragt wird nach dem optimalen Wettbetrag x_j^* für das j-te Pferd, $j = 1, \ldots, n$. Eine Wette besteht aus einem Vektor $x = (x_1, \ldots, x_n)^T$, dessen j-te Komponente angibt, wieviel auf das j-te Pferd gewettet wird. Sie ist zulässig, wenn $x \geq 0$ und $e^T x = K$ (hier ist e der Vektor dessen Komponenten sämtlich gleich 1 sind). Der Gesamtwettbetrag für alle Pferde des Rennens ist $K + \sum_{j=1}^n s_j$ und daher $c(K + \sum_{j=1}^n s_j)$ der auszuschüttende Gesamtbetrag. Folglich ist $c(K + \sum_{j=1}^n s_j) x_j/(s_j + x_j)$ der Betrag, den man gewinnt, wenn das j-te Pferd gewinnt. Der zu erwartende Gewinn zur Wette x ist also

$$R(x) := c\Big(K + \sum_{j=1}^n s_j\Big) \sum_{j=1}^n \frac{p_j x_j}{s_j + x_j} - K.$$

Das Problem, eine optimale Wette zu finden, besteht also darin, $R(\cdot)$ unter den Nebenbedingungen $x \geq 0$, $e^T x = K$ zu maximieren, bzw. die Aufgabe

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

zu lösen. Da M kompakt ist, besitzt (P) eine Lösung $x^* \in M$. Wegen Satz 2.1 existiert ein Paar $(u^*, v^*) \in \mathbb{R}^n \times \mathbb{R}$ mit

$$u^* \geq 0, \qquad \nabla f(x^*) - u^* + v^* e = 0, \qquad (x^*)^T u^* = 0.$$

Wegen $(\nabla f(x^*))_j = -p_j s_j/(s_j + x_j^*)^2$ existiert also ein $v^* \in \mathbb{R}$ derart, dass

$$\Big(-\frac{p_j s_j}{(s_j + x_j^*)^2} + v^*\Big) \geq 0, \qquad x_j^* \Big(-\frac{p_j s_j}{(s_j + x_j^*)^2} + v^*\Big) = 0 \qquad (j = 1, \ldots, n).$$

Hieraus schließen wir, dass v^* positiv ist und

$$x_j^* = \begin{cases} 0, & \text{falls } p_j/s_j \leq v^*, \\ \sqrt{p_j s_j / v^*} - s_j, & \text{falls } p_j/s_j > v^*. \end{cases}$$

Der noch unbekannte Multiplikator v^* ist zu bestimmen aus der Gleichung

$$e^T x^* = \sum_{j=1}^n x_j^* = \sum_{j : p_j/s_j > v^*} \Big(\sqrt{\frac{p_j s_j}{v^*}} - s_j\Big) = K.$$

Daher muss man $v^* > 0$ als Nullstelle der auf $(0, \infty)$ definierten Funktion

$$h(v) := \sum_{j : p_j/s_j > v} \Big(\sqrt{\frac{p_j s_j}{v}} - s_j\Big) - K$$

bestimmen. Es ist $\lim_{v\to 0+} h(v) = +\infty$, $h(v) = -K < 0$ für $v \geq \max_{j=1,\ldots,n}(p_j/s_j) =: b$ und $h(\cdot)$ auf $(0, b)$ monoton fallend. Daher besitzt $h(\cdot)$ genau eine Nullstelle $v^* > 0$ in $(0, b)$.

Für spezielle Daten wollen wir die optimale Wette numerisch berechnen. Es sei $K = 100$, $n = 6$ und

j	1	2	3	4	5	6
p_j	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
s_j	2800	1500	1400	1900	2000	1800

In Abbildung 3.6 haben wir $h(\cdot)$ auf $[0.000085, 0.0001]$ aufgetragen. Man erkennt, dass

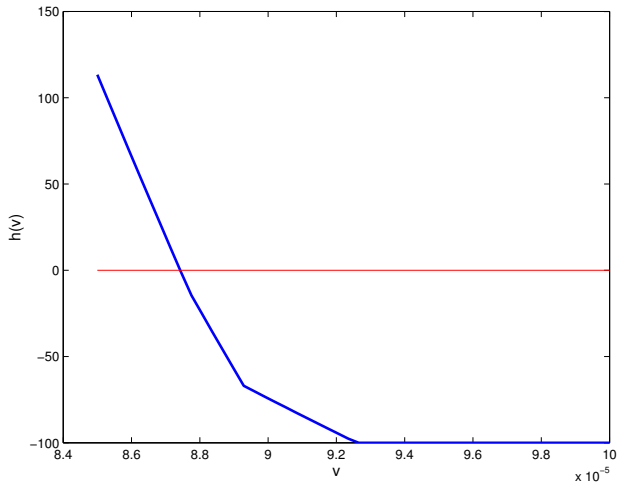


Abbildung 3.6: Die Funktion $h(\cdot)$ beim Wettproblem

die Nullstelle von $h(\cdot)$ etwa bei $8.7 \cdot 10^{-5}$ liegt. Mit dem Bisektionsverfahren erhalten wir $v^* = 8.7427 \cdot 10^{-5}$. Die optimale Wette ist

$$x^* = (29.61, 0, 14.80, 3.17, 0, 52.41)^T.$$

Rundungsfehler sind der Grund dafür, dass sich als Summe der einzelnen Wetteinsätze nicht genau 100 ergibt. Es wird bei diesem Wettproblem außer Acht gelassen, dass i. Allg. Wettbeträge nur in gewissen Stückelungen möglich sind. □

Beispiel: Notwendige Optimalitätsbedingungen können natürlich zum Nachweis dafür benutzt werden, dass ein gewisser zulässiger Punkt *keine* Lösung ist. Als Beispiel hierfür betrachten wir die Optimierungsaufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := x_1^2 - 6x_1 + x_2^3 - 3x_2 \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^2 : \begin{pmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\}. \end{array} \right.$$

Wir wollen nachweisen, dass $x^* := (\frac{1}{2}, \frac{1}{2})^T$ *keine* Lösung von (P) ist. Da nur die erste Restriktion aktiv ist, müsste andernfalls ein $u_1^* \geq 0$ existieren mit

$$0 = \nabla f(x^*) + u_1^* \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -5 \\ -\frac{9}{4} \end{pmatrix} + u_1^* \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

was natürlich nicht möglich ist.Daher ist $(\frac{1}{2}, \frac{1}{2})^T$ keine lokale Lösung von (P). In Abbildung 3.7 geben wir den zulässigen Bereich M und einige Höhenlinien an. An dieser

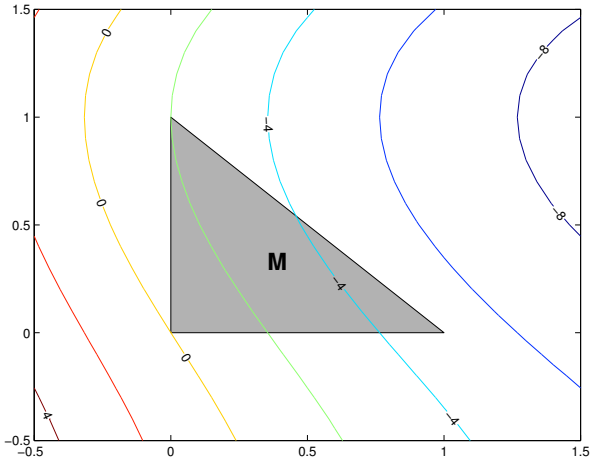


Abbildung 3.7: Zulässige Menge M und Höhenlinien

Abbildung erkennt man, dass $x^* = (1, 0)^T$ die Lösung von (P) ist. Wir wollen wenigstens nachweisen, dass x^* eine stationäre Lösung von (P) ist, also den notwendigen Optimalitätsbedingungen erster Ordnung genügt. Offenbar sind die erste und die dritte Restriktion aktiv. Daher sind nichtnegative u_1^* und u_3^* mit

$$u_1^* \begin{pmatrix} 1 \\ 1 \end{pmatrix} + u_3^* \begin{pmatrix} 0 \\ -1 \end{pmatrix} = -\nabla f(x^*) = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$$

zu bestimmen, was auf $u_1^* = 4$, $u_3^* = 1$ führt. □

Schwierigkeiten bei der Herleitung notwendiger Optimalitätsbedingungen machen vor allem nichtlineare Gleichungsrestriktionen. Um diesen Schwierigkeiten zunächst aus dem Weg zu gehen, betrachten wir im folgenden Satz Optimierungsaufgaben mit nichtlinearen Ungleichungsrestriktionen aber nach wie vor (affin) linearen Gleichungsrestriktionen.

Ist $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$, so wird mit $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ die i-te Komponentenabbildung bezeichnet, d. h. es ist $g(x) = (g_1(x), \ldots, g_l(x))^T$. Ist $g(\cdot)$ in x^* differenzierbar, so wird mit

$$g'(x^*) = \Big(\frac{\partial g_i}{\partial x_j}(x^*)\Big)_{\substack{i=1,\ldots,l \\ j=1,\ldots,n}} = \begin{pmatrix} \nabla g_1(x^*)^T \\ \vdots \\ \nabla g_l(x^*)^T \end{pmatrix} \in \mathbb{R}^{l \times n}$$

die *Funktionalmatrix* oder auch *Jacobi*[5]*-Matrix* (engl.: Jacobian) von g in x^* bezeichnet. Entsprechende Bezeichnungen werden natürlich auch für eine Abbildung $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ benutzt.

Satz 2.2 *Gegeben sei die Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad f(x) \quad \textit{auf} \quad M := \{x \in \mathbb{R}^n : g(x) \leq 0,\ A_0 x = b_0\}.$$

Hierbei seien $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $A_0 \in \mathbb{R}^{m\times n}$, $b_0 \in \mathbb{R}^m$, gegeben. Sei $x^* \in M$ eine lokale Lösung von (P) und $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ in x^* stetig differenzierbar. Ferner sei

$$\text{(CQ)} \qquad L_+(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p < 0,\ i \in I^*,\ A_0 p = 0\} \neq \emptyset,$$

wobei

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

die Menge der in x^* aktiven Ungleichungsrestriktionen bezeichnet. Dann existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$, sogenannte Lagrange[6]*-Multiplikatoren, mit*

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* + A_0^T v^* = 0, \qquad g(x^*)^T u^* = 0.$$

Beweis: Wir wollen uns überlegen, dass

$$L_0(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p \leq 0,\ i \in I^*,\ A_0 p = 0\} \subset T(M; x^*).$$

Da das System

$$\nabla f(x^*)^T p < 0, \qquad p \in L_0(M; x^*)$$

keine Lösung besitzt, folgt dann die Behauptung genau wie im letzten Satz.

Offenbar ist $L_+(M; x^*) \subset F(M; x^*)$. Sei $p \in L_0(M; x^*)$ und $\hat{p} \in L_+(M; x^*)$. Für alle $t \in (0, 1]$ ist

$$(1 - t)p + t\hat{p} \in L_+(M; x^*) \subset F(M; x^*)$$

und daher

$$p = \lim_{t\to 0+} ((1 - t)p + t\hat{p}) \in \text{cl } F(M; x^*) \subset T(M; x^*),$$

wobei wir am Schluss die Abgeschlossenheit des Tangentialkegels benutzt haben. □

Beispiel: Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{l} \text{Minimiere} \quad f(x) := -x_1 \quad \text{unter den Nebenbedingungen} \\ \qquad g(x) := \begin{pmatrix} (x_1 - 1)^2 + (x_2 - \frac{1}{2})^2 - 2 \\ (x_1 - 1)^2 + (x_2 + \frac{1}{2})^2 - 2 \end{pmatrix} \leq 0. \end{array} \right.$$

In Abbildung 3.8 links stellen wir den zulässigen Bereich dar. Die Lösung von (P) ist

[5]Carl Gustav Jacob Jacobi (1804–1851) war ein deutscher Mathematiker. Nach ihm benannt sind die Jacobimatrix, die Jacobi-Polynome, das Jacobi-Verfahren, die Jacobi-Identität, das Jacobi-Symbol und ein Mondkrater.

[6]Joseph Louis Lagrange (1736–1813) war ein italienischer Mathematiker und Astronom. Lagrange wurde als Giuseppe Luigi Lagrangia geboren. In der Analysis ist die Lagrangesche Darstellung des Restgliedes der Taylor-Formel oder auch die Lagrangesche Multiplikatorenregel bekannt. Lagrange ist namentlich auf dem Eiffelturm verewigt.

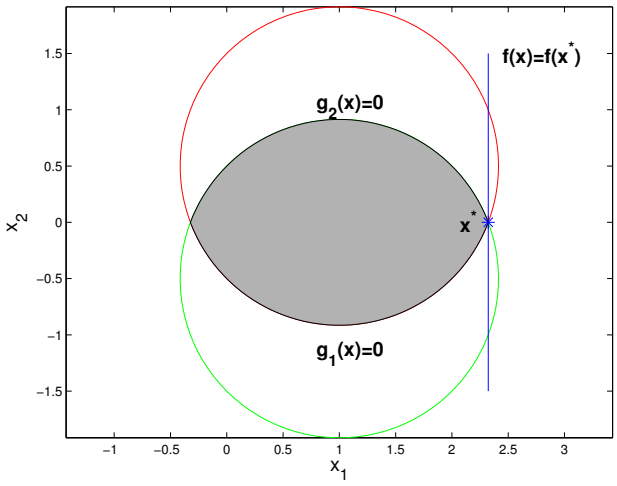


Abbildung 3.8: Ein nichtlinear restringiertes Optimierungsproblem

offenbar $x^* = (1 + \frac{1}{2}\sqrt{7}, 0)$. Das liest man aus Abbildung 3.8 ab. Beide Ungleichungen sind in x^* aktiv. □

Bemerkung: Sieht man einmal von den linearen Gleichungsnebenbedingungen ab, so besagt Satz 2.2, dass das Negative des Gradienten der Zielfunktion in einer lokalen Lösung x^* sich als nichtnegative Linearkombination derjenigen Gradienten der Ungleichungsnebenbedingungen schreiben lässt, die zu in x^* aktiven Restriktionen gehören. Man überprüfe diese Aussage für die Optimierungsaufgabe in Abbildung 3.8! □

Bemerkung: Die Zusatzbedingung (CQ) in Satz 2.2 nennt man eine *Constraint Qualification*. Ohne eine solche Zusatzbedingung ist die Aussage bei nichtlinear restringierten Problemen i. Allg. nicht richtig. Hierzu betrachte man die Aufgabe

$$\text{Minimiere} \quad f(x) := -x_1 \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^2 : g(x) := \begin{pmatrix} -x_1 \\ -x_2 \\ x_2 - (1-x_1)^3 \end{pmatrix} \leq 0 \right\},$$

die offenbar die eindeutige Lösung $x^* = (1, 0)^T$ besitzt. In Abbildung 3.9 stellen wir die Menge M der zulässigen Lösungen dar. Offenbar ist $I^* = \{2, 3\}$ die Menge der in x^* aktiven Restriktionen. Es ist

$$\nabla f(x^*) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \qquad \nabla g_2(x^*) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \qquad \nabla g_3(x^*) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Daher ist $-\nabla f(x^*)$ *keine* nichtnegative Linearkombination von $\nabla g_2(x^*)$, $\nabla g_3(x^*)$. Dies liegt natürlich daran, dass die Constraint Qualification (CQ) in Satz 2.2 nicht erfüllt ist. □

Nur mit einer Beweisidee geben wir notwendige Optimalitätsbedingungen auch für nichtlineare Gleichungsnebenbedingungen an. Der folgende Satz (und auch die letzten beiden Sätze) sind nach Karush-Kuhn-Tucker (KKT) benannt.

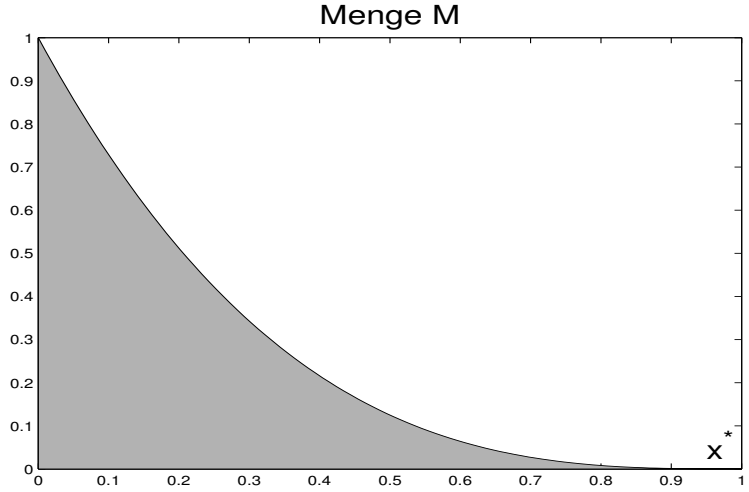


Abbildung 3.9: Eine Constraint Qualification ist nötig!

Satz 2.3 *Sei x^* eine lokale Lösung von*

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

Die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildungen $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sowie $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ seien auf einer Umgebung von x^* stetig differenzierbar. Mit $I^* := \{i \in \{1, \ldots, l\} : g_i(x^*) = 0\}$ werde die Indexmenge der in x^* aktiven Ungleichungsrestriktionen bezeichnet. Es sei

$$L_+(M; x^*) := \{p \in \mathbb{R}^n : \nabla g_i(x^*)^T p < 0,\ i \in I^*,\ h'(x^*)p = 0\} \neq \emptyset,$$

ferner sei Rang $h'(x^*) = m$. *Dann existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$, sogenannte Lagrange-Multiplikatoren, mit*

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

Beweisidee: Entsprechend Satz 2.2 bzw. dessen Beweis definiere man

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

Mit Hilfe des Satzes über implizite Funktionen zeige man:

- Zu vorgegebenem $p \in L_+(M; x^*)$ existieren ein $\epsilon > 0$ und eine Abbildung $x : (-\epsilon, \epsilon) \longrightarrow \mathbb{R}^n$ mit $x(t) \in M$ für alle $t \in (0, \epsilon)$ und $\lim_{t \to 0}[x(t) - x^*]/t = p$.

Ist dies gelungen, 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 $r(t) = o(t)$ bzw. $r(t)/t \to 0$ mit $t \to 0+$. Wegen der Abgeschlossenheit von Tangentialkegeln ist

$$L_0(M; x^*) \subset \mathrm{cl}\, L_+(M; x^*) \subset T(M; x^*)$$

und man erhält die Behauptung mit Hilfe des Farkas-Lemmas wie im Beweis von Satz 2.1. □

Bemerkungen: Die Zusatzbedingung in Satz 2.3, also

$$L_+(M;x^*) \neq \emptyset, \qquad \text{Rang } h'(x^*) = m,$$

nennt man eine *Constraint Qualification* oder auch *Regularitätsbedingung*. Sie wird gelegentlich nach Arrow-Hurwicz-Uzawa benannt.

Die Bedingung $g(x^*)^T u^* = 0$ in den Sätzen 2.2, 2.3 bzw. die Bedingung $(b - Ax^*)^T u^* = 0$ in Satz 2.1 heißt *Komplementaritätsbedingung* oder *Gleichgewichtsbedingung*. Sie besagt: Ist eine Ungleichungsrestriktion in einer lokalen Lösung inaktiv, so verschwindet der entsprechende Lagrange-Multiplikator. *Strikte Komplementarität* liegt vor, wenn außerdem komponentenweise $-g(x^*) + u^* > 0$, die zu aktiven Ungleichungsrestriktionen gehörenden Lagrange-Multiplikatoren also positiv sind. □

Definition 2.4 Ein zulässiger Punkt $x^* \in M$, in dem die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind, also ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

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

existiert, heißt eine *stationäre* oder *kritische* Lösung der zugehörigen Optimierungsaufgabe (P). Eine stationäre Lösung x^* zusammen mit den Lagrange-Multiplikatoren (u^*, v^*) heißt ein *Kuhn-Tucker (KT)* oder auch *Karush-Kuhn-Tucker (KKT)* Tripel.

Bei notwendigen Optimalitätsbedingungen zweiter Ordnung wollen wir uns auf linear restringierte Optimierungsaufgaben beschränken.

Satz 2.5 *Gegeben sei die Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad \textit{auf} \quad M := \{x \in \mathbb{R}^n : Ax \leq b,\ A_0 x = b_0\},$$

wobei $A \in \mathbb{R}^{l \times n}$, $A_0 \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^l$ und $b_0 \in \mathbb{R}^m$. 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

$$u^* \geq 0, \qquad \nabla f(x^*) + A^T u^* + A_0^T v^* = 0, \qquad (b - Ax^*)^T u^* = 0$$

und der Eigenschaft, dass $\nabla^2 f(x^*)$ positiv semidefinit auf

$$L^0(M;x^*) := \left\{ p \in \mathbb{R}^n : \begin{array}{l} a_i^T p = 0,\ i \in I_+^*, \\ a_i^T p \leq 0,\ i \in I^* \setminus I_+^*, \end{array} \ A_0 p = 0 \right\}$$

ist, wobei a_i^T, die i-te Zeile von A bezeichnet, $i = 1, \ldots, l$, I^* die Indexmenge der in x^* aktiven Ungleichungsrestriktionen ist und $I_+^* := \{i \in \{1, \ldots, l\} : u_i^* > 0\}$.

Beweis: Wegen Satz 2.1 existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u^* \geq 0, \qquad \nabla f(x^*) + A^T u^* + A_0^T v^* = 0, \qquad (b - Ax^*)^T u^* = 0.$$

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. Offenbar ist $x^* + tp \in M$ für alle hinreichend kleinen $t > 0$. Für diese t ist mit einem $\theta(t) \in (0, 1)$:

$$\begin{aligned}
0 &\leq \frac{f(x^* + tp) - f(x^*)}{t} \\
&= \nabla f(x^*)^T p + \tfrac{1}{2} t p^T \nabla^2 f(x^* + \theta(t)tp)p \\
&= \Big(-\sum_{i \in I_+^*} u_i^* a_i - A_0^T v^*\Big)^T p + \tfrac{1}{2} t p^T \nabla^2 f(x^* + \theta(t)tp)p \\
&= -\sum_{i \in I_+^*} u_i^* \underbrace{a_i^T p}_{=0} - (v^*)^T \underbrace{A_0 p}_{=0} + \tfrac{1}{2} t p^T \nabla^2 f(x^* + \theta(t)tp)p \\
&= \tfrac{1}{2} t p^T \nabla^2 f(x^* + \theta(t)tp)p,
\end{aligned}$$

woraus nach Division durch t und Grenzübergang $t \to 0+$ die Behauptung folgt. □

3.2.2 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 2.6 *Gegeben sei die Optimierungsaufgabe*

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

Hierbei seien $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ *und* $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ *(komponentenweise) konvex sowie* $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *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*

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

d. h. ist (x^*, u^*, v^*) *ein Kuhn-Tucker-Tripel, 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 bewiesen ist. □

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 $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ mit $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv definit in x^* ein isoliertes, lokales Minimum besitzt.

Satz 2.7 *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 $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ sowie die Restriktionsabbildungen $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ seien in $x^* \in M$ zweimal stetig differenzierbar. Mit I^* sei die Indexmenge der in x^* aktiven Ungleichungsrestriktionen bezeichnet. Es existiere ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit:

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

und

$$p^T \Big[\nabla^2 f(x^*) + \sum_{i=1}^{l} u_i^* \nabla^2 g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla^2 h_i(x^*)\Big] p > 0 \qquad \textit{für alle } p \in L^0(M; x^*) \setminus \{0\},$$

wobei

$$L^0(M; x^*) := \Big\{p \in \mathbb{R}^n : \begin{array}{l} \nabla g_i(x^*)^T p = 0,\ i \in I_+^*, \\ \nabla g_i(x^*)^T p \leq 0,\ i \in I^* \setminus I_+^*, \end{array} \quad h'(x^*)p = 0\Big\}$$

mit[7]

$$I_+^* := \{i \in \{1, \ldots, l\} : u_i^* > 0\}.$$

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 $f(x_k) \leq f(x^*)$ 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

$$\nabla f(x^*)^T p \leq 0, \qquad \nabla g_i(x^*)^T p \leq 0 \quad (i \in I^*), \qquad h'(x^*)p = 0.$$

Wir unterscheiden zwei Fälle und zeigen, dass sich jeweils ein Widerspruch ergibt.

[7] Man beachte, dass I_+^* offenbar eine Teilmenge von I^*, der Indexmenge der in x^* aktiven Ungleichungsrestriktionen, ist.

Angenommen, es ist $\nabla g_i(x^*)^T p < 0$ für wenigstens ein $i \in I_+^*$. Dann ist

$$0 \geq \nabla f(x^*)^T p = -\sum_{i \in I_+^*} \underbrace{u_i^*}_{>0} \underbrace{\nabla g_i(x^*)^T p}_{\leq 0} - (v^*)^T \underbrace{h'(x^*)p}_{=0} > 0,$$

ein Widerspruch.

Sei $\nabla g_i(x^*)^T p = 0$ für alle $i \in I_+^*$. Durch eine Entwicklung nach Taylor erhält man

$$\begin{aligned}
0 \;\geq\; f(x^* + t_k p_k) - f(x^*) &= t_k \nabla f(x^*)^T p_k + \tfrac{1}{2}\, t_k^2 p_k^T \nabla^2 f(x_k^{(0)}) p_k, \\
0 \;\geq\; g_i(x^* + t_k p_k) &= g_i(x^*) + t_k \nabla g_i(x^*)^T p_k + \tfrac{1}{2}\, t_k^2 p_k^T \nabla^2 g_i(x_{i,k}^{(1)}) p_k, \\
0 \;=\; h_i(x^* + t_k p_k) &= \underbrace{h_i(x^*)}_{=0} + t_k \nabla h_i(x^*)^T p_k + \tfrac{1}{2}\, t_k^2 p_k^T \nabla^2 h_i(x_{i,k}^{(2)}) p_k.
\end{aligned}$$

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 $g(x^*)^T u^* = 0$ 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: Wir kommen auf ein Beispiel aus der Einführung zurück. Um ein Blättern zu vermeiden, wiederholen wir es hier: 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. Der Boden ($x_1 x_2$ m^2) und die beiden Seiten ($2x_1 x_3$ m^2) müssen aus einem Material hergestellt sein, das zwar nichts kostet, von dem aber nur 4 m^2 zur Verfügung steht. Das Material für die beiden Enden ($2x_2 x_3$ m^2) kostet 200 € pro m^2. Ein Transport der Box kostet 1 €. Wie hat man die Box zu konstruieren?

Wie wir uns schon in der Einführung überlegten, haben wir die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := \dfrac{1}{x_1 x_2 x_3} + x_2 x_3 \quad \text{unter den Nebenbedingungen} \\ g(x) := x_1 x_2 + 2x_1 x_3 - 4 \leq 0, \qquad x_1, x_2, x_3 > 0, \end{cases}$$

zu lösen. Wir wollen den Satz 2.2 von Kuhn-Tucker anwenden und nehmen hierzu an, x^* sei eine lokale Lösung von (P). Die Constraint Qualification (CQ) ist erfüllt. Daher existiert ein $u^* \in \mathbb{R}$ mit

$$u^* \geq 0, \quad \begin{pmatrix} -1/((x_1^*)^2 x_2^* x_3^*) \\ -1/(x_1^*(x_2^*)^2 x_3^*) + x_3^* \\ -1/(x_1^* x_2^*(x_3^*)^2) + x_2^* \end{pmatrix} + u^* \begin{pmatrix} x_2^* + 2x_3^* \\ x_1^* \\ 2x_1^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

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. □

3.2.3 Aufgaben

1. Man löse das folgende, auf S. Lhuillier (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.

2. Man zeige, dass $x^* := (1, 1, 2)^T$ die Lösung von

$$(\mathrm{P})\quad \left\{\begin{array}{l} \text{Minimiere} \quad -5x_2 + \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \quad \text{unter den Nebenbedingungen} \\ \begin{array}{rrrcr} -4x_1 & -3x_2 & & \geq & -8 \\ 2x_1 & +\ x_2 & & \geq & 2 \\ & -2x_2 & +x_3 & \geq & 0 \\ x_1 & -2x_2 & +x_3 & = & 1 \end{array} \end{array}\right.$$

ist.

3. Für die Aufgabe

$$(\mathrm{P}) \quad \left\{\begin{array}{ll} \text{Minimiere} & f(x) := x_1^2 + 4x_2^2 + 16x_3^2 \quad \text{unter der Nebenbedingung} \\ & h(x) := x_1x_2x_3 - 1 = 0 \end{array}\right.$$

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.

4. Gegeben sei die Optimierungsaufgabe

$$(\mathrm{P}) \quad \left\{\begin{array}{ll} \text{Minimiere} & f(x) := -(x_1x_2 + x_2x_3 + x_1x_3) \quad \text{u. d. NB.} \\ & h(x) := x_1 + x_2 + x_3 - 3 = 0. \end{array}\right.$$

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[8] sei die Optimierungsaufgabe

$$(\mathrm{P}_\gamma) \quad \left\{\begin{array}{l} \text{minimiere} \quad f(x) := -(x_1+1)^2 - (x_2+1)^2 \quad \text{auf} \\ M_\gamma := \left\{ x \in \mathbb{R}^2 : g(x) := \begin{pmatrix} x_1^2 + x_2^2 - 2 \\ x_1 - \gamma \end{pmatrix} \leq 0 \right\}, \end{array}\right.$$

wobei $\gamma \geq -\sqrt{2}$ fest vorgegeben sei.

[8] Diese Aufgabe findet man bei C. Geiger, C. Kanzow (2002, S. 72).

(a) Man ermittle anhand einer Skizze die Lösung $x^* = x^*(\gamma)$ von (P_γ), wobei man die Fälle $\gamma = -\sqrt{2}$, $-\sqrt{2} < \gamma \leq 1$ und $\gamma > 1$ unterscheide.

(b) Ist die Regularitätsbedingung (CQ) aus Satz 2.2 erfüllt?

(c) Gibt es zu x^* ein $u^* \in \mathbb{R}^2$, so dass (x^*, u^*) ein Kuhn-Tucker-Punkt zu (P_γ) ist, also

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* = 0, \qquad g(x^*)^T u^* = 0$$

gilt?

6. Als Hoffman-Theorem (siehe A. J. Hoffman (1952)) 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 $(y_+)_i = \max(y_i, 0)$, $i = 1, \ldots, l$.

   Sei

$$P := \{x \in \mathbb{R}^n : Ax \leq b,\ Cx = d\} \neq \emptyset,$$

   wobei $A \in \mathbb{R}^{l \times n}$, $b \in \mathbb{R}^l$, $C \in \mathbb{R}^{m \times n}$, $d \in \mathbb{R}^m$. Dann existiert eine (von b und d unabhängige) 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.

7. 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.

8. Mit Hilfe des Hoffman-Theorems zeige man: Gegeben sei das lineare Programm

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

   Hierbei seien $A \in \mathbb{R}^{l \times n}$, $b \in \mathbb{R}^l$ und $c \in \mathbb{R}^n$. Es wird $M \neq \emptyset$ und $\inf(\mathrm{P}) > -\infty$ vorausgesetzt. Also ist die Menge M_{opt} der Lösungen von (P) nichtleer. Man zeige die Existenz einer Konstanten $c_0 = c_0(A, c) > 0$ derart, dass

$$\operatorname{dist}(x, M_{\mathrm{opt}}) \leq c_0[c^T x - \min(\mathrm{P})] \qquad \text{für alle } x \in M.$$

   Hinweis: Man beachte, dass $M_{\mathrm{opt}} = M \cap \{x \in \mathbb{R}^n : c^T x - \min(\mathrm{P}) = 0\}$.

3.3 Dualität bei konvexen Optimierungsaufgaben

3.3.1 Definition des dualen Programms, schwacher Dualitätssatz

Wir betrachten im folgenden eine Optimierungsaufgabe der Form

(P) $\qquad$ Minimiere $\quad f(x) \quad$ 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, $f : C \longrightarrow \mathbb{R}$ und $g : C \longrightarrow \mathbb{R}^l$ sind (komponentenweise) konvex, $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ist affin linear.

Unter der Voraussetzung (V) handelt es sich bei (P) um eine *konvexe Optimierungsaufgabe*, d. h. sowohl die Zielfunktion als auch die Menge der zulässigen Lösungen von (P) ist konvex. Statt von einer Optimierungsaufgabe sprechen wir hier auch von einem *Programm*.

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

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

definiert. Schließlich ist die zu (P) *Lagrange-duale* Optimierungsaufgabe 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 oder Ungleichungen 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 oder Ungleichungen als Restriktionen bezieht, ist irrevalent. □

Beispiel: Gegeben sei das lineare Programm (in sogenannter Normal- oder Standardform)

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

In der allgemeinen Formulierung eines konvexen Programms sei also $f(x) := c^T x$, $C := \mathbb{R}^n_{\geq 0}$ der sogenannte nichtnegative Orthant im $\mathbb{R}^n$, also die Menge aller (in allen Komponenten) nichtnegativen Vektoren des $\mathbb{R}^n$, und $h(x) := b - Ax$. Die zugehörige Lagrange-Funktion ist

$$L(x, u, v) = c^T x + (b - Ax)^T v = b^T v + (c - A^T v)^T x.$$

Daher ist

$$\phi(u, v) := \inf_{x \geq 0} L(x, u, v) > -\infty \iff c - A^T v \geq 0.$$

Das zu (P) duale lineare Programm ist daher

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

Auch diese Aufgabe ist ein konvexes Programm und kann dualisiert werden. Wegen

$$\inf_{v \in \mathbb{R}^n} [-b^T v + x^T (A^T v - c)] > -\infty \iff Ax = b$$

stimmt das zu (D) duale Programm genau mit dem Ausgangsprogramm überein. □

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

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

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. Wir schreiben $\max(\mathrm{D})$ statt $\sup(\mathrm{D})$, wenn (D) lösbar ist.

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

Satz 3.1 *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. □

3.3.2 Starker Dualitätssatz

Wir begnügen uns mit einem einzigen starken Dualitätssatz. Diesen werden wir später auf lineare Optimierungsaufgaben anwenden.

Satz 3.2 *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 (Beweis?). Wir zeigen nun, dass (D) zulässig und $\sup(\mathrm{D}) = \min(\mathrm{P})$ ist. 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 bzw. sein Korollar 1.4 sichert die Existenz eines Tripels (q^*, u^*, v^*) und einer Zahl $\gamma \in \mathbb{R}$ mit

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

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

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

hieraus $(u^*, v^*) \in N$, so dass (D) zulässig ist, und $\alpha < \phi(u^*, v^*) \le \sup(\mathrm{D})$. Da $\alpha < \min(\mathrm{P})$ beliebig ist, folgt $\min(\mathrm{P}) \le \sup(\mathrm{D})$. 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 bzw. seines Korollars 1.4 liefert die Existenz von (q^*, u^*, v^*) und $\gamma \in \mathbb{R}$ mit

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

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

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

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

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

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

3.3.3 Dualität in der linearen Optimierung

Bei einem linearen Programm[9] sind in der obigen Formulierung eines konvexen Programms die Zielfunktion linear, die Restriktionsabbildungen g und h jeweils affin linear, ferner können durch die Menge C Vorzeichenbedingungen an die Variablen beschieben werden. Ein lineares Programm ist in *Normal-* oder *Standardform*, wenn alle Variablen nichtnegativ sind und nur Gleichungen als weitere Restriktionen vorkommen, es also die Form

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

hat. Durch die Einführung von *Schlupfvariablen* und die Darstellung nicht vorzeichenbeschränkter Variablen als Differenz von nichtnegativen Variablen kann man jedes lineare Programm auf äquivalente Normalform bringen.

Beispiel: Beim Produktionsplanungsproblem will ein Betrieb unter Kapazitätsbeschränkungen an m benötigte Hilfsmittel seinen Gesamtgewinn bei der Herstellung von n Produkten maximieren. Als lineares Programm lautet es:

$$\text{Maximiere} \quad p^T x \quad \text{unter den Nebenbedingungen} \quad x \geq 0, \quad Ax \leq b.$$

Durch die Einführung einer Schlupfvariablen z kann man diese Aufgabe in äquivalenter Normalform

$$\begin{cases} \text{Minimiere} \quad \begin{pmatrix} -p \\ 0 \end{pmatrix}^T \begin{pmatrix} x \\ z \end{pmatrix} \quad \text{unter den Nebenbedingungen} \\ \qquad\qquad \begin{pmatrix} x \\ z \end{pmatrix} \geq 0, \quad \begin{pmatrix} A & I \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} = b \end{cases}$$

schreiben. Als duales Problem erhalten wir

$$\text{Maximiere} \quad b^T v \quad \text{unter den Nebenbedingungen} \quad \begin{pmatrix} A^T \\ I \end{pmatrix} v \leq \begin{pmatrix} -p \\ 0 \end{pmatrix}$$

bzw. nach der Variablentransformation $v \mapsto -y$ die Aufgabe

$$\text{Minimiere} \quad b^T y \quad \text{unter den Nebenbedingungen} \quad y \geq 0, \quad A^T y \geq p.$$

□

Daher 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 (siehe früheres Beispiel)

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

Das Dualisieren von (D) führt zum Ausgangsproblem (P) zurück, wie wir uns auch schon überlegt haben.

[9] Wir benutzen weiter oft das Wort *Programm* statt *Optimierungsaufgabe*.

Satz 3.3 *Das lineare Programm (P) sei zulässig und* $\inf(\mathrm{P}) > -\infty$. *Dann besitzt (P) eine Lösung.*

Beweis: Wir wollen die Existenzaussage in Satz 3.2 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: Da man *jedes* lineare Programm in äquivalente Normalform überführen kann, gilt ganz allgemein für lineare Programme: Ist die Menge der zulässigen Lösungen nichtleer und die (lineare) Zielfunktion auf ihr nach unten beschränkt, so besitzt das zugehörige lineare Programm eine Lösung. □

Satz 3.4 *Gegeben sei das lineare Programm*

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

und das dazu duale lineare Programm

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

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 3.2 (die Menge Λ ist abgeschlossen) ist $\max(\mathrm{D}) = \min(\mathrm{P})$. Auch die beiden weiteren Aussagen folgen direkt aus dem starken Dualitätssatz 3.2. □

Bemerkung: Gegeben sei das lineare Programm

$$\text{(P)} \qquad \text{Minimiere} \quad c^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Ax \leq b,\ A_0 x = b_0\},$$

wobei die Daten die üblichen Dimensionen haben. Mit der Lagrange-Funktion

$$L(x, u, v) := c^T x + u^T(Ax - b) + v^T(A_0 x - b_0)$$

erhält man als zugehöriges duales Problem die Aufgabe

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad -b^T u - b_0^T v \quad \text{auf} \\ N := \{(u,v) \in \mathbb{R}^l \times \mathbb{R}^m : u \geq 0,\ c + A^T u + A_0^T v = 0\}. \end{cases}$$

Ist nun $x^* \in M$ eine Lösung von (P), so sagt Satz 2.1 aus, dass ein Paar $(u^*, v^*) \in N$ mit $(b - Ax^*)^T u^* = 0$ existiert. Dies bedeutet aber, dass

$$c^T x^* = -(Ax^*)^T u^* - (A_0 x^*)^T v^* = -b^T u^* - b_0^T v^*.$$

Dies impliziert, dass (u^*, v^*) eine Lösung von (D) ist. Die Lagrange-Multiplikatoren im ersten Kuhn-Tucker-Satz 2.1 ergeben also eine Lösung des dualen Programms (D). □

Bemerkung: Für lineare Programme kann man zeigen, 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 $x^* \in M_{\text{opt}}$ und $y^* \in N_{\text{opt}}$, so ist

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

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. Man kann aber zeigen, dass ein Paar $(x^*, y^*) \in M_{\text{opt}} \times N_{\text{opt}}$ mit $x^* + c - A^T y^* > 0$. existiert. Siehe auch A. Schrijwer (1986, S. 95).

□

3.3.4 Quadratisch restringierte quadratische Programme

Wir werden uns sehr kurz halten, Näheres findet man unter `http://www.num.math.uni-goettingen.de/werner/opti.pdf`. Gegeben sei eine Aufgabe der Form

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

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

Der folgende Satz ist ein Existenzsatz, der vollständig dem Existenzsatz der linearen Optimierung entspricht. Er stammt von E. L. Peterson, J. G. Ecker (1969, 1970).

Satz 3.5 *Das konvexe, quadratisch restringierte quadratische Programm (P) sei zulässig, ferner sei* $\inf(\text{P}) > -\infty$. *Dann besitzt (P) eine Lösung.*

Bemerkungen: Die Existenz einer Lösung von (P) ist trivial, wenn (P) zulässig bzw. $M \neq \emptyset$ ist und mindestens eine der Matrizen $Q_0, \ldots, Q_l$ sogar positiv definit ist. Denn ist Q_0 positiv definit, so ist jede Niveaumenge zu (P) kompakt, während aus der positiven Definitheit eines Q_i mit $i \in \{1, \ldots, l\}$ die Kompaktheit von M folgt.

Als Spezialfall von Satz 3.5 erhält man: Ist eine konvexe, quadratische Funktion auf einem nichtleeren *Polyeder* (das ist der Durchschnitt von endlich vielen Halbräumen) nach unten beschränkt, so nimmt sie auf diesem Polyeder ihr Minimum an. Dieses Ergebnis, der Existenzsatz der quadratischen Optimierung, wurde zuerst von E. Barankin, R. Dorfman (1958) bewiesen. □

Man kann den starken Dualitätssatz 3.2 auf konvexe, quadratisch restringierte quadratische Programme anwenden und erhält einen entsprechenden starken Dualitätssatz. Hierauf wollen wir aber nicht mehr eingehen.

3.3.5 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. Gegeben sei das quadratische Programm

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

   wobei $A \in \mathbb{R}^{l \times n}$, $b \in \mathbb{R}^l$, $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit ist. Es wird vorausgesetzt, dass (P) zulässig, also $M \neq \emptyset$ ist.

   (a) Man begründe, weshalb (P) eine eindeutige Lösung $x^* \in M$ besitzt.

   (b) Man stelle das zu (P) duale Programm (D) auf und zeige, dass dieses eine Lösung u^* besitzt und $\min(\mathrm{P}) = \max(\mathrm{D})$ gilt.

   (c) Man zeige: Ist u^* eine Lösung von (D), so ist $x^* := -Q^{-1}(c + A^T u^*)$ die Lösung von (P).

   Hinweis: Zum Nachweis von 3b kann man zeigen, dass ein Lagrange-Multiplikator u^* zur Lösung x^* von (P) eine Lösung von (D) ist.

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.

Hinweis: Zum Nachweis von $N_{\mathrm{opt}} \neq \emptyset$ definiere man

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

zeige, dass Λ_+ konvex (und nichtleer) ist und $(\inf(\mathrm{P}), 0) \notin \Lambda_+$ gilt. Anschließend wende man den Trennungssatz (Satz 1.7) für konvexe Mengen an. Hiernach existiert ein Paar $(q^*, u^*) \in \mathbb{R} \times \mathbb{R}^l \setminus \{(0,0)\}$ mit

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

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

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

Wegen der Slaterschen Constraint Qualification ist $u^* = 0$, ein Widerspruch zu $(q^*, u^*) \neq (0,0)$. O. B. d. A. können wir dann $q^* = 1$ annehmen und haben

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

Also ist $u^* \in N$ dual zulässig und $\inf(\mathrm{P}) \leq \phi(u^*)$. Aus dem schwachen Dualitätssatz (Satz 3.1) $u^* \in N_{\mathrm{opt}}$ und $\max(\mathrm{D}) = \inf(\mathrm{P})$.

5. Seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ gegeben. Die Aufgabe

(P) Minimiere $f(x) := \|Ax - b\|_\infty, \quad x \in \mathbb{R}^n,$

nennt man das *diskrete, lineare Tschebyscheffsche Approximationsproblem*. Hierbei ist $\|\cdot\|_\infty$ die Maximum- (oder auch Tschebyscheff-) Norm auf dem $\mathbb{R}^m$, also $\|y\|_\infty := \max_{i=1,\ldots,m} |y_i|$. Der Aufgabe (P) ordne man die lineare Optimierungsaufgabe

(Q) Minimiere $g(x,\delta) := \delta$ auf $M := \{(x,\delta) \in \mathbb{R}^n \times \mathbb{R} : -\delta e \leq Ax - b \leq \delta e\}$

zu, wobei e der Vektor des $\mathbb{R}^m$ ist, dessen Komponenten sämtlich gleich 1 sind.

(a) Man zeige: Ist $x^* \in \mathbb{R}^n$ eine Lösung von (P), so ist $(x^*, \|Ax^* - b\|_\infty)$ eine Lösung von (Q). Ist umgekehrt $(x^*, \delta^*) \in M$ eine Lösung von (Q), so ist x^* eine Lösung von (P) und $\delta^* = \|Ax^* - b\|_\infty$.

(b) Man begründe, weshalb (Q) und damit auch (P) mindestens eine Lösung besitzt.

(c) Man führe die lineare Optimierungsaufgabe (Q) in Standardform über und berechne die hierzu duale Optimierungsaufgabe.

Kapitel 4

Innere-Punkt-Verfahren bei linearen Optimierungsaufgaben

Seit der Aufsehen erregenden Arbeit von N. KARMARKAR (1984) sind Innere-Punkt-Verfahren, insbesondere bei linearen Optimierungsproblemen, außerordentlich gründlich untersucht worden. Inzwischen bilden diese Verfahren eine ernste Konkurrenz für das Simplexverfahren (auf das wir nicht eingehen werden). In der Optimization Toolbox von MATLAB gibt es zur Lösung linearer Programme die Funktion `linprog`. Für hochdimensionale (large-scale) Probleme ist dort ein sogenanntes primal-duales Innere-Punkt-Verfahren implementiert. Wir werden versuchen, die theoretischen Grundlagen von Innere-Punkt-Verfahren bei linearen Optimierungsaufgaben zu legen und etwas zur Implementation zu sagen, wobei wir uns kurz fassen werden.

4.1 Grundlagen

4.1.1 Kompaktheit von Lösungsmengen

Gegeben sei das lineare Programm in Normalform

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

Wie üblich seien hierbei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$. Das zu (P) duale lineare Programm ist

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

I. Allg. werden wir im folgenden voraussetzen, dass

(A1) Es ist $M_0 := \{x \in \mathbb{R}^n : x > 0,\ Ax = b\} \neq \emptyset$.

(A2) Es ist $N_0 := \{y \in \mathbb{R}^m : A^T y < c\} \neq \emptyset$.

(A3) Es ist Rang $(A) = m$.

Dann gilt:

Lemma 1.1 *Sind die Voraussetzungen (A1), (A2) und (A3) erfüllt, so sind die Mengen M_{opt} bzw. N_{opt} der Lösungen von (P) bzw. (D) nichtleer und kompakt.*

Beweis: Da insbesondere wegen (A1) und (A2) die Probleme (P) und (D) zulässig sind, folgt aus dem starken Dualitätssatz der linearen Optimierung, dass (P) und (D) lösbar sind, bzw. $M_{\text{opt}} \neq \emptyset$ und $N_{\text{opt}} \neq \emptyset$. Da die Lösungsmengen trivialerweise abgeschlossen sind, bleibt ihre Beschränktheit zu zeigen.

Angenommen,

$$M_{\text{opt}} = \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b,\ c^T x = \min\,(\text{P})\}$$

wäre nicht beschränkt. Dann existiert eine Folge $\{x_k\} \subset M_{\text{opt}}$ mit $\|x_k\| \to \infty$. Man definiere $p_k := x_k/\|x_k\|$ und beachte, dass man wegen der Kompaktheit der Einheitssphäre aus $\{p_k\}$ eine gegen ein $p \neq 0$ konvergente Teilfolge auswählen kann. Notwendigerweise ist offenbar

$$p \geq 0, \quad Ap = 0, \quad c^T p = 0.$$

Nach Voraussetzung (A2) existiert ein $y_0 \in N_0$. Aus

$$0 < p^T(c - A^T y_0) = \underbrace{c^T p}_{=0} - \underbrace{(Ap)^T y_0}_{=0} = 0$$

erhalten wir einen Widerspruch.

Nun nehmen wir an,

$$N_{\text{opt}} = \{y \in \mathbb{R}^m : A^T y \leq c,\ b^T y = \max\,(\text{D})\}$$

sei unbeschränkt. Analog der eben angegebenen Argumentation existiert ein $q \neq 0$ mit

$$A^T q \leq 0, \quad b^T q = 0.$$

Dann ist $A^T q = 0$, da andernfalls mit einem nach Voraussetzung (A1) existierenden $x_0 \in M_0$ die Ungleichung

$$0 > x_0^T(A^T q) = (Ax_0)^T q = b^T q = 0$$

gelten würde, ein Widerspruch. Wegen Voraussetzung (A3) folgt aus $A^T q = 0$, dass $q = 0$, ein Widerspruch. Damit ist das Lemma bewiesen. □

4.1.2 Logarithmische Barrieren, zentraler Pfad

Mit $\mu > 0$ ordnen wir dem linearen Programm in Normalform (P) und seinem dualen Programm (D) die (im wesentlichen unrestringierten) Optimierungsaufgaben

$$(\text{P}_\mu) \qquad \text{Minimiere} \quad f_\mu(x) := c^T x - \mu \sum_{j=1}^{n} \log x_j, \quad x \in M_0$$

und

$$(\text{D}_\mu) \qquad \text{Maximiere} \quad g_\mu(y) := b^T y + \mu \sum_{j=1}^{n} \log(c - A^T y)_j, \quad y \in N_0$$

zu, wobei wir natürlich zumindestens (A1) und (A2) voraussetzen. Hierbei werden also der Zielfunktion von (P) bzw. (D) logarithmische Terme bzw. Barrieren hinzugefügt, die Zielfunktion f_μ von (P_μ) heißt *logarithmische Barriere-Funktion*.

Beispiel: Gegeben sei

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^2 : x \geq 0, \ \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 1 \right\} \end{cases}$$

und das hierzu duale Programm

$$\text{(D)} \qquad \text{Maximiere} \quad y \quad \text{auf} \quad N := \left\{ y \in \mathbb{R} : \begin{pmatrix} 1 \\ -1 \end{pmatrix} y \leq \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} = [0, 1].$$

Die Lösungen sind offenbar $x^* = (1, 0)^T$ bzw. $y^* = 1$. Als Lösung von (P_μ) erhält man

$$x_\mu = \begin{pmatrix} \mu + \frac{1}{2}(1 + \sqrt{4\mu^2 + 1}) \\ \mu - \frac{1}{2}(1 - \sqrt{4\mu^2 + 1}) \end{pmatrix},$$

die Lösung von (D_μ) ist

$$y_\mu = -\mu + \frac{1}{2}(1 + \sqrt{4\mu^2 + 1}).$$

In Abbildung 4.1 links haben wir die beiden Komponenten von x_μ, rechts y_μ für $\mu \in (0, 0.5]$ dargestellt. Offenbar ist $\lim_{\mu \to 0+} x_\mu = x^*$ und $\lim_{\mu \to 0+} y_\mu = y^*$. □

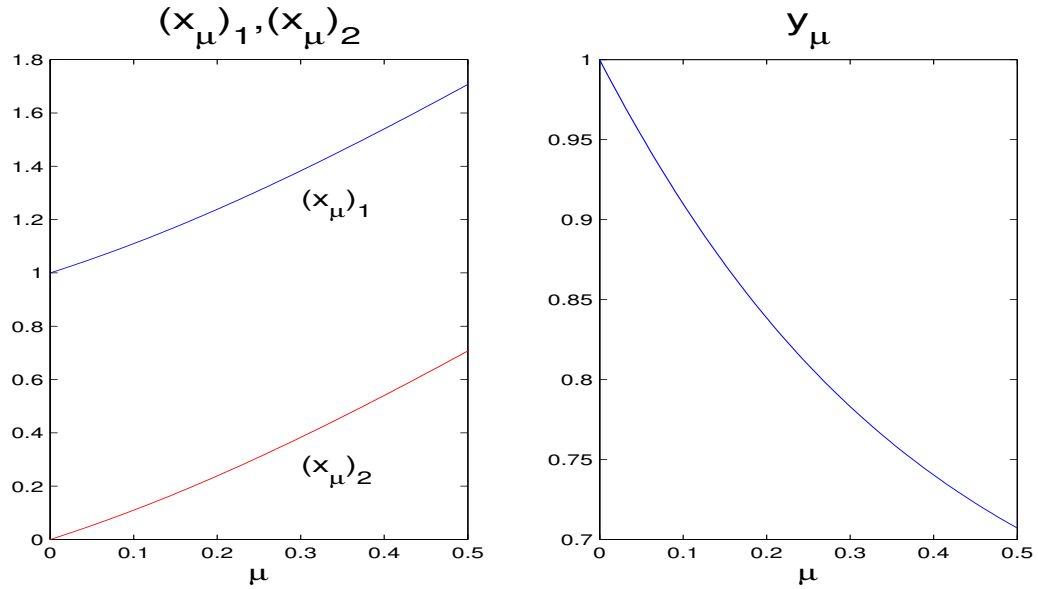


Abbildung 4.1: Die Lösungen x_μ und y_μ von (P_μ) bzw. (D_μ)

Den folgenden Satz geben wir ohne Beweis an[1].

Satz 1.2 *Die Voraussetzungen (A1)–(A3) seien erfüllt. Dann gilt: Die Optimierungsaufgaben (P_μ) und (D_μ) besitzen für jedes $\mu > 0$ genau eine Lösung $x_\mu \in M_0$ bzw. $y_\mu \in N_0$. Ferner existieren $x^* := \lim_{\mu\to 0+} x_\mu$ und $y^* := \lim_{\mu\to 0+} y_\mu$ und sind (gewisse) Lösungen von (P) bzw. (D).*

Im folgenden benutzen wir die folgende Bezeichnung. Ist $u \in \mathbb{R}^p$, so sei $U \in \mathbb{R}^{p\times p}$ diejenige Diagonalmatrix, die die Komponenten von u der Reihe nach in der Diagonalen als Einträge hat. Weiter sei e stets ein Vektor passender Länge, dessen Komponenten alle gleich 1 sind. Also[2] ist $u = Ue$.

Satz 1.3 *Die Voraussetzungen (A1)–(A3) seien erfüllt. Dann besitzt bei vorgegebenem $\mu > 0$ das nichtlineare "Gleichungssystem"*

$$(*) \qquad F_\mu(x,y,z) := \begin{pmatrix} Xz - \mu e \\ Ax - b \\ A^T y + z - c \end{pmatrix} = 0, \qquad \begin{array}{rcl} x & > & 0, \\ z & > & 0 \end{array}$$

genau eine Lösung $(x_\mu, y_\mu, z_\mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$. Hierbei ist $x_\mu \in M_0$ die Lösung von (P_μ), $y_\mu \in N_0$ die Lösung von (D_μ) und $c^T x_\mu - b^T y_\mu = \mu n$.

Beweis: Im ersten Teil des Beweises zeigen wir die Existenz einer Lösung von $(*)$. Genauer gilt:

- Sei $x_\mu \in M_0$ die Lösung von (P_μ), $y_\mu \in N_0$ die Lösung von (D_μ) und $z_\mu := c - A^T y_\mu$. Dann ist (x_μ, y_μ, z_μ) eine Lösung von $(*)$.

Denn: Offensichtlich ist $x_\mu > 0$, $z_\mu > 0$, $Ax_\mu - b = 0$ und $A^T y_\mu + z_\mu - c = 0$. Zu zeigen bleibt also nur noch $X_\mu z_\mu - \mu e = 0$ bzw. $c - \mu X_\mu^{-1} e - A^T y_\mu = 0$. Da x_μ Lösung von (P_μ) ist, folgt aus der Lagrangeschen Multiplikatorenregel die Existenz von $\tilde{y}_\mu \in \mathbb{R}^m$ mit $c - \mu X_\mu^{-1} e - A^T \tilde{y}_\mu = 0$. Wir haben zu zeigen, dass $\tilde{y}_\mu = y_\mu$. Offensichtlich ist $\tilde{y}_\mu \in N_0$. Mit $\tilde{z}_\mu := c - A^T \tilde{y}_\mu$ ist ferner

$$\nabla g_\mu(\tilde{y}_\mu) = b - \mu A \tilde{Z}_\mu^{-1} e = b - AX_\mu e = b - Ax_\mu = 0.$$

Da $g_\mu(\cdot)$ auf N_0 konkav ist, ist $\tilde{y}_\mu$ Lösung von (D_μ). Nun ist aber y_μ die eindeutige Lösung von (D_μ), folglich ist $\tilde{y}_\mu = y_\mu$ und nachgewiesen, dass (x_μ, y_μ, z_μ) Lösung von $(*)$ ist.

Nachdem eben die Existenz eine Lösung von $(*)$ nachgewiesen wurde, folgt jetzt die Eindeutigkeit.

- Sei (x, y, z) eine Lösung von $(*)$. Dann ist $x = x_\mu$ die Lösung von (P_μ), $y = y_\mu$ die Lösung von (D_μ) und $z = c - A^T y_\mu$. Insbesondere ist also eine Lösung von $(*)$ eindeutig bestimmt.

[1] Eine ausführliche Darstellung findet man in einem Skript zu einer Vorlesung über Operations Research, welches man unter `http://www.num.math.uni-goettingen.de/werner/opres.pdf` finden kann.

[2] Konsequenterweise müsste man die Einheitsmatrix dann allerdings mit E bezeichnen. Wir bleiben bei der (für uns) üblichen Bezeichnung I.

Denn: Wegen $x > 0$ und $Ax = b$ ist $x \in M_0$. Weiter ist

$$0 = c - A^T y - z = c - \mu X^{-1} e - A^T y = \nabla f_\mu(x) - A^T y.$$

Wegen der Konvexität von $f_\mu(\cdot)$ auf M_0 folgt, dass x Lösung von (P_μ) ist. Wegen $z = c - A^T y > 0$ ist $y \in N_0$. Weiter ist

$$\nabla g_\mu(y) = b - \mu A Z^{-1} e = b - AXe = b - Ax = 0,$$

woraus wegen der Konkavität von $g_\mu(\cdot)$ auf N_0 folgt, dass y eine Lösung von (D_μ) ist, also $y = y_\mu$ gilt. Damit ist auch der zweite Beweisschritt abgeschlossen.

Nun kommt der letzte Schritt.

- Ist (x_μ, y_μ, z_μ) die eindeutige Lösung von $(*)$, so ist $c^T x_\mu - b^T y_\mu = \mu n$.

Denn:

$$c^T x_\mu = (A^T y_\mu + z_\mu)^T x_\mu = b^T y_\mu + z_\mu^T x_\mu = b^T y_\mu + \mu e^T e = b^T y_\mu + \mu n.$$

Damit ist der Bewies abgeschlossen. □

Als Korollar zum letzten Satz könnten wir formulieren, dass das nichtlineare Gleichungssystem $(*)$ für jedes $\mu > 0$ eine eindeutige Lösung (x_μ, y_μ, z_μ) besitzt und für $\mu \to 0+$ in den ersten beiden Komponenten Konvergenz gegen eine Lösung von (P) bzw. (D) vorliegt. Die Menge

$$\mathcal{C} := \{(x_\mu, y_\mu, z_\mu) : \mu > 0\}$$

heißt der zu (P) und (D) gehörende *zentrale Pfad*.

4.1.3 Aufgaben

1. Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Die Menge $M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax \geq b\}$ sei nichtleer. Man zeige, dass M genau dann beschränkt ist, wenn es ein $u \in \mathbb{R}^m$ mit $u \geq 0$ und $A^T u < 0$ gibt.

2. 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 $Q_0, Q_1, \ldots, Q_l \in \mathbb{R}^{n \times n}$, Vektoren $c_0, c_1, \ldots, c_l \in \mathbb{R}^n$, $\beta_1, \ldots, \beta_l \in \mathbb{R}$ sowie $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Wir setzen

voraus, dass (P) zulässig bzw. $M \neq \emptyset$ ist. Man zeige, dass die Menge M_{opt} der Lösungen von (P) genau dann nichtleer und kompakt ist, 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.

Hinweis: Es darf der (hier unbewiesene) Existenzsatz für konvexe, quadratisch restringierte quadratische Programme (Satz 3.5 in Abschnitt 3.3) benutzt werden.

4.2 Das primal-duale Innere-Punkt-Verfahren

Die Idee, (P_μ) und (D_μ) für immer kleinere μ (mehr oder weniger) exakt zu lösen, ist wegen Instabilität und schlechter Konvergenz keine gute Idee. Wir werden ein wesentlich besseres Verfahren, nämlich das primal-duale Innere-Punkt-Verfahren, kennenlernen. Als Literatur zu diesem Abschnitt sei S. J. Wright (1997) empfohlen. Informationen kann man unter der Adresse `http://www.siam.org/books/swright/` erhalten.

4.2.1 Beschreibung des Verfahrens

Wir benutzen die im vorigen Abschnitt benutzten Bezeichnungen. Seien also weiter (P) und (D) die gegebenen zueinander dualen linearen Programme. Die Voraussetzungen (A1)–(A3) werden auch weiterhin wichtig sein. Nach wie vor sei bei gegebenem $x \in \mathbb{R}^n$ die Matrix $X \in \mathbb{R}^{n \times n}$ durch $X := \operatorname{diag}(x)$ definiert[3].

Wenn man ein nichtlineares Gleichungssystem lösen will, denkt man zunächst meistens an das Newton-Verfahren. Das gilt insbesondere für das nichtlineare Gleichungssystem $(*)$ in Satz 1.3. Um das Newton-Verfahren anwenden zu können, ist es wichtig, die Nichtsingularität der Funktionalmatrix zu sichern. Dies geschieht im folgenden Lemma.

Lemma 2.1 *Die Voraussetzung (A3) sei erfüllt. Bei vorgegebenem $\mu > 0$ sei die Abbildung $F_\mu : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ durch*

$$F_\mu(x, y, z) := \begin{pmatrix} Xz - \mu e \\ Ax - b \\ A^T y + z - c \end{pmatrix}$$

definiert. Dann ist die (von dem Parameter μ unabhängige) Funktionalmatrix

$$F'_\mu(x, y, z) = \begin{pmatrix} Z & 0 & X \\ A & 0 & 0 \\ 0 & A^T & I \end{pmatrix}$$

in jedem Tripel $(x, y, z) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ mit $x > 0$ und $z > 0$ nichtsingulär.

[3] In MATLAB ist `diag` ein Befehl, der genau dieser Beschreibung entspricht.

Beweis: Sei (p, q, r) aus dem Kern von $F'_\mu(x, y, z)$, also

$$Zp + Xr = 0, \qquad Ap = 0, \qquad A^T q + r = 0.$$

Nun ist

$$0 = q^T A(\underbrace{p + Z^{-1}Xr}_{=0}) = q^T \underbrace{Ap}_{=0} - q^T AZ^{-1}XA^T q = -q^T AZ^{-1}XA^T q.$$

Da $Z^{-1}X$ eine positiv definite (Diagonal-)Matrix ist und wegen (A3) Rang $(A) = m$ gilt, ist $AZ^{-1}XA^T$ positiv definit. Daher ist $q = 0$, folglich auch $r = 0$ und $p = 0$. Das einfache Lemma ist bewiesen. □

Ein Schritt des primal-dualen Innere-Punkt-Verfahrens besteht darin, zu einer aktuellen Näherung (x, y, z) (hier ist x Näherung für eine Lösung des primalen Problems, y Näherung für eine Lösung des dualen Problems und z eine Näherung für den optimalen Schlupf) mit $x > 0$ und $z > 0$ (hieran liegt es, dass man von *Innere-Punkt*-Verfahren spricht) ein $\mu > 0$ zu bestimmen, i. Allg. wird

$$\mu := \sigma \frac{x^T z}{n}$$

mit einem $\sigma \in (0, 1)$ gesetzt, dann einen Schritt des Newton-Verfahrens zur Lösung von $F_\mu = 0$ durchzuführen und diesen Schritt so zu dämpfen, dass auch für die neue Näherung (x_+, y_+, z_+) die Positivitätsbedingungen $x_+ > 0$ und $z_+ > 0$ erhalten bleiben. Wir beschreiben ein *unzulässiges* primal-duales Innere-Punkt-Verfahren und geben erst danach in einer Bemerkung die Vereinfachungen an, die sich bei einem *zulässigen* primal-dualen Innere-Punkt-Verfahren ergeben.

- Gegeben sei ein Tripel $(x, y, z) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ mit $x > 0$ und $z > 0$ sowie ein $\mu > 0$, etwa $\mu := \sigma x^T z/n$ mit $\sigma \in (0, 1)$.

- Berechne die Newton-Richtung (p, q, r) durch

$$\begin{pmatrix} p \\ q \\ r \end{pmatrix} := -F'_\mu(x, y, z)^{-1} F_\mu(x, y, z)$$

  bzw. als Lösung von

$$\begin{pmatrix} Z & 0 & X \\ A & 0 & 0 \\ 0 & A^T & I \end{pmatrix} \begin{pmatrix} p \\ q \\ r \end{pmatrix} = - \begin{pmatrix} Xz - \mu e \\ Ax - b \\ A^T y + z - c \end{pmatrix}.$$

- Mit Schrittweiten $\alpha_p > 0$ und $\alpha_d > 0$ berechne man das neue Tripel (x_+, y_+, z_+) durch

$$x_+ := x + \alpha_p p, \qquad \begin{pmatrix} y_+ \\ z_+ \end{pmatrix} := \begin{pmatrix} y \\ z \end{pmatrix} + \alpha_d \begin{pmatrix} q \\ r \end{pmatrix}.$$

Hierbei sind $\alpha_p > 0$ und $\alpha_d > 0$ zumindestens so zu wählen, dass $x_+ > 0$ und $z_+ > 0$. Häufig wird man im Prinzip folgendermaßen vorgehen. Man wähle $\tau_p, \tau_d \in (0,1)$ (z. B. $\tau_p = \tau_d = 0.995$) und bestimme

$$\alpha_{p,\max} := \sup\{\alpha > 0 : x + \alpha p > 0\}, \qquad \alpha_{d,\max} := \sup\{\alpha > 0 : z + \alpha r > 0\}$$

und setze anschließend $\alpha_p := \min(1, \tau_p \alpha_{p,\max})$ sowie $\alpha_d := \min(1, \tau_d \alpha_{d,\max})$. Offenbar ist

$$\alpha_{p,\max} = \min_{j:p_j<0}\Big(-\frac{x_j}{p_j}\Big), \qquad \alpha_{d,\max} = \min_{j:r_j<0}\Big(-\frac{z_j}{r_j}\Big).$$

Hierdurch ist ein Schritt des (unzulässigen) primal-dualen Innere-Punkt-Verfahrens beschrieben.

4.2.2 Einzelheiten zum Verfahren

Die Hauptarbeit besteht in jedem Schritt des oben geschilderten primal-dualen Innere-Punkt-Verfahrens in der Berechnung der Newton-Richtung (p, q, r), welche aus den Gleichungen

$$Zp + Xr = -(Xz - \mu e), \qquad Ap = -(Ax - b), \qquad A^T q + r = -(A^T y + z - c)$$

zu erhalten ist. Dies erlaubt die sukzessive Berechnung von q, danach von r und schließlich von p. Zur Abkürzung definiere man die positive Diagonalmatrix $D := Z^{-1}X$. Dann ist

$$\begin{aligned} ADA^T q &= -AD(A^T y + z - c) - ADr \\ &= -AD(A^T y + z - c) - AZ^{-1}Xr \\ &= -AD(A^T y + z - c) - AZ^{-1}[-Zp - (Xz - \mu e)] \\ &= -AD(A^T y + z - c) - (Ax - b) + ADX^{-1}(Xz - \mu e) \\ &= -[Ax - b + AD((A^T y + z - c) - X^{-1}(Xz - \mu e))] \end{aligned}$$

Also ist q aus dem linearen Gleichungssystem

$$(*) \qquad ADA^T q = -[Ax - b + AD((A^T y + z - c) - X^{-1}(Xz - \mu e))]$$

zu berechnen. Anschließend berechnet man p und r durch

$$r := -A^T q - (A^T y + z - c), \qquad p := -Dr - Z^{-1}(Xz - \mu e).$$

Unter der Rangvoraussetzung (A3) an A ist die Koeffizientenmatrix ADA^T des linearen Gleichungssystems $(*)$ symmetrisch und positiv definit, was natürlich bei einer effizienten Lösung zu beachten ist. Leider folgt aus einer Dünnbesetztheit von A i. Allg. nicht die von ADA^T. Denn ist etwa $A = (\ a_1 \ \cdots \ a_n\ )$ und $D = \operatorname{diag}(d_1, \ldots, d_n)$, so ist

$$ADA^T = \sum_{j=1}^{n} d_j a_j a_j^T.$$

Wenn A also nur eine voll besetzte Spalte hat, so ist ADA^T voll besetzt. Allerdings wird meistens nur ein kleiner Teil der Spalten dicht besetzt sein, was ausgenutzt werden kann, wie man in dem Report von Y. Zhang, siehe

`http://www.caam.rice.edu/~zhang/lipsol/`

nachlesen kann. Die Güte einer aktuellen Näherung wird bei Zhang gemessen durch den Defekt

$$d(x,y,z) := \frac{\|Ax-b\|}{\max(1,\|b\|)} + \frac{\|A^Ty+z-c\|}{\max(1,\|c\|)} + \frac{|c^Tx-b^Ty|}{\max(1,|c^Tx|,|b^Ty|)}$$

und die Rechnung abgebrochen, wenn dieser hinreichend klein (etwa 10^{-8}) ist.

Bemerkung: Ein Schritt eines *zulässigen* primal-dualen Innere-Punkt Verfahrens sieht im Prinzip folgendermaßen aus.

- Gegeben sei ein Paar $(x,y) \in M_0 \times N_0$, setze $z := c - A^Ty$, $D := Z^{-1}X$ und setze $\mu := \sigma x^Tz/n$ mit $\sigma \in (0,1)$.
- Berechne q als Lösung von

  $$ADA^Tq = X^{-1}(Xz - \mu e),$$

  anschließend p durch

  $$p = DA^T[q - X^{-1}(Xz-\mu e)].$$

- Berechne Schrittweite[4] $\alpha > 0$ mit $x + \alpha p > 0$ (wegen $Ap = 0$ ist dann auch $x + \alpha p \in M_0$) und $z - \alpha A^Tq > 0$ (dann ist auch $y + \alpha q \in N_0$). Etwa wähle man $\tau \in (0,1)$ und setze $\alpha := \min(1, \tau\alpha_{\max})$, wobei

  $$\alpha_{\max} := \min\Big[\min_{j:p_j<0}\Big(-\frac{x_j}{p_j}\Big), \min_{j:(A^Tq)_j>0}\Big(\frac{z_j}{(A^Tq)_j}\Big)\Big].$$

- Setze $(x_+, y_+) := (x,y) + \alpha(p,q)$.

Wegen der Zulässigkeit von x und y für (P) bzw. (D) wird die Güte der Näherung (x,y) im wesentlichen durch die Dualitätslücke $c^Tx - b^Ty = x^Tz$ gemessen. Daher ist es interessant, wie sich die Dualitätslücke von Schritt zu Schritt verändert. Es ist

$$\begin{aligned}
c^T(x+\alpha p) - b^T(y+\alpha q) &= c^Tx - b^Ty + \alpha(c-A^Ty)^Tp - \alpha(Ax)^Tq \\
&\qquad \text{(wegen } Ap = 0 \text{ und } Ax = b\text{)} \\
&= c^Tx - b^Ty + \alpha e^T(Zp - XA^Tq) \\
&= c^Tx - b^Ty + \alpha e^T(\mu e - XZe) \\
&= c^Tx - b^Ty + \alpha\mu n - \alpha x^Tz \\
&= [1-\alpha(1-\sigma)](c^Tx - b^Ty),
\end{aligned}$$

wobei wir davon ausgehen, dass $\mu = \sigma x^Tz/n$ mit einem $\sigma \in (0,1)$. Daher verkleinert sich die Dualitätslücke von Schritt zu Schritt. Eigentlich hängen die Dualitätslücke, die Schrittweite und auch σ von dem Iterationsindex ab. Es ist also

$$\frac{x_{k+1}^Tz_{k+1}}{x_k^Tz_k} = 1 - (1-\sigma_k)\alpha_k.$$

[4] In der Praxis wird man mit einer primalen *und* einer dualen Schrittweite arbeiten.

Wird also die Dualitätslücke gleichmäßig verkleinert, existiert also ein $c_0 \in (0,1)$ mit $1-(1-\sigma_k)\alpha_k \leq c_0$, so konvergiert die Dualitätslücke Q-linear gegen Null. Weiter konvergiert die Folge der Dualitätslücken sogar superlinear gegen Null, wenn[5] z. B. $\sigma_k \to 0$ und $\alpha_k \to 1$. □

Die *zulässigen* primal-dualen Innere-Punkt-Verfahren sind von eher historischer Bedeutung. Daher wollen wir auf die hier noch in Reichweite liegenden Konvergenzergebnisse nicht eingehen. Die *unzulässigen* primal-dualen Innere-Punkt-Verfahren haben sich in der Praxis durchgesetzt. Andererseits sind Konvergenzanalysen nach wie vor eher unerfreulich. Einen kleinen Eindruck hiervon erhält man in Chapter 6 bei S. J. Wright (1997).

Beispiel: In der Einführung haben wir erläutert, was man unter einem Netzwerkflussproblem versteht und wie eine Formulierung als lineare Optimierungsaufgabe aussieht. Hier wollen wir ein spezielles Beispiel betrachten und hierauf das (unzulässige) primal-duale Innere-Punkt-Verfahren anwenden, wobei wir allerdings die spezielle Struktur des Problems nicht ausnutzen und daher ganz naiv vorgehen werden.

Gegeben sei der in Abbildung 4.2 angegebene Digraph, wobei rechts verdeutlicht ist,

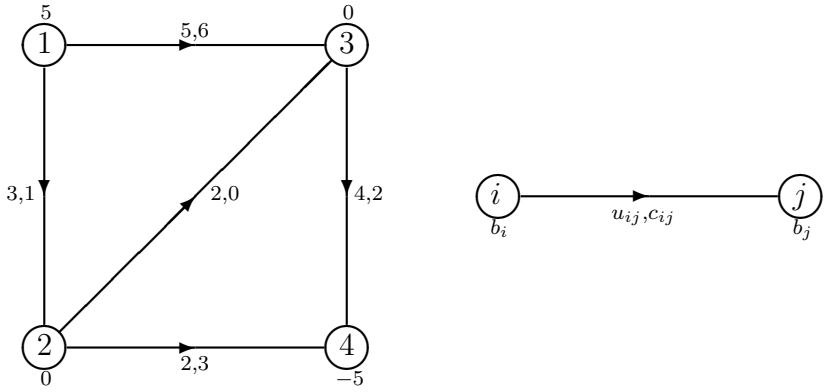


Abbildung 4.2: Ein spezielles Netzwerkflussproblem

welche Bedeutung die angegebenen Zahlen haben. Z. B. sind die Knoten 2 und 3 Umladeknoten, der Knoten 1 ein Angebots- und der Knoten 4 ein Bedarfsknoten. Als Knoten-Pfeil-Inzidenzmatrix hat man

	$(1,2)$	$(1,3)$	$(2,3)$	$(2,4)$	$(3,4)$
1	1	1	0	0	0
2	−1	0	1	1	0
3	0	−1	−1	0	1
4	0	0	0	−1	−1

[5] Numerische Experimente zeigen, dass es keine gute Idee ist, die σ_k konstant zu halten. Das gilt auch für das unzulässige Verfahren.

Das zugehörige lineare Programm lautet also

$$\text{Minimiere} \quad \begin{pmatrix} 1 \\ 6 \\ 0 \\ 3 \\ 2 \end{pmatrix}^T \begin{pmatrix} x_{12} \\ x_{13} \\ x_{23} \\ x_{24} \\ x_{34} \end{pmatrix} \quad \text{unter den Nebenbedingungen}$$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \leq \begin{pmatrix} x_{12} \\ x_{13} \\ x_{23} \\ x_{24} \\ x_{34} \end{pmatrix} \leq \begin{pmatrix} 3 \\ 5 \\ 2 \\ 2 \\ 4 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & -1 \end{pmatrix} \begin{pmatrix} x_{12} \\ x_{13} \\ x_{23} \\ x_{24} \\ x_{34} \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \\ 0 \\ -5 \end{pmatrix}.$$

Wir benutzen eine MATLAB-Funktion `primdual`, die wir zur Lösung einer Übungsaufgabe geschrieben haben. Ihr Kopf sieht folgendermaßen aus:

```
function [x_p,y_p,z_p,def]=primdual(A,b,c,x,y,z,sigma);
%*****************************************************************
%Es wird ein Schritt eines unzulaessigen primal-dualen Innere Punkt
%Verfahrens durchgefuehrt.
%Input:    Problem-Daten (A,b,c), ferner Tripel (x,y,z) mit x>0,z>0
%          und sigma>0.
%Output:   Neues Naeherungstripel (x_p,y_p,z_p) und Defekt def zu
%          Ausgangstripel (x,y,z)
%*****************************************************************
```

Wir beachten, dass die Knoten-Pfeil-Inzidenzmatrix nicht vollen Rang hat, da das Aufsummieren ihrer Zeilen den Nullvektor ergibt. Daher streichen wir eine der Gleichungen, etwa die letzte, führen Schlupfvariablen zur Überführung auf Normalform ein und starten das unzulässige primal-duale Innere-Punkt-Verfahren mit dem Tripel $(x, y, z) = (e, 0, e)$. Wir haben genau 6 Schritte des Verfahrens durchgeführt, wobei wir im k-ten Schritt $\sigma := 1/(k+1)^2$ benutzten. Mit def wird der erhaltene Defekt bezeichnet. Mit `format short` erhielten wir die folgenden Ergebnisse:

k	x_{12}	x_{13}	x_{23}	x_{24}	x_{34}	def
1	2.8794	1.7739	1.2211	1.7739	2.8794	2.3973
2	2.9773	2.0227	1.3453	1.6320	3.3680	0.9782
3	2.9999	2.0001	1.9414	1.0584	3.9416	0.3079
4	2.9997	2.0003	1.9994	1.0003	3.9997	0.0908
5	3.0000	2.0000	1.9999	1.0001	3.9999	0.0006
6	3.0000	2.0000	2.0000	1.0000	4.0000	0.0000

Hieraus liest man den optimalen Fluss ab. □

4.2.3 Aufgaben

1. **Programmieraufgabe:** Man programmiere einen Schritt des unzulässigen primal-dualen Innere-Punkt-Verfahrens für ein lineares Programm in Standardform

mit den Daten (A, b, c). Anschließend wende man das Verfahren (indem man etwa 10 Schritte durchführt) auf ein Beispiel mit den Daten

$$\begin{array}{|c|c|}\hline c^T & \\ \hline A & b \\ \hline \end{array} := \left[\begin{array}{rrrrrrr|r} 5 & 3 & 3 & 6 & 0 & 0 & 0 & \\ \hline -6 & 1 & 2 & 4 & 1 & 0 & 0 & 14 \\ 3 & -2 & -1 & -5 & 0 & 1 & 0 & -25 \\ -2 & 1 & 0 & 2 & 0 & 0 & 1 & 14 \end{array}\right]$$

an, wobei man mit $(x, y, z) := (e, 0, e)$ starte. Insbesondere beobachte man, wie sich der Defekt

$$d(x, y, z) := \frac{\|Ax - b\|}{\max(1, \|b\|)} + \frac{\|A^T y + z - c\|}{\max(1, \|c\|)} + \frac{|c^T x - b^T y|}{\max(1, |c^T x|, |b^T y|)}$$

verändert. Man sollte verschiedene Strategien bei der Wahl von σ ausprobieren, etwa $\sigma_k := 0.5$, $\sigma_k := 1/(k+1)$ und $\sigma_k := 1/(k+1)^2$.

2. Formulieren Sie die folgende Aufgabenstellung aus dem praktischen Leben als lineare Optimierungsaufgabe und führen Sie diese durch Einführung von Schlupfvariablen auf Normal- bzw. Standardform über. Wenn Aufgabe 1 erfolgreich gelöst wurde, können Sie anschließend die gesuchte Lösung mit Hilfe des primal-dualen Innere-Punkt-Verfahrens berechnen.

   Sie wollen Ihrer Tante (vielleicht eine reiche Erbtante?) zum Geburtstag eine Freude machen. Ihre Tante trinkt gerne einen süßen Wein und da Ihnen eine Beerenauslese zu teuer ist, kommen Sie auf die Idee, ihr einen Liter Wein zukommen zu lassen, den Sie selbst zusammengestellt haben.

   Hierzu können Sie einen Landwein für €1.00 pro Liter, zur Anhebung der Süße Diäthylenglykol-haltiges Frostschutzmittel für €1.20 pro Liter und für eine Verbesserung der Lagerungsfähigkeit eine Natriumacid-Lösung für €1.80 pro Liter kaufen. Verständlicherweise wollen Sie eine möglichst billige Mischung herstellen, wobei allerdings folgende Nebenbedingungen zu beachten sind: Um eine hinreichende Süße zu garantieren, muss die Mischung mindestens 1/3 Frostschutzmittel enthalten. Andererseits muss (z. B. wegen gesetzlicher Bestimmungen) mindestens halb so viel Wein wie Frostschutzmittel enthalten sein. Der Natriumacid-Anteil muss mindestens halb so groß, darf aber andererseits höchstens so groß wie der Glykol-Anteil sein und darf die Hälfte des Weinanteils nicht unterschreiten.

Kapitel 5

Quadratische Optimierungsaufgaben

In diesem Kapitel wollen wir auf quadratische Programme eingehen, also Optimierungsaufgaben, bei denen eine quadratische Zielfunktion unter (affin) linearen Nebenbedingungen zu minimieren ist. I. Allg. werden wir in diesem Kapitel die Aufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{l} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \le b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\} \end{array} \right.$$

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 die symmetrische Matrix $Q \in \mathbb{R}^{n \times n}$ gegeben, ferner sei m_0, die Anzahl der Ungleichungsrestriktionen, eine nichtnegative ganze Zahl kleiner oder gleich m. Zur Abkürzung setzen wir

$$A := \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix} \in \mathbb{R}^{m \times n}, \qquad b := \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m.$$

Wir weichen also geringfügig von den früheren Bezeichnungen ab. Später werden wir die dualen Variablen einheitlich mit y (statt (u, v)) bezeichnen.

Die Schwierigkeiten, mit denen man bei der numerischen Lösung von (P) zu rechnen hat, hängen stark von der Matrix Q ab. Insbesondere ist das Problem verhältnismäßig harmlos, wenn Q positiv definit ist. Dagegen hat z. B. die Aufgabe

$$\text{Minimiere} \quad f(x) := -\tfrac{1}{2} x^T x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : -e \le x \le e\},$$

wobei $e \in \mathbb{R}^n$ der Vektor ist, dessen Komponenten alle gleich 1 sind, jedes $z = (z_i) \in \mathbb{R}^n$ mit $|z_i| = 1$, $i = 1, \ldots, n$, als Lösung. Das sind also insgesamt 2^n Lösungen. Damit hat auch die MATLAB-Funktion `quadprog` Schwierigkeiten. Nach

```
Q=-eye(6);c=zeros(6,1);e=ones(6,1);
x=quadprog(Q,c,[],[],[],[],-e,e);
```

wird

$$x = \begin{pmatrix} -1.00000000000000 \\ -1.00000000000000 \\ -1.00000000000000 \\ -1.00000000000000 \\ -1.00000000000000 \\ -1.00000000000000 \end{pmatrix}$$

ausgegeben, immerhin eine der Lösungen. Hier wird per default mit einem sogenannten large scale Verfahren gerechnet. Nach

```
options=optimset('Largescale','off');
x=quadprog(Q,c,[],[],[],[],-e,e,[],options);
```

erhält man den sechsten Einheitsvektor des $\mathbb{R}^6$ als Ausgabe. In diesem sind zwar die notwendigen Optimalitätbedingungen erster Ordnung erfüllt, er ist aber kein globales Minimum von f. Man kann aber auch einen Startwert vorgeben. So erhält man z. B. nach

```
x=quadprog(Q,c,[],[],[],[],-e,e,0.5*e,options);
```

die Lösung $x = (1, 1, 1, 1, 1, 1)^T$.

Im ersten Abschnitt gehen wir auf das primale Verfahren von Fletcher, in dem darauf folgenden Abschnitt sehr kurz auf das duale Verfahren von Goldfarb-Idnani ein (hier wird die positive Definitheit von Q vorausgesetzt).

5.1 Das primale Verfahren von Fletcher

Wir beginnen mit einem primalen Verfahren bei quadratischen Programmen. Hier wird eine Folge zulässiger Lösungen mit monoton fallenden oder zumindestens monoton nicht wachsenden 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).

5.1.1 Gleichungen als Restriktionen

Wir betrachten den Spezialfall eines quadratischen Programms, bei welchem alle Restriktionen in der Form von m (linearen) Gleichungen vorliegen:

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

Hierbei seien $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und die symmetrische Matrix $Q \in \mathbb{R}^{n \times n}$ gegeben. Es gilt:

Satz 1.1 *Gegeben sei das durch lineare Gleichungen restringierte quadratische Programm (P). Dann gilt:*

1. *Die folgenden Aussagen sind äquivalent:*
   (a) $x^* \in M$ *ist eine lokale Lösung.*
   (b) *Es existiert* $y^* \in \mathbb{R}^m$ *mit* $c+Qx^*+A^Ty^* = 0$ *und es ist* Q *positiv semidefinit auf* $\mathrm{Kern}\,(A)$.
   (c) $x^* \in M$ *ist eine globale Lösung von (P).*
2. *Ist* $x^* \in M$, *existiert ein* $y^* \in \mathbb{R}^m$ *mit* $c + Qx^* + A^Ty^* = 0$ *und ist* Q *positiv definit auf* $\mathrm{Kern}\,(A)$, *so ist* x^* *eindeutige globale Lösung von (P).*
3. *Ist* Q *positiv definit auf* $\mathrm{Kern}\,(A)$ *und* $\mathrm{Rang}\,(A) = m$, *so ist*
$$K := \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix}$$
nichtsingulär. Das lineare Gleichungssystem
$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}$$
besitzt also eine eindeutige Lösung $(x^*, y^*) \in \mathbb{R}^n \times \mathbb{R}^m$, *wobei* x^* *die Lösung von (P) und* y^* *der zugehörige Lagrange-Multiplikator ist.*

Beweis:

1. Die Implikation (a)$\Longrightarrow$(b) folgt aus den notwendigen Bedingungen zweiter Ordnung in Satz 2.5 in Abschnitt 3.2. Zum Nachweis der Implikation (b)$\Longrightarrow$(c) geben wir uns $x \in M$ beliebig vor. Dann ist $x - x^* \in \mathrm{Kern}\,(A)$ und daher
$$\begin{aligned} f(x) &= f(x^*) + (c + Qx^*)^T(x - x^*) + \tfrac{1}{2}\underbrace{(x - x^*)^TQ(x - x^*)}_{\geq 0} \\ &\geq f(x^*) - (y^*)^T\underbrace{A(x - x^*)}_{=0} \\ &= f(x^*), \end{aligned}$$
   also $x^* \in M$ globale Lösung von (P). Die Implikation (c)$\Longrightarrow$(a) ist trivial.
2. Aus der obigen Gleichungs-Ungleichungskette folgt die Behauptung sofort.
3. Wir zeigen, dass der Kern von K trivial ist. Angenommen, es ist
$$Qx + A^Ty = 0, \qquad Ax = 0.$$
   Eine Multiplikation der ersten Gleichung von links mit x^T liefert $x^TQx = 0$, wegen $x \in \mathrm{Kern}\,(A)$ ist $x = 0$. Da die m Zeilen von A bzw. die m Spalten von A^T linear unabhängig sind, folgt aus der ersten Gleichung $y = 0$. Insgesamt ist K nichtsingulär. Der Rest folgt aus dem gerade eben bewiesenen zweiten Teil.

Der Satz ist bewiesen. □

Bemerkung: Sei Q auf Kern (A) positiv definit und Rang $(A) = m$. Wir schildern, wie man das lineare Gleichungssystem

$$(*) \qquad \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}$$

effizient löst[1].

- Man berechne eine QR-Zerlegung von A^T, also eine Darstellung

  $$A^T = Z \begin{pmatrix} R \\ 0 \end{pmatrix} \begin{matrix} \} \, m \\ \} \, n-m \end{matrix}$$

  mit einer orthogonalen Matrix $Z \in \mathbb{R}^{n \times n}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{m \times m}$. Man denke sich Z durch

  $$Z = (\underbrace{Z^{(1)}}_{m} \quad \underbrace{Z^{(2)}}_{n-m})$$

  partitioniert.

Dann ist R nichtsingulär, was sofort aus der Darstellung $A^T = Z^{(1)}R$ und Rang $(A) = m$ folgt. Ferner ist Kern $(A) =$ Bild $(Z^{(2)})$. Denn wegen $AZ^{(2)} = 0$ ist Bild $(Z^{(2)}) \subset$ Kern (A). Die $n - m$ Spalten von $Z^{(2)}$ sind linear unabhängig, so dass Bild $(Z^{(2)})$ ein $(n - m)$-dimensionaler linearer Teilraum des $\mathbb{R}^n$ ist. Da auch Kern (A) ein $(n - m)$-dimensionaler (wegen Rang $(A) = m$) linearer Teilraum des $\mathbb{R}^n$ ist, ist Kern $(A) =$ Bild $(Z^{(2)})$. Die $n - m$ Spalten von $Z^{(2)}$ bilden also eine Basis von Kern (A). Da Z orthogonal ist, ist

$$Z^T Z = \begin{pmatrix} (Z^{(1)})^T Z^{(1)} & (Z^{(1)})^T Z^{(2)} \\ (Z^{(2)})^T Z^{(1)} & (Z^{(2)})^T Z^{(2)} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

- Durch die folgenden Schritte berechne man die Lösung $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ von $(*)$:
  - Berechne $x^{(1)} \in \mathbb{R}^m$ durch Vorwärtseinsetzen aus

    $$R^T x^{(1)} = b.$$

  - Berechne

    $$\begin{pmatrix} c^{(1)} \\ c^{(2)} \end{pmatrix} := Z^T c, \qquad \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} := Z^T Q Z.$$

  - Berechne $x^{(2)} \in \mathbb{R}^{n-m}$ als Lösung von

    $$\underbrace{(Z^{(2)})^T Q Z^{(2)}}_{=B_{22}} x^{(2)} = -\underbrace{(Z^{(2)})^T c}_{c^{(2)}} - \underbrace{(Z^{(2)})^T Q Z^{(1)}}_{B_{21}} x^{(1)}.$$

[1] Siehe auch W. Alt (2002, S. 225 ff.).

– Berechne

$$x := Z \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix} = Z^{(1)}x^{(1)} + Z^{(2)}x^{(2)}.$$

– Berechne $y \in \mathbb{R}^m$ durch Rückwärtseinsetzen aus

$$Ry = -\underbrace{(Z^{(1)})^T c}_{c^{(1)}} - \underbrace{(Z^{(1)})^T Q Z^{(1)}}_{B_{11}} x^{(1)} - \underbrace{(Z^{(1)})^T Q Z^{(2)}}_{B_{12}} x^{(2)}.$$

Diese Schritte sind durchführbar, da R (und damit auch R^T) nichtsingulär ist, ferner $(Z^{(2)})^T Q Z^{(2)}$ positiv definit ist (da Q positiv definit auf $\text{Kern}\,(A) = \text{Bild}\,(Z^{(2)})$). Zu zeigen bleibt, dass das berechnete Paar (x, y) die Lösung von $(*)$ ergibt. Zunächst ist

$$Ax = \underbrace{AZ^{(1)}}_{=R^T} x^{(1)} + \underbrace{AZ^{(2)}}_{=0} x^{(2)} = R^T x^{(1)} = b.$$

Weiter ist

$$\begin{aligned} Z^T(Qx + A^T y) &= \begin{pmatrix} (Z^{(1)})^T \\ (Z^{(2)})^T \end{pmatrix} [Qx + Z^{(1)} R y] \\ &= \begin{pmatrix} (Z^{(1)})^T Q x + Ry \\ (Z^{(2)})^T Q Z^{(1)} x^{(1)} + (Z^{(2)})^T Q Z^{(2)} x^{(2)} \end{pmatrix} \\ &= -\begin{pmatrix} (Z^{(1)})^T c \\ (Z^{(2)})^T c \end{pmatrix} \\ &= -Z^T c, \end{aligned}$$

folglich auch $Qx + A^T y = -c$. Damit ist schließlich gezeigt, dass das Verfahren durchführbar ist und die gesuchte Lösung berechnet. □

5.1.2 Das Verfahren von Fletcher

Gegeben sei das zu Beginn des Kapitels angegebene quadratische Programm (P) mit der Menge M zulässiger Lösungen, welche durch m_0 Ungleichungen und $m - m_0$ Gleichungen beschrieben ist. In diesem Unterabschnitt werden wir ein Verfahren von R. Fletcher (1971) 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) angegeben worden. Von P. E. Gill, W. Murray (1978) 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).

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. Daher enthält $I(x)$ insbesondere die Indizes aller Gleichungsrestriktionen, also $\{m_0 + 1, \ldots, m\}$.

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 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.

Der Einfachheit halber setzen wir im folgenden voraus:

(V) Es ist Rang $(A_{I(x)}) = \#(I(x))$ für jedes $x \in M$. Ferner ist die symmetrische Matrix $Q \in \mathbb{R}^{n \times n}$ positiv definit auf Kern $(A_{\{m_0+1,\ldots,m\}})$ und damit auf Kern $(A_{I(x)})$ für jedes $x \in M$.

Außerdem wird natürlich stets die Zulässigkeit von (P) bzw. $M \neq \emptyset$ vorausgesetzt (andernfalls macht ein primales Verfahren keinen Sinn). Dann besitzt (P) eine eindeutige Lösung $x^* \in M$. Mit $I^* := I(x^*)$ ist x^* auch die eindeutige Lösung von

$$\text{Minimiere} \quad c^T x + \tfrac{1}{2} x^T Q x \quad \text{unter der Nebenbedingung} \quad A_{I^*} x = b_{I^*}.$$

Würden wir also die richtige Indexmenge in der Lösung aktiver Restriktionen kennen, so könnten wir die Lösung von (P) nach der im letzten Unterabschnitt beschriebenen Methode bestimmen.

Beispiel: Wir betrachten das quadratische Programm

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} & \begin{pmatrix} -2 \\ -6 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ & \text{unter der Nebenbedingung} \\ & \begin{pmatrix} 1 & 1 \\ -1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq \begin{pmatrix} 2 \\ 2 \\ 3 \end{pmatrix}. \end{cases}$$

Angenommen, wir würden wissen, dass $I^* = \{1, 2\}$ die optimale Indexmenge ist, dann könnten wir die Lösung x^* und den zur Indexmenge I^* gehörenden Multiplikator $y_{I^*}^*$ durch Lösen des linearen Gleichungssystems

$$\begin{pmatrix} Q & A_{I^*}^T \\ A_{I^*} & 0 \end{pmatrix} \begin{pmatrix} x^* \\ y_{I^*}^* \end{pmatrix} = \begin{pmatrix} -c \\ b_{I^*} \end{pmatrix}$$

bzw.

$$\left(\begin{array}{cc|cc} 1 & -1 & 1 & -1 \\ -1 & 2 & 1 & 2 \\ \hline 1 & 1 & 0 & 0 \\ -1 & 2 & 0 & 0 \end{array}\right) \left(\begin{array}{c} x^* \\ \hline y_{I^*}^* \end{array}\right) = \left(\begin{array}{c} 2 \\ 6 \\ \hline 2 \\ 2 \end{array}\right).$$

Als Lösung erhält man

$$x^* = \begin{pmatrix} \frac{2}{3} \\ \frac{4}{3} \end{pmatrix}, \qquad y_{I^*}^* = \begin{pmatrix} \frac{28}{9} \\ \frac{4}{9} \end{pmatrix}.$$

Da die Komponenten von $y_{I^*}^*$ nichtnegativ sind, ist die Annahme $I^* = \{1, 2\}$ im nachhinein gerechtfertigt und x^* wirklich die Lösung von (P). □

Die Idee des primalen Verfahrens von Fletcher, auch *Aktive-Mengen-Methode* genannt, besteht darin, eine Folge von Indexmengen und zugehörigen Lösungen gleichungsrestringierter quadratischer Programme mit monoton nicht wachsenden Kosten zu bestimmen, dabei stets zu kontrollieren, ob die zu Ungleichungen gehörenden Multiplikatoren nichtnegativ sind und gegebenenfalls mit der Lösung abzubrechen.

Genauer sieht ein Schritt des Verfahrens von Fletcher folgendermaßen aus:

(0) Gegeben sei ein $x \in M$ und eine Indexmenge I mit $\{m_0+1,\dots,m\} \subset I \subset I(x)$ und Rang $(A_I) = q$, wobei $q := \#(I)$.

(1) Berechne $p \in \mathbb{R}^n$ und $y_I = (y_i)_{i\in I} \in \mathbb{R}^q$ mit

$$c + Qx + Qp + A_I^T y_I = 0, \qquad A_I p = 0,$$

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[2] $x + p \in M$, dann:

Setze $x_+ := x + p$.

Bestimme $l \in I \cap \{1,\dots,m_0\}$ mit $y_l = \min_{i\in I\cap\{1,\dots,m_0\}} y_i$.

Falls $y_l \geq 0$, dann:

STOP, $x^* := x_+$ ist die Lösung von (P), y_I ist zugehöriger Lagrange-Multiplikator.

Andernfalls:

Setze $I_+ := I \setminus \{l\}$.

Andernfalls:

Berechne[3] die maximale Schrittweite

$$s(x,p) := \min\Big\{ \frac{b_i - a_i^T x}{a_i^T p} : i \notin I,\ a_i^T p > 0 \Big\} = \frac{b_r - a_r^T x}{a_r^T p}.$$

Setze[4] $x_+ := x + s(x,p)p$ und $I_+ := I \cup \{r\}$.

(3) Setze $(x, I) := (x_+, I_+)$, gehe nach (1).

Die Durchführbarkeit dieses Verfahrens ist unter der Voraussetzung (V) gesichert und das neue Paar (x_+, I_+) genügt den in (0) gemachten Voraussetzungen. Wir berechnen

[2] Hier müssen natürlich nur die Restriktionen mit einem Index aus $\{1,\dots,m\} \setminus I$ getestet werden.
[3] Man beachte, dass $\{i \in \{1,\dots,m\} : i \notin I,\ a_i^T p > 0\} \neq \emptyset$, da andernfalls $x + p \in M$.
[4] Dann ist $x_+ \in M$ und $a_r^T x_+ = b_r$.

die Kosten $f(x_+) = f(x+tp)$, wobei $t := 1$, falls $x + p \in M$, und $t := s(x,p) \in [0,1)$, falls $x + p \notin M$. Es ist

$$\begin{aligned} f(x_+) &= f(x) + t(c+Qx)^T p + \tfrac{1}{2}t^2 p^T Q p \\ &= f(x) - t(A_I^T y_I + Qp)^T p + \tfrac{1}{2}t^2 p^T Q p \\ &= f(x) + (\tfrac{1}{2}t^2 - t) p^T Q p. \end{aligned}$$

Daher ist $f(x+tp) \le f(x)$ für alle $t \in [0,2]$ und $f(x+tp)$ minimal für $t = 1$. Es ist also vernünftig, die neue Näherung x_+ in der angegebenen Weise zu definieren, da die Zielfunktion auf $\{x + tp : t \ge 0\} \cap M$ gerade in x_+ minimal wird. Jedenfalls ist $f(x_+) < f(x)$ außer in dem entarteten Fall $x_+ = x$. Dieser Fall tritt entweder ein wenn $p = 0$ oder wenn $s(x,p) = 0$. Wir wollen keinen Beweis für die Endlichkeit des Fletcher-Verfahrens bringen (siehe Bemerkungen bei C. Geiger, C. Kanzow (2002, S. 204), die ebenfalls nicht die Endlichkeit beweisen).

Beispiel: Wir reproduzieren ein Beispiel aus dem Skript der Vorlesung Optimierung von H. J. Oberle aus dem WS 2005/2006. Die Aufgabe lautet:

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := \frac{1}{2}(x_1^2 + x_2^2) \quad \text{unter den Nebenbedingungen} \\ \qquad\qquad x_1 \le 2, \quad x_2 \le 1, \quad x_1 + 3x_2 \ge 2. \end{cases}$$

Man hat also $n = 2$ Variable, $m_0 = m = 3$ Ungleichungs- und keine Gleichungsrestriktionen, ferner ist

$$c = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -3 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix}.$$

Wir starten mit $(x_0, I_0) := ((2,1)^T, \{1,2\})$, der zugehörige Zielfunktionswert ist $f(x_0) = \frac{5}{2}$. Anschließend berechne man die Lösung des linearen Gleichungssystems

$$\left(\begin{array}{c|c} Q & A_{I_0}^T \\ \hline A_{I_0} & 0 \end{array}\right) \begin{pmatrix} p \\ \hline y_{I_0} \end{pmatrix} = \begin{pmatrix} -(c+Qx_0) \\ \hline 0 \end{pmatrix}$$

bzw.

$$\left(\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right) \begin{pmatrix} p_1 \\ p_2 \\ \hline y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \\ \hline 0 \\ 0 \end{pmatrix}.$$

Wir erhalten $p = (0,0)^T$, $y_{I_0} = (-2,-1)^T$. Dann ist $x_1 := x_0$ und die Restriktion $l = 1$ wird aus I_0 entfernt, d. h. es ist $I_1 := \{2\}$. Anschließend ist das Gleichungssystem

$$\left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 1 \\ \hline 0 & 1 & 0 \end{array}\right) \begin{pmatrix} p_1 \\ p_2 \\ \hline y_2 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \\ \hline 0 \end{pmatrix}$$

zu lösen. Wir erhalten $p = (-2, 0)^T$, $y_{I_1} = (-1)$. Es ist $x_2 := x_1 + p = (0, 1)^T$ zulässig, $f(x_2) = \frac{1}{2}$ und $I_2 := \emptyset$. Im nächsten Schritt hat man

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

zu lösen und erhält $p = (0, -1)^T$. Da $x_2 + p = 0$ nicht zulässig ist, muss die maximale Schrittweite

$$s(x_2, p) = \frac{-2+3}{3} = \frac{1}{3} = \frac{b_3 - a_3^T x_2}{a_3^T p}$$

berechnet werden. Daher ist $x_3 := x_2 + \frac{1}{3}p = (0, \frac{2}{3})^T$, $I_3 = \{3\}$, $f(x_3) = \frac{2}{9}$. Das nächste zu lösende lineare Gleichungssystem ist

$$\left(\begin{array}{cc|c} 1 & 0 & -1 \\ 0 & 1 & -3 \\ \hline -1 & -3 & 0 \end{array}\right) \begin{pmatrix} p_1 \\ p_2 \\ y_3 \end{pmatrix} = \left(\begin{array}{c} 0 \\ -\frac{2}{3} \\ \hline 0 \end{array}\right)$$

mit der Lösung $p = (\frac{1}{5}, -\frac{1}{15})^T$, $y_{I_3} = \frac{1}{5}$. Man stellt fest, dass $x_4 := x_3 + p = (\frac{1}{5}, \frac{3}{5})^T$ zulässig und das STOP-Kriterium erfüllt ist, da der zugehörige Multiplikator nichtnegativ ist. Daher ist x_4 mit $f(x_4) = \frac{1}{5}$ die eindeutige globale Lösung von (P). In Abbildung 5.1 stellen wir das Problem, die Folge der Iterierten und die Lösung dar. □

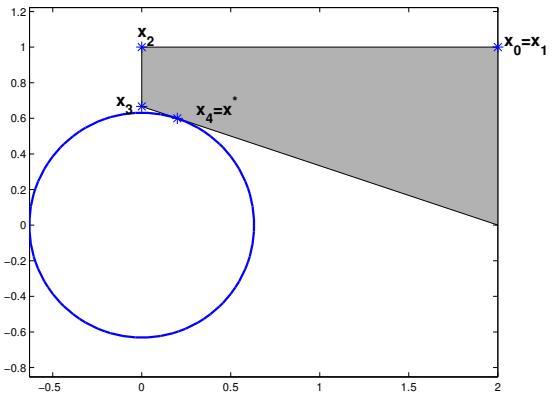


Abbildung 5.1: Das Verfahren von Fletcher bei einem quadratischen Programm

Wir schreiben ein einfaches MATLAB function file `Fletcher.m`, in welchem das primale Fletcher-Verfahren realisiert wird. Wir beschränken uns auf den Fall, dass nur Ungleichungen als Restriktionen auftreten. Es sei betont, dass die folgende Funktion einige Mängel hat. So wird die Berechnung von (p, y_I) in Schritt (1) auf die einfachste Art durchgeführt, ohne die im letzten Unterabschnitt angegebene Methode zu berücksichtigen. Weiter werden falsche Daten oder unzulässige Eingaben nicht abgefangen, ferner ist der Test auf Zulässigkeit recht naiv.

```
function [x,y_I,I,info]=Fletcher(Q,c,A,b,x,I,max_it);
%*********************************************************************
%Es wird das primale Verfahrens von Fletcher auf
%das quadratische Programm
%  Minimiere c'*x+0.5*x'*Q*x unter den Nebenbedingungen
%                  A*x<=b
%angewandt.
%*********************************************************************
%Input-Parameter:
%       Q,c          Definieren Zielfunktion
%       A,b          Definieren Restriktionen
%       x            zulaessiger Startvektor (wird nicht geprueft)
%       I            Indexmenge: A_Ix=b_I, rang(A_I)=#(I),
%                    Zeilenvektor
%       max_it       maximale Anzahl der Iterationen
%*********************************************************************
%Output-Parameter:
%       x            Loesung
%       y_I          Lagrange-Multiplikator
%       I            Indexmenge
%       info         Ist info>0, so ist x Loesung, info gibt die Anzahl
%                    der Iterationen an. Andernfalls ist info=0
%*********************************************************************
[m,n]=size(A);q=length(I);iter=0;info=0;
while(info==0)&(iter<max_it)
    iter=iter+1; A_I=A(I,:);
    z=-[Q A_I';A_I zeros(q)]\[c+Q*x;zeros(q,1)];
    p=z(1:n);y_I=z(n+1:n+q);
    not_I=setdiff(1:m,I);
    if length(not_I)>0
        A_not_I=A(not_I,:);b_not_I=b(not_I);
        defekt=b_not_I-A_not_I*x; u=A_not_I*p;
        defekt_p=defekt-u;
    else
        defekt_p=0;
    end;
    if min(defekt_p)>=0
        x=x+p;
        if (q==0)
            info=iter;
        else
            [y_l,l]=min(y_I);
            if y_l>=0
                 info=iter;
            else
                 I=setdiff(I,I(l));q=q-1;
            end;
        end;
    else
        P=find(u>0);
        [s,r]=min(defekt(P)./u(P));
        x=x+s*p; I=[I not_I(P(r))];q=q+1;
    end;
end;
```

Beispiel: Wir betrachten ein Beispiel bei C. Geiger, C. Kanzow (2002, S. 204), nämlich das quadratische Programm

$$\text{Minimiere} \quad \begin{pmatrix} 2 \\ 1 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \tfrac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

unter der Nebenbedingung

$$\begin{pmatrix} -1 & -1 \\ 0 & 1 \\ 1 & 1 \\ -1 & 1 \\ 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 2 \\ 5 \\ 2 \\ 5 \\ 1 \end{pmatrix}.$$

Wie bei Geiger-Kanzow starten wir mit $x = (5, 0)^T$ und der Indexmenge $I = (3, 5)$.

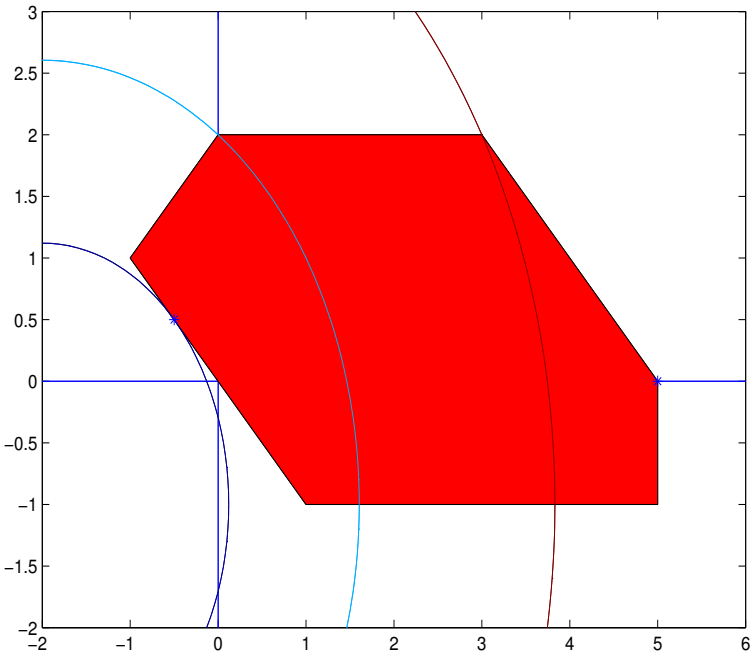


Abbildung 5.2: Zulässige Menge, Zielfunktion eines quadratischen Programms

Nach 8 Iterationen hat man die Lösung $x^* := (-0.5, 0.5)^T$ gefunden, die optimale Indexmenge ist $I^* := (1)$, der zugehörige Multiplikator ist $y_{I^*} = (1.5)$. Wir erhalten dieselbe Folge von Näherungslösungen wie bei C. Geiger, C. Kanzow (2002, S. 205). Startet man übrigens mit derselben Näherungslösung $x := (5, 0)^T$, setzt aber `I=[]` die leere Indexmenge, so ist man schon nach zwei Iterationen am Ziel. Zum Vergleich wollen wir auch noch die Funktion `quadprog` aus der Optimization Toolbox von MATLAB benutzen. Hierzu schreiben wir ein Script M-file `Test.m`, in dem die Daten festgelegt werden, der Startpunkt festgelegt (was nicht nötig ist) und `quadprog` aufgerufen wird.

```
Q=[1,0;0,1]; c=[2;1];
A=[-1,-1;0,1;1,1;-1,1;1,0;0,-1]; b=[0;2;5;2;5;1];
x=[5;0];
[x,fval,exitflag,output,lambda]=quadprog(Q,c,A,b,[],[],[],[],x);
```

In `x` steht die Lösung, `fval` gibt den zugehörigen Zielfunktionswert an, in `lambda` stehen die zugehörigen Lagrange-Multiplikatoren, wobei zu unterscheiden ist zwischen Multiplikatoren bezüglich unterschiedlicher Restriktionen: Ungleichungen, Gleichungen, unterer und oberer Schranken. Die Multiplikatoren zu den Ungleichungsrestriktionen stehen z. B. in `lambda.ineqlin`. Natürlich stimmen die Ergebnisse mit den oben erhaltenen überein. □

In einem Iterationsschritt wird in der Indexmenge I entweder ein Index entfernt oder es kommt ein Index hinzu. Bezogen auf das zu lösende lineare Gleichungssystem (und genau hierin besteht die wesentliche Arbeit) bedeutet dies, dass entweder eine Spalte und eine Zeile gestrichen werden oder jeweils eine hinzukommt. Eine stabile, effiziente Implementation sollte das berücksichtigen. Genaueres hierzu findet man in dem oben zitierten Aufsatz von Gill-Murray.

5.1.3 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

$$N := (AQ^{-1}A^T)^{-1}AQ^{-1}, \qquad H := Q^{-1}(I - A^T N)$$

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. **Programmieraufgabe:** Gegeben sei das durch lineare Gleichungen restringierte quadratische Programm

(P) Minimiere $f(x) := c^T x + \frac{1}{2}x^T Qx$ auf $M := \{x \in \mathbb{R}^n : Ax = b\}$.

Hierbei seien $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und die symmetrische Matrix $Q \in \mathbb{R}^{n\times n}$ gegeben, die auf Kern (A) positiv definit sei. Zur Lösung von (P) bzw. des linearen Gleichungssystems

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}$$

wurde in Unterabschnitt 5.1.1 das folgende Verfahren angegeben:

- Berechne eine QR-Zerlegung von A^T, also eine Darstellung

$$A^T = Z \begin{pmatrix} R \\ 0 \end{pmatrix} \begin{matrix} \} \, m \\ \} \, n-m \end{matrix}$$

mit einer orthogonalen Matrix $Z \in \mathbb{R}^{n \times n}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{m \times m}$. Man denke sich Z durch

$$Z = (\underbrace{Z^{(1)}}_{m} \quad \underbrace{Z^{(2)}}_{n-m})$$

partitioniert.

- Berechne $x^{(1)} \in \mathbb{R}^m$ durch Vorwärtseinsetzen aus

$$R^T x^{(1)} = b.$$

- Berechne

$$\begin{pmatrix} c^{(1)} \\ c^{(2)} \end{pmatrix} := Z^T c, \qquad \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} := Z^T Q Z.$$

- Berechne $x^{(2)} \in \mathbb{R}^{n-m}$ als Lösung von

$$\underbrace{(Z^{(2)})^T Q Z^{(2)}}_{=B_{22}} x^{(2)} = - \underbrace{(Z^{(2)})^T c}_{c^{(2)}} - \underbrace{(Z^{(2)})^T Q Z^{(1)}}_{B_{21}} x^{(1)}.$$

- Berechne

$$x := Z \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix} = Z^{(1)} x^{(1)} + Z^{(2)} x^{(2)}.$$

- Berechne $y \in \mathbb{R}^m$ durch Rückwärtseinsetzen aus

$$Ry = - \underbrace{(Z^{(1)})^T c}_{c^{(1)}} - \underbrace{(Z^{(1)})^T Q Z^{(1)}}_{B_{11}} x^{(1)} - \underbrace{(Z^{(1)})^T Q Z^{(2)}}_{B_{12}} x^{(2)}.$$

Man implementiere dieses Verfahren und erprobe die Implementation an der Aufgabe mit den Daten

$$A := \begin{pmatrix} 1 & 3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & -2 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix}, \quad b := \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad c := \begin{pmatrix} 0 \\ -2 \\ -2 \\ -1 \\ -1 \end{pmatrix}$$

sowie

$$Q := \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

3. Bei gegebenen $n \in \mathbb{N}$ und $K > 0$ bestimme man[5] die Lösung der Aufgabe

$$\text{Minimiere} \quad f(x) := \tfrac{1}{2}\sum_{j=1}^{n} j x_j^2 \quad \text{auf} \quad M := \Big\{x \in \mathbb{R}^n : \sum_{j=1}^{n} x_j = K\Big\}.$$

4. Man wende das primale Verfahren von Fletcher auf das quadratische Programm

$$\left\{ \begin{array}{c} \text{Minimiere} \quad \begin{pmatrix} -3 \\ 0 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \tfrac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \text{unter der Nebenbedingung} \\ \begin{pmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \le \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \end{array} \right.$$

an. Hierbei starte man mit $(x_0, I_0) := ((0,0)^T, \{1,2\})$.

5. In dem Polyeder

$$P := \{x \in \mathbb{R}^3 : x_1 + 2x_2 - x_3 \ge 4,\ -x_1 + x_2 - x_3 \le 2\}$$

bestimme man den Punkt, der den kleinsten euklidischen Abstand zum Nullpunkt im $\mathbb{R}^3$ besitzt[6].

6. Gegeben sei die Aufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := \tfrac{1}{2}\|x - z\|^2 \quad \text{unter der Nebenbedingung} \\ g(x) := \tfrac{1}{2} x^T A x + b^T x - \alpha \le 0. \end{array} \right.$$

Hierbei sind $\alpha \in \mathbb{R}$, $z, b \in \mathbb{R}^n$ und die symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ vorgegeben, ferner ist $\|\cdot\|$ die euklidische Norm im $\mathbb{R}^n$. Es sei $g(z) > 0$ bzw. $z \notin M$ (andernfalls ist die Aufgabe (P) trivial) und $-\frac{1}{2} b^T A^{-1} b - \alpha < 0$ bzw. $\min_{x\in\mathbb{R}^n} g(x) < 0$. Dies impliziert, dass $M \neq \emptyset$.

   (a) Man zeige, dass (P) genau eine Lösung $x^* \in M$ besitzt.

   (b) Man zeige, dass $g(x^*) = 0$ und $\nabla g(x^*) \neq 0$. Hieraus schließe man, dass die Lösung $x^* \in M$ von (P) charakterisiert ist durch die Existenz eines $\lambda^* > 0$ mit $x^* - z + \lambda^*(Ax^* + b) = 0$.

   (c) Für den Spezialfall $n := 3$, $z := (1,1,1)^T$, $A := \operatorname{diag}(1,2,3)$, $b := (0,0,0)^T$ und $\alpha := 2$ bestimme man auf effiziente Weise die Lösung x^* von (P).

7. Seien die symmetrische und positiv definite Matrix $Q \in \mathbb{R}^{n\times n}$, die positive Zahl C und der Vektor $f \in \mathbb{R}^n$ gegeben. Hiermit betrachte man die Aufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x,\eta) := \tfrac{1}{2} x^T Q x + C\eta \quad \text{auf} \\ M := \{(x,\eta) \in \mathbb{R}^n \times \mathbb{R} : -\eta e \le Qx - f \le \eta e\}, \end{array} \right.$$

[5] Siehe R. Fletcher (1987, S. 255).
[6] Siehe R. Fletcher (1987, S. 257).

wobei $e \in \mathbb{R}^n$ der Vektor ist, dessen Komponenten sämtlich gleich 1 sind[7]. Man zeige, dass (P) genau eine Lösung $(x^*, \eta^*) \in M$ besitzt.

8. Man entwickle ein Verfahren zur Bestimmung der Projektion eines Vektors $z \in \mathbb{R}^n$ auf das Simplex $\Sigma := \{x \in \mathbb{R}^n : x \geq 0,\ e^T x \leq 1\}$. Anschließend wende man das Verfahren auf den Spezialfall $n := 4$, $z := (1, 5, 3, 2)^T$ an.

   Hinweis: Bei gegebenem $\lambda \in \mathbb{R}$ bestimme man $x(\lambda) \in \mathbb{R}^n$ mit

   $$x(\lambda) \geq 0, \qquad x(\lambda) - z + \lambda e \geq 0, \qquad (x(\lambda) - z + \lambda e)^T x(\lambda) = 0.$$

   Man zeige: Ist $e^T x(0) \leq 1$, so ist $x^* := x(0)$ die Lösung von (P). Andernfalls bestimme man $\lambda^* > 0$ als positive Nullstelle von $h(\lambda) := e^T x(\lambda) - 1$ und zeige, dass $x^* = x(\lambda^*)$ die gesuchte Projektion auf das Simplex ist. Wie λ^* berechnet werden kann, zeige man wenigstens für den angegebenen Spezialfall.

5.2 Das duale Verfahren von Goldfarb-Idnani

In diesem Abschnitt betrachten wir wieder das zu Beginn dieses Kapitels angegebene quadratische Programmn (P), wobei die Matrix Q als positiv definit und $a_i \in \mathbb{R}^n \setminus \{0\}$, $i = 1, \ldots, m$, vorausgesetzt wird. Ziel wird es sein, das duale Verfahren von Goldfarb-Idnani, siehe D. Goldfarb, A. Idnani (1982, 1983) in seinen Grundzügen darzustellen. 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 nächsten Unterabschnitt werden wir einige für das Verfahren wichtige Begriffe einführen und das Verfahren in seiner Grundform beschreiben. In dem zweiten Unterabschnitt erfolgt eine genauere Beschreibung des Verfahrens von Goldfarb-Idnani, wobei wir uns zu Gunsten einer einfacheren Notation auf den Fall beschränken werden, dass keine Gleichungsrestriktionen auftreten.

5.2.1 Prinzipielle Beschreibung des Verfahrens

Bei einer gegebenen Indexmenge $I \subset \{1, \ldots, m\}$ definieren wir das (im Vergleich zu (P)) relaxierte[8] quadratische Programm (P_I) durch

$$(\mathrm{P}_I) \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x \quad \text{auf} \\ M_I := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \leq b_i & (i \in I \cap \{1, \ldots, m_0\}) \\ a_i^T x = b_i & (i \in I \cap \{m_0 + 1, \ldots, m\}) \end{array} \right\}. \end{array} \right.$$

Als grundlegend wird sich die folgende Definition herausstellen.

[7] Probleme dieser Art treten bei Problemen des Maschinellen Lernens auf. Siehe R. Schaback, J. Werner (2006).

[8] Bei einem relaxierten Problem werden Restriktionen aufgegeben, die Menge der zulässigen Lösungen wird also vergrößert.

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. $\{a_i\}_{i \in I} \subset \mathbb{R}^n$ linear unabhängig,
2. $M_I \neq \emptyset$,
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, da es höchstens so viele Lösungspaare wie Teilmengen von $\{1, \ldots, m\}$ 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, wie wir in einem Satz im Anschluss zeigen werden.

Ist $I := \{i_1, \ldots, i_q\} \subset \{1, \ldots, m\}$ eine Indexmenge mit $q := \#(I)$ Elementen[9], 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.

Wegen der notwendigen Optimalitätsbedingungen in Satz 2.1 (Kuhn-Tucker) und der hinreichenden Optimalitätsbedingung in Satz 2.6 in Abschnitt 3.2 gilt: 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 $\mathrm{Rang}\,(A_I) = q$, $A_I x = b_I$ und ein $y_I \in \mathbb{R}^q$ mit $c + Qx + A_I^T y_I = 0$ und $y_i \geq 0$ für alle $i \in I \cap \{1, \ldots, m_0\}$ existiert.

Satz 2.2 *Gegeben sei das obige quadratische Programm (P) mit symmetrischem, positiv definiten $Q \in \mathbb{R}^{n \times n}$ und $a_i \in \mathbb{R}^n \setminus \{0\}$, $i = 1, \ldots, m$. Sei (P) zulässig, $x^* \in M$ sei die eindeutige Lösung von (P). Dann existiert eine Indexmenge $I^* \subset \{1, \ldots, m\}$ derart, dass (x^*, I^*) ein Lösungspaar für (P) ist.*

Beweis: 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

$$-(c + Qx^*) \in \Big\{ A_I^T y_I : y_i \geq 0 \quad (i \in I \cap \{1, \ldots, m_0\}) \Big\}.$$

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:

[9] Diese Bezeichnung werden wir beibehalten: Ist $I \subset \{1, \ldots, m\}$ eine Indexmenge, so sei grundsätzlich $q := \#(I)$ die Anzahl der Elemente von I.

1. Es ist Rang $(A_{I^*}) = \#(I^*)$.

Wäre Rang $(A_{I^*}) < \#(I^*)$, so existierten $z_{I^*} \neq 0$ mit $A_{I^*}^T z_{I^*} = 0$. Nach Voraussetzung existiert y_{I^*} mit $c + Qx^* + A_{I^*}^T y_{I^*} = 0$ und $y_i \geq 0$ 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 $y_i^* + tz_i = 0$ für wenigstens ein $i \in I^*$ und $y_i^* + tz_i \geq 0$ für alle $i \in I^* \cap \{1, \ldots, m_0\}$. Dies ist aber ein Widerspruch zur Minimalität von I^*.

2. Es ist $A_{I^*} x^* = b_{I^*}$.

Dies ist richtig, da $A_{I_0} x^* = b_{I_0}$ und $I^* \subset I_0$.

3. 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. □

Bemerkung: Das zu (P) duale Programm[10] lautet

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \phi(y) := -b^T y - \frac{1}{2}(c + A^T y)^T Q^{-1}(c + A^T y) \quad \text{auf} \\ N := \{y \in \mathbb{R}^m : y_i \geq 0 \quad (i = 1, \ldots, m_0)\}. \end{cases}$$

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

$$\begin{aligned} \phi(y) &= -b^T y - \tfrac{1}{2}(c + A^T y)^T Q^{-1}(c + A^T y) \\ &= -b_I^T y_I - \tfrac{1}{2}(c + A_I^T y_I)^T Q^{-1}(c + A_I^T y_I) \\ &= -(A_I^T y_I)^T x - \tfrac{1}{2}(Qx)^T Q^{-1}(Qx) \\ &= (c + Qx)^T x - \tfrac{1}{2} x^T Q x \\ &= f(x). \end{aligned}$$

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 geben wir an, wie im Prinzip ein Schritt des Goldfarb-Idnani-Verfahrens zur Lösung des quadratischen Programms (P) mit strikt konvexer Zielfunktion aussieht.

- Sei (x, I) ein Lösungspaar[11].

[10] Die Lagrange-Funktion zu (P) ist

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

Da Q positiv definit, nimmt $L(\cdot, y)$ für jedes $y \in \mathbb{R}^m$ sein unrestringiertes Minimum an, und zwar dort, wo der Gradient $\nabla_x L(\cdot, y)$ verschwindet. Daher ist

$$\phi(y) := \inf_{x \in \mathbb{R}^n} L(x, y) = -b^T y - \tfrac{1}{2}(c + A^T y)^T Q^{-1}(c + A^T y).$$

[11] Zu Beginn des Verfahrens ist $(x, I) := (-Q^{-1}c, \emptyset)$.

- Falls $x \in M$, dann: STOP, da x die Lösung von (P) ist.
- Andernfalls:
  - Bestimme eine verletzte Restriktion $p \in \{1, \ldots, m\} \setminus I$.
  - Falls $M_{I \cup \{p\}} = \emptyset$, dann: STOP, da (P) nicht zulässig.
  - Andernfalls: Bestimme Lösungspaar (x_+, I_+) mit $I_+ = \overline{I} \cup \{p\}$, $\overline{I} \subset I$ und $f(x_+) > f(x)$.

Bemerkung: 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. □

Im folgenden Satz geben wir eine prinzipielle (also ohne weitere Angaben nicht praktikable) Methode an, ein Lösungspaar (x_+, I_+) mit Kosten $f(x_+) > f(x)$ zu gewinnen.

Satz 2.3 *Gegeben sei das obige quadratische Programm (P) mit symmetrischem, positiv definiten $Q \in \mathbb{R}^{n \times n}$ und $a_i \in \mathbb{R}^n \setminus \{0\}$, $i = 1, \ldots, m$. Sei (x, I) ein Lösungspaar, $p \in \{1, \ldots, m\} \setminus I$ eine durch x verletzte Restriktion und $a_p \notin \operatorname{span}\{a_i : i \in I\}$ (und damit $\{a_i\}_{i \in I \cup \{p\}}$ linear unabhängig und $M_{I \cup \{p\}} \neq \emptyset$). Sei x_+ die (eindeutige) Lösung von $(P_{I \cup \{p\}})$, $\overline{I} := \{i \in I : a_i^T x_+ = b_i\}$ und $I_+ := \overline{I} \cup \{p\}$. Dann ist (x_+, I_+) ein Lösungspaar mit $f(x_+) > f(x)$ ist.*

Beweis: Klar ist, da $I_+ \subset I \cup \{p\}$, 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 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

$$c + Q x_+ + A_{I \cup \{p\}}^T y_+ = 0, \qquad (A_{I \cup \{p\}} x_+ - b_{I \cup \{p\}})^T y_+ = 0$$

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. Nach Definition von $\overline{I}$ folgt aus der Gleichgewichtsbedingung, dass $(y_+)_i = 0$ für $i \in I \setminus \overline{I}$ und daher

$$c + Q x_+ = - \sum_{i \in I \cup \{p\}} (y_+)_i a_i = - \sum_{i \in \overline{I} \cup \{p\}} (y_+)_i a_i = - \sum_{i \in I_+} (y_+)_i a_i.$$

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 $y_I \in \mathbb{R}^q$:

$$\begin{aligned} f(x_+) &= f(x) + (c+Qx)^T(x_+ - x) + \tfrac{1}{2}\underbrace{(x_+ - x)^T Q(x_+ - x)}_{>0} \\ &> f(x) + (c+Qx)^T(x_+ - x) \\ &= f(x) - (A_I^T y_I)^T(x_+ - x) \\ &= f(x) - y_I^T(A_I x_+ - b_I) \\ &= f(x) - \sum_{i \in I \cap \{1,\ldots,m_0\}} \underbrace{y_i}_{\geq 0}\underbrace{(a_i^T x_+ - b_i)}_{\leq 0} \\ &\geq f(x), \end{aligned}$$

womit auch $f(x_+) > f(x)$ bewiesen ist. Hierbei haben wir am Schluss beachtet, dass $x_+ \in M_{I\cup\{p\}}$, insbesondere also $a_i^T x_+ \leq b_i$, $i \in I \cap \{1, \ldots, m_0\}$. □

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 und folglich $M_{I\cup\{p\}} \neq \emptyset$. 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. Ferner wird getestet, ob $M_{I\cup\{p\}} = \emptyset$.

Beide Fälle werden im folgenden Unterabschnitt getrennt untersucht, wobei wir der einfacheren Notation wegen uns auf den Fall beschränken werden, dass alle Restriktionen Ungleichungsrestriktionen sind.

5.2.2 Genauere Beschreibung des Verfahrens

Gegeben sei das durch lineare Ungleichungen restringierte quadratische Programm

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

wobei $A \in \mathbb{R}^{m\times n}$ nicht verschwindende Zeilen $a_1^T, \ldots, a_m^T$ habe, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit ist. Das Verfahren von Goldfarb-Idnani wird im wesentlichen durch die Aussagen in den folgenden beiden Lemmata beschrieben.

Lemma 2.4 *Sei (x, I) ein Lösungspaar für das quadratische Programm (P) mit einem zugehörigen Langrange-Vektor $y_I = (y_i)_{i\in I} \in \mathbb{R}^q$, $p \in \{1, \ldots, m\} \setminus I$ eine durch x verletzte Restriktion, also $a_p^T x > b_p$, 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

$$r_I := \begin{cases} (A_I Q^{-1} A_I^T)^{-1} A_I Q^{-1} a_p, & \text{falls } I \neq \emptyset, \\ 0 & \text{sonst,} \end{cases}$$

sowie

$$z := Q^{-1}(a_p - A_I^T r_I), \qquad t_1 := \frac{a_p^T x - b_p}{a_p^T z}.$$

(2) Falls $I = \emptyset$ oder $y_I - t_1 r_I \geq 0$, 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 neues Lösungspaar (x_+, I_+) mit zugehörigem Lagrange-Vektor y_{I_+} und dem Zielfunktionswert

$$f_+ := f(x_+) = f + t_1(\tfrac{1}{2}t_1 + \theta)a_p^T z > f(x) = f$$

gegeben.

(3) Andernfalls: Berechne

$$t_2 := \min\Big\{\frac{y_i}{r_i} : i \in I,\ r_i > 0\Big\} = \frac{y_l}{r_l}.$$

Anschließend setze

$$x_- := x - t_2 z, \qquad I_- := I \setminus \{l\}, \qquad y_{I_-} := T_l(y_I - t_2 r_I),$$

wobei T_l die Komponente mit dem Index l entferne, sowie

$$f_- := f + t_2(\tfrac{1}{2}t_2 + \theta)a_p^T z, \qquad \theta_- := \theta + t_2.$$

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 $(A_I) = \#(I)$ und p der Index einer durch x verletzten Restriktion mit $a_p \notin \text{span}\{a_i : i \in I\}$. Ferner gelte

$$A_I x = b_I, \qquad c + Qx + A_I^T y_I + \theta a_p = 0, \qquad y_I \geq 0$$

sowie

$$f = f(x), \qquad \theta \geq 0.$$

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 $a_p \notin \text{span}\{a_i : i \in I\}$ 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 $t_1 > 0$. 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

$$(*) \qquad a_p^T x(t) = a_p^T x - t a_p^T z = b_p + a_p^T z \Big(\underbrace{\frac{a_p^T x - b_p}{a_p^T z}}_{=t_1} - t \Big).$$

Ferner ist nach einfacher Rechnung

$$c + Qx(t) + A_I^T(y_I - t r_I) + (\theta + t) a_p = 0, \qquad f(x(t)) = f(x) + t(\tfrac{1}{2}t + \theta) a_p^T z.$$

Hieraus folgt: Ist $I = \emptyset$ oder $y_I - t_1 r_I \geq 0$, 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. Der zugehörige Funktionwert ist

$$f_+ = f(x(t_1)) = f(x) + \underbrace{t_1}_{>0}(\tfrac{1}{2}t_1 + \underbrace{\theta}_{\geq 0}) \underbrace{a_p^T z}_{>0} > f(x) = f.$$

Ist dagegen $I \neq \emptyset$ und $y_I - t_1 r_I \not\geq 0$, so wird t_2 in Schritt (3) berechnet, es ist $0 \leq t_2 < t_1$, $y_I - t_2 r_I \geq 0$ und $y_l - t_2 r_l = 0$. Im Algorithmus wird in Schritt (3) das neue 5-Tupel $(x_-, I_-, y_{I_-}, f_-, \theta_-)$ berechnet. Wegen $I_- := I \setminus \{l\}$ und Rang $(A_I) = \#(I)$

ist Rang $(A_{I_-}) = \#(I_-)$. Ferner verletzt auch $x_- := x - t_2 z$ die p-te Restriktion wegen (benutze $(*)$)

$$a_p^T x_- = a_p^T x(t_2) = b_p + \underbrace{a_p^T z}_{>0}\underbrace{(t_1 - t_2)}_{>0} > b_p.$$

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 $l \in I$ bestimmt, $I_- := I \setminus \{l\}$ gesetzt und ein Quintupel $(x_-, I_-, y_{I_-}, f_-, \theta_-)$ berechnet, mit dem in in Schritt (1) des Verfahrens aus Lemma 2.4 eingestiegen werden kann.

Lemma 2.5 *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 $r_I \leq 0$, so ist $(P_{I\cup\{p\}})$ und damit auch (P) nicht zulässig.*
2. *Ist $r_I \not\leq 0$, so bestimme man $l \in I$ mit $r_l > 0$ und*

$$t_2 := \min\Big\{\frac{y_i}{r_i} : i \in I,\ r_i > 0\Big\} = \frac{y_l}{r_l} \geq 0.$$

Setzt man anschließend

$$x_- := x, \qquad I_- := I \setminus \{l\}, \qquad y_{I_-} := T_l(y_I - t_2 r_I), \qquad \theta_- := t_2,$$

wobei T_l wieder aus einem Vektor die Komponente mit dem Index l entfernt, so ist $a_p \notin \operatorname{span}\{a_i : i \in I_-\}$ und

$$A_{I_-} x_- = b_{I_-}, \qquad c + Q x_- + A_{I_-}^T y_{I_-} + \theta_- a_p = 0$$

sowie

$$y_{I_-} \geq 0, \qquad \theta_- \geq 0.$$

Beweis: Nach Voraussetzung ist $a_p \in \operatorname{span}\{a_i : i \in I\}$ und Rang $(A_I) = \#(I)$. Daher besitzt a_p eine eindeutige Darstellung der Form $a_p = A_I^T \lambda_I$. 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 an, es sei $r_I \leq 0$. 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 $a_i^T z \leq 0$, $i \in I$. Das Erfülltsein der p-ten Restriktion durch $x + z$ impliziert $a_p^T z \leq b_p - a_p^T x < 0$. Andererseits ist

$$a_p^T z = r_I^T A_I z = \sum_{i\in I} \underbrace{r_i}_{\leq 0} \underbrace{a_i^T z}_{\leq 0} \geq 0.$$

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 $r_I \not\leq 0$ und bestimmen, wie angegeben, $l \in I$ sowie (x_-, I_-, y_{I_-}). Wegen $r_l \neq 0$ (genauer ist $r_l > 0$), $I_- := I \setminus \{l\}$ 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

$$c + Qx_- = -A_I^T y_I = -\sum_{i\in I^-} \Big(y_i - \frac{y_l}{r_l} r_i\Big)\, a_i - \frac{y_l}{r_l}\, a_p = -A_{I_-}^T y_{I_-} - \theta_-\, a_p.$$

Die restlichen Aussagen gelten nach Wahl des Index l. Das Lemma ist bewiesen. □

Damit ist das Verfahren von Goldfarb-Idnani genauer beschrieben. Wir fassen die Schritte zusammen.

- Input: Gegeben sind die Daten (A, b, c, Q) des quadratischen Programms

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

  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 eine von x verletzte Restriktion $p \in \{1, \ldots, m\} \setminus I$ und setze $\theta := 0$.

  (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 $z := H_I a_p$ und, falls $I \neq \emptyset$, $r_I := N_I a_p$.

(3) Bestimme primale und duale Schrittweiten.
Setze

$$t_1 := \begin{cases} \dfrac{a_p^T x - b_p}{a_p^T z} & \text{für} \quad z \neq 0, \\ \infty & \text{für} \quad z = 0. \end{cases}$$

Setze

$$t_2 := \begin{cases} \infty & \text{falls } I = \emptyset \text{ oder } r_I \leq 0, \\ \dfrac{y_l}{r_l} & \text{sonst,} \end{cases}$$

wobei

$$\frac{y_l}{r_l} = \min\Big\{ \frac{y_i}{r_i} : i \in I,\ r_i > 0 \Big\}.$$

Anschließend berechne

$$t := \min(t_1, t_2).$$

(4) Test auf Unzulässigkeit.
Ist $t = \infty$, dann: STOP, (P) ist nicht zulässig.

(5) Dualer Schritt.
Falls $t_1 = \infty$, 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

$$y_{I\cup\{p\}} := \begin{pmatrix} y_I - t r_I \\ \theta \end{pmatrix}, \qquad I := I \cup \{p\}, \qquad q := q + 1$$

und gehe nach (1).

(b) Ist $t = t_2$, so setze

$$y_{I\setminus\{l\}} := T_l(y_I - t r_I), \qquad I := I \setminus \{l\}, \qquad q := q - 1$$

und gehe nach (2).

- Output: Das Verfahren bricht (bei exakter Arithmetik) nach einer endlichen Zahl von Schritten mit der Lösung von (P) und einem zugehörigen Lagrange-Vektor ab oder es liefert die Information, dass (P) nicht zulässig und daher nicht lösbar ist.

5.2.3 Hinweise zur Implementation des Verfahrens

Um die Notation möglichst einfach zu halten, betrachten wir wie im letzten Unterabschnitt quadratische Programme, bei denen alle Restriktionen in Ungleichungsform vorliegen, die also die Form

$$\text{Minimiere} \quad f(x) := c^T x + \tfrac{1}{2} x^T Q x \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Ax \leq b\}$$

haben, wobei $Q \in \mathbb{R}^{n \times n}$ als symmetrisch und positiv definit vorausgesetzt wird. Zu Beginn des Verfahrens von Goldfarb-Idnani wird die Cholesky-Zerlegung $Q = U^T U$ von Q berechnet, wodurch gleichzeitig getestet wird, ob Q positiv definit ist. Anschließend berechnen wir $Z := U^{-1}$. Insbesondere ist dann $ZZ^T = Q^{-1}$, was sich später als wichtig herausstellen wird. Danach 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 wird zunächst getestet, ob x zulässig ist. Ist dies nicht der Fall, so wird eine durch x verletzte Restriktion $p \in \{1, \ldots, m\} \setminus I$ bestimmt. Wie schon früher wird mit q die Anzahl der Elemente der Indexmenge I bezeichnet. 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 $r_I := N_I a_p$. Von Schritt zu Schritt verändert sich die Indexmenge I um genau ein Element, was für eine effiziente Implementation ausgenutzt werden sollte. Zu unterscheiden sind 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. Sind z. B. N_I und H_I bekannt, damit auch z und r_I, ist ferner $z \neq 0$ und damit $\{a_i\}_{i \in I \cup \{p\}}$ linear unabhängig, so ist

$$N_{I \cup \{p\}} = \begin{pmatrix} N_I - \dfrac{r_I z^T}{a_p^T z} \\ \dfrac{z^T}{a_p^T z} \end{pmatrix}, \qquad H_{I \cup \{p\}} = H_I - \frac{z z^T}{a_p^T z}.$$

Nur von der Notation her etwas schwieriger ist der Fall, dass aus I ein Index l entfernt wird. 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. 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, 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 getrennt untersucht, wobei jeweils 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 die erste Bedingung in $(*)$ automatisch erfüllt.

In beiden Fällen spielen Givens-Rotationen, also spezielle orthogonale Matrizen, eine besondere Rolle. 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}. Bei einer Umsetzung in MATLAB ist die Funktion `planerot` wichtig, auf die wir schon im Zusammenhang mit der Implementation des BFGS-Verfahrens hingewiesen hatten.

Zunächst betrachten wir den Fall, dass $I_+ := I \cup \{p\}$ 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 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

$$R_{I\cup\{p\}} := \left.\begin{pmatrix} R_I & d_I^{(1)} \\ 0^T & \delta_I \end{pmatrix}\right\} \; {\scriptstyle q+1}$$

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.

Nun betrachten wir den Fall, dass $I_+ := I \setminus \{l\}$ mit einem $l \in I$. Sei l das k-te Element von I mit $1 \le k \le q := \#(I)$. 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 $R_I \in \mathbb{R}^{q\times q}$ durch Streichen der k-ten Spalte entsteht. 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:

$$Z_I^{(1)} = (\ \underbrace{Z_{k-1}^{(1)}}_{k-1} \ \underbrace{Z_{q-k+1}^{(1)}}_{q-k+1}\ ).$$

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

$$d_{I\setminus\{l\}} = \Omega_I d_I = \begin{pmatrix} I_{k-1} & 0 & 0 \\ 0 & \Omega_I^{(2)} & 0 \\ 0 & 0 & I_{n-q} \end{pmatrix} \begin{pmatrix} d_{k-1}^{(1)} \\ d_{q-k+1}^{(1)} \\ d_I^{(2)} \end{pmatrix} \begin{pmatrix} d_{k-1}^{(1)} \\ \Omega_I^{(2)} d_{q-k+1}^{(1)} \\ d_I^{(2)} \end{pmatrix}.$$

Wir haben eine MATLAB-Funktion `GO_Id` geschrieben, die mit dem Verfahren von Goldfarb-Idnani das quadratische Programm (P) löst (nach wie vor nehmen wir an, dass keine Gleichungen als Restriktionen auftreten, wobei diese Vereinfachung eigentlich nur den vorigen Unterabschnitt betrifft). Der Kopf dieser Funktion kann folgendermaßen aussehen:

```
function [x,f_min,I,y_I,info]=Go_Id(Q,c,A,b);
%******************************************************
%Die Funktion Go_Id loest das quadratische Programm
%Minimiere c'*x+0.5*x'*Q*x unter der Nebenbedingung
%               Ax<=b
%******************************************************
%Input-Parameter:
% Q     symmetrische, positiv definite n x n-Matrix
% c     n-Vektor
% A     m x n-Matrix
% b     m-Vektor
%Output-Parameter:
% x     Bei erfolgreichem Ausgang: eindeutige L"osung
% f_min minimaler Funktionswert
% I     optimale Indexmenge (Zeilenvektor): (x,I) ist
%       optimales Loesungspaar
% y_I   optimaler Lagrange-Vektor
% info  Ist info>0, so ist das Verfahren erfolgreich und
%       info gibt die Anzahl der berechneten
%       Loesungspaare an. Ist info=-1, so ist das
%       Problem nicht zulaessig.
%******************************************************
```

Die Funktion selber wollen wir hier nicht angeben. Wir haben sie an verschiedenen Beispielen getestet. U. a. an einem Beispiel von D. Goldfarb, A. Idnani (1982), das durch die Daten

$$Q := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad c := \begin{pmatrix} 0 \\ -5 \\ 0 \end{pmatrix}, \quad A := \begin{pmatrix} 4 & 3 & 0 \\ -2 & -1 & 0 \\ 0 & 2 & -1 \end{pmatrix}, \quad b := \begin{pmatrix} 8 \\ -2 \\ 0 \end{pmatrix}$$

gegeben ist. Der Aufruf

```
[x,f_min,I,y_I,info]=Go_Id(Q,c,A,b);
```

liefert (`format long`):

$$x = \begin{pmatrix} 0.476190476190476 \\ 1.047619047619048 \\ 2.095238095238095 \end{pmatrix}, \qquad f_{\min} = -2.380952380952381$$

sowie

$$I = \{3, 2\}, \qquad y_I = \begin{pmatrix} 2.095238095238095 \\ 0.238095238095238 \end{pmatrix}, \qquad \text{info} = 3.$$

Dies stimmt vorzüglich mit der exakten Lösung $x^* = \frac{1}{21}(10, 22, 44)^T$ überein. Weiter wurden einige Beispiele aus der Testsammlung von W. Hock, K. Schittkowski (1981) nachgerechnet, z. B. die auf S. 58, 96. Ist bei der Anwendung von `Go_Id` die Matrix Q positiv semidefinit, aber nicht positiv definit, so steigt unsere Funktion bei der Berechnung der Cholesky-Zerlegung mit einer Fehlermeldung aus. Häufig genügt es, ein kleines Vielfaches der Einheitsmatrix zu Q zu addieren.

5.2.4 Aufgaben

1. Mit Hilfe des Verfahrens von Goldfarb-Idnani löse man die quadratische Optimierungsaufgabe

$$
(\mathrm{P}) \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := \begin{pmatrix} -2 \\ -6 \end{pmatrix}^T x + \frac{1}{2} x^T \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} x \\ \text{unter der Nebenbedingung} \\ \begin{pmatrix} 1 & 1 \\ -1 & 2 \\ 2 & 1 \end{pmatrix} x \leq \begin{pmatrix} 2 \\ 2 \\ 3 \end{pmatrix}. \end{array} \right.
$$

Kapitel 6

Linear restringierte Optimierungsaufgaben

In diesem Kapitel betrachten wir Optimierungsaufgaben der Form

$$\text{(P)} \quad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \leq b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$

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 und zunächst auf die *Methode der aktiven Mengen*, danach auf *Verfahren der zulässigen Richtungen* eingehen.

6.1 Die Methode der aktiven Mengen

Den Darstellungen bei P. E. Gill, W. Murray, M. H. Wright (1981, S. 155 ff.) und R. Fletcher (1987, S. 259 ff.) folgend schildern wir in diesem Abschnitt die *Methode der aktiven Mengen*. Ähnlich wie in der quadratischen Optimierung (siehe das Verfahren von Fletcher in Abschnitt 5.1) wird hierbei eine Folge von Optimierungsaufgaben mit linearen Gleichungsrestriktionen gelöst. Daher beschäftigen wir uns mit diesen im nächsten Unterabschnitt.

6.1.1 Lineare Gleichungsrestriktionen

In diesem Unterabschnitt betrachten wir Aufgaben der Form

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

wobei $A \in \mathbb{R}^{m\times n}$ mit $\operatorname{Rang}(A) = m$. Natürlich kann man bei nichtquadratischem $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ nicht erwarten, dass man (P) in endlich vielen Schritten lösen kann. Es liegt aber nahe, dass man (P) auf eine unrestringierte Optimierungsaufgabe zurückführen kann. Wir nehmen hierzu an (gegenüber Unterabschnitt 5.1.1 verändern wir die Bezeichnungen), 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 $\operatorname{Kern}(A)$. Dann ist (P) äquivalent zu der *reduzierten Optimierungsaufgabe*

$$(\mathrm{P}_{\hat{x}}) \qquad \text{Minimiere} \quad \psi_{\hat{x}}(u) := f(\hat{x} + Zu), \quad u \in \mathbb{R}^{n-m}$$

wobei $\hat{x}$ eine spezielle Lösung von $Ax = b$ ist, z. B. $\hat{x} = Yb$. Zur numerischen Lösung der reduzierten Optimierungsaufgabe $(\mathrm{P}_{\hat{x}})$ kann im Prinzip jedes Verfahren der unrestringierten Optimierung herangezogen werden, also etwa Quasi-Newton-Verfahren und hier insbesondere das BFGS-Verfahren. Der Gradient von $\psi_{\hat{x}}(\cdot)$ ist durch

$$\nabla\psi_{\hat{x}}(u) = Z^T\nabla f(\hat{x} + Zu),$$

den sogenannten *reduzierten Gradienten*, die Hessesche durch

$$\nabla^2\psi_{\hat{x}}(u) = Z^T\nabla^2 f(\hat{x} + Zu)Z,$$

die *reduzierte Hessesche*, gegeben. Für den Zusammenhang zwischen (P) und $(\mathrm{P}_{\hat{x}})$ gilt:

- Es ist $u^* \in \mathbb{R}^{n-m}$ genau dann eine stationäre Lösung von $(\mathrm{P}_{\hat{x}})$, wenn $x^* := \hat{x} + Zu^* \in M$ der notwendigen Optimalitätsbedingung erster Ordnung für (P) genügt, also ein $y^* \in \mathbb{R}^m$ mit $\nabla f(x^*) + A^T y^* = 0$ existiert.

- In $u^* \in \mathbb{R}^{n-m}$ sind genau dann die notwendigen Optimalitätsbedingungen zweiter Ordnung für $(\mathrm{P}_{\hat{x}})$ erfüllt (also $\nabla\psi_{\hat{x}}(u^*) = 0$ und $\nabla^2\psi_{\hat{x}}(u^*)$ positiv semidefinit), wenn $x^* := \hat{x} + Zu^* \in M$ den notwendigen Optimalitätsbedingungen zweiter Ordnung für (P) genügt (es existiert $y^* \in \mathbb{R}^m$ mit $\nabla f(x^*) + A^T y^* = 0$ und $\nabla^2 f(x^*)$ ist positiv semidefinit auf $\operatorname{Kern}(A)$).

Denn: Es ist

$$\nabla\psi_{\hat{x}}(u^*) = 0 \iff \nabla f(x^*) \in \operatorname{Kern}(Z^T) = \operatorname{Bild}(Z)^\perp = \operatorname{Kern}(A)^\perp = \operatorname{Bild}(A^T)$$

und wegen $\operatorname{Kern}(A) = \operatorname{Bild}(Z)$ ist $\nabla^2\psi(u^*)$ 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 stationären Lösung x^* von (P) gehörenden Lagrange-Vektor (oder zumindestens eine Näherung) zu berechnen, also einen Vektor $y^* \in \mathbb{R}^m$ mit $\nabla f(x^*) + A^T y^* = 0$. Das ist einfach, wenn man die nichtsinguläre Matrix $(\ Y \quad Z\ )$ mit $AY = I$ und $AZ = 0$ bestimmt hat. Denn wir wissen schon,

dass $\nabla f(x^*) \in \text{Bild}\,(A^T)$, also $\nabla f(x^*) + A^T y^* = 0$ 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 daher eine Näherung y_k für y^* durch $y_k := -Y^T \nabla f(x_k)$.

Bemerkung: Die empfehlenswerteste Methode zur Berechnung einer nichtsingulären Matrix $\begin{pmatrix} Y & Z \end{pmatrix} \in \mathbb{R}^{n\times n}$ mit $Y \in \mathbb{R}^{n\times m}$, $Z \in \mathbb{R}^{n\times(n-m)}$ mit $AY = I$ und $AZ = 0$ ist die sogenannte *Orthogonalzerlegungsmethode*. Diese besteht darin, eine QR-Zerlegung 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. □

Beispiel: Eine einfache MATLAB-Funktion zur Orthogonalzerlegungsmethode könnte folgendermaßen aussehen:

```
function [Y,Z]=Orth_Zer(A);
%*********************************************
%Input: A    m x n-Matrix mit m<=n und rang(A)=m
%Output:Y    n x m-Matrix
%       Z    n x (n-m)-Matrix mit
%            AY=I  und AZ=0
%*********************************************
[m,n]=size(A);
[Q,R]=qr(A');R=R(1:m,:);
Y=Q(:,1:m)*(R'\eye(m));Z=Q(:,m+1:n);
%*********************************************
```

Eine Anwendung[1] auf $A := \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$ liefert (`format short`)

$$Y = \begin{pmatrix} 0.3333 \\ 0.3333 \\ 0.3333 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}$$

und

$$Z = \begin{pmatrix} -0.5774 & -0.5774 \\ 0.7887 & -0.2113 \\ -0.2113 & 0.7887 \end{pmatrix} = \begin{pmatrix} -\frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} \\ \frac{3+\sqrt{3}}{6} & -\frac{3-\sqrt{3}}{6} \\ -\frac{3-\sqrt{3}}{6} & \frac{3+\sqrt{3}}{6} \end{pmatrix}.$$

[1] Siehe P. E. Gill et al (1981, S. 156).

□

Wie sieht die Anwendung des BFGS-Verfahrens auf die unrestringierte Optimierungsaufgabe ($P_{\hat{x}}$) aus? Wir beschreiben einen Schritt, wobei es zweckmäßig ist, $\hat{x}$ stets als aktuelle Näherung im x-Raum zu nehmen. Wir gehen davon aus, dass eine nichtsinguläre Matrix $\begin{pmatrix} Y & Z \end{pmatrix}$ mit $AY = I$ und $AZ = 0$ bekannt ist.

- Sei $x \in M$ gegeben (zum Start z. B. $x := Yb$). Ferner sei $H \in \mathbb{R}^{(n-m)\times(n-m)}$ symmetrisch und positiv definit, $H \approx \nabla^2\psi_x(0) = Z^T\nabla^2 f(x)Z$.
- Berechne $\nabla\psi_x(0) = Z^T\nabla f(x)$. Falls $\nabla\psi_x(0) = 0$, STOP.
- Berechne die Quasi-Newton-Richtung p im u-Raum durch

$$p := -H^{-1}\nabla\psi_x(0) = -H^{-1}Z^T\nabla f(x).$$

- Berechne Schrittweite $t > 0$, z. B. die Wolfe-Schrittweite. Bei gegebenen $\alpha \in (0, \frac{1}{2})$, $\beta \in (\alpha, 1)$ ist hier $t > 0$ so zu bestimmen, dass

$$\psi_x(0 + tp) \le \psi_x(0) + \alpha t\nabla\psi_x(0)^T p, \qquad \nabla\psi_x(0 + tp) \ge \beta\nabla\psi_x(0)^T p$$

  bzw.

$$f(x + tZp) \le f(x) + \alpha t\nabla f(x)^T(Zp), \qquad \nabla f(x + tp)^T(Zp) \ge \beta\nabla f(x)^T(Zp).$$

- Sei

$$u_+ := 0 + tp = -tH^{-1}Z^T\nabla f(x)$$

  die neue Näherung im u-Raum,

$$x_+ := x + tZp = x - tZH^{-1}Z^T\nabla f(x)$$

  die neue Näherung im x-Raum.

- Sei $H_+ \in \mathbb{R}^{(n-m)\times(n-m)}$ das BFGS-Update. Mit

$$s := u_+ - 0 = u_+, \qquad y = \nabla\psi_x(u_+) - \nabla\psi_x(0) = Z^T[\nabla f(x_+) - \nabla f(x)]$$

  sei also

$$H_+ := H - \frac{(Hs)(Hs)^T}{s^T Hs} + \frac{yy^T}{y^T s}.$$

Im Verfahren macht man den Update $(x, H) := (x_+, H_+)$ und steigt wieder in den Anfang ein.

Wir haben ein einfaches MATLAB function file geschrieben, welches das eben gesagte umsetzt. Diese benutzt eine Funktion `Wolfe` zur Berechnung der Wolfe-Schrittweite und die Funktion `CholBFGS` zum Updaten der Cholesky-Zerlegung der BFGS-Matrizen.

```
function [x_stern,y_stern,f_stern,iter]=Lin_Glei(fun,A,b,max_iter,tol);
%************************************************************************
%Diese Funktion loest die Optimierungsaufgabe, die
%Funktion fun unter der Restriktion  A*x=b zu minimieren
%************************************************************************
%Input-Parameter
%  A           m x n Matrix mit Rang(A)=m
%  b           m-Vektor
%  fun         [f,grad]=fun(x) liefern Funktionswert und
%              Gradient von fun
%  max_iter    maximale Zahl der Iterationen
%  tol
%Output-Parameter
%  x_stern     Loesung
%  y_stern     Lagrange-Multiplikator
%  f_stern     Funktionswert in x_stern
%  iter        Anzahl der Iterationen
%************************************************************************
[m,n]=size(A);[Q,R]=qr(A');R=R(1:m,:);Y=Q(:,1:m)*(R'\eye(m));Z=Q(:,m+1:n);
x=Y*b;[f,g]=feval(fun,x);iter=0;red_grad=Z'*g;
L=sqrt(abs(f))*eye(n-m);
while (norm(red_grad)>tol)&(iter<max_iter)
   p=-(L'\(L\red_grad));%Richtung im u-Raum
   q=Z*p;               %Richtung im x-Raum
   t=Wolfe(fun,x,q);x_plus=x+t*q;
   [f_plus,g_plus]=feval(fun,x_plus); red_grad_plus=Z'*g_plus;
   s=t*p;y=red_grad_plus-red_grad;
   L=CholBFGS(L,y,s);red_grad=red_grad_plus;x=x_plus;
   iter=iter+1;
end;
if (norm(red_grad)<=tol)
   x_stern=x;[f,g]=feval(fun,x);
   y_stern=-Y'*g;f_stern=f;
end;
```

Beispiel: Wir wollen obige Funktion an einem Beispiel testen. Dieses ist in der am Schluss von Kapitel 5 erwähnten Testsammlung von W. Hock, K. Schittkowski (1981, S. 73) enthalten. Es lautet:

$$\begin{cases} \text{Minimiere} \quad f(x) := (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_4)^4 + (x_4 - x_5)^2 \\ \begin{pmatrix} 1 & 2 & 3 & 0 & 0 \\ 0 & 1 & 2 & 3 & 0 \\ 0 & 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \\ 6 \end{pmatrix}. \end{cases}$$

Die eindeutige Lösung ist $x^* = (1, 1, 1, 1, 1)^T$. Der Aufruf

```
[x,y,f,iter]=Lin_Glei('myfun',A,b,50,1e-10);
```

liefert, nachdem `myfun`, `A` und `b` entsprechend erklärt sind, die Werte (`format long`)

$$x = \begin{pmatrix} 1.000000000000124 \\ 1.000000000000056 \\ 0.999999999999922 \\ 1.000000000000034 \\ 1.000000000000003 \end{pmatrix}, \quad y = 10^{-13} \begin{pmatrix} 0.340939397925798 \\ -0.195298322968153 \\ 0.273114864057789 \end{pmatrix}, \quad \text{iter} = 8.$$

Zum Vergleich wenden wir die Funktion `fmincon` aus der Optimization-Toolbox von MATLAB an. Nachdem wir

```
options=optimset('GradObj','on');
x=fmincon('myfun',[35;-31;11;5;-5],[],[],A,b,[],[],[],options);
```

eingegeben haben, wobei durch `myfun` Funktionswert und Gradient obiger Zielfunktion geliefert und A, b entsprechend besetzt sind (die Funktion `fmincon` *verlangt* einen Startwert, wir nehmen denselben wie Hock-Schittkowski), erhalten wir

$$x = \begin{pmatrix} 0.999999999999999 \\ 0.999999999999999 \\ 1.000000000000002 \\ 0.999999999999999 \\ 1.000000000000000 \end{pmatrix},$$

also noch etwas bessere Werte als unsere obigen. □

6.1.2 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 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 $x \in M$, $\{m_0 + 1, \ldots, m\} \subset I \subset I(x)$ und $\text{Rang}\,(A_I) = \#(I)$.

2. Bestimme eine Lösung $p^* \in \mathbb{R}^n$ und einen zugehörigen Lagrange-Multiplikator $y_I^* \in \mathbb{R}^q$ der Optimierungsaufgabe

   $$\text{Minimiere} \quad \psi_x(p) := f(x+p) \quad \text{unter der Nebenbedingung} \quad A_I p = 0.$$

   Insbesondere ist also $\nabla f(x+p^*) + A_I^T y_I^* = 0$, $A_I p^* = 0$.

   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 stationäre 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 $t^* \in (0, s(x, p^*)]$ 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.

Bemerkungen: In Schritt 2 braucht der Lagrange-Vektor nur dann ausgerechnet zu werden, wenn $p^* = 0$.

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 stationäre 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, $\psi_x(p)$ 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

$$0 < c \, \|p^{**}\|^2 \leq [\nabla f(x + p^{**}) - \nabla f(x)]^T p^{**} = y_l^* \, a_l^T p^{**},$$

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 $\nabla f(x+p^*) + A_I^T y_I^* = 0$, woraus man mit $A_I p^* = 0$ erhält, dass $\nabla f(x+p^*)^T p^* = 0$. Mit einer positiven Konstanten c ist daher

$$0 < c\,\|p^*\|^2 \le [\nabla f(x+p^*) - \nabla f(x)]^T p^* = -\nabla f(x)^T p^*$$

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) eine bisher inaktive Ungleichungsrestriktion aktiv. □

Beispiel: Wir betrachten die Aufgabe

$$\begin{cases} \text{Minimiere} \quad f(x) := (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2 \quad \text{unter der Nebenbedingung} \\ \qquad\qquad \begin{pmatrix} 2 & -1 \\ -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \le \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}. \end{cases}$$

In Abbildung 6.1 stellen wir die Menge M der zulässigen Lösungen und einige Niveaulinien dar. Die Lösung ist offenbar $x^* = (2.5, 1)^T$.

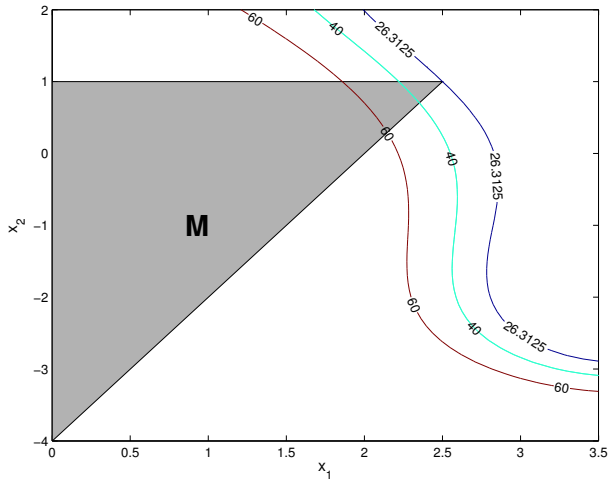


Abbildung 6.1: Eine linear restringierte nichtlineare Optimierungsaufgabe

Nun wenden wir obiges Verfahren der aktiven Mengen an.

1. Wir starten mit $(x, I) := ((0,1)^T, \{2\})$, der zugehörige Zielfunktionswert ist $f(x) = 136$. Als Lösung p^* der Aufgabe

$$\begin{cases} \text{Minimiere} \quad \psi_x(p) = (p_1^2 + 1 + p_2 - 11)^2 + (p_1 + (1+p_2)^2 - 7)^2 \\ \qquad\qquad \text{unter der Nebenbedingung } -p_1 = 0 \end{cases}$$

erhalten wir

$$p^* = (0, -\frac{1}{2} + \frac{1}{2}\sqrt{23})^T \approx (0, 1.8979)^T.$$

Da $p^* \neq 0$, braucht der zugehörige Lagrange-Multiplikator nicht ausgerechnet zu werden. Es ist

$$s(x, p^*) = \frac{b_3 - a_3^T x}{a_3^T p^*} = \frac{1-1}{p_2^*} = 0.$$

2. Im zweiten Schritt ist daher $x := x + s(x, p^*)p^* = (0, 1)$, $I := I \cup \{3\} = \{2, 3\}$. Im nächsten Schritt hat man $\psi_x(p)$ unter der Nebenbedingung $A_I p = 0$ zu minimieren. Da $A_{\{2,3\}}$ nichtsingulär ist, ist natürlich $p^* = 0$. Den zugehörigen Lagrange-Multiplikator y_I^* erhalten wir aus

$$0 = \nabla f(x) + A_I^T y_I^* = \begin{pmatrix} -12 \\ -44 \end{pmatrix} + \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} y_I^*,$$

es ist also $y_I^* = (-12, 44)^T$.

3. Im dritten Schritt ist nach wie vor $x = (0, 1)^T$ und $I := I \setminus \{2\} = \{3\}$. Als Lösung p^* der Aufgabe

$$\begin{cases} \text{Minimiere} \quad \psi_x(p) = (p_1^2 + 1 + p_2 - 11)^2 + (p_1 + (1 + p_2)^2 - 7)^2 \\ \qquad\qquad \text{unter der Nebenbedingung } p_2 = 0 \end{cases}$$

erhalten wir $p^* \approx (3.2294, 0)^T$. Da $p^* \neq 0$, braucht der zugehörige Lagrange-Multiplikator wieder nicht berechnet zu werden. Die maximale Schrittweite ist

$$s(x, p^*) = \frac{b_1 - a_1^T x}{a_1^T p^*} = \frac{4 - (-1)}{2p_1^*} = \frac{5}{2 \cdot 3.2294} \approx 0.7741.$$

4. Da $f(x + tp^*)$ auf $[0, s(x, p^*)]$ monoton fallend ist, ist im vierten Schritt

$$x := x + s(x, p^*)p^* = (\frac{5}{2}, 1)^T, \qquad I := I \cup \{1\} = \{3, 1\},$$

der zugehörige Funktionswert ist $f(x) = 26.3125$. Als Lösung der Aufgabe, $\psi_x(p)$ unter Nebenbedingung $A_I p = 0$ zu minimieren, erhält man $p^* = 0$, da $A_{\{3,1\}}$ nichtsingulär ist. Aus

$$0 = \nabla f(x) + A_I^T y_I^* = \begin{pmatrix} -\frac{89}{2} \\ -\frac{43}{2} \end{pmatrix} + \begin{pmatrix} 0 & 2 \\ 1 & -1 \end{pmatrix} y_I^*$$

erhalten wir den zugehörigen Multiplikator $y_I^* = (\frac{89}{4}, \frac{175}{4})^T$. Da dieser nichtnegativ ist, ist $x^* = (\frac{5}{2}, 1)^T$ zumindestens eine stationäre Lösung der gegebenen linear restringierten Optimierungsaufgabe.

□

6.1.3 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

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

Hierbei seien $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$ und $b \in \mathbb{R}^m$ gegeben. Wie in Unterabschnitt 6.1.1 geschildert ordne man (P) die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}_x) \qquad \text{Minimiere} \quad \psi(u) := f(x + Zu), \quad u \in \mathbb{R}^{n-m},$$

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 $\mathbb{R}^{n-m}$.

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 6.1.1 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.

6.2 Verfahren der zulässigen Richtungen

6.2.1 Einige grundlegende Begriffe

Der Kegel $F(M;x)$ der zulässigen Richtungen an eine Menge $M \subset \mathbb{R}^n$ in einem Punkt $x \in M$ ist zu Beginn von Kapitel 3 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 worden. 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

$$F(M;x) = \{p \in \mathbb{R}^n : a_i^T p \le 0 \ (i \in I(x)),\ a_i^T p = 0 \ (i = m_0 + 1, \ldots, m)\}.$$

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 Abstiegsrichtung, ist also $\nabla f(x)^T p \ge 0$ für alle $p \in F(M;x)$, so sind in x die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt, d. h. x ist eine *stationäre* oder auch *kritische* Lösung von (P), siehe Beweis von Satz 2.1 in Abschnitt 3.2. Auch die Umkehrung der obigen Aussage ist richtig: Ist $x \in M$ eine stationäre Lösung von (P), so gibt es in x keine zulässige Abstiegsrichtung[2]. 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

$$s(x,p) = \min\Big\{ \frac{b_i - a_i^T x}{a_i^T p} : i \in \{1, \ldots, m_0\} \setminus I(x),\ a_i^T p > 0 \Big\}.$$

[2]Denn ist $x \in M$ eine stationäre Lösung von (P), so existiert $y \in \mathbb{R}^m$ mit

$$y_i \ge 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x) + \sum_{i=1}^{m} y_i a_i = 0, \qquad y_i(b_i - a_i^T x) = 0 \quad (i = 1, \ldots, m_0).$$

Ist $p \in F(M;x)$, so ist daher

$$\nabla f(x)^T p = -\sum_{i \in I(x)} \underbrace{y_i a_i^T p}_{\le 0} \ge 0,$$

also p keine Abstiegsrichtung.

Ist $a_i^T p \leq 0$ für alle $i \in \{1, \ldots, m_0\} \setminus I(x)$, so ist natürlich $s(x,p) = +\infty$.

Ein Schritt eines Verfahrens der zulässigen Richtungen, angewandt auf das zu Beginn dieses Kapitels formulierte linear restringierte, nichtlineare Optimierungsaufgabe (P) sieht im Prinzip folgendermaßen aus:

- Sei $x \in M$ gegeben.
- Falls $F(M;x) \cap \{p \in \mathbb{R}^n : \nabla f(x)^T p < 0\} = \emptyset$, dann: STOP, x ist stationäre Lösung von (P).
- Andernfalls:
  - Wähle $p \in F(M;x)$ mit $\nabla f(x)^T p < 0$.
  - Berechne die maximale Schrittweite $s(x,p)$ und bestimme $t \in (0, s(x,p)]$ mit $f(x+tp) < f(x)$.
- Setze $x_+ := x + tp$.

Genau wie die Schrittweiten-Verfahren für unrestringierte Optimierungsaufgaben (siehe Abschnitt 2.1) ist ein Verfahren der zulässigen Richtungen durch die Spezifikation der Schrittweiten- und der Richtungsstrategie gegeben.

6.2.2 Schrittweitenstrategien

Es ist einfach, die aus der unrestringierten Optimierung her bekannten Schrittweitenstrategien zu übertragen. Fast wörtlich wie zu Beginn von Unterabschnitt 2.1.1 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

$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma \|x - y\| \qquad \text{für alle } x, y \in L_0.$$

Bemerkung: Sei $x \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine zulässige Abstiegsrichtung für f in x. Mit $s(x,p) > 0$ wird die maximale Schrittweite in x in Richtung p bezeichnet. Wie in der unrestringierten Optimierung besteht eine naheliegende Schrittweitenstrategie darin, als Schrittweite $t^*(x,p) > 0$ eine Lösung der eindimensionalen Minimierungsaufgabe

$$\text{Minimiere} \quad \phi(t) := f(x+tp), \quad t \in [0, s(x,p)],$$

zu wählen. Alternativ sei $t^*(x,p) > 0$ die erste positive Nullstelle von $\phi'(\cdot)$ auf dem Intervall $(0, s(x,p)]$, wenn eine solche existiert, andernfalls sei $t^* = s(x,p)$. In beiden Fällen spricht man von einer *exakten Schrittweite*. Unter der Voraussetzung (V) ist

die Existenz dieser Schrittweite jeweils gesichert. In beiden Fällen kann die Existenz einer Konstanten $\theta > 0$ mit

$$f(x) - f(x + t^*(x,p)p) \geq \theta \, \min\Big[-s(x,p)\,\nabla f(x)^T p, \Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2\Big]$$

für alle nichtstationären $x \in L_0$ und alle in x zulässigen Abstiegsrichtungen $p \in \mathbb{R}^n$. Im unrestringierten Fall ist $s(x,p) = +\infty$ und man erhält das schon bekannte Resultat. Einen Beweis dieser Aussage überlassen wir als Übungsaufgabe. □

Auch die Wolfe-Schrittweite der unrestringierten Optimierung ist einfach zu übertragen. Wieder sei $x \in L_0$ keine stationäre Lösung von (P), $p \in \mathbb{R}^n$ eine zulässige Abstiegsrichtung für f in x und $s(x,p) > 0$ die maximale Schrittweite in x in Richtung p. Es seien Konstanten $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ vorgegeben. Eine Wolfe-Schrittweite $t_W(x,p)$ in x in Richtung p ist folgendermaßen definiert: Man setze $t_W(x,p) := s(x,p)$, 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 $t_W(x,p) \in (0, s(x,p))$ beliebig mit

(a) $$f(x + t_W(x,p)p) \leq f(x) + \alpha t_W(x,p)\,\nabla f(x)^T p$$

und

(b) $$\nabla f(x + t_W(x,p)p)^T p \geq \beta\,\nabla f(x)^T p.$$

Der folgende Satz entspricht völlig Satz 1.1 in Abschnitt 2.1.

Satz 2.1 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x \in L_0$ keine stationäre Lösung von (P) und p eine zulässige Abstiegsrichtung für f in x. Seien $\alpha \in (0, \frac{1}{2})$, $\beta \in (\alpha, 1)$ gegeben $T_W(x,p)$ die oben definierte Menge der Wolfe-Schrittweiten in x in Richtung p. Dann gilt:*

1. *Es ist $T_W(x,p) \neq \emptyset$.*
2. *Es existiert eine Konstante $\theta > 0$, die nur von α, β und γ (der Lipschitzkonstanten von $\nabla f(\cdot)$ auf L_0) abhängt, nicht aber von x oder p, mit*

$$f(x) - f(x + tp) \geq \theta \, \min\Big[-s(x,p)\,\nabla f(x)^T p, \Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2\Big]$$

für alle $t \in T_W(x,p)$

Beweis: Natürlich können wir annehmen, dass $s(x,p) < \infty$, da die Aussage des Satzes andernfalls aus Satz 1.1 in Abschnitt 2.1 folgt. Wie dort setzen wir zur Abkürzung

$$\Phi(t) := f(x) + \alpha t \nabla f(x)^T p - f(x + tp)$$

und können annehmen, dass $\Phi(s(x,p)) < 0$ (andernfalls ist $t = s(x,p)$ die Wolfe-Schrittweite und die Abschätzung über die Verminderung der Zielfunktion gilt mit

$\theta = \alpha$). Es ist $\Phi(0) = 0$ und $\Phi'(0) = -(1-\alpha)\nabla f(x)^T p > 0$, also ist $\Phi(\cdot)$ für alle hinreichend kleinen $t > 0$ positiv. Da aber $\Phi(s(x,p)) < 0$, gibt es eine erste positive Nullstelle $\hat{t}$ von $\Phi(\cdot)$. Alle Punkte aus $[0, \hat{t}]$ genügen der ersten Bedingung (a) für eine Wolfe-Schrittweite. Wegen des Satzes von Rolle existiert ein $t \in (0, \hat{t})$ mit $\Phi'(t) = 0$. Wegen $\alpha < \beta$ und

$$0 = \Phi'(t) = \alpha \nabla f(x)^T p - \nabla f(x+tp)^T p \geq \beta \nabla f(x)^T p - \nabla f(x+tp)^T p$$

genügt t auch der zweiten Bedingung (b) für eine Wolfe-Schrittweite. Also ist $t \in T_W(x,p)$, der erste Teil des Satzes ist bewiesen.

Auch für den zweiten Teil des Satzes können wir $s(x,p) < \infty$ und $\Phi(s(x,p)) < 0$ annehmen. Sei $t \in T_W(x,p)$ gegeben. Dann ist $f(x+tp) \leq f(x)$ (wegen (a)) und $x + tp \in M$ (wegen $t < s(x,p)$), also $x + tp \in L_0$. Aus der Bedingung (b) einer Wolfe-Schrittweite und der Lipschitzstetigkeit des Gradienten auf der Niveaumenge L_0 folgt

$$-(1-\beta)\nabla f(x)^T p \leq [\nabla f(x+tp) - \nabla f(x)]^T p \leq \gamma t \, \|p\|^2$$

und daher

$$f(x) - f(x+tp) \geq -\alpha t \nabla f(x)^T p \geq \frac{\alpha(1-\beta)}{\gamma} \Big(\frac{\nabla f(x)^T p}{\|p\|} \Big)^2 .$$

Die Behauptung ist mit $\theta := \alpha \min(1, (1-\beta)/\gamma)$ bewiesen. □

Bemerkungen: Eine geringfügige Modifikation der Wolfe-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. Setzt man $\tilde{s}(x,p) := \min(1, s(x,p))$, so kann bei vorgegebenen Konstanten $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ die (modifizierte) Wolfe-Schrittweite $t_W(x,p) := \tilde{s}(x,p)$ gesetzt werden, falls

$$f(x + \tilde{s}(x,p)p) \leq f(x) + \alpha \, \tilde{s}(x,p) \, \nabla f(x)^T p,$$

andernfalls bestimme man $t_W(x,p) \in (0, \tilde{s}(x,p))$ mit

$$f(x + t_W(x,p)p) \leq f(x) + \alpha t_W(x,p) \, \nabla f(x)^T p, \qquad \nabla f(x + t_W(x,p)p)^T p \geq \beta \, \nabla f(x)^T p.$$

Natürlich existiert auch diese (modifizierte) Wolfe-Schrittweite, ferner gilt eine Satz 2.1 entsprechende Aussage.

Die Berechnung einer Wolfe-Schrittweite kann fast genau wie bei unrestringierten Optimierungsaufgaben erfolgen. Hierbei kann $s(x,p) < \infty$ angenommen werden (sonst ist kein Unterschied zum unrestringierten Fall), ferner kann angenommen werden, dass die Bedingung (a) für (wir gehen nur auf die modifizierte Wolfe-Schrittweite ein) $t = \tilde{s}(x,p)$ nicht gilt (andernfalls ist $t = \tilde{s}(x,p)$). Das Verfahren zur Bestimmung der Wolfe-Schrittweite, das wir in Unterabschnitt 2.1.1 angaben, bestand aus zwei Phasen. In der ersten wurde ein nichtleeres Intervall $[t_{\min}, t_{\max}]$ mit der Eigenschaft bestimmt, das in $t_{\min}$ (a) aber nicht (b) und in $t_{\max}$ (a) nicht erfüllt ist. Da (a) für $\tilde{s}(x,p)$ nicht erfüllt ist, können wir $t_{\max} := \tilde{s}(x,p)$ setzen. Für alle hinreichend kleinen $t > 0$ gilt (a), aber nicht (b). Durch sukzessive Halbierung kann man daher, ausgehend von $\tilde{s}(x,p)$, das

gesuchte $t_{\min}$ finden. Die zweite Phase ist völlig identisch mit der entsprechenden bei unrestringierten Optimierungsaufgaben. □

Auf die Armijo-Schrittweite, die in naheliegender Weise definiert werden kann, wollen wir nicht mehr eingehen.

6.2.3 Richtungsstrategien

Gegeben sei wieder die linear restringierte Optimierungsaufgabe (P). Für eine gegebene aktuelle Näherung $x \in M$ und $\epsilon \geq 0$ sei

$$I_\epsilon(x) := \{i \in \{1, \ldots, m_0\} : b_i - a_i^T x \leq \epsilon\}$$

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. Die zu Beginn von Unterabschnitt 6.2.2 angegebenen Voraussetzungen (V) (a)–(c) seien erfüllt.

Wir stellen uns das folgende Problem: Gegeben sei ein nichtstationäres $x \in L_0$, etwa eine aktuelle Näherung für eine (lokale oder stationäre) 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 2.2 *Sei $H \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, $\epsilon \geq 0$ und $x \in M$. Sei p die eindeutige Lösung des quadratischen Programms*

$$(\mathrm{P}_\epsilon(x)) \qquad \begin{cases} \text{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T H p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \leq b_i - a_i^T x \quad (i \in I_\epsilon(x)), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m). \end{cases}$$

Dann gilt: Ist $p \neq 0$, so ist p eine in x zulässige Abstiegsrichtung mit $0 < p^T H p \leq -\nabla f(x)^T p$, andernfalls ist x eine stationäre 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 Matrix H symmetrisch und positiv definit 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

$$y_i \geq 0 \quad (i \in I_\epsilon(x)), \qquad \nabla f(x) + Hp + \sum_{i \in I_\epsilon(x)} y_i a_i + \sum_{i=m_0+1}^{m} y_i a_i = 0$$

sowie

$$y_i(b_i - a_i^T x - a_i^T p) = 0 \quad (i \in I_\epsilon(x))$$

genügen. Ist $p = 0$ und definiert man $y_i := 0$ für alle $i \in \{1, \ldots, m_0\} \setminus I_\epsilon(x)$, so ist

$$y_i \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x) + A^T y = 0, \qquad y^T(b - Ax) = 0.$$

Das wiederum bedeutet, dass in x die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind bzw. x eine stationäre Lösung von (P) ist. Sei daher nun $p \neq 0$. Offensichtlich ist $p \in F(M;x)$. Ferner ist

$$\nabla f(x)^T p + p^T H p = -\sum_{i \in I_\epsilon(x)} y_i a_i^T p = -\sum_{i \in I_\epsilon(x)} \underbrace{y_i}_{\geq 0} \underbrace{(b_i - a_i^T x)}_{\geq 0} \leq 0,$$

so dass, wie behauptet, $0 < p^T H p \leq -\nabla f(x)^T p$. Insgesamt ist p eine zulässige Abstiegsrichtung. □

Bemerkung: Die Motivation für die in Lemma 2.2 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, H als eine Approximation an die Hessesche $\nabla^2 f(x)$ zu wählen. Für $H := \nabla^2 f(x)$ 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

$$a_i^T p \leq b_i - a_i^T x \quad (i = 1, \ldots, m_0), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m).$$

Für große ϵ sind dies genau die Restriktionen des quadratischen Programms ($\mathrm{P}_\epsilon(x)$) in Lemma 2.2. 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. □

6.2.4 Konvergenzaussagen

Nun stellt sich naheliegenderweise die Frage, ob die in Lemma 2.2 angegebene Richtungsstrategie, kombiniert z. B. mit der Wolfe-Schrittweitenstrategie ein konvergentes Verfahren ergibt. Die einfachste Aussage hierzu formulieren wir in dem folgenden Satz. Das dort angegebene Verfahren gehört zu den SQP-Verfahren, wobei SQP für **S**equential **Q**uadratic **P**rogramming steht, was bedeutet, dass in einem SQP-Verfahren eine Folge von quadratischen Programmen zu lösen ist.

Satz 2.3 *Gegeben sei die linear restringierte Optimierungsaufgabe* (P)*, die Voraussetzungen* (V) (a)–(c) *seien erfüllt. Sei* $\{H_k\} \subset \mathbb{R}^{n \times n}$ *eine Folge symmetrischer Matrizen, die gleichmäßig positiv definit und beschränkt ist, d. h. es existieren positive Konstanten* μ *und* η *mit*

$$\mu \|p\|^2 \leq p^T H_k p \leq \eta \|p\|^2 \qquad \text{für alle } p \in \mathbb{R}^n,\ k = 0, 1, \ldots.$$

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*

$$\begin{cases} \text{Minimiere} \quad \nabla f(x_k)^T p + \frac{1}{2} p^T H_k p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \le b_i - a_i^T x_k \quad (i \in I_\epsilon(x_k)), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m). \end{cases}$$

  - *Falls $p_k = 0$, dann: STOP, x_k ist eine stationäre Lösung von* (P).
  - *Berechne die Wolfe-Schrittweite $t_k := t_W(x_k, p_k)$.*
  - *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 stationäre Lösung von (P) *ist. Besitzt* (P) *genau eine stationäre Lösung x^* in der kompakten Niveaumenge L_0, so konvergiert die gesamte Folge $\{x_k\}$ gegen x^*.*

Beweis: Wegen Lemma 2.2 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 und eine Folge von Schrittweiten $\{t_k\} \subset \mathbb{R}_+$ erzeugt. Der Beweis dafür, dass jeder Häufungspunkt x^* von $\{x_k\}$ eine stationäre 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 $\|p_k\| \le c_0$, $k = 0, 1, \ldots$.

Denn: Wegen Lemma 2.2 und der gleichmäßigen positiven Definitheit der Folge $\{H_k\}$ symmetrischer Matrizen ist

$$0 < \mu \|p_k\|^2 \le p_k^T H_k p_k \le -\nabla f(x_k)^T p_k \le C \|p_k\|, \qquad k = 0, 1, \ldots,$$

mit einer Konstanten $C > 0$, die 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 möglich ist. Damit ist

$$\|p_k\| \le \frac{C}{\mu} =: c_0, \qquad k = 0, 1, \ldots,$$

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 $s(x_k, p_k) \ge \delta$, $k = 0, 1, \ldots$. 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 $i \in \{1, \ldots, m_0\} \setminus I_\epsilon(x_k)$ mit $a_i^T p_k > 0$ ist

$$\frac{b_i - a_i^T x_k}{a_i^T p_k} > \frac{\epsilon}{a_i^T p_k} \geq \frac{\epsilon}{\|a_i\|\,\|p_k\|} \geq c_1$$

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

$$\frac{b_i - a_i^T x_k}{a_i^T p_k} \geq 1.$$

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 $\|p_k\| \leq c_0$, wegen (b) existiert eine Konstante $\delta > 0$ mit $\tilde{s}(x_k, p_k) \geq \delta$ für $k = 0, 1, \ldots$. Aus Satz 2.1 erhält man die Existenz einer von k unabhängigen Konstanten $\theta > 0$ mit

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \theta \min\Big[-\tilde{s}(x_k, p_k)\, \nabla f(x_k)^T p_k, \Big(\frac{\nabla f(x_k)^T p_k}{\|p_k\|} \Big)^2 \Big] \\ &\geq \theta \min\Big[-\delta\, \nabla f(x_k)^T p_k, \frac{1}{c_0^2} \left(\nabla f(x_k)^T p_k \right)^2 \Big]. \end{aligned}$$

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

$$\mu\, \|p_k\|^2 \leq p_k^T H_k p_k \leq -\nabla f(x_k)^T p_k$$

mit $\mu > 0$, gilt auch $\lim_{k\to\infty} p_k = 0$.

(d) Jeder Häufungspunkt x^* von $\{x_k\}$ ist eine stationäre Lösung von (P).

Denn[3]: 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 $\nabla f(x^*)^T p^* \geq 0$ gilt. Damit wird gezeigt sein, dass es in x^* keine zulässige Abstiegsrichtung gibt, bzw. dass x^* eine stationäre Lösung von (P) ist.

[3]Der Beweis wäre einfach und kurz, wenn wir wüssten (oder voraussetzen würden), dass die zu der Lösung p_k des quadratischen Hilfsprogramms gehörende Folge von Lagrange-Multiplikatoren beschränkt ist.

Nach Konstruktion ist p_k die Lösung von

$$(\mathrm{P}_k) \qquad \begin{cases} \text{Minimiere} \quad \nabla f(x_k)^T p + \frac{1}{2}\, p^T H_k p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \le b_i - a_i^T x_k \quad (i \in I_\epsilon(x_k)), \qquad a_i^T p = 0 \quad (i = m_0+1, \ldots, m). \end{cases}$$

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

$$a_i^T (s_0 p^*) \le b_i - a_i^T x_k \quad (i \in I_\epsilon(x_k)), \qquad a_i^T (s_0 p^*) = 0 \quad (i = m_0+1, \ldots, m)$$

für alle hinreichend großen $k \in K$ gilt. Eine Einschränkung an $s_0 > 0$ ergibt sich offenbar nur für die $i \in I_\epsilon(x_k)$, für die $a_i^T p^* > 0$. Alle übrigen Restriktionen sind für alle $s_0 > 0$ erfüllt. Nach Definition der Indexmenge $I(x^*)$ der in x^* aktiven Ungleichungsrestriktionen existiert ein $\zeta > 0$ mit $b_i - a_i^T x^* \ge \zeta$ für alle $i \in \{1, \ldots, m_0\} \setminus I(x^*)$. Für alle hinreichend großen $k \in K$, etwa $k \ge k_0$, ist daher $b_i - a_i^T x_k \ge \frac{1}{2}\zeta$ für alle $i \in \{1, \ldots, m_0\} \setminus I(x^*)$. Nun wähle man $s_0 > 0$ so klein, dass $\frac{1}{2}\zeta \ge a_i^T (s_0 p^*)$ für alle $i \in \{1, \ldots, m_0\}$ mit $a_i^T p^* > 0$. Sei $k \in K$ und $k \ge k_0$, ferner $i \in I_\epsilon(x_k)$ und $a_i^T p^* > 0$. Dann ist $i \notin I(x^*)$. Nach Definition von ζ ist dann

$$b_i - a_i^T x_k \ge \tfrac{1}{2}\zeta \ge a_i^T (s_0 p^*)$$

sogar für alle $i \in \{1, \ldots, m_0\} \setminus I(x^*)$, erst recht also für alle $i \in I_\epsilon(x_k) \setminus I(x^*)$. Für alle hinreichend großen $k \in K$ ist damit $s_0 p^*$ zulässig für (P_k). Da $p = 0$ trivialerweise zulässig ist, ist aus Konvexitätsgründen $s p^*$ 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

$$\begin{aligned} \nabla f(x_k)^T p_k + \tfrac{1}{2}\,\mu \,\|p_k\|^2 &\le \nabla f(x_k)^T p_k + \tfrac{1}{2}\, p_k^T H_k p_k \\ &\le s\, \nabla f(x_k)^T p^* + \tfrac{1}{2}\, s^2\, (p^*)^T H_k p^* \\ &\le s\, \nabla f(x_k)^T p^* + \tfrac{1}{2}\, s^2 \eta\, \|p^*\|^2 \end{aligned}$$

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 $0 \le \nabla f(x^*)^T p^* + \frac{1}{2}\,\mu s\, \|p^*\|^2$ für alle $s \in (0, s_0]$. Mit $s \to 0+$ folgt $\nabla f(x^*)^T p^* \ge 0$, womit schließlich auch (d) bewiesen ist.

(e) Besitzt (P) genau eine stationäre Lösung x^* in der Niveaumenge L_0, so konvergiert die gesamte Folge $\{x_k\}$ gegen x^*.

Denn[4]: Angenommen, $\{x_k\}$ würde nicht gegen x^* konvergieren. Dann existiert eine unendliche Teilmenge $K \subset \mathbb{N}$ und ein $\delta > 0$ mit $\|x_k - x^*\| \ge \delta$ 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 stationäre Lösung von (P). Da aber $\|x^{**} - x^*\| \ge \delta$ ergibt sich ein Widerspruch zur Voraussetzung, dass (P) genau eine stationäre Lösung in L_0 besitzt. □

[4]Es folgt ein Routineargument, siehe auch den Schluss des Beweises von Satz 1.3 in Unterabschnitt 2.1.2.

Bemerkungen: Natürlich stellt sich die Frage, wie die Matrizen H_k im obigen Verfahren gewählt werden sollten. Setzt man $H_k := \nabla^2 f(x_k)$ (natürlich muss die Zielfunktion dann entsprechend glatt sein), so spricht man vom *Newton-Verfahren*. Die positive Definitheit ist natürlich nicht gesichert. Eine andere Möglichkeit besteht darin, am Anfang etwa $H_0 := I$ zu setzen und H_{k+1} aus H_k durch einen BFGS-Update zu gewinnen, also durch die Vorschrift

$$H_{k+1} := H_k - \frac{(H_k s_k)(H_k s_k)^T}{s_k^T H_k s_k} + \frac{y_k y_k^T}{y_k^T s_k},$$

wobei

$$s_k := x_{k+1} - x_k, \qquad y_k := \nabla f(x_{k+1}) - \nabla f(x_k).$$

Mit H_k ist auch H_{k+1} positiv definit, wenn $y_k^T s_k > 0$. Im Gegensatz zur unrestringierten Optimierung ist diese Bedingung bei Wahl der Wolfe-Schrittweite allerdings nicht automatisch gesichert.

Setzt man im obigen Verfahren der zulässigen Richtungen $\epsilon := +\infty$, so ist in jedem Schritt ein quadratisches Programm der Form

$$\begin{cases} \text{Minimiere} \quad \nabla f(x_k)^T p + \frac{1}{2} p^T H_k p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \le b_i - a_i^T x_k \quad (i = 1, \ldots, m_0), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m) \end{cases}$$

zu lösen. Wegen $x_k + p_k \in M$ ist die maximale Schrittweite $s(x_k, p_k)$ in jedem Schritt mindestens 1 und daher $\tilde{s}(x_k, p_k) = \min(1, s(x_k, p_k) = 1$. Die modifizierte Wolfe-Schrittweite wird in diesem Falle also einfacher, da die Schrittweite 1 akzeptiert wird, wenn nur $f(x_k + p_k) \le f(x_k) + \alpha \nabla f(x_k)^T p_k$. □

Beispiel: Wir wenden das BFGS-Verfahren mit der Wolfe-Schrittweite und $\epsilon := +\infty$ auf die folgende Aufgabe (siehe W. Hock, K. Schittkowski (1981, S. 47)) an:

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := ((x_1 - 3)^2 - 9) x_2^3 \quad \text{unter der Nebenbedingung} \\ \begin{pmatrix} -1/\sqrt{3} & 1 \\ 1 & \sqrt{3} \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \le \begin{pmatrix} 0 \\ 0 \\ 6 \\ 0 \\ 0 \end{pmatrix}. \end{cases}$$

In Abbildung 6.2 veranschaulichen wir uns die Aufgabe. Man erkennt, dass die Lösung $x^* = (3, \sqrt{3})^T$ ist. Wie Hock-Schittkowski nehmen wir den Startwert $x_0 := (1, 0.5)^T$. Am Anfang sei ferner $H_0 := I$. Die auftretenden quadratischen Hilfsprobleme lösen wir mit der Funktion `quadprog` aus der Optimization-Toolbox von MATLAB. Schon nach zwei Schritten[5] hat man die Lösung bestimmt:

k	x_k^T	
0	1.000000000000000	0.500000000000000
1	2.777803983041933	1.603765877365275
2	3.000000000000000	1.732050807568877

[5]Das gilt aber erstaunlicherweise auch, wenn man $H_k = I$, $k = 1, \ldots$, wählt.

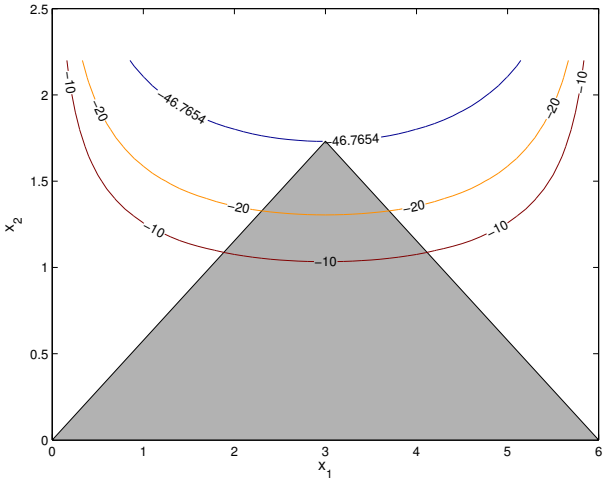


Abbildung 6.2: Noch eine linear restringierte nichtlineare Optimierungsaufgabe

Es wird jeweils die Schrittweite 1 genommen. Zu beachten ist, dass jeder Punkt $x = (x_1, 0)$ mit $0 \leq x_1 \leq 6$ wegen $\nabla f(x) = 0$ eine (zulässige) stationäre Lösung von (P) ist. Startet man also mit einem solchen Punkt, so erfolgt sofort ein Abbruch. Auch die Funktion `fmincon` aus der Optimization-Toolbox von MATLAB ist zufrieden, wenn die Optimalitätsbedingungen erster Ordnung erfüllt sind. □

Etwas genauer wollen wir jetzt noch auf das Newton-Verfahren, angewandt auf die linear restringierte Optimierungsaufgabe (P), eingehen und formulieren hierzu eine einfache lokale Konvergenzaussage.

Satz 2.4 *Gegeben sei die linear restringierte nichtlineare Optimierungsaufgabe (P). Sei $x^* \in M$ eine lokale Lösung von (P), die Zielfunktion f auf einer Umgebung U^* von x^* zweimal stetig differenzierbar, $\nabla^2 f(x^*)$ positiv definit und die Hessesche $\nabla^2 f$ auf U^* lipschitzstetig mit einer Lipschitzkonstanten $L > 0$. Mit einem $x_0 \in U^*$ sei die Folge $\{x_k\}$ durch $x_{k+1} := x_k + p_k$ definiert, wobei p_k die Lösung von*

$$(\mathrm{P}_k) \quad \begin{cases} \text{Minimiere} \quad \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \leq b_i - a_i^T x_k \quad (i = 1, \ldots, m_0), \qquad a_i^T p = 0 \quad (i = m_0 + 1, \ldots, m) \end{cases}$$

ist. Wir setzen voraus, die Folge $\{x_k\}$ sei in U^* enthalten und konvergiere gegen x^*. Dann konvergiert $\{x_k\}$ sogar quadratisch gegen x^*, d. h. es existiert eine Konstante $C > 0$ mit $\|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2$, $k = 0, 1, \ldots$.

Beweis: Wir können annehmen, die Umgebung U^* von x^* sei so klein, dass

$$\lambda_{\min}(\nabla^2 f(x)) \geq \mu > 0 \quad \text{für alle } x \in U^*,$$

dass also der kleinste Eigenwert der Hesseschen auf U^* gleichmäßig durch die positive Konstante μ gegen Null beschränkt ist. Dann ist

$$\mu \|x_{k+1} - x^*\|^2 \quad \leq \quad (x_{k+1} - x^*)^T \nabla^2 f(x_k)(x_{k+1} - x^*)$$

$$\begin{aligned} &= (p_k + x_k - x^*)^T \nabla^2 f(x_k)(x_{k+1} - x^*) \\ &= [\nabla^2 f(x_k) p_k]^T (x_{k+1} - x^*) - [\nabla^2 f(x_k)(x^* - x_k)]^T (x_{k+1} - x^*). \end{aligned}$$

Da p_k die Lösung des quadratischen Hilfsproblems (P_k) ist, existieren $y_i^{(k)}$, $i = 1, \ldots, m$, mit

$$y_i^{(k)} \geq 0 \quad (i = 1, \ldots, m_0), \qquad \nabla f(x_k) + \nabla^2 f(x_k) p_k + \sum_{i=1}^{m} y_i^{(k)} a_i = 0$$

sowie

$$y_i^{(k)} (b_i - a_i^T x_k - a_i^T p_k) = 0 \quad (i = 1, \ldots, m_0).$$

Daher ist

$$\begin{aligned} [\nabla^2 f(x_k) p_k]^T (x_{k+1} - x^*) &= -\nabla f(x_k)^T (x_{k+1} - x^*) - \sum_{i=1}^{m} y_i^{(k)} a_i^T (x_{k+1} - x^*) \\ &= -\nabla f(x_k)^T (x_{k+1} - x^*) - \sum_{i=1}^{m_0} y_i^{(k)} a_i^T (x_k + p_k - x^*) \\ &= -\nabla f(x_k)^T p_k - \sum_{i=1}^{m_0} \underbrace{y_i^{(k)} (b_i - a_i^T x^*)}_{\geq 0} \\ &\leq -\nabla f(x_k)^T (x_{k+1} - x^*). \end{aligned}$$

Folglich ist

$$\begin{aligned} \mu \, \|x_{k+1} - x^*\|^2 &\leq -\nabla f(x_k)^T (x_{k+1} - x^*) - [\nabla^2 f(x_k)(x^* - x_k)]^T (x_{k+1} - x^*) \\ &\leq [\nabla f(x^*) - \nabla f(x_k) - \nabla^2 f(x_k)(x^* - x_k)]^T (x_{k+1} - x^*) \\ &\qquad (\text{wegen } \nabla f(x^*)^T (x_{k+1} - x^*) \geq 0) \end{aligned}$$

und nach Anwendung der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned} \mu \, \|x_{k+1} - x^*\| &\leq \|\nabla f(x^*) - \nabla f(x_k) - \nabla^2 f(x_k)(x^* - x_k)\| \\ &= \left\| \int_0^1 [\nabla^2 f(x_k + t(x^* - x_k)) - \nabla^2 f(x_k)](x^* - x_k) \, dt \right\| \\ &\leq \int_0^1 \|\nabla^2 f(x_k + t(x^* - x_k)) - \nabla^2 f(x_k)\| \, dt \, \|x_k - x^*\| \\ &\leq \frac{L}{2} \|x_k - x^*\|^2, \end{aligned}$$

womit gezeigt ist, dass die Folge $\{x_k\}$ sogar quadratisch gegen x^* konvergiert. □

Beispiel: Wir wenden das ungedämpfte Newton-Verfahren auf die folgende Aufgabe (siehe W. Hock, K. Schittkowski (1981, S. 60)) an:

$$\left\{ \text{Minimiere} \quad f(x) := -x_1 x_2 x_3 \quad \text{unter der Nebenbedingung} \right.$$
$$\left. \begin{pmatrix} 1 & 2 & 2 \\ -1 & -2 & -2 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \leq \begin{pmatrix} 72 \\ 0 \\ 0 \\ 0 \\ 0 \\ 42 \\ 42 \\ 42 \end{pmatrix} . \right.$$

Als Startwert nehmen wir wie Hock-Schittkowski $x_0 := (10, 10, 10)^T$. Die quadratischen Hilfsprobleme lösen wir mit der Funktion `quadprog` der Optimization-Toolbox von MATLAB. Wir erhalten die folgenden Ergebnisse[6]:

k	x_k^T		
0	10.000000000000000	10.000000000000000	10.000000000000000
1	25.142857142857153	11.714285714285714	11.714285714285715
2	23.971428571428572	12.007142857142856	12.007142857142856
3	23.999983013419403	12.000004246645148	12.000004246645148
4	23.999999999993992	12.000000000001499	12.000000000001501
5	24.000000000000007	11.999999999999998	12.000000000000000
6	24.000000000000004	12.000000000000000	12.000000000000000

Wir haben auf dasselbe Beispiel auch noch das BFGS-Verfahren mit der Wolfe-Schrittweite angewandt. Nur in einem Schritt wird nicht die Schrittweite $t_k = 1$ genommen:

k	x_k^T			t_k
0	10.000000000000000	10.000000000000000	10.000000000000000	1.000000000000000
1	42.000000000000000	7.500000000000002	7.500000000000005	1.000000000000000
2	31.114571234396763	10.221357191400850	10.221357191400768	1.000000000000000
3	18.404200002946038	13.398949999260706	13.398949999266272	1.000000000000000
4	24.856511730698099	11.785872067348684	11.785872067302263	1.000000000000000
5	24.090878517660517	11.977280370102607	11.977280371067131	1.000000000000000
6	23.998345713088913	12.000413584276398	12.000413559179142	1.000000000000000
7	24.000003137898098	11.999998932296547	11.999999498754400	1.000000000000000
8	24.000002889667424	11.999999532145763	11.999999023020523	0.079266344750289
9	23.999997434807749	12.000000419293178	12.000000863302947	1.000000000000000
10	24.000000000000153	11.999999999999929	11.999999999999995	1.000000000000000

Das ist nun erstaunlich, dass genau einmal eine ziemlich kleine Schrittweite genommen wird, sonst immer die Schrittweite 1. Wählt man durchgehend die Schrittweite 1, so erhält man wesentlich schlechtere Ergebnisse. □

[6]Bei einem ganz ähnlichen Problem (siehe W. Hock, K. Schittkowski (1981, S. 59) ist das ungedämpfte Newton-Verfahren ausgehend von $x_0 := (10, 10, 10)^T$ sogar in einem Schritt bei der Lösung $x^* = (20, 11, 15)^T$.

6.2.5 Aufgaben

1. Gegeben sei eine linear restringierte nichtlineare Optimierungsaufgabe mit einer stetig differenzierbaren Zielfunktion. Man zeige, dass eine zulässige Lösung genau dann eine stationäre Lösung ist, wenn es in ihr keine zulässige Abstiegsrichtung gibt.

2. Gegeben sei die linear restringierte nichtlineare Optimierungsaufgabe (siehe W. Hock, K. Schittkowski (1981, S. 78))

$$
\text{(P)} \quad \left\{ \begin{array}{l} \text{Minimiere} \quad f(x) := x_1 + 2x_2 + 4x_5 + \exp(x_1 x_4) \quad \text{unter den NBen} \\ \begin{pmatrix} 1 & 2 & 0 & 0 & 5 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \\ 2 \\ 1 \\ 2 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \leq \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}, \\ \begin{array}{rcl} x_1 & \leq & 1, \\ x_4 & \leq & 1. \end{array} \end{array} \right.
$$

   Man löse diese Aufgabe mit der Matlab-Funktion `fmincon`, wobei man wie bei Hock-Schittkowski den Startwert $x_0 = (1, 2, 0, 0, 0, 2)^T$ nehme. Anschließend wiederhole man das Experiment mit dem Startwert $x_0 = (0, 2, 0, 1, 0, 2)^T$. Welche der erhaltenen Lösungen ist besser?

3. Man zeige: Genügt die Zielfunktion f von (P) den Voraussetzungen (V) (a)–(c) in Unterabschnitt 6.2.2, so existiert eine Konstante $\theta_C > 0$ derart, dass

$$
\begin{aligned} f(x) - f(x + t_M(x,p)p) &\geq f(x) - f(x + t_C(x,p)p) \\ &\geq \theta_C \, \min\Big[-s(x,p)\, \nabla f(x)^T p, \Big(\frac{\nabla f(x)^T p}{\|p\|} \Big)^2 \Big] \end{aligned}
$$

   für alle nicht stationären $x \in L_0$ 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. Gegeben sei das linear restringierte Programm

$$
\text{(P)} \quad \text{Minimiere } f(x) \text{ auf } M := \Big\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \Big\}.
$$

   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

$$\left\{ \begin{array}{c} \text{Minimiere} \quad \nabla f(x_k)^T p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \geq b_i - a_i^T x_k \;\; (i = 1, \ldots, m_0), \quad a_i^T p = 0 \;\; (i = m_0 + 1, \ldots, m). \end{array} \right.$$

– Falls $\nabla f(x_k)^T p_k = 0$, dann: STOP, x_k ist stationäre Lösung von (P).
– Berechne $t_k := t_M(x_k, p_k)$, $t_C(x_k, p_k)$ oder $t_k \in t_W(x_k, p_k)$.
– Setze $x_{k+1} := x_k + t_k p_k$.

Dann gilt: Bricht das Verfahren nicht vorzeitig mit einer stationären Lösung von (P) ab, so liefert es eine Folge $\{x_k\}$ mit der Eigenschaft, dass jeder Häufungspunkt von $\{x_k\}$ eine stationäre Lösung von (P) ist.

5. Gegeben sei das linear restringierte Programm

$$\text{(P)} \quad \text{Minimiere } f(x) \text{ auf } M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$

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

$$\text{(P}(x)\text{)} \quad \left\{ \begin{array}{l} \text{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T B p \quad \text{unter den Nebenbedingungen} \\ \begin{array}{rcll} a_i^T p & \geq & b_i - a_i^T x & (i = 1, \ldots, m_0), \\ a_i^T p & = & 0 & (i = m_0 + 1, \ldots, m), \end{array} \qquad \|p\|_\infty \leq 1. \end{array} \right.$$

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.

Kapitel 7

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 $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ sowie die Restriktionsabbildungen $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ 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 denselben 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.

7.1 Penalty- und Barriere-Verfahren

7.1.1 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 Idee bei durch nichtlineare Gleichungen restringierte Optimierungsaufgaben genauer untersuchen (siehe R. Fletcher (1987, S. 277 ff.) oder auch C. Geiger, C. Kanzow (2002, S. 206 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

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \frac{\sigma}{2}\,\|h(x)\|^2, \quad x \in \mathbb{R}^n,$$

zugeordnet, wobei $\|\cdot\|$ die euklidische Norm auf dem $\mathbb{R}^m$ bedeutet. Die bei differenzierbaren $f(\cdot)$ und $h(\cdot)$ differenzierbare Zielfunktion $\Phi_\sigma(\cdot)$ von (P_σ) heißt (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 sieht folgendermaßen aus:

- Wähle $\sigma_0 > 0$.
- Für $k = 0, 1, \ldots$:
  - Bestimme eine globale Lösung x_k von (P_{σ_k}).
  - Ist $h(x_k) = 0$, STOP: x_k ist globale Lösung von (P).
  - Wähle $\sigma_{k+1} > \sigma_k$, z. B. $\sigma_{k+1} := 10\sigma_k$.

Beispiel: Wir betrachten die Aufgabe[1]

$$(\mathrm{P}) \quad \text{Minimiere} \quad f(x) := -x_1 - x_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^2 : h(x) := 1 - x_1^2 - x_2^2 = 0\}.$$

Die Lösung x^* und den zugehörigen Lagrange-Multiplikator y^* erhält man leicht aus den notwendigen Bedingungen erster Ordnung. Ist x^* eine lokale Lösung, so ist $\nabla h(x^*) \neq 0$, die Constraint Qualification in Satz 2.3 in Unterabschnitt 3.2 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 $x^* = (1/\sqrt{2}, 1/\sqrt{2})^T$, der zugehörige Multiplikator ist $y^* = -1/\sqrt{2}$. Es ist

$$\Phi_\sigma(x) := -x_1 - x_2 + \frac{\sigma}{2}(1 - x_1^2 - x_2^2)^2.$$

Mit Hilfe von

$$\nabla \Phi_\sigma(x) = \begin{pmatrix} -1 \\ -1 \end{pmatrix} - 2\sigma \begin{pmatrix} (1 - x_1^2 - x_2^2)x_1 \\ (1 - x_1^2 - x_2^2)x_2 \end{pmatrix}$$

erhält man aus $\nabla \Phi_\sigma(x_\sigma) = 0$, dass $(x_\sigma)_1 = (x_\sigma)_2$ als Lösung von $(1 - 2x^2)x = -1/(2\sigma)$ zu bestimmen ist, was bei gegebenem $\sigma > 0$ zumindestens numerisch leicht möglich ist. Bei R. Fletcher (1987, S. 280) findet man einige numerische Ergebnisse. Wir erhalten die folgenden Ergebnisse, wobei wir in der vorletzten Spalte die Näherung $y_\sigma := \sigma h(x_\sigma)$ für den Lagrange-Multiplikator y^* (Vorgriff: Siehe Satz 1.2) und in einer

[1] Siehe R. Fletcher (1987, S. 279 ff.) und C. Geiger, C. Kanzow (2002, S. 212).

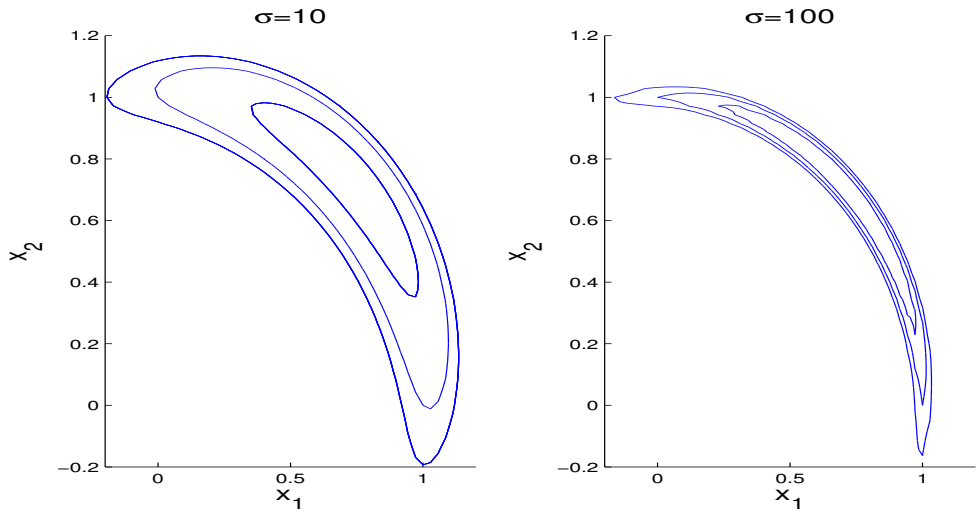


Abbildung 7.1: Höhenlinien von Φ_σ für $\sigma = 10$ und $\sigma = 100$

letzten Spalte die Kondition (bezüglich der euklidischen Norm bzw. Spektralnorm) der Hesseschen von Φ_σ in x_σ eintragen.

σ	$(x_\sigma)_1 = (x_\sigma)_2$	$y_\sigma = \sigma h(x_\sigma)$	$\kappa(\nabla^2 \Phi_\sigma(x_\sigma))$
1	0.884646177119316	−0.565197717383639	6.53858470847726
10	0.730893103186221	−0.684094565703688	32.2357241274489
100	0.709593646502773	−0.704628631420601	286.837458601107
1 000	0.707356648728881	−0.706857001907757	2832.42659491601
10 000	0.707131779860847	−0.707081783395402	28288.2711944907
100 000	0.707109281173289	−0.707104281216786	282846.712463874

Zum Vergleich ist $x_1^* = x_2^* = 1/\sqrt{2} = 0.707106781186547$, $y^* = -1/\sqrt{2}$. Die Kondition der Hesseschen der Zielfunktion Φ_σ in der Lösung x_σ von (P_σ) wächst mit σ stark an, was ein Indiz dafür ist, dass die Berechnung von x_σ zunehmend schwieriger wird, weil man in einem lang gestreckten Tal den niedrigsten Punkt sucht. In Abbildung 7.1 geben wir für $\sigma = 10$ und $\sigma = 100$ einige Höhenlinien an. □

Im folgenden Satz nehmen wir an (ohne es genau vorauszusetzen, siehe auch Theorem 12.1.1 bei R. Fletcher (1987, S. 281) oder Satz 5.6 bei C. Geiger, C. Kanzow (2002, S. 208)), die Aufgabe (P_σ) besitze für jedes $\sigma > 0$ eine globale Lösung x_σ, ferner sei (P) zulässig.

Satz 1.1 *Die Funktionen $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ seien stetig und $\{\sigma_k\} \subset \mathbb{R}_+$ streng monoton wachsend mit $\sigma_k \to +\infty$. Dann gilt:*

1. *$\{\Phi_{\sigma_k}(x_k)\}$ ist monoton wachsend.*
2. *$\{\|h(x_k)\|\}$ ist monoton fallend.*
3. *$\{f(x_k)\}$ ist monoton wachsend.*

4. *Es ist* $\lim_{k\to\infty} h(x_k) = 0$.

5. *Jeder Häufungspunkt von* $\{x_k\}$ *ist eine globale Lösung von (P).*

6. *Ist* x_σ *eine Lösung von* (P_σ) *mit* $h(x_\sigma) = 0$, *so ist* x_σ *eine globale Lösung von (P).*

Beweis: Da $\sigma_k < \sigma_{k+1}$ und x_k eine globale Lösung von (P_{σ_k}) ist, gilt

$$\Phi_{\sigma_k}(x_k) \le \Phi_{\sigma_k}(x_{k+1}) \le \Phi_{\sigma_{k+1}}(x_{k+1}).$$

Die beiden Ungleichungen $\Phi_{\sigma_k}(x_k) \le \Phi_{\sigma_k}(x_{k+1})$ und $\Phi_{\sigma_{k+1}}(x_{k+1}) \le \Phi_{\sigma_{k+1}}(x_k)$ addiere man und subtrahiere anschließend den auf beiden Seiten auftretenden Term $f(x_k) + f(x_{k+1})$. Man erhält (multipliziere die Ungleichung noch mit 2)

$$\sigma_k \|h(x_k)\|^2 + \sigma_{k+1} \|h(x_{k+1})\|^2 \le \sigma_k \|h(x_{k+1})\|^2 + \sigma_{k+1} \|h(x_k)\|^2.$$

Dies impliziert

$$\underbrace{(\sigma_k - \sigma_{k+1})}_{<0} (\|h(x_k)\|^2 - \|h(x_{k+1}\|^2) \le 0.$$

Hieraus folgt $\|h(x_k)\| \ge \|h(x_{k+1})\|$, womit die zweite Aussage bewiesen ist.

Unter Benutzung der gerade eben bewiesenen Aussage erhalten wir

$$0 \le \Phi_{\sigma_k}(x_{k+1}) - \Phi_{\sigma_k}(x_k) = f(x_{k+1} - f(x_k) + \frac{\sigma_k}{2}(\underbrace{\|h(x_{k+1})\|^2 - \|h(x_k)\|^2}_{\le 0})$$

und hieraus $f(x_k) \le f(x_{k+1})$.

Da wir vorausgesetzt haben, dass (P) zulässig bzw. $M \neq \emptyset$ ist, ist

$$\Phi_{\sigma_k}(x_k) \le \inf_{x\in M} \Phi_{\sigma_k}(x) \le \inf_{x\in M} f(x) = \inf(\mathrm{P}) < +\infty.$$

Als monoton fallende, nach unten beschränkte Folge ist $\{\|h(x_k)\|\}$ konvergent. Angenommen, es sei $0 < c := \lim_{k\to\infty} \|h(x_k)\|$. Dann ist

$$+\infty > \inf(\mathrm{P}) \ge \Phi_{\sigma_k}(x_k) = f(x_k) + \frac{\sigma_k}{2} \|h(x_k)\|^2 \ge f(x_0) + \frac{\sigma_k}{2} c^2,$$

was mit $\sigma_k \to \infty$ einen Widerspruch ergibt.

Sei x^* ein Häufungspunkt von $\{x_k\}$, also Limes einer Teilfolge $\{x_k\}_{k\in K}$. Wegen $\lim_{k\to\infty} h(x_k) = 0$ ist $h(x^*) = 0$ bzw. $x^* \in M$. Dann ist

$$\inf(\mathrm{P}) \le f(x^*) = \lim_{k\in K, k\to\infty} f(x_k) \le \lim_{k\in K, k\to\infty} \Phi_{\sigma_k}(x_k) \le \inf(\mathrm{P})$$

bzw. $x^* \in M$ eine globale Lösung von (P).

Ist x_σ eine Lösung von (P_σ) mit $h(x_\sigma) = 0$, so ist

$$\inf(\mathrm{P}) \le f(x_\sigma) \le \Phi_\sigma(x_\sigma) \le \inf_{x\in M} \Phi_\sigma(x) = \inf_{x\in M} f(x) = \inf(\mathrm{P}),$$

also $x_\sigma \in M$ eine Lösung von (P). Damit ist der Satz bewiesen. □

Bemerkung: Hat man eine nichtlinear restringierte Optimierungsaufgabe vorliegen, bei der auch Ungleichungen als Restriktionen auftreten, also

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

wobei $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$, so ist diese Aufgabe identisch mit

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

wobei

$$\max(g(x), 0) = (\max(g_i(x), 0), \ldots, \max(g_l(x), 0))^T.$$

Eine zu (P) gehörende quadratische Straffunktion ist also

$$\phi_\sigma(x) := f(x) + \frac{\sigma}{2}(\| \max(g(x), 0)\|^2 + \|h(x)\|^2).$$

Die Aussagen von Satz 1.1 gelten sinngemäß. □

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_σ von

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := x^2 + \frac{\sigma}{2} \max(1 - x, 0)^2, \quad x \in \mathbb{R}$$

bestimmen. Es ist

$$\Phi'_\sigma(x) = \begin{cases} 2x + \sigma(x - 1), & x < 1, \\ 2x, & x \geq 1. \end{cases}$$

Daher ist

$$2x_\sigma + \sigma(x_\sigma - 1) = 0,$$

bzw.

$$x_\sigma = \frac{\sigma}{2 + \sigma}.$$

Auch hier ist $\lim_{\sigma \to \infty} x_\sigma = x^*$. In Abbildung 7.2 links haben wir die Straffunktion Φ_σ für $\sigma = 1$ und $\sigma = 10$ eingetragen, rechts findet man x_σ für $\sigma \in [0, 100]$. □

Jetzt sei (P) wieder die Optimierungsaufgabe mit einer Gleichungsrestriktion $h(x) = 0$. Ist x_k eine Lösung der unrestringierten Optimierungsaufgabe (P_{σ_k}) und sind f und h sogar stetig differenzierbar, so ist

$$\nabla \Phi_{\sigma_k}(x_k) = \nabla f(x_k) + \sigma_k h'(x_k)^T h'(x_k) = \nabla f(x_k) + h'(x_k)^T y_k$$

mit

$$y_k := \sigma_k h(x_k).$$

Wir wissen schon, dass jeder Häufungspunkt x^* von $\{x_k\}$ eine Lösung von (P) ist. Unter der Constraint Qualification Rang $(h'(x^*)) = m$ gibt es zu x^* einen Lagrange-Multiplikator y^* mit

$$\nabla f(x^*) + h'(x^*)^T y^* = 0,$$

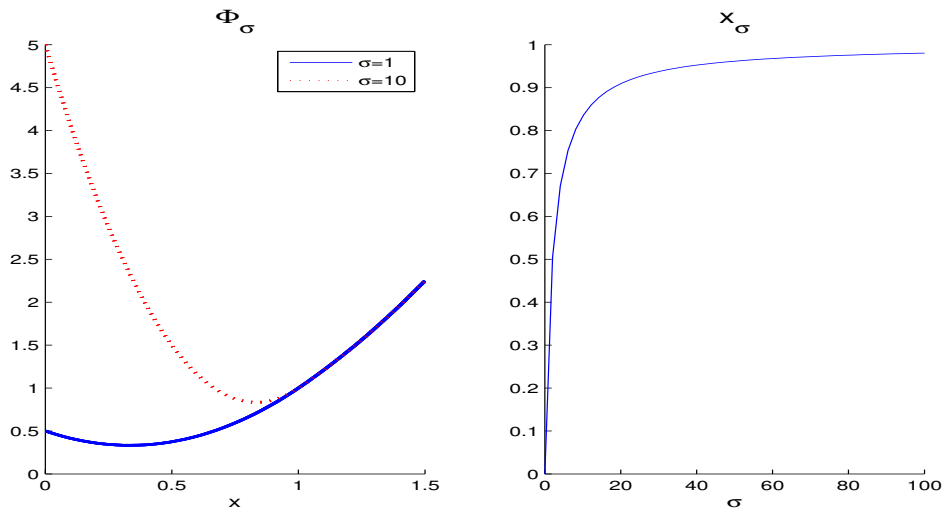


Abbildung 7.2: Φ_σ und x_σ

siehe Satz 2.3 in Abschnitt 3.2. Man vermutet daher, dass die Folge $\{y_k\}$ den (wegen der Rangvoraussetzung eindeutig bestimmten) Lagrange-Multiplikator y^* approximiert. Dies ist richtig, wie der folgende Satz (siehe R. Fletcher (1987, S. 282) und C. Geiger, C. Kanzow (2002, S. 210)) zeigt.

Satz 1.2 *Seien $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ und $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ stetig differenzierbar. Sei $\{x_k\}$ durch das Penalty-Verfahren gewonnen und $\lim_{k\to\infty} x_k = x^*$. Sei* Rang $(h'(x^*)) = m$ *(bzw. die Gradienten $\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ linear unabhängig). Dann gilt:*

1. *Die Folge $\{y_k\}$ mit $y_k := \sigma_k h(x_k)$ konvergiert gegen ein y^*.*

2. *(x^*, y^*) ist ein Kuhn-Tucker-Paar zu (P), d. h. es ist $\nabla f(x^*) + h'(x^*)^T y^* = 0$, wobei y^* der eindeutig bestimmte Lagrange-Multiplikator zu der Lösung x^* ist.*

Beweis: Da[2] Rang $(h'(x^*)) = m$ und $\lim_{k\to\infty} x_k = x^*$, ist auch Rang $(h'(x_k)) = m$ für alle hinreichend großen k. Wegen $\nabla f(x_k) + h'(x_k)^T y_k = 0$ folgt

$$y_k = -(h'(x_k)h'(x_k)^T)^{-1}h'(x_k)\nabla f(x_k) \to -(h'(x^*)h'(x^*)^T)^{-1}h'(x^*)\nabla f(x^*) =: y^*.$$

Damit ist der erste Teil des Satzes bewiesen.

Da x^* als Lösung von (P) insbesondere eine stationäre Lösung ist, existiert ein $v^* \in \mathbb{R}^m$ mit $\nabla f(x^*) + h'(x^*)^T v^* = 0$. Da $h'(x^*)h'(x^*)^T$ nichtsingulär ist, folgt hieraus

$$v^* = -(h'(x^*)h'(x^*)^T)^{-1}h'(x^*)\nabla f(x^*) = y^*.$$

Damit ist auch der zweite Teil des Satzes bewiesen. □

[2]Beachte: Ist $A \in \mathbb{R}^{m\times n}$, so ist Rang $(A) = m$ genau dann, wenn $AA^T \in \mathbb{R}^{m\times m}$ nichtsingulär ist.

7.1.2 Barriere-Funktionen

Die im letzten Unterabschnitt angegebenen Penalty-Verfahren heißen *äußere Penalty-Verfahren*, da die Iterierten x_k außerhalb der zulässigen Menge liegen. Dagegen spricht man von einem *inneren Penalty-Verfahren*, wenn die Iterierten im Inneren der zulässigen Menge liegen und durch eine Barriere-Funktion verhindert wird, dass die Iterierten sich zu schnell dem Rand der zulässigen Menge nähern. Wir begnügen uns damit, sehr kurz und ohne Beweise[3] auf die wichtigste Barriere-Funktion, nämlich die *logarithmische Barriere-Funktion* und das zugehörige Barriere-Verfahren einzugehen. Bei linearen Optimierungsaufgaben sind wir schon auf logarithmische Barriere-Funktionen eingegangen (siehe Unterabschnitt 4.1.2), hier betrachten wir allgemeiner konvexe Optimierungsaufgaben.

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 setzen wir voraus:

(V1) Die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ und die Restriktionsabbildung $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ sind stetig differenzierbar und (komponentenweise) konvex, die Abbildung $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ist affin linear, also (P) eine konvexe Optimierungsaufgabe. Ferner habe die (konstante) Funktionalmatrix $h' \in \mathbb{R}^{m\times n}$ den Rang m.

(V2) Das relative Innere

$$M_0 := \{x \in \mathbb{R}^m : g(x) < 0,\ h(x) = 0\}$$

von M ist nichtleer.

(V3) Die Menge M_{opt} der Lösungen von (P) ist nichtleer und kompakt.

Beispiel: Von J. J. Sylvester[4] (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

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x,\delta) := \delta \quad \text{auf} \\ M := \{(x,\delta) \in \mathbb{R}^n \times \mathbb{R} : g_i(x,\delta) := \frac{1}{2}\|x - a_i\|^2 - \delta \leq 0,\ i = 1, \ldots, l\}. \end{array} \right.$$

Hierbei bedeute $\|\cdot\|$ natürlich die euklidische Norm auf dem $\mathbb{R}^n$. Bei (P) handelt es sich um ein konvexes Problem, ferner ist

$$M_0 := \{(x,\delta) \in \mathbb{R}^n \times \mathbb{R} : g_i(x,\delta) := \tfrac{1}{2}\|x - a_i\|^2 - \delta < 0,\ i = 1, \ldots, l\} \neq \emptyset,$$

[3] Als Literatur nennen wir A. V. Fiacco, G. P. McCormick (1968) und M. H. Wright (1992).

[4] James Joseph Sylvester (1814–1897) war ein englischer Mathematiker. Er forschte zusammen mit Arthur Cayley auf dem Gebiet der Invariantentheorie. Ein weiteres Arbeitsgebiet war die Theorie von Matrizen und Determinanten. Die Bezeichnung "Matrix" wurde 1850 von Sylvester eingeführt, ebenso ist der Trägheitssatz von Sylvester nach ihm benannt. Dieser besagt bekanntlich: Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch und $X \in \mathbb{R}^{n\times n}$ nichtsingulär, so haben A und $X^T AX$ dieselbe Anzahl positiver, negativer und verschwindender Eigenwerte.

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 M_{opt} einpunktig und folglich 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

$$\sqrt{2\delta^*} = \max_{i=1,\ldots,l} \|\tfrac{1}{2}(x_1^* - a_i) + \tfrac{1}{2}(x_2^* - a_i)\| = \|\tfrac{1}{2}(x_1^* - a_j) + \tfrac{1}{2}(x_2^* - a_j)\|$$

mit einem $j \in \{1, \ldots, l\}$. Dann ist aber

$$\begin{aligned} \sqrt{2\delta^*} &= \|\tfrac{1}{2}(x_1^* - a_j) + \tfrac{1}{2}(x_2^* - a_j)\| \\ &\leq \tfrac{1}{2}\|x_1^* - a_j\| + \tfrac{1}{2}\|x_2^* - a_j\| \\ &\leq \tfrac{1}{2}\max_{i=1,\ldots,l} \|x_1^* - a_i\| + \tfrac{1}{2}\max_{i=1,\ldots,l} \|x_2^* - a_i\| \\ &= \sqrt{2\delta^*}. \end{aligned}$$

Da die euklidische Norm strikt konvex ist, folgt hieraus $x_1^* = x_2^*$, insgesamt also die eindeutige Lösbarkeit von (P). Damit erfüllt (P) die Voraussetzungen (V1)–(V3). In Abbildung 7.3 geben wir zu 5 Punkten in der Ebene den kleinsten Kreis (d. h. den Kreis mit dem kleinsten Radius) an, der diese Punkte enthält. Drei der Punkte liegen

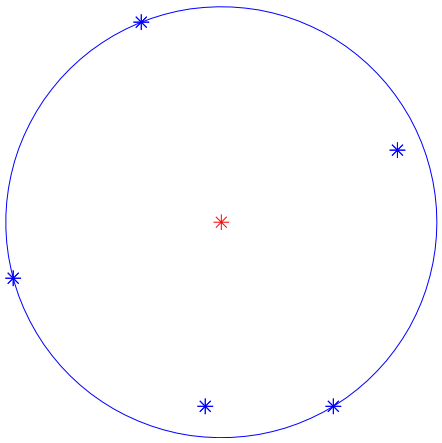


Abbildung 7.3: Kleinster Kreis zu 5 Punkten

auf dem Rand des Kreises. Ist das immer so? □

Der Aufgabe (P) wird mittels eines Parameters $\sigma > 0$ eine Schar von Optimierungsaufgaben

$$(\mathrm{P}_\sigma) \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad \Psi_\sigma(x) := f(x) - \dfrac{1}{\sigma}\displaystyle\sum_{i=1}^{l} \log(-g_i(x)) \quad \text{auf} \\ M_0 := \{x \in \mathbb{R}^n : g(x) < 0,\ h(x) = 0\} \end{array} \right.$$

zugeordnet. Man nennt Ψ_σ eine *logarithmische Barriere-Funktion*: Ist $\sigma > 0$ fest und $\{x_k\} \subset M_0$ eine Folge, die gegen einen Punkt des Randes von M konvergiert, so ist $\Psi_\sigma(x_k) \to +\infty$. Mit wachsendem σ wird diese Barriere allerdings immer schwächer.

Beispiel: Wir kommen auf die schon früher angegebene Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := x^2 \quad \text{auf} \quad M := \{x \in \mathbb{R} : g(x) := 1 - x \leq 0\}$$

zurück. Die Aufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := x^2 - \frac{1}{\sigma}\log(x-1), \qquad x > 1,$$

besitzt die eindeutige Lösung

$$x_\sigma := \frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{2}{\sigma}}$$

und es ist offenbar $\lim_{\sigma\to\infty} x_\sigma = x^*$, wobei $x^* = 1$ die Lösung von (P) ist. □

Wir begnügen uns mit einem einzigen Satz zu den logarithmischen Barriere-Funktionen und dem zugehörigen Barriere-Verfahren. Für den Beweis verweisen wir auf das angegebene Optimierungsskript von J. Werner.

Satz 1.3 *Gegeben sei die konvexe Optimierungsaufgabe (P), die Voraussetzungen (V1), (V2) und (V3) seien erfüllt. Dann gilt:*

1. *Für jedes $\sigma > 0$ ist die Menge der Lösungen von (P_σ) nichtleer und kompakt.*
2. *Ist $\{\sigma_k\} \subset \mathbb{R}_+$ eine Folge mit $\sigma_k \to \infty$ und $x_k \in M_0$ eine Lösung von (P_{σ_k}), so ist die Folge $\{x_k\}$ beschränkt, ferner ist jeder Häufungspunkt von $\{x_k\}$ eine Lösung[5] von (P). Schließlich gilt $\lim_{k\to\infty} \min(\mathrm{P}_{\sigma_k}) = \min(\mathrm{P})$.*
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 $Q_0, Q_1, \ldots, Q_l \in \mathbb{R}^{n\times n}$. Dann besitzt die Aufgabe (P_σ) für jedes $\sigma > 0$ genau eine Lösung $x_\sigma \in M_0$ und $x^* = \lim_{\sigma\to\infty} x_\sigma$ existiert und gehört zu M_{opt}.

Bemerkung: Man könnte auf die Idee kommen, eine gegebene konvexe Optimierungsaufgabe dadurch numerisch zu lösen, indem man für wachsende $\sigma_1 < \sigma_2 < \ldots$ (im wesentlichen) unrestringierte Optimierungsaufgaben (P_{σ_k}) etwa mit dem Newton-Verfahren löst, wobei man die (gefundene) Lösung x_k des Problems (P_{σ_k}) als Startwert zur Lösung des Problems $(\mathrm{P}_{\sigma_{k+1}})$ nimmt. Dieses sogenannte *Barriere-Verfahren* ist aus mindestens zwei Gründen i. Allg. nicht empfehlenswert. Zum einen werden die Probleme wie bei der quadratischen Straffunktion mit wachsendem Parameter σ immer schlechter konditioniert (die Kondition der Hesseschen wird immer größer), zum anderen ist die

[5] Besitzt (P) insbesondere eine eindeutige Lösung x^*, so konvergiert die ganze Folge $\{x_k\}$ gegen x^*.

Konvergenz schlecht. Trotzdem wollen wir schildern, wie man das Newton-Verfahren zur Lösung von (P_σ) einsetzen kann, wobei wir bequemlichkeitshalber davon ausgehen, dass keine (affin linearen) Gleichungsrestriktionen auftreten. Das zu lösende Problem ist also

$$(\mathrm{P}_\sigma) \qquad \begin{cases} \text{Minimiere} \quad \Psi_\sigma(x) := f(x) - \dfrac{1}{\sigma}\displaystyle\sum_{i=1}^{l} \log(-g_i(x)) \quad \text{auf} \\ \qquad\qquad M_0 := \{x \in \mathbb{R}^n : g(x) < 0\}, \end{cases}$$

wobei f und g (komponentenweise) als konvex und zweimal stetig differenzierbar vorausgesetzt werden. Der Gradient und die Hessesche von Ψ_σ sind gegeben durch

$$\begin{aligned} \nabla\Psi_\sigma(x) &= \nabla f(x) - \frac{1}{\sigma}\sum_{i=1}^{l} \frac{1}{g_i(x)}\nabla g_i(x), \\ \nabla^2\Psi_\sigma(x) &= \nabla^2 f(x) + \frac{1}{\sigma}\sum_{i=1}^{l}\Big[\frac{1}{g_i(x)^2}\nabla g_i(x)\nabla g_i(x)^T - \frac{1}{g_i(x)}\nabla^2 g_i(x)\Big]. \end{aligned}$$

Wir schildern jetzt einen Schritt des Newton-Verfahrens zur Lösung von (P_σ).

- Gegeben seien (unabhängig vom Iterationsschritt) $\alpha \in (0, \frac{1}{2})$ und $\rho \in (0,1)$ für die Armijo-Schrittweite (z. B. $\alpha = 0.0001$ und $\rho = 0.5$) sowie $\tau \in (0,1)$ (z. B. $\tau = 0.99$).
- Gegeben $x \in M_0$.
- Berechne die Newton-Richtung $p := -\nabla^2\Psi_\sigma(x)^{-1}\nabla\Psi_\sigma(x)$.
- Falls $p = 0$, dann: STOP, da x Lösung von (P_σ).
- Berechne maximale Schrittweite $s(x,p) := \sup\{t > 0 : g(x+tp) \le 0\}$ und setze $t := \min(1, \tau s(x,p))$.
- Berechne Armijo-Schrittweite:

  Solange $\Psi_\sigma(x+tp) > \Psi_\sigma(x) + \alpha t \nabla\Psi_\sigma(x)^T p$

  $t := \rho t$.
- Setze $x_+ := x + tp$.

Wir wollen nichts zur Konvergenz dieses Verfahrens aussagen (auch wenn es möglich wäre), sondern es im Anschluss an einem Beispiel erproben. Dazu bemerken wir noch, dass bei einer konvexen, quadratischen Restriktion die maximale Schrittweite geschlossen berechnet werden kann. Denn sei g gegeben durch $g(x) := c^T x + \frac{1}{2}x^T Q x$ mit $c \in \mathbb{R}^n$ und symmetrischem, positiv semidefinitem $Q \in \mathbb{R}^{n\times n}$. Sei $g(x) < 0$ und $p \in \mathbb{R}^n \setminus \{0\}$. Wegen

$$g(x+tp) = \underbrace{g(x)}_{<0} + t\nabla g(x)^T p + \tfrac{1}{2}t^2 \underbrace{p^T Q p}_{\ge 0}$$

unterscheiden wir zwei Fälle. Ist $p^TQp > 0$ (z. B. ist dies der Fall, wenn Q sogar positiv definit ist), so ist die maximale Schrittweite die positive Nullstelle von

$$t\nabla f(x)^T p + \tfrac{1}{2}t^2 p^T Q p = -g(x),$$

d. h. es ist

$$s(x,p) = \frac{\sqrt{(\nabla g(x)^T p)^2 - 2g(x)\, p^T Q p} - \nabla g(x)^T p}{p^T Q p}.$$

Ist dagegen $p^TQp = 0$, damit $Qp = 0$ und $\nabla g(x)^T p = c^T p$, so ist

$$s(x,p) = \begin{cases} +\infty, & \text{falls } c^T p \le 0, \\ -g(x)/c^T p, & \text{falls } c^T p > 0. \end{cases}$$

Bei mehr als einer konvexen, quadratischen Restriktion muss man natürlich das Minimum entsprechender Größen nehmen. □

Beispiel: Wir betrachten die Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := 2x_1^2 + x_2^2 - 4x_1 + 4x_2 \quad \text{auf} \\ M := \{x \in \mathbb{R}^2 : g(x) := 6x_1^2 - x_1x_2 + 2x_2^2 - 1 \le 0\}. \end{cases}$$

Die Voraussetzungen (V1)–(V3) sind für (P) offenbar erfüllt. In Abbildung 7.4 stellen wir die Menge M der zulässigen Lösungen und einige Höhenlinien dar. Beim ersten

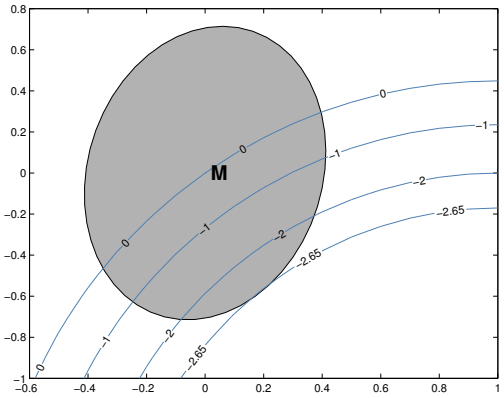


Abbildung 7.4: Eine konvexe, nichtlinear restringierte Optimierungsaufgabe

Problem (P_1) starten wir mit $(0,0)^T$, danach nehmen wir als Startwert für das Newton-

Verfahren den gerade vorher gefundenen Wert. Wir erhalten die folgenden Werte:

k	σ_k	x_k		$\|\nabla\Psi_{\sigma_k}(x_k)\|$	$\kappa(\nabla^2\Psi_{\sigma_k}(x_k))$	iter_k
1	1	0.1239344	−0.3998605	$1.256074 \cdot 10^{-15}$	3.483388	9
2	5	0.1727696	−0.5343143	$2.248073 \cdot 10^{-12}$	10.46055	8
3	25	0.1864910	−0.5702867	$1.478890 \cdot 10^{-14}$	46.77111	8
4	125	0.1894581	−0.5779692	$4.499349 \cdot 10^{-14}$	228.5566	8
5	625	0.1900612	−0.5795267	$2.702942 \cdot 10^{-13}$	1137.527	9
6	3125	0.1901822	−0.5798390	$1.299999 \cdot 10^{-12}$	5682.387	8
7	15625	0.1902064	−0.5799015	$6.990949 \cdot 10^{-12}$	28406.69	8
8	78125	0.1902113	−0.5799140	$1.165032 \cdot 10^{-11}$	142028.2	8

Hierbei brechen wir das Newton-Verfahren ab, wenn $\|\nabla\Psi_{\sigma_k}(x_k)\| \leq 10^{-10}$, durch iter_k wird die Anzahl der Iterationen gezählt. Das starke Ansteigen der Kondition der Hesseschen der logarithmischen Barriere-Funktion ist gut zu beobachten. □

7.1.3 Nichtdifferenzierbare, exakte Straffunktionen

Gegeben sei 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, stationäre) Lösung von (P). Ferner wird wieder angenommen, die 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

$$\Phi_\sigma(x) := f(x) + \frac{\sigma}{2}\Big(\sum_{i=1}^{l} \max(g_i(x), 0)^2 + \|h(x)\|^2\Big)$$

hat den Nachteil, dass die zugehörige unrestringierte Optimierungsaufgabe mit wachsendem σ immer schlechter konditioniert ist. Man stellt sich daher die Frage, ob man 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 möglich ist, die dabei auftretenden Straffunktionen (die dann auch *exakt* genannt werden) aber nichtdifferenzierbar sind. Die bekannteste nichtdifferenzierbare exakte Straffunktion ist die (exakte) L_1-Straffunktion, welche durch

$$\Psi_\sigma(x) := f(x) + \sigma\Big(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1\Big)$$

definiert ist. 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.

Beispiel: Zu

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) := x^2 \quad \text{auf} \quad M := \{x \in \mathbb{R} : h(x) := x - 1 = 0\}$$

gehört die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := x^2 + \sigma\,|x - 1|, \quad x \in \mathbb{R}.$$

Natürlich ist $x^* := 1$ die einzige zulässige Lösung und damit die Lösung von (P). In Abbildung 7.5 links geben wir die Abbildung Ψ_σ für $\sigma = 0.5$ an. Man erkennt, dass

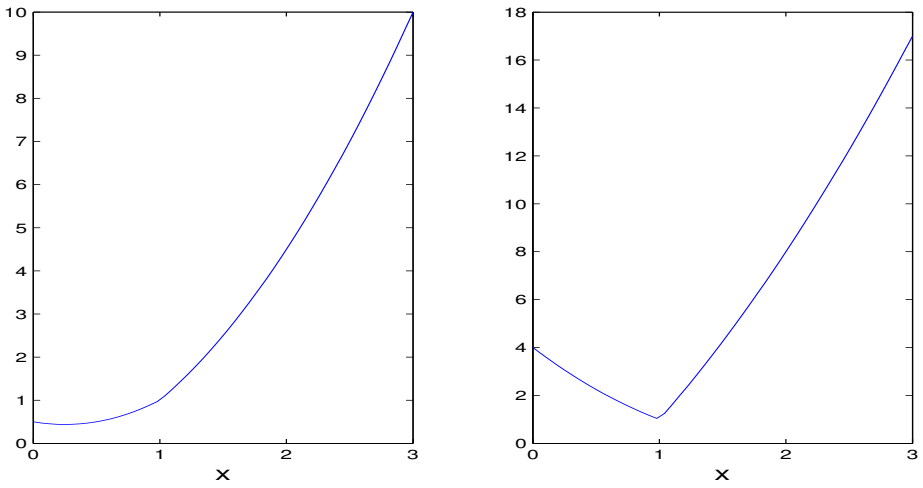


Abbildung 7.5: Die exakte L_1-Straffunktion für $\sigma = 0.5$ und $\sigma = 4$

$x^* = 1$ keine Lösung von $(\mathrm{P}_{0.5})$ ist. Im Gegensatz hierzu zeichnen wir in Abbildung 7.5 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 $\sigma \geq 2$ ist daher $x^* = 1$ die Lösung der unrestringierten Optimierungsaufgabe (P_σ). □

Jetzt wollen wir die Exaktheit der L_1-Straffunktion für gewisse Prolemklassen nachweisen. Wir haben also unter gewissen Voraussetzungen zu zeigen:

- Ist $x^* \in M$ eine (stationäre, lokale, globale) Lösung von (P), so existiert ein $\sigma^* > 0$ derart, dass x^* für alle $\sigma \geq \sigma^*$ eine (stationäre, lokale, globale) Lösung von (P) ist.

Für eine konvexe Optimierungsaufgabe mit differenzierbaren Daten fallen die eben genannten drei Lösungsbegriffe zusammen. Daher ist zu vermuten, dass dieser Fall am einfachsten zu klären ist. Dies ist richtig, wie der nächste Satz (siehe auch C. GEIGER, C. KANZOW (2002, S. 220)) zeigt.

Satz 1.4 *Gegeben sei die konvexe Optimierungsaufgabe*

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

Es seien also $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ und $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ (komponentenweise) konvex, $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear. Weiter seien f und g stetig differenzierbar. Ist $x^* \in M$ eine stationäre (und damit auch globale) Lösung von (P), so existiert ein $\sigma^* > 0$ derart, dass x^* eine globale Lösung von (P_σ) für alle $\sigma \ge \sigma^*$ ist.

Beweis: Da $x^* \in M$ eine stationäre Lösung von (P) ist bzw. die notwendigen Optimalitätsbedingungen erster Ordnung in x^* erfüllt sind, existiert ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

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

Wir definieren $\sigma^* := \max(\|u^*\|_\infty, \|v^*\|_\infty)$. Für beliebige $\sigma \ge \sigma^*$ und $x \in \mathbb{R}^n$ ist dann

$$\begin{aligned}
\Psi_\sigma(x^*) &= f(x^*) \\
&= f(x^*) + \underbrace{g(x^*)^T u^*}_{=0} + \underbrace{h(x^*)}_{=0}^T v^* \\
&= f(x^*) + g(x^*)^T u^* + h(x^*)^T v^* + [\underbrace{\nabla f(x^*) + g'(x^*)^T u^* + (h')^T v^*}_{=0}]^T (x - x^*) \\
&= \underbrace{f(x^*) + \nabla f(x^*)^T (x - x^*)}_{\le f(x)} + [\underbrace{g(x^*) + g'(x^*)(x - x^*)}_{\le g(x)}]^T \underbrace{u^*}_{\ge 0} \\
&\qquad + [\underbrace{h(x^*) + h'(x - x^*)}_{=h(x)}]^T v^* \\
&\le f(x) + g(x)^T u^* + h(x)^T v^* \\
&\qquad \text{(wegen der Konvexitätsvoraussetzungen, siehe Satz 1.5 in 2.1.2)} \\
&\le f(x) + \sum_{i=1}^{l} u_i^* \max(g_i(x), 0) + \sum_{j=1}^{m} v_j^* h_j(x) \\
&\le f(x) + \sum_{i=1}^{l} u_i^* \max(g_i(x), 0) + \sum_{j=1}^{m} |v_j^*|\, |h_j(x)| \\
&\le f(x) + \sigma^* \Big(\sum_{i=1}^{l} \max(g_i(x), 0) + \sum_{j=1}^{m} |h_j(x)| \Big) \\
&= \Psi_{\sigma^*}(x) \\
&\le \Psi_\sigma(x),
\end{aligned}$$

also ist x^* ein globales Minimum von Ψ_σ für alle $\sigma \ge \sigma^*$. □

Bemerkung: Die L_1-Straffunktion $\Psi_\sigma(\cdot)$ ist zwar nicht im üblichen Sinne differenzierbar (siehe Abbildung 7.5 rechts), sie ist aber in jedem Punkt x^*, in dem f, g und h stetig differenzierbar sind, in jede Richtung p *richtungsdifferenzierbar* bzw. besitzt eine *Richtungsableitung*. Das bedeutet, dass der einseitige Grenzwert

$$\Psi'_\sigma(x^*;p) := \lim_{t\to 0+} \frac{\Psi_\sigma(x^*+tp)-\Psi_\sigma(x^*)}{t}$$

für jedes $p \in \mathbb{R}^n$ existiert. Genauer kann man ziemlich einfach zeigen, dass

$$\begin{aligned}\Psi'_\sigma(x^*;p) \;=\; & \nabla f(x^*)^T p + \sigma\Big(\sum_{i\in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{i\notin I^*} \tau_i \nabla g_i(x^*)^T p \\ & + \sum_{j\in J^*} |\nabla h_j(x^*)^T p| + \sum_{j\notin J^*} \operatorname{sign}[h_j(x^*)]\nabla h_j(x^*)^T p\Big).\end{aligned}$$

Hierbei ist

$$I^* := \{i \in \{1,\dots,l\} : g_i(x^*) = 0\}, \qquad J^* := \{j \in \{1,\dots,m\} : h_j(x^*) = 0\},$$

ferner sind 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. □

Auch ohne Konvexitätsvoraussetzungen kann ein Satz 1.4 sehr ähnliches Resultat bewiesen werden. Es gilt nämlich:

Satz 1.5 *Ist $x^* \in M$ eine stationäre Lösung von*

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

und sind f, g und h in x^* stetig differenzierbar, existiert also ein Paar $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ mit

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

so ist x^* für alle $\sigma \ge \sigma^* := \max(\|u^*\|_\infty, \|v^*\|_\infty)$ eine stationäre Lösung von

$$(\mathrm{P}_\sigma) \qquad \textit{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma\Big(\sum_{i=1}^{l} \max(g_i(x), 0) + \sum_{j=1}^{n} |h_j(x)|\Big), \quad x \in \mathbb{R}^n,$$

d. h. für jedes $p \in \mathbb{R}^n$ ist $\Psi'_\sigma(x^*;p) \ge 0$.

Beweis: Sei $\sigma \ge \sigma^* := \max(\|u^*\|_\infty, \|v^*\|_\infty)$, also insbesondere $\sigma \ge u_i^*$, $i = 1,\dots,l$, und $\sigma \ge |v_j^*|$, $j = 1,\dots,m$. Mit den Bezeichnungen I^*, J^* und τ_i, $i \in \{1,\dots,l\} \setminus I^*$,

der letzten Bemerkung ist wegen $x^* \in M$ (also $g_i(x^*) \le 0$, $i = 1, \dots, l$, und $h_j(x^*) = 0$, $j = 1, \dots, m$)

$$\begin{aligned}
\Psi'_\sigma(x^*; p) &= \nabla f(x^*)^T p + \sigma\Big(\sum_{i \in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{j=1}^{m} |\nabla h_j(x^*)^T p|\Big) \\
&= -\sum_{i \in I^*} u_i^* \nabla g_i(x^*)^T p + \sigma \sum_{i \in I^*} \max(\nabla g_i(x^*)^T p, 0) \\
&\qquad - \sum_{j=1}^{n} v_j^* \nabla h_j(x^*)^T p + \sigma \sum_{j=1}^{m} |\nabla h_j(x^*)^T p| \\
&\ge \sum_{i \in I^*} \underbrace{u_i^*}_{\ge 0} \underbrace{[\max(\nabla g_i(x^*)^T p, 0) - \nabla g_i(x^*)^T p]}_{\ge 0} \\
&\qquad + \sum_{j=1}^{m} \underbrace{[|v_j^*|\, |\nabla h_j(x^*)^T p| - v_j^* \nabla h_j(x^*)^T p]}_{\ge 0}.
\end{aligned}$$

Damit ist der Satz bewiesen. □

Ein letztes Ergebnis zur Exaktheit der L_1-Straffunktion ist der folgende Satz.

Satz 1.6 *Gegeben sei die Optimierungsaufgabe*

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

Die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ *sowie die Restriktionsabbildungen* $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ *und* $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *seien in* $x^* \in M$ *zweimal stetig differenzierbar. In* $x^* \in M$ *seien die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt*[6]*, d. h. es existiere ein Paar* $(u^*, v^*) \in \mathbb{R}^l \times \mathbb{R}^m$ *mit:*

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

und

$$p^T W^* p > 0 \qquad \text{für alle } p \in L^0(M; x^*) \setminus \{0\},$$

wobei

$$W^* := \nabla^2 f(x^*) + \sum_{i=1}^{l} u_i^* \nabla^2 g_i(x^*) + \sum_{i=1}^{m} v_i^* \nabla^2 h_i(x^*)$$

und

$$L^0(M; x^*) := \Big\{p \in \mathbb{R}^n : \begin{array}{l} \nabla g_i(x^*)^T p = 0,\ i \in I_+^*, \\ \nabla g_i(x^*)^T p \le 0,\ i \in I^* \setminus I_+^*, \end{array} \ h'(x^*)p = 0\Big\}$$

mit

$$I^* := \{i \in \{1, \dots, l\} : g_i(x^*) = 0\}, \qquad I_+^* := \{i \in \{1, \dots, l\} : u_i^* > 0\}.$$

Dann ist x^* *für alle* $\sigma > \sigma^* := \max(\|u^*\|_\infty, \|v^*\|_\infty)$ *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* $x \in U^* \setminus \{x^*\}$.

Der Beweis wird in der Vorlesung nicht gebracht.

Beweis: Sei $\sigma > \sigma^* := \max(\|u^*\|_\infty, \|v^*\|_\infty)$ vorgegeben. Angenommen, die Behauptung sei falsch. Dann gibt es eine gegen x^* konvergierende Folge $\{x_k\}$, $x_k \neq x^*$ für alle k, mit $\Psi_\sigma(x_k) \leq \Psi_\sigma(x^*)$. 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 $\|p\| = 1$ 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 Beginn des Beweises von Satz 1.5)

$$\begin{aligned}
0 &\geq \Psi'_\sigma(x^*; p) \\
&= \nabla f(x^*)^T p + \sigma\Big(\sum_{i \in I^*} \max(\nabla g_i(x^*)^T p, 0) + \sum_{j=1}^m |\nabla h_j(x^*)^T p|\Big) \\
&= \sum_{i \in I^*} \underbrace{[\sigma \max(\nabla g_i(x^*)^T p, 0) - u_i^* \nabla g_i(x^*)^T p]}_{\geq 0} + \sum_{j=1}^m \underbrace{[\sigma |\nabla h_j(x^*)^T p| - v_j^* \nabla h_j(x^*)^T p]}_{\geq 0}
\end{aligned}$$

Hieraus folgt

$$\max(\nabla g_i(x^*)^T p, 0) = \frac{u_i^*}{\sigma} \nabla g_i(x^*)^T p, \qquad i \in I^*,$$

und

$$|\nabla h_j(x^*)^T p| = \frac{v_j^*}{\sigma} \nabla h_j(x^*)^T p, \qquad j = 1, \ldots, m.$$

Hieraus wollen wir schließen, dass $p \in L^0(M; x^*)$. Aus der ersten Beziehung erhalten wir: Ist $i \in I_+^*$, also $i \in I^*$ mit $u_i^* > 0$, so ist zunächst $\nabla g_i(x^*)^T p \geq 0$ und dann $\nabla g_i(x^*)^T p = 0$ wegen $u_i^* < \sigma$. Ist dagegen $i \in I^* \setminus I_+^*$, also $u_i^* = 0$, so ist $\nabla g_i(x^*)^T p \leq 0$. Aus der zweiten Beziehung folgt wegen $|v_j^*| < \sigma$, dass $\nabla h_j(x^*)^T p = 0$. Wegen $p \neq 0$ ist daher insgesamt $p \in L^0(M; x^*) \setminus \{0\}$. Für alle hinreichend großen k ist $g_i(x_k) < 0$ für alle $i \in \{1, \ldots, l\} \setminus I^*$ und daher

$$\begin{aligned}
0 &\geq \Psi_\sigma(x_k) - \underbrace{\Psi_\sigma(x^*)}_{=f(x^*)} \\
&= f(x_k) - f(x^*) + \sigma\Big(\sum_{i=1}^l \max(g_i(x_k), 0) + \sum_{j=1}^m |h_j(x_k)|\Big) \\
&= f(x_k) - f(x^*) + \sigma\Big(\sum_{i \in I^*} \max(g_i(x_k), 0) + \sum_{j=1}^m |h_j(x_k)|\Big)
\end{aligned}$$

[6] Siehe Satz 2.7 in Abschnitt 3.2.

$$
\begin{aligned}
&\geq f(x_k) - f(x^*) + \sum_{i\in I^*} u_i^* \max(g_i(x_k), 0) + \sum_{j=1}^{m} |v_j^*|\,|h_j(x_k)| \\
&\geq f(x_k) - f(x^*) + \sum_{i\in I^*} u_i^* g_i(x_k) + \sum_{j=1}^{m} v_j^* h_j(x_k) \\
&= f(x_k) - f(x^*) + \sum_{i\in I^*} u_i^* [g_i(x_k) - g_i(x^*)] + \sum_{j=1}^{m} v_j^* [h_j(x_k) - h_j(x^*)] \\
&= t_k \underbrace{\Big[\nabla f(x^*) + \sum_{i\in I^*} u_i^* \nabla g_i(x^*) + \sum_{j=1}^{m} v_j^* \nabla h_j(x^*)\Big]^T}_{=0} p_k + \tfrac{1}{2} t_k^2 p_k^T W_k p_k \\
&= \tfrac{1}{2} t_k^2 p_k^T W_k p_k
\end{aligned}
$$

mit

$$W_k = \nabla^2 f(x_k^{(0)}) + \sum_{i\in I^*} u_i^* \nabla^2 g_i(\hat{x}_k^{(i)}) + \sum_{j=1}^{m} v_j^* \nabla^2 h_j(\overline{x}_k^{(j)}),$$

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 $p_k^T W_k p_k \leq 0$ nach dem Grenzübergang $k \to \infty$, dass $p^T W^* p \leq 0$, womit der gesuchte Widerspruch erhalten ist. □

Beispiel: Wir betrachten die Aufgabe

(P) Minimiere $f(x) := x_1 x_2^2$ auf $M := \{x \in \mathbb{R}^2 : g(x) := x_1^2 + x_2^2 - 2 \leq 0\}$.

Die zulässige Menge M und einige Höhenlinien haben wir in Abbildung 7.6 links wiedergegeben. Sei $x^* \in M$ eine lokale Lösung. Wegen Satz 2.2 in Abschnitt 3.2 (die

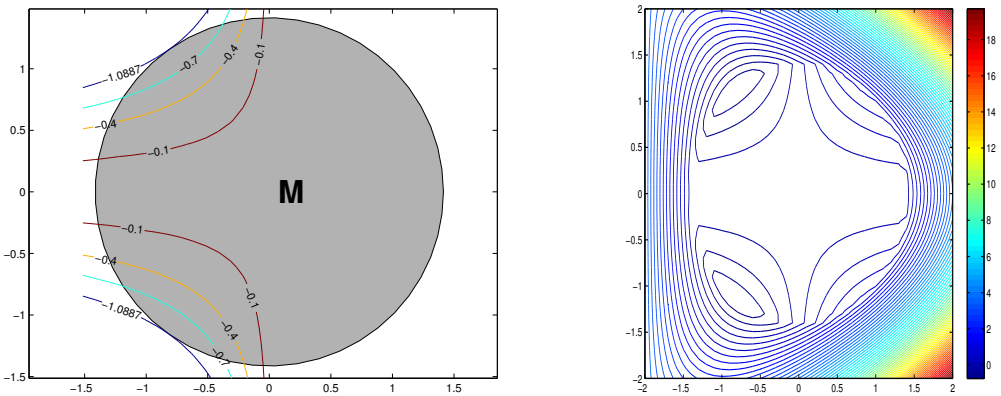


Abbildung 7.6: Höhenlinien und zulässiger Bereich, Höhenlinienplot von Ψ_2

Constraint Qualification ist offensichtlich erfüllt) existiert ein $u^* \in \mathbb{R}$ mit

$$u^* \geq 0, \qquad \begin{pmatrix} (x_2^*)^2 \\ 2x_1^* x_2^* \end{pmatrix} + u^* \begin{pmatrix} 2x_1^* \\ 2x_2^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad [(x_1^*)^2 + (x_2^*)^2 - 2]u^* = 0.$$

Wir nehmen an[7], es sei $u^* > 0$ und folglich $x_2^* \neq 0$. Dann ist $u^* = -x_1^*$ und $(x_2^*)^2 - 2(x_1^*)^2 = 0$. Als Lösung von

$$-2(x_1^*)^2 + (x_2^*)^2 = 0, \qquad (x_1^*)^2 + (x_2^*)^2 = 2$$

erhalten wir $x_\pm^* = (-\sqrt{\frac{2}{3}}, \pm 2\sqrt{\frac{1}{3}})^T$ als stationäre Lösungen von (P) mit zugehörigem Lagrange-Multiplikator $u^* = \sqrt{\frac{2}{3}}$. Es ist

$$\begin{aligned} W_\pm^* &:= \nabla^2 f(x_\pm^*) + u^* \nabla^2 g(x_\pm^*) \\ &= \begin{pmatrix} 0 & \pm 4\sqrt{\frac{1}{3}} \\ \pm 4\sqrt{\frac{1}{3}} & -2\sqrt{\frac{2}{3}} \end{pmatrix} + \sqrt{\tfrac{2}{3}} \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \\ &= 2\sqrt{\tfrac{1}{3}} \begin{pmatrix} \sqrt{2} & \pm 2 \\ \pm 2 & 0 \end{pmatrix}. \end{aligned}$$

Ist $\nabla g(x^*)^T p = 0$ bzw. $(-1, \pm\sqrt{2})^T p = 0$, so ist $p = \alpha(\pm\sqrt{2}, 1)^T$ mit $\alpha \in \mathbb{R}$ und $p^T W_\pm^* p > 0$ für $p \neq 0$. Daher sind die hinreichenden Optimalitätsbedingungen zweiter Ordnung in $x_\pm^*$ erfüllt. Wegen Satz 1.6 wissen wir, dass $x_\pm^*$ eine isolierte, lokale Lösung von

$$(\mathrm{P}_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := x_1 x_2^2 + \sigma \max(x_1^2 + x_2^2 - 2, 0), \quad x \in \mathbb{R}^2,$$

für alle $\sigma > u^* = \sqrt{\frac{2}{3}} \approx 0.8165$ ist. In Abbildung 7.6 rechts haben wir einen Höhenlinienplot von Ψ_2 über dem Quadrat $[-2, 2] \times [-2, 2]$ wiedergegeben. □

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 den SQP-Methoden, wobei SQP für **S**equential **Q**uadratic **P**rogramming steht.

7.1.4 Erweitertes Lagrange-Verfahren

In diesem Unterabschnitt betrachten wir die Aufgabe

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

wobei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ und $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ zweimal stetig differenzierbar sind. Die Lagrange-Funktion $L : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$ zu (P) ist bekanntlich gegeben durch

$$L(x, y) := f(x) + h(x)^T y.$$

[7] Offenbar ist $x^* = (x_1^*, 0)^T$ mit $|x_1^*| \leq \sqrt{2}$ eine stationäre Lösung von (P) mit zugehörigem Lagrange-Multiplikator $u^* = 0$ und Zielfunktionswert $f(x^*) = 0$. Ist x^* sogar eine lokale Lösung von (P)?

Dagegen heißt bei vorgegebenem $\sigma > 0$ die Funktion $L_\sigma : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$, definiert durch

$$L_\sigma(x, y) := f(x) + \frac{\sigma}{2}\|h(x)\|^2 + h(x)^T y,$$

die zu (P) gehörige *erweiterte Lagrange-Funktion* (engl.: *augmented Lagrangian function*). Dies ist offenbar genau die Lagrange-Funktion zu der Aufgabe

$$\text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \frac{\sigma}{2}\|h(x)\|^2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : h(x) = 0\}.$$

Wir werden im weiteren annehmen, dass in einem Punkt $x^* \in M$ die hinreichenden Optimalitätsbedingungen zweiter Ordnung für (P) erfüllt sind, also ein $y^* \in \mathbb{R}^m$ existiert mit $\nabla f(x^*) + h'(x^*)^T y^* = 0$ und der Eigenschaft, dass

$$\nabla^2_{xx} L(x^*, y^*) = \nabla^2 f(x^*) + \sum_{i=1}^{m} y_i^* \nabla^2 h_i(x^*)$$

auf Kern $(h'(x^*))$ positiv definit ist. Dann ist

$$\nabla_x L_\sigma(x, y) = \nabla f(x) + \sigma h'(x)^T h(x) + h'(x)^T y$$

und folglich $\nabla_x L_\sigma(x^*, y^*) = 0$. Also ist x^* ein stationärer Punkt von $L_\sigma(\cdot, y^*)$. Wir werden in Kürze nachweisen, dass $\nabla^2_{xx} L_\sigma(x^*, y^*)$ für alle hinreichend großen σ positiv definit ist. Dies impliziert, dass x^* für alle hinreichend großen σ ein isoliertes, lokales Minimum von $L_\sigma(\cdot, y^*)$ ist.

Beispiel: Wir betrachten[8] die Aufgabe

$$\text{(P)} \quad \text{Minimiere} \quad f(x) := -x_1 x_2^2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^2 : h(x) := x_1^2 + x_2^2 - 1 = 0\}.$$

Stationäre Lösungen von (P), in denen die hinreichenden Optimalitätsbedingungen zweiter Ordnung erfüllt sind, sind $x^*_\pm = (\sqrt{\frac{1}{3}}, \pm\sqrt{\frac{2}{3}})^T$ mit dem Lagrange-Multiplikator $y^* = \sqrt{\frac{1}{3}}$. Die erweiterte Lagrange-Funktion zu (P) ist

$$L_\sigma(x, y) = -x_1 x_2^2 + \frac{\sigma}{2}(x_1^2 + x_2^2 - 1)^2 + y(x_1^2 + x_2^2 - 1).$$

Dann ist

$$\nabla_x L_\sigma(x, y) = \begin{pmatrix} -x_2^2 \\ -2x_1 x_2 \end{pmatrix} + 2\sigma(x_1^2 + x_2^2 - 1)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + 2y\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

und

$$\nabla^2_{xx} L_\sigma(x, y) = 2\begin{pmatrix} \sigma(3x_1^2 + x_2^2 - 1) + y & -x_2 + 2\sigma x_1 x_2 \\ -x_2 + 2\sigma x_1 x_2 & -x_1 + \sigma(x_1^2 + 3x_2^2 - 1) + y \end{pmatrix}.$$

[8] Siehe C. Geiger, C. Kanzow (2002, S. 305).

Insbesondere ist

$$\nabla^2_{xx} L_\sigma(x^*_\pm, y^*) = 2 \begin{pmatrix} \sigma\frac{2}{3} + \sqrt{\frac{1}{3}} & \mp\sqrt{\frac{2}{3}} \pm \sigma\frac{2}{3}\sqrt{2} \\ \mp\sqrt{\frac{2}{3}} \pm \sigma\frac{2}{3}\sqrt{2} & \sigma\frac{4}{3} \end{pmatrix}.$$

Diese (symmetrische) 2×2-Matrix hat für $\sigma > 0$ positive Diagonalelemente und ist daher positiv definit, wenn die Determinante positiv ist. Wegen

$$\det(\nabla^2_{xx} L_\sigma(x^*_\pm, y^*)) = 4(4\sigma\sqrt{\tfrac{1}{3}} - \tfrac{2}{3})$$

ist $\nabla^2_{xx} L_\sigma(x^*_\pm, y^*)$ für alle $\sigma > \frac{1}{2}\sqrt{\frac{1}{3}}$ positiv definit. □

Nun wollen wir das vermutete und im letzten Beispiel in einem Spezialfall nachgewiesene Ergebnis beweisen. Hierzu formulieren und beweisen wir zunächst einen Hilfssatz[9].

Lemma 1.7 *Seien $P, Q \in \mathbb{R}^{n\times n}$ symmetrisch, Q positiv semidefinit und P positiv definit auf* Kern (Q)*. Dann existiert ein $\sigma^* > 0$ derart, dass $P + \sigma Q$ für alle $\sigma \geq \sigma^*$ positiv definit ist.*

Beweis: Der Beweis erfolgt durch Widerspruch. Angenommen, zu jedem $k \in \mathbb{N}$ existiert ein $x_k \in \mathbb{R}^n$ mit $x_k \neq 0$ und

$$(*) \qquad x_k^T P x_k + k x_k^T Q x_k \leq 0.$$

O. B. d. A. ist $\|x_k\| = 1$. Es existiert eine konvergente Teilfolge $\{x_k\}_{k\in K}$, für den Limes x^* gilt $\|x^*\| = 1$. Aus $(*)$ folgt

$$x_k^T Q x_k \leq -\frac{1}{k} x_k^T P x_k.$$

Die rechte Seite konvergiert gegen 0, mit dem Grenzübergang $k \in K,\ k \to \infty$, folgt $(x^*)^T Q x^* \leq 0$ und damit $(x^*)^T Q x^* = 0$, da Q als positiv semidefinit vorausgesetzt ist. Hieraus folgt aber[10] $Qx^* = 0$. Aus $(*)$ folgt aber auch $x_k^T P x_k \leq 0$, da Q positiv semidefinit ist. Mit dem Grenzübergang $k \in K,\ k \to \infty$, folgt $(x^*)^T P x^* \leq 0$, was ein Widerspruch dazu ist, dass P auf Kern (Q) positiv definit ist. □

Als Folgerung aus dem letzten Lemma erhalten wir das gewünschte Ergebnis.

Satz 1.8 *Gegeben sei die gleichungsrestringierte Optimierungsaufgabe (P) mit zweimal stetig differenzierbaren f, h. Die zugehörige Lagrange-Funktion sei L, die erweiterte Lagrange-Funktion L_σ. Sei $x^* \in M$ eine stationäre Lösung von (P), in dem die hinreichende Optimalitätsbedingung zweiter Ordnung erfüllt ist, d. h. es existiere ein $y^* \in \mathbb{R}^m$*

[9] Siehe D. P. Bertsekas (1999, S. 298) oder C. Geiger, C. Kanzow (2002, S. 229).

[10] Hier kann man z. B. folgendermaßen schließen: Aus

$$0 = (x^*)^T Q x^* = (x^*)^T Q^{1/2} Q^{1/2} x^* = (Q^{1/2} x^*)^T (Q^{1/2} x^*) = \|Q^{1/2} x^*\|^2$$

folgt zunächst $Q^{1/2} x^* = 0$ und dann $Qx^* = 0$.

mit $\nabla f(x^*) + h'(x^*)^T y^* = 0$ *und der Eigenschaft, dass* $\nabla^2_{xx} L(x^*, y^*)$ *auf* $\operatorname{Kern}(h'(x^*))$ *positiv definit ist. Dann existiert ein* $\sigma^* > 0$ *derart, dass* $\nabla^2_{xx} L_\sigma(x^*, y^*)$ *für alle* $\sigma \geq \sigma^*$ *positiv definit ist. Insbesondere ist* $x^* \in M$ *für alle* $\sigma \geq \sigma^*$ *ein isoliertes, lokales Minimum von* $L_\sigma(\cdot, y^*)$.

Beweis: Wegen $L_\sigma(x, y) = L(x, y) + \frac{1}{2}\sigma\, h(x)^T h(x)$ ist

$$\nabla^2_{xx} L_\sigma(x, y) = \nabla^2_{xx} L(x, y) + \sigma\Big(\sum_{i=1}^{m} h_i(x)\nabla^2 h_i(x) + h'(x)^T h'(x)\Big).$$

Wegen $h(x^*) = 0$ bzw. $h_i(x^*) = 0$, $i = 1, \ldots, m$, ist

$$\nabla^2_{xx} L_\sigma(x^*, y^*) = \nabla^2_{xx} L(x^*, y^*) + \sigma h'(x^*)^T h'(x^*).$$

Mit $P := \nabla^2_{xx} L(x^*, y^*)$ und $Q := h'(x^*)^T h'(x^*)$ sind die Voraussetzungen von Lemma 1.7 erfüllt und die Behauptung folgt. □

Bemerkung: Die Voraussetzungen des vorigen Satzes 1.8 seien erfüllt. Sei $y \in \mathbb{R}^m$ eine Näherung für den Lagrange-Multiplikator y^* und $\sigma > 0$ hinreichend groß. Ist x_+ eine stationäre Lösung der unrestringierten Optimierungsaufgabe, $L_\sigma(\cdot, y)$ auf dem $\mathbb{R}^n$ zu minimieren, so ist

$$\nabla_x L_\sigma(x_+, y) = \nabla f(x_+) + h'(x_+)^T[y + \sigma h(x_+)] = 0.$$

Andererseits ist

$$\nabla_x L(x^*, y^*) = \nabla f(x^*) + h'(x^*)^T y^* = 0.$$

Dies legt es nahe, eine neue Näherung y_+ für den Lagrange-Multiplikator y^* gemäß

$$y_+ := y + \sigma h(x_+)$$

zu bestimmen. □

Ein Schritt des erweiterten (augmented) Lagrange-Verfahrens (bei C. Geiger, C. Kanzow (2002, S. 228 ff.) *Multiplier-Penalty-Methode* genannt), angewandt auf die gleichungsrestringierte Optimierungsaufgabe (P), sieht folgendermaßen aus:

- Gegeben seien (vom Iterationsschritt unabhängige) positive Konstanten $\beta > 1$ und $\gamma < 1$, z. B. $\beta = 10$ und $\gamma = 0.25$.
- Gegeben $x \in \mathbb{R}^n$ (Näherung für Lösung x^* von (P)), $y \in \mathbb{R}^m$ (Näherung für Lagrange-Multiplikator y^*) und $\sigma > 0$.
- Bestimme (stationäre, lokale, globale) Lösung x_+ von

$$\text{Minimiere} \quad L_\sigma(x, y), \quad x \in \mathbb{R}^n.$$

- Berechne $y_+ := y + \sigma h(x_+)$.

- Berechne

$$\sigma_+ := \begin{cases} \beta\sigma, & \text{falls } \|h(x_+)\| \geq \gamma \|h(x)\|, \\ \sigma, & \text{falls } \|h(x_+)\| < \gamma \|h(x)\|. \end{cases}$$

Wir haben eine MATLAB-Funktion `Erw_Lag.m` geschrieben, durch die das erweiterte Lagrange-Verfahren implementiert wurde. Diese benutzt die Funktion `BFGS`, bzw. das BFGS-Verfahren mit einer Schrittweitensteuerung durch die Wolfe-Schrittweite:

```
function [x,y,iter]=Erw_Lag(Ziel_fun,Res_fun,x,y,sigma,max_iter,tol);
%*************************************************************************
%Diese Funktion loest die Aufgabe
%         Minimiere Ziel_fun(x) u.d.N. Res_fun(x)=0
%mit dem erweiterten Lagrange-Verfahren
%Input-Parameter:
%        Ziel_fun        [f,g]=Ziel_fun(x) liefert Zielfunktion und
%                        Gradienten in x.
%        Res_fun         [h,J]=Res_fun(x) liefert Wert der Gleichungs-
%                        restriktion und Jacobi-Matrix in x
%        x,y,sigma       Ausgangswerte wie in Vorlesung
%        max_iter        maximale Zahl der Iterationsschritte
%        tol             Es wird abgebrochen, wenn
%                        max(||grad L_x||,||h(x)||)<=tol
%Output-Parameter:
%        x               Loesung
%        y               Lagrange-Multiplikator
%        iter            Anzahl der Iterationen
%*************************************************************************
iter=0;gamma=0.25;beta=10;
[h,J]=feval(Res_fun,x);[f,g]=feval(Ziel_fun,x);nabla_L=g+J'*y;
while (max(norm(nabla_L),norm(h))>tol)&(iter<max_iter)
    x=BFGS(@(x)L_sigma(x,y,sigma,'Ziel_fun','Res_fun'),x,50,0.1*tol);
    [h_p,J_p]=feval(Res_fun,x);[f_p,g_p]=feval(Ziel_fun,x);
    y=y+sigma*h_p;
    if (norm(h_p)>=gamma*norm(h))
        sigma=beta*10;
    end;
    h=h_p;nabla_L=g_p+J_p'*y;iter=iter+1;
end;
function [f,g]=L_sigma(x,y,sigma,Ziel_fun,Res_fun);
%*************************************************************************
%L_sigma ist die erweiterte Lagrange-Funktion zu Ziel_fun, Res_fun
%*************************************************************************
[f,g]=feval(Ziel_fun,x); [h,J]=feval(Res_fun,x);
f=f+0.5*sigma*norm(h)^2+h'*y;
if nargout>1
     g=g+sigma*J'*h+J'*y;
end;
```

Beispiel: Wir kommen auf das letzte Beispiel zurück und betrachten wieder die Aufgabe

$$\text{(P)} \quad \text{Minimiere} \quad f(x) := -x_1 x_2^2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^2 : h(x) := x_1^2 + x_2^2 - 1 = 0\}.$$

Diese hat die Lösungen $x_\pm^* = (\sqrt{\frac{1}{3}}, \pm\sqrt{\frac{2}{3}})^T$ mit dem Lagrange-Multiplikator $y^* = \sqrt{\frac{1}{3}}$. Z. B. erhalten wir durch den Aufruf

```
[x,y,iter]=Erw_Lag('Ziel_fun','Res_fun',[0.3;0.3],[0.1],5,20,1e-8);
```

(vorher haben wir `Ziel_fun` und `Res_fun` entsprechend bereitgestellt) das Ergebnis

$$x = \begin{pmatrix} 0.577350270703048 \\ 0.816496583068009 \end{pmatrix}, \quad y = 0.577350270721633, \quad \text{iter} = 7.$$

Zum Vergleich hier die exakte Lösung

$$x_+^* = \begin{pmatrix} \sqrt{\frac{1}{3}} \\ \sqrt{\frac{2}{3}} \end{pmatrix} = \begin{pmatrix} 0.577350269189626 \\ 0.816496580927726 \end{pmatrix}, \qquad y^* = \sqrt{\tfrac{1}{3}} = 0.577350269189626.$$

□

Beispiel: Wir testen die obige MATLAB-Funktion noch an einem Beispiel bei C. Geiger, C. Kanzow (2002, S. 212 und S. 232):

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := -x_1^2 - 2x_2^2 - x_3^2 - x_1x_2 - x_1x_3 \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^3 : h(x) := \begin{pmatrix} x_1^2 + x_2^2 + x_3^2 - 25 \\ 8x_1 + 14x_2 + 7x_3 - 56 \end{pmatrix} = 0 \right\}. \end{cases}$$

Wie Geiger-Kanzow nehmen wir die Startwerte $x_0 = (3, 0.2, 3)^T$, $y_0 = (0, 0)^T$, $\sigma_0 = 1$. Wir erhalten die folgenden Werte:

k	x_k			y_k		σ_k
0	3.0000000	0.2000000	3.0000000	0	0	1.0
1	3.5761018	0.1442584	3.6632707	1.2288677	0.2713276	1.0
2	3.5119278	0.2176700	3.5515528	1.2234128	0.2750005	1.0
3	3.5121226	0.2169794	3.5521777	1.2234643	0.2749362	1.0
4	3.5121214	0.2169881	3.5521710	1.2234635	0.2749371	1.0

Startet man dagegen mit $x_0 = (0, 0, 0)^T$, $y_0 = (0, 0)^T$, $\sigma_0 = 1$, so erhält man:

k	x_k			y_k		σ_k
0	0	0	0	0	0	1.0
1	0.2732494	4.7972846	−1.8597387	1.5472338	0.3298104	1.0
2	0.3312857	4.6780148	−1.7357682	1.5536958	0.3219229	1.0
3	0.3319989	4.6776595	−1.7347492	1.5537710	0.3219009	1.0
4	0.3320037	4.6776543	−1.7347410	1.5537715	0.3219006	1.0

Der Penalty-Parameter wird in beiden Beispielen nicht erhöht. Der Zielfunktionswert im zweiten Durchlauf ist übrigens kleiner als der im ersten. □

Bemerkung: Hat man in einer nichtlinearen Optimierungsaufgabe auch Ungleichungen als Restriktionen, so kann man diese durch Schlupfvariable auf Gleichungsrestriktionen zurückführen und auf dieses so erhaltene gleichungsrestringierte Problem das erweiterte Lagrange-Verfahren anwenden, siehe C. Geiger, C. Kanzow (2002, S. 231 ff.).

Gegeben sei die Aufgabe

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

wobei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ und $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ hinreichend glatt sind. Wir benutzen im folgenden eine MATLAB-Notation. Ist nämlich $z = (z_i) \in \mathbb{R}^l$, so sei $z.^2 = (z_i^2) \in \mathbb{R}^l$. Dann ist (P) äquivalent zu

$$(\overline{\text{P}}) \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad \overline{f}(x,z) := f(x) \quad \text{auf} \\ \overline{M} := \left\{ (x,z) \in \mathbb{R}^n \times \mathbb{R}^l : \overline{h}(x,z) := \begin{pmatrix} g(x) + z.^2 \\ h(x) \end{pmatrix} = 0 \right\}. \end{array} \right.$$

Als erweiterte Lagrange-Funktion $\overline{L}_\sigma : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$ zu $(\overline{\text{P}})$ erhalten wir

$$\begin{aligned} \overline{L}_\sigma(x,z,u,v) &:= f(x) + \begin{pmatrix} g(x) + z.^2 \\ h(x) \end{pmatrix}^T \begin{pmatrix} u \\ v \end{pmatrix} + \frac{\sigma}{2} \left\| \begin{pmatrix} g(x) + z.^2 \\ h(x) \end{pmatrix} \right\|^2 \\ &= f(x) + (g(x) + z.^2)^T u + h(x)^T v + \frac{\sigma}{2}(\|g(x) + z.^2\|^2 + \|h(x)\|^2). \end{aligned}$$

Bei festen $\sigma > 0$ sowie $(x,u,v) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ kann die Aufgabe

$$\text{Minimiere} \quad \overline{L}_\sigma(x,z,u,v), \quad z \in \mathbb{R}^l$$

geschlossen gelöst und damit die Schlupfvariable z wieder eliminiert werden, wie wir gleich sehen werden. Die Funktion $L_\sigma : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \longrightarrow \mathbb{R}$, definiert durch

$$L_\sigma(x,u,v) := \min_{z \in \mathbb{R}^l} \overline{L}_\sigma(x,z,u,v),$$

bezeichnen wir als *erweiterte Lagrange-Funktion* zur Optimierungsaufgabe (P), welche auch Ungleichungen als Restriktionen enthalten kann. In der Tat gilt für die Lösung z_i^* von

$$\text{Minimiere} \quad \phi_i(z_i) := u_i(g_i(x) + z_i^2) + \frac{\sigma}{2}(g_i(x) + z_i^2)^2, \quad z_i \in \mathbb{R},$$

die Beziehung (Beweis?)

$$(z_i^*)^2 = \max(-(u_i/\sigma + g_i(x)), 0), \qquad i = 1, \ldots, l.$$

Damit wird nach einfacher Rechnung

$$\begin{aligned} L_\sigma(x,u,v) &= \overline{L}_\sigma(x, z^*, u, v) \\ &= f(x) + h(x)^T v + \frac{\sigma}{2}\|h(x)\|^2 + \sum_{i=1}^{l} \phi_i(z_i^*) \\ &= f(x) + h(x)^T v + \frac{\sigma}{2}\|h(x)\|^2 + \frac{1}{2\sigma} \sum_{i=1}^{l} [\max(u_i + \sigma g_i(x), 0)^2 - u_i^2]. \end{aligned}$$

Man beachte, dass $L_\sigma(\cdot, u, v)$ i. Allg. nur einmal stetig differenzierbar ist. Ist x_+ (stationäre, lokale, globale) Lösung von

$$\text{Minimiere} \quad L_\sigma(x, u, v), \quad x \in \mathbb{R}^n,$$

so ist

$$\nabla_x L_\sigma(x_+, u, v) = \nabla f(x_+) + h'(x_+)^T(v + \sigma h(x_+)) + g'(x_+)^T \max(u + \sigma g(x_+), 0) = 0.$$

Hierbei operiere max auf Vektoren komponentenweise. Als Update-Formeln für die Lagrange-Multiplikatoren erhalten wir diesmal

$$u_+ := \max(u + \sigma g(x_+), 0), \qquad v_+ = v + \sigma h(x_+).$$

□

Beispiel: Gegeben sei die Aufgabe[11]

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := -x_1 x_2 x_3 \quad \text{auf} \\ M := \{x \in \mathbb{R}^3 : g(x) := x_1^2 + 2x_2^2 + 4x_3^2 - 48 \le 0\}. \end{cases}$$

Mit einer einfachen Modifikation der obigen Funktion `Erw_Lag` (die Abbruchbedingung und die Update-Formel für die Lagrange-Multiplikatoren müssen modifiziert werden) erhalten wir mit $x_0 = (1, 1, 1)^T$, $u_0 = 0.5$ und $\sigma_0 = 1$ die folgenden Ergebnisse:

k	x_k			u_k	σ_k
0	1.0000000	1.0000000	1.0000000	0.5000000	1.0
1	4.0086842	2.8345675	2.0043421	0.7086419	1.0
2	3.9999356	2.8283815	1.9999678	0.7070954	1.0
3	4.0000005	2.8284276	2.0000002	0.7071069	1.0
4	4.0000000	2.8284271	2.0000000	0.7071068	1.0

Die exakten Lösungen von (P) sind $x_1^* = (4, 2\sqrt{2}, 2)^T$, $x_2^* = (4, -2\sqrt{2}, -2)^T$, $x_3^* = (-4, 2\sqrt{2}, -2)^T$ und $x_4^* = (-4, -2\sqrt{2}, 2)^T$. Mit anderen Startwerten werden andere Lösungen approximiert, wobei nicht recht vorhersehbar zu sein scheint, welche angenähert wird. □

7.1.5 Aufgaben

1. Man betrachte die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := (x_1 - x_2)^2 + (x_2 - x_3)^4 \\ \text{unter der Nebenbedingung} \\ h(x) := (1 + x_2^2)x_1 + x_3^4 - 3 = 0. \end{cases}$$

   (a) Man bestimme alle Lösungen.

[11] Siehe W. Hock, K. Schittkowski (1981, S. 52).

(b) Man löse die Aufgabe numerisch mit Hilfe von `fmincon`. Wie bei Hock-Schittkowski, S. 49, nehme man den Startwert $x_0 := (-2.6, 2, 2)$.

2. Gegeben sei das quadratische Programm

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : h(x) := Ax - b = 0\} \end{cases}$$

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_σ 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.

3. Gegeben sei das quadratische Programm

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : h(x) := Ax - b = 0\} \end{cases}$$

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

$$(\text{P}_\sigma^*) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + (y^*)^T h(x) + \tfrac{1}{2}\sigma \|h(x)\|^2, \quad x \in \mathbb{R}^n,$$

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.

4. 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

$$(\text{P}_\sigma) \qquad \text{Minimiere} \quad \Phi_\sigma(x) := f(x) + \tfrac{1}{2}\sigma h(x)^2, \quad x \in \mathbb{R}^2$$

zu. Man bestimme die Lösung x_σ von (P_σ) und bestätige die Aussage von Aufgabe 2, 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 $\lim_{\sigma\to\infty} \sigma h(x_\sigma) = y^*$.

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

$$(\text{P}_\sigma) \quad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma \underbrace{\Big(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1 \Big)}_{=:S(x)}, \quad x \in \mathbb{R}^n.$$

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), der Beweis ist einfach.

7.2 SQP-Verfahren

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. Das Newton-Verfahren zur Lösung eines nichtlinearen Gleichungssystems beruht darauf, in einer Näherung die nichtlineare Abbildung nach Taylor zu linearisieren, und (oft berechtigt) zu hoffen, dass eine Nullstelle dieser linearisierten Abbildung eine verbesserte Näherung für die Nullstellenaufgabe liefert. Entsprechend werden bei einem SQP-Verfahren zur Lösung von (P) Zielfunktion und Restriktionsabbildungen durch eine quadratische bzw. lineare Approximation ersetzt, und eine Lösung des hierdurch gewonnenen quadratischen Programms als (hoffentlich verbesserte) neue Näherung genommen. Genau wie beim Newton-Verfahren bei unrestringierten Optimierungsaufgaben unterscheiden wir zwischen einem lokalen bzw. ungedämpften SQP-Verfahren und einem durch eine Schrittweitensteuerung globalisierten SQP-Verfahren.

7.2.1 Ungedämpftes SQP-Verfahren

Das ungedämpfte SQP-Verfahren zur Lösung von (P) ist ganz einfach. Wir schildern, um uns lästige Iterationsindizes zu ersparen, nur einen Schritt.

- Sei $(x, u, v) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ als Näherung für ein Kuhn-Tucker-Tripel (x^*, u^*, v^*) zu (P) gegeben. Sei ferner $H \in \mathbb{R}^{n \times n}$ symmetrisch.

- Ist (x, u, v) ein Kuhn-Tucker-Tripel, ist also $x \in M$ und

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

  dann: STOP.

- Berechne eine (stationäre) Lösung $p^* \in \mathbb{R}^n$ und zugehörige Lagrange-Multiplikatoren $(u_+, v_+) \in \mathbb{R}^l \times \mathbb{R}^m$ des quadratischen Hilfsproblems

$$(P_{(x,H)}) \quad \begin{cases} \text{Minimiere} & \nabla f(x)^T p + \frac{1}{2} p^T H p \quad \text{unter den Nebenbedingungen} \\ & g(x) + g'(x)p \leq 0, \quad h(x) + h'(x)p = 0. \end{cases}$$

  Bestimme also $(p^*, u_+, v_+) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u_+ \geq 0, \quad \nabla f(x) + Hp^* + g'(x)^T u_+ + h'(x)^T v_+ = 0, \quad (g(x) + g'(x)p^*)^T u_+ = 0$$

sowie

$$g(x) + g'(x)p^* \leq 0, \qquad h(x) + h'(x)p^* = 0.$$

- Falls $p^* = 0$, dann: STOP, da (x, u_+, v_+) ein Kuhn-Tucker-Tripel ist.
- Setze $x_+ := x + p^*$ und wähle eine symmetrische Matrix $H_+ \in \mathbb{R}^{n\times n}$.

Bemerkung: Das quadratische Hilfsproblem $(\mathrm{P}_{(x,H)})$ ist nicht notwendig zulässig. In einem speziellen Fall kann die Zulässigkeit nachgewiesen werden (siehe C. Geiger, C. Kanzow (2002, S. 263)):

- Ist $g : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ (komponentenweise) konvex, $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear und (P) zulässig, so ist das quadratische Hilfsproblem $(\mathrm{P}_{(x,H)})$ zulässig. Ist zusätzlich $H \in \mathbb{R}^{n\times n}$ positiv definit, so besitzt $(\mathrm{P}_{(x,H)})$ eine eindeutige Lösung p^*.

Denn sei $\hat{x} \in M$. Dann ist $p := \hat{x} - x$ zulässig für $(\mathrm{P}_{(x,H)})$. Ist H positiv definit, so ist die Zielfunktion in $(\mathrm{P}_{(x,H)})$ gleichmäßig konvex und die Existenz und Eindeutigkeit einer Lösung folgt, wie wir uns schon in der Einführung überlegt haben. □

Bemerkung: In dieser Bemerkung betrachten wir den Spezialfall, dass nur Gleichungen als Restriktionen vorliegen, das zu Grunde liegende Problem also durch

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

gegeben ist. Das sogenannte *Lagrange-Newton-Verfahren* (siehe z. B. C. Geiger, C. Kanzow (2002, S. 239 ff.)) besteht darin, das Newton-Verfahren auf das nichtlineare Gleichungssystem

$$\Phi(x, y) := \begin{pmatrix} \nabla f(x) + h'(x)^T y \\ h(x) \end{pmatrix} = 0$$

anzuwenden. Bei einer gegebenen Näherung (x, y) bestimmt man also die Lösung $(p^*, q^*) \in \mathbb{R}^n \times \mathbb{R}^m$ des linearen Gleichungssystems

$$\Phi'(x, y) \begin{pmatrix} p^* \\ q^* \end{pmatrix} = -\Phi(x, y)$$

bzw.

$$\begin{pmatrix} \nabla^2 f(x) + \sum_{i=1}^m y_i \nabla^2 h_i(x) & h'(x)^T \\ h'(x) & 0 \end{pmatrix} \begin{pmatrix} p^* \\ q^* \end{pmatrix} = - \begin{pmatrix} \nabla f(x) + h'(x)^T y \\ h(x) \end{pmatrix}$$

und setzt anschließend $x_+ := x + p^*$, $y_+ := y + q^*$. Die lokale Konvergenz des Lagrange-Newton-Verfahrens ist einfach zu zeigen, wenn man entsprechende Aussagen für das Newton-Verfahren benutzt. Bei entsprechender Glattheit von f und h wissen wir dann nämlich, dass lokale superlineare oder gar quadratische Konvergenz vorliegt, wenn $\Phi(x^*, y^*) = 0$ und $\Phi'(x^*, y^*)$ nichtsingulär ist. Letzteres ist offenbar der Fall, wenn Rang $(h'(x^*)) = m$ und in x^* zusammen mit dem Multiplikator y^* die hinreichende Optimalitätsbedingung zweiter Ordnung erfüllt ist.

Wählt man im SQP-Verfahren bei Vorliegen einer Näherung (x, y) die symmetrische Matrix H durch

$$H := \nabla^2_{xx} L(x, y) = \nabla^2 f(x) + \sum_{i=1}^{m} y_i \nabla^2 h_i(x),$$

wobei $L(x, y) := f(x) + h(x)^T y$ die zu (P) gehörende Lagrange-Funktion ist, so wird eine stationäre Lösung p^* von $(\mathrm{P}_{(x,H)})$ mit Multiplikator y_+ durch Lösen des linearen Gleichungssystems

$$\begin{pmatrix} \nabla^2 f(x) + \sum_{i=1}^m y_i \nabla^2 h_i(x) & h'(x)^T \\ h'(x) & 0 \end{pmatrix} \begin{pmatrix} p^* \\ y_+ \end{pmatrix} = - \begin{pmatrix} \nabla f(x) \\ h(x) \end{pmatrix}$$

gewonnen und dann $x_+ := x + p^*$ gesetzt. Genau dieses Resultat (x_+, y_+) erhält man, wenn man einen Schritt des Lagrange-Newton-Verfahrens durchführt. Daher ist das SQP-Verfahren mit $H := \nabla^2_{xx} L(x, y)$ äquivalent zum Lagrange-Newton-Verfahren. □

Beispiel: Wir wollen das Lagrange-Newton-Verfahren bei einem Beispiel testen, das wir schon auf Seite 200 betrachteten. Gegeben sei also die Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := -x_1^2 - 2x_2^2 - x_3^2 - x_1 x_2 - x_1 x_3 \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^3 : h(x) := \begin{pmatrix} x_1^2 + x_2^2 + x_3^2 - 25 \\ 8x_1 + 14x_2 + 7x_3 - 56 \end{pmatrix} = 0 \right\}. \end{cases}$$

Wieder nehmen wir die Startwerte $x_0 = (3, 0.2, 3)^T$ und $y_0 = (0, 0)^T$. Wir erhalten die folgenden Ergebnisse:

k	x_k			y_k		$\|\Phi(x_k, y_k)\|$
0	3.0000000	0.2000000	3.0000000	0	0	17.1977215
1	3.6332135	0.1591222	3.5295117	1.4754113	0.2628241	2.5325367
2	3.5311730	0.2145396	3.5352945	1.2278141	0.2739328	0.0591377
3	3.5122564	0.2169302	3.5521324	1.2234626	0.2749389	0.0006839
4	3.5121214	0.2169879	3.5521712	1.2234635	0.2749371	0.0000000
5	3.5121214	0.2169879	3.5521712	1.2234635	0.2749371	0.0000000

Mit den Startwerten $x_0 = (0.3, 5, -2)^T$, $y_0 = (0, 0)^T$ erhalten wir:

k	x_k			y_k		$\|\Phi(x_k, y_k)\|$
0	0.3000000	5.0000000	−2.0000000	0	0	21.4762211
1	0.3952224	4.6692300	−1.7901423	1.4134592	0.3526821	1.1511548
2	0.3215196	4.6811891	−1.7298288	1.5534095	0.3218583	0.0284016
3	0.3319997	4.6776643	−1.7347566	1.5537705	0.3219015	0.0001468
4	0.3320037	4.6776543	−1.7347410	1.5537716	0.3219006	0.0000000
5	0.3320037	4.6776543	−1.7347410	1.5537716	0.3219006	0.0000000

□

In der vorigen Bemerkung haben wir uns klar gemacht, dass das Lagrange-Newton-Verfahren für eine gleichungsrestringierte nichtlineare Optimierungsaufgabe lokal superlinear oder quadratisch konvergent ist, wenn der zulässige Punkt x^* regulär ist bzw.

Rang $(h'(x^*)) = m$ gilt und mit dem Multiplikator $y^* \in \mathbb{R}^m$ die hinreichende Optimalitätsbedingung zweiter Ordnung erfüllt ist, also mit $L(x,y) := f(x) + h(x)^T y$ einerseits $\nabla_x L(x^*, y^*) = 0$ gilt und andererseits $\nabla^2_{xx} L(x^*, y^*)$ auf Kern $(h'(x^*))$ positiv definit ist. Dieses Ergebnis kann auf das ungedämpfte bzw. lokale SQP-Verfahren für die allgemeine (d. h. es können Gleichungen und Ungleichungen als Restriktionen auftreten) nichtlineare Optimierungsaufgabe (P) verallgemeinert werden, wobei die Matrix H als Hessesche der Lagrange-Funktion gewählt, also

$$H := \nabla^2_{xx} L(x,u,v) = \nabla^2 f(x) + \sum_{i=1}^{l} u_i \nabla^2 g_i(x) + \sum_{i=1}^{m} v_i \nabla^2 h_i(x)$$

gesetzt wird. Siehe z. B. C. Geiger, C. Kanzow (2002, S. 245).

Beispiel: Wir betrachten die Optimierungsaufgabe[12]

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := (x_1 - x_2)^2 + (x_2 - 1)^2 \quad \text{auf} \\ M := \left\{ x \in \mathbb{R}^2 : g(x) := \begin{pmatrix} x_1^2 + x_2 - 1 \\ x_1^2 - x_2 - 1 \end{pmatrix} \leq 0 \right\}. \end{cases}$$

In Abbildung 7.7 haben wir die zulässige Menge M und einige Höhenlinien eingetragen. Die quadratischen Hilfsprobleme (man beachte: $H_k := \nabla^2_{xx} L(x_k, u_k)$ ist positiv definit)

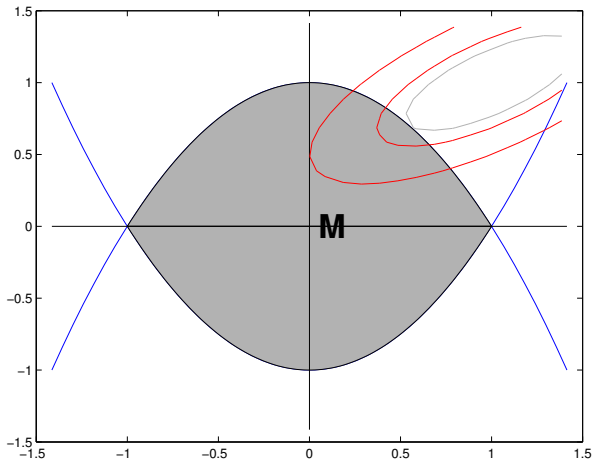


Abbildung 7.7: Zulässige Menge und Höhenlinien der Zielfunktion

lösen wir mit dem Verfahren von Goldfarb-Idnani, wobei wir die selbst geschriebene MATLAB-Funktion `Go_Id` benutzen. Wir erhalten die folgenden Ergebnisse, die mit

[12] Siehe P. Spellucci (1993, S. 457).

den Startwerten $x_0 = (1,1)^T$ und $u_0 = (1,1)^T$ gewonnen wurden.

k	x_k		u_k		$\|\Phi_k\|$
0	1.0000000	1.0000000	1.0000000	1.0000000	4.2426405
1	0.6153846	0.7692308	0.1538462	0	1.0575174
2	0.5456139	0.7071734	0.2625342	0	0.1842120
3	0.5461217	0.7017514	0.2852378	0	0.0260730
4	0.5460955	0.7017797	0.2850723	0	0.0001981
5	0.5460969	0.7017782	0.2850808	0	0.0000102
6	0.5460968	0.7017783	0.2850804	0	0.0000005
7	0.5460968	0.7017783	0.2850804	0	0.0000000

Hierbei ist

$$\Phi_k = \begin{pmatrix} \nabla f(x_k) + g'(x_k)^T u_k \\ \min(-g(x_k), u_k) \end{pmatrix}.$$

□

7.2.2 Gedämpftes SQP-Verfahren

Von dem ungedämpften SQP-Verfahren kann man nur lokale Konvergenz erwarten. Genau wie in der unrestringierten Optimierung erhält man eine Globalisierung durch die Einführung einer Schrittweite. Da die Näherungen i. Allg. unzulässig sind, ist für die Bewertung der Güte einer Näherung nicht alleine der Zielfunktionswert ausreichend, auch die Abweichung von der Zulässigkeit muss berücksichtigt werden. Als wichtig für die Bewertung einer Näherung der nichtlinearen Optimierungsaufgabe

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

wird sich die exakte L_1-Straffunktion (auch *merit function* genannt)

$$\Psi_\sigma(x) := f(x) + \sigma\Big(\sum_{i=1}^{l} \max(g_i(x), 0) + \sum_{j=1}^{m} |h_j(x)|\Big)$$

herausstellen, die wir in Unterabschnitt 7.1.3 untersucht haben. Im folgenden Lemma wird angegeben, wie man zu einer aktuellen Näherung $x \in \mathbb{R}^n$ für eine (globale, lokale, stationäre) Lösung durch Lösen eines quadratischen Hilfsproblems wie beim ungedämpften SQP-Verfahren eine Suchrichtung $p \in \mathbb{R}^n$ bestimmen kann, die für die exakte Straffunktion Ψ_σ für alle hinreichend großen σ eine Abstiegsrichtung in x ist (wenn nicht x schon eine zulässige stationäre Lösung von (P) ist). Hierbei ist p eine *Abstiegsrichtung* in x für Ψ_σ, wenn die Richtungsableitung $\Psi'_\sigma(x;p)$ negativ ist, denn dies bedeutet, dass $\Psi_\sigma(x+tp) < \Psi_\sigma(x)$ für alle hinreichend kleinen $t > 0$. Die Richtungsableitung ist (siehe Bemerkung im Anschluss an Satz 1.4, siehe auch C. Geiger, C. Kanzow (2002, S. 252)) gegeben durch

$$\begin{aligned} \Psi'_\sigma(x;p) \;=\; & \nabla f(x)^T p + \sigma\Big(\sum_{i\in I} \max(\nabla g_i(x)^T p, 0) + \sum_{i\notin I} \tau_i \nabla g_i(x)^T p \\ & + \sum_{j\in J} |\nabla h_j(x)^T p| + \sum_{j\notin J} \operatorname{sign}[h_j(x)] \nabla h_j(x)^T p\Big), \end{aligned}$$

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. Dann gilt (siehe auch C. Geiger, C. Kanzow (2002, S. 253)):

Lemma 2.1 *Gegeben sei ein Paar $(x, H) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}$, wobei $H \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit ist. Es wird vorausgesetzt, dass das quadratische Programm*

$$(\mathrm{P}_{(x,H)}) \qquad \begin{cases} \text{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T H p \quad \text{unter den Nebenbedingungen} \\ g(x) + g'(x)p \le 0, \qquad h(x) + h'(x)p = 0 \end{cases}$$

zulässig ist. Die dann eindeutige Lösung von $(P_{(x,H)})$ werde mit p^* bezeichnet, ferner seien (u_+, v_+) zugehörige Lagrange-Multiplikatoren. Dann gilt:

1. *Ist $p^* = 0$, so ist x eine zulässige, stationäre Lösung von (P).*
2. *Ist $p^* \neq 0$, so ist $\Psi'_\sigma(x; p^*) \le -(p^*)^T H p^*$ für alle $\sigma \ge \max(\|u_+\|_\infty, \|v_+\|_\infty)$, insbesondere also p^* eine Abstiegsrichtung für Ψ_σ in x.*

Beweis: Die Lösung p^* von $(\mathrm{P}_{(x,H)})$ ist durch die Existenz von $(u_+, v_+) \in \mathbb{R}^l \times \mathbb{R}^n$ mit

$$u_+ \ge 0, \qquad \nabla f(x) + Hp^* + g'(x)^T u_+ + h'(x)^T v_+ = 0, \qquad [g(x) + g'(x)p^*]^T u_+ = 0$$

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 stationäre Lösung von (P). Daher sei jetzt $p^* \neq 0$. Für $\sigma \ge \max(\|u_+\|_\infty, \|v_+\|_\infty)$ ist

$$\begin{aligned} \Psi'_\sigma(x; p^*) &= \nabla f(x)^T p^* + \sigma\Big(\sum_{i\in I} \max(\nabla g_i(x)^T p^*, 0) + \sum_{i\notin I} \tau_i \nabla g_i(x)^T p^* \\ &\qquad + \sum_{j\in J} |\nabla h_j(x)^T p^*| + \sum_{j\notin J} \operatorname{sign}[h_j(x)] \nabla h_j(x)^T p^*\Big) \\ &= \underbrace{-(p^*)^T H p^* - u_+^T g'(x) p^* + v_+^T h'(x) p^*}_{=\nabla f(x)^T p^*} \\ &\qquad + \sigma\Big(\sum_{i\in I} \max(\nabla g_i(x)^T p^*, 0) + \sum_{i\notin I} \tau_i \nabla g_i(x)^T p^* \\ &\qquad + \sum_{j\in J} |\nabla h_j(x)^T p^*| + \sum_{j\notin J} \operatorname{sign}[h_j(x)] \nabla h_j(x)^T p^*\Big) \\ &= -(p^*)^T H p^* + u_+^T g(x) + v_+^T h(x) \\ &\qquad + \sigma\Big(\sum_{i\in I} \max(\underbrace{\nabla g_i(x)^T p^*}_{\le 0}, 0) + \sum_{i\notin I} \tau_i \underbrace{\nabla g_i(x)^T p^*}_{\le -g_i(x)} \end{aligned}$$

$$
\begin{aligned}
&\quad + \sum_{j\in J} |\underbrace{\nabla h_j(x)^T p^*}_{=0}| + \sum_{j\notin J} \operatorname{sign}[h_j(x)] \underbrace{\nabla h_j(x)^T p^*}_{=-h_j(x)}\Big) \\
&\leq -(p^*)^T H p^* - \sum_{i\notin I} (\sigma\tau_i - (u_+)_i) g_i(x) - \sum_{j\notin J} (\sigma - (v_+)_j \operatorname{sign}[h_j(x)]) |h_j(x)| \\
&\leq -(p^*)^T H p^* - \sum_{i\notin I, g_i(x)>0} (\underbrace{\sigma - (u_+)_i}_{\geq 0}) g_i(x) \\
&\qquad - \sum_{j\notin J} (\underbrace{\sigma - (v_+)_j \operatorname{sign}[h_j(x)]}_{\geq 0}) |h_j(x)| \\
&\leq -(p^*)^T H p^*.
\end{aligned}
$$

Damit ist das Lemma bewiesen. □

Durch Lemma 2.1 haben wir gezeigt, wie man durch Lösen eines quadratischen Programms eine Abstiegsrichtung p^* für die exakte L_1-Straffunktion Ψ_σ in der aktuellen Näherung x bestimmen oder feststellen kann, dass x eine stationäre Lösung ist. Die neue Näherung in einem gedämpften SQP-Verfahren wird $x_+ := x + tp$ mit einer gewissen Schrittweite $t > 0$ sein. Es liegt nahe, von dieser Schrittweite $t > 0$ analog zu unrestringierten Optimierungsaufgaben zu fordern, dass mit einem vorgegebenen $\alpha \in (0, \frac{1}{2})$, z. B. $\alpha = 0.0001$, die Bedingung

$$\Psi_\sigma(x + tp^*) \leq \Psi_\sigma(x) + \alpha t \Psi'_\sigma(x; p^*)$$

erfüllt ist. Wir ziehen es vor (im Gegensatz zu C. Geiger, C. Kanzow (2002, S. 255)), die *Armijo-Schrittweite* für die vorliegende Situation folgendermaßen zu spezifizieren:

- Seien $x \in \mathbb{R}^n$ und die symmetrische, positiv definite Matrix $H \in \mathbb{R}^{n\times n}$ gegeben. Das quadratische Programm $(\mathrm{P}_{(x,H)})$ in Lemma 2.1 sei zulässig, $p^* \neq 0$ die Lösung, (u_+, v_+) Lagrange-Multiplikatoren zu p^* und $\sigma \geq \max(\|u_+\|_\infty, \|v_+\|_\infty)$. Seien $\alpha \in (0, \frac{1}{2})$ und $\rho \in (0, 1)$ vorgegeben. Dann heißt $t := \rho^j$, wobei j die kleinste nichtnegative ganze Zahl mit

  $$\Psi_\sigma(x + \rho^j p^*) \leq \Psi_\sigma(x) - \alpha \rho^j (p^*)^T H p^*$$

  ist, die *Armijo-Schrittweite* für Ψ_σ in x in Richtung p^*. Eine Realisierung der Armijo-Schrittweite kann erfolgen durch:

  $t := 1$

  Solange $\Psi_\sigma(x + tp^*) > \Psi_\sigma(x) - \alpha t (p^*)^T H p^*$:

  $\quad t := \rho t$

Es ist klar, dass die Armijo-Schrittweite existiert bzw. die obige Schleife nach endlich vielen Schritten verlassen wird, denn andernfalls wäre

$$\Psi_\sigma(x + \rho^j p^*) > \Psi_\sigma(x) - \alpha \rho^j (p^*)^T H p^*, \qquad j = 0, 1, \ldots,$$

mit der Nullfolge $\{\rho^j\} \subset \mathbb{R}_+$. Hieraus folgt

$$-\alpha(p^*)^T H p^* \leq \lim_{j\to\infty} \frac{\Psi_\sigma(x+\rho^j p^*) - \Psi_\sigma(x)}{\rho^j} = \Psi'_\sigma(x;p^*) \leq -(p^*)^T H p^*,$$

wobei wir bei der letzten Ungleichung die Aussage von Lemma 2.1 benutzt haben. Damit ist

$$\underbrace{(1-\alpha)}_{>0}\,\underbrace{(p^*)^T H p^*}_{>0} \leq 0,$$

ein Widerspruch. Damit ist die Existenz der Armijo-Schrittweite nachgewiesen.

Ein Schritt des durch die Armijo-Schrittweite gedämpften SQP-Verfahrens sieht daher folgendermaßen aus (siehe C. Geiger, C. Kanzow (2002, S. 255)):

- Gegeben seien (unabhängig von dem Iterationsschritt) Konstanten $\alpha \in (0, \frac{1}{2})$, $\rho \in (0,1)$ zur Bestimmung der Armijo-Schrittweite, z. B. $\alpha := 0.0001$, $\rho := 0.5$.
- Sei $(x,u,v) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ als Näherung für ein Kuhn-Tucker-Tripel (x^*,u^*,v^*) zu (P) gegeben. Sei ferner $H \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit.
- Ist (x,u,v) ein Kuhn-Tucker-Tripel, ist also $x \in M$ und

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

  dann: STOP.
- Berechne eine Lösung $p^* \in \mathbb{R}^n$ und Lagrange-Multiplikatoren $(u_+, v_+) \in \mathbb{R}^l \times \mathbb{R}^m$ des quadratischen Hilfsproblems

  $$(\mathrm{P}_{(x,H)}) \quad \begin{cases} \text{Minimiere} & \nabla f(x)^T p + \frac{1}{2} p^T H p \quad \text{unter den Nebenbedingungen} \\ & g(x) + g'(x)p \leq 0, \quad h(x) + h'(x)p = 0. \end{cases}$$

  Bestimme also $(p^*, u_+, v_+) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ mit

  $$u_+ \geq 0, \quad \nabla f(x) + Hp^* + g'(x)^T u_+ + h'(x)^T v_+ = 0, \quad (g(x) + g'(x)p^*)^T u_+ = 0$$

  sowie

  $$g(x) + g'(x)p^* \leq 0, \qquad h(x) + h'(x)p^* = 0.$$
- Falls $p^* = 0$, dann: STOP, da (x, u_+, v_+) ein Kuhn-Tucker-Tripel ist.
- Setze $\sigma := \max(\|u_+\|_\infty, \|v_+\|_\infty)$ und berechne die Armijo-Schrittweite $t := \rho^j$, wobei j die kleinste nichtnegative ganze Zahl mit

  $$\Psi_\sigma(x+\rho^j p^*) \leq \Psi_\sigma(x) - \alpha\rho^j (p^*)^T H p^*$$

  ist.
- Setze $x_+ := x + tp^*$ und wähle symmetrische und positiv definite Matrix $H_+ \in \mathbb{R}^{n\times n}$.

Nun muss noch etwas zur Wahl der Matrizen H bzw. H_+ gesagt werden. Mit

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

sei wieder die Lagrange-Funktion zu (P) bezeichnet. Man könnte

$$H := \nabla^2_{xx} L(x,u,v)$$

setzen, zumindestens wenn diese Matrix positiv definit ist. Aber selbst wenn dies der Fall ist, müssten dann zweite Ableitungen der Zielfunktion und/oder der Restriktionsabbildungen berechnet werden, was schwierig sein kann. Oft ist es besser, einen Quasi-Newton-Ansatz zu machen und sich zu überlegen, wie man aus (x,u,v) und der symmetrischen, positiv definiten Matrix $H \in \mathbb{R}^{n\times n}$ die neue symmetrische, positiv definite Matrix $H_+ \in \mathbb{R}^{n\times n}$ bestimmt. Naheliegend wäre es,

$$s := x_+ - x, \qquad y := \nabla_x L(x_+,u,v) - \nabla_x L(x,u,v)$$

zu setzen und anschließend H_+ durch den BFGS-Update

$$H_+ := H - \frac{(Hs)(Hs)^T}{s^T H s} + \frac{yy^T}{y^T s}$$

zu definieren. Da wir nicht sichern können, dass $y^T s > 0$ gilt, ist die positive Definitheit von H_+ nicht gewährleistet. Daher schlägt M. J. D. Powell (1978)[13] die folgende Vorgehensweise vor: Wie oben seien

$$s := x_+ - x, \qquad y := \nabla_x L(x_+,u,v) - \nabla_x L(x,u,v)$$

berechnet. Anschließend setze man

$$z := \theta y + (1-\theta) Hs,$$

wobei $\theta \in [0,1]$ möglichst nahe bei 1 (damit z möglichst nahe bei y) unter der Nebenbedingung $y^T z \geq 0.2 s^T H s$ gewählt wird. Dies führt auf

$$\theta := \begin{cases} 1, & \text{falls} \quad y^T s \geq 0.2 s^T H s, \\ \dfrac{0.8 s^T H s}{s^T H s - y^T s}, & \text{falls} \quad y^T s < 0.2 s^T H s. \end{cases}$$

Anschließend macht man den BFGS-Update

$$H_+ := H - \frac{(Hs)(Hs)^T}{s^T H s} + \frac{zz^T}{z^T s}$$

und ist sich durch diese Konstruktion sicher, dass $s^T z > 0$ und folglich mit H auch H_+ positiv definit ist (siehe Satz 1.10 in Abschnitt 2.1). Dies ist für $y^T s \geq 0.2 s^T H s$ klar, da dann $z = y$. Andernfalls ist nach einfacher Rechnung $z^T s = 0.2 s^T H s > 0$.

[13] M. J. D. Powell (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.

Beispiel: Zunächst wollen wir erläutern, wie man die Funktion `fmincon` aus der Optimization-Toolbox von MATLAB einsetzen kann, um eine nichtlineare Optimierungsaufgabe zu lösen. Gegeben sei die Aufgabe[14]

$$\begin{cases} \text{Minimiere} \quad f(x) := x_1x_4(x_1+x_2+x_3)+x_3 \quad \text{unter den Nebenbedingungen} \\ -x_1x_2x_3x_4+25 \le 0, \quad x_1^2+x_2^2+x_3^2+x_4^2-40=0, \quad 1 \le x_i \le 5, \quad i=1,\ldots,4. \end{cases}$$

Man schreibe eine Funktion `Ziel_Fun`, durch welche der Wert und der Gradient der Zielfunktion bestimmt wird:

```
function [f,grad_f]=Ziel_Fun(x);
f=x(1)*x(4)*(x(1)+x(2)+x(3))+x(3);
grad_f=[x(4)*(2*x(1)+x(2)+x(3));x(1)*x(4);...
                  x(1)*x(4)+1;x(1)*(x(1)+x(2)+x(3))];
```

Dann schreibe man eine Funktion `Res_Fun`, durch die der Wert der Restriktionsabbildungen und die *Transponierten* der Funktionalmatrizen berechnet werden:

```
function [g,h,J_g,J_h]=Res_Fun(x);
g=-x(1)*x(2)*x(3)*x(4)+25;
h=x(1)^2+x(2)^2+x(3)^2+x(4)^2-40;
J_g=[-x(2)*x(3)*x(4),-x(1)*x(3)*x(4),-x(1)*x(2)*x(4),-x(1)*x(2)*x(3)]';
J_h=[2*x(1),2*x(2),2*x(3),2*x(4)]';
```

Danach schreibe man ein script file `Test` mit dem Inhalt:

```
x_0=[1;5;5;1]       %Startwert wie bei Hock-Schittkowski
options = optimset('LargeScale','off');
options = optimset(options,'GradObj','on','GradConstr','on');
options = optimset(options,'Display','iter');
l=ones(4,1);u=5*l;
[x,fval,exitflag,output,lambda]=fmincon(@Ziel_Fun,x_0,...
                                    [],[],[],[],l,u,@Res_Fun,options);
```

Hierdurch erhält man Informationen über das verwendete Verfahren (SQP, Quasi-Newton, line search), die Anzahl der Iterationen und Funktionsauswertungen, die benutzten Schrittweiten und einiges mehr. Als Werte erhalten wir (nach `format long`)

$$x = \begin{pmatrix} 1.000000000000000 \\ 4.742920256889164 \\ 3.821253703267071 \\ 1.379393935718463 \end{pmatrix}, \qquad \text{fval} = 17.014017264063018,$$

was sehr gut mit den bei Hock-Schittkowski angegebenen Werten übereinstimmt. Die Multiplikatoren sind in `lambda` enthalten. Z. B. ist

$$\text{lambda.lower} = \begin{pmatrix} 1.087851411104378 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \text{lambda.upper} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

[14] Siehe W. Hock, K. Schittkowski (1981, S. 92).

Dass die Lagrange-Multiplikatoren zu den oberen box-constraints sämtlich verschwinden, ist natürlich kein Wunder, da die Lösung hier inaktiv ist. Die Multiplikatoren zu den nichtlinearen Nebenbedingungen sind

lambda.ineqnonlin = 0.552289501508602, lambda.eqnonlin = 0.161465713172384.

□

Nun wollen wir noch eine einfache MATLAB-Funktion `Min_Res` zum durch die Armijo-Schrittweite gedämpften SQP-Verfahren angeben. Zur Lösung der auftretenden quadratischen Programme benutzen wir die Funktion `quadprog` aus der Optimization-Toolbox. Es wäre nicht schwierig, die Funktion `Min_Res` so zu modifizieren, dass auch box-constraints oder sogar lineare Ungleichungs- und Gleichungsnebenbedingungen angegeben werden können. Darauf verzichten wir. Ein weiterer Schönheitsfehler der folgenden Funktion besteht darin, dass davon ausgegangen wird, dass Ungleichungs- *und* Gleichungsrestriktionen vorkommen. Ist dies nicht der Fall, so hat man entsprechende dummy-Restriktionen einzuführen.

```
function [x,u,v,iter]=Min_Res(Ziel_Fun,Res_Fun,x,u,v,max_iter,tol);
%*****************************************************************************
% Diese Funktion loest mit dem gedaempften SQP-Verfahren die Aufgabe:
%     Minimiere Ziel_Fun unter durch Res_Fun gegebene Ungleichungen
%     und Gleichungen
%*****************************************************************************
%Input-Parameter:
%        Ziel_Fun          [f,grad_f]=Ziel_Fun(x) geben Wert und Gradient
%                          (Spaltenvektor!) der Zielfunktion
%        Res_Fun           [g,h,J_g,J_h]=Res_Fun(x) geben Vektor der
%                          Ungleichungs- und Gleichungsrestriktionen sowie
%                          ihrer Jacobi-Matrizen
%        x,u,v             Startwerte
%        max_iter          maximale Zahl der Iterationen
%        tol               Toleranz. Es wird abgebrochen, wenn
%                          norm_inf (nabla_x L(x,u,v),h(x),min(-g(x),u))<=tol
%Output-Parameter:
%       x,u,v              Loesung und zugehoerige Lagrange-Multiplikatoren
%       iter               Anzahl der benoetigten Iterationen
%*****************************************************************************
options=optimset('Display','off','LargeScale','off');
alpha=0.0001;rho=0.5;                          %Parameter fuer Armijo-Schrittweite
[f,grad_f]=feval(Ziel_Fun,x);[g,h,J_g,J_h]=feval(Res_Fun,x);
H=eye(length(x));iter=0;Phi=[grad_f+J_g'*u+J_h'*v;h;min(-g,u)];
while (norm(Phi,inf)>tol)&(iter<max_iter)
     [p_stern,fval,exitflag,output,lambda]=quadprog(H,grad_f,J_g,-g,...
                                                    J_h,-h,[],[],[],options);
     u_p=lambda.ineqlin;v_p=lambda.eqlin;sigma=max(norm(u_p,inf),norm(v_p,inf));
     Psi_strich=-p_stern'*H*p_stern;
     t=1;
     Psi=f+sigma*(sum(max(g,zeros(size(g))))+norm(h,1));
     f_t=feval(Ziel_Fun,x+p_stern);[g_t,h_t]=feval(Res_Fun,x+p_stern);
     Psi_t=f_t+sigma*(sum(max(g_t,zeros(size(g))))+norm(h_t,1));
     while (Psi_t>Psi+alpha*t*Psi_strich)
```

```
          t=rho*t;
          f_t=feval(Ziel_Fun,x+t*p_stern);[g_t,h_t]=feval(Res_Fun,x+t*p_stern);
          Psi_t=f_t+sigma*(sum(max(g_t,zeros(size(g))))+norm(h_t,1));
     end;
     x_p=x+t*p_stern;
     [f_p,grad_f_p]=feval(Ziel_Fun,x_p);[g_p,h_p,J_g_p,J_h_p]=feval(Res_Fun,x_p);
     s=x_p-x; y=grad_f_p-grad_f+(J_g_p-J_g)'*u+(J_h_p-J_h)'*v;
     y_s=y'*s; H_s=H*s; s_H_s=s'*H_s;
     if y_s>=0.2*s_H_s
          theta=1;
     else
          theta=0.8*s_H_s/(s_H_s-y_s);
     end;
     z=theta*y+(1-theta)*H_s;
     H=H-(H_s)*(H_s)'/s_H_s+z*z'/(z'*s);
     f=f_p; grad_f=grad_f_p;g=g_p;J_g=J_g_p;h=h_p;J_h=J_h_p;x=x_p;u=u_p;v=v_p;
     Phi=[grad_f+J_g'*u+J_h'*v;h;min(-g,u)];
     iter=iter+1;
end;
```

Beispiel: Wir wenden die obige MATLAB-Funktion `Min_Res` auf ein Beispiel an, das man im Tutorial der Optimization-Toolbox findet. Gegeben sei die Aufgabe

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad f(x) := e^{x_1}(4x_1^2 + 2x_2^2 + 4x_1x_2 + 2x_2 + 1) \quad \text{auf} \\ M := \{x \in \mathbb{R}^2 : g(x) := -x_1x_2 - 10 \leq 0,\ h(x) := x_1^2 + x_2 - 1 = 0\}. \end{cases}$$

Die Zielfunktion und die Restriktionsabbildungen mitsamt ihrer Gradienten legen wir in Files `Ziel_Fun.m` sowie `Res_Fun.m` mit dem Inhalt

```
function [f,grad_f]=Ziel_Fun(x);
f=exp(x(1))*(4*x(1)^2+2*x(2)^2+4*x(1)*x(2)+2*x(2)+1);
grad_f=[f+exp(x(1))*(8*x(1)+4*x(2));exp(x(1))*(4*x(2)+4*x(1)+2)];
```

sowie

```
function [g,h,J_g,J_h]=Res_Fun(x);
g=-x(1)*x(2)-10;h=x(1)^2+x(2)-1;
J_g=[-x(2),-x(1)];J_h=[2*x(1),1];
```

ab. Wir starten mit $x = (1,1)^T$, $u = 0$, $v = 0$. Mit dem Aufruf

```
x=[1;1];u=[0];v=[0];
max_iter=20;tol=1e-10;
[x,u,v,iter]=Min_Res(@Ziel_Fun,@Res_Fun,x,u,v,max_iter,tol);
```

bestimmen wir die gesuchte Lösung. In der folgenden Tabelle geben wir auch noch Zwischenresultate an:

k	x_k		u_k	v_k	t_k	$\|\Phi_k\|_\infty$
0	1.0000000	1.0000000	0	0	0.2500000	67.9570465
1	0.2204295	2.3091409	0	−32.4193802	1.0000000	22.4987755
2	0.4762689	0.8386217	0	−0.9116790	1.0000000	21.2753639
3	0.4573773	0.7911628	0	−8.5987663	1.0000000	12.5817680
4	0.1517658	1.0703655	0	−6.2013373	1.0000000	11.6997957
5	−0.1639282	1.0727903	0	1.3196247	1.0000000	6.2144308
6	−0.5515335	0.8460487	0	1.7289487	1.0000000	3.5597205
7	−0.6221060	0.6179647	0	−0.7671815	1.0000000	1.2256302
8	−0.7001541	0.5158758	0	−0.5412115	1.0000000	0.5306457
9	−0.7466103	0.4447312	0	−0.3397169	1.0000000	0.0358969
10	−0.7525138	0.4337578	0	−0.3391687	1.0000000	0.0024259
11	−0.7528743	0.4331805	0	−0.3397161	1.0000000	0.0000736
12	−0.7528791	0.4331731	0	−0.3396801	1.0000000	0.0000002
13	−0.7528791	0.4331731	0	−0.3396800	1.0000000	0.0000000

Hierbei ist

$$\Phi(x,u,v) := \begin{pmatrix} \nabla f(x) + g'(x)^T u + h'(x)^T v \\ h(x) \\ \min(-g(x), u) \end{pmatrix}$$

und $\Phi_k := \Phi(x_k, u_k, v_k)$. Man beachte, dass $\Phi(x,u,v) = 0$ genau dann, wenn (x,u,v) ein Kuhn-Tucker-Tripel ist. Daher brechen wir ab, wenn die Maximumnorm von Φ in einer aktuellen Näherung kleiner oder gleich einer vorgegebenen Toleranz ist. □

Beispiel: Wir wenden die Funktion `Min_Res` auf das Beispiel auf Seite 213 an. Diesmal können wir die box-constraints nicht getrennt behandeln, da wir dies bequemlichkeitshalber in der Funktion `Min_Res` nicht vorgesehen haben.

Mit den Startwerten $x_0 = (1,5,5,1)^T$ wie oben, $u_0 = 0$ und $v_0 = 0$ sowie `tol=1e-10` erhalten wir die folgenden Werte:

k	x_k				t_k	$\|\Phi_k\|_\infty$
0	1.0000000	5.0000000	5.0000000	1.0000000	1.0000000	12.0000000
1	1.0000000	4.8750000	3.8750000	1.2500000	1.0000000	2.0774128
2	1.0000000	4.9013491	3.6204784	1.3987553	1.0000000	0.5942721
3	1.0000000	4.7557416	3.8145182	1.3754168	1.0000000	0.3299005
4	1.0000000	4.7424445	3.8219166	1.3792822	1.0000000	0.0084346
5	1.0000000	4.7429881	3.8211651	1.3794061	1.0000000	0.0002799
6	1.0000000	4.7429996	3.8211501	1.3794082	1.0000000	0.0000002
7	1.0000000	4.7429996	3.8211501	1.3794082	1.0000000	0.0000000

Mit den Startwerten $x_0 = (3, 5, 5, 3)^T$ erhalten wir:

k	x_k				t_k	$\|\Phi_k\|_\infty$
0	3.0000000	5.0000000	5.0000000	3.0000000	1.0000000	48.0000000
1	1.0000000	4.5625000	3.5625000	3.4583333	1.0000000	16.1962147
2	1.0000000	4.9889588	2.0914853	3.4759228	1.0000000	6.0013642
3	1.0000000	5.0000000	1.7429260	3.3323328	1.0000000	3.3188188
4	1.0000000	5.0000000	1.6939994	3.3365817	0.2500000	0.3603283
5	1.0000000	5.0000000	1.6703137	3.3485167	0.0625000	0.7445744
6	1.0000000	5.0000000	1.6555928	3.3558364	0.1250000	1.0315189
7	1.0000000	5.0000000	1.6282573	3.3692734	0.1250000	0.7554409
8	1.0000000	5.0000000	1.6047682	3.3805652	0.2500000	0.1237112
9	1.0000000	5.0000000	1.5642729	3.3996589	0.5000000	1.5908016
10	1.0000000	5.0000000	1.5051421	3.4265261	1.0000000	12.1421251
11	1.0000000	4.9986854	1.4481860	3.4525094	1.0000000	22.6767235
12	1.0000000	5.0000000	1.4494874	3.4494925	1.0000000	6.9272494
13	1.0000000	5.0000000	1.4494897	3.4494898	1.0000000	0.0019848
14	1.0000000	5.0000000	1.4494897	3.4494898	1.0000000	0.0000000
15	1.0000000	5.0000000	1.4494897	3.4494898	1.0000000	0.0000000

Offenbar liegt hier Konvergenz gegen eine weitere stationäre Lösung (mit einem größeren Zielfunktionswert) vor. Wendet man dagegen das *ungedämpfte* SQP-Verfahren an, so erhält man Konvergenz gegen $x^* = (1, 4.7429996, 3.8211501, 1.3794082)^T$. □

Beispiel: In diesem Beispiel wollen wir den sogenannten *Maratos-Effekt* demonstrieren. Gegeben sei die Aufgabe (siehe auch C. Geiger, C. Kanzow (2002, S. 259) und J. Nocedal, S. J. Wright (1999, S. 567))

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := 2(x_1^2 + x_2^2 - 1) - x_1 \quad \text{auf} \\ \quad M := \{x \in \mathbb{R}^2 : h(x) := x_1^2 + x_2^2 - 1 = 0\}. \end{cases}$$

Die Lösung ist offenbar $x^* = (1, 0)^T$ mit zugehörigem Lagrange-Multiplikator $y^* = -\frac{3}{2}$. Wir rufen die Funktion `Min_Res` mit den sehr guten Startwerten $x_0 = (0.995, 0.005)^T$, $y_0 = -1.495$ auf[15]. Als erste Iterationen erhält man die folgenden Werte:

k	x_k		y_k	t_k	$\|\Phi_k\|_\infty$
0	0.9950000	0.0050000	−1.4950000	0.5000000	0.0099500
1	0.9975126	0.0025001	−1.5000126	0.5000000	0.0049624
2	0.9987594	0.0012625	−1.5000156	0.5000000	0.0024781
3	0.9993805	0.0006312	−1.5000007	0.5000000	0.0012383
4	0.9996904	0.0003156	−1.5000002	0.5000000	0.0006189
5	0.9998453	0.0001578	−1.5000000	0.5000000	0.0003094

Hierbei ist

$$\Phi_k := \begin{pmatrix} \nabla f(x_k) + h'(x_k)^T y_k \\ h(x_k) \end{pmatrix}.$$

[15] In der Funktion zur Beschreibung der Restriktionen fügen wir eine triviale Ungleichungsrestriktion ein.

Trotz der sehr guten Startwerte hat man nur lineare Konvergenz. Es fällt ferner auf, dass die Schrittweite $t = 1$ nicht akzeptiert wird. Genau das ist das Problem. Denn wenden wir mit denselben Startwerten das ungedämpfte SQP-Verfahren an, erzwingen wir also die Schrittweite $t = 1$, so erhalten wir (diesmal benutzen wir `format long`):

k	x_k		y_k	$\|\Phi_k\|_\infty$
0	0.995000000000000	0.005000000000000	−1.495000000000000	0.009950000000000
1	1.000025124993687	0.000000126256250	−1.500012625624968	0.000050250618655
2	1.000000000315648	−0.000000124374969	−1.499999937657844	0.000000124999961
3	1.000000000000008	−0.000000000624853	−1.499999999996876	0.000000000624853
4	1.000000000000000	−0.000000000000000	−1.500000000000000	0.000000000000000

Das ungedämpfte SQP-Verfahren liefert also bei diesem Beispiel wesentlich bessere Ergebnisse als das durch die Armijo-Schrittweite gedämpfte SQP-Verfahren. Dies kann man folgendermaßen verstehen. Es ist

$$\nabla^2_{xx} L(x^*, y^*) = \nabla^2 f(x^*) + y^* \nabla^2 h(x^*) = 4I - \tfrac{3}{2} 2I = I.$$

Bei gegebenem *zulässigem* $x \in \mathbb{R}^2$ sei $p^* \in \mathbb{R}^2$ die Lösung von

$$(\mathrm{P}_{(x,H)} \quad \text{Minimiere} \quad \nabla f(x)^T p + \tfrac{1}{2}\|p\|^2 \quad \text{unter der Nebenbedingung} \quad \nabla h(x)^T p = 0.$$

Dann ist $f(x + p^*) = f(x) + \|p^*\|^2$ und $h(x + p^*) = \|p^*\|^2$, folglich $\Psi_\sigma(x + p^*) > \Psi_\sigma(x)$ für jedes zulässige $x \neq x^*$, die Schrittweite 1 wird also nicht akzeptiert. □

Methoden zur Vermeidung des Maratos-Effektes werden bei C. Geiger, C. Kanzow (2002, S. 260 ff.) und J. Nocedal, S. J. Wright (1999, S. 569 ff.) geschildert. Im wesentlichen gibt es zwei Ansätze. Beim ersten werden Informationen zweiter Ordnung auch für die Nebenbedingungen berücksichtigt, beim zweiten wird nicht mehr darauf bestanden, dass die merit function von Schritt zu Schritt monoton fällt. Wir wollen nur auf den ersten Ansatz etwas genauer eingehen und folgen dabei M. Fukushima (1986)[16]. Gegeben sei die durch Gleichungen und Ungleichungen restringierte nichtlineare Optimierungsaufgabe (P).

Wie im gedämpften SQP-Verfahren sei ein Paar $(x, H) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}$ gegeben, wobei die symmetrische, positiv definite Matrix H als Approximation an die Hessesche $\nabla^2_{xx} L(x^*, u^*, v^*)$ der Lagrange-Funktion zu verstehen ist und (x^*, u^*, v^*) das zu approximierende Kuhn-Tucker-Tripel ist. Wir nehmen an, das quadratische Hilfsproblem

$$(\mathrm{P}_{(x,H)}) \qquad \begin{cases} \text{Minimiere} & \nabla f(x)^T p + \frac{1}{2} p^T H p \quad \text{unter den Nebenbedingungen} \\ & g(x) + g'(x)p \le 0, \quad h(x) + h'(x)p = 0 \end{cases}$$

sei zulässig, $p^* \in \mathbb{R}^n$ sei die dann eindeutige Lösung und $(u_+, v_+) \in \mathbb{R}^l \times \mathbb{R}^m$ zugehörige Lagrange-Multiplikatoren. Berücksichtigen wir zweite Ableitungen der Zielfunktion f sowie der Restriktionabbildungen g und h, so erhalten wir eine Optimierungsaufgabe

[16] M. Fukushima (1986) A successive quadratic programming algorithm with global and superlinear convergence propertie. Mathematical Programming 35, 253–264.

mit einer quadratischen Zielfunktion und quadratischen Ungleichungs- und Gleichungsrestriktionen, nämlich

$$(\mathrm{P}_x^2) \qquad \begin{cases} \text{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p \quad \text{unter den Nebenbedingungen} \\ g_i(x) + \nabla g_i(x)^T p + \frac{1}{2} p^T \nabla^2 g_i(x) p \;\leq\; 0, \qquad i = 1, \ldots, l, \\ h_j(x) + \nabla h_j(x)^T p + \frac{1}{2} p^T \nabla^2 h_j(x) p \;=\; 0, \qquad j = 1, \ldots, m. \end{cases}$$

Wegen der quadratischen Restriktionen und der zweiten Ableitungen ist diese Aufgabe wesentlich schwieriger als die Aufgabe $(\mathrm{P}_{(x,H)})$. Wir überlegen uns, welche Veränderungen man vornehmen muss, um auf ein gewöhnliches quadratisches Programm mit bekannten Daten zu kommen.

Ist p eine lokale (insbesondere zulässige) Lösung von (P_x^2) und ist eine entsprechende Constraint Qualification erfüllt, so existieren $(u, v) \in \mathbb{R}^l \times \mathbb{R}^m$ derart, dass die Kuhn-Tucker-Bedingungen erfüllt sind, also

$$u \geq 0$$

ist, sowie die Lagrangesche Multiplikatorenregel

$$\nabla f(x) + \nabla^2 f(x) p + \sum_{i=1}^{l} u_i [\nabla g_i(x) + \nabla^2 g_i(x) p] + \sum_{j=1}^{m} v_j [\nabla h_j(x) + \nabla^2 h_j(x) p] = 0$$

und die Gleichgewichtsbedingung

$$u_i [g_i(x) + \nabla g_i(x)^T p + \tfrac{1}{2} p^T \nabla^2 g_i(x) p] = 0, \qquad i = 1, \ldots, l,$$

gelten. In diesen Kuhn-Tucker-Bedingungen ersetzen wir unbekannte Terme (es sind weder die Lösung p von (P_x^2) noch die zugehörigen Lagrange-Multiplikatoren (u, v) bekannt, außerdem ist man nicht bereit, zweite Ableitungen der Zielfunktion bzw. der Restriktionsabbildungen zu berechnen) durch bekannte (die Lösung p^* von $(\mathrm{P}_{(x,H)})$ und die zugehörigen Lagrange-Multiplikatoren (u_+, v_+)) um auf ein System von Gleichungen und Ungleichungen zu kommen, das man als Kuhn-Tucker-Bedingungen eines quadratischen Programms entlarvt, das nur bekannte Größen enthält. Die Lagrangesche Multiplikatorenregel schreiben wir umständlicher in der Form

$$\begin{aligned} & \nabla f(x) - \tfrac{1}{2} \sum_{i=1}^{l} u_i \nabla^2 g_i(x) p - \tfrac{1}{2} \sum_{j=1}^{m} v_j \nabla^2 h_j(x) p \\ & \quad + \underbrace{\Big(\nabla^2 f(x) + \sum_{i=1}^{l} u_i \nabla^2 g_i(x) + \sum_{j=1}^{m} v_j \nabla^2 h_j(x) \Big)}_{= \nabla^2_{xx} L(x,u,v) \approx H} p \\ & \quad + \sum_{i=1}^{l} u_i [\nabla g_i(x) + \tfrac{1}{2} \nabla^2 g_i(x) p] + \sum_{j=1}^{m} v_j [\nabla h_j(x) + \tfrac{1}{2} \nabla^2 h_j(x) p] \\ = \;& 0. \end{aligned}$$

In der ersten Zeile ersetze man u_i durch $(u_+)_i$, v_j durch $(v_+)_j$, ferner $\nabla^2 g_i(x)p$ durch $\nabla g_i(x+p^*) - \nabla g_i(x)$ sowie $\nabla^2 h_j(x)p$ durch $\nabla h_j(x+p^*) - \nabla h_j(x)$. Insgesamt wird also die erste Zeile ersetzt durch

$$\begin{aligned} q &:= \nabla f(x) - \tfrac{1}{2}\sum_{i=1}^{l}(u_+)_i[\nabla g_i(x+p^*) - \nabla g_i(x)] - \tfrac{1}{2}\sum_{j=1}^{m}(v_+)_j[\nabla h_j(x+p^*) - \nabla h_j(x)] \\ &= \nabla f(x) - \tfrac{1}{2}[g'(x+p^*) - g'(x)]^T u_+ - \tfrac{1}{2}[h'(x+p^*) - h'(x)]^T v_+. \end{aligned}$$

Auch in der dritten Zeile ersetze man $\nabla^2 g_i(x)p$ durch $\nabla g_i(x+p^*) - \nabla g_i(x)$ und entsprechend $\nabla^2 h_j(x)p$ durch $\nabla h_j(x+p^*) - \nabla h_j(x)$. Die veränderte Lagrangesche Multiplikatorenregel (für ein noch nicht definiertes quadratisches Programm) lautet

$$\begin{aligned} 0 &= q + Hp + \tfrac{1}{2}\sum_{i=1}^{l} u_i[\nabla g_i(x) + \nabla g_i(x+p^*)] + \tfrac{1}{2}\sum_{j=1}^{m} v_j[\nabla h_j(x) + \nabla h_j(x+p^*)] \\ &= q + Hp + \tfrac{1}{2}[g'(x) + g'(x+p^*)]^T u + \tfrac{1}{2}[h'(x) + h'(x+p^*)]^T v. \end{aligned}$$

Dies ist die Lagrangesche Multiplikatorenregel zu dem quadratischen Programm

$$(\mathrm{P}^2_{(x,H)}) \quad \begin{cases} \text{Minimiere} \quad q^T p + \tfrac{1}{2}p^T H p \quad \text{unter den Nebenbedingungen} \\ g(x) + \tfrac{1}{2}[g'(x+p^*) + g'(x)]p \le 0, \qquad h(x) + \tfrac{1}{2}[h'(x+p^*) + h'(x)]p = 0. \end{cases}$$

Wir nehmen an, $(\mathrm{P}^2_{(x,H)})$ sei zulässig und besitze daher eine eindeutige Lösung $\hat{p}$. Wie M. Fukushima (1986) setzen wir anschließend

$$x(t) := x + tp^* + t^2(\hat{p} - p^*).$$

Wählt man $\sigma \ge \max(\|u_+\|_\infty, \|v_+\|_\infty)$, so ist (siehe Lemma 2.1)

$$\lim_{t\to 0+} \frac{\Psi_\sigma(x(t)) - \Psi_\sigma(x)}{t} = \Psi'_\sigma(x; p^*) \le -(p^*)^T H p^*.$$

Bei vorgegebenen $\alpha \in (0, \frac{1}{2})$ (z. B. $\alpha := 0.0001$) und $\rho \in (0,1)$ bestimmen wir daher die Armijo-Schrittweite t so, dass $t = \rho^j$ und j die kleinste nichtnegative ganze Zahl mit

$$\Psi_\sigma(x(\rho^j)) \le \Psi_\sigma(x) - \alpha\rho^j (p^*)^T H p^*$$

ist[17]. Wir werden am letzten Beispiel erproben, ob die folgende Modifikation des gedämpften SQP-Verfahrens den Maratos-Effekt vermeidet. Wir schildern einen Schritt.

- Gegeben seien (unabhängig von dem Iterationsschritt) Konstanten $\alpha \in (0, \frac{1}{2})$, $\rho \in (0,1)$ zur Bestimmung der Armijo-Schrittweite, z. B. $\alpha := 0.0001$, $\rho := 0.5$.
- Sei $(x, u, v) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ als Näherung für ein Kuhn-Tucker-Tripel (x^*, u^*, v^*) zu (P) gegeben. Sei ferner $H \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit.

[17] Bei M. Fukushima (1986) wird eine etwas andere Abschätzung für die Richtungsableitung $\Psi'_\sigma(x; p^*)$ benutzt, was uns hier aber nicht kümmern soll.

- Ist (x, u, v) ein Kuhn-Tucker-Tripel, ist also $x \in M$ und

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

dann: STOP.

- Berechne eine Lösung $p^* \in \mathbb{R}^n$ und Lagrange-Multiplikatoren $(u_+, v_+) \in \mathbb{R}^l \times \mathbb{R}^m$ des quadratischen Hilfsproblems

$$(\mathrm{P}_{(x,H)}) \quad \begin{cases} \text{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T H p \quad \text{unter den Nebenbedingungen} \\ g(x) + g'(x)p \leq 0, \quad h(x) + h'(x)p = 0. \end{cases}$$

Bestimme also $(p^*, u_+, v_+) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ mit

$$u_+ \geq 0, \quad \nabla f(x) + Hp^* + g'(x)^T u_+ + h'(x)^T v_+ = 0, \quad [g(x) + g'(x)p^*]^T u_+ = 0$$

sowie

$$g(x) + g'(x)p^* \leq 0, \qquad h(x) + h'(x)p^* = 0.$$

- Falls $p^* = 0$, dann: STOP, da (x, u_+, v_+) ein Kuhn-Tucker-Tripel ist.
- Setze $\sigma := \max(\|u_+\|_\infty, \|v_+\|_\infty)$.
- Falls $\Psi_\sigma(x + p^*) \leq \Psi_\sigma(x) - \alpha(p^*)^T H p^*$:
  - Setze $x_+ := x + p^*$.
- Andernfalls:
  - Berechne

$$q := \nabla f(x) - \tfrac{1}{2}[g'(x + p^*) - g'(x)]^T u_+ - \tfrac{1}{2}[h'(x + p^*) - h'(x)]^T v_+$$

und berechne die Lösung $\hat{p}$ von

$$(\mathrm{P}^2_{(x,H)}) \quad \begin{cases} \text{Minimiere} \quad q^T p + \frac{1}{2} p^T H p \quad \text{unter den Nebenbedingungen} \\ g(x) + \frac{1}{2}[g'(x + p^*) + g'(x)]p \leq 0, \\ h(x) + \frac{1}{2}[h'(x + p^*) + h'(x)]p = 0. \end{cases}$$

  - Bestimme $t = \rho^j$, wobei j die kleinste nichtnegative ganze Zahl mit

$$\Psi_\sigma(x + \rho^j p^* + \rho^{2j}(\hat{p} - p^*)) \leq \Psi_\sigma(x) - \alpha\rho^j (p^*)^T H p^*$$

ist und setze $x_+ := x + tp^* + t^2(\hat{p} - p)$.

- Bestimme symmetrischen und positiv definiten Update H_+, z. B. genau so, wie es oben geschildert wurde.

Beispiel: Wir wenden das obige modifizierte gedämpfte SQP-Verfahren auf das letzte Beispiel an, bei dem der Maratos-Effekt auftrat, also auf

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := 2(x_1^2 + x_2^2 - 1) - x_1 \quad \text{auf} \\ \quad M := \{x \in \mathbb{R}^2 : h(x) := x_1^2 + x_2^2 - 1 = 0\}. \end{cases}$$

Zunächst nehmen wir die sehr guten Startwerte $x_0 = (0.995, 0.005)^T$, $y_0 = -1.495$ wie oben. Das Resultat ist überzeugend. Es wird von Anfang an die Schrittweite 1 genommen und man erhält die folgenden Werte:

k	x_k		y_k	$\|\Phi_k\|_\infty$
0	0.995000000000000	0.005000000000000	−1.495000000000000	0.009950000000000
1	0.999999937189094	0.000000063128121	−1.500012625624968	0.000025314059257
2	1.000000000000000	0.000000000626562	−1.500000000313283	0.000000000626567
3	1.000000000000000	−0.000000000000008	−1.499999999999996	0.000000000000008

Wir haben unsere Implementation des obigen Verfahrens auch noch mit wesentlich schlechteren Startwerten gestestet. So erhalten wir z. B. die folgenden Werte, wobei durch das Verfahren stets die Schrittweite 1 gewählt wurde,

k	x_k		y_k	$\|\Phi_k\|_\infty$
0	0	1.000000000000000	0	4.000000000000000
1	0.800000000000000	0.600000000000000	−2.000000000000000	1.000000000000000
2	0.896152623211447	0.445087440381558	−1.705882352941176	0.472851398110913
3	0.960866027217312	−0.197108071093728	−1.624882063794531	0.279123837800588
4	0.995805546508375	0.101301236492962	−1.549365277926211	0.102510888619322
5	0.999624158508554	−0.029763034960412	−1.510350316495995	0.029146921296969
6	0.999995024645091	0.003213653483585	−1.500602971831193	0.003209777998533
7	0.999999997745248	−0.000067718245309	−1.500010272694933	0.000067716854012
8	0.999999999999997	0.000000081711973	−1.500000017557886	0.000000081711970
9	1.000000000000000	−0.000000000001679	−1.500000000000417	0.000000000001679
10	1.000000000000000	0.000000000000000	−1.500000000000000	0.000000000000000

□

Kapitel 8

Lösungen zu den Aufgaben

8.1 Aufgaben zu Kapitel 1

1. Gegeben sei die durch
$$f(x) := 2x_1^2 + x_1x_2^2 + x_2^2$$
definierte Funktion $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$. Man bestimme alle stationären Punkte von f und entscheide, welche davon lokale oder gar globale Extrema von f sind.

   Lösung: Es ist
$$\nabla f(x) = \begin{pmatrix} 4x_1 + x_2^2 \\ 2x_1x_2 + 2x_2 \end{pmatrix}.$$
Stationäre Punkte von f sind Punkte, in denen der Gradient von f verschwindet, also Lösungen des nichtlinearen Gleichungssystems
$$4x_1 + x_2^2 = 0, \quad 2x_1x_2 + 2x_2 = 0.$$
Aus der zweiten Gleichung erhält man, dass für einen stationären Punkt (x_1, x_2) die zweite Komponente x_2 verschwindet oder die erste Komponente x_1 gleich -1 ist. Wegen der ersten Gleichung ist im ersten Fall notwendigerweise auch die erste Komponente x_1 gleich 0, während im zweiten Fall die zweite Komponente x_2 gleich ± 2 ist. Es gibt also drei stationäre Lösungen $(0, 0)$, $(-1, 2)$ und $(-1, -2)$. Die Hessesche von f ist gegeben durch
$$\nabla^2 f(x) = \begin{pmatrix} 4 & 2x_2 \\ 2x_2 & 2x_1 + 2 \end{pmatrix}.$$
Die Matrix
$$\nabla^2 f(0, 0) = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$$
ist positiv definit, während die Matrix
$$\nabla^2 f(-1, \pm 2) = \begin{pmatrix} 4 & \pm 4 \\ \pm 4 & 0 \end{pmatrix}$$

die Eigenwerte $\lambda_{1,2} = 2 \pm \sqrt{20}$ besitzt, also einen positiven und einen negativen Eigenwert, also weder positiv noch negativ semidefinit ist. In $(-1, \pm 2)$ ist also kein lokales Extremum von f. In $(0, 0)$ ist *kein* globales Minimum von f, da z. B. $f(-2, x_2) = 8 - x_2^2 \to -\infty$ mit $x_2 \to \pm\infty$. In Abbildung 8.1 geben wir links einen Flächen- und rechts einen Höhenlinienplot über dem Quadrat $[-3, 3] \times [-3, 3]$ an.

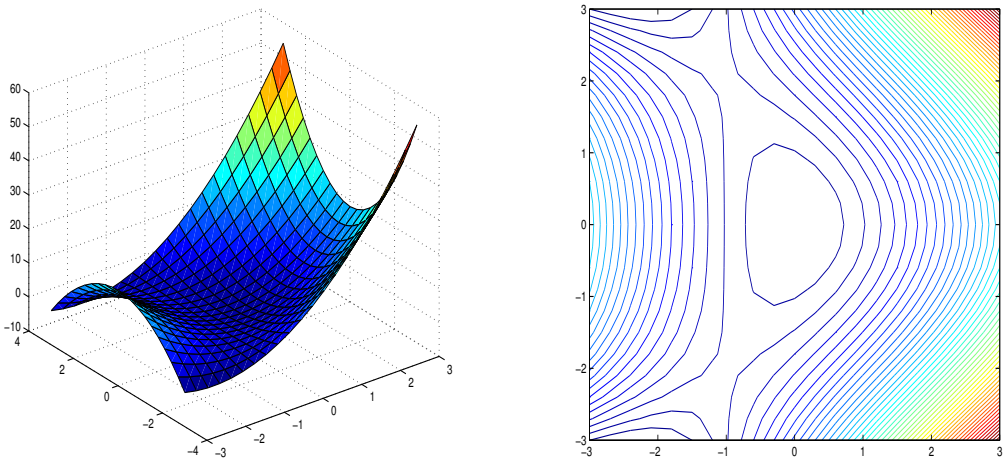


Abbildung 8.1: Flächenplot, Höhenlinienplot zu $f(x) := 2x_1^2 + x_1x_2^2 + x_2^2$

2. Gegeben sei die durch

$$f(x) := 2x_1^3 - 3x_1^2 - 6x_1x_2(x_1 - x_2 - 1)$$

definierte Funktion $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$.

   (a) Man bestimme alle stationären Punkte von f und entscheide, welche lokale Minima bzw. Maxima sind.

   (b) Über dem Quadrat $[-2, 2] \times [-2, 2]$ mache man mit MATLAB einen Flächen- und einen Höhenlinienplot von f.

Lösung:

   (a) Der Gradient von f an der Stelle x ist gegeben durch

$$\begin{aligned} \nabla f(x) &= \begin{pmatrix} 6x_1^2 - 6x_1 - 12x_1x_2 + 6x_2^2 + 6x_2 \\ -6x_1^2 + 12x_1x_2 + 6x_1 \end{pmatrix} \\ &= 6 \begin{pmatrix} (x_1 - x_2)(x_1 - x_2 - 1) \\ x_1(1 - x_1 + 2x_2) \end{pmatrix}. \end{aligned}$$

   In einem stationären Punkt x ist notwendig (betrachte die zweite Komponente des Gradienten) $x_1 = 0$ und/oder $1 - x_1 + 2x_2 = 0$. Ist $x_1 = 0$, so ist

(betrachte erste Komponente von $\nabla f(x)$) notwendig $x_2 = 0$ oder $x_2 = -1$. Daher sind $(0,0)$ und $(0,-1)$ genau die stationären Punkte von f mit verschwindender erster Komponente. Ist x ein stationärer Punkt mit $x_1 \neq 0$, so ist $x_1 = 1 + 2x_2$ und daher (betrachte erste Komponente von $\nabla f(x)$) notwendig $x_2 = 0$ oder $x_2 = -1$. Daher sind $(0,0)$, $(0,-1)$ und $(1,0)$, $(-1,-1)$ die einzigen stationären Punkte von f. Dies hätten wir z. B. mit dem Computeralgebrasystem Maple auch als Antwort auf die Eingabe

```
solve({(x_1-x_2)*(x_1-x_2-1),x_1*(1-x_1+2*x_2)},{x_1,x_2});
```

erhalten können. Als Hessesche von f in x erhält man

$$\nabla^2 f(x) = 6 \begin{pmatrix} 2(x_1 - x_2) - 1 & -2(x_1 - x_2) + 1 \\ -2(x_1 - x_2) + 1 & 2x_1 \end{pmatrix}.$$

Die Matrix $\nabla^2 f(0,0)$ besitzt die Eigenwerte $3(-1 \pm \sqrt{5})$, ist also indefinit. Der stationäre Punkt $(0,0)$ ist weder ein lokales Minimum, noch ein lokales Maximum von f. Entsprechendes gilt für den stationären Punkt $(0,-1)$, da $\nabla^2 f(0,-1) = -\nabla^2 f(0,0)$. Die Matrix $\nabla^2 f(1,0)$ besitzt die Eigenwerte $3(3 \pm \sqrt{5}$, ist also positiv definit. Hier sind die hinreichenden Bedingungen zweiter Ordnung erfüllt, in $(1,0)$ liegt also ein lokales Minimum von f. Dies ist aber natürlich kein globales Minimum, da $f(x_1, 0) \to -\infty$ mit $x_1 \to -\infty$. Dagegen besitzt $\nabla^2 f(-1,-1)$ die negativen Eigenwerte $6(-3 \pm \sqrt{5})/2$, ist also negativ definit. Daher besitzt f in $(-1,-1)$ ein lokales Maximum.

(b) Wir erhalten auf üblichem Wege die Plots in Abbildung 8.2.

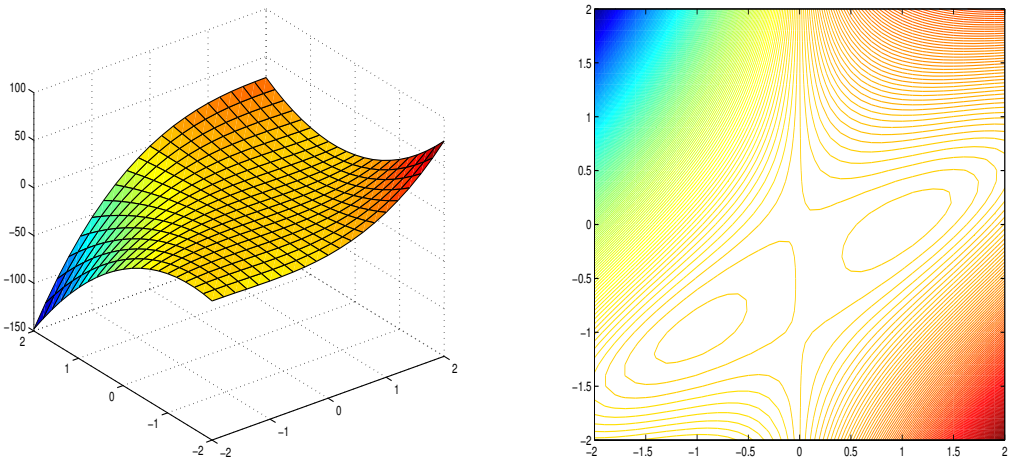


Abbildung 8.2: Flächenplot, Höhenlinienplot zu $f(x) := 2x_1^3 - 3x_1^2 - 6x_1x_2(x_1 - x_2 - 1)$

3. 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. Wie hat man r und h bei vorgegebenen positiven V, c_1 und c_2 zu bestimmen?

Lösung: Die Kosten zur Herstellung einer Dose mit dem Radius r und der Höhe h sind durch $2\pi r^2 c_1 + 2\pi r h c_2$ gegeben. Da das Volumen $V = \pi r^2 h$ vorgegeben ist, ist der Radius als Lösung von

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

zu bestimmen. Es gibt genau ein $r^* > 0$ mit $f'(r^*) = 0$, nämlich

$$r^* = \left(\frac{c_2 V}{2c_1 \pi}\right)^{1/3}$$

und dies ist das eindeutige globale Minimum von f auf $\mathbb{R}_+$, da f'' auf $\mathbb{R}_+$ offenbar positiv ist.

4. Man löse Tartaglia's Problem: Eine Strecke der Länge 8 ist so in zwei Teile zu zerlegen, dass das Produkt aus dem Produkt und der Differenz der beiden Strecken maximal ist.

   Lösung: Die bei der Teilung auftretende längere Strecke sei x, die andere Strecke hat die Länge $8 - x$. Es kommt darauf an, die Funktion

   $$g(x) := x(8 - x)(2x - 8)$$

   auf $(4, 8)$ zu maximieren. Diese haben wir in Abbildung 8.3 dargestellt. Es ist

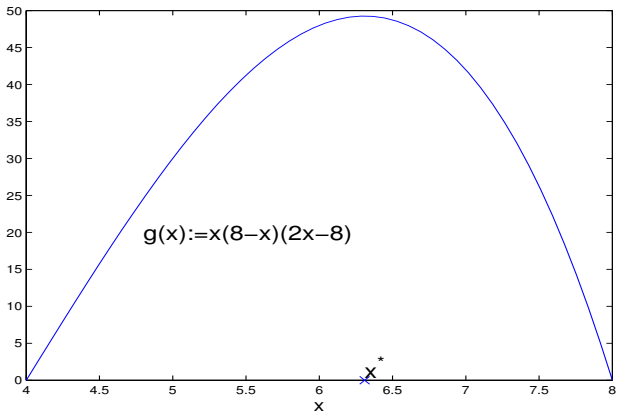


Abbildung 8.3: Tartaglia's Problem

$$g(x) = -64x + 24x^2 - 2x^3, \qquad g'(x) = -64 + 48x - 6x^2.$$

In $(4, 8)$ hat g' genau eine Nullstelle, nämlich

$$x^* := 4 + \frac{4}{3}\sqrt{3}.$$

Dies ist die optimale längere Strecke, die andere hat die Länge $8 - x^*$.

5. Sei $M \subset \mathbb{R}^n$ konvex und $f\colon M \longrightarrow \mathbb{R}$ auf M konvex. Man zeige, dass eine lokale Lösung $x^* \in M$ der konvexen Optimierungsaufgabe

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

sogar eine globale Lösung von (P) ist.

Lösung: Sei $x^* \in M$ eine lokale Lösung von (P). Dann existiert eine Umgebung U^* von x^* mit $f(x^*) \leq f(x)$ für alle $x \in U^* \cap M$. Sei $z \in M$ beliebig. Mit einem hinreichend kleinen $\lambda \in (0, 1]$ ist $x := (1-\lambda)x^* + \lambda z \in U^*$. Wegen der Konvexität von M ist $x \in M$, folglich $x \in U^* \cap M$ und daher

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

Dies impliziert $\lambda f(x^*) \leq \lambda f(z)$ und damit $f(x^*) \leq f(z)$, womit die Behauptung bewiesen ist.

6. Gegeben sei die konvexe Optimierungsaufgabe

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

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

   (a) 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.

   (b) 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: Der erste Teil der Aufgabe ist einfach einzusehen. Ist nämlich mit einem $x_0 \in M$ die Niveaumenge $L_0 := \{x \in M : f(x) \leq f(x_0)\}$ kompakt, so nimmt die stetige Funktion f auf L_0 ihr Minimum in einem $x^* \in L_0$ an. Offenbar ist $x^* \in M_{\text{opt}}$. Für den zweiten Teil wird vorausgesetzt, die Menge der (globalen) Lösungen M_{opt} von (P) sei nichtleer und kompakt. Für ein beliebiges $x_0 \in M$ ist die Niveaumenge $L_0 := \{x \in M : f(x) \leq f(x_0)\}$ wegen der Abgeschlossenheit von M und der Stetigkeit von f abgeschlossen. Zu zeigen bleibt die Beschränktheit von L_0. Angenommen, dies sei nicht der Fall. Dann gibt es eine Folge $\{x_k\}_{k\in\mathbb{N}} \subset L_0$ mit $\|x_k\| \to \infty$, wobei $\|\cdot\|$ eine beliebige Norm auf dem $\mathbb{R}^n$ sei. Da man notfalls zu einer Teilfolge übergehen kann, können wir o. B. d. A. annehmen, dass der Grenzwert $p := \lim_{k\to\infty} x_k/\|x_k\|$ existiert. Mit einem beliebigen $x^* \in M_{\text{opt}}$ ist dann, wie wir uns gleich überlegen werden, $x^* + tp \in M_{\text{opt}}$ für alle $t \geq 0$, was

wegen $p \neq 0$ zu einem Widerspruch zu der vorausgesetzten Kompaktheit von M_{opt} führt. Wir geben uns ein $t \geq 0$ vor und können annehmen, dass $t/\|x_k\| \in [0,1]$ für alle k. Wegen der Konvexität von M ist

$$\Big(1 - \frac{t}{\|x_k\|}\Big)x^* + \frac{t}{\|x_k\|}x_k \in M,$$

wegen der Abgeschlossenheit von M ist

$$\lim_{k\to\infty}\Big[\Big(1 - \frac{t}{\|x_k\|}\Big)x^* + \frac{t}{\|x_k\|}x_k\Big] = x^* + tp \in M.$$

Weiter ist auch $f(x^* + tp) = \min(\mathrm{P})$ und damit $x^* + tp \in M_{\text{opt}}$, wie man aus

$$\begin{aligned} f\Big(\underbrace{\Big(1 - \frac{t}{\|x_k\|}\Big) + \frac{t}{\|x_k\|}x_k}_{\to x^*+tp}\Big) &\leq \Big(1 - \frac{t}{\|x_k\|}\Big)f(x^*) + \frac{t}{\|x_k\|}f(x_k) \\ &\leq \underbrace{\Big(1 - \frac{t}{\|x_k\|}\Big)}_{\to 1}\min(\mathrm{P}) + \underbrace{\frac{t}{\|x_k\|}f(x_0)}_{\to 0} \end{aligned}$$

erkennt.

7. Sei $A \in \mathbb{R}^{m\times n}$ eine Matrix mit $\mathrm{Rang}\,(A) = n$, die n Spalten von A seien also linear unabhängig. Ferner sei $b \in \mathbb{R}^m$ gegeben. Man begründe, weshalb dann das vorzeichenbeschränkte lineare Ausgleichsproblem

$$\text{Minimiere} \quad \|Ax - b\|_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0\}$$

eindeutig lösbar ist.

Hinweis: Man betrachte das äquivalente Problem

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) := \frac{1}{2}\|Ax - b\|_2^2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : x \geq 0\}.$$

Lösung: Es ist

$$f(x) := \frac{1}{2}\|b\|_2^2 - (A^T b)^T x + \frac{1}{2}x^T A^T A x.$$

Die Matrix $Q := A^T A \in \mathbb{R}^{n\times n}$ ist symmetrisch und positiv definit. Ersteres ist offensichtlich, letzteres erkennt man an

$$x^T A^T A x = \|Ax\|_2^2 \geq 0, \qquad x^T A^T A x = 0 \iff Ax = 0 \iff x = 0,$$

wobei wir am Schluss die Rangvoraussetzung an A benutzt haben. Also ist die quadratische Funktion f strikt konvex. Da schließlich der nichtnegative Orthant nichtleer, konvex und abgeschlossen ist, folgt die Behauptung aus Bemerkungen zur Existenz und Eindeutigkeit einer Lösung am Schluss von Kapitel 1.

8.2 Aufgaben zu Kapitel 2

8.2.1 Aufgaben zu Abschnitt 2.1

1. Eine Funktion $\phi\colon [a,b] \longrightarrow \mathbb{R}$ heißt *unimodal*, wenn es genau ein $t^* \in (a,b)$ gibt mit $\phi(t^*) = \min_{t\in[a,b]} \phi(t)$, und wenn ϕ auf $[a,t^*]$ monoton fallend und auf $[t^*,b]$ monoton wachsend ist. Zur Lokalisierung des Minimums t^* der auf $[a,b]$ unimodularen Funktion ϕ betrachte man die *Methode vom goldenen Schnitt*:

   - Sei $\epsilon > 0$ (gewünschte Genauigkeit) gegeben, setze $F := (\sqrt{5}-1)/2$.
   - Berechne $\left\{ \begin{array}{lcllcl} s & := & a+(1-F)(b-a), & \phi_s & := & \phi(s), \\ t & := & a+F(b-a), & \phi_t & := & \phi(t). \end{array} \right.$
   - Solange $b-a > \epsilon$:
     - Falls $\phi_s > \phi_t$, dann:

       $$a := s, \quad s := t, \quad t := a+F(b-a), \quad \phi_s := \phi_t, \quad \phi_t := \phi(t)$$

     - Andernfalls:

       $$b := t, \quad t := s, \quad s := a+(1-F)(b-a), \quad \phi_t := \phi_s, \quad \phi_s := \phi(s).$$

   - $t^* \approx (a+b)/2$.

   Man beweise, dass dieser Algorithmus nach endlich vielen Schritten mit einem Intervall $[a,b]$ abbricht, das t^* enthält.

   Lösung: Es genügt den ersten Schritt zu betrachten. Durch s und t wird das Intervall $[a,b]$ im goldenen Schnitt geteilt. Ist $\phi(s) > \phi(t)$, so kann t^* nicht in $[a,s]$ liegen, so dass man sich bei der Minimumsuche auf das Intervall $[s,b]$ beschränken kann. Der Witz besteht nun darin, dass t einer der beiden Punkte ist, welche dieses Intervall im goldenen Schnitt teilt, in diesem ist der Funktionswert von ϕ schon bekannt. Dies motiviert die Setzungen $a := s$, $s := t$, $\phi_s := \phi_t$. Ist im zweiten Fall $\phi(s) \leq \phi(t)$, so kann man sich bei der Suche nach dem Minimum t^* auf das Intervall $[a,t]$ beschränken. In jedem Schritt wird die Intervalllänge mit F multipliziert, diese geht daher gegen Null und die Aussage ist bewiesen.

2. Man gebe eine MATLAB-Implementation der Methode des goldenen Schnittes an und erprobe sie an den Funktionen[1]

   (a) $\phi(t) := -t/(t^2+c)$ mit $c := 2$,

   (b) $\phi(t) := (t+c)^5 - 2(t+c)^4$ mit $c := 0.004$.

   Hierbei veranschauliche man sich die Funktionen durch einen Plot.

   Lösung: Zunächst geben wir Plots der beiden Funktionen an, siehe Abbildung 8.4. Nun eine MATLAB-Implementation der oben angegebenen Methode.

[1] Siehe C. Geiger, C. Kanzow (1999, S. 52).

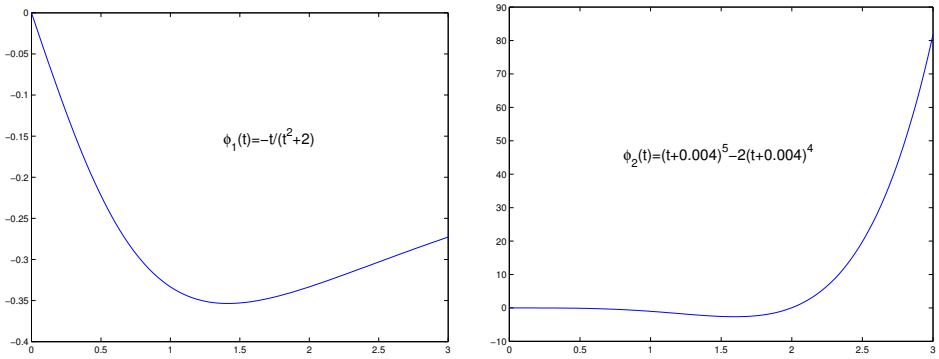


Abbildung 8.4: Zwei unimodale Funktionen

```
function t_stern=GoldenSection(fun,a,b,epsi);
%Input-Parameter:
%               fun    real valued function
%               a,b    interval [a,b], fun should be
%                      unimodal on [a,b]
%               epsi   desired accuracy
%************************************************************
F=0.5*(sqrt(5)-1);G=1-F;
s=a+G*(b-a);t=a+F*(b-a);phi_s=feval(fun,s);phi_t=feval(fun,t);
while (b-a>epsi)
   if (phi_s>phi_t)
       a=s;s=t;t=a+F*(b-a);phi_s=phi_t;phi_t=feval(fun,t);
   else
       b=t;t=s;s=a+G*(b-a);phi_t=phi_s;phi_s=feval(fun,s);
   end;
end;
t_stern=0.5*(a+b);
%************************************************************
function out=phi_1(t);
out=-t/(t^2+2);
%************************************************************
function out=phi_2(t);
out=(t+0.004)^5-2*(t+0.004);
```

Der Aufruf

```
t_stern=GoldenSection('phi_1',0,3,1e-10);
```

liefert (nach `format long g`) $t^* = 1.41421355853631$, mit der Genauigkeit `3*eps` (hierbei ist `eps` die Maschinengenauigkeit) erhält man $t^* = 1.4142135585311$. Bei der Minimierung von ϕ_2 erhält man $t^* = 0.791270726350009$ bzw. (mit der Genauigkeit `3*eps`) $t^* = 0.791270726342191$. In MATLAB gibt es die Funktion `fminbnd` (diese ersetzt die frühere Funktion `fmin` (jeweils ist nicht die Optimization Toolbox nötig). Wir können die zu minimierende Funktion auch `inline` übergeben (das hätten wir auch bei obiger Funktion `GoldenSection` machen können, es hat bei einfachen Funktionen den Vorteil, dass man nicht extra ein M-file zu schreiben braucht):

```
>> phi=inline('-t/(t^2+2)');
>>t_stern=fminbnd(phi,0,3);
```

Als Ergebnis erhalten wir $t^* = 1.41421729307232$, wobei wir darauf hingewiesen werden, dass hier nur mit einer Toleranz von `1e-4` gerechnet wurde. Diese kann man erhöhen, indem man z. B.

```
>> options=optimset('TolX',1e-10);
>> t_stern=fminbnd(phi,0,3,options);
```

eingibt, das Ergebnis ist $t^* = 1.41421356247266$. Man erkennt, dass man lange nicht allen Dezimalen trauen darf.

3. Sei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ definiert durch $f(x) := c^T x + \frac{1}{2} x^T Q x$ mit $c \in \mathbb{R}^n$ und symmetrischer, positiv definiter Matrix $Q \in \mathbb{R}^{n \times n}$. Sei $x \in \mathbb{R}^n$ kein stationärer Punkt von f und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x.

   (a) Man berechne die exakte Schrittweite t^* und eine (von x und p unabhängige, möglichst große) Konstante $\theta > 0$ mit

   $$f(x) - f(x + t^* p) \geq \theta \Big(\frac{\nabla f(x)^T p}{\|p\|} \Big)^2 .$$

   (b) Zu $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ berechne man die Menge der Wolfe-Schrittweiten und zeige, dass diese ein nichtleeres Intervall ist.

Lösung:

(a) Man definiere $\phi : [0, \infty) \longrightarrow \mathbb{R}$ durch $\phi(t) := f(x + tp)$. Die Funktion $\phi(\cdot)$ besitzt auf $[0, \infty)$ ein globales Minimum, welches durch die Gleichung $\phi'(t^*) = 0$ bestimmt ist. Wegen

$$0 = \phi'(t^*) = \nabla f(x + t^* p)^T p = [c + Q(x + t^* p)]^T p = (c + Qx)^T p + t^* p^T Q p$$

ist die exakte Schrittweite durch

$$t^* := -\frac{(c + Qx)^T p}{p^T Q p} = -\frac{\nabla f(x)^T p}{p^T Q p}$$

gegeben. Nach Taylor ist

$$f(x + t^* p) = f(x) + t^* \nabla f(x)^T p + \frac{1}{2} (t^*)^2 p^T Q p$$

und daher

$$\begin{aligned} f(x) - f(x + t^* p) &= -t^* \nabla f(x)^T p - \frac{1}{2} (t^*)^2 p^T Q p \\ &= \frac{1}{2} \frac{(\nabla f(x)^T p)^2}{p^T Q p} \\ &\geq \frac{1}{2 \lambda_{\max}(Q)} \Big(\frac{\nabla f(x)^T p}{\|p\|} \Big)^2 , \end{aligned}$$

so dass $\theta := 1/(2\lambda_{\max}(Q))$ die gesuchte Konstante ist.

(b) Es ist die Menge aller positiven t mit

$$f(x+tp) \le f(x) + \alpha t \nabla f(x)^T p, \qquad \nabla f(x+tp)^T p \ge \beta \nabla f(x)^T p$$

zu bestimmen. Es ist

$$f(x) - f(x+tp) + \alpha t \nabla f(x)^T p = -(1-\alpha) t \nabla f(x)^T p - \frac{1}{2} t^2 p^T Q p.$$

Weiter ist

$$\begin{aligned} \nabla f(x+tp)^T p - \beta \nabla f(x)^T p &= (c + Q(x+tp))^T p - \beta (c+Qx)^T p \\ &= (1-\beta)(c+Qx)^T p + t p^T Q p. \end{aligned}$$

Die Menge der Wolfe-Schrittweiten ist also die Menge aller t mit

$$-(1-\beta)\frac{(c+Qx)^T p}{p^T Q p} \le t \le -2(1-\alpha)\frac{(c+Qx)^T p}{p^T Q p}.$$

Wegen $1-\beta < 1-\alpha < 2(1-\alpha)$ handelt es sich hier um ein nichtleeres Intervall.

4. Sei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ wie in Aufgabe 3.

(a) Zur Minimierung von f auf dem $\mathbb{R}^n$ betrachte man das Gradientenverfahren mit exakter Schrittweite. Ausgehend von einem Startwert $x_0 \in \mathbb{R}^n$ erzeuge dieses die Folge $\{x_k\}$. Man begründe, weshalb die Folge $\{x_k\}$ gegen das eindeutige Minimum x^* von f konvergiert.

(b) Man zeige, dass mit $g_k := \nabla f(x_k)$ die Beziehung

$$\frac{f(x_k) - f(x^*)}{f(x_k) - f(x_{k+1})} = \frac{(g_k^T Q g_k)(g_k^T Q^{-1} g_k)}{\|g_k\|^4}$$

gilt (hierbei sei $\|\cdot\|$ weiterhin die euklidische Norm). Hieraus schließe man, dass

$$f(x_{k+1}) - f(x^*) \le \Big(1 - \frac{1}{\kappa(Q)}\Big)[f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots.$$

Hierbei ist $\kappa(Q) := \|Q\|\,\|Q^{-1}\| = \lambda_{\max}(Q)/\lambda_{\min}(Q)$ die *Kondition* von Q bezüglich der Spektralnorm.

(c) Man informiere sich (etwa bei C. Geiger, C. Kanzow (1999, S. 70 ff.)), wie die letzte Abschätzung mit Hilfe der Ungleichung von Kantorowitsch verbessert werden kann.

Lösung:

(a) Die Voraussetzungen (V) (a)–(c) sind mit $\gamma := \lambda_{\max}(Q)$ erfüllt. Da f genau eine stationäre Lösung besitzt und Satz 1.3 mit $\sigma := 1$ anwendbar ist, folgt die Konvergenz der durch das Gradientenverfahren erzeugten Folge $\{x_k\}$.

(b) Mit der exakten Schrittweite $t_k^* = \|g_k\|^2/g_k^T Q g_k$ ist $x_{k+1} = x_k - t_k^* g_k$ und daher (Taylor!)

$$f(x_k) - f(x_{k+1}) = \underbrace{g_{k+1}^T(x_k - x_{k+1})}_{=0} + \frac{1}{2}(t_k^*)^2 g_k^T Q g_k = \frac{1}{2}\,\frac{\|g_k\|^4}{g_k^T Q g_k}.$$

Weiter ist (ebenfalls Taylor!)

$$\begin{aligned} f(x_k) - f(x^*) &= \underbrace{\nabla f(x^*)}_{=0}^T (x_k - x^*) + \frac{1}{2}(x_k - x^*)^T Q (x_k - x^*) \\ &= \frac{1}{2}(x_k + Q^{-1}c)^T Q (x_k + Q^{-1}c) \\ &= \frac{1}{2}(Qx_k + c)^T Q^{-1}(Qx_k + c) \\ &= \frac{1}{2} g_k^T Q^{-1} g_k. \end{aligned}$$

Hieraus erhält man die behauptete Beziehung und anschließend mit der Cauchy-Schwarzschen Ungleichung

$$\frac{f(x_k) - f(x^*)}{f(x_k) - f(x_{k+1})} = \frac{(g_k^T Q g_k)(g_k^T Q^{-1} g_k)}{\|g_k\|^4} \le \|Q\|\,\|Q^{-1}\| = \kappa(Q).$$

Daher ist

$$\begin{aligned} f(x_{k+1}) - f(x^*) &= f(x_k) - f(x^*) - [f(x_k) - f(x_{k+1})] \\ &= \Big(1 - \frac{f(x_k) - f(x_{k+1})}{f(x_k) - f(x_{k+1})}\Big)[f(x_k) - f(x^*)] \\ &\le \Big(1 - \frac{1}{\kappa(Q)}\Big)[f(x_k) - f(x^*)]. \end{aligned}$$

(c) Für eine symmetrische, positiv definite Matrix $Q \in \mathbb{R}^{n\times n}$ mit kleinstem (positivem) Eigenwert $\lambda_{\min}(Q)$ und größtem Eigenwert $\lambda_{\max}(Q)$ lautet die Ungleichung von Kantorowitsch:

$$\frac{(x^T Q x)(x^T Q^{-1} x)}{\|x\|^4} \le \frac{(\lambda_{\min}(Q) + \lambda_{\max}(Q))^2}{4\lambda_{\min}(Q)\lambda_{\max}(Q)} = \frac{(1 + \kappa(Q))^2}{4\kappa(Q)}$$

für alle $x \in \mathbb{R}^n$ mit $x \neq 0$. Die obige Abschätzung verbessert sich zu

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\le \Big(1 - \frac{4\kappa(Q)}{(1 + \kappa(Q))^2}\Big)[f(x_k) - f(x^*)] \\ &= \Big(\frac{\kappa(Q) - 1}{\kappa(Q) + 1}\Big)^2 [f(x_k) - f(x^*)]. \end{aligned}$$

5. Die Voraussetzungen von Satz 1.3 (erster Konvergenzsatz für den Modellalgorithmus) seien erfüllt. Die Zielfunktion f besitze in der (kompakten) Niveaumenge

L_0 nur endlich viele stationäre Punkte. Der Modellalgorithmus erzeuge eine Folge $\{x_k\} \subset L_0$ mit $\lim_{k\to\infty}(x_{k+1} - x_k) = 0$. Man zeige, dass dann die gesamte Folge $\{x_k\}$ gegen einen der stationären Punkte von f konvergiert.

Hinweis: Siehe J. M. Ortega, W. C. Rheinboldt (1970, S. 476). Man überlege sich, dass auch die Menge der Häufungspunkte von $\{x_k\}$ endlich ist. Wenn es genau einen Häufungspunkt gibt, so ist die Behauptung richtig. Die Annahme, dass es mehr als einen Häufungspunkt gibt, führe man zum Widerspruch.

Lösung: Sei $H \subset L_0$ die Menge der Häufungspunkte von $\{x_k\}$. Da jeder Häufungspunkt von $\{x_k\}$ wegen Satz 1.3 eine stationäre Lösung ist und es von diesen nach Voraussetzung nur endlich viele gibt, ist H ebenfalls endlich. Sei etwa $H = \{z_1, \ldots, z_m\}$. Ist $m = 1$, so konvergiert die gesamte Folge $\{x_k\}$ trivialerweise gegen den einzigen Häufungspunkt. Wir nehmen daher an, es sei $m > 1$, und definieren die positive Zahl

$$\delta := \min\{\|z_i - z_j\| : i \neq j,\ i, j = 1, \ldots, m\},$$

also den kleinsten Abstand zwischen zwei Häufungspunkten. Es existiert ein $k_0 \in \mathbb{N}$ mit

$$x_k \in \bigcup_{i=1}^{m} B[z_i; \delta/4], \qquad \|x_{k+1} - x_k\| \leq \delta/4 \qquad \text{für alle } k \geq k_0,$$

wobei $B[z_i; \delta/4]$ die (abgeschlossene euklidische) Kugel um z_i mit Radius $\delta/4$ bedeutet. Ist nun $x_{k_1} \in B[z_1; \delta/4]$ für ein $k_1 \geq k_0$, so ist

$$\begin{aligned} \|z_i - x_{k_1+1}\| &\geq \|z_i - z_1\| - (\|z_1 - x_{k_1}\| + \|x_{k_1} - x_{k_1+1}\|) \\ &\geq \delta - 2\delta/4 \\ &= \delta/2, \qquad i = 2, \ldots, m. \end{aligned}$$

Folglich ist notwendigerweise $x_{k_1+1} \in B[z_1; \delta/4]$. Durch vollständige Induktion folgt $x_k \in B[z_1; \delta/4]$ für alle $k \geq k_1$, was der Annahme widerspricht, dass $z_2, \ldots, z_m$ Häufungspunkte von $\{x_k\}$ sind. Also ist $m = 1$ und die Aussage bewiesen.

6. Seien $y, s \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $H \in \mathbb{R}^{n\times n}$ gegeben. Es sei $(y - Hs)^T s \neq 0$. Man bestimme $\gamma \in \mathbb{R}$ und $u \in \mathbb{R}^n$ so, dass die Matrix $H_+ = H + \gamma uu^T$ der Quasi-Newton-Gleichung $H_+ s = y$ genügt. Unter welchen Voraussetzungen an H, y und s ist die so bestimmte Matrix H_+ positiv definit?

   Hinweis: Zu der symmetrischen, positiv definiten Matrix $H \in \mathbb{R}^{n\times n}$ gibt es eine symmetrische, positiv definite Matrix $H^{1/2}$ mit $H^{1/2}H^{1/2} = H$ (siehe z. B. C. Geiger, C. Kanzow (1999, S. 331)). Man benutze, dass die Matrix H_+ genau dann positiv definit ist, wenn $H^{-1/2}H_+H^{-1/2}$ es ist, wobei $H^{-1/2} := (H^{1/2})^{-1}$.

   Lösung: Eine Matrix $H_+ = H + \gamma uu^T$ genügt der Quasi-Newton-Gleichung $H_+ s = y$, wenn $Hs + \gamma(u^T s)u = y$. Es liegt daher nahe, $u := y - Hs$ zu wählen

und γ aus $\gamma(u^T s) = 1$ zu bestimmen. Insgesamt kommen wir auf die sogenannte SR1-Update-Formel[2]

$$H_+ := H + \frac{(y - Hs)(y - Hs)^T}{(y - Hs)^T s}.$$

Da wir $(y - Hs)^T s \neq 0$ vorausgesetzt haben, ist H_+ wohldefiniert. Jetzt untersuchen wir, unter welchen Voraussetzungen H_+ positiv definit ist . Es ist

$$H^{-1/2} H_+ H^{-1/2} = I + \frac{(H^{-1/2}y - H^{1/2}s)(H^{-1/2}y - H^{1/2}s)}{(y - Hs)^T s}.$$

Hieraus liest man ab, dass 1 ein $(n-1)$-facher Eigenwert mit Eigenvektoren aus dem orthogonalen Komplement zu span $\{H^{-1/2}y - H^{1/2}s\}$ ist, während der andere Eigenwert

$$\lambda := 1 + \frac{\|H^{-1/2}y - H^{1/2}s\|^2}{(y - Hs)^T s} = \frac{y^T H^{-1} y - y^T s}{y^T s - s^T H s}$$

mit dem Eigenvektor $H^{-1/2}y - H^{1/2}s$ ist. Daher ist H_+ genau dann positiv definit, wenn

$$\frac{y^T H^{-1} y - y^T s}{y^T s - s^T H s} > 0.$$

7. Man gebe eine MATLAB-Implementation des CG-Verfahrens von Hestenes-Stiefel für lineare Gleichungssysteme mit symmetrischer, positiv definiter Koeffizientenmatrix an und löse hiermit die Aufgabe, die Funktion

$$f(x) := x_1^2 + 0.3x_1x_2 + 0.975x_2^2 + 0.01x_1x_3 + x_3^2 + 3x_1 - 4x_2 + x_3$$

auf dem $\mathbb{R}^3$ zu minimieren[3].

Lösung: Wir geben das folgende function M-file `ConGrad.m` an:

```
function [x,error,iter] = ConGrad(A,b,x,max_iter,tol)
%**********************************************************************
% Diese Funktion loest das  symmetrische, positiv definite lineare System
% Ax=b mit dem CG-Verfahren ohne  Praekonditionierung.
% Input-Parameter
%         A        symmetrische, positiv definite Matrix
%         b        Vektor der rechten Seite
%         x        Startvektor
%         max_iter maximale Zahl der Iterationen
%         tol      Fehlertoleranz: Abbruch wenn ||grad f||/||b||<=tol
%
% Output-Parameter
```

[2]SR1 steht hier für **S**ymmetrischer **R**ang **1** Update, siehe auch C. Geiger, C. Kanzow (1999, S. 176). Gibt man SR1 in Google ein, so lernt man u. a., dass SR1 auch Abkürzung für **S**chul**R**echner 1 ist, einem Taschenrechner in der DDR, der in der Erweiterten Oberschule zum Schuljahr 1984/85 eingeführt wurde.

[3]Siehe P. Spellucci (1993, S. 164).

```
%         x        Loesung
%         error    Norm des Fehlers: ||A*x-b||/||b||
%         iter     Zahl der ausgefuehrten Iterationen
%*************************************************************************
iter=0; [n,n]=size(A);
bnrm2=norm(b);
if bnrm2==0
    x=zeros(n,1); error=0; return
end;
g=A*x-b; p=-g; rho=-g'*p;
error=norm(g)/bnrm2;
while (error>tol)&(iter<max_iter)
    q=A*p; t=rho/(p'*q); x=x+t*p; g=g+t*q;
    rho_plus=g'*g; beta=rho_plus/rho;
    p=-g+beta*p; rho=rho_plus;
    error=norm(g)/bnrm2; iter=iter+1;
end;
```

Wir geben ein:

```
>>A=[2,0.3,0.01;0.3,1.95,0;0.01,0,2];
>>b=[-3;4;-1];L=eye(3);x_0=zeros(3,1);
>>[x,error,iter]=ConGrad(A,b,x_0,L,5,1e-8);
```

Nach `format long` erhalten wir

$$x = \begin{pmatrix} -1.847881933986500 \\ 2.335571579587667 \\ -0.490760590330068 \end{pmatrix}, \qquad \text{error} = 5.303306391588063 \cdot 10^{-18}$$

und

$$\text{iter} = 3.$$

8. Gegeben sei die quadratische Zielfunktion $f(x) := \frac{1}{2}x^T Ax - b^T x$ mit einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$. Seien $p_0, \ldots, p_{n-1} \in \mathbb{R}^n$ konjugiert bezüglich A, d. h. $p_0, \ldots, p_{n-1}$ sind vom Nullvektor verschieden und es ist $p_i^T A p_j = 0$, $0 \le i < j \le n-1$. Man betrachte das folgende Verfahren zur Minimierung von $f(x)$ auf dem $\mathbb{R}^n$:

   - Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := Ax_0 - b$.
   - Für $k = 0, 1, \ldots$:
     - Falls $g_k = 0$, dann: $m := k$, f nimmt in x_m das Minimum an. STOP.
     - Andernfalls berechne

$$t_k := -\frac{g_k^T p_k}{p_k^T A p_k}, \qquad x_{k+1} := x_k + t_k p_k, \qquad g_{k+1} := g_k + t_k A p_k.$$

Durch vollständige Induktion nach k zeige man: Sind $g_0, \ldots, g_k \neq 0$, ist das Verfahren im k-ten Schritt also noch nicht abgebrochen, so ist x_{k+1} die Lösung der Aufgabe

$$(\mathrm{P}_k) \qquad \text{Minimiere} \quad f(x), \quad x \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}.$$

Wegen $x_0 + \operatorname{span}\{p_0, \ldots, p_{n-1}\} = \mathbb{R}^n$ bricht das Verfahren also nach $m \leq n$ Schritten mit dem Minimum von f ab.

Hinweis: Nach Konstruktion ist klar, dass $x_{k+1} \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$. Man zeige, dass $g_{k+1}^T p_i = 0$, $i = 0, \ldots, k$, und überlege sich, dass dies die Behauptung impliziert.

Lösung: Durch vollständige Induktion nach k zeigen wir: Sind $g_0, \ldots, g_k \neq 0$, so ist $x_{k+1} \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$ und $g_{k+1}^T p_i = 0$, $i = 0, \ldots, k$. Der Induktionsanfang liegt bei $k = 0$. Es ist

$$x_1 = x_0 + t_0 p_0 \in x_0 + \operatorname{span}\{p_0\},$$

ferner

$$g_1^T p_0 = g_0^T p_0 + t_0 p_0^T A p_0 = g_0^T p_0 - \frac{g_0^T p_0}{p_0^T A p_0} p_0^T A p_0 = 0.$$

Der Induktionsanfang ist also gelegt. Nun nehmen wir an, die Behauptung sei für $k-1$ richtig. Es seien also $g_0, \ldots, g_k \neq 0$ und $x_k \in x_0 + \operatorname{span}\{p_0, \ldots, p_{k-1}\}$ und $g_k^T p_i = 0$, $i = 0, \ldots, k-1$. Dann ist

$$x_{k+1} = x_k + t_k p_k \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$$

und

$$g_{k+1}^T p_i = (g_k + t_k A p_k)^T p_i = g_k^T p_i + t_k p_k^T A p_i.$$

Wegen der A-Konjugiertheit der Richtungen und der Induktionsannahme ist $g_{k+1}^T p_i = 0$, $i = 0, \ldots, k-1$. Weiter ist

$$g_{k+1}^T p_k = g_k^T p_k - \frac{g_k^T p_k}{p_k^T A p_k} p_k^T A p_k = 0,$$

so dass der Induktionsbeweis abgeschlossen ist. Sei nun $x \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$ beliebig. Dann ist

$$\begin{aligned} f(x) - f(x_{k+1}) &= \underbrace{\nabla f(x_{k+1})}_{=g_{k+1}}{}^T (x - x_{k+1}) + \underbrace{\frac{1}{2}(x - x_{k+1})^T A (x - x_{k+1})}_{\geq 0} \\ &\geq g_{k+1}^T (x - x_{k+1}) \\ &= 0, \end{aligned}$$

da $x - x_{k+1} \in \operatorname{span}\{p_0, \ldots, p_k\}$ und $g_{k+1}^T p_i = 0$, $i = 0, \ldots, k$. Folglich ist x_{k+1} Lösung von (P_k). Es ist die eindeutige Lösung, wie man aus obiger Gleichungs-Ungleichungskette abliest. Damit ist die Aufgabe gelöst.

9. Das Verfahren der konjugierten Gradienten von Polak-Ribière unterscheidet sich von dem Fletcher-Reeves-Verfahren nur darin, dass

$$\beta_k := \frac{g_{k+1}^T(g_{k+1} - g_k)}{\|g_k\|^2}$$

(statt $\beta_k := \|g_{k+1}\|^2/\|g_k\|^2$) gesetzt wird. Man betrachte das dann definierte Polak-Ribière-Verfahren mit exakter Schrittweitenstrategie zur Lösung von

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

Man zeige: Sind die (gleichmäßigen) Konvexitätsvoraussetzungen (K) (a)–(c) aus Lemma 1.6 erfüllt, so liefert das Verfahren, wenn es nicht vorzeitig mit der Lösung x^* von (P) abbricht, eine Folge $\{x_k\}$, die R-linear gegen x^* konvergiert.

Hinweis: Man setze

$$\delta_k := \Big(\frac{g_k^T p_k}{\|g_k\|\,\|p_k\|}\Big)^2 = \frac{\|g_k\|^2}{\|p_k\|^2},$$

zeige $\|p_{k+1}\| \le (1+\gamma/c)\|g_{k+1}\|$, $k = 0, 1, \ldots$, folgere hieraus auf die Existenz einer Konstanten $\delta > 0$ mit $\delta_k \ge \delta$, $k = 0, 1, \ldots$, und wende Satz 1.7 an (siehe auch J. Werner (1992b, S. 234)).

Lösung: Wie angegeben definiere man δ_k. Es ist

$$\begin{aligned}
\|p_{k+1}\| &= \| - g_{k+1} + \beta_k p_k\| \\
&\le \|g_{k+1}\| + \frac{|g_{k+1}^T(g_{k+1} - g_k)|}{\|g_k\|^2}\|p_k\| \\
&\le \|g_{k+1}\| + \frac{\|g_{k+1}\|\,\gamma\,\|x_{k+1} - x_k\|}{\|g_k\|^2}\|p_k\| \\
&= \Big(1 + \gamma t_k \frac{\|p_k\|^2}{\|g_k\|^2}\Big)\|g_{k+1}\| \\
&\le \Big(1 + \frac{\gamma}{c}\Big)\|g_{k+1}\|.
\end{aligned}$$

Hierbei haben wir am Schluss benutzt, dass

$$0 = g_{k+1}^T p_k = g_k^T p_k + (g_{k+1} - g_k)^T p_k \ge g_k^T p_k + c t_k \|p_k\|^2 = -\|g_k\|^2 + c t_k \|p_k\|^2,$$

also

$$t_k \le \frac{\|g_k\|^2}{c\,\|p_k\|^2}$$

gilt. Folglich ist

$$\delta_{k+1} = \frac{\|g_{k+1}\|^2}{\|p_{k+1}\|^2} \ge \frac{1}{(1+\gamma/c)^2} =: \delta > 0, \qquad k = 0, 1, \ldots.$$

Damit ist die Behauptung bewiesen.

8.2.2 Aufgaben zu Abschnitt 2.2

1. Gegeben sei die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T H p, \quad \|p\|_2 \leq \Delta,$$

wobei

$$g := \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad H := \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \Delta := \frac{1}{2}.$$

Man definiere $p : [0, \infty) \longrightarrow \mathbb{R}$ durch $p(\lambda) := -(H + \lambda I)^{-1} g$. Auf dem Intervall $[0, 2]$ plotte man $\|p(\cdot)\|_2$ und $1/\|p(\cdot)\|_2$. Anschließend berechne man die Lösung von (P) numerisch, indem man z.B. auf

$$\chi(\lambda) := \frac{1}{\|p(\lambda)\|_2} - \frac{1}{\Delta} = 0$$

das Newton-Verfahren mit Startwert $\lambda_0 := 0$ anwendet. Ist $\chi(\lambda^*) = 0$, so ist $p^* := p(\lambda^*)$ die Lösung von (P).

Lösung: Offensichtlich ist H positiv definit, ferner ist

$$p(\lambda) = \frac{1}{(2 + \lambda)(1 + \lambda) - 1} \begin{pmatrix} -\lambda \\ 1 + \lambda \end{pmatrix}.$$

Daher ist

$$\psi(\lambda) := \|p(\lambda)\|_2 = \frac{\sqrt{\lambda^2 + (1 + \lambda)^2}}{\lambda^2 + 3\lambda + 1}.$$

In Abbildung 8.5 skizzieren wir $\|p(\cdot)\|_2$ und $1/\|p(\cdot)\|_2$ über dem Intervall $[0, 2]$. Man erkennt, dass $1/\|p(\cdot)\|_2$ fast linear ist, so dass das auf

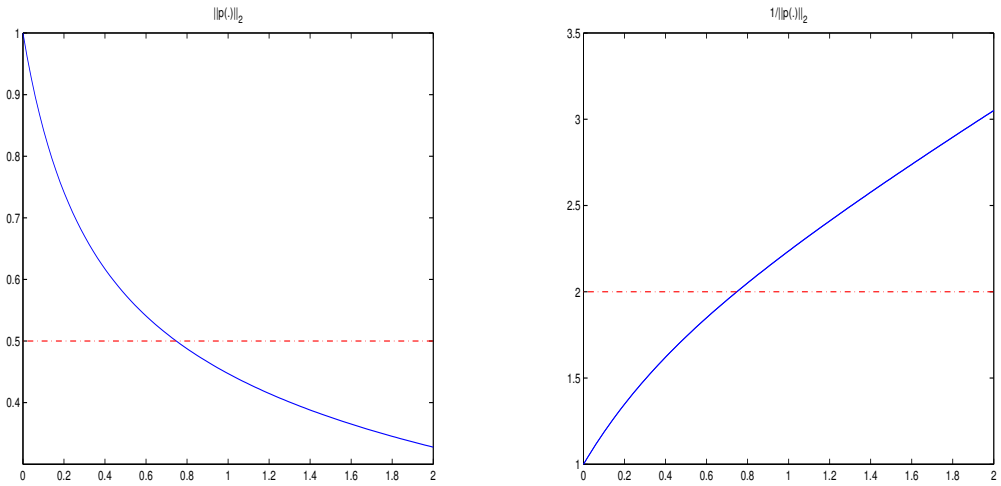


Abbildung 8.5: Plot von $\|p(\cdot)\|_2$ (links) und $1/\|p(\cdot)\|_2$ (rechts)

$$\chi(\lambda) := \frac{1}{\psi(\lambda)} - \frac{1}{\Delta} = 0$$

angewandte Newton-Verfahren schnell konvergieren wird. Mit

$$\psi'(\lambda) = \frac{2\lambda+1}{\sqrt{\lambda^2+(1+\lambda)^2}(\lambda^2+3\lambda+1)} - \frac{\sqrt{\lambda^2+(1+\lambda)^2}(2\lambda+3)}{(\lambda^2+3\lambda+1)^2}$$

und wegen

$$\chi'(\lambda) = -\frac{\psi'(\lambda)}{\psi(\lambda)^2}$$

lautet das Newton-Verfahren:

$$\lambda_{k+1} := \lambda_k - \frac{\chi(\lambda_k)}{\chi'(\lambda_k)} = \lambda_k + \Big(1 - \frac{\psi(\lambda_k)}{\Delta}\Big)\frac{\psi(\lambda_k)}{\psi'(\lambda_k)}$$

Mit dem Startwert $\lambda_0 := 0$ erhält man die Werte

k	λ_k
0	0
1	0.500000000000000
2	0.729043144537989
3	0.747453772907109
4	0.747535239909947
5	0.747535241461013
6	0.747535241461014

$$p^* = p(\lambda^*) = \begin{pmatrix} -0.196646592914588 \\ 0.459706555853932 \end{pmatrix}.$$

Natürlich könnte man auch die numerischen Fähigkeiten eines Computeralgebra-Systems (MuPAD, Maple oder Mathematica) ausnutzen.

2. Sei $f \in \mathbb{R}$, $g \in \mathbb{R}^n \setminus \{0\}$ und $\Delta > 0$. Man gebe eine Lösung von

(P) $$\text{Minimiere} \quad \phi(p) := f + g^T p, \quad \|p\|_\infty \le \Delta$$

an und begründe dies.

Lösung: Man definiere $p^* = (p_j^*)$ durch

$$p_j^* := -\text{sign}\,(g_j)\Delta, \qquad j = 1, \ldots, n.$$

Dann ist

$$\phi(p^*) = f + g^T p^* = f + \sum_{j=1}^n g_j p_j^* = f - \Delta \sum_{j=1}^n |g_j| = f - \Delta\|g\|_1.$$

Für ein beliebiges $p \in \mathbb{R}^n$ mit $\|p\|_\infty \le \Delta$ ist andererseits

$$\phi(p) = f + g^T p \ge f - \sum_{j=1} |g_j|\,|p_j| \ge f - \Delta \sum_{j=1} |g_j| = f - \Delta\|g\|_1 = \phi(p^*).$$

Daher ist das angegebene p^* eine Lösung von (P).

3. Man betrachte die unrestringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad p \in \mathbb{R}^n,$$

wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^n$ und $B \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix ist. Man zeige:

(a) (P) besitzt genau dann eine Lösung, wenn B positiv semidefinit und $g \in \text{Bild}\,(B)$ ist.

(b) (P) besitzt genau dann eine eindeutige Lösung, wenn B positiv definit ist.

Lösung: Angenommen, (P) besitzt eine lokale Lösung p^*. Diese muss den notwendigen Bedingungen zweiter Ordnung genögen, d. h. es ist $\nabla\phi(p^*) = 0$ bzw. $g + Bp^* = 0$, folglich $g \in \text{Bild}\,(B)$, und $\nabla^2\phi(p^*) = B$ ist positiv semidefinit. Die Umkehrung ist trivial, denn ist B positiv semidefinit, so ist ϕ konvex. Wegen $g \in \text{Bild}\,(B)$ existiert ferner ein p^* mit $\nabla\phi(p^*) = 0$, dieses p^* ist Lösung von (P).

Besitzt (P) eine eindeutige Lösung, so ist wegen des ersten Teils B zumindestens positiv semidefinit und $g \in \text{Bild}\,(B)$. Jedes p^* mit $g + Bp^*$ ist eine Lösung, ist diese eindeutig, so ist der Kern von B notwendigerweise trivial und damit B positiv definit. Die Umkehrung ist natürlich trivial.

4. Sei

$$g := \begin{pmatrix} -2 \\ -20 \end{pmatrix}, \qquad B := \begin{pmatrix} 42 & 0 \\ 0 & 20 \end{pmatrix}.$$

Für $\Delta := \frac{1}{2}$, 1, 2 berechne man eine Lösung von

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_\infty \le \Delta.$$

Lösung: Sei zunächst $\Delta = 0.5$. Wir zeigen, dass $p^* = (\frac{1}{21}, \frac{1}{2})^T$ eine Lösung ist. Denn für ein beliebiges $p \in \mathbb{R}^2$ mit $\|p\|_\infty \le \frac{1}{2}$ ist

$$\begin{aligned}
\phi(p) - \phi(p^*) &= \nabla\phi(p^*)^T(p - p^*) + \underbrace{\frac{1}{2}(p - p^*)^T B (p - p^*)}_{\ge 0} \\
&\ge \nabla\phi(p^*)^T(p - p^*) \\
&= \begin{pmatrix} 0 \\ -10 \end{pmatrix}^T \begin{pmatrix} p_1 - \frac{1}{21} \\ p_2 - \frac{1}{2} \end{pmatrix} \\
&= 10\underbrace{(\tfrac{1}{2} - p_2)}_{\ge 0} \\
&\ge 0,
\end{aligned}$$

womit die Behauptung bewiesen ist. Für $\Delta = 1$ ist entsprechend $p^* = (\frac{1}{21}, 1)^T$, während man für $\Delta = 2$ dieselbe Lösung $p^* = (\frac{1}{21}, 1)^T$ erhält. Steht die Optimization Toolbox zur Verfügung, so erhält man nach

```
options=optimset('LargeScale','off');
g=[-2;-20];B=[42,0;0,20];Delta=0.5;
l=-Delta*ones(2,1);u=-l;
[p,v]=quadprog(B,g,[],[],[],[],l,u,[],options);
```

und `format long`

$$p = \begin{pmatrix} 0.04761904761905 \\ 0.50000000000000 \end{pmatrix}, \qquad v = -7.54761904761905.$$

Die leeren Klammern `[]` bedeuten hierbei jeweils, dass keine Ungleichungen (Matrix und rechte Seite unbesetzt), keine Gleichungen (Matrix und rechte Seite unbesetzt) in der Optimierungsaufgabe vorkommen und dass kein Startwert vorgegeben wird.

5. Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch mit kleinstem Eigenwert λ_1 und $g \in \mathbb{R}^n \setminus \{0\}$. Man zeige, dass die durch
$$\psi(\lambda) := \|(B+\lambda I)^{-1} g\|_2$$
definierte Funktion $\psi\colon (-\lambda_1, \infty) \longrightarrow \mathbb{R}$ auf $(-\lambda_1, \infty)$ monoton fallend und konvex ist.

Lösung: Seien $\lambda_1 \le \cdots \le \lambda_n$ die Eigenwerte von B und $\{u_1, \ldots, u_n\}$ ein zugehöriges Orthonormalsystem von Eigenvektoren. Mit $\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ und $U := (\ u_1 \ \cdots \ u_n\ )$ ist dann

$$\begin{aligned} \psi(\lambda) &= \|(B+\lambda I)^{-1} g\|_2 \\ &= \|U(\Lambda+\lambda I)^{-1} U^T g\|_2 \\ &= \|(\Lambda+\lambda I)^{-1} U^T g\|_2 \\ &= \Big(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^2}\Big)^{1/2} \end{aligned}$$

für $\lambda \in (-\lambda_1, \infty)$. Als Ableitung berechnet man

$$\psi'(\lambda) = -\Big(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^2}\Big)^{-1/2} \Big(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3}\Big) = -\frac{1}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3}.$$

Wegen $g \neq 0$ ist $u_i^T g \neq 0$ für wenigstens ein i und folglich $\psi'(\lambda) < 0$ für alle $\lambda \in (-\lambda_1, \infty)$, also $\psi(\cdot)$ auf $(-\lambda_1, \infty)$ monoton fallend. Erneutes Differenzieren liefert

$$\begin{aligned} \psi''(\lambda) &= \frac{3}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_t^T g)^2}{(\lambda_i+\lambda)^4} + \frac{\psi'(\lambda)}{\psi(\lambda)^2} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3} \\ &= \frac{3}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^4} - \frac{1}{\psi(\lambda)^3} \Big(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3}\Big)^2 \\ &= \frac{2}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^4} \end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{1}{\psi(\lambda)^3}\Big[\Big(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^2}\Big)\Big(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^4}\Big) - \Big(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3}\Big)^2\Big] \\
&\geq \frac{2}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^4} \\
&> 0,
\end{aligned}
$$

insbesondere ist $\psi(\cdot)$ auf $(-\lambda_1, \infty)$ (strikt) konvex.

6. **Programmieraufgabe:** Sei $H \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $g \in \mathbb{R}^n$ und $\Delta > 0$. Man schreibe eine MATLAB-Funktion `TrustStep` zur Berechnung der Lösung des Trust-Region-Hilfsproblems

$$
\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T H p, \quad \|p\|_2 \leq \Delta.
$$

Anschließend teste man die Funktion für den Spezialfall, dass

$$
H := \begin{pmatrix}
2 & -1 & & & \\
-1 & 2 & -1 & & \\
& \ddots & \ddots & \ddots & \\
& & -1 & 2 & -1 \\
& & & -1 & 2
\end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad g := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^n
$$

mit $n := 10$ und $\Delta := 1$.

Lösung: Wir schreiben die folgende Funktion:

```
function [p,lambda,iter]=TrustStep(H,g,Delta,max_iter,tol);
%****************************************************************
%Loese das Trust-Region-Hilfsproblem
%Minimiere g^Tp+0.5*p^THp,   ||p||_2<=Delta
%****************************************************************
%Input-Parameter:
%  H        symmetrische, positiv definite n-mal-n Matrix
%  g        n-Vector
%  Delta    positiver Trust-Region-Radius
%  max_iter maximale Zahl der Iterationen
%  tol      Toleranz: STOP, wenn |||p||_2-Delta|<=tol
%Output-Parameter:
%  p        approximative Loesung
%  lambda   zugehoeriger Multiplikator
%  iter     Zahl der Iterationen
%****************************************************************
lambda=0; n=length(g);
for k=1:max_iter
    L=chol(H+lambda*eye(n))';
    p=-L'\(L\g);normp=norm(p);
    if ((k==1)&(normp<=Delta)|((k>1))&(abs(normp-Delta)<=tol))
        iter=k-1;return
    end;
```

```
    w=L\p;
    lambda=lambda+((normp-Delta)/Delta)*(normp/norm(w))^2;
end;
iter=max_iter;
```

Zunächst setzen wir

```
>> H=2*eye(10)+diag(-ones(9,1),1)+diag(-ones(9,1),-1);
>> g=ones(10,1);
```

Mit dem Aufruf

```
[p,lambda,iter]=TrustStep(H,g,1.0,100,1e-12);
```

erhalten wir (mit `format long`)

$$p = \begin{pmatrix} -0.263308042110668 \\ -0.318188480444401 \\ -0.329626173531830 \\ -0.332005798923651 \\ -0.332481128902764 \\ -0.332481128902764 \\ -0.332005798923651 \\ -0.329626173531830 \\ -0.318188480444401 \\ -0.263308042110668 \end{pmatrix}, \qquad \lambda = 3.006259853963632, \qquad \text{iter} = 3.$$

7. Gegeben sei das Trust-Region Hilfsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T H p, \quad \|p\|_2 \leq \Delta,$$

wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^{n \times n} \setminus \{0\}$, die symmetrische Matrix $H \in \mathbb{R}^{n \times n}$ und $\Delta > 0$ gegeben sind. Der sogenannte *Cauchy-Punkt* p^C ist definiert durch

$$p^C := -\tau \frac{\Delta}{\|g\|_2} g,$$

wobei

$$\tau := \begin{cases} 1, & \text{falls} \quad g^T H g \leq 0, \\ \min(\|g\|_2^3/(\Delta g^T H g), 1), & \text{sonst.} \end{cases}$$

Man zeige, dass p^C eine Lösung der Aufgabe

$$\text{Minimiere} \quad \phi(p), \quad \|p\|_2 \leq \Delta, \; p \in \text{span}\,\{g\}$$

ist.

Hinweis: Man stelle $p \in \operatorname{span}\{g\}$ mit $\|p\|_2 \leq \Delta$ dar als $p = -\sigma(\Delta/\|g\|_2)g$ mit $|\sigma| \leq 1$.

Lösung: Die angebene Transformation führt auf die Aufgabe

$$\text{Minimiere} \quad \psi(\sigma) := f - \sigma\Delta\|g\|_2 + \frac{\sigma^2}{2}\frac{\Delta^2}{\|g\|_2^2} g^T H g, \quad |\sigma| \leq 1.$$

Ist $g^T H g \leq 0$, so ist $\psi(\cdot)$ eine konkave Funktion, die ihr Minimum an einem der Intervallenden, also für $\sigma = 1$ annimmt. Sei daher jetzt $g^T H g > 0$ und daher $\psi(\cdot)$ konvex. Ist die Nullstelle $\|g\|_2^3/(\Delta g^T H g)$ von $\psi'(\cdot)$ kleiner oder gleich 1, so ist sie eine Lösung, andernfalls ist es 1. Insgesamt ist die Behauptung bewiesen.

8. Gegeben sei das Trust-Region Hilfsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T H p, \quad \|p\|_2 \leq \Delta,$$

wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^{n\times n} \setminus \{0\}$, die symmetrische Matrix $H \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben sind. Sei p^C der in Aufgabe 7 definierte Cauchy-Punkt. Man zeige, dass

$$f - \phi(p^C) \geq \frac{1}{2}\|g\|_2 \min\Big(\Delta, \frac{\|g\|_2}{\|H\|_2}\Big),$$

der Cauchy-Punkt p^C also derselben Abschätzung wie eine globale Lösung p^* von (P) genügt, siehe Satz 2.1.

Hinweis: Man unterscheide zwischen den Fällen $g^T H g \leq 0$ und $g^T H g > 0$. Für $g^T H g > 0$ betrachte man die beiden Fälle $\|g\|_2^3/g^T H g \leq 1$ und $\|g\|_2^3/g^T H g > 1$.

Lösung: Wir nehmen zunächst an, es sei $g^T H g \leq 0$. Dann ist

$$\begin{aligned}
f - \phi(p^C) &= -g^T p^C - \frac{1}{2}(p^C)^T H p^C \\
&= \Delta\|g\|_2 - \frac{1}{2}\frac{\Delta^2}{\|g\|_2^2}\underbrace{g^T H g}_{\leq 0} \\
&\geq \Delta\|g\|_2 \\
&\geq \|g\|_2 \min\Big(\Delta, \frac{\|g\|_2}{\|H\|_2}\Big) \\
&\geq \frac{1}{2}\|g\|_2 \min\Big(\Delta, \frac{\|g\|_2}{\|H\|_2}\Big).
\end{aligned}$$

In diesem Fall ist die Behauptung also richtig. Nun nehmen wir $g^T H g > 0$ an. Auch für diesen Fall machen wir eine Fallunterscheidung und setzen zunächst $\|g\|_2^3/(\Delta g^T H g) \leq 1$ voraus. Dann ist

$$\begin{aligned}
f - \phi(p^C) &= -g^T p^C - \frac{1}{2}(p^C)^T H p^C \\
&= \frac{1}{2}\frac{\|g\|_2^4}{g^T H g}
\end{aligned}$$

$$\begin{aligned} &\geq \frac{1}{2}\frac{\|g\|_2^4}{\|H\|_2\,\|g\|_2^2} \\ &= \frac{1}{2}\frac{\|g\|_2^2}{\|H\|_2} \\ &\geq \frac{1}{2}\|g\|_2 \min\Big(\Delta, \frac{\|g\|_2}{\|H\|_2}\Big). \end{aligned}$$

Ist dagegen $\|g\|_2^3/(\Delta g^T H g) > 1$, so ist

$$\begin{aligned} f - \phi(p^C) &= -g^T p^C - \frac{1}{2}(p^C)^T H p^C \\ &= \Delta\|g\|_2 - \frac{1}{2}\frac{\Delta^2}{\|g\|_2^2} g^T H g \\ &\geq \Delta\|g\|_2 - \frac{1}{2}\Delta\|g\|_2 \\ &= \frac{1}{2}\Delta\|g\|_2 \\ &\geq \frac{1}{2}\|g\|_2 \min\Big(\Delta, \frac{\|g\|_2}{\|B\|_2}\Big). \end{aligned}$$

Die Aufgabe ist gelöst.

8.3 Aufgaben zu Kapitel 3

8.3.1 Aufgaben zu Abschnitt 3.1

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
   $$\|P_K(x) - P_K(y)\| \leq \|x - y\| \qquad \text{für alle } x, y \in \mathbb{R}^n.$$
   (Hierbei bedeutet $\|\cdot\|$ natürlich die euklidische Norm auf dem $\mathbb{R}^n$.) Die Projektionsabbildung ist also *nicht expandierend* auf dem $\mathbb{R}^n$.

   (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: Mit vorgegebenen $x, y \in \mathbb{R}^n$ liefert eine Anwendung der notwendigen und hinreichenden Optimalitätsbedingungen des Projektionssatzes 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

$$[P_K(x) - P_K(y) - (x-y)]^T[P_K(x) - P_K(y)] \le 0$$

bzw. mit der Cauchy-Schwarzschen Ungleichung

$$\|P_K(x) - P_K(y)\|^2 \le (x-y)^T[P_K(x) - P_K(y)] \le \|x-y\|\,\|P_K(x) - P_K(y)\|,$$

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 $x \in L$ (Beweis?). Ist daher $[z_1 - P_L(z_1)]^T x = 0$ und $[z_2 - P_L(z_2)]^T x = 0$ jeweils 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 $l \le u$. Hiermit definiere man den Quader

$$Q := \{x \in \mathbb{R}^n : l \le x \le u\}.$$

Man zeige, dass für $x \in \mathbb{R}^n$ die Projektion $P_Q(x)$ von x auf Q durch

$$P_Q(x)_j = \begin{cases} l_j, & \text{falls} \quad x_j < l_j, \\ x_j, & \text{falls} \quad l_j \le x_j \le u_j, \\ u_j, & \text{falls} \quad u_j < x_j, \end{cases} \qquad j = 1, \ldots, n,$$

gegeben ist.

Lösung: Für $x \in \mathbb{R}^n$ ist $P_Q(x) \in Q$. Wegen des Projektionssatzes bleibt die charakteriserende Eigenschaft

$$(P_Q(x) - x)^T (z - P_Q(x)) \ge 0 \qquad \text{für alle } z \in Q$$

nachzuprüfen. Wir definieren die Indexmengen

$$\begin{aligned} J_- &:= \{j \in \{1,\dots,n\} : x_j < u_j\}, \\ J_0 &:= \{j \in \{1,\dots,n\} : l_j \le x_j \le u_j\}, \\ J_+ &:= \{j \in \{1,\dots,n\} : u_j < x_j\}. \end{aligned}$$

Für ein gegebenes $z \in Q$ ist dann

$$\begin{aligned} (P_Q(x) - x)^T (z - P_Q(x)) &= \sum_{j=1}^{n} (P_Q(x)_j - x_j)(z_j - P_Q(x)_j) \\ &= \sum_{j \in J_-} (\underbrace{l_j - x_j}_{>0})(\underbrace{z_j - l_j}_{\ge 0}) + \sum_{j \in J_+} (\underbrace{u_j - x_j}_{<0})(\underbrace{z_j - u_j}_{\le 0}) \\ &\ge 0. \end{aligned}$$

Damit ist die Behauptung nachgewiesen.

3. 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

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

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 $C = \operatorname{cl}(\operatorname{int}(C))$.

Lösung: Wie im Hinweis angegeben, zeigen wir zunächst, dass mit C auch $\operatorname{int}(C)$ konvex ist. Seien hierzu $x_1, x_2 \in \operatorname{int}(C)$ sowie $\lambda \in (0,1)$. Wir zeigen, dass auch $(1-\lambda)x_1 + \lambda x_2 \in \operatorname{int}(C)$. Da $x_1 \in \operatorname{int}(C)$, 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 $\|(1-\lambda)x_1 + \lambda x_2 - x\| \le (1-\lambda)\epsilon_1$, 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

$$\|x_1 - z\| = \left\| x_1 - \frac{1}{1-\lambda}(x - \lambda x_2) \right\| = \frac{1}{1-\lambda} \underbrace{\|(1-\lambda)x_1 + \lambda x_2 - x\|}_{\le (1-\lambda)\epsilon_1} \le \epsilon_1.$$

Die disjunkten, konvexen Mengen $\{x^*\}$ und $\operatorname{int}(C)$ lassen sich nach dem Trennungssatz 1.7 trennen, es existiert also ein $y \in \mathbb{R}^n \setminus \{0\}$ mit $y^T x^* \le y^T x$ für alle $x \in \operatorname{int}(C)$ und damit auch für alle $x \in \operatorname{cl}(\operatorname{int}(C))$. Nun ist jedes $x \in C$ Limes einer Folge aus $\operatorname{int}(C)$, wie z. B. sofort aus dem Beweis des ersten Teiles der Aufgabe folgt: Ist $x_1 \in \operatorname{int}(C)$ beliebig, ferner $\{\lambda_k\} \subset (0,1)$ eine beliebige Folge mit $\lambda_k \to 1$, so ist $x_k := (1-\lambda_k)x_1 + \lambda_k x \in \operatorname{int}(C)$ und $x_k \to x$, also $x \in \operatorname{cl}(\operatorname{int}(C))$.

4. Man zeige, dass zwei nichtleere, konvexe Mengen $A, B \subset \mathbb{R}^n$ genau dann stark trennbar sind, wenn $0 \notin \mathrm{cl}\,(B-A)$. Für eine Menge $C \subset \mathbb{R}^n$ bedeutet $\mathrm{cl}(C) \subset \mathbb{R}^n$ hierbei den *Abschluss* der Menge C, es ist also

$$\mathrm{cl}(C) := \{x \in \mathbb{R}^n : \text{Es existiert eine Folge } \{x_k\} \subset C \text{ mit } x = \lim_{k\to\infty} x_k\}.$$

Hinweis: Man überlege sich zunächst, dass mit konvexen Mengen A, B auch $B-A$ und der Abschluss $\mathrm{cl}\,(B-A)$ konvex sind.

Lösung: Mit A und B ist auch $B-A$ und dann auch $\mathrm{cl}\,(B-A)$ konvex. Denn sind $b_1-a_1, b_2-a_2 \in B-A$ und $\lambda \in [0,1]$, so ist

$$(1-\lambda)(b_1-a_1)+\lambda(b_2-a_2) = [\underbrace{(1-\lambda)b_1+\lambda b_2}_{\in B}] - [\underbrace{(1-\lambda)a_1+\lambda a_2}_{\in A}] \in B-A.$$

Ist ferner $C \subset \mathbb{R}^n$ konvex, so ist auch $\mathrm{cl}(C)$ konvex. Sind nämlich $x, y \in \mathrm{cl}(C)$, existieren also Folgen $\{x_k\}, \{y_k\} \subset C$ mit $x = \lim_{k\to\infty} x_k$, $y = \lim_{k\to\infty} y_k$, und ist $\lambda \in [0,1]$, so ist $(1-\lambda)x+\lambda y \in \mathrm{cl}(C)$, da $\{(1-\lambda)x_k+\lambda y_k\} \subset C$ und $(1-\lambda)x+\lambda y = \lim_{k\to\infty}(1-\lambda)x_k+\lambda y_k$. Ist $0 \notin \mathrm{cl}\,(B-A)$, so folgt aus dem Korollar zum starken Trennungssatz 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 $0 < \gamma \le y^T b - y^T a$, $a \in A$, $b \in B$, 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

$$y^T a_k \le \sup_{a\in A} y^T a < \inf_{b\in B} y^T b \le y^T b_k,$$

offensichtlich ein Widerspruch.

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

(I) $\qquad Ax = 0, \quad x \ge 0, \quad x \ne 0 \qquad$ 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: Denn ist $x \in \mathbb{R}^n$ eine Lösung von (I) und $y \in \mathbb{R}^m$ eine Lösung von (II), so ist

$$0 < x^T A^T y = (Ax)^T y = 0,$$

ein Widerspruch.

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.

6. Man beweise den folgenden Satz von Fan-Glicksburg-Hoffman (siehe z. B. O. L. Mangasarian (1969, S. 63)):

   Sei $C \subset \mathbb{R}^n$ nichtleer und konvex, die Abbildung $g : C \longrightarrow \mathbb{R}^l$ (komponentenweise) konvex, die Abbildung $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ 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 für konvexe Mengen 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.

8.3.2 Aufgaben zu Abschnitt 3.2

1. Man löse das folgende, auf S. Lhuillier (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

$$f(b_1) := \frac{1}{(\hat{l} - b_1)^2} - \frac{1}{b_1^2} - 4\Big[\frac{1}{a_2^2} - \frac{1}{a_1^2}\Big]$$

   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 $b_2 = 0.696485$.

2. Man zeige, dass $x^* := (1, 1, 2)^T$ die Lösung von

$$(\mathrm{P}) \quad \left\{ \begin{array}{lrrrcr} \text{Minimiere} & -5x_2 + \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) & \text{unter den Nebenbedingungen} & & & \\ & -4x_1 & -3x_2 & & \geq & -8 \\ & 2x_1 & + \ x_2 & & \geq & 2 \\ & & -2x_2 & +x_3 & \geq & 0 \\ & x_1 & -2x_2 & +x_3 & = & 1 \end{array} \right.$$

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.

3. Für die Aufgabe

$$(\mathrm{P}) \quad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := x_1^2 + 4x_2^2 + 16x_3^2 \quad \text{unter der Nebenbedingung} \\ h(x) := x_1 x_2 x_3 - 1 = 0 \end{array} \right.$$

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

$$x^{(1)} := \begin{pmatrix} 2 \\ 1 \\ \frac{1}{2} \end{pmatrix}, \quad x^{(2)} := \begin{pmatrix} -2 \\ -1 \\ \frac{1}{2} \end{pmatrix}, \quad x^{(3)} := \begin{pmatrix} 2 \\ -1 \\ -\frac{1}{2} \end{pmatrix}, \quad x^{(4)} := \begin{pmatrix} -2 \\ 1 \\ -\frac{1}{2} \end{pmatrix}.$$

Diese vier Vektoren genügen jeweils den notwendigen Opyimalitätsbedingungen (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.

4. Gegeben sei die Optimierungsaufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} & f(x) := -(x_1x_2 + x_2x_3 + x_1x_3) \quad \text{u. d. NB.} \\ & h(x) := x_1 + x_2 + x_3 - 3 = 0. \end{cases}$$

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.

Lösung: Da die Nebenbedingungen affin linear sind, braucht keine Constraint Qualification nachgeprüft zu werden. Zu einer lokalen Lösung $x^* = (x_1^*, x_2^*, x_3^*)^T$ existiert daher ein $v^* \in \mathbb{R}$ mit $\nabla f(x^*) + v^* \nabla h(x^*) = 0$ bzw.

$$-\begin{pmatrix} x_2^* + x_3^* \\ x_1^* + x_3^* \\ x_1^* + x_2^* \end{pmatrix} + v^* \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Zusammen mit der Nebenbedingung $h(x^*) = 0$ hat man zur Bestimmung von x_1^*, x_2^*, x_3^* und v^* das lineare Gleichungssystem

$$\begin{pmatrix} 0 & -1 & -1 & 1 \\ -1 & 0 & -1 & 1 \\ -1 & -1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1^* \\ x_2^* \\ x_3^* \\ v^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \end{pmatrix}$$

zu lösen. Hieraus erhält man

$$\begin{pmatrix} x_1^* \\ x_2^* \\ x_3^* \\ v^* \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \end{pmatrix}.$$

Um nachzuweisen, dass x^* eine strikte lokale Lösung von (P) ist, haben wir wegen Satz 2.6 nachzuprüfen, ob $\nabla^2 f(x^*)$ auf Kern $(h'(x^*))$ positiv definit ist. Zu überprüfen ist also, ob

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}^T \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = 0, \quad p \neq 0 \Longrightarrow p^T \begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix} p > 0.$$

Wegen

$$\text{span} \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}^{\perp} = \text{span} \left\{ \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \right\}$$

ist zu zeigen, dass

$$\begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix} \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

positiv definit ist, was offenbar der Fall ist.

5. Gegeben[4] sei die Optimierungsaufgabe

$$(\mathrm{P}_\gamma) \qquad \begin{cases} \text{minimiere} \quad f(x) := -(x_1+1)^2 - (x_2+1)^2 \quad \text{auf} \\ M_\gamma := \left\{ x \in \mathbb{R}^2 : g(x) := \begin{pmatrix} x_1^2 + x_2^2 - 2 \\ x_1 - \gamma \end{pmatrix} \leq 0 \right\}, \end{cases}$$

wobei $\gamma \geq -\sqrt{2}$ fest vorgegeben sei.

   (a) Man ermittle anhand einer Skizze die Lösung $x^* = x^*(\gamma)$ von (P_γ), wobei man die Fälle $\gamma = -\sqrt{2}$, $-\sqrt{2} < \gamma \leq 1$ und $\gamma > 1$ unterscheide.

   (b) Ist die Regularitätsbedingung (CQ) aus Satz 2.2 erfüllt?

   (c) Gibt es zu x^* ein $u^* \in \mathbb{R}^2$, so dass (x^*, u^*) ein Kuhn-Tucker-Punkt zu (P_γ) ist, also

$$u^* \geq 0, \qquad \nabla f(x^*) + g'(x^*)^T u^* = 0, \qquad g(x^*)^T u^* = 0$$

   gilt?

[4] Diese Aufgabe findet man bei C. Geiger, C. Kanzow (2002, S. 72).

Lösung: Geometrisch bedeutet die Aufgabe (P_γ), in M_γ, dem Durchschnitt des Kreises um den Nullpunkt mit dem Radius $\sqrt{2}$ und dem Halbraum $H^- := \{x \in \mathbb{R}^2 : x_1 \leq \gamma\}$, denjenigen Punkt zu finden, der maximalen euklidischen Abstand zu $z := (-1,-1)^T$ besitzt.

(a) Ist $\gamma = -\sqrt{2}$, so ist $M_\gamma = \{(-\sqrt{2},0)^T)\}$, und daher $x^* = (-\sqrt{2},0)^T$ die Lösung von (P_γ).
Für $-\sqrt{2} < \gamma \leq 1$ skizzieren wir den zulässigen Bereich in Abbildung 8.6 links. Offenbar ist $x^*(\gamma) = (\gamma, \sqrt{2-\gamma^2})^T$.

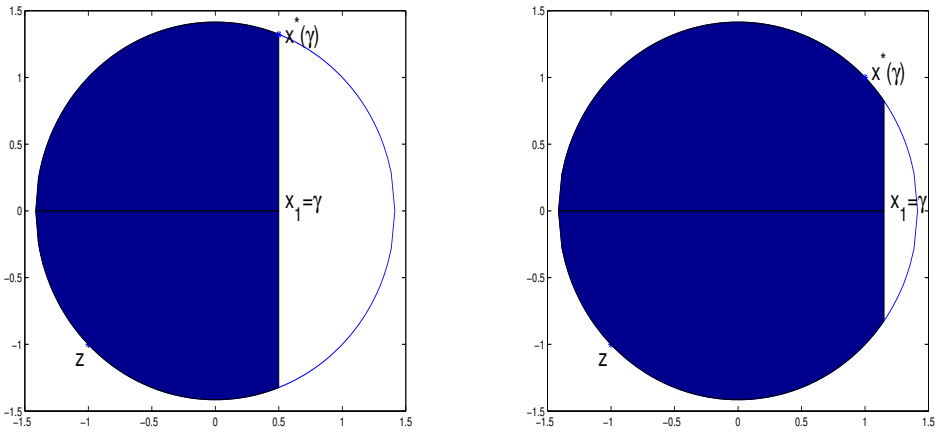


Abbildung 8.6: Menge der zulässigen Lösungen M_γ

Nun betrachten wir noch den Fall $\gamma > 1$, siehe Abbildung 8.6 rechts. In diesem Fall ist $x^* = (1,1)^T$ die Lösung von (P_γ).

(b) Da der Fall $\gamma = -\sqrt{2}$ trivial ist, betrachten wir nur die beiden Fälle $-\sqrt{2} < \gamma \leq 1$ und $\gamma > 1$. Im ersten Fall sind beide Restriktionen aktiv, also $I^* = \{1,2\}$, und

$$\nabla g_1(x^*) = 2 \begin{pmatrix} \gamma \\ \sqrt{2-\gamma^2} \end{pmatrix}, \qquad \nabla g_2(x^*) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Die Constraint Qualification (CQ) in Satz 2.2 ist genau dann erfüllt, wenn $p = (p_1,p_2)^T \in \mathbb{R}^2$ mit

$$\gamma p_1 + \sqrt{2-\gamma^2} p_2 < 0, \qquad p_1 < 0$$

existiert. Dies ist aber offenbar der Fall (setze z. B. $p_1 = -1$ und $p_2 < \gamma/\sqrt{2-\gamma^2}$). Im zweiten Fall ist nur die erste Restriktion aktiv, also $I^* = \{1\}$, und

$$\nabla g_1(x^*) = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

Auch in diesem Fall ist die Regularitätsbedingung (CQ) in Satz 2.2 erfüllt.

(c) Wegen des Erfülltseins der Regularitätsbedingung und Satz 2.2 ist klar, dass ein $u^* \in \mathbb{R}^2$ rexistiert derart, dass (x^*, u^*) ein Kuhn-Tucker-Punkt zu (P_γ) ist. Wir wollen den Multiplikator u^* ausrechnen. Im ersten Fall ist $u^* = (u_1^*, u_2^*)^T$ zu bestimmen aus

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

bzw.

$$\begin{pmatrix} 2\gamma & 1 \\ 2\sqrt{2-\gamma^2} & 0 \end{pmatrix} \begin{pmatrix} u_1^* \\ u_2^* \end{pmatrix} = \begin{pmatrix} 2(\gamma+1) \\ 2(\sqrt{2-\gamma^2}+1) \end{pmatrix}.$$

Die Lösung dieses linearen Gleichungssystems ist

$$u_1^* = 1 + \frac{1}{\sqrt{2-\gamma^2}}, \qquad u_2^* = 2\Big(1 - \frac{\gamma}{\sqrt{2-\gamma^2}}\Big).$$

Wegen $-\sqrt{2} < \gamma \leq 1$ sind u_1^* und u_2^* nichtnegativ. Im zweiten Fall ist $u_2^* = 0$ und u_1^* ist zu bestimmen aus

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

bzw.

$$\begin{pmatrix} -2 \\ -2 \end{pmatrix} + u_1^* \begin{pmatrix} 2 \\ 2 \end{pmatrix} = 0,$$

was auf $u_1^* = 1$ führt. Damit hat man in beiden Fällen ein Kuhn-Tucker-Paar gefunden.

6. Als Hoffman-Theorem (siehe A. J. Hoffman (1952)) 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 $(y_+)_i = \max(y_i, 0)$, $i = 1, \ldots, l$.

Sei

$$P := \{x \in \mathbb{R}^n : Ax \leq b,\ Cx = d\} \neq \emptyset,$$

wobei $A \in \mathbb{R}^{l \times n}$, $b \in \mathbb{R}^l$, $C \in \mathbb{R}^{m \times n}$, $d \in \mathbb{R}^m$. Dann existiert eine (von b und d unabhängige) 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 bedeute $\|\cdot\|$ jeweils die euklidische Norm auf dem entsprechenden Raum.

Lösung: Für eine Indexmenge $I \subset \{1, \ldots, l\}$ 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 $I \subset \{1, \ldots, l\}$, $N_I := \{x \in \mathbb{R}^n : A_I x \geq 0,\ Cx = 0\}$. Sei

$$N_I^+ := \{y \in \mathbb{R}^n : x^T y \geq 0 \ \text{ für alle } x \in N_I\}.$$

Dann existiert eine Konstante $d_I > 0$ mit

$$d_I \|y\| \leq \left\| \begin{pmatrix} (A_I y)_+ \\ Cy \end{pmatrix} \right\| \qquad \text{für alle } y \in N_I^+.$$

Wir können annehmen, dass $N_I^+ \neq \{0\}$, 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 $-\|y\|^2 \geq 0$ 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

$$u(z) \geq 0, \qquad x(z) - z + A^T u(z) + C^T v(z) = 0, \qquad u(z)^T (Ax(z) - b) = 0.$$

Mit $I(z) \subset \{1, \ldots, l\}$ werde die Indexmenge der in $x(z)$ aktiven Ungleichungsrestriktionen bezeichnet. Es ist also

$$u_{I(z)} \geq 0, \qquad x(z) - z + A_{I(z)}^T u_{I(z)} + C^T v(z) = 0.$$

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 $A_{I(z)} x \geq 0$ und $Cx = 0$) ist

$$x^T(z - x(z)) = x^T [A_{I(z)}^T u_{I(z)} + C^T v(z)] = \underbrace{(A_{I(z)} x)}_{\geq 0}{}^T \underbrace{u_{I(z)}}_{\geq 0} + \underbrace{(Cx)}_{=0}{}^T v(z) \geq 0.$$

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.

7. 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 $b \in \mathrm{Bild}\,(A)$ ein $x^* \in \mathbb{R}^n$ mit $Ax^* = b$ und $\|x^*\| \leq c_0 \, \|b\|$ gibt.

Lösung: Mit vorgegebenem $b \in \text{Bild}\,(A)$ sei $P_b := \{x \in \mathbb{R}^n : Ax = b\}$. Wegen des Hoffman-Theorems existiert eine (von b unabhängige) Konstante $c_0 = c_0(A)$ mit

$$\text{dist}(z, P_b) \le c_0 \,\|Az - b\| \qquad \text{für alle } z \in \mathbb{R}^n \text{ und alle } b \in \text{Bild}\,(A).$$

Setzt man hier nun $z := 0$, so erhält man die Behauptung.

8. Mit Hilfe des Hoffman-Theorems zeige man: Gegeben sei das lineare Programm

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

Hierbei seien $A \in \mathbb{R}^{l \times n}$, $b \in \mathbb{R}^l$ und $c \in \mathbb{R}^n$. Es wird $M \neq \emptyset$ und $\inf(\text{P}) > -\infty$ vorausgesetzt. Also ist die Menge M_{opt} der Lösungen von (P) nichtleer. Man zeige die Existenz einer Konstanten $c_0 = c_0(A, c) > 0$ derart, dass

$$\text{dist}(x, M_{\text{opt}}) \le c_0[c^T x - \min(\text{P})] \qquad \text{für alle } x \in M.$$

Hinweis: Man beachte, dass $M_{\text{opt}} = M \cap \{x \in \mathbb{R}^n : c^T x - \min(\text{P}) = 0\}$.

Lösung: Da (P) lösbar, ist $M_{\text{opt}} = \{x \in \mathbb{R}^n : Ax \le b,\ c^T x = \min(\text{P})\}$ nichtleer. Das Hoffman-Theorem liefert die Existenz einer Konstanten $c_0 = c_0(A, c)$ mit

$$\text{dist}(x, M_{\text{opt}}) \le c_0 \left\| \begin{pmatrix} (Ax - b)_+ \\ c^T x - \min(\text{P}) \end{pmatrix} \right\| \qquad \text{für alle } x \in \mathbb{R}^n.$$

Insbesondere ist

$$\text{dist}(x, M_{\text{opt}}) \le c_0 \left\| \begin{pmatrix} 0 \\ c^T x - \min(\text{P}) \end{pmatrix} \right\| = c_0[c^T x - \min(\text{P})] \qquad \text{für alle } x \in M.$$

Damit ist die Aufgabe gelöst.

8.3.3 Aufgaben zu Abschnitt 3.3

1. Gegeben sei das lineare Programm

(P) $\qquad$ Minimiere $\quad c^T x \quad$ auf $\quad M := \{x : x \ge 0,\ b - Ax \le 0\}$.

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 $y \ge 0$ ist der Wert der dualen Zielfunktion daher

$$\phi(y) := \inf_{x \ge 0} L(x, y) = \begin{cases} b^T y, & \text{falls} \quad c - A^T y \ge 0, \\ -\infty, & \text{sonst.} \end{cases}$$

Das zu (P) duale lineare Programm ist daher

(D) $\qquad$ Maximiere $\quad b^T y \quad$ auf $\quad N := \{y : y \ge 0,\ c - A^T y \ge 0\}$.

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. Gegeben sei das quadratische Programm

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

wobei $A \in \mathbb{R}^{l \times n}$, $b \in \mathbb{R}^l$, $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit ist. Es wird vorausgesetzt, dass (P) zulässig, also $M \neq \emptyset$ ist.

(a) Man begründe, weshalb (P) eine eindeutige Lösung $x^* \in M$ besitzt.

(b) Man stelle das zu (P) duale Programm (D) auf und zeige, dass dieses eine Lösung u^* besitzt und $\min(\text{P}) = \max(\text{D})$ gilt.

(c) Man zeige: Ist u^* eine Lösung von (D), so ist $x^* := -Q^{-1}(c + A^T u^*)$ die Lösung von (P).

Hinweis: Zum Nachweis von 3b kann man zeigen, dass ein Lagrange-Multiplikator u^* zur Lösung x^* von (P) eine Lösung von (D) ist.

Lösung:

(a) Die gleichmäßige Konvexität der Zielfunktion sichert, zusammen mit der Zulässigkeit von (P), die Existenz und Eindeutigkeit einer Lösung x^* von (P).

(b) Die Lagrange-Funktion zu (P) ist

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

Die Aufgabe, $L(\cdot, u)$ auf dem $\mathbb{R}^n$ zu minimieren, besitzt eine eindeutige Lösung $x = x(u)$, welche durch

$$0 = \nabla_x L(x(u), u) = c + Qx(u) + A^T u$$

charakterisiert ist und daher durch $x(u) = -Q^{-1}(c + A^T u)$ gegeben ist. Nach einfacher Rechnung ist

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

Das zu (P) duale Programm ist daher

$$\text{(D)} \quad \left\{ \begin{array}{c} \text{Maximiere} \quad \phi(u) := -b^T u - \frac{1}{2}(c + A^T u)^T Q^{-1}(c + A^T u) \quad \text{auf} \\ N := \{u \in \mathbb{R}^l : u \geq 0\}. \end{array} \right.$$

Wegen des Kuhn-Tucker-Satzes 2.1 existiert zu der Lösung x^* von (P) ein $u^* \in \mathbb{R}^l$ mit

$$u^* \geq 0, \qquad c + Qx^* + A^T u^* = 0, \qquad (b - Ax^*)^T u^* = 0.$$

Dieses u^* ist dual zulässig und es gilt

$$\begin{aligned} \phi(u^*) &= -b^T u^* - \frac{1}{2}(c + A^T u^*)^T Q^{-1}(c + A^T u^*) \\ &= -(Ax^*)^T u^* - \frac{1}{2}(x^*)^T Q x^* \\ &= (x^*)^T (c + Qx^*) \\ &= c^T x^* + \frac{1}{2}(x^*)^T Q x^* \\ &= f(x^*). \end{aligned}$$

Der schwache Dualitätssatz zeigt, dass u^* eine Lösung von (D) ist.

(c) Ist $u^* \in N$ eine Lösung von (D), so zeigt der Kuhn-Tucker-Satz 2.1 die Existenz eines $z^* \in \mathbb{R}^l$ mit (beachte, dass (D) eine Maximierungsaufgabe ist!)

$$z^* \geq 0, \qquad -\nabla\phi(u^*) - z^* = 0, \qquad (u^*)^T z^* = 0$$

bzw. die Gültigkeit von

$$\nabla\phi(u^*) \leq 0, \qquad (u^*)^T \nabla\phi(u^*) = 0.$$

Mit $x^* := -Q^{-1}(c + A^T u^*)$ ist

$$0 \geq \nabla\phi(u^*) = -b - AQ^{-1}(c + A^T u^*) = -b + Ax^*,$$

und das zeigt, dass $x^* \in M$ zulässig für (P) ist. Weiter ist

$$u^* \geq 0, \qquad c + Qx^* + A^T u^* = 0, \qquad (b - Ax^*)^T u^* = 0.$$

Da die notwendigen Optimalitätsbedingungen erster Ordnung bei einer konvexen Optimierungsaufgabe auch hinreichend für Optimalität sind (siehe Satz 2.5), ist $x^* \in M$ eine Lösung und wegen der Eindeutigkeit sogar *die* Lösung von (P).

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.

   Hinweis: Zum Nachweis von $N_{\mathrm{opt}} \neq \emptyset$ definiere man

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

   zeige, dass Λ_+ konvex (und nichtleer) ist und $(\inf(\mathrm{P}), 0) \notin \Lambda_+$ gilt. Anschließend wende man den Trennungssatz (Satz 1.7) für konvexe Mengen an. Hiernach existiert ein Paar $(q^*, u^*) \in \mathbb{R} \times \mathbb{R}^l \setminus \{(0,0)\}$ mit

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

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

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

   Wegen der Slaterschen Constraint Qualification ist $u^* = 0$, ein Widerspruch zu $(q^*, u^*) \neq (0,0)$. O. B. d. A. können wir dann $q^* = 1$ annehmen und haben

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

   Also ist $u^* \in N$ dual zulässig und $\inf(\mathrm{P}) \leq \phi(u^*)$. Aus dem schwachen Dualitätssatz (Satz 3.1) folgt $u^* \in N_{\mathrm{opt}}$ und $\max(\mathrm{D}) = \inf(\mathrm{P})$.

   Lösung: Der Hinweis liefert $N_{\mathrm{opt}} \neq \emptyset$. 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. Seien $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ gegeben. Die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|Ax - b\|_\infty, \quad x \in \mathbb{R}^n,$$

nennt man das *diskrete, lineare Tschebyscheffsche Approximationsproblem*. Hierbei ist $\|\cdot\|_\infty$ die Maximum- (oder auch Tschebyscheff-) Norm auf dem $\mathbb{R}^m$, also $\|y\|_\infty := \max_{i=1,\ldots,m} |y_i|$. Der Aufgabe (P) ordne man die lineare Optimierungsaufgabe

$$\text{(Q)} \quad \text{Minimiere} \quad g(x,\delta) := \delta \quad \text{auf} \quad M := \{(x,\delta) \in \mathbb{R}^n \times \mathbb{R} : -\delta e \leq Ax - b \leq \delta e\}$$

zu, wobei e der Vektor des $\mathbb{R}^m$ ist, dessen Komponenten sämtlich gleich 1 sind.

(a) Man zeige: Ist $x^* \in \mathbb{R}^n$ eine Lösung von (P), so ist $(x^*, \|Ax^* - b\|_\infty)$ eine Lösung von (Q). Ist umgekehrt $(x^*, \delta^*) \in M$ eine Lösung von (Q), so ist x^* eine Lösung von (P) und $\delta^* = \|Ax^* - b\|_\infty$.

(b) Man begründe, weshalb (Q) und damit auch (P) mindestens eine Lösung besitzt.

(c) Man führe die lineare Optimierungsaufgabe (Q) in Standardform über und berechne die hierzu duale Optimierungsaufgabe.

Lösung: Sei $x^* \in \mathbb{R}^n$ eine Lösung von (P). Dann ist $(x^*, \|Ax^* - b\|_\infty) \in M$, für beliebiges $(x,\delta) \in M$ ist ferner

$$\delta^* := \|Ax^* - b\|_\infty \leq \|Ax - b\|_\infty \leq \delta.$$

Daher ist $(x^*, \|Ax^* - b\|_\infty)$ eine Lösung von (Q). Ist umgekehrt $(x^*, \delta^*) \in M$ eine Lösung von (Q) und $x \in \mathbb{R}^n$ beliebig, so ist $(x, \|Ax - b\|_\infty) \in M$ und daher

$$\|Ax^* - b\|_\infty \leq \delta^* \leq \|Ax - b\|_\infty.$$

Hieraus liest man ab, dass x^* eine Lösung von (P) ist, setzt man $x = x^*$ so folgt auch $\delta^* = \|Ax^* - b\|_\infty$.

(Q) ist eine lineare Optimierungsaufgabe, die zulässig ist (für ein beliebiges $x \in \mathbb{R}^n$ ist $(x, \|Ax - b\|_\infty) \in M$) und deren Optimalwert nach unten durch 0 beschränkt ist. Der Existenzsatz der linearen Optimierung zeigt, dass (Q) und damit auch (P) lösbar ist.

Ist $(x,\delta) \in M$, so ist notwendig $\delta \geq 0$. Wir stellen daher nur x als Differenz nichtnegativer Variabler x_+, x_- dar und führen zwei Schlupfvariable $y, z \in \mathbb{R}^m$

ein. Als äquivalentes Problem in Standardform erhalten wir (man hat hier diverse Möglichkeiten, wir ziehen nichtnegative Schlupfvariablen ab, damit die dualen Variablen automatisch nichtnegativ sind)

$$\begin{cases} \text{Minimiere} \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} x_+ \\ x_- \\ y \\ z \\ \delta \end{pmatrix} \quad \text{unter den Nebenbedingungen} \\ \begin{pmatrix} x_+ \\ x_- \\ y \\ z \\ \delta \end{pmatrix} \geq 0, \quad \begin{pmatrix} -A & A & -I & 0 & e \\ A & -A & 0 & -I & e \end{pmatrix} \begin{pmatrix} x_+ \\ x_- \\ y \\ z \\ \delta \end{pmatrix} = \begin{pmatrix} -b \\ b \end{pmatrix}. \end{cases}$$

Das hierzu duale lineare Programm ist

$$\begin{cases} \text{Maximiere} \quad \begin{pmatrix} -b \\ b \end{pmatrix}^T \begin{pmatrix} u \\ v \end{pmatrix} \quad \text{auf} \\ N := \{(u,v) \in \mathbb{R}^m : u, v \geq 0,\ A^T(u-v) = 0,\ e^T(u+v) = 1\}. \end{cases}$$

8.4 Aufgaben zu Kapitel 4

8.4.1 Aufgaben zu Abschnitt 4.1

1. Sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Die Menge $M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax \geq b\}$ sei nichtleer. Man zeige, dass M genau dann beschränkt ist, wenn es ein $u \in \mathbb{R}^m$ mit $u \geq 0$ und $A^T u < 0$ gibt.

   Lösung: Wir nehmen zunächst an, es sei M beschränkt. Dann gibt es kein $p \in \mathbb{R}^n \setminus \{0\}$ mit $Ap \geq 0$, $p \geq 0$ (da mit einem $x \in M$ andernfalls $x + tp \in M$ für alle $t \geq 0$ wäre, ein Widerspruch zur Beschränktheit von M). Dies impliziert, dass das System $Ap \geq 0$, $p \geq 0$, $-e^T p < 0$ (hierbei ist e der Vektor, dessen Komponenten alle gleich 1 sind) nicht lösbar ist. Das Farkas-Lemma wiederum zeigt, dass es ein $u \in \mathbb{R}^m$ mit $-e - A^T u \geq 0$, $u \geq 0$ gibt, was zu zeigen war.

   Nun existiere ein $u \in \mathbb{R}^m$ mit $u \geq 0$ und $A^T u < 0$. Im Widerspruch zur Behauptung nehmen wir an, die Menge M sei nicht beschränkt. Dann existiert eine Folge $\{x_k\} \subset M$ mit $\|x_k\| \to \infty$. Die Folge $\{p_k\}$ mit $p_k := x_k/\|x_k\|$ besitzt einen Häufungspunkt p, für den offenbar $p \neq 0$ und $p \geq 0$, $Ap \geq 0$ gilt. Dann wäre $(A^T u)^T p$ als Skalarprodukt eines negativen Vektors mit einem nichtverschwindenden nichtnegativen Vektor einerseits negativ, andererseits ist $(A^T u)^T p = u^T Ap$ als Skalarprodukt zweier nichtnegativer Vektoren nichtnegativ. Das ist der gewünschte Widerspruch.

2. 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 $Q_0, Q_1, \ldots, Q_l \in \mathbb{R}^{n \times n}$, Vektoren $c_0, c_1, \ldots, c_l \in \mathbb{R}^n$, $\beta_1, \ldots, \beta_l \in \mathbb{R}$ sowie $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Wir setzen voraus, dass (P) zulässig bzw. $M \neq \emptyset$ ist. Man zeige, dass die Menge M_{opt} der Lösungen von (P) genau dann nichtleer und kompakt ist, 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.

Hinweis: Es darf der (hier unbewiesene) Existenzsatz für konvexe, quadratisch restringierte quadratische Programme (Satz 3.5 in Abschnitt 3.3) benutzt werden.

Lösung: Zunächst nehmen wir an, dass das System $(*)$ nur trivial lösbar sei. Wir zeigen, dass

$$\inf(\mathrm{P}) := \inf_{x \in M} f(x) > -\infty.$$

Da (P) als zulässig vorausgesetzt ist, folgt dann aus dem im Hinweis angegebenen Satz $M_{\text{opt}} \neq \emptyset$. Wäre $\inf(\mathrm{P}) = -\infty$, so existierte eine Folge $\{x_k\} \subset M$ mit $\|x_k\| \to \infty$ und $f(x_k) \to -\infty$. Wir können annehmen (notfalls gehe man zu einer Teilfolge über), dass die Folge $\{p_k\}$ mit $p_k := x_k / \|x_k\|$ gegen einen Vektor p konvergiert, der natürlich vom Nullvektor verschieden ist. Für alle hinreichend großen k ist

$$\frac{f(x_k)}{\|x_k\|^2} = \frac{1}{\|x_k\|} c_0^T p_k + \frac{1}{2} p_k^T Q_0 p_k \to \frac{1}{2} p^T Q_0 p \leq 0.$$

Mit $k \to \infty$ folgt $p^T Q_0 p \leq 0$, hieraus $p^T Q_0 p = 0$ (da Q_0 positiv semidefinit ist) und $Q_0 p = 0$ (Beweis?). Für alle hinreichend großen k ist ebenso

$$0 \geq \frac{f(x_k)}{\|x_k\|} \geq c_0^T p_k \to c_0^T p,$$

also $c_0^T p \leq 0$. Entsprechend folgt aus $g_i(x_k) \leq 0$, dass $c_i^T p \leq 0$ und $Q_i p = 0$. Aus $h(x_k) = Ax_k - b = 0$ folgt nach Division mit $\|x_k\|$ und Grenzübergang $k \to /infty$, dass $Ap = 0$. Im Widerspruch zur Annahme ist also das Gleichungs-Ungleichungssystem $(*)$ nichttrivial lösbar. Damit ist $M_{\text{opt}} \neq \emptyset$ bewiesen. Da M_{opt} natürlich abgeschlossen ist, bleibt für die Kompaktheit von M_{opt} zu zeigen, dass M_{opt} beschränkt ist. Angenommen, dies wäre nicht der Fall. Dann existiert eine Folge $\{x_k\} \subset M_{\text{opt}}$ mit $\|x_k\| \to \infty$. Wieder kann angenommen werden, dass die Folge $\{p_k\}$ mit $p_k := x_k / \|x_k\|$ gegen ein p konvergiert, wobei natürlich $p \neq 0$. Fast genau (wir nutzen aus, dass $f(x_k) = \min(\mathrm{P})$) wie eben erhält man, dass p eine nichttriviale Lösung von $(*)$ ist, ein Widerspruch.

Nun setzen wir voraus, M_{opt} sei nichtleer und kompakt. Sei p eine Lösung von $(*)$. Mit einem $x^* \in M_{\text{opt}}$ ist dann, wie wir gleich sehen werden, $x^* + tp \in M_{\text{opt}}$

für alle $t \geq 0$. Wegen der vorausgesetzten Beschränktheit von M_{opt} folgt $p = 0$, das System $(*)$ ist also nur trivial lösbar. Zunächst zeigen wir, dass $x^* + tp$ für alle $t \geq 0$ zulässig für (P) ist. Für alle $t \geq 0$ und $i = 1, \ldots, l$ ist

$$\begin{aligned} g_i(x^* + tp) &= g_i(x^*) + t\nabla g_i(x^*)^T p + \frac{1}{2}t^2 p^T \nabla^2 g_i(x^*) p \\ &= \underbrace{g_i(x^*)}_{\leq 0} + t(c_i + Q_i x^*)^T p + \frac{1}{2}t^2 p^T \underbrace{Q_i p}_{=0} \\ &\leq t[\underbrace{c_i^T p}_{\leq 0} + (x^*)^T \underbrace{Q_i p}_{=0}] \\ &\leq 0. \end{aligned}$$

Da außerdem

$$h(x^* + tp) = A(x^* + tp) - b = \underbrace{Ax^*}_{=b} + t\underbrace{Ap}_{=0} - b = 0,$$

ist $x^* + tp \in M$ für alle $t \geq 0$. Weiter ist

$$\begin{aligned} f(x^* + tp) &= f(x^*) + t\nabla f(x^*)^T p + \frac{1}{2}t^2 p^T \nabla^2 f(x^*) p \\ &= \min(\text{P}) + t(c_0 + Q_0 x^*)^T p + \frac{1}{2}t^2 p^T Q_0 p \\ &= \min(\text{P}) + t\underbrace{c_0^T p}_{\leq 0} \\ &\leq \min(\text{P}) \end{aligned}$$

für alle $t \geq 0$. Hieraus folgt $x^* + tp \in M_{\text{opt}}$ für alle $t \geq 0$ und damit die Behauptung.

8.4.2 Aufgaben zu Abschnitt 4.2

1. **Programmieraufgabe:** Man programmiere einen Schritt des unzulässigen primal-dualen Innere-Punkt-Verfahrens für ein lineares Programm in Standardform mit den Daten (A, b, c). Anschließend wende man das Verfahren (indem man etwa 10 Schritte durchführt) auf ein Beispiel mit den Daten

$$\begin{array}{|c|c|}\hline c^T & \\ \hline A & b \\ \hline \end{array} := \begin{array}{|rrrrrrr|r|}\hline 5 & 3 & 3 & 6 & 0 & 0 & 0 & \\ \hline -6 & 1 & 2 & 4 & 1 & 0 & 0 & 14 \\ 3 & -2 & -1 & -5 & 0 & 1 & 0 & -25 \\ -2 & 1 & 0 & 2 & 0 & 0 & 1 & 14 \\ \hline \end{array}$$

an, wobei man mit $(x, y, z) := (e, 0, e)$ starte. Insbesondere beobachte man, wie sich der Defekt

$$d(x, y, z) := \frac{\|Ax - b\|}{\max(1, \|b\|)} + \frac{\|A^T y + z - c\|}{\max(1, \|c\|)} + \frac{|c^T x - b^T y|}{\max(1, |c^T x|, |b^T y|)}$$

verändert. Man sollte verschiedene Strategien bei der Wahl von σ ausprobieren, etwa $\sigma_k := 0.5$, $\sigma_k := 1/(k+1)$ und $\sigma_k := 1/(k+1)^2$.

Lösung: Wir haben in MATLAB eine sehr einfache Funktion geschrieben, durch die ein Schritt des primal-dualen Innere-Punkt-Verfahrens durchgeführt wird. Hierbei benutzen wir im wesentlichen die Bezeichnungen aus der Vorlesung und verzichten auf alle programmtechnischen Feinheiten.

```
function [x_p,y_p,z_p,def]=primdual(A,b,c,x,y,z,sigma);
%*****************************************************************
%Es wird ein Schritt eines unzulaessigen primal-dualen Innere Punkt
%Verfahrens durchgef"uhrt. Es werden die Bezeichnungen der Vorlesung
%benutzt.
%Input:    Problem-Daten (A,b,c), ferner Tripel (x,y,z) mit x>0,z>0
%          und sigma>0.
%Output:   Neues N"aherungstripel (x_p,y_p,z_p) und Defekt def zu
%          Ausgangstripel (x,y,z)
%*****************************************************************
[m,n]=size(A);tau_p=0.995;tau_d=0.995;mu=sigma*x'*z/n;
X=diag(x);Z=diag(z);D=diag(x./z);e=ones(size(x));
%Berechne Residuen:
r_p=A*x-b; r_d=A'*y+z-c; r_xz=X*z-mu*e;
%Berechne Defekt:
def=norm(r_p)/max(1,norm(b))+norm(r_d)/max(norm(c))+...
      abs(c'*x-b'*y)/max([1;abs(c'*x);abs(b'*y)]);
%Berechne Newton-Richtungen:
q=-A*D*A'\(r_p+A*D*(r_d-r_xz./x));
r=-A'*q-r_d;  p=-D*r-r_xz./z;
%Berechne primale Schrittweite:
P=find(p<0);if isempty(P), alpha_p=1; else
alpha_pmax=min(-x(P)./p(P));alpha_p=min(1,tau_p*alpha_pmax);end;
%Berechne duale Schrittweite:
R=find(r<0);if isempty(R), alpha_d=1; else
alpha_dmax=min(-z(R)./r(R));alpha_d=min(1,tau_d*alpha_dmax);end;
%Berechne neue N"aherung:
x_p=x+alpha_p*p;y_p=y+alpha_d*q;z_p=z+alpha_d*r;
%*****************************************************************
```

In der Tabelle 8.1 (wir benutzen `format long`) geben wir für 10 Iterationen bei Wahl verschiedener σ-Strategien den erhaltenen Defekt an: Man erkennt sehr deutlich, dass die Folge $\{\sigma_k\}$ hinreichend schnell gegen Null konvergieren sollte. Es wäre sehr interessant, hierfür genauere theoretische Begründungen zu geben. Weiter müsste man sich über die Wahl einer geeigneten Startnäherung (mehr) Gedanken machen.

2. Formulieren Sie die folgende Aufgabenstellung aus dem praktischen Leben als lineare Optimierungsaufgabe und führen Sie diese durch Einführung von Schlupfvariablen auf Normal- bzw. Standardform über. Wenn Aufgabe 1 erfolgreich gelöst

$\sigma_k = 1/(k+1)^2$	$\sigma_k = 1/(k+1)$	$\sigma_k = 0.5$
2.657961204783823	2.657961204783823	2.657961204783823
1.279135898658486	1.237228789828415	1.237228789828415
0.951738073993734	0.897453424125422	0.877199287357384
0.319096617458393	0.196882471958517	0.087000879748754
0.000278928125730	0.001212857148388	0.007012670631776
0.000007749548295	0.000202031216207	0.003512756739985
0.000000158154734	0.000028864455854	0.001758154269761
0.000000002471168	0.000003608109094	0.000879518919492
0.000000000030508	0.000000400901751	0.000439869995117
0.000000000000307	0.000000040090182	0.000219962641834

Tabelle 8.1: Defekt bei verschiedenen Strategien

wurde, können Sie anschließend die gesuchte Lösung mit Hilfe des primal-dualen Innere-Punkt-Verfahrens berechnen.

Sie wollen Ihrer Tante (vielleicht eine reiche Erbtante?) zum Geburtstag eine Freude machen. Ihre Tante trinkt gerne einen süßen Wein und da Ihnen eine Beerenauslese zu teuer ist, kommen Sie auf die Idee, ihr einen Liter Wein zukommen zu lassen, den Sie selbst zusammengestellt haben.

Hierzu können Sie einen Landwein für €1.00 pro Liter, zur Anhebung der Süße Diäthylenglykol-haltiges Frostschutzmittel für €1.20 pro Liter und für eine Verbesserung der Lagerungsfähigkeit eine Natriumacid-Lösung für €1.80 pro Liter kaufen. Verständlicherweise wollen Sie eine möglichst billige Mischung herstellen, wobei allerdings folgende Nebenbedingungen zu beachten sind: Um eine hinreichende Süße zu garantieren, muss die Mischung mindestens 1/3 Frostschutzmittel enthalten. Andererseits muss (z. B. wegen gesetzlicher Bestimmungen) mindestens halb so viel Wein wie Frostschutzmittel enthalten sein. Der Natriumacid-Anteil muss mindestens halb so groß, darf aber andererseits höchstens so groß wie der Glykol-Anteil sein und darf die Hälfte des Weinanteils nicht unterschreiten.

Lösung: Die gesuchte Mischung bestehe aus x_1 Liter Frostschutzmittel, x_2 Liter Natriumacid-Lösung und x_3 Liter Wein. Für eine "zulässige" Mischung erhalten wir die Nebenbedingungen

$$x_1 + x_2 + x_3 = 1, \qquad x_1 \geq 0, \quad x_2 \geq 0, \quad x_3 \geq 0$$

sowie

$$x_1 \geq \frac{1}{3}, \quad x_3 \geq \frac{1}{2}x_1, \quad \frac{1}{2}x_1 \leq x_2 \leq x_1, \quad \frac{1}{2}x_3 \leq x_2.$$

Die Kosten ergeben sich zu $1.2x_1 + 1.8x_2 + x_3$ €. Zu lösen ist also die Optimie-

rungsaufgabe

$$\text{Minimiere} \quad \tfrac{6}{5}x_1 + \tfrac{9}{5}x_2 + x_3 \quad \text{unter den Nebenbedingungen}$$

$$\begin{array}{rcrcrcl} x_1 & & & - & 2x_3 & \leq & 0, \\ x_1 & - & 2x_2 & & & \leq & 0, \\ -x_1 & + & x_2 & & & \leq & 0, \\ & - & 2x_2 & + & x_3 & \leq & 0, \\ 3x_1 & & & & & \geq & 1, \\ x_1 & + & x_2 & + & x_3 & = & 1, \end{array} \qquad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Nach Einführung von Schlupfvariablen hat man die Aufgabe

$$\text{Minimiere} \quad c^T x \quad \text{unter den Nebenbedingungen} \quad x \geq 0, \;\; Ax = b$$

mit

$$\begin{array}{|c|c|} \hline c^T & \\ \hline A & b \\ \hline \end{array} := \begin{array}{|cccccccc|c|} \hline \frac{6}{5} & \frac{9}{5} & 1 & 0 & 0 & 0 & 0 & 0 & \\ \hline 1 & 0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline \end{array}$$

zu lösen. Mit dem Startwert $(x, y, z) := (e, 0, e)$ führen wir 12 Schritte des unzulässigen primal-dualen Innere-Punkt-Verfahrens durch, wobei wir (abhängig vom Iterationsschritt) $\sigma_k := 1/(k+1)^2$ wählen. Als Defekt erhalten wir

k	Defekt
1	3.732050807568877
2	1.430904387194137
3	0.160280404809486
4	0.076764875220459
5	0.014608881511480
6	0.002948192997206
7	0.000324674452536
8	0.000005073263606
9	0.000000062632950
10	0.000000000626330
11	0.000000000005176
12	0.000000000000037

Nach 12 Iterationen hat man als Näherungen für die Lösung x^* erhalten:

$$x^* = \begin{pmatrix} 0.399999999999999 \\ 0.200000000000000 \\ 0.400000000000000 \\ 0.400000000000002 \\ 0.000000000000001 \\ 0.199999999999999 \\ 0.000000000000000 \\ 0.199999999999999 \end{pmatrix}.$$

Die kostenminimale Mischung enthält $\frac{2}{5}$ Liter Frostschutzmittel, $\frac{1}{5}$ Liter Natriumacid-Lösung und $\frac{2}{5}$ Liter Wein., das Geburtstagspräsent kostet €1.24.

8.5 Aufgaben zu Kapitel 5

8.5.1 Aufgaben zu Abschnitt 5.1

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

$$N := (AQ^{-1}A^T)^{-1}AQ^{-1}, \qquad H := Q^{-1}(I - A^TN)$$

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.

$$Qu + A^Tv = 0, \qquad Au = 0.$$

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

$$HQ + N^T A = (I - Q^{-1}A^T(AQ^{-1}A^T)^{-1}A) + Q^{-1}A^T(AQ^{-1}A^T)^{-1}A = I$$

und

$$NQ - NQN^T A = (AQ^{-1}A^T)^{-1}A - (AQ^{-1}A^T)^{-1}AQ^{-1}A^T(AQ^{-1}A^T)^{-1}A = 0.$$

Damit ist die Aufgabe gelöst.

2. **Programmieraufgabe:** Gegeben sei das durch lineare Gleichungen restringierte quadratische Programm

   (P) Minimiere $f(x) := c^T x + \frac{1}{2}x^T Qx$ auf $M := \{x \in \mathbb{R}^n : Ax = b\}$.

   Hierbei seien $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und die symmetrische Matrix $Q \in \mathbb{R}^{n\times n}$ gegeben, die auf Kern (A) positiv definit sei. Zur Lösung von (P) bzw. des linearen Gleichungssystems

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}$$

   wurde in Unterabschnitt 5.1.1 das folgende Verfahren angegeben:

   - Berechne eine QR-Zerlegung von A^T, also eine Darstellung

$$A^T = Z\begin{pmatrix} R \\ 0 \end{pmatrix}\begin{matrix} \} \, m \\ \} \, n-m \end{matrix}$$

     mit einer orthogonalen Matrix $Z \in \mathbb{R}^{n\times n}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{m\times m}$. Man denke sich Z durch

$$Z = (\underbrace{Z^{(1)}}_{m} \quad \underbrace{Z^{(2)}}_{n-m})$$

     partitioniert.

   - Berechne $x^{(1)} \in \mathbb{R}^m$ durch Vorwärtseinsetzen aus

$$R^T x^{(1)} = b.$$

   - Berechne

$$\begin{pmatrix} c^{(1)} \\ c^{(2)} \end{pmatrix} := Z^T c, \qquad \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} := Z^T QZ.$$

   - Berechne $x^{(2)} \in \mathbb{R}^{n-m}$ als Lösung von

$$\underbrace{(Z^{(2)})^T QZ^{(2)}}_{=B_{22}} x^{(2)} = -\underbrace{(Z^{(2)})^T c}_{c^{(2)}} - \underbrace{(Z^{(2)})^T QZ^{(1)}}_{B_{21}} x^{(1)}.$$

- Berechne
$$x := Z \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix} = Z^{(1)}x^{(1)} + Z^{(2)}x^{(2)}.$$

- Berechne $y \in \mathbb{R}^m$ durch Rückwärtseinsetzen aus
$$Ry = -\underbrace{(Z^{(1)})^T c}_{c^{(1)}} - \underbrace{(Z^{(1)})^T Q Z^{(1)}}_{B_{11}} x^{(1)} - \underbrace{(Z^{(1)})^T Q Z^{(2)}}_{B_{12}} x^{(2)}.$$

Man implementiere dieses Verfahren und erprobe die Implementation an der Aufgabe mit den Daten

$$A := \begin{pmatrix} 1 & 3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & -2 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix}, \quad b := \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad c := \begin{pmatrix} 0 \\ -2 \\ -2 \\ -1 \\ -1 \end{pmatrix}$$

sowie

$$Q := \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Lösung: Wir schreiben die folgende einfache Funktion:

```
function [x,y]=Quad_Gleich(Q,A,b,c);
%************************************************************
%Input-Parameter:
%         Q    symmetrische, positiv definite n x n-Matrix
%         A    m x n-Matrix mit Rang(A)=m
%         b    m-Vektor
%         c    n-Vektor
%Output-Parameter:
%         x,y  Loesung von
%                (Q   A')(x)  (-c)
%                (      )    =
%                (A   0 )(y)  (b)
%************************************************************
[m,n]=size(A);[Z,R]=qr(A');R=R(1:m,:);
x_1=R'\b;
a=Z'*c;c_1=a(1:m);c_2=a(m+1:n);
B=Z'*Q*Z;B_11=B(1:m,1:m);B_12=B(1:m,m+1:n);
B_21=B(m+1:n,1:m);B_22=B(m+1:n,m+1:n);
x_2=B_22\(-c_2-B_21*x_1);
x=Z*[x_1;x_2];y=R\(-c_1-B_11*x_1-B_12*x_2);
%************************************************************
```

Für die angegebenen Daten erhalten wir

$$x^* = \begin{pmatrix} -0.767441860465116 \\ 0.255813953488372 \\ 0.627906976744187 \\ -0.116279069767442 \\ 0.255813953488372 \end{pmatrix}, \qquad y^* = \begin{pmatrix} 1.023255813953488 \\ 1.116279069767441 \\ -2.976744186046510 \end{pmatrix}.$$

Dies stimmt vorzüglich mit der exakten Lösung

$$x^* = \frac{1}{43}\begin{pmatrix} -33 \\ 11 \\ 27 \\ -5 \\ 11 \end{pmatrix}, \qquad y^* = \frac{1}{43}\begin{pmatrix} 44 \\ 50 \\ -128 \end{pmatrix}$$

überein.

3. Bei gegebenen $n \in \mathbb{N}$ und $K > 0$ bestimme man[5] die Lösung der Aufgabe

$$\text{Minimiere} \quad f(x) := \tfrac{1}{2}\sum_{j=1}^{n} j x_j^2 \quad \text{auf} \quad M := \Big\{x \in \mathbb{R}^n : \sum_{j=1}^{n} x_j = K\Big\}.$$

Lösung: Wegen der notwendigen Optimalitätsbedingung erster Ordnung existiert zu der eindeutigen Lösung $x^* \in M$ ein $\lambda^* \in \mathbb{R}$ mit

$$j x_j^* - \lambda^* = 0, \qquad j = 1, \ldots, n.$$

Also ist $x_j^* = (1/j)\lambda^*$, $j = 1, \ldots, n$. Aus $\sum_{k=1}^{n} x_k^* = K$ erhält man $\lambda^* = K/(\sum_{k=1}^{n}(1/k))$, so dass die Lösung x^* durch

$$x_j^* = K/(\sum_{k=1}^{n}(1/k))\frac{1}{j}, \qquad j = 1, \ldots, n,$$

gegeben ist.

4. Man wende das primale Verfahren von Fletcher auf das quadratische Programm

$$\left\{\begin{array}{c} \text{Minimiere} \quad \begin{pmatrix} -3 \\ 0 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \tfrac{1}{2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \text{unter der Nebenbedingung} \\ \begin{pmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \le \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \end{array}\right.$$

an. Hierbei starte man mit $(x_0, I_0) := ((0,0)^T, \{1,2\})$.

[5] Siehe R. Fletcher (1987, S. 255).

Lösung: Zunächst haben wir das lineare Gleichungssystem

$$\left(\begin{array}{rr|rr} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ \hline -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{array}\right) \left(\begin{array}{c} p \\ \hline y_{I_0} \end{array}\right) = \left(\begin{array}{c} 3 \\ 0 \\ \hline 0 \\ 0 \end{array}\right)$$

zu lösen. Dies ergibt $p = 0$, $y_{I_0} = (-3, 0)^T$. Dann ist $x_1 := x_0$ und die Restriktion $l = 1$ wird aus I_0 entfernt, es ist also $I_1 := \{2\}$. Dann ist das Gleichungssystem

$$\left(\begin{array}{rr|r} 2 & -1 & 0 \\ -1 & 2 & -1 \\ \hline 0 & -1 & 0 \end{array}\right) \left(\begin{array}{c} p \\ \hline y_{I_1} \end{array}\right) = \left(\begin{array}{c} 3 \\ 0 \\ \hline 0 \end{array}\right)$$

zu lösen. Man erhält $p = (\frac{3}{2}, 0)^T$, $y_{I_1} = (-\frac{3}{2})$. Es ist $x_2 := x_1 + p = (\frac{3}{2}, 0)^T$ zulässig und $I_2 := \emptyset$. Anschließend ist

$$\left(\begin{array}{rr} 2 & -1 \\ -1 & 2 \end{array}\right) \left(\begin{array}{c} p \end{array}\right) = \left(\begin{array}{c} 0 \\ \frac{3}{2} \end{array}\right)$$

zu lösen. Man erhält $p = (\frac{1}{2}, 1)^T$. Da $x_2 + p = (2, 1)^T$ nicht zulässig ist, hat man die maximale Schrittweite zu berechnen. Es ist $s(x_2, p) = \frac{1}{3}$ und daher $x_3 := x_2 + s(x_2, p)p = (\frac{5}{3}, \frac{1}{3})^T$, $I_3 := \{3\}$. Das nächste zu lösende lineare Gleichungssystem ist

$$\left(\begin{array}{rr|r} 2 & -1 & 1 \\ -1 & 2 & 1 \\ \hline 1 & 1 & 0 \end{array}\right) \left(\begin{array}{c} p \\ \hline y_{I_3} \end{array}\right) = \left(\begin{array}{c} 0 \\ 1 \\ \hline 0 \end{array}\right).$$

Man erhält $p = (-\frac{1}{6}, \frac{1}{6})^T$ und $y_{I_3} = (\frac{1}{2})$. Da $x_4 := x_3 + p = (\frac{3}{2}, \frac{1}{2})^T$ zulässig und $y_{I_3} \geq 0$, ist x_4 die gesuchte Lösung.

5. In dem Polyeder

$$P := \{x \in \mathbb{R}^3 : x_1 + 2x_2 - x_3 \geq 4,\ -x_1 + x_2 - x_3 \leq 2\}$$

bestimme man den Punkt, der den kleinsten euklidischen Abstand zum Nullpunkt im $\mathbb{R}^3$ besitzt[6].

Lösung: Zu lösen ist das quadratische Programm

$$\left\{\begin{array}{l} \text{Minimiere} \quad \frac{1}{2}\|x\|_2^2 \quad \text{unter der Nebenbedingung} \\ \qquad \left(\begin{array}{rrr} -1 & -2 & 1 \\ -1 & 1 & -1 \end{array}\right) \left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) \leq \left(\begin{array}{r} -4 \\ 2 \end{array}\right). \end{array}\right.$$

Wir zeigen, dass $x^* = (\frac{2}{3}, \frac{4}{3}, -\frac{2}{3})^T$ der gesuchte Punkt ist. Dies wollen wir mit Hilfe der notwendigen und hinreichenden Optimalitätsbedingungen erster Ordnung

[6]Siehe R. Fletcher (1987, S. 257).

nachweisen. Nur die erste Restriktion ist aktiv. Daher ist x^* der gesuchte Punkt, wenn ein $u_1^* \geq 0$ mit

$$0 = x^* + u_1^* \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ \frac{4}{3} \\ -\frac{2}{3} \end{pmatrix} + u_1^* \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix}$$

existiert. Mit $u_1^* = \frac{2}{3}$ ist das der Fall.

6. Gegeben sei die Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := \frac{1}{2}\|x - z\|^2 \quad \text{unter der Nebenbedingung} \\ \qquad\qquad\qquad g(x) := \frac{1}{2}x^T Ax + b^T x - \alpha \leq 0. \end{cases}$$

Hierbei sind $\alpha \in \mathbb{R}$, $z, b \in \mathbb{R}^n$ und die symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ vorgegeben, ferner ist $\|\cdot\|$ die euklidische Norm im $\mathbb{R}^n$. Es sei $g(z) > 0$ bzw. $z \notin M$ (andernfalls ist die Aufgabe (P) trivial) und $-\frac{1}{2}b^T A^{-1}b - \alpha < 0$ bzw. $\min_{x\in\mathbb{R}^n} g(x) < 0$. Dies impliziert, dass $M \neq \emptyset$.

(a) Man zeige, dass (P) genau eine Lösung $x^* \in M$ besitzt.

(b) Man zeige, dass $g(x^*) = 0$ und $\nabla g(x^*) \neq 0$. Hieraus schließe man, dass die Lösung $x^* \in M$ von (P) charakterisiert ist durch die Existenz eines $\lambda^* > 0$ mit $x^* - z + \lambda^*(Ax^* + b) = 0$.

(c) Für den Spezialfall $n := 3$, $z := (1, 1, 1)^T$, $A := \text{diag}\,(1, 2, 3)$, $b := (0, 0, 0)^T$ und $\alpha := 2$ bestimme man auf effiziente Weise die Lösung x^* von (P).

Lösung: Bei der Aufgabe (P) handelt es sich darum, den Punkt z auf das nichtleere, abgeschlossene und konvexe Ellipsoid M zu projizieren. Wegen des Projektionssatzes für konvexe Mengen besitzt diese Aufgabe eine eindeutige Lösung $x^* \in M$.

Angenommen, es wäre $g(x^*) < 0$. Wegen $g(z) > 0$ existiert ein $t \in (0, 1)$ mit $g((1 - t)x^* + tz) = 0$. Dann ist $(1 - t)x^* + tz \in M$ und

$$\|(1 - t)x^* + tz - z\|_2 = (1 - t)\|x^* - z\|_2 < \|x^* - z\|_2,$$

ein Widerspruch dazu, dass x^* Lösung von (P). Weiter ist $\nabla g(x^*) = Ax^* + b \neq 0$, denn andernfalls wäre $x^* = -A^{-1}b$ und $g(x^*) = -\frac{1}{2}b^T A^{-1}b - \alpha < 0$, im Gegensatz zu dem, was wir gerade bewiesen haben. Die Constraint Qualification für die notwendige Bedingung erster Ordnung ist erfüllt, da $\nabla g(x^*) \neq 0$. Daher existiert ein $\lambda^* \geq 0$ mit $\nabla f(x^*) + \lambda^* \nabla g(x^*) = 0$ bzw. $x^* - z + \lambda^*(Ax^* + b) = 0$. Wegen $x^* \neq a$ ist $\lambda^* > 0$. Existiert umgekehrt zu x^* mit $g(x^*) = 0$ ein $\lambda^* > 0$ mit $x^* - z + \lambda^*(Ax^* + b) = 0$, so ist x^* die Lösung von (P).

Offenbar ist

$$x_1^* = \frac{1}{1 + \lambda^*}, \quad x_2^* = \frac{1}{1 + 2\lambda^*}, \quad x_3^* = \frac{1}{1 + 3\lambda^*}.$$

Daher ist $\lambda^* > 0$ zu bestimmen als positive Lösung der Gleichung

$$h(\lambda) := \tfrac{1}{2}\Big(\frac{1}{(1+\lambda)^2} + \frac{2}{(1+2\lambda)^2} + \frac{3}{(1+3\lambda)^2}\Big) - 2 = 0.$$

Eine Skizze zeigt, dass die Lösung in der Nähe von 0.1 liegt. Mit diesem Startwert und dem Newton-Verfahren erhalten wir die Werte

k	λ_k
0	0.100000000000000
1	0.099335655479481
2	0.099336946295961
3	0.099336946300852
4	0.099336946300852

$$x^* = \begin{pmatrix} 0.909639217862085 \\ 0.834255260060361 \\ 0.770409591375586 \end{pmatrix}.$$

7. Seien die symmetrische und positiv definite Matrix $Q \in \mathbb{R}^{n\times n}$, die positive Zahl C und der Vektor $f \in \mathbb{R}^n$ gegeben. Hiermit betrachte man die Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x,\eta) := \frac{1}{2}x^TQx + C\eta \quad \text{auf} \\ M := \{(x,\eta) \in \mathbb{R}^n \times \mathbb{R} : -\eta e \le Qx - f \le \eta e\}, \end{cases}$$

wobei $e \in \mathbb{R}^n$ der Vektor ist, dessen Komponenten sämtlich gleich 1 sind[7]. Man zeige, dass (P) genau eine Lösung $(x^*, \eta^*) \in M$ besitzt.

Lösung: Das Problem (P) ist zulässig. Um dies einzusehen wähle man $x_0 \in \mathbb{R}^n$ beliebig und $\eta_0 \ge \|Qx_0 - f\|_\infty$. Offenbar ist dann $(x_0, \eta_0) \in M$. Da (P) zulässig und $\inf(\text{P}) > -\infty$ (es ist sogar $\inf(\text{P}) \ge 0$) folgt aus dem Satz von Barankin-Dorfman die Existenz einer Lösung von (P). Da dieser Satz in der Vorlesung aber nur erwähnt, aber nicht bewiesen wurde, wollen wir einen unabhängigen Beweis angeben. Hierzu bilde man mit einem $(x_0, \eta_0) \in M$ die Niveaumenge

$$L_0 := \{(x,\eta) \in \mathbb{R}^n \times \mathbb{R} : \|Qx - f\|_\infty \le \eta,\ f(x,\eta) \le f(x_0,\eta_0)\}.$$

Offenbar ist L_0 abgeschlossen und beschränkt, also kompakt. Nur die Beschränktheit ist nicht völlig offensichtlich. Für $(x,\eta) \in L_0$ ist

$$\|x\|_2 \le (2f(x_0,\eta_0)/\lambda_{\min}(Q))^{1/2}, \qquad 0 \le \eta \le f(x_0,\eta_0)/C$$

und das zeigt die Beschränktheit von L_0. Angenommen, (x_1,η_1) und (x_2,η_2) seien zwei Lösungen von (P). Da (P) eine konvexe Optimierungsaufgabe ist, ist auch $(x_3,\eta_3) := \frac{1}{2}(x_1,\eta_1) + \frac{1}{2}(x_2,\eta_2)$ eine Lösung und es ist $f(x_1,\eta_1) = f(x_2,\eta_2) = f(x_3,\eta_3)$. Da auch $f(x_3,\eta_3) = \frac{1}{2}f(x_1,\eta_1) + \frac{1}{2}f(x_2,\eta_2)$, erhält man zunächst $x_1^TQx_1 = x_2^TQx_2 = x_3^TQx_3$, hieraus wegen der strikten Konvexität des quadratischen Anteils der Zielfunktion $x_1 = x_2$ und hieraus schließlich auch $\eta_1 = \eta_2$.

[7] Probleme dieser Art treten bei Problemen des Maschinellen Lernens auf. Siehe R. Schaback, J. Werner (2006).

8. Man entwickle ein Verfahren zur Bestimmung der Projektion eines Vektors $z \in \mathbb{R}^n$ auf das Simplex $\Sigma := \{x \in \mathbb{R}^n : x \geq 0,\ e^T x \leq 1\}$. Anschließend wende man das Verfahren auf den Spezialfall $n := 4$, $z := (1, 5, 3, 2)^T$ an.

   Hinweis: Bei gegebenem $\lambda \in \mathbb{R}$ bestimme man $x(\lambda) \in \mathbb{R}^n$ mit

   $$x(\lambda) \geq 0, \qquad x(\lambda) - z + \lambda e \geq 0, \qquad (x(\lambda) - z + \lambda e)^T x(\lambda) = 0.$$

   Man zeige: Ist $e^T x(0) \leq 1$, so ist $x^* := x(0)$ die Lösung von (P). Andernfalls bestimme man $\lambda^* > 0$ als positive Nullstelle von $h(\lambda) := e^T x(\lambda) - 1$ und zeige, dass $x^* = x(\lambda^*)$ die gesuchte Projektion auf das Simplex ist. Wie λ^* berechnet werden kann, zeige man wenigstens für den angegebenen Spezialfall.

   Lösung: Zu lösen ist die quadratische Optimierungsaufgabe

   $$\text{(P)} \qquad \text{Minimiere} \quad \tfrac{1}{2}\|x - z\|^2 \quad \text{auf} \quad \Sigma := \{x \in \mathbb{R}^n : x \geq 0,\ e^T x \leq 1\},$$

   wobei $\|\cdot\|$ die euklidische Norm und $e := (1, \ldots, 1)^T \in \mathbb{R}^n$ ist. Die Lösung $x^* \in M$ ist charakterisiert durch die Existenz eines $\lambda^* \geq$ mit

   $$x^* - z + \lambda^* e \geq 0, \qquad (x^* - z + \lambda^* e)^T x^* = 0, \qquad \lambda^*(e^T x^* - 1) = 0.$$

   In Abhängigkeit von $\lambda \geq 0$ definieren wir $x(\lambda) \in \mathbb{R}^n$ durch

   $$x_j(\lambda) := \begin{cases} z_j - \lambda, & \text{falls } \lambda < z_j, \\ 0, & \text{falls } \lambda \geq z_j, \end{cases} \qquad j = 1, \ldots, n.$$

   Für jedes $\lambda \geq 0$ ist

   $$x(\lambda) \geq 0, \qquad x(\lambda) - z + \lambda e \geq 0, \qquad (x(\lambda) - z + \lambda e)^T x(\lambda) = 0.$$

   Ist daher $e^T x(0) \leq 1$, so ist $x^* := x(0)$ die Lösung von (P), andernfalls ist $\lambda^* > 0$ als Nullstelle von $h(\lambda) := e^T x(\lambda) - 1$ zu bestimmen. Es ist

   $$h(\lambda) := \sum_{j:\lambda<z_j} (z_j - \lambda) - 1.$$

   Wir nehmen an, es sei $e^T x(0) > 1$ bzw. $h(0) > 0$, weil man andernfalls die gesuchte Lösung schon bestimmt hat. Für alle $\lambda > 0$ mit $\lambda > \max_{j=1,\ldots,n} z_j$ ist dagegen $h(\lambda) = -1$. Ferner ist $h(\cdot)$ stetig, stückweise linear und auf $[0, \max_j z_j]$ monoton fallend, es existiert also genau eine positive Nullstelle λ^* von $h(\cdot)$. Für den angegebenen Spezialfall stellen wir die Funktion $h(\cdot)$ in Abbildung 8.7 dar. An hand der Abbildung erkennt man, dass die gesuchte Nullstelle ziemlich genau bei 4 liegt, außerdem erkennt man (was wir schon wissen), dass $h(\lambda) = -1$ für $\lambda \geq 5$. Für $\lambda \in [3, 5]$ ist $h(\lambda) = (5 - \lambda) + (3 - \lambda) = 8 - 2\lambda$, es ist also $\lambda^* := 4$ die gesuchte Nullstelle. Die gesuchte Projektion ist $x^* := (0, 1, 0, 0)^T$.

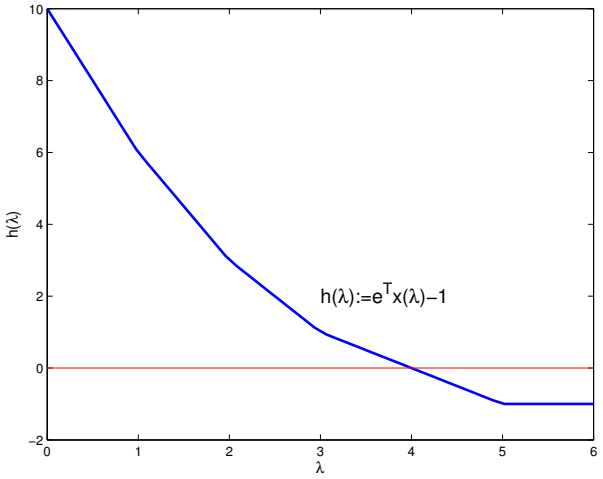


Abbildung 8.7: Die Funktion $h(\lambda) := e^T x(\lambda) - 1$

8.5.2 Aufgaben zu Abschnitt 5.2

1. Mit Hilfe des Verfahrens von Goldfarb-Idnani löse man die quadratische Optimierungsaufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := \begin{pmatrix} -2 \\ -6 \end{pmatrix}^T x + \frac{1}{2} x^T \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} x \\ \text{unter der Nebenbedingung} \\ \begin{pmatrix} 1 & 1 \\ -1 & 2 \\ 2 & 1 \end{pmatrix} x \leq \begin{pmatrix} 2 \\ 2 \\ 3 \end{pmatrix}. \end{array} \right.$$

Lösung: 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 verletzte Restriktion, setzen $\theta := 0$ und 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

$$t_1 = \tfrac{16}{5}, \qquad t_2 = +\infty, \qquad t = \tfrac{16}{5}.$$

Anschließend wird der primale Schritt gemacht und

$$x_1 = \begin{pmatrix} \frac{2}{5} \\ \frac{8}{5} \end{pmatrix}, \qquad f_1 = -\tfrac{42}{5}, \qquad \theta = \tfrac{16}{5}$$

berechnet, der duale Schritt liefert

$$y_1 = (\tfrac{16}{5}), \qquad I_1 = \{1\}.$$

Damit ist der erste Schritt abgeschlossen. Im zweiten Schritt ist nur die zweite Restriktion verletzt, es ist also $p = 2$, ferner ist wieder $\theta = 0$. Wir berechnen

$$N_{I_1} = \begin{pmatrix} \frac{3}{5} & \frac{2}{5} \end{pmatrix}, \qquad H_{I_1} = \begin{pmatrix} \frac{1}{5} & -\frac{1}{5} \\ -\frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

und die primalen bzw. dualen Richtungen

$$z = \begin{pmatrix} \frac{3}{5} \\ -\frac{3}{5} \end{pmatrix}, \qquad r_{I_1} = (\tfrac{1}{5}).$$

Anschließend berechnet man die Schrittweiten

$$t_1 = \tfrac{4}{9}, \qquad t_2 = 16, \qquad t = \tfrac{4}{9}.$$

Im primalen und dualen Schritt wird zunächst

$$x_2 = \begin{pmatrix} \frac{2}{3} \\ \frac{4}{3} \end{pmatrix}, \qquad f_2 = -\tfrac{74}{9}, \qquad \theta = \tfrac{4}{9},$$

danach

$$y_2 = \begin{pmatrix} \frac{28}{9} \\ \frac{4}{9} \end{pmatrix}, \qquad I_2 = \{1, 2\}$$

berechnet. Da x_2 zulässig ist, hat man die Lösung gefunden.

8.6 Aufgaben zu Kapitel 6

8.6.1 Aufgaben zu Abschnitt 6.1

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

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

Hierbei seien $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = m$ und $b \in \mathbb{R}^m$ gegeben. Wie in Unterabschnitt 6.1.1 geschildert ordne man (P) die unrestringierte Optimierungsaufgabe

$$(\mathrm{P}_x) \qquad \text{Minimiere} \quad \psi(u) := f(x + Zu), \quad u \in \mathbb{R}^{n-m},$$

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 $\mathbb{R}^{n-m}$.

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 6.1.1 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

$$Lw = \Big(\frac{y^T s}{s^T B s}\Big)^{1/2} Bs.$$

Es ist nämlich

$$\begin{aligned}
J_+ J_+^T &= \Big(L + \frac{(y-Lw)w^T}{y^T s}\Big)\Big(L^T + \frac{w(y-Lw)^T}{y^T s}\Big) \\
&= LL^T + \frac{Lw(y-Lw)^T}{y^T s} + \frac{(y-Lw)(Lw)^T}{y^T s} \\
&\qquad + \Big(\frac{\|w\|}{y^T s}\Big)^2 (y-Lw)(y-Lw)^T \\
&= LL^T + \frac{Lw(y-Lw)^T}{y^T s} + \frac{(y-Lw)(Lw)^T}{y^T s} + \frac{(y-Lw)(y-Lw)^T}{y^T s} \\
&= LL^T - \frac{(Lw)(Lw)^T}{y^T s} + \frac{yy^T}{y^T s} \\
&= B - \frac{(Bs)(Bs)^T}{s^T B s} + \frac{yy^T}{y^T s} \\
&= B_+.
\end{aligned}$$

Damit ist die erste Aussage bewiesen. Wegen

$$\sigma := 1 + \frac{w^T(L^{-1}y - w)}{y^T s} = \frac{w^T L^{-1} y}{y^T s} = \Big(\frac{y^T s}{s^T B s}\Big)^{1/2} \neq 0$$

ist J_+ nach der Sherman-Morrison-Lemma 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 $m := \max\{i \in \{1,\dots,n\} : u_i \neq 0\}$. Zunächst führt man den Vektor u durch sukzessive Multiplikation mit $m-1$ geeigneten Givensrotationen $G_{m-1,m}, \dots, G_{12}$, welche der Reihe nach die Komponenten mit den Indizes $m, \dots, 2$ annullieren, in ein Vielfaches $u_1 e_1$ des ersten Einheitsvektors über. Die parallel hierzu durchgeführte Multiplikation der oberen Dreiecksmatrix R mit den Givensrotationen $G_{m-1,m}, \dots, G_{12}$ transformiert diese in eine obere Hessenberg-Matrix, die wir mit H bezeichnen. Nach Abschluss dieses ersten Schrittes ist $G_{12} \cdots G_{m-1,m} A_+ = H + u_1 e_1 v^T$ mit einer oberen Hessenberg-Matrix (deren Subdiagonalelemente in den Spalten $m, \dots, n-1$ verschwinden. In einem

Zwischenschritt berechnet man $H := H + u_1 e_1 v^T$, 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.

8.6.2 Aufgaben zu Abschnitt 6.2

1. Gegeben sei eine linear restringierte nichtlineare Optimierungsaufgabe mit einer stetig differenzierbaren Zielfunktion. Man zeige, dass eine zulässige Lösung genau dann eine stationäre Lösung ist, wenn es in ihr keine zulässige Abstiegsrichtung gibt.

   Lösung: Gegeben sei die Optimierungsaufgabe

   $$\text{(P)}\quad \text{Minimiere } f(x) \text{ auf } M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0) \\ a_i^T x = b_i & (i = m_0+1, \ldots, m) \end{array} \right\},$$

   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 stationäre 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 stationäre Lösung. Gäbe es eine in x zulässige Abstiegsrichtung p, so wäre

   $$0 > \nabla f(x)^T p = (A^T y)^T p = y^T A p = \sum_{i \in I(x)} y_i a_i^T p \geq 0,$$

   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 Farkas-Lemma (bzw. eine entsprechende Variante) 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 stationäre Lösung von (P).

2. Gegeben sei die linear restringierte nichtlineare Optimierungsaufgabe (siehe W. Hock, K. Schittkowski (1981, S. 78))

$$
\text{(P)} \quad \left\{ \begin{array}{l} \text{Minimiere} \quad f(x) := x_1 + 2x_2 + 4x_5 + \exp(x_1 x_4) \quad \text{unter den NBen} \\ \begin{pmatrix} 1 & 2 & 0 & 0 & 5 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \\ 2 \\ 1 \\ 2 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \leq \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}, \\ \begin{array}{rcl} x_1 & \leq & 1, \\ x_4 & \leq & 1. \end{array} \end{array} \right.
$$

Man löse diese Aufgabe mit der Matlab-Funktion `fmincon`, wobei man wie bei Hock-Schittkowski den Startwert $x_0 = (1, 2, 0, 0, 0, 2)^T$ nehme. Anschließend wiederhole man das Experiment mit dem Startwert $x_0 = (0, 2, 0, 1, 0, 2)^T$. Welche der erhaltenen Lösungen ist besser?

Lösung: Wir geben die Zielfunktion und ihre Ableitung durch

```
function [f,g]=Obj(x);
temp=exp(x(1)*x(4));
f=x(1)+2*x(2)+4*x(5)+temp;
if nargout > 1
   g=[1+x(4)*temp;2;0;x(1)*temp;4;0];
end;
```

an. Anschließend schreiben wir ein Script file `Test.m` mit dem Inhalt

```
A=[1 2 0 0 5 0;1 1 1 0 0 0;0 0 0 1 1 1;...
 1 0 0 1 0 0; 0 1 0 0 1 0;0 0 1 0 0 1];
b=[6;3;2;1;2;2];x_0=[1;2;0;0;0;2];
l=zeros(6,1);u=inf*ones(6,1);u(1)=1;u(4)=1;
options=optimset('GradObj','on');
[x,f_min,flag,output,y]=fmincon(@Obj,x_0,[],[],A,b,...
                                              l,u,[],options);
```

Als Resultat erhalten wir

$$
x^* = \begin{pmatrix} 1.00000000000000 \\ 1.66666666666667 \\ 0.33333333333333 \\ 0.00000000000000 \\ 0.33333333333333 \\ 1.66666666666667 \end{pmatrix}, \qquad f_{\min} = 6.66666666666667.
$$

Dies ist keine globale Lösung! Mit dem Startwert $x_0 = (0, 2, 0, 1, 0, 2)^T$ erhalten wir nämlich

$$x^* = \begin{pmatrix} 0 \\ 1.33333333333333 \\ 1.66666666666667 \\ 1.00000000000000 \\ 0.66666666666667 \\ 0.33333333333333 \end{pmatrix}, \qquad f_{\min} = 6.33333333333333,$$

also einen zulässigen Punkt mit einem kleineren Zielfunktionswert.

3. Man zeige: Genügt die Zielfunktion f von (P) den Voraussetzungen (V) (a)–(c) in Unterabschnitt 6.2.2, so existiert eine Konstante $\theta_C > 0$ derart, dass

$$\begin{aligned} f(x) - f(x + t_M(x,p)p) &\geq f(x) - f(x + t_C(x,p)p) \\ &\geq \theta_C \, \min\Big[-s(x,p)\,\nabla f(x)^T p, \Big(\frac{\nabla f(x)^T p}{\|p\|}\Big)^2\Big] \end{aligned}$$

für alle nicht stationären $x \in L_0$ 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

$$\phi(t) := f(x + tp)$$

und

$$\psi(t) := \tfrac{1}{2}\,\phi'(0) - \phi'(t) = \tfrac{1}{2}\,\nabla f(x)^T p - \nabla f(x + tp)^T p.$$

Sei $\tilde{t}(x,p)$ die erste Nullstelle von $\psi(\cdot)$ in $(0, t_C(x,p)]$, falls eine 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

$$\tfrac{1}{2}\,\nabla f(x)^T p = [\nabla f(x) - \nabla f(x + \tilde{t}(x,p)p)]^T p \geq -\tilde{t}(x,p)\,\gamma\,\|p\|^2,$$

woraus

$$\tilde{t}(x,p) \geq -\frac{1}{2\gamma}\Big(\frac{\nabla f(x)^T p}{\|p\|^2}\Big)$$

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. Gegeben sei das linear restringierte Programm

$$\text{(P)} \quad \text{Minimiere } f(x) \text{ auf } M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$

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

$$\left\{ \begin{array}{c} \text{Minimiere} \quad \nabla f(x_k)^T p \quad \text{unter den Nebenbedingungen} \\ a_i^T p \geq b_i - a_i^T x_k \;\; (i = 1, \ldots, m_0), \quad a_i^T p = 0 \;\; (i = m_0 + 1, \ldots, m). \end{array} \right.$$

  - Falls $\nabla f(x_k)^T p_k = 0$, dann: STOP, x_k ist stationäre Lösung von (P).
  - Berechne $t_k := t_M(x_k, p_k)$, $t_C(x_k, p_k)$ oder $t_k \in t_W(x_k, p_k)$.
  - Setze $x_{k+1} := x_k + t_k p_k$.

Dann gilt: Bricht das Verfahren nicht vorzeitig mit einer stationären Lösung von (P) ab, so liefert es eine Folge $\{x_k\}$ mit der Eigenschaft, dass jeder Häufungspunkt von $\{x_k\}$ eine stationäre 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

$$y_i \geq 0 \quad (i = 1, \ldots, m_0), \quad \nabla f(x_k) = \sum_{i=1}^{m} y_i a_i$$

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

$$y_i \geq 0 \ \ (i = 1, \ldots, m_0), \quad \nabla f(x_k) = \sum_{i=1}^{m} y_i a_i, \quad y_i(a_i^T x_k - b_i) = 0 \ \ (i = 1, \ldots, m_0),$$

also x_k eine stationäre 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 der Aufgabe 3 (Minimum- und Curry-Schrittweite) und dem Lemma 2.1 (Wolfe-Schrittweite) erhält man die Existenz einer Konstanten $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \, \min\Big[-\nabla f(x_k)^T p_k, \Big(\frac{\nabla f(x_k)^T p_k}{\|p_k\|}\Big)^2\Big].$$

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 2.3 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 $i \in \{1, \ldots, m_0\}$ 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 $i \in \{1, \ldots, m_0\} \setminus I(x^*)$ und $a_i^T p^* \geq 0$. Daher können

wir jetzt annehmen, es sei $i \in \{1, \ldots, m_0\} \setminus I(x^*)$ 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.

5. Gegeben sei das linear restringierte Programm

$$(\mathrm{P}) \quad \text{Minimiere } f(x) \text{ auf } M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} a_i^T x \geq b_i & (i = 1, \ldots, m_0), \\ a_i^T x = b_i & (i = m_0 + 1, \ldots, m) \end{array} \right\}.$$

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

$$(\mathrm{P}(x)) \quad \left\{ \begin{array}{l} \text{Minimiere} \quad \nabla f(x)^T p + \frac{1}{2} p^T B p \quad \text{unter den Nebenbedingungen} \\ \begin{array}{rcll} a_i^T p & \geq & b_i - a_i^T x & (i = 1, \ldots, m_0), \\ a_i^T p & = & 0 & (i = m_0 + 1, \ldots, m), \end{array} \qquad \|p\|_\infty \leq 1. \end{array} \right.$$

Sei p^* eine Lösung von $(\mathrm{P}(x))$. 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

$$y_i \geq 0 \quad (i = 1, \ldots, m_0), \qquad u, v \geq 0, \qquad \nabla f(x) + B p^* = \sum_{i=1}^{m} y_i a_i - u + v$$

und

$$y_i(a_i^T p^* + a_i^T x - b_i) = 0 \quad (i = 1, \ldots, m_0), \quad u^T(p^* - e) = 0, \quad v^T(p^* + e) = 0.$$

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 $B p^* = 0$ 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

$$y_i \geq 0 \ \ (i = 1, \ldots, m_0), \quad \nabla f(x) = \sum_{i=1}^{m} y_i a_i, \quad y_i(b_i - a_i^T x) = 0 \ \ (i = 1, \ldots, m_0),$$

d. h. $x \in M$ ist eine kritische Lösung von (P).

8.7 Aufgaben zu Kapitel 7

8.7.1 Aufgaben zu Abschnitt 7.1

1. Man betrachte die Optimierungsaufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad f(x) := (x_1 - x_2)^2 + (x_2 - x_3)^4 \\ \text{unter der Nebenbedingung} \\ h(x) := (1 + x_2^2)x_1 + x_3^4 - 3 = 0. \end{array} \right.$$

   (a) Man bestimme alle Lösungen.

   (b) Man löse die Aufgabe numerisch mit Hilfe von `fmincon`. Wie bei Hock-Schittkowski, S. 49, nehme man den Startwert $x_0 := (-2.6, 2, 2)$.

Lösung: Zunächst ist $x^* = (1, 1, 1)^T$ eine Lösung mit dem (nicht zu unterbietenden) Wert $f(x^*) = 0$. Bei jeder anderen Lösung müssen alle Komponenten gleich sein. Der Ansatz $x^* = (a, a, a)^T$ führt (damit dieser Punkt zulässig ist) auf die Gleichung $\phi(a) := a^4 + a^3 + a - 3 = 0$. Wir plotten ϕ über dem Intervall $[-2, 2]$: Man erkennt, dass neben $a = 1$ noch eine weitere reelle Lösung existiert, die etwa

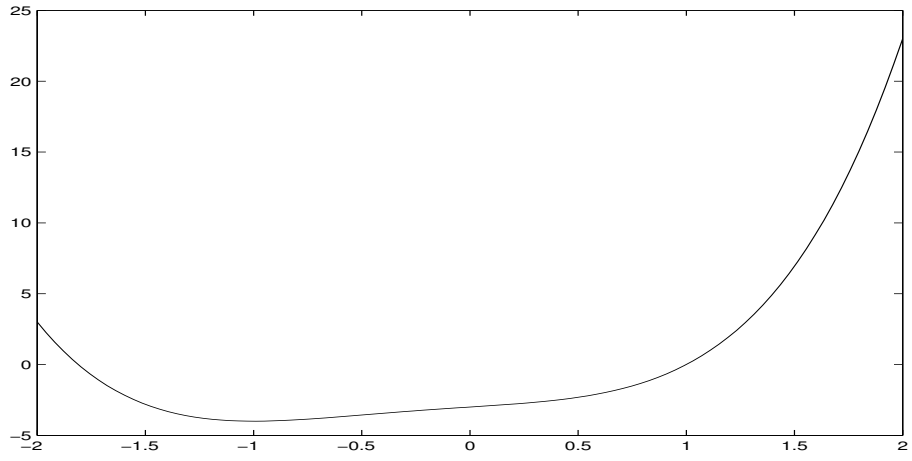


Abbildung 8.8: Die Funktion $\phi(a) := a^4 + a^3 + a - 3$

bei -1.8 liegt. Nach

```
a=fzero(inline('a^4+a^3+a-3'),-1.8)
```

erhalten wir $a = -1.8105$.

Wir schreiben die beiden function files

```
function [f,g]=MyObj_3(x);
f=(x(1)-x(2))^2+(x(2)-x(3))^4;
g=2*[x(1)-x(2);x(2)-x(1)+2*(x(2)-x(3))^3;2*(x(3)-x(2))^3];
```

und

```
function [g,h,Jg,Jh]=MyCon_3(x);
g=[];Jg=[];
h=(1+x(2)^2)*x(1)+x(3)^4-3;
Jh=[1+x(2)^2;2*x(1)*x(2);4*x(3)^3];
```

Anschließend geben wir

```
x_0_0=[-2.6;2;2];
options=optimset('GradObj','on','GradConstr','on');
[x,f,ex,out,y]=fmincon(@MyObj_3,x_0,[],[],[],[],[],[],@MyCon_3,options);
```

ein. Wir erhalten

$$x = \begin{pmatrix} 1.0192 \\ 1.0192 \\ 0.9800 \end{pmatrix}.$$

Erniedrigen wir die Parameter `TolFun` und `TolCon` (per default sind sie gleich 10^{-6}) auf 10^{-10}, so erhalten wir

$$x = \begin{pmatrix} 1.0001 \\ 1.0001 \\ 0.9999 \end{pmatrix}.$$

2. Gegeben sei das quadratische Programm

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x \quad \text{auf} \\ M := \{x \in \mathbb{R}^n : h(x) := Ax - b = 0\} \end{cases}$$

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_σ 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

$$\Phi_\sigma(x) := f(x) + \frac{\sigma}{2}\|h(x)\|^2 = c^T x + \frac{1}{2} x^T Q x + \frac{\sigma}{2}\|Ax - b\|^2$$

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

$$\begin{aligned} x^* &= -Q^{-1}c + Q^{-1}A^T(AQ^{-1}A^T)^{-1}(AQ^{-1}c + b), \\ y^* &= -(AQ^{-1}A^T)^{-1}(AQ^{-1}c + b). \end{aligned}$$

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

$$\begin{aligned} x_\sigma &= (Q + \sigma A^T A)^{-1}(\sigma A^T b - c) \\ &= Q^{-1/2}(I + \sigma Q^{-1/2}A^T A Q^{-1/2})^{-1}(\sigma Q^{-1/2}A^T b - Q^{-1/2}c) \\ &= Q^{-1/2}\left[I + \sigma U \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} \underbrace{V^T V}_{=I} (\ \hat\Sigma \quad 0\ ) U^T\right]^{-1}\left[\sigma U \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} V^T b - Q^{-1/2}c\right] \\ &= Q^{-1/2}\left[U \begin{pmatrix} I + \sigma\hat\Sigma^2 & 0 \\ 0 & I \end{pmatrix} U^T\right]^{-1}\left[\sigma U \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} V^T b - Q^{-1/2}c\right] \\ &= Q^{-1/2}U \begin{pmatrix} (I + \sigma\hat\Sigma^2)^{-1} & 0 \\ 0 & I \end{pmatrix} \sigma \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} V^T b \\ &\qquad - Q^{-1/2}U \begin{pmatrix} (I + \sigma\hat\Sigma^2)^{-1} & 0 \\ 0 & I \end{pmatrix} U^T Q^{-1/2} c \\ &\to Q^{-1/2}U \begin{pmatrix} \hat\Sigma^{-1} \\ 0 \end{pmatrix} V^T b - Q^{-1/2}U \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} U^T Q^{-1/2} c \qquad \text{mit } \sigma \to \infty. \end{aligned}$$

Andererseits ist

$$\begin{aligned} x^* &= Q^{-1}A^T(AQ^{-1}A^T)^{-1}(AQ^{-1}c+b)-Q^{-1}c \\ &= Q^{-1/2}Q^{-1/2}A^T(AQ^{-1/2}Q^{-1/2}A^T)^{-1}(AQ^{-1/2}Q^{-1/2}c+b)-Q^{-1/2}Q^{-1/2}c \\ &= Q^{-1/2}U\begin{pmatrix} I & 0\\ 0 & 0\end{pmatrix}U^TQ^{-1/2}c+Q^{-1/2}U\begin{pmatrix}\hat\Sigma^{-1}\\ 0\end{pmatrix}V^Tb-Q^{-1/2}Q^{-1/2}c \\ &= Q^{-1/2}U\begin{pmatrix}\hat\Sigma^{-1}\\ 0\end{pmatrix}V^Tb-Q^{-1/2}U\begin{pmatrix}0 & 0\\ 0 & I\end{pmatrix}U^TQ^{-1/2}c. \end{aligned}$$

Damit ist $\lim_{\sigma\to\infty} x_\sigma = x^*$ nachgewiesen. Weiter ist

$$\begin{aligned} \sigma[Ax_\sigma-b] &= \sigma[A(Q+\sigma A^TA)^{-1}(\sigma A^Tb-c)-b] \\ &= \sigma\Big[AQ^{-1/2}U\begin{pmatrix}\sigma\hat\Sigma(I+\sigma\hat\Sigma^2)^{-1}\\ 0\end{pmatrix}V^Tb-b \\ &\qquad -AQ^{-1/2}U\begin{pmatrix}(I+\sigma\hat\Sigma^2)^{-1} & 0\\ 0 & I\end{pmatrix}U^TQ^{-1/2}c\Big] \\ &= -\sigma V(I+\sigma\hat\Sigma^2)^{-1}V^Tb-\sigma V(\ \hat\Sigma(I+\sigma\hat\Sigma^2)^{-1}\quad 0\ )U^TQ^{-1/2}c \\ &\to -V\hat\Sigma^{-2}V^Tb-V(\ \hat\Sigma^{-1}\quad 0\ )U^TQ^{-1/2}c \quad \text{mit } \sigma\to\infty. \end{aligned}$$

Andererseits ist

$$\begin{aligned} y^* &= -(AQ^{-1}A^T)^{-1}(AQ^{-1}c+b) \\ &= -(AQ^{-1/2}Q^{-1/2}A^T)^{-1}(AQ^{-1/2}Q^{-1/2}c+b) \\ &= -\left[V(\ \hat\Sigma\quad 0\ )U^TU\begin{pmatrix}\hat\Sigma\\ 0\end{pmatrix}V^T\right]^{-1}\left[V(\ \hat\Sigma\quad 0\ )U^TQ^{-1/2}c+b\right] \\ &= -V(\ \hat\Sigma^{-1}\quad 0\ )U^TQ^{-1/2}c-V\hat\Sigma^{-2}V^Tb. \end{aligned}$$

Damit ist auch $\lim_{\sigma\to\infty}\sigma h(x_\sigma) = y^*$ nachgewiesen.

3. Gegeben sei das quadratische Programm

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := c^Tx+\frac{1}{2}x^TQx \quad \text{auf} \\ \quad M := \{x\in\mathbb{R}^n : h(x) := Ax-b=0\} \end{cases}$$

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

$$(\mathrm{P}^*_\sigma) \qquad \text{Minimiere} \quad \Psi_\sigma(x) := f(x)+(y^*)^Th(x)+\tfrac{1}{2}\sigma\|h(x)\|^2, \quad x\in\mathbb{R}^n,$$

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).

4. Gegeben sei (siehe P. SPELLUCCI (1993, S. 394)) die Optimierungsaufgabe

$$(\mathrm{P}) \qquad \begin{cases} \text{Minimiere} \quad f(x) := x_1^2 + 4x_1x_2 + 5x_2^2 - 10x_1 - 20x_2 \quad \text{auf} \\ M := \{x \in \mathbb{R}^2 : h(x) := x_1 + x_2 - 2 = 0\}. \end{cases}$$

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_σ von (P_σ) und bestätige die Aussage von Aufgabe 2, 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:

$$\begin{cases} \text{Minimiere} \quad f(x) := \begin{pmatrix} -10 \\ -20 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 2 & 4 \\ 4 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \qquad\qquad \text{auf} \quad M := \left\{ x \in \mathbb{R}^2 : \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 2 = 0 \right\}. \end{cases}$$

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 2 bestätigt.

5. Gegeben sei die zulässige, restringierte Optimierungsaufgabe

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

   und hierzu die unrestringierte Optimierungsaufgabe

   $$(\mathrm{P}_\sigma) \quad \text{Minimiere} \quad \Psi_\sigma(x) := f(x) + \sigma \underbrace{\Big(\sum_{i=1}^{l} \max(g_i(x), 0) + \|h(x)\|_1 \Big)}_{=:S(x)}, \quad x \in \mathbb{R}^n.$$

   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), 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

   $$S(x^*) = \sum_{i=1}^{l} \max(g_i(x^*), 0) + \|h(x^*)\|_1 > 0.$$

   Sei $x \in M$ ein beliebiger, für (P) zulässiger Punkt und

   $$\sigma > \max\Big(\frac{f(x) - f(x^*)}{S(x^*)}, \sigma^* \Big).$$

   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).

Literaturverzeichnis

[1] Alt, W. (2002) *Nichtlineare Optimierung. Eine Einführung in Theorie, Verfahren und Anwendungen*. Vieweg, Braunschweig-Wiesbaden.

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

[3] Bertsekas, D. P. (1999) *Nonlinear Programming. Second Edition*. Athena Scientific, Belmont.

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

[5] Cantor, M. (1880) *Vorlesungen über Geschichte der Mathematik Bd. 1: Von den ältesten Zeiten bis zum Jahre 1200 n. Chr.*, B. G. Teubner, Leipzig.

[6] Conn, A. R., N. I. M. Gould and Ph. L. Toint (2000) *Trust-Region Methods*. SIAM-MPS, Philadelphia.

[7] Dennis, J. E. and R. B. Schnabel (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs.

[8] Fiacco, A. V. and G. P. McCormick (1968) *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. John Wiley, New York.

[9] Fletcher, R. (1971) A general quadratic programming algorithm. Journal of the Institute of Mathematics and its Applications 7, 76–91.

[10] Fletcher, R. (1987) *Practical Methods of Optimization. Second Edition.* John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.

[11] Fletcher, R. and C. M. Reeves (1964) Function minimization by conjugate gradients. Computer Journal 7, 149–154.

[12] Fukushima, M. (1986) A successive quadratic programming algorithm with global and superlinear convergence propertie. Mathematical Programming 35, 253–264.

[13] Geiger, C. und C. Kanzow (1999) *Numerische Verfahren zur Lösung unrestringierte Optimierungsaufgaben*. Springer, Berlin-Heidelberg.

[14] Geiger, C. und C. Kanzow (2002) *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, Berlin-Heidelberg.

[15] Gill, P. E. and W. Murray (1978) Numerically stable methods for quadratic programming. Mathematical Programming 14, 349–372.

[16] Gill, P. E., W. Murray and M. H. Wright (1981) *Practical Optimization.* Academic Press, London-New York.

[17] 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.

[18] 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.

[19] Goldfarb, D. and A. Idnani (1983) A numerically stable dual method for solving strictly convex quadratic programs. Mathematical Programming 27, 1–33.

[20] Han, S.-P. and O. L. Mangasarian (1979) Exact penalty functions in nonlinear programming. Mathematical Programming 17, 251–269.

[21] Himmelblau, D. M. (1972) *Applied Nonlinear Programming*. McGraw Hill, New York.

[22] Hock, W. and K. Schittkowski (1981) *Test Examples for Nonlinear Programming Codes*. Lecture Notes in Economics and Mathematical Systems Vol. 187, Springer-Verlag, Berlin-Heidelberg-New York.

[23] Hoffman, A. J. On approximate solutions of systems of linear inequalities. J. Res. Natl. Bur. Standards, 49 (1952), pp. 263–265.

[24] Jarre, F. und J. Stoer (2004) *Optimierung*. Springer, Berlin-Heidelberg.

[25] Kosmol, P. (1989) *Methoden numerischer Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben.* B. G. Teubner, Stuttgart.

[26] Karmarkar, N. (1984) A new polynomial time algorithm for linear programming. Combinatorica 4, 373–395.

[27] Luenberger, D. G. (1969) *Optimization by Vector Space Methods*. John Wiley, New York.

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

[29] Nocedal, J. and S. J. Wright (1999) *Numerical Optimization*. Springer, Berlin-Heidelberg-New York.

[30] Ortega, J. M. and W. C. Rheinboldt (1970) *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York-London.

[31] Peterson, E. L. and 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.

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

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

[34] 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.

[35] Schaback, R. und J. Werner (2006) Linearly constrained reconstruction of functions by kernels with applications to machine learning. Advances in Computational Mathematics 25, pp. 237–258.

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

[37] Spelluci, P. (1993) *Numerische Verfahren der nichtlinearen Optimierung*. Birkhäuser, Basel-Boston-Berlin.

[38] Warth, W. und J. Werner (1977) Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben. Computing 19, 59–72.

[39] Werner, J. (1992a) *Numerische Mathematik 1*. Vieweg, Braunschweig-Wiesbaden.

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

[41] Werner, J. (2000) `http://www.num.math.uni-goettingen.de/werner/opti.ps`, Vorlesung über Optimierung.

[42] Werner, J. (2001) `http://www.num.math.uni-goettingen.de/werner/opres.ps`, Vorlesung über Operations Research.

[43] Werner, J. (2002) `http://www.num.math.uni-goettingen.de/werner/uncopt.pdf`, Vorlesung über unrestringierte Optimierungsaufgaben.

[44] Wright, M. H. (1992) Interior methods for constrained optimization. In: A. Iseles (ed.) Acta Numerica 1992. Cambridge University Press, 341–407.

[45] Wright, S. J. (1997) *Primal-Dual Interior-Point Methods*. SIAM, Philadelphia.

Index