Optimierung

Skript zur Vorlesung

Hans Joachim Oberle

Sommersemester 2012

Inhalt

1. Einführung

Optimierungsaufgaben spielen in allen Anwendungsbereichen von Mathematik eine wichtige Rolle. Hauptanwendungsgebiete sind Wirtschaftswissenschaften (Operations Research), Technik und Naturwissenschaften.

In Nocedal, Wright findet man beispielsweise folgende Formulierung:

People optimize: Airline companies schedule crews and aircraft to minimize cost. Investors seek to creat portfolios that avoid risks while achieving a high rate of return. Manufacturers aim for maximizing efficiency in the design and operation of their production processes.

Nature optimizes: Physical systems tend to a state of minimum energy. The molecules in an isolated chemical system react with each other until the total potential energy of their electrons is minimized, Rays of light follow paths that minimize their travel time.

Die Optimierung steht in enger Beziehung zur *Modellierung*, d.h. Optimierungtechniken werden auf mathematische Modelle angewendet, für die dann gewisse unbekannte Modellparameter oder -funktionen so zu bestimmen sind, dass eine *Zielfunktion* unter vorgegebenen *Nebenbedingungen* minimiert (oder maximiert) wird.

Problem (1.1)

Sei $(V, \|\cdot\|)$ ein reeller normierter Raum, $X \subset V$ nichtleer und $f : X \to \mathbb{R}$.
Gesucht ist: $x^* \in X$ mit $\forall x \in X : f(x^*) \leq f(x)$.

Das Problem wird meist folgendermaßen formuliert: Minimiere $f(x)$ unter der Nebenbedingung $x \in X$.

Je nach Problemstellung unterscheidet man die Begriffe

- *infinite Optimierung* (V ist unendlich dimensionaler Funktionenraum),
- *finite Optimierung* ($V = \mathbb{R}^n$),
- *kontinuierliche Optimierung* ($\text{int}\,(X) \neq \emptyset$),
- *diskrete Optimierung* ($X \subset \mathbb{Z}^n$).

Definition (1.2)

a) X heißt der *zulässige Bereich*, die $x \in X$ heißen *zulässige Punkte* und f das *Zielfunktional* bzw. die *Zielfunktion* der Optimierungsaufgabe (1.1).

b) x^* heißt ein *globales Minimum* von f über X, falls die Bedingung aus (1.1) erfüllt ist, also $\forall x \in X: \ f(x^*) \leq f(x)$.

c) Gilt sogar $\forall x \in X, x \neq x^*: \ f(x^*) < f(x)$, so heißt x ein *striktes globales Minimium* von f auf X.

d) x^* heißt ein *lokales Minimum* von f über X, falls es eine (offene) Umgebung U von x^* in V gibt, so dass x^* ein globales Minimum von f auf $X \cap U$ ist. Analog wird der Begriff *striktes lokales Minimum* von f auf X erklärt.

Die Aufgabe, eine Funktion f zu *maximieren*, ist äquivalent dazu, die Funktion $(-f)$ zu minimieren. Damit sind auch die Begriffe *lokales/globales Maximum* bzw. *striktes lokales/globales Maximum* definiert.

Generell versteht man unter *Optimierung* die Maximierung oder Minimierung einer Funktion auf einem (nichtleeren) zulässigen Bereich.

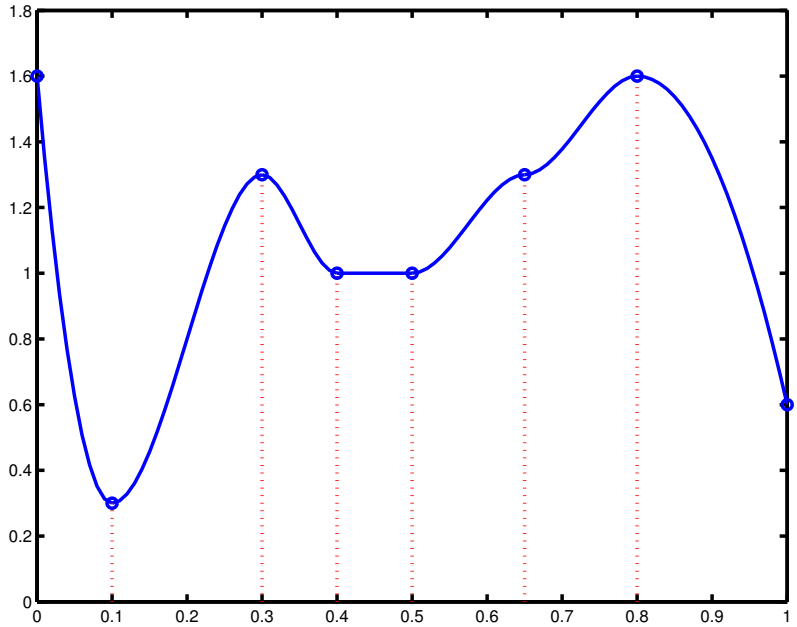


Abb. 1.1 Maxima, Minima und Sattelpunkte.

$x_1 = 0$ und $x_7 = 0.8$ sind globale Maxima und zugleich strikte lokale Maxima, $x_2 = 0.1$ ist striktes globales Minimum, $x_3 = 0.3$ ist ein striktes lokales Maximum, $x_8 = 1$ ein striktes lokales Minimum. Alle Punkte im Intervall $[x_5, x_6] = [0.4, 0.5]$ sind lokale Minima, jedoch nicht strikt, $x_6 = 0.65$ ist ein stationärer Punkt, der zugleich Sattelpunkt ist.

In dieser Vorlesung betrachten wir den Spezialfall der *endlich dimensionalen, kontinuierlichen Optimierung*. Hierbei ist $V := \mathbb{R}^n$ und die zulässige Menge ist gegeben durch ein endliches System von Gleichungs- und Ungleichungsrestriktionen

$$X \;=\; \{x \in \mathbb{R}^n : \quad g_i(x) \le 0 \,(i = 1, \ldots, m), \quad h_j(x) = 0 \,(j = 1, \ldots, p)\}. \qquad (1.3)$$

Problem (1.4)

Zu vorgegebenen (hinreichend glatten) Funktionen $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, \ldots, m)$ und $h_j : \mathbb{R}^n \to \mathbb{R}$ $(j = 1, \ldots, p)$ ist ein Punkt $x^* \in \mathbb{R}^n$ gesucht, für den $f(x)$ minimiert wird unter den Nebenbedingungen

$$g_i(x) \le 0,\ i = 1, \ldots, m, \quad h_j(x) = 0,\ j = 1, \ldots, p.$$

Wir fassen die Funktionen g_i und h_j auch zu vektorwertigen Funktionen $g := (g_1, \ldots, g_m)^{\mathrm{T}}$ bzw. $h := (h_1, \ldots, h_p)^{\mathrm{T}}$ zusammen.
Dann lautet die Optimierungsaufgabe: Minimiere $f(x)$, $x \in \mathbb{R}^n$, unter den Nebenbedingungen $g(x) \le 0 \in \mathbb{R}^m$ und $h(x) = 0 \in \mathbb{R}^p$.

Beispiel (1.5) (Flugzeugnase nach Gill, Murray und Wright)
Gesucht sind die Konstruktionsdaten einer Flugzeugnase, bestehend aus einer Kugelkappe und Kegelstümpfen, so dass der Luftwiderstand D (drag) minimiert wird.

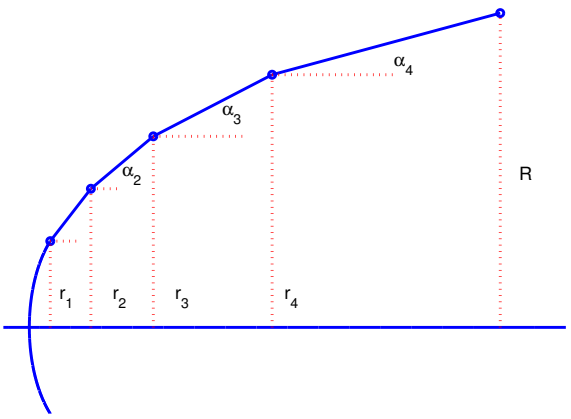


Abb. 1.2 Flugzeugnase.

Mathematische Formulierung: Minimiere $D = D(\alpha_1, \ldots, \alpha_4, r_1, \ldots, r_4)$ unter den Nebenbedingungen $0 \le r_i \le R,\ i = 1, \ldots, 4, \quad 0 \le \alpha_i \le \pi/2,\ i = 1, \ldots, 4,$
$\alpha_4 \le \alpha_3 \le \alpha_2 \le \alpha_1, \quad \mathrm{Vol}(\alpha_1, \ldots, \alpha_4, r_1, \ldots, r_4) - V = 0,$
$\mathrm{Länge}(\alpha_1, \ldots, \alpha_4, r_1, \ldots, r_4) - L \le 0.$

Feinere Klassifizierung:

- *Unrestringierte Optimierung:* $m = p = 0$

  Restringierte Optimierung: $m > 0$ oder $p > 0$

- *Lineare Optimierung:* f linear, g_i, h_j affin-linear

- *Quadratische Optimierung:* $f(x) = \frac{1}{2}\, x^{\mathrm{T}} A\, x + b^{\mathrm{T}} x + c$, g_i, h_j affin-linear

- *Konvexe Optimierung:* f konvex, g_i konvex, h_j affin-linear

- *L_2–Probleme:* Die Zielfunktion f hat die spezielle Form

$$f(x) \;=\; \sum_{i=1}^{m} w_i \,(f_i(x))^2$$

  mit glatten Funktionen f_i und Gewichten $w_i > 0$. Man vergleiche hierzu auch die Problemstellung für (nichtlineare) Ausgleichsaufgaben.

- *Minimax Probleme:* Die Zielfunktion f hat die spezielle Form

$$f(x) \;=\; \max_{i=1,\ldots,m} f_i(x)$$

  wiederum mit glatten Funktionen f_i.
  Man beachte, dass Minimax Probleme i. Allg. auf eine nichtdifferenzierbare Zielfunktion führen.

Beispiel (1.6) (Standortproblem)

Zu vorgegebenen Punkten $a_i \in \mathbb{R}^n$, $i = 1, \ldots, m$ (Standorten) ist ein neuer Standort $x \in \mathbb{R}^n$ so zu bestimmen, dass der gewichtete mittlere Abstand zu den a_i minimal ist. Die positiven Gewichte ω_i sind dabei - je nach Wichtigkeit der einzelnen Standortentfernungen - vorgegeben . Der gewichtete mittlere Abstand (das ist die zu minimierende Zielfunktion) lässt sich dabei auf verschiedene Arten festlegen. Gebräuchlich sind die Zielfunktionen

(i) $\quad f_1(x) \;:=\; \sum_{i=1}^{m} \omega_i \,\|x - a_i\|_2$ $\qquad$ (L_1-Abstand)

(ii) $\quad f_2(x) \;:=\; \left(\sum_{i=1}^{m} \omega_i \,\|x - a_i\|_2^2 \right)^{1/2}$ $\qquad$ (L_2-Abstand)

(iii) $\quad f_\infty(x) \;:=\; \max\{\, \omega_i \,\|x - a_i\|_2 :\; i =, 1, \ldots, m\}$ $\qquad$ (L_∞-Abstand)

Beispiel (1.7) (Angebotsproblem)

Ein Unternehmen will eine bestimmte Menge M einer Ware einkaufen und holt dazu Angebote von n Lieferanten ein. Der Lieferant Nr. i gibt an, welche Menge $x_{i,\max}$ er maximal liefern kann und welchen Preis $f_i(x_i)$ er - in Abhängigkeit von der bestellten Menge x_i - für seine Ware verlangt. f_i wird dabei eine monoton wachsende, aber im Allg. keine lineare Funktion sein.

Zur Minimierung der Gesamtkosten hat das Unternehmen dann die folgende Optimierungsaufgabe zu lösen:

$$\begin{array}{ll} \text{Minimiere} & f(x) \;:=\; \sum_{i=1}^{n} f_i(x_i) \\ \text{Nebenbedingungen:} & \sum_{i=1}^{n} x_i \;=\; M, \quad 0 \;\le\; x_i \;\le\; x_{i,\max},\;\; i = 1, \ldots, n \end{array}$$

Dass selbst sehr einfach aussehende nichtlineare Optimierungsaufgaben beliebig schwierig sein können, mag das folgende, etwas ungewöhnliche Beispiel zeigen.

Beispiel (1.8) (Der große Fermatsche Satz)

Zu bestimmen sei ein (globales) Minimum $x \in \mathbb{R}^4$ der Funktion

$$f(x) \;:=\; (x_1^{x_4} \;+\; x_2^{x_4} \;-\; x_3^{x_4})^2 \;+\; \sum_{i=1}^{4} (1 \;-\; \cos(2\,\pi\, x_i))$$

unter den Nebenbedingungen $x_1,\, x_2,\, x_3 \;\ge\; 1$ und $x_4 \;\ge\; 3$.

Hätte man diese Aufgabe gelöst, oder gezeigt, dass es kein globales Minimum gibt, so hätte man damit auch die Frage beantwortet, ob es natürliche Zahlen $x_1,\, x_2,\, x_3$ und $n \ge 3$ gibt mit $x_1^n + x_2^n \;=\; x_3^n$.

Der große Fermatsche Satz besagt gerade, dass es solche Zahlen nicht gibt; diese Aussage, von Fermat vermutet, wurde jedoch erst 1994 von Wiles und Taylor mit Mitteln der Zahlentheorie bewiesen.

Die folgenden Fragen sind im Zusammenhang mit allgemeinen Optimierungaufgaben zu untersuchen

- Gibt es zulässige Punkte (auch zulässige Lösungen genannt)?
- Existieren optimale Lösungen? Eindeutigkeit?
- Wie hängt die (eine) optimale Lösung von den Problemparametern ab (Stabilität)?
- Lassen sich notwendige und hinreichende Bedingungen überprüfen?

- Welche Algorithmen lassen sich zur Berechnung einer optimalen Lösung einsetzen? Welche numerischen Eigenschaften haben diese (Konvergenz, Konvergenzgeschwindigkeit, numerische Stabilität)?
- Lassen sich Fehlerabschätzungen für die berechneten Näherungslösungen angeben?

Notation (1.9)

Vektoren $x \in \mathbb{R}^n$ bezeichnen stets *Spaltenvektoren*; mit x^{T} wird der zugehörige Zeilenvektor bezeichnet.

Mit $\|\cdot\|$ wird eine Norm auf dem $\mathbb{R}^n$ bezeichnet, in der Regel ist dabei die Euklidische Norm gemeint:

$$\|x\| \;=\; \|x\|_2 \;=\; (x^{\mathrm{T}}x)^{1/2} \;=\; \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}.$$

Ist $\|\cdot\|$ eine Norm auf dem $\mathbb{R}^n$, so wird durch

$$\|A\| \;:=\; \max\{\|Ax\| :\; \|x\| \le 1\}$$

eine *Matrixnorm* (*Operatornorm*) für $A \in \mathbb{R}^{(n,n)}$ definiert. Für die Euklidische Vektornorm ist $\|A\| = \left(\lambda_{\max}(A^{\mathrm{T}}A)\right)^{1/2}$.

Zu $x \in \mathbb{R}$ und $\varepsilon > 0$ wird mit $K_\varepsilon(x)$ die offene Kugel um x mit Radius ε bezeichnet,

$$K_\varepsilon(x) \;:=\; \{y \in \mathbb{R}^n :\; \|y - x\| \;<\; \varepsilon\}.$$

Zu $X \subset \mathbb{R}^n$ bezeichnet $\mathrm{int}\,(X)$ den *offenen Kern* von X, also $x \in \mathrm{int}\,(X)$ genau dann, wenn es ein $\varepsilon > 0$ gibt mit $K_\varepsilon(x) \subset X$.
Mit $\overline{X}$ wird die *abgeschlossene Hülle* von X bezeichnet, das ist die Menge aller Grenzwerte von Folgen aus X.
Schließlich bezeichnet $\partial X := \overline{X} \setminus \mathrm{int}\,(X)$ den *Rand* der Menge X.

Eine Teilmenge $X \subset \mathbb{R}^n$ heißt *offen*, falls $\mathrm{int}\,(X) = X$, *abgeschlossen*, falls $\overline{X} = X$, und *beschränkt*, falls $X \subset K_R(0)$ für ein hinreichend großes $R > 0$ gilt.

X heißt *kompakt*, falls jede offene Überdeckung von X eine endliche Teilüberdeckung besitzt, und *folgenkompakt*, falls jede Folge von Elementen aus X eine in X konvergente Teilfolge besitzt. Für metrische Räume stimmen die Begriffe Kompaktheit und Folgenkompaktheit überein (siehe z.B. J. Dieudonne: Foundations of Modern Analysis, Hesperides Press, 2006).

Eine Teilmenge $X \subset \mathbb{R}^n$ ist ferner genau dann kompakt, wenn sie beschränkt und abgeschlossen ist. Diese letzte Aussage lässt sich nicht auf unendlich dimensionale normierte Räume übertragen.

Ist $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, also eine C^1-Funktion, so heißt

$$\nabla f(x) \;=\; \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(x) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(x) \end{pmatrix}$$

der Gradient von f in x (Spaltenvektor!).

Ist f zweifach stetig differenzierbar, $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, so heißt

$$\nabla^2 f(x) \;=\; \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1}(x) & \cdots & \dfrac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{pmatrix}$$

die Hesse–Matrix von f in x. $\nabla^2 f(x)$ ist eine *symmetrische* Matrix.

Satz von Taylor (nur linear und quadratisch)

a) Ist $f \in \mathrm{C}^1(X, \mathbb{R})$ wobei $X \subset \mathbb{R}^n$ offen sei, so gilt für $x \in X$, $d \in \mathbb{R}^n$, $\|d\|$ hinreichend klein:

$$f(x+d) \;=\; f(x) \;+\; \nabla f(x)^{\mathrm{T}} d \;+\; o(\|d\|).$$

Dabei ist das Landau-Symbol definiert durch

$$\phi(d) = o(\|d\|^k) \;:\Leftrightarrow\; \lim_{d \to 0} \phi(d)/\|d\|^k \;=\; 0.$$

Ist $f \in \mathrm{C}^2(X, \mathbb{R})$, so lässt sich ferner der Fehler darstellen durch

$$f(x+d) \;=\; f(x) \;+\; \nabla f(x)^{\mathrm{T}} d \;+\; \frac{1}{2}\, d^{\mathrm{T}} \nabla^2 f(x+\theta d)\, d, \;\; 0 < \theta < 1.$$

b) Ist $f \in \mathrm{C}^2(X, \mathbb{R})$, $X \subset \mathbb{R}^n$ offen, so gilt für $x \in X$, $d \in \mathbb{R}^n$, $\|d\|$ hinreichend klein:

$$f(x+d) \;=\; f(x) \;+\; \nabla f(x)^{\mathrm{T}} d \;+\; \frac{1}{2}\, d^{\mathrm{T}} \nabla^2 f(x)\, d \;+\; o(\|d\|^2).$$

Ist $f \in \mathrm{C}^3(X, \mathbb{R})$, so lässt sich der Fehler wiederum nach Lagrange darstellen durch

$$R_2(x+d; x) \;=\; \frac{1}{3!} \sum_{i,j,k=1}^{n} \frac{\partial^3 f(x+\theta\, d)}{\partial x_i \partial x_j \partial x_k}\, d_i\, d_j\, d_k, \;\; 0 < \theta < 1.$$

2. Optimalitätskriterien für unrestringierte Optimierungsaufgaben.

A. Globale Minima.

Satz (2.1) (Kriterium von Weierstraß)

Ist $X \subset \mathbb{R}^n$ nichtleer und kompakt und $f : \mathbb{R}^n \to \mathbb{R}$ stetig, so existiert ein globales Minimum von f über X.

Beweis: Es gibt eine Folge (x^k) in X mit $f(x^k) \to \inf f(X)$. Da X kompakt ist und somit auch folgenkompakt, gibt es eine konvergente Teilfolge von (x^k) mit $x^{k_j} \to x^* \in X$ $(j \to \infty)$.

Aufgrund der Stetigkeit von f folgt hiermit aber auch $f(x^{k_j}) \to f(x^*)$ $(j \to \infty)$ und damit $f(x^*) = \inf f(X)$. □

Folgerung (2.2)

Ist $f : \mathbb{R}^n \to \mathbb{R}$ stetig, $X \subset \mathbb{R}^n$ nichtleer und abgeschlossen und ist die so genannte *untere Levelmenge (Niveaumenge)*

$$L_f(x^0) := \{x \in X : f(x) \leq f(x^0)\} \tag{2.3}$$

für ein $x^0 \in X$ beschränkt, so besitzt f über X ein globales Minimum.

Beweis: Aufgrund der Stetigkeit von f ist $L_f(x^0)$ abgeschlossen und damit auch kompakt. Der Satz von Weierstraß besagt, dass dann f über $L_f(x^0)$ ein globales Minimum besitzt. Dieses ist wegen der Definition der unteren Levelmenge dann auch ein globales Minimum von f über X. □

B. Lokale Minima für unrestringierte Aufgaben

Satz (2.4) (Notwendige Bedingungen)

a) Ist $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, wir schreiben hierfür $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$, und ist $x^* \in \mathrm{int}\,(X)$ ein lokales Minimum von f über $X \subset \mathbb{R}^n$, so gilt

$$\nabla f(x^*) = 0.$$

b) Ist f sogar zweifach stetig differenzierbar, $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, so folgt unter den obigen Annahmen zudem:

$$\nabla^2 f(x^*) \geq 0 \quad \text{(positiv semidefinit)}.$$

Beweis:

zu a): Wir wenden den Taylorschen Satz auf die Funktion $\Phi(t) := f(x^* + t\,d)$ an. Dabei ist $d \in \mathbb{R}^n \setminus \{0\}$ eine beliebige Richtung und $|t|$ hinreichend klein (so dass $x^* + td \in \text{int}\,(X)$ für alle t mit $|t| < t_0$). Demnach ist

$$\Phi(t) \;=\; \Phi(0) \;+\; t\,\Phi'(0) \;+\; o(t).$$

Ausführung der Differentiation (Kettenregel!) ergibt unter Verwendung der Definition von Φ

$$f(x^* + t\,d) \;=\; f(x^*) \;+\; t\,\left(\nabla f(x^*)^T d\right) \;+\; o(t)$$

Wäre nun $\nabla f(x^*) \neq 0$, so setzen wir in der obigen Relation $d := -\nabla f(x^*)$ ein und finden

$$f(x^* + t\,d) \;=\; f(x^*) \;-\; t\,\|\nabla f(x^*)\|_2^2 \;+\; o(t) \;<\; f(x^*)$$

für alle hinreichend kleinen $t > 0$. Widerspruch!

zu b): Die Taylor–Entwicklung wird um einen Term weitergeführt:

$$\Phi(t) \;=\; \Phi(0) \;+\; t\,\Phi'(0) \;+\; (1/2)t^2\,\Phi''(0) \;+\; o(t^2).$$

Die Berechnung der zweiten Ableitung ergibt nun bei Verwendung von $\Phi'(0) = 0$

$$f(x^* + t\,d) \;=\; f(x^*) \;+\; (1/2)\,t^2\,d^T\,\nabla^2 f(x^*)\,d \;+\; o(t^2).$$

Wäre $\nabla^2 f(x^*)$ nicht positiv semidefinit, so gäbe es ein $d \in \mathbb{R}^n \setminus \{0\}$ mit $d^T\,\nabla^2 f(x^*)\,d < 0$. Widerspruch zur Minimaleigenschaft von x^*. □

Es sei angemerkt, dass die obigen notwendigen Bedingungen *nicht hinreichend* für ein lokales Minimum sind. So ist beispielsweise $x^* = 0$ ein Sattelpunkt von $f(x_1, x_2) := x_1^2 - x_2^4$, der jedoch die notwendigen Bedingungen aus (2.4) erfüllt.

Satz (2.5) (Hinreichende Bedingungen)
Ist $f \in \text{C}^2(\mathbb{R}^n, \mathbb{R})$ und $x^* \in X \subset \mathbb{R}^n$ mit

$$\nabla f(x^*) \;=\; 0, \quad \nabla^2 f(x^*) \;>\; 0 \quad \text{(positiv definit)},$$

so ist x^* ein striktes lokales Minimum von f auf X.

Beweis:

Wir verwenden das Rayleighsche Prinzip, wonach für eine *symmetrische* Matrix $A \in \mathbb{R}^{(n,n)}$ und $d \in \mathbb{R}^n \setminus \{0\}$ gilt

$$\lambda_{\min}(A) \;\leq\; \frac{d^T\,A\,d}{d^T\,d} \;\leq\; \lambda_{\max}(A),$$

wobei das Minimum bzw. Maximum jeweils für die entsprechenden Eigenvektoren d zu einem minimalen/maximalen Eigenwert angenommen wird.

Für die positiv definite, symmetrische Matrix $A = \nabla^2 f(x^*)$ gilt demnach mit $\lambda := \lambda_{\min}(A) > 0$:

$$d^T \nabla^2 f(x^*)\, d \;\geq\; \lambda \, \|d\|_2^2, \qquad \forall\, d \in \mathbb{R}^n. \tag{2.6}$$

Die Taylor–Entwicklung von f zum Entwicklungspunkt x^* (mit Restglied) ergibt

$$f(x^* + d) \;=\; f(x^*) \;+\; \nabla f(x^*)^T\, d \;+\; (1/2)\, d^T \nabla^2 f(\tilde{x})\, d$$

mit $\tilde{x} = x^* + \theta d,\ 0 < \theta < 1$.

Mit den gegebenen Voraussetzungen des Satzes und (2.6) ergibt sich

$$\begin{aligned} f(x^* + d) \;&=\; f(x^*) \;+\; (1/2)\, d^T \nabla^2 f(x^*)\, d \;+\; (1/2)\, d^T \left(\nabla^2 f(\tilde{x}) - \nabla^2 f(x^*)\right) d \\ &\geq\; f(x^*) \;+\; (1/2)\big(\lambda - \|\nabla^2 f(\tilde{x}) - \nabla^2 f(x^*)\|_2\big)\, \|d\|_2^2 \end{aligned}$$

Aus der Stetigkeit der Hesse–Matrix folgt hiermit $f(x^* + d) > f(x^*)$ für alle $d \neq 0$ aus einer hinreichend kleinen Nullumgebung. □

Anmerkungen (2.7)

a) Da unter den Voraussetzungen des Satzes (2.5) die Hesse-Matrix von f auch in einer ganzen Umgebung $K_\varepsilon(x^*)$ positiv definit ist, folgt die Existenz einer (von ε abhängigen) Konstanten $\mu > 0$ mit

$$f(x^* + d) - f(x^*) \;=\; (1/2)\, d^T \nabla^2 f(\tilde{x})\, d \;\geq\; (1/2)\, \mu \,\|\, d \,\|_2^2$$

für alle $d \in K_\varepsilon(0)$ (vgl. (2.6)).

Die Funktion f wächst also bei x^* wenigstens *quadratisch* mit $\|x - x^*\|$ an!

b) Die hinreichenden Bedingungen aus (2.5) sind i. Allg. *nicht notwendig*. Ein Gegenbeispiel liefert die Funktion $f(x_1, x_2) := x_1^2 + x_2^4$. Hier ist $x = 0$ ein striktes globales Minimum, die Hesse-Matrix ist jedoch nicht positiv definit.

c) Ist x^* ein *stationärer Punkt* (d.h. $\nabla f(x^*) = 0$) mit indefiniter Hesse-Matrix, so ist x^* ein *Sattelpunkt*, d.h. in jeder Umgebung von x^* existieren Punkte x^1, x^2 mit $f(x^1) < f(x^*) < f(x^2)$.

Beweis: Es gibt Eigenwerte $\lambda_1 < 0$ und $\lambda_2 > 0$. Für zugehörige Eigenvektoren d^1, d^2 gilt dann

$$d^{1\,T} \nabla^2 f(x^*)\, d^1 < 0, \quad d^{2\,T} \nabla^2 f(x^*)\, d^2 > 0.$$

Weiterer Schluss wie im Beweis zu (2.5). □

d) In einem stationären Punkt x^* sind durch Eigenvektoren zu positiven bzw. negativen Eigenwerten von $\nabla^2 f(x^*)$ Richtungen gegeben, in denen f wächst bzw. fällt.

Beispiel (2.8)

Für die Funktion $f(x,y) := y^2(x-1) + x^2(x+1)$ berechnet man

$$\nabla f(x,y) = (y^2 + 3\,x^2 + 2\,x\,,\; 2\,y\,(x-1))^{\mathrm{T}}.$$

Somit ergeben sich die stationären Punkte $(x^0,y^0) = (0,0)$ und $(x^1,y^1) = (-2/3,0)$.

Für die zugehörigen Hesse-Matrizen erhält man

$$\nabla^2 f(x^0,y^0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \text{ indefinit } \Rightarrow \begin{pmatrix} x^0 \\ y^0 \end{pmatrix} \text{ ist Sattelpunkt}$$

$$\nabla^2 f(x^1,y^1) = \begin{pmatrix} -2 & 0 \\ 0 & -\frac{10}{3} \end{pmatrix} \text{ neg. def. } \Rightarrow \begin{pmatrix} x^1 \\ y^1 \end{pmatrix} \text{ ist striktes lokes Maximum}$$

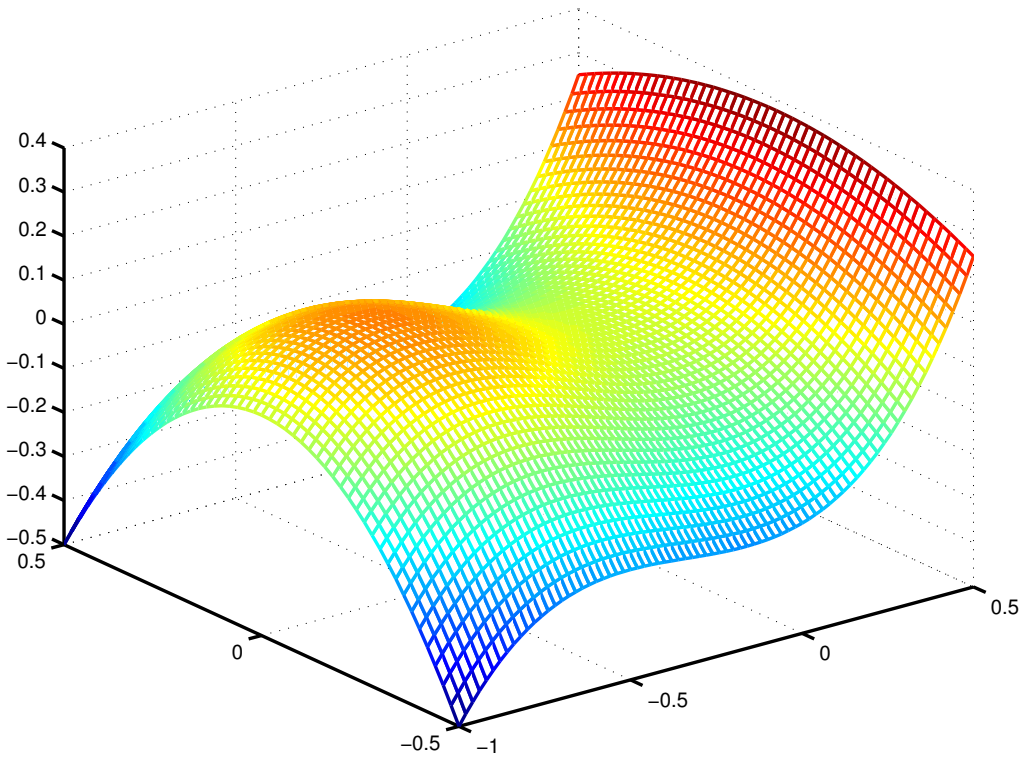


Abb. 2.1 3-D-Darstellung der Funktion f.

MATLAB Programm (2.9)

```
%
%  3-D-Darstellung der Funktion z = f(x,y)
%
[x,y] = meshgrid(-1:0.02:0.5,-0.5:0.02:0.5);
%
z = (y.^2).*(x-1) + (x.^2).*(x+1);
mesh(x,y,z)
%
print -depsc abb2_1
%
```

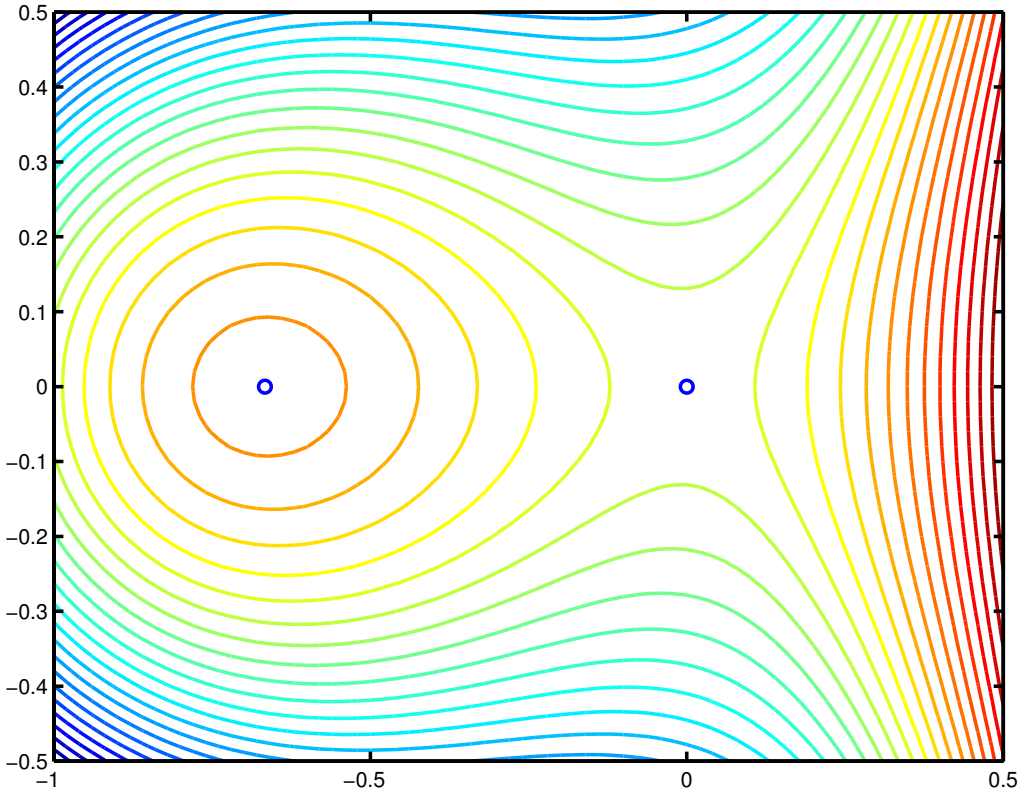


Abb. 2.2 Höhenlinien der Funktion f.

MATLAB Programm (2.10)

```
%
%  Hoehenlinie einer Funktion z = f(x,y)
%
[x,y] = meshgrid(-1:0.02:0.5,-0.5:0.02:0.5);
%
z = (y.^2).*(x-1) + (x.^2).*(x+1);
contour(x,y,z,30)
%
hold
plot(0,0,'bo')
plot(-2/3,0,'bo')
%
print -depsc abb2_2
%
```

Aufgabe (2.11)

Ein zum Testen von Optimierungsalgorithmen häufig verwendetes Beispiel ist die so genannte *Rosenbrock-Funktion*

$$f(x,y) \;=\; 100\,(y-x^2)^2 \;+\; (1-x)^2.$$

a) Zeigen Sie, dass f genau einen stationären Punkt (x^*, y^*) besitzt und dass dieser ein striktes globales Minimum von f ist.

b) Berechnen Sie die Hesse-Matrix $\nabla^2 f(x^*, y^*)$ und bestimmen Sie ihre Eigenwerte und Eigenvektoren.

c) Zeichnen Sie mit Hilfe der MATLAB Programme CONTOUR und MESH ein Höhenlinien-Diagramm (Bereich: $[-0.5, 1.5] \times [-0.5, 1.5]$) sowie eine 3D-Darstellung des Funktionsgraphen (Bereich: $[0.5, 1.5] \times [0.5, 1.5] \times [0, 10]$).

d) Ein relativ einfaches Verfahren für unrestringierte Optimierungsaufgaben ist das so genannte Simplexverfahren nach Nelder und Mead. Dieses ist im Programm FMINSEARCH unter MATLAB realisiert.
Bestimmen Sie hiermit numerisch das Minimum der Rosenbrock-Funktion.

Startvektor: $(x^0, y^0) := (-1, 1)$, Genauigkeitsforderung: TolX $:= 1e-4$.

Wie viele Iterationen benötigt das Programm? Wie groß ist die Anzahl der Funktionsauswertungen?

3. Konvexe Funktionen

Definition (3.1) Eine Menge $X \subset \mathbb{R}^n$ heißt *konvex*, falls mit zwei Punkten stets auch deren gesamte Verbindungsstrecke zu X gehört, also

$$\forall\, x, y \in X :\ \forall\, \lambda \in [0,1] :\ x \,+\, \lambda\,(y - x) \,\in\, X.$$

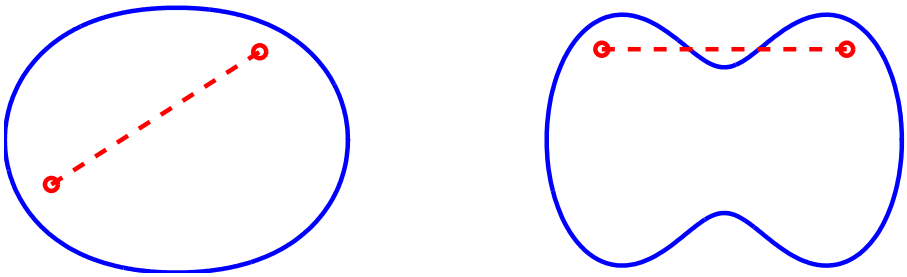


Abb. 3.1 Konvexe und nicht konvexe Menge.

Bemerkungen (3.2)

a) Beliebiger Durchschnitt konvexer Mengen ist konvex.

b) Zu jeder Menge $X \subset \mathbb{R}^n$ gibt es eine (eindeutig bestimmte) kleinste konvexe Menge $\mathrm{co}(X)$, die diese umfasst, $X \subset \mathrm{co}(X) \subset \mathbb{R}^n$. Diese heißt die *konvexe Hülle* von X.

Sie besitzt die folgenden Darstellungen:

$$\begin{aligned} \mathrm{co}(X) &= \bigcap\{K \subset \mathbb{R}^n :\ X \subset K \wedge K \text{ konvex}\} \\ &= \{\sum_{i=1}^{m} \lambda_i\, x_i :\ x_i \in X,\ \lambda_i \geq 0,\ \sum_{i=1}^{m} \lambda_i = 1\} \end{aligned}$$

Definition (3.3)

a) Sei $X \subset \mathbb{R}^n$ konvex. Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt *konvex* auf X, falls

$$\forall\, x, y \in X :\ \forall\, \lambda \in [0,1] :\ f(x \,+\, \lambda\,(y - x)) \,\leq\, f(x) \,+\, \lambda\,(f(y) - f(x)).$$

b) Die Funktion f heißt *strikt konvex* auf X, falls

$$\forall\, x \neq y \in X :\ \forall\, \lambda \in]0,1[:\ f(x \,+\, \lambda\,(y - x)) \,<\, f(x) \,+\, \lambda\,(f(y) - f(x)).$$

c) Die Funktion f heißt *gleichmäßig konvex* auf X, falls es ein $\mu > 0$ gibt mit

$$\forall\, x, y \in X : \forall\, \lambda \in [0,1] :$$
$$f(x + \lambda\,(y - x)) + \mu\,\lambda\,(1-\lambda)\,\|y - x\|^2 \leq f(x) + \lambda\,(f(y) - f(x)).$$

Aus den obigen Definitionen geht hervor, dass eine gleichmäßig konvexe Funktion auch strikt konvex ist, und eine strikt konvexe Funktion auch konvex ist.

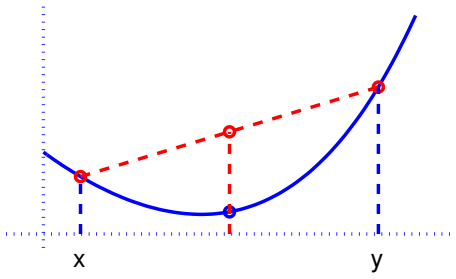


Abb. 3.2 Konvexe Funktion.

Satz (3.4) (Eindeutigkeit)

a) Ist $f : \mathbb{R}^n \to \mathbb{R}$ konvex auf einer nichtleeren, konvexen Menge X, so ist die Menge der globalen Minima konvex.

Ist f sogar strikt konvex, so gibt es höchstens ein globales Minimum von f über X (*Eindeutigkeit*).

b) Jedes lokale Minimum einer konvexen Funktion f über X (konvex) ist zugleich global.

Ist f konvex und stetig differenzierbar auf einer offenen und konvexen Menge X, so ist jeder stationäre Punkt bereits ein globales Minimum von f über X.

Beweis:

zu a): Aus $f(x) = f(y) = \inf f(X)$, $x, y \in X$ und $\lambda \in [0,1]$ folgt

$$f(x + \lambda\,(y - x)) \leq f(x) + \lambda\,(f(y) - f(x)) = \inf f(X),$$

d.h. auch $x + \lambda\,(y - x)$ ist ein globales Minimum von f über X.

Da in der obigen Ungleichung also Gleichheit gilt, folgt für strikt konvexes f somit $x = y$ und damit die behauptete Eindeutigkeit.

zu b): Nehmen wir an, x^* sei ein lokales, aber kein globales Minimum. Dann gibt es ein $y \in X$ mit $f(y) < f(x^*)$. Für $\lambda \in]0,1]$ folgt

$$f(x^* + \lambda(y - x^*)) \leq f(x^*) + \lambda(f(y) - f(x^*)) < f(x^*) + \lambda(f(x^*) - f(x^*)) = f(x^*)$$

Damit gibt es in jeder Umgebung von x^* Punkte mit kleinerem Zielfunktionswert als dem in x^*. Damit kann x^* aber kein lokales Minimum sein. Widerspruch!

Zur zweiten Aussage betrachten wir zu $y \in X$ die Hilfsfunktion

$$\Psi(\lambda) \;:=\; f(x^*) + \lambda(f(y) - f(x^*)) \;-\; f(x^* + \lambda(y - x^*)) \;\geq\; 0.$$

Ψ ist stetig differenzierbar und nicht negativ auf dem Intervall $[0, 1]$. Ferner gilt $\Psi(0) = 0$. Hiermit folgt

$$\Psi'(0) \;=\; (f(y) - f(x^*)) \;-\; \nabla f(x^*)^{\mathrm{T}}(y - x^*) \;\geq\; 0.$$

Gilt also $\nabla f(x^*) = 0$, so ist $f(y) - f(x^*) \geq 0$ für alle $y \in X$. Damit ist gezeigt, dass x^* ein globale Minimum von f über X ist. □

Die Definition (3.3) der Konvexität einer Funktion f eignet sich meist nicht so gut zur konkreten Überprüfung dieser Eigenschaft. Falls f jedoch gewisse Glattheitseigenschaften besitzt, lassen sich die folgenden Charakterisierungen der Konvexität verwenden.

Satz (3.5)

a) Für $f \in \mathrm{C}^1(\mathbb{R}^n)$ und $X \subset \mathbb{R}^n$ konvex und nichtleer gilt:

$$f \text{ konvex auf } X \iff \forall\, x, y \in X :\; f(y) \;\geq\; f(x) \;+\; \nabla f(x)^{\mathrm{T}}(y - x)$$

Ferner ist f genau dann strikt konvex, wenn die obige Ungleichung für $x \neq y$ strikt erfüllt ist.

b) Für $f \in \mathrm{C}^2(\mathbb{R}^n)$ und $X \subset \mathbb{R}^n$ offen, konvex und nichtleer gilt:

$$f \text{ konvex auf } X \iff \forall\, x \in X :\; \nabla^2 f(x) \text{ positiv semidefinit}$$

Ferner: Ist die Hesse-Matrix $\nabla^2 f(x)$ sogar positiv definit auf X, so ist f strikt konvex.

Beweis:

zu a):

(i) $\Rightarrow$: Der Beweis erfolgt wie der zu Satz (3.4) b). Die Funktion

$$\Psi(\lambda) \;:=\; f(x) + \lambda(f(y) - f(x)) \;-\; f(x + \lambda(y - x)) \;\geq\; 0,\;\; \lambda \in [0, 1],$$

ist stetig differenzierbar mit $\Psi(0) = 0$. Daher ist auch

$$\Psi'(0) \;=\; f(y) - f(x) \;-\; \nabla f(x)^{\mathrm{T}}(y - x) \;\geq\; 0.$$

(i) $\Leftarrow$**:** Mit $\bar{x} := x + \lambda(y - x)$ gelten nach Voraussetzung

$$\begin{aligned} f(x) &\geq f(\bar{x}) + \nabla f(\bar{x})^{\mathrm{T}}(x - \bar{x}) \\ f(y) &\geq f(\bar{x}) + \nabla f(\bar{x})^{\mathrm{T}}(y - \bar{x}) \end{aligned}$$

Multipliziert man die erste Ungleichung mit $(1 - \lambda)$, die zweite mit λ und addiert, so ergibt sich die Behauptung

$$(1 - \lambda)\, f(x) + \lambda\, f(y) \geq f(\bar{x}) + 0 = f(x + \lambda(y - x)).$$

(ii) $\Rightarrow$**:** Für $x \neq y$ sei $z := (1/2)(x + y) = x + (1/2)(y - x)$, $\lambda = 1/2$.
Mit (i) und der strikten Konvexität folgt dann:

$$\begin{aligned} \nabla f(x)^{\mathrm{T}}(y - x) &= 2\, \nabla f(x)^{\mathrm{T}}(z - x) \leq 2\,(f(z) - f(x)) \\ &< 2\,\big[f(x) + (1/2)(f(y) - f(x)) - f(x)\big] = f(y) - f(x) \end{aligned}$$

(ii) $\Leftarrow$**:** Analog zum Beweis von (i) $\Leftarrow$.

zu b):

(i) $\Rightarrow$**:** Nach a) und dem Taylorschen Satz gilt für $x \in X$, $d \in \mathbb{R}^n$ und $t > 0$ hinreichend klein

$$0 \leq f(x + t\,d) - f(x) - t\, \nabla f(x)^{\mathrm{T}} d = \frac{t^2}{2}\, d^{\mathrm{T}} \nabla^2 f(x)\, d + o(t^2).$$

Division durch $t^2/2 > 0$ und Grenzwertbildung $t \downarrow 0$ ergibt $d^{\mathrm{T}} \nabla^2 f(x)\, d \geq 0$, also die positive Semidefinitheit der Hesse-Matrix.
Man beachte, dass mit $x \in X$ (offen!) auch $x + td \in X$ gilt für alle hinreichend kleinen $t > 0$.

(i) $\Leftarrow$**:** Nach dem Taylorschen Satz gilt für $x, y \in X$:

$$f(y) = f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + (1/2)(y - x)^{\mathrm{T}}\, \nabla^2 f(x + \theta(y - x))\, (y - x)$$

mit einem (von x, y abhängigen) $\theta \in]0, 1[$.

Ist die Hesse-Matrix nun positiv semidefinit, so folgt

$$f(y) \geq f(x) + \nabla f(x)^{\mathrm{T}}(y - x)$$

und damit die Konvexität von f über X nach a).

(ii): Aus der positiven Definitheit der Hesse-Matrix folgt analog für $x \neq y$:

$$f(y) > f(x) + \nabla f(x)^{\mathrm{T}}(y - x)$$

und damit die strikte Konvexität von f über X. □

Bemerkungen (3.6)

a) Man beachte, dass die Umkehrung der Zusatzaussage in Satz (3.5) b) i. Allg. falsch ist, d.h. für eine strikt konvexe C^2–Funktion muss die Hesse-Matrix nicht überall positiv definit sein. Ein Gegenbeispiel ist etwa $f(x) := x^4$.

b) Mit der gleichen Beweistechnik wie in Satz (3.5) lässt sich auch eine Charakterisierung der gleichmäßigen Konvexität beweisen.

Für $f \in C^1(\mathbb{R}^n, \mathbb{R})$ und $\emptyset \neq X \subset \mathbb{R}^n$ konvex gilt: f ist genau dann gleichmäßig konvex, wenn es ein $\mu > 0$ gibt, so dass

$$\forall\, x, y \in X: \quad f(y) \;\geq\; f(x) \;+\; \nabla f(x)^{\mathrm{T}}(y - x) \;+\; \mu\, \|y - x\|^2.$$

Für $f \in C^2(\mathbb{R}^n, \mathbb{R})$ und $\emptyset \neq X \subset \mathbb{R}^n$ konvex und offen gilt: f ist genau dann gleichmäßig konvex, wenn es ein $\mu > 0$ gibt, so dass

$$\forall\, x \in X,\, d \in \mathbb{R}^n: \quad d^{\mathrm{T}}\, \nabla^2 f(x)\, d \;\geq\; 2\,\mu\, \|d\|^2.$$

Beispiel (3.7)

Eine quadratische Funktion $f: \mathbb{R}^n \to \mathbb{R}$ hat die Form

$$f(x) \;=\; \frac{1}{2}\, x^{\mathrm{T}} A\, x \;+\; b^{\mathrm{T}} x \;+\; c$$

wobei $A \in \mathbb{R}^{(n,n)}$ eine *symmetrische* Matrix ist, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$.

f ist offenbar eine C^∞-Funktion mit

$$\nabla f(x) \;=\; A\, x \;+\; b, \qquad \nabla^2 f(x) \;=\; A.$$

Damit ist f genau dann konvex, wenn A positiv semidefinit ist. Ist A sogar positiv definit, so ist f strikt konvex und besitzt ein eindeutig bestimmtes striktes, globales Minimum x^*, das sich über das lineare Gleichungssystem

$$\nabla f(x^*) \;=\; A\, x^* \;+\; b \;=\; 0$$

berechnen lässt.

Satz (3.8) (Existenz und Eindeutigkeit)

Ist für $f \in C^1(\mathbb{R}^n, \mathbb{R})$ und $x^0 \in \mathbb{R}^n$ die untere Levelmenge $L_f(x^0)$ konvex und ist f gleichmäßig konvex auf $L_f(x^0)$, so besitzt f ein eindeutig bestimmtes (striktes) globales Minimum.

Beweis:

Aus der Definition der gleichmäßigen Konvexität folgt für $x \in L_f(x^0)$ und $\lambda = 1/2$

$$f(x^0 + \frac{1}{2}(x - x^0)) + \frac{\mu}{4}\|x - x^0\|^2 \leq f(x^0) + \frac{1}{2}\left(f(x) - f(x^0)\right).$$

Aus $f(x) \leq f(x^0)$, der Konvexität von f auf $L_f(x^0)$, (3.5a) sowie der Cauchy-Schwarzschen Ungleichung ergibt sich hieraus

$$\begin{aligned}
\frac{\mu}{4}\|x - x^0\|^2 &\leq f(x^0) + \frac{1}{2}\left(f(x) - f(x^0)\right) - f(x^0 + \frac{1}{2}(x - x^0)) \\
&\leq f(x^0) - f(x^0 + \frac{1}{2}(x - x^0)) \\
&\leq -\frac{1}{2}\nabla f(x^0)^{\mathrm{T}}(x - x^0) \\
&\leq \frac{1}{2}\|\nabla f(x^0)\|\,\|x - x^0\|,
\end{aligned}$$

und damit die Abschätzung $\|x - x^0\| \leq (2/\mu)\|\nabla f(x^0)\|$.

Somit ist die untere Levelmenge $L_f(x^0)$ also beschränkt und damit auch kompakt. Die Behauptung ergibt sich dann als Folgerung aus (2.2) und (3.4). □

Aufgabe (3.9)

Sei $X \subset \mathbb{R}^n$ eine nichtleere und abgeschlosse Menge. Zu festem $y \in \mathbb{R}^n$ werde definiert $f(x) := \|x - y\|_2, \quad x \in \mathbb{R}^n$.

Zeigen Sie: f besitzt auf X ein globales Minimum. Ist X konvex, so ist das globale Minimum eindeutig bestimmt.

4. Simplexmethode nach Nelder, Mead

In der Regel verlangen Optimierungsverfahren eine gewisse Glattheit, etwa zweifache stetige Differenzierbarkeit von Zielfunktion und Restriktionen. Eine Ausnahme hiervon bilden die so genannten *direkten Suchmethoden*, die keine Glattheit der Zielfunktion verlangen, mit einfachen Heuristiken arbeiten, aber i. Allg. auch wenig effizient sind.

Wir beschreiben die auf Nelder und Mead (1965) zurückgehende *Simplexmethode* (auch *Polyedermethode* genannt) zur Lösung einer unrestringierten Optimierungsaufgabe. Dieses Verfahren ist auch unter MATLAB realisiert (Programm FMINSEARCH) und erfreut sich unter Anwendern einer gewissen Beliebtheit.

Problem (4.1) Minimiere $f(x),\ x \in \mathbb{R}^n,$ mit nur stetigem $f \in \mathrm{C}(\mathbb{R}^n, \mathbb{R})$, $n \geq 2$.

Definition (4.2) Unter einem *Simplex (Polyeder)* verstehen wir eine $(n+1)$–elementige Punktmenge (das sind die Ecken des Simplex)

$$\Delta \ = \ \{x^1, \ldots, x^{n+1}\} \ \subset \ \mathbb{R}^n$$

in allgemeiner Lage, d.h. die von x^1 ausgehenden Kanten $(x^2 - x^1, \ldots, x^{n+1} - x^1)$ sollen linear unabhängig sein.

In der Geometrie wird i. Allg. die konvexe Hülle von Δ als Simplex bezeichnet. Für $n = 2$ ist dies ein nichtentartetes Dreieck, für $n = 3$ ein nichtentartetes Tetraeder.

Iterationsschritt der Simplexmethode: $(n \geq 2)$

a) Man bestimme Indizes $s,\ a,\ b \in \{1, \ldots, n+1\}$

$x^s \ := \ \mathrm{argmax}\{f(x) : \ x \in \Delta\}$ (Schlechtester Punkt)

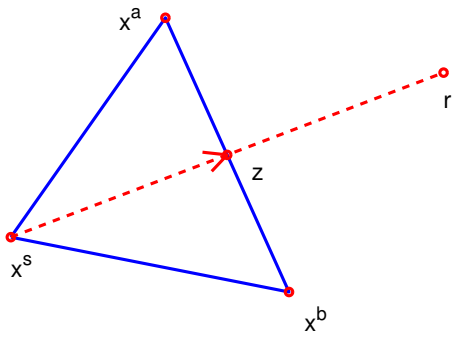
$x^a \ := \ \mathrm{argmax}\{f(x) : \ x \in \Delta,\ a \neq s\}$ (Zweitschlechtester Punkt)

$x^b \ := \ \mathrm{argmin}\{f(x) : \ x \in \Delta,\ b \neq s,\ a\}$ (Bester Punkt)

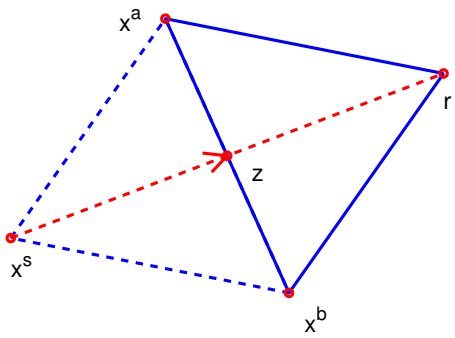
b) Weiterhin bestimme man das Zentrum der n besten Punkte: $z \ := \ \dfrac{1}{n} \displaystyle\sum_{i \neq s} x^i,$

und den am Zentrum reflektierten schlechtesten Punkt: $r \ := \ z \ + \ \alpha\,(z - x^s),$

Hierbei setzt man i. Allg. $\alpha = 1$.

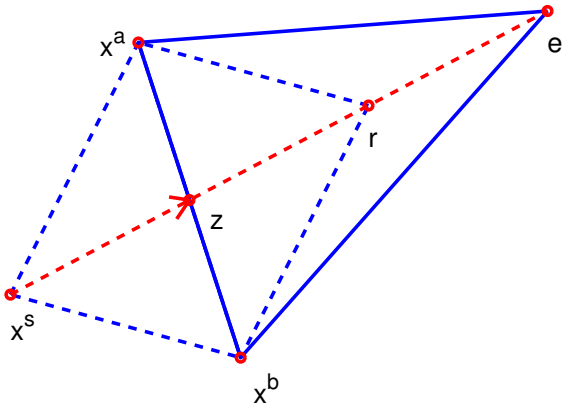


c) Falls $f(x^b) \leq f(r) \leq f(x^a)$ gilt, so ersetze man x^s durch r.



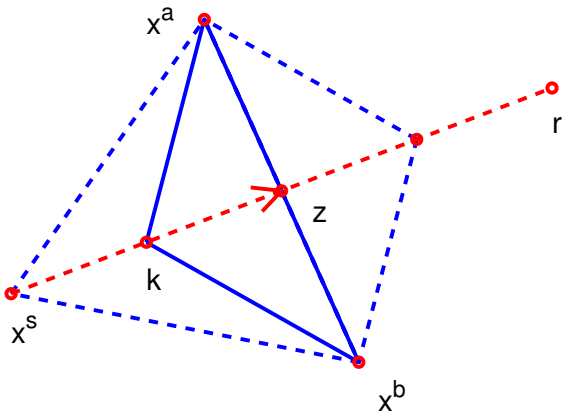
d) Falls sogar $f(r) < f(x^b)$ gilt, berechne man den *expandierten Punkt* $e := z + \beta\,(z - x^s)$, dabei sei $\beta > \alpha$. I. Allg. setzt man $\beta = 2$.

Falls $f(e) < f(r)$ gilt, ersetze man x^s durch e, andernfalls ersetze man wie in c) x^s durch r.



e) Falls $f(r) > f(x^a)$: Berechne einen *kontrahierten Punkt* k:

Falls $f(r) > f(x^s)$:	$k := z + \gamma\,(x^s - z)$,
andernfalls:	$k := z + \gamma\,(z - x^s)$,
dabei ist $0 < \gamma < \alpha$,	i. Allg. setzt man $\gamma = 1/2$.
Falls $f(k) < f(x^s)$:	Ersetze x^s durch k,
andernfalls:	Kontrahiere den Simplex um x^b: $x^i := (x^i + x^b)/2$.



Damit ist ein Schritt des Simplexverfahrens beschrieben. Dieser wird nun iteriert bis sich die Simplexe (hoffentlich) einem lokalen Minimum annähern.

Abbruchkriterien (4.3)

- $\forall\, j,\, k: \quad \|x^j - x^k\| \;\leq\; \mathrm{TOL}.$
- $\dfrac{1}{n+1}\displaystyle\sum_{i=1}^{n+1}(f(x^i) - f(z))^2 \;\leq\; \mathrm{TOL}^2.$
- $\dfrac{1}{n+1}\displaystyle\sum_{i=1}^{n+1}(f(x^i) - \bar{f})^2 \;\leq\; \mathrm{TOL}^2, \quad \bar{f} := \dfrac{1}{n+1}\displaystyle\sum_{i=1}^{n+1} f(x^i).$

Startsimplex (4.4)

Ist eine Ecke x^1 und die Kantenlänge c vorgegeben, so lässt sich ein regelmäßiges Simplex mit der Kantenlänge c folgendermaßen konstruieren:

$$\begin{aligned} p &:= \frac{c}{\sqrt{2}} \frac{\sqrt{n+1} - 1}{n} \\ q &:= p\,(1,\ldots,1)^{\mathrm{T}} \in \mathbb{R}^n \\ x^j &:= x^1 + q + \frac{c}{\sqrt{2}}\, e^{j-1},\ j = 2,\ldots,n+1. \end{aligned}$$

Die $e^1,\ldots,e^n$ bezeichnen dabei die kanonischen Einheitsvektoren im $\mathbb{R}^n$.

Beispiele (4.5)

- $f(x,y) = x^2\,y - 0.25\,(2x^2 - y^2) + 0.5\,(2 - x^2 - y^2)^2$
  Startdreieck: $(-3,-2)$, $(-2,-2)$, $(-3,-1.5)$
- $f(x,y) = 100\,(y - x^2)^2 + (1-x)^2$ (Rosenbrock-Funktion)
  Startdreieck: Gemäß (4.4) mit $x^1 = (-2,2)^{\mathrm{T}}$ und $c = 1$.

Literatur:

Nelder, J.A. und R. Mead: *A Simplex Method for Function Minimization.* Computer Journal, 7; 308-313, 1965.

Lagarius, J.C, Reeds, J.A., Wright, M.H. und P.E. Wright: *Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions.* SIAM J. Optim. 9, 112-147, 1998.

Price, C.J., Coope, I.D. und D. Byatt: *A Convergent Variant of the Nelder-Mead Algorithm.* J. of Optim. Theory and Appl. 113, 5-19, 2002.

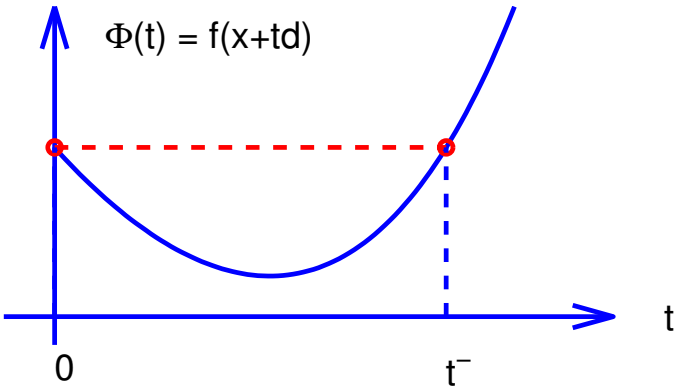
5. Abstiegsverfahren

Wir betrachten wieder eine unrestringierte Minimierungsaufgabe

Problem (5.1) Minimiere $f(x), \quad x \in \mathbb{R}^n,$

wobei $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$ eine stetig differenzierbare Zielfunktion sei.

Definition (5.2) Ein Vektor $d \in \mathbb{R}^n,$ heißt eine *Abstiegsrichtung* von f in einem Punkt $x \in \mathbb{R}^n,$ falls es ein $\bar{t} > 0$ gibt mit $f(x + t\,d) < f(x)$ für alle $t \in]0, \bar{t}[.$



Satz (5.3) Gilt $\nabla f(x)^{\mathrm{T}} d < 0,$ so ist d eine Abstiegsrichtung von f in $x.$

Beweis: Für die C^1–Funktion $\Phi(t) := f(x + t\,d)$ gilt $\Phi'(0) = \nabla f(x)^{\mathrm{T}} d < 0.$ Hieraus folgt unmittelbar die Behauptung. □

Beispiel: Ist $B \in \mathbb{R}^{(n,n)}$ symmetrisch und positiv definit, so ist $d = -B\,\nabla f(x)$ eine Abstiegsrichtung von f in $x,$ falls $\nabla f(x) \neq 0.$

Algorithmus (5.4)

1.) Wähle Startvektor $x^0 \in \mathbb{R}^n, \quad k = 0;$

2.) Genügt x^k einem Abbruchkriterium: STOP;

3.) Bestimme eine Abstiegsrichtung d^k mit $\nabla f(x^k)^{\mathrm{T}} d^k < 0;$

4.) Bestimme eine Schrittweite t_k mit $f(x^k + t_k\,d^k) < f(x^k);$

5.) $x^{k+1} := x^k + t_k\,d^k; \quad k := k + 1;$ gehe zu Punkt 2.);

Der obige Abstiegsalgorithmus lässt viele Freiheiten für die Bestimmung der Abstiegsrichtung d^k und der Schrittweite t_k.

Eine Standardwahl für die Wahl von d^k, nämlich

$$d^k \;=\; d^k_G \;:=\; -\nabla f(x^k), \tag{5.5}$$

liefert das *Verfahren des steilsten Abstiegs (steepest descend)* oder auch *Gradientenverfahren.*

Beim *Newton-Verfahren* (Nullstellenbestimmung für $\nabla f(x)$) wird dagegen

$$d^k \;=\; d^k_N \;:=\; -\left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) \tag{5.6}$$

gewählt. In allen Punkten x^k, in denen die Hesse-Matrix $\nabla^2 f(x^k)$ positiv definit ist, ist die Newton-Richtung auch eine Abstiegsrichtung, sofern $\nabla f(x^k) \neq 0$ vorausgesetzt wird.

Schließlich setzt man beim *Quasi-Newton-Verfahren*

$$d^k \;=\; d^k_{QN} \;:=\; -H_k^{-1}\, \nabla f(x^k), \tag{5.7}$$

wobei H_k eine (geeignet gewählte) positiv definite Matrix bezeichnet.

Satz (5.8) Ist $H \in \mathbb{R}^{(n,n)}$ symmetrisch und positiv definit, so minimiert die (skalierte und normierte) Gradientenrichtung

$$d \;:=\; -\frac{H^{-1}\nabla f(x)}{\|H^{-1}\nabla f(x)\|_H}$$

den Anstieg $\nabla f(x)^{\mathrm{T}} d$ über alle Richtungen $d \in \mathbb{R}^n$ mit $\|d\|_H = 1$.
Dabei sei $\nabla f(x) \neq 0$ und die (skalierte) Norm gegeben durch $\|x\|_H := (x^{\mathrm{T}} H\, x)^{1/2}$.

Beweis: (für $H = I_n$)

Mit der Cauchy-Schwarzschen Ungleichung folgt für $\|d\|_2 = 1$:

$$|\nabla f(x)^{\mathrm{T}} d| \;\leq\; \|\nabla f(x)\|_2\, \|d\|_2 \;=\; \|\nabla f(x)\|_2.$$

Diese Schranke wird gerade für $d = \pm\nabla f(x)/\|\nabla f(x)\|_2$ angenommen. □

Beispiel (5.9)

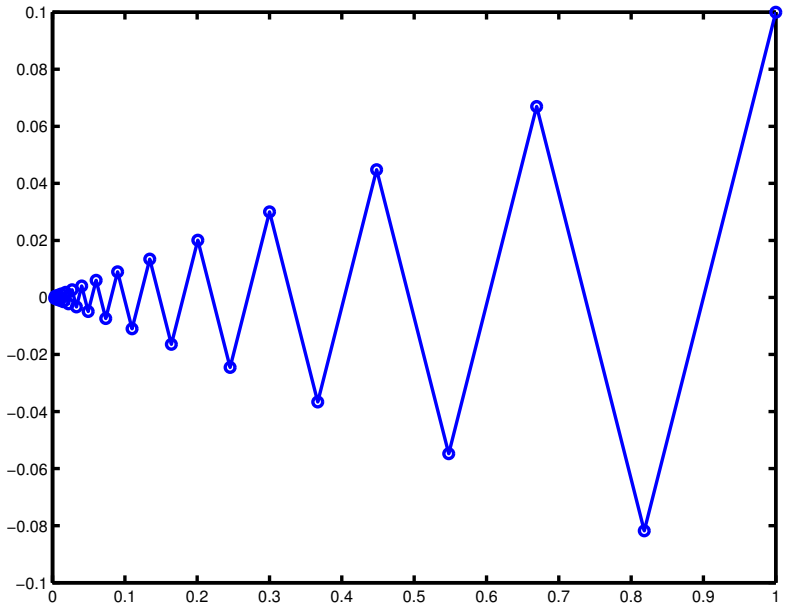
Wir verwenden das (unskalierte) Gradientenverfahren zur Minimierung der Funktion $f(x_1, x_2) := x_1^2 + 10\, x_2^2$. Die Bestimmung der Schrittweite erfolge mit *exakter Liniensuche*, d.h. $t := \mathrm{argmin}\{f(x + t\, d) : t > 0\}$.

Wir erhalten

$$\begin{aligned}
\nabla f(x_1, x_2) &= (2x_1, 20x_2)^{\mathrm{T}} \\
d &= -(2x_1, 20x_2)^{\mathrm{T}} \\
\Phi(t) &= (x_1 + td_1)^2 \; + \; 10\,(x_2 + td_2)^2 \\
\Phi'(t) &= 2\,(x_1 + td_1)\,d_1 \; + \; 20\,(x_2 + td_2)\,d_2 \; = \; 0(!)
\end{aligned}$$

Damit: $t := -\dfrac{x_1 d_1 + 10\, x_2 d_2}{d_1^2 + 10 d_2^2}, \quad x_i^{\text{neu}} := x_i + t\, d_i, \; (i = 1, 2).$

Mit dem Startvektor $x^0 := (1, 0.1)^{\mathrm{T}}$ benötigt das Verfahren 63 Iterationen um das Abbruchkriterium $\|\nabla f(x)\|_2 \leq 10^{-5}$ zu erfüllen.



Würden wir dagegen die in Satz 5.7 angegebene skalierte Gradientenrichtung verwenden mit einer geschickt gewählten Matrix

$$H \;=\; \begin{pmatrix} 2 & 0 \\ 0 & 20 \end{pmatrix},$$

so würde sich ergeben

$$d \;=\; -H^{-1}\nabla f(x) \;=\; -\begin{pmatrix} 1/2 & 0 \\ 0 & 1/20 \end{pmatrix}\begin{pmatrix} 2\,x_1 \\ 20\,x_2 \end{pmatrix} \;=\; -\,x,$$

sowie (bei exakter Liniensuche) $t = 1$, so dass schon der erste Iterationsschritt bereits die exakte Lösung $x^* = 0$ liefert.

Die Winkelbedingung.

Die geometrische Bedeutung der (hinreichenden) Abstiegsbedingung aus Satz (5.3) besagt, dass der Winkel γ zwischen Abstiegsrichtung und negativer Gradientenrichtung kleiner als $\pi/2$ ist. Um Konvergenzaussagen zu gewinnen, ist es daher naheliegend zu verlangen, dass der Winkel γ während der gesamten Iteration von $\pi/2$ weg beschränkt ist.

Wegen $\cos\gamma = -\nabla f(x)^{\mathrm{T}} d/(\|\nabla f(x)\|\|d\|)$ besagt dies, dass wir für alle Iterierten x^k, d^k des Algorithmus (5.4) verlangen, dass die folgende **Winkelbedingung** erfüllt ist.

$$\exists\, c > 0 :\ \forall k \in \mathbb{N} :\ \ c_k \ := \ -\,\frac{\nabla f(x^k)^{\mathrm{T}} d^k}{\|\nabla f(x^k)\|\ \|d^k\|} \ \ \geq \ \ c. \tag{5.10}$$

Mitunter wird anstelle von (5.10) auch die etwas schwächere *Zoutendijk–Bedingung* $\sum_{k=0}^{\infty} c_k^2 = \infty$ verwendet.

Schrittweitenstrategie.

Wir verlangen, dass durch eine geeignete Schrittweitenwahl $t(x,d)$ der Funktionswert $f(x + t\,d)$ in ausreichendem Maße verkleinert wird. Genauer: Wir suchen eine Aussage darüber, um wieviel man $f(x)$ in Richtung $x + t\,d$ (d vorgegebene Abstiegsrichtung) verkleinern kann.

Dazu nehmen wir an, dass $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$ ist und dass die Levelmenge $L_f(x^0)$ zum Startvektor x^0 kompakt ist, vgl. (2.2). $x = x^k$ sei eine Iterierte aus $L_f(x^0)$. Da die Hesse-Matrix $\nabla^2 f(x)$ auf $L_f(x^0)$ stetig ist, gibt es eine Konstante $C > 0$ mit

$$\|\nabla^2 f(x)\|_2 \ \leq \ C, \quad \forall\, x \in L_f(x^0).$$

Mittels Taylor-Entwicklung folgt

$$\begin{aligned} f(x + t\,d) \ &= \ f(x) \ + \ t\,\nabla f(x)^{\mathrm{T}} d \ + \ (t^2/2)\, d^{\mathrm{T}} \nabla^2 f(z)\, d \\ &\leq \ f(x) \ + \ t\,\nabla f(x)^{\mathrm{T}} d \ + \ (t^2/2)\, C\, \|d\|_2^2\,. \end{aligned} \tag{5.11}$$

Dabei ist $z = x + \Theta\, t\, d$ eine Zwischenstelle, $0 < \Theta < 1$. Die Abschätzung (5.11) gilt für alle $t > 0$ mit $x + [0,t]\, d \subset L_f(x^0)$.

Nun ist die rechte Seite dieser Abschätzung ein Polynom $p(t)$ zweiten Grades in t. Dieses besitzt in

$$t^* \ := \ -\,\frac{\nabla f(x)^{\mathrm{T}} d}{C\, \|d\|_2^2} \ > \ 0$$

ein striktes globales Minimum.

Ist $\hat{t}$ die (eindeutig bestimmte) maximale Schrittweite mit $\forall t \in [0, \hat{t}] :\ x + t\,d \in L_f(x)$, so folgt

$$p(\hat{t}) \ \geq \ f(x + \hat{t}\, d) \ = \ f(x) \ = \ p(0).$$

Wegen $p'(0) < 0$ folgt hieraus, dass t^* im offenen Intervall $]0, \hat{t}[$ liegt, insbesondere also zulässig ist, $x + t^* d \in L_f(x)$.

Mit der Abschätzung (5.11) folgt hieraus

$$f(x + t^* d) \;\leq\; p(t^*) \;=\; f(x) \;-\; \frac{1}{2\,C} \left(\frac{\nabla f(x)^{\mathrm{T}} d}{\|d\|} \right)^2 .$$

Diese Abschätzung zeigt, dass ein Mindestabstieg der obigen Gestalt unter den genannten Voraussetzungen möglich ist. Da man aber i. Allg. die Schranke C für die Hesse-Matrix nicht kennt, definiert man etwas vorsichtiger:

Definition (5.12)

Eine Schrittweitenstrategie $t = t(x, d)$ heißt **effizient**, wenn es zu $x^0 \in \mathbb{R}^n$ ein $\Theta > 0$ gibt mit

$$f(x + t(x,d)\, d) \;\leq\; f(x) \;-\; \Theta \left(\frac{\nabla f(x)^{\mathrm{T}} d}{\|d\|} \right)^2$$

für alle $x \in L_f(x^0)$ und d Abstiegsrichtung von f in x mit $\nabla f(x)^{\mathrm{T}} d < 0$.

Bemerkung (5.13)

Unter den obigen Voraussetzungen ($f \in \mathrm{C}^2$, $L_f(x^0)$ kompakt) ist die exakte Schrittweitenstrategie $t := \operatorname{argmin}\{ f(x + t\, d) : t > 0 \}$ effizient.

Satz (5.14) (Konvergenzsatz 1)

Sei $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$, $x^0 \in \mathbb{R}^n$. Das Abstiegsverfahren (5.4) erfülle die Winkelbedingung und arbeite mit einer effizienten Schrittweitenstrategie. Dann trifft eine der folgenden drei Möglichkeiten zu:

- Abbruch nach endlich vielen Iterationen mit $\nabla f(x^k) = 0$.
- $f(x^k) \to -\infty$ $(k \to \infty)$.
- $\nabla f(x^k) \to 0$ $(k \to \infty)$, d.h. jeder Häufungspunkt der Folge (x^k) ist ein stationärer Punkt von f.

Beweis:

Zunächst folgt aus der Effizienz und der Winkelbedingung (5.10)

$$f(x^{k+1}) - f(x^k) \;\leq\; -\,\Theta \left(\frac{\nabla f(x^k)^{\mathrm{T}} d^k}{\|d^k\|} \right)^2 \;=\; -\,\Theta\, c_k^2\, \|\nabla f(x^k)\|_2^2 .$$

Bricht die Iteration nun nicht nach endlich vielen Schritten ab, so folgt für beliebige $N \in \mathbb{N}$:

$$f(x^N) - f(x^0) \;=\; \sum_{k=0}^{N-1} \left(f(x^{k+1}) - f(x^k)\right) \;\leq\; -\,\Theta \sum_{k=0}^{N-1} c_k^2 \,\|\nabla f(x^k)\|_2^2$$

Ist f zudem nach unten beschränkt, so ergibt sich

$$\sum_{k=0}^{\infty} c_k^2 \,\|\nabla f(x^k)\|^2 \;<\; \infty$$

Mit der Winkelbedingung (5.10) folgt hieraus $\|\nabla f(x^k)\| \to 0$. □

Wir zitieren ohne Beweis eine Variante des obigen Konvergenzsatzes, die mit stärkeren Voraussetzungen arbeitet, vgl. Geiger, Kanzow.

Satz (5.15) (Konvergenzsatz 2)

Sei $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$, $x^0 \in \mathbb{R}^n$, die Levelmenge $L_f(x^0)$ sei konvex und f sei gleichmäßig konvex auf $L_f(x^0)$, d.h.

$$f(\lambda x + (1-\lambda)y) \;\leq\; \lambda f(x) + (1-\lambda) f(y) - \mu\lambda(1-\lambda)\|x-y\|^2$$

für ein festes $\mu > 0$, und allen $x, y \in L_f(x^0)$ und $\lambda \in [0,1]$.
Erfüllt das Abstiegsverfahren (5.4) nun die Zoutendijk Bedingung und arbeitet es mit einer effizienten Schrittweitenstrategie, so konvergiert die Folge (x^k) gegen das eindeutig bestimmte globale Minimum von f, vgl. auch (3.8).

Abbruchkriterien.

Von Gill, Murray und Wright werden die folgenden Abbruchkriterien für unrestrigierte Optimierungsaufgaben vorgeschlagen. Hierbei bezeichne TOL eine, vom Benutzer vorzugebende, *relative Genauigkeitsschranke*:

$$\mathrm{TOL} \;\approx\; 10^{-r},$$

wenn etwa r gültige Dezimalziffern gewünscht werden. Man beachte,

- dass es sinnlos ist, TOL zu klein zu wählen. So sollte zumindest TOL $\geq$ EPMACH sein, wobei EPMACH die relative Maschinengenauigkeit bezeichnet.
- dass sowohl die Lösung x (bzw. gewisse Komponenten von x) wie auch die Zielfunktion $f(x)$ im Lösungspunkt verschwinden können. In diesem Fall gibt es keine *relative* sondern nur eine *absolute* Genauigkeit.

Abbruchkriterien (5.16)

(K1) $f(x^{k-1}) - f(x^k) \leq \mathrm{TOL}\,(1 + |f(x^k)|)$

(K2) $\| x^{k-1} - x^k \| \leq \sqrt{\mathrm{TOL}}\,(1 + \| x^k \|)$

(K3) $\| \nabla f(x^k) \| \leq \sqrt[3]{\mathrm{TOL}}\,(1 + |f(x^k)|)$

(K4) $\| \nabla f(x^k) \| \leq \mathrm{EPMACH}$

(K5) $k \geq k_{\max}$

Erläuterungen (5.17)

- (K4) und (K5) sind *Notbremsen*. Man sollte diese unbedingt einbauen und auch eine entsprechende Warnung ausgeben, falls eine dieser Bedingungen aktiv wird.
- (K1), (K2), (K3) sind Kriterien für ein erfolgreich behandeltes Problem (bei nicht zu kleinem TOL). Alle Abfragen sind relativ. Durch die Fakoren $(1+\ldots)$ soll von einer relativen auf eine absolute Genauigkeit umgeschaltet werden, falls $\|x\|$ bzw. $|f(x)|$ klein sind. Nach Gill et al. sollten alle drei Bedingungen bei Abbruch erfüllt werden.
- Die Wurzel $\sqrt{\mathrm{TOL}}$ in (K2) erklärt sich durch die Taylor-Entwicklung in der Nähe des Lösungspunktes ($\nabla f(x^k) \approx 0$):
$$f(x^{k-1}) \approx f(x^k) + O(\| x^{k-1} - x^k \|^2)$$
- Die dritte Wurzel in der Bedingung (K3) soll laut Gill et al. die Bedingung gegenüber der theoretisch begündbaren zweiten Wurzel abschwächen. Die zweite Wurzel wäre zu restriktiv.

Aufgabe (5.18)

a) Modifizieren Sie den Beweis von Satz (5.8) für eine beliebige symmetrische und positiv definite Matrix $H \in \mathbb{R}^{(n,n)}$.

b) Zeigen Sie, dass die Suchrichtung $d := -H^{-1}\nabla f(x)$ die Winkelbedingung erfüllt:
$$\exists\, c > 0 : \forall\, x \in \mathbb{R}^n : \nabla f(x) \neq 0 \implies -\frac{\nabla f(x)^{\mathrm{T}} d}{\|\nabla f(x)\|_2 \, \|d\|_2} \geq c\,.$$
Hinweis: Normäquivalenzsatz

6. Das Gradientenverfahren

Wir untersuchen hier das Gradientenverfahren bzw. das Verfahren des steilsten Abstiegs zur (unrestringierten) Minimierung einer Funktion $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$. Da die exakte Liniensuche i. Allg. zu aufwendig ist, ersetzen wir diese durch eine einfache Heuristik, der **Schrittweitenbestimmung nach Armijo**. Das Verfahren lautet hiermit

Algorithmus (6.1)

1.) Wähle Startvektor $x^0 \in \mathbb{R}^n$, $\sigma := 0.1$, $\alpha := 1/2$, $k = 0$;

2.) Falls $\|\nabla f(x^k)\| \leq \mathrm{TOL}$: STOP;

3.) Wähle die Abstiegsrichtung: $d^k := -\nabla f(x^k)$;

4.) Bestimme die Schrittweite t_k als größten Wert $t_k = \alpha^\ell$, $\ell = 0, 1, 2, \ldots,$ mit

$$f(x^k + \alpha^\ell\, d^k) \;\leq\; f(x^k) \;+\; \sigma\, \alpha^\ell\, (\nabla f(x^k)^{\mathrm{T}} d^k)$$

5.) $x^{k+1} := x^k + t_k\, d^k$; $\quad k := k + 1$; $\quad$ gehe zu Punkt 2.);

Bemerkungen (6.2)

a) Bei der folgenden Konvergenzuntersuchung kann von beliebigen Parametern $\sigma \in]0, 1[$ und $\alpha \in]0, 1[$ ausgegangen werden.

b) Bei numerischer Realisierung des Algorithmus ist die Genauigkeitsschranke $\mathrm{TOL} > 0$ geeignet vorzugeben. Für die folgende Konvergenzuntersuchung setzen wir jedoch $\mathrm{TOL} = 0$.

c) Die Schrittweitenstrategie nach Armijo erfüllt i. Allg. nicht die Effizienzbedingung, so dass der allgemeine Konvergenzsatz (5.14) nicht ohne zusätzliche Voraussetzungen angewendet werden kann.

Satz (6.3) (Konvergenzsatz 3)

Wird durch den Algorithmus (6.1) für ein $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$, $x^0 \in \mathbb{R}^n$, und $\mathrm{TOL} = 0$ eine nicht abbrechende Folge (x^k) definiert, so gilt für jeden Häufungspunkt x^* dieser Folge $\nabla f(x^*) = 0$.

Beweis: (indirekt)

Wir nehmen an, es gäbe eine Teilfolge (x^{k_j}) von (x^k) mit $x^{k_j} \to x^*$ $(j \to \infty)$ und $\nabla f(x^*) \neq 0$.

Da die Folge $(f(x^k))$ nach Konstruktion streng monoton fällt, folgt aus der Stetigkeit von f, dass $f(x^k) \to f(x^*)$ konvergiert. Somit gilt nach (6.1)

$$\sigma\, t_k\, \|\nabla f(x^k)\|^2 \;=\; -\sigma\, t_k\, \nabla f(x^k)^T d_k \;\leq\; f(x^k) - f(x^{k+1}) \;\to\; 0.$$

Speziell für die Teilfolge $k = k_j$ folgt somit aus $\nabla f(x^*) \neq 0$, dass

$$t_{k_j} \to 0 \quad (j \to \infty).$$

Für hinreichend große j und $k = k_j$ kann daher $t_k < 1$ angenommen werden. Aus der Schrittweitensteuerung nach Armijo ergibt sich nun:

$$f(x^k + \alpha^{\ell_k - 1}\, d^k) \;>\; f(x^k) \;+\; \sigma\, \alpha^{\ell_k - 1}\, \nabla f(x^k)^T d^k,$$

oder umgeformt

$$\frac{f(x^k + \alpha^{\ell_k - 1}\, d^k) \;-\; f(x^k)}{\alpha^{\ell_k - 1}} \;>\; \sigma\, \nabla f(x^k)^T d^k$$

Der Mittelwertsatz liefert nun mit einem Zwischenpunkt $z^k = x^k + \Theta_k\, \alpha^{\ell_k - 1}\, d^k$, $\Theta_k \in]0,1[$: $\nabla f(z^k)^T d^k > \sigma\, \nabla f(x^k)^T d^k$.

Für $j \to \infty$ gilt nun $\alpha^{\ell_k - 1} = t_k/\alpha \to 0$ und $d^k \to -\nabla f(x^*) \neq 0$. Damit folgt auch $z^k \to x^*$ und somit im Grenzwert

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

was der Voraussetzung $0 < \sigma < 1$ widerspricht. □

Quadratische Zielfunktion.

Zur Untersuchung der Konvergenzgeschwindigkeit des Gradientenverfahrens beschränken wir uns auf den Fall einer quadratischen Zielfunktion

$$f(x) \;=\; \frac{1}{2}\, (x - x^*)^{\mathrm{T}} A\, (x - x^*), \tag{6.4}$$

wobei $A \in \mathbb{R}^{(n,n)}$ symmetrisch und positiv definit sei. Ferner wählen wir die exakte Schrittweite.

Bemerkung (6.5)

Eine *beliebige* quadratische Funktion $\tilde{f}(x) = \frac{1}{2}\, x^{\mathrm{T}} A\, x + b^{\mathrm{T}} x + c$ mit symmetrischer und positiv definiter Matrix A, lässt sich (formal) stets in die Form (6.4) transformieren.

Dazu sei x^* Lösung des linearen Gleichungssystems $\nabla\tilde{f}(x) = A\,x + b = 0$. Wir formen nun um

$$\begin{aligned}\tilde{f}(x) &= \frac{1}{2}\,(x-x^*)^{\mathrm{T}}A\,(x-x^*) + (x^*)^{\mathrm{T}}A\,x - \frac{1}{2}\,(x^*)^{\mathrm{T}}A\,x^* \\ &+ (-A\,x^*)^{\mathrm{T}}x + c \\ &= \frac{1}{2}\,(x-x^*)^{\mathrm{T}}A\,(x-x^*) + c - \frac{1}{2}\,(x^*)^{\mathrm{T}}A\,x^* \\ &= f(x) + c - \frac{1}{2}\,(x^*)^{\mathrm{T}}A\,x^*.\end{aligned}$$

Die beiden Zielfunktionen $\tilde{f}$ und f unterscheiden sich also nur durch eine additive Konstante, $\tilde{f}(x) = f(x) + \tilde{f}(x^*)$.

Für die quadratische Zielfunktion f lässt sich die *exakte* Schrittweite explizit angeben (vergleiche Beispiel (5.9)).

Wir haben:

$$\begin{aligned}f(x) &= \frac{1}{2}\,(x-x^*)^{\mathrm{T}}A\,(x-x^*) \\ g(x) &:= \nabla f(x) = A\,(x-x^*)\end{aligned}$$

Damit folgt

$$\begin{aligned}\frac{d}{d\,t}\,\big(f(x-t\,g)\big) &= \big\{A\,(x-t\,g-x^*)\big\}^{\mathrm{T}}(-g) \\ &= -\,(x-x^*)^{\mathrm{T}}A\,g + t\,(g^{\mathrm{T}}A\,g) \\ &= -\,g^{\mathrm{T}}g + t\,(g^{\mathrm{T}}A\,g)\end{aligned}$$

und für die exakte Schrittweite ergibt sich

$$t_k := \frac{g_k^{\mathrm{T}}g_k}{g_k^{\mathrm{T}}A\,g_k}, \quad g_k := \nabla f(x^k) \tag{6.6}$$

Im Gradientenverfahren (6.1) wird also die Schrittweitenbestimmung ersetzt durch die exakte Liniensuche (6.6).

Setzt man nun x^{k+1} in f ein, so erhält man nach etwas Rechnung die folgende Relation für den Abstieg der Zielfunktion:

$$f(x^{k+1}) = \Big(1 - \frac{(g_k^{\mathrm{T}}g_k)^2}{(g_k^{\mathrm{T}}A\,g_k)\,(g_k^{\mathrm{T}}A^{-1}g_k)}\Big)\,f(x^k). \tag{6.7}$$

Zur Abschätzung des Verkleinerungsfaktors von f aus der obigen Relation verwenden wir nun die so genannte Kantorowitsch–Ungleichung.

Satz (6.8) (Kantorowitsch–Ungleichung)

Sei $A \in \mathbb{R}^{(n,n)}$ eine symmetrische und positiv definite Matrix mit Eigenwerten $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Dann gilt für alle $x \in \mathbb{R}^n$:

$$\frac{(x^{\mathrm{T}}x)^2}{(x^{\mathrm{T}}A\,x)\,(x^{\mathrm{T}}A^{-1}x)} \;\geq\; 4\;\frac{\lambda_1\,\lambda_n}{(\lambda_1\,+\,\lambda_n)^2}\,, \qquad \text{für } x \neq 0.$$

Fasst man nun (6.7) und (6.8) zusammen, so ergibt sich die folgende Aussage über die Konvergenzgeschwindigkeit des Gradientenverfahrens

Satz (6.9) (Konvergenzgeschwindigkeit)

Wendet man auf eine quadratische Zielfunktion

$$f(x) \;=\; \frac{1}{2}\,x^{\mathrm{T}}A\,x \;+\; b^{\mathrm{T}}x \;+\; c$$

mit symmetrischer und positiv definiter Matrix $A \in \mathbb{R}^{(n,n)}$ das Gradientenverfahren (6.1) mit exakter Schrittweitenbestimmung (6.6) an, so gilt für alle $k \in \mathbb{N}$:

$$(f(x^{k+1}) \;-\; f(x^*)) \;\leq\; \Big(\frac{\lambda_{\max} \;-\; \lambda_{\min}}{\lambda_{\max} \;+\; \lambda_{\min}}\Big)^2\,(f(x^k) \;-\; f(x^*)).$$

Dabei bezeichnet $\lambda_{\max}$ bzw. $\lambda_{\min}$ den größten bzw. kleinsten Eigenwert der Hesse–Matrix $A = \nabla^2 f(x^*)$.

Beweis:

Nach (6.5) gilt $f(x) = \frac{1}{2}\,(x - x^*)^T A\,(x - x^*) \;+\; f(x^*)$, also folgt mit (6.7) auch

$$\begin{aligned}
(f(x^{k+1}) \;-\; f(x^*)) \;&=\; \Big(1 \;-\; \frac{(g_k^T g_k)^2}{(g_k^T A\,g_k)\,(g_k^T A^{-1} g_k)}\Big)\;(f(x^k) \;-\; f(x^*)) \\
&\leq\; \Big(1 \;-\; 4\;\frac{\lambda_1\,\lambda_n}{(\lambda_1\,+\,\lambda_n)^2}\Big)\;(f(x^k) \;-\; f(x^*)) \\
&=\; \Big(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\Big)^2\;(f(x^k) \;-\; f(x^*)). \qquad \square
\end{aligned}$$

Bemerkungen (6.10)

a) Die Abschätzung in (6.9) wird für kritische Beispiele mit Gleichheit erfüllt.

b) Satz (6.9) gilt sinngemäß auch für skalierte negative Gradientenrichtungen der Form $d = -H^{-1}\nabla f(x)$, H symmetrisch und positiv definit. Dabei

bezeichnet dann $\lambda_{\max}$ bzw. $\lambda_{\min}$ den größten bzw. kleinsten Eigenwert von $H^{-1}A$.
Man beachte, dass $H^{-1}A$ ähnlich ist zur Matrix $H^{-1/2}\,A\,H^{-1/2}$ und dass diese Matrix symmetrisch und positiv definit ist.

c) $\kappa := \lambda_{\max}/\lambda_{\min}$ ist die spektrale Konditionszahl der Matrix A. Nach (6.9) gilt

$$(f(x^{k+1}) - f(x^*)) \leq \Big(\frac{\kappa - 1}{\kappa + 1}\Big)^2 (f(x^k) - f(x^*)),$$

d.h. die Konvergenz ist umso langsamer, je größer die Kondition κ ist. Für $\kappa = 100$ ergibt sich z.B. der Verkleinerungsfaktor ≈ 0.96.

d) Die Aussage des Satzes (6.9) lässt sich lokal auch auf den Fall nichtquadratischer Funktionen $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$ übertragen (vgl. z.B. Luenberger, 1973). Dabei ist $\lambda_{\max}$ bzw. $\lambda_{\min}$ der größte bzw. kleinste Eigenwert der (als positiv definit vorausgesetzten) Hesse–Matrix $\nabla^2 f(x^*)$ im Lösungspunkt.

Beispiel (6.11) (Luenberger[1])

Wir betrachten die quadratische Zielfunktion $f(x) = 0.5x^{\mathrm{T}}Ax - b^{\mathrm{T}}x$ mit

$$A = \begin{pmatrix} 0.78 & -0.02 & -0.12 & -0.14 \\ -0.02 & 0.86 & -0.04 & 0.06 \\ -0.12 & -0.04 & 0.72 & -0.08 \\ -0.14 & 0.06 & -0.08 & 0.74 \end{pmatrix}, \quad b = \begin{pmatrix} 0.76 \\ 0.08 \\ 1.12 \\ 0.68 \end{pmatrix}.$$

Die Matrix A ist symmetrisch und diagonaldominant und damit auch positiv definit. Beispielsweise mit der MATLAB–Routine eig findet man $\lambda_{\min}(A) \approx 0.52$ und $\lambda_{\max}(A) \approx 0.94$. Damit wird $\kappa \approx 1.8$ und

$$\Big(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\Big)^2 \approx 0.083.$$

Mit den Startvektor $x^0 := 0$ wird das Abbruchkriterium $\|\nabla f(x^k)\| \leq 10^{-5}$ bereits nach 8 Iterationen erfüllt. Für die numerische Lösung findet man

$$x^8 \approx (1.53496, 0.12201, 1.97515, 1.41295)^{\mathrm{T}}.$$

[1] D.G. Luenberger: Linear and nonlinear programming, Springer, 2008.

Beweis der Kantorowitsch–Ungleichung:

Zu zeigen ist: $\forall\, x \in \mathbb{R}^n \setminus \{0\}: \ \dfrac{(x^{\mathrm{T}}x)^2}{(x^{\mathrm{T}}Ax)(x^{\mathrm{T}}A^{-1}x)} \ \geq \ \dfrac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}.$

Dabei seien $0 < \lambda_1 \leq \ldots \leq \lambda_n$ die Eigenwerte der symmetrischen und positiv definiten Matrix A.

Es sei (u_j) eine Orthonormalbasis aus zugehörigen Eigenvektoren von A. Mit der Darstellung $x = \sum_{i=1}^n \xi_i u_i,\ \xi_i \in \mathbb{R},\ \sum_{i=1}^n \xi_i^2 > 0$ folgt dann

$$\begin{aligned} F(x) \ &:= \ \frac{(x^{\mathrm{T}}x)^2}{(x^{\mathrm{T}}Ax)(x^{\mathrm{T}}A^{-1}x)} \ = \ \frac{(\sum \xi_j^2)^2}{(\sum \lambda_i\xi_i^2)\,(\sum \dfrac{1}{\lambda_i}\xi_i^2)} \\ &= \ \frac{1}{(\sum \lambda_i \dfrac{\xi_i^2}{\sum \xi_j^2})\,(\sum \dfrac{1}{\lambda_i}\dfrac{\xi_i^2}{\sum \xi_j^2})} \\ &= \ \frac{(\sum \lambda_i\,\gamma_i)^{-1}}{(\sum \lambda_i^{-1}\,\gamma_i)}, \quad \text{mit} \quad \gamma_i := \xi_i^2 / \sum \xi_j^2. \end{aligned}$$

Die γ_i erfüllen die Voraussetzung der Koeffizienten einer Konvexkombination $\gamma_i \geq 0,\ \sum \gamma_i = 1$.

Betrachtet man die folgenden Punkte in der Ebene

$$P_i \ := \ (\lambda_i, \frac{1}{\lambda_i}), \quad Q \ := \ (\sum \gamma_i\lambda_i, \frac{1}{\sum \gamma_i\lambda_i}), \quad R \ := \ (\sum \gamma_i\lambda_i, \sum \gamma_i\frac{1}{\lambda_i}),$$

so stellt man fest:

Die Punkte $P_1, \ldots, P_n$ und Q liegen auf dem Hyperbelast $y = 1/x,\ x > 0$.

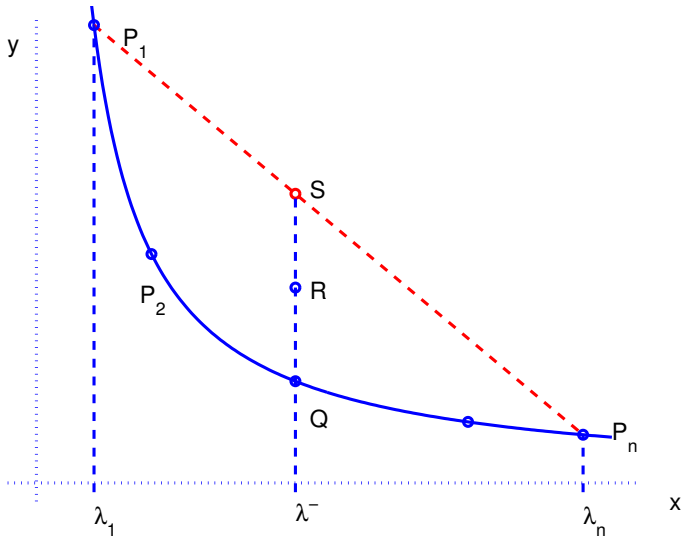
Die x–Koordinate $\bar{\lambda} := \sum \gamma_i\lambda_i$ von Q ist eine Konvexkombination der $\lambda_1, \ldots, \lambda_n$; daher gilt $\lambda_1 \leq \bar{\lambda} \leq \lambda_n$.

R hat die gleiche x–Koordinate wie Q und ist eine Konvexkombination von $P_1, \ldots, P_n$. Deshalb muss R in der konvexen Hülle von $P_1, \ldots, P_n$ liegen, also oberhalb von Q und unterhalb der Sekante zwischen P_1 und P_n.

Die Sekantengleichung (Zweipunkteform der Geradengleichung) lautet

$$\frac{y - 1/\lambda_1}{x - \lambda_1} \ = \ \frac{1/\lambda_n - 1/\lambda_1}{\lambda_n - \lambda_1} \quad \text{oder} \quad y \ = \ \frac{\lambda_1 + \lambda_n - x}{\lambda_1\,\lambda_n}.$$

Damit folgt $\quad \sum \gamma_i \dfrac{1}{\lambda_i} \ \leq \ \dfrac{\lambda_1 + \lambda_n - \bar{\lambda}}{\lambda_1\,\lambda_n} \quad$ und somit



$$F(x) \;=\; \frac{\bar\lambda^{-1}}{\sum \gamma_i \lambda_i^{-1}} \;\geq\; \frac{\lambda_1\,\lambda_n}{\bar\lambda\,(\lambda_1+\lambda_n-\bar\lambda)} \;\geq\; \min\{\frac{\lambda_1\,\lambda_n}{\mu\,(\lambda_1+\lambda_n-\mu)} :\; \lambda_1 \leq \mu \leq \lambda_n\}.$$

Das Minimum wird in $\mu = (\lambda_1+\lambda_n)/2$ angenommen (der Nenner ist eine nach unten geöffnete Parabel) und somit folgt

$$F(x) \;\geq\; \frac{4\,\lambda_1\,\lambda_n}{(\lambda_1+\lambda_n)^2}\,. \qquad \square$$

Aufgabe (6.12)

Für die Funktion $f(x_1,x_2) = x_1^2 + 100\,x_2^2$ und Startvektor $x^0 = (1, 0.01)^T$ soll das Verfahren des steilsten Abstiegs mit optimaler Schrittweitenstrategie durchzuführen werden.

Geben Sie eine explizite Darstellung der Iterierten x^k an und zeigen Sie, dass die Abschätzung der Konvergenzgeschwindigkeit aus Satz (6.9) mit Gleichheit erfüllt ist.

7. Schrittweitenstrategien

Zur Bestimmung geeigneter Schrittweiten im k-ten Iterationsschritt eines Abstiegsverfahrens gehen wir von folgender Situation aus: $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$ sei die zu minimierende Zielfunktion, $x \in \mathbb{R}^n$ die aktuelle Näherung und $d \in \mathbb{R}^n$ eine Abstiegsrichtung, die die hinreichende Bedingung $\nabla f(x)^{\mathrm{T}} d < 0$ erfüllt.

Ziel ist es, eine Schrittweite $t > 0$ zu bestimmen, so dass die Hilfsfunktion

$$\Phi(t) := f(x + t\,d), \quad t \geq 0 \tag{7.1}$$

näherungsweise minimiert wird, bzw. im ausreichenden Maße verkleinert wird.

Wegen $\Phi'(t) = \nabla f(x + t\,d)^{\mathrm{T}} d$ gilt $\Phi'(0) < 0$. Daher fällt Φ für kleine t-Werte:

$$\exists\, t_0 > 0 : \ \forall\, t \in]0, t_0[: \quad \Phi(t) < \Phi(0). \tag{7.2}$$

Gehen wir ferner von einer *kompakten* Levelmenge $L_f(x^0)$ aus, so wissen wir auch, dass $\Phi(t)$ für hinreichend große t wieder wächst.

Die Armijo–Schrittweite.

Zu Parametern $\sigma \in]\,0, 1[$, $\alpha \in]\,0, 1[$ wähle man die Schrittweite $t := \alpha^\ell$ mit

$$\ell := \min\{j \in \mathbb{N}_0 : \ f(x + \alpha^j d) \leq f(x) + \sigma\alpha^j (\nabla f(x)^{\mathrm{T}} d)\}. \tag{7.3}$$

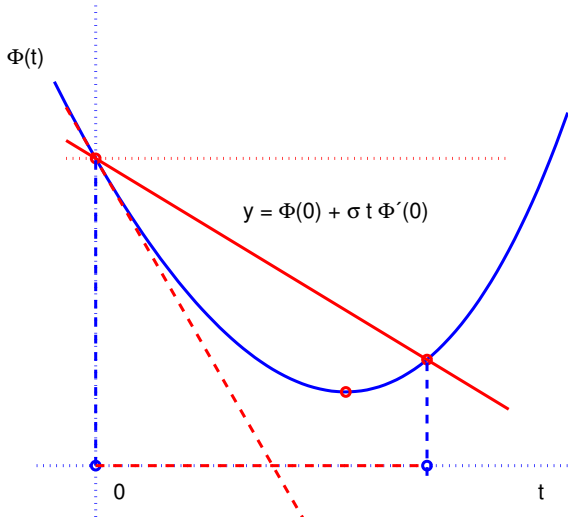


Abb.7.1 Armijo–Schrittweite

Ist $\Psi(t) := \Phi(t) - (f(x) + \sigma\, t\, (\nabla f(x)^{\mathrm{T}} d))$, so folgt nach Obigem $\Psi(0) = 0$ und $\Psi'(0) = (1 - \sigma)\, (\nabla f(x)^{\mathrm{T}} d) < 0$. Damit ist klar, dass es ein eindeutig bestimmtes

minimales $\ell \in \mathbb{N}_0$ und damit auch ein (maximales) $t = \alpha^\ell > 0$ gibt, welches die Armijo-Bedingung (7.3), $\Psi(t) \le 0$, erfüllt. Das gesuchte $t = \alpha^\ell$ lässt sich durch einfaches Absuchen $j = 0, 1, 2, \ldots$ bestimmen.

Anmerkungen (7.4)

a) Die Armijo–Schrittweite ist i. Allg. nicht effizient. Allerdings erfüllt sie für $0 < \sigma < 0.5$ die (schwächere) Semi-Effizienzbedingung

$$f(x + t\,d) \;\le\; f(x) \;-\; \Theta \,\min\left[-\nabla f(x)^{\mathrm{T}} d, \left(\frac{\nabla f(x)^{\mathrm{T}} d}{\|d\|}\right)^2\right]$$

b) Die *skalierte Armijo-Schrittweite* arbeitet mit einem Skalierungsfaktor $s > 0$ und verlangt anstelle von (7.3)

$$\ell \;:=\; \min\{j \in \mathbb{N}_0 :\; f(x + s\,\alpha^j d) \;\le\; f(x) + \sigma\, s\, \alpha^j (\nabla f(x)^{\mathrm{T}} d)\}.$$

Bei kompakter Levelmenge $L_f(x^0)$ und $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$ ist die skalierte Armijo–Schrittweite für hinreichend große Skalierungsparameter s effizient.

c) Die *Armijo–Schrittweite mit Aufweitung* sucht dagegen in $\mathbb{Z}$

$$\ell \;:=\; \min\{j \in \mathbb{Z} :\; f(x + \alpha^j d) \;\le\; f(x) + \sigma\, \alpha^j (\nabla f(x)^{\mathrm{T}} d)\}.$$

Damit sind auch Schrittweiten $t > 1$ möglich. Bei kompakter Levelmenge $L_f(x^0)$ und $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$ ist die Armijo–Schrittweite mit Aufweitung effizient.

Die Goldstein–Schrittweite.

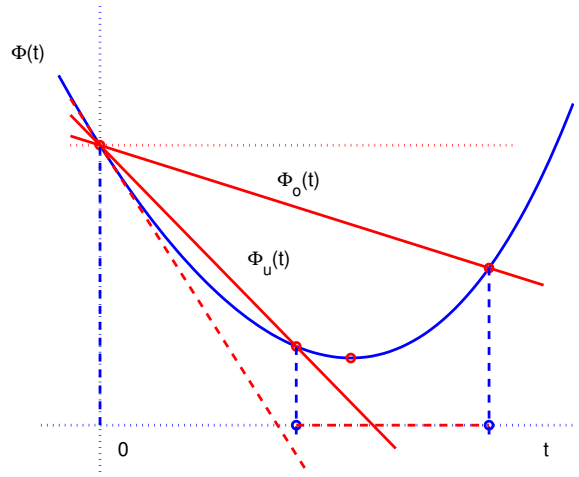


Abb.7.2 Goldstein–Schrittweite

Um kleine Schrittweiten zu vermeiden wird der zulässige Bereich auch nach unten eingeschränkt. Wir sagen, $t > 0$ erfüllt die *Goldstein–Bedingung*, wenn zu einem festen Parameter $\sigma \in]0, 0.5[$ gilt

$$\Phi_u(t) \;\leq\; \Phi(t) \;\leq\; \Phi_o(t), \tag{7.5}$$

wobei die Geraden Φ_u und Φ_o wie folgt erklärt sind

$$\Phi_o(t) \;:=\; \Phi(0) \;+\; \sigma\, t\, \Phi'(0), \quad \Phi_u(t) \;:=\; \Phi(0) \;+\; (1-\sigma)\, t\, \Phi'(0). \tag{7.6}$$

Satz (7.7) (Effizienz)

Ist $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, $x \in \mathbb{R}^n$ mit $L_f(x)$ kompakt, d Abstiegsrichtung mit $\nabla f(x)^{\mathrm{T}} d < 0$ und $C := \max\{\| \nabla^2 f(y) \| : y \in L_f(x)\}$, so gilt für jede Schrittweite t, die die Goldstein–Bedingung erfüllt,

$$f(x + t\, d) \;\leq\; f(x) \;-\; \frac{\sigma}{C} \left(\frac{\nabla f(x)^{\mathrm{T}} d}{\|d\|} \right)^2 .$$

Beweis:

Sei $t^* > 0$ das Infimum der positiven lokalen Minima von Φ. Ein solches existiert aufgrund der Voraussetzungen. t^* ist dann ein stationärer Punkt (als Häufungspunkt stationärer Punkte) und Φ ist streng monoton fallend in $[0, t^*]$ (Satz von Rolle).

Fall 1: $t \leq t^*$.

Mit $\Phi_u(t) = \Phi(0) + (1-\sigma)\, t\, \Phi'(0) \leq \Phi(t)$ folgt mittels des Taylorschen Satzes

$$(1-\sigma)\, t\, \Phi'(0) \;\leq\; \Phi(t) \;-\; \Phi(0) \;=\; t\, \Phi'(0) \;+\; \frac{t^2}{2} \Phi''(\tilde{t}) \;\leq\; t\, \Phi'(0) \;+\; \frac{t^2}{2}\, C\, \|d\|^2$$

und somit $-\sigma\, t\, \Phi'(0) \;\leq\; \dfrac{t^2}{2}\, C\, \|d\|^2$, bzw. $t \;\geq\; -\dfrac{2\,\sigma}{C}\, \dfrac{\Phi'(0)}{\|d\|^2} \;=:\; \hat{t} \;>\; 0.$

Wegen der Monotonie von Φ auf $0 < \hat{t} \leq t \leq t^*$ folgt weiter

$$\begin{aligned} \Phi(t) \;&\leq\; \Phi(\hat{t}) \;\leq\; \Phi(0) \;+\; \hat{t}\, \Phi'(0) \;+\; \frac{\hat{t}^2}{2}\, C\, \|d\|^2 \\ &=\; \Phi(0) \;-\; \frac{2\,\sigma\,(1-\sigma)}{C} \left(\frac{\Phi'(0)}{\|d\|} \right)^2 \\ &\leq\; \Phi(0) \;-\; \frac{\sigma}{C} \left(\frac{\Phi'(0)}{\|d\|} \right)^2 . \end{aligned}$$

Die Gleichheit in der zweiten Zeile folgt durch Einsetzen von $\hat{t}$, die letzte Ungleichung mit $0 < \sigma < 0.5$.

Fall 2: $t > t^*$.

Wieder folgt zunächst mit dem Taylorschen Satz

$$0 \;=\; \Phi'(t^*) \;=\; \Phi'(0) \;+\; t^*\Phi''(\tilde{t}) \;\leq\; \Phi'(0) \;+\; t^*\, C\, \|d\|^2,$$

also $t^* \;\geq\; -\,\dfrac{\Phi'(0)}{C\,\|d\|^2}$. Hieraus ergibt sich weiter

$$\begin{aligned} \Phi(t) \;&\leq\; \Phi_o(t) \;=\; \Phi(0) \;+\; \sigma\, t\, \Phi'(0) \\ &\leq\; \Phi(0) \;+\; \sigma\, t^*\, \Phi'(0) \;\leq\; \Phi(0) \;-\; \frac{\sigma}{C}\left(\frac{\Phi'(0)}{\|d\|}\right)^2. \quad \square \end{aligned}$$

Algorithmus (7.8)

1.) Wähle Parameter $t_u := 0,\;\; t_o > 0$;

2.) Solange $\Phi(t_o) < \Phi_u(t_o)$: $\;t_u := t_o,\;\; t_o := 2\, t_o$;

3.) Falls $\Phi(t_o) \leq \Phi_o(t_o)$: $\;t := t_o$; Stop.

4.) Wiederhole $t := (t_u + t_o)/2$,

Falls $\Phi(t) < \Phi_u(t)$: $\;t_u := t$,

Falls $\Phi(t) > \Phi_o(t)$: $\;t_o := t$,

bis gilt: $\;\Phi_u(t) \leq \Phi(t) \leq \Phi_o(t)$; Stop.

Bemerkungen (7.9)

a) Der obige Algorithmus bricht nach endlich vielen Schritten mit einer zulässigen Goldstein–Schrittweite ab.

b) Alternativ zu der Halbierungsstrategie in (7.8) lassen sich Interpolationstechniken zu einer effizienteren Wahl von t verwenden.

c) Bei der iterierten Anwendung des Algorithmus in einem Abstiegsverfahren könnte man als Startwert für t_o die letzte erfolgreiche Schrittweite wählen.

d) Im Zusammenhang mit Newton– und Quasi–Newton–Verfahren sollte man stets zunächst die Schrittweite $t = 1$ testen, da nur diese die quadratische bzw. superlineare Konvergenz des Verfahrens sicherstellt.

Die Wolfe–Powell Schrittweitenregel.

Je nach Wahl des Parameters σ kann die untere Goldstein–Bedingung das Minimum von Φ ausschließen, d.h. die exakte Schrittweite wäre demnach nicht zulässig. Um diesen Nachteil zu vermeiden, wird bei der Wolfe–Powell Bedingung die untere Ungleichung ersetzt durch eine Bedingung an die Steigung von Φ.

Wir sagen, eine Schrittweite $t > 0$ erfüllt die *Wolfe–Powell–Bedingung*, wenn zu festen Parametern $\sigma \in]0, 0.5[$ und $\sigma < \rho < 1$ gilt

$$\Phi'(t) \;\geq\; \rho\,\Phi'(0), \quad \Phi(t) \;\leq\; \Phi_o(t), \tag{7.10}$$

wobei die Gerade Φ_o erklärt ist wie in (7.6): $\Phi_o(t) := \Phi(0) + \sigma\, t\, \Phi'(0)$.

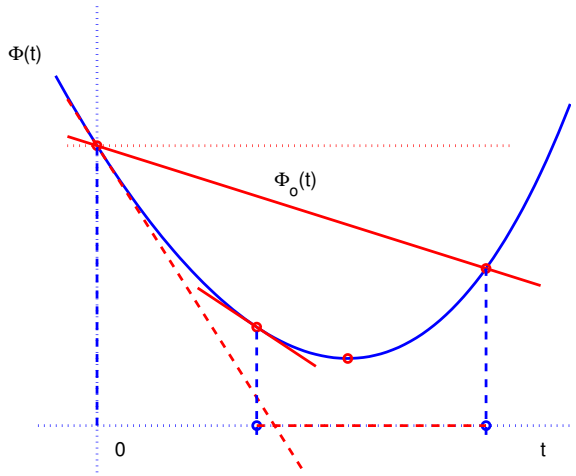


Abb.7.3 Wolfe–Powell Schrittweite

Satz (7.11) (Effizienz)

Ist $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, $x \in \mathbb{R}^n$ mit $L_f(x)$ kompakt, d Abstiegsrichtung mit $\nabla f(x)^{\mathrm{T}} d < 0$ und $C := \max\{\| \nabla^2 f(y) \| : y \in L_f(x)\}$, so gilt für jede Schrittweite t, die die Wolfe-Powell–Bedingungen (7.10) erfüllt,

$$f(x + t\,d) \;\leq\; f(x) \;-\; \frac{\sigma\,(1-\rho)}{C} \left(\frac{\nabla f(x)^{\mathrm{T}} d}{\|d\|} \right)^2 .$$

Beweis:

Zur Existenz: Nach Voraussetzung existiert ein $t_0 > 0$ mit $\Phi(t_0) = \Phi_o(t_0)$ und $\forall\, 0 < t < t_0 :\ \Phi(t) < \Phi_o(t)$.

Nach dem Mittelwertsatz existiert damit ein $t \in]0, t_0[$ mit

$$\Phi'(t) \;=\; \frac{\Phi(t_0) - \Phi(0)}{t_0 - 0} \;=\; (\Phi_o(t_0) - \Phi_o(0))/t_0 \;=\; \sigma\,\Phi'(0) \;>\; \rho\,\Phi'(0).$$

Dieses t erfüllt also die Wolfe-Powell Bedingungen (7.10).

Zur Effizienz: Es bezeichne t^* wieder das Infimum der positiven lokalen Minima von Φ. Wie im Beweis von Satz (7.7), Fall 2, folgt zunächst für alle zulässigen Schrittweiten $t > t^*$:

$$\Phi(t) \;<\; \Phi(0) \;-\; \frac{\sigma}{C}\left(\frac{\Phi'(0)}{\|d\|}\right)^2 \;\leq\; \Phi(0) \;-\; \frac{\sigma\,(1-\rho)}{C}\left(\frac{\Phi'(0)}{\|d\|}\right)^2.$$

Für zulässige Schrittweiten $0 < t \leq t^*$ schließt man wie folgt:

$$\begin{aligned} & \Phi'(t) \;\geq\; \rho\,\Phi'(0) \;=\; \Phi'(0) - (1-\rho)\,\Phi'(0) \\ \Rightarrow \qquad & (1-\rho)\,\Phi'(0) \;\geq\; \Phi'(0) - \Phi'(t) \;=\; -t\,\Phi''(\tilde{t}) \;\geq\; -t\,C\,\|d\|^2. \end{aligned}$$

Hiermit ist

$$t \;\geq\; -\,\frac{(1-\rho)\,\Phi'(0)}{C\,\|d\|^2}$$

und damit

$$\begin{aligned} \Phi(t) \;&\leq\; \Phi_o(t) \;=\; \Phi(0) \;+\; \sigma\,t\,\Phi'(0) \\ &\leq\; \Phi(0) \;-\; \frac{\sigma\,(1-\rho)}{C}\left(\frac{\Phi'(0)}{\|d\|}\right)^2. \qquad \square \end{aligned}$$

Im Folgenden beschreiben wir einen Algorithmus zur Realisierung einer Wolfe-Powell Schrittweite. Analog zur Algorithums (7.8) tritt eine Expansionsphase gefolgt von einer Kontraktionsphase auf.

Algorithmus (7.12)

1.) Wähle Parameter $\alpha_1,\ \alpha_2 \in\ ]0, 0.5]$, $\gamma > 1$ sowie eine Startschrittweite $t_0 > 0$; $t := t_0$.

2.) Falls $\Phi(t) > \Phi_o(t)$: Gehe nach 3.);
Falls $\Phi(t) \leq \Phi_o(t)$ und $\Phi'(t) < \rho\,\Phi'(0)$: $t := \gamma\,t$, Gehe nach 2.);
Falls $\Phi(t) \leq \Phi_o(t)$ und $\Phi'(t) \geq \rho\,\Phi'(0)$: Stop.

3.) $t_u := 0$; $t_o := t$;

4.) Wähle $t \in [t_u + \alpha_1\,(t_o - t_u), t_o - \alpha_2\,(t_o - t_u)]$;

5.) Falls $\Phi(t) > \Phi_o(t)$: $t_o := t$, Gehe zu 4.);
Falls $\Phi(t) \leq \Phi_o(t)$ und $\Phi'(t) < \rho\,\Phi'(0)$: $t_u := t$, Gehe zu 4.);
Falls $\Phi(t) \leq \Phi_o(t)$ und $\Phi'(t) \geq \rho\,\Phi'(0)$: Stop.

Erläuterung:

In Schritt 2.) findet eine *Expansionsphase* statt. Das Intervall $[0, t]$ wird solange vergrößert, bis $\Phi(t) \geq \Phi_o(t)$ gilt, bzw. bis eine zulästige Schrittweite gefunden wurde.

Nach dem obigen Beweis zu (7.11) ist dann klar, dass es zulässige Schrittweiten t in diesem Intervall gibt.

In den *Kontraktionsphasen* 4.) und 5.) wird das Intervall $[t_u, t_o]$ schrittweise verkleinert, wobei stets $t_u < t_o$ und

$$\Phi(t_u) \leq \Phi_o(t_u), \quad \Phi'(t_u) < \rho\,\Phi'(0) \quad \text{und} \quad \Phi(t_o) > \Phi_o(t_o)$$

gelten müssen. In jedem dieser Intervalle gibt es demnach zulässige Schrittweiten (Beweis analog zu (7.11)). Da die Intervalllänge aber gegen Null konvergieren würde, bricht die Schleife notwendigerweise nach *endlich* vielen Kontraktionsschritten mit einer zulässigen Schrittweite t ab.

Bemerkungen (7.13)

a) Zur Festlegung der Schrittweite t in Schritt 4.) lassen sich *Interpolationstechniken* verwenden. Dazu kann man beispielsweise das Hermite-Interpolationspolynom $p \in \Pi_3$ zu den Interpolationsdaten $(t_u, \Phi(t_u), \Phi'(t_u))$ und $(t_o, \Phi(t_o), \Phi'(t_o))$ berechnen. Besitzt p nun in einem inneren Punkt des Intervalls $[t_u + \alpha_1\,(t_o - t_u), t_o - \alpha_2\,(t_o - t_u)]$ ein (lokales) Minimum, so wähle man dieses als neue Versuchs-Schrittweite t; andernfalls nehme man den Mittelpunkt dieses Intervalls.

b) Der Algorithmus (7.12) zur Berechnung einer Wolfe-Powell Schrittweite benötigt i. Allg. weniger Iterationen als der Algorithmus (7.8) zur Bestimmung einer Goldstein Schrittweite, allerdings wird in jeden Teilschritt von (7.12) eine Gradientenauswertung $\Phi'(t) = \nabla f(x + td)^{\mathrm{T}} d$ benötigt, während in Algorithmus (7.8) lediglich die einmalige Auswertung von $\Phi'(0)$ notwendig ist.

Die strenge Wolfe–Powell Schrittweitenregel.

Mit der Wolfe-Powell Bedingung (7.10) werden Schrittweiten mit zu kleiner (negativen) Steigung der Hilfsfunktion $\Phi(t) := f(x+td)$ ausgeschlossen. Es ist jedoch auch sinnvoll, Schrittweiten mit einer großen (positiven) Steigung von Φ auszuschliessen. Nach Fletcher wird demnach die Wolfe-Powell Bedingung (7.10) ersetzt durch die *strenge Wolfe-Powell-Bedingung*

$$|\Phi'(t)| \;\leq\; -\rho\,\Phi'(0), \quad \Phi(t) \;\leq\; \Phi_o(t), \tag{7.14}$$

wobei wie zuvor $\Phi_o(t) = \Phi(0) + \sigma\,t\,\Phi'(0)$, $0 < \sigma < 0.5$ und $\sigma < \rho < 1$.

Es lässt sich zeigen, dass zu einer C^1-Funktion f, $x \in \mathbb{R}^n$ mit beschränktere Levelmenge $L_f(x)$ und einer Abstiegsrichtung d mit $\nabla f(x)^T d < 0$ stets Schrittweiten $t > 0$ existieren, die den strengen Wolfe-Powell Bedingungen genügen. Ferner gilt analog die Effizienzaussage von Satz (7.11) auch für diese Schrittweiten.

Zur numerischen Realisierung lässt sich eine Variante des obigen Algorithmus (7.12) verwenden.

Aufgabe (7.15)

Zur Minimierung einer Funktion $f \in C^1(\mathbb{R}^n, \mathbb{R})$ sei $x \in \mathbb{R}^n$ gegeben mit kompakter unteren Levelmenge $L_f(x)$ und eine Abstiegsrichtung $d \in \mathbb{R}^n$ mit $\nabla f(x)^T d < 0$.

Zeigen Sie, dass es Schrittweiten $t > 0$ gibt, die den strengen Wolfe–Powell–Bedingungen genügen.

Genauer ist zu zeigen, dass mit $\Phi(t) := f(x + td)$, $\Phi_o(t) := \Phi(0) + t\,\sigma\,\Phi'(0)$, $t \geq 0$, $0 < \sigma < 1/2$ und $\sigma < \rho < 1$ gilt:

Sind $0 \leq t_u < t_o$ Schrittweiten mit $\Phi(t_u) \leq \Phi_o(t_u)$, $\Phi'(t_u) < \rho\,\Phi'(0)$ und $\Phi(t_o) > \Phi_o(t_o)$ *oder* $\Phi'(t_o) > -\rho\,\Phi'(0)$, so gibt es eine Schrittweite $t \in]t_u, t_o[$ mit

$$\Phi(t) \leq \Phi_o(t) \text{ und } |\Phi'(t)| \leq -\rho\,\Phi'(0).$$

Aufgabe (7.16)

Geben Sie eine Modifikation des Algorithmus (7.12) an, mit der eine Schrittweite t berechnet werden kann, die den strengen Wolfe–Powell–Bedingungen genügt.

Zeigen Sie insbesondere, dass der Algorithmus – unter den Voraussetzungen von Aufgabe (7.15) – in endlich vielen Schritten mit einer zulässigen Schrittweite abbricht.

Hinweis: Zur Konstruktion des Algorithmus orientiere man sich an Aufgabe (7.15).

8. Das Newton–Verfahren

Ein generelles Prinzip zur Bestimmung einer Abstiegsrichtung ist *das Prinzip der lokalen Approximation*. Man ersetzt hierzu die Zielfunktion f in der Nähe einer aktuellen Näherung $x = x^k$ durch eine Approximation $f(z) \to \tilde{f}(z)$ und minimiert anstelle von f nun $\tilde{f}$. (Wobei vorausgesetzt wird, dass auch das Minimum von $\tilde{f}$ in der Nähe von x liegt.

Beispiel (8.1) (Lineares Modell)

Wir setzen nach Taylor $\tilde{f}(z) := f(x) + \nabla f(x)^{\mathrm{T}}(z - x)$. Da eine lineare Funktion aber i.Allg. kein Minimum auf $\mathbb{R}^n$ besitzt, macht ein solches Modell nur Sinn, wenn man es auf einen Bereich um x, z.B. auf $\|z - x\|_2 \leq 1$ einschränkt.

Nach Früherem ist das Ergebnis der Minimimierung von $\tilde{f}$ über diesem Bereich aber gerade $d_G = z^* - x = -\nabla f(x)/\|\nabla f(x)\|$. D.h. die hierdurch definierte Abstiegsrichtung ergibt gerade das *Gradientenverfahren* für das Ausgangsproblem.

Beispiel (8.2) (Quadratisches Modell)

Wir führen die Taylor–Entwicklung einen Term weiter und setzen

$$\tilde{f}(z) \;:=\; f(x) \;+\; \nabla f(x)^{\mathrm{T}}(z - x) \;+\; \frac{1}{2}\,(z - x)^{\mathrm{T}} \nabla^2 f(x)\,(z - x).$$

Nach (3.7) wissen wir, dass eine quadratische Funktion ein eindeutig bestimmtes globales Minimum besitzt, wenn die (symmetrische) Koeffizientenmatrix der quadratischen Terme positiv definit ist. Falls die Hesse-Matrix $\nabla^2 f(x)$ also positiv definit ist, so gibt es genau ein globales Minimum z^* von $\tilde{f}$ und dieses ist gegeben durch das lineare Gleichungssystem

$$\nabla \tilde{f}(z^*) \;=\; \nabla f(x) \;+\; \nabla^2 f(x)\,(z^* - x) \;=\; 0 \;\in\; \mathbb{R}^n.$$

Mit $d_N := z^* - x$ lautet dieses lineare Gleichungssystem also

$$\nabla^2 f(x)\, d_N \;=\; -\;\nabla f(x), \tag{8.3}$$

d.h. d_N ist die *Newton-Richtung* und das Verfahren $x^{k+1} := x^k + d_N$ ist das *Newton-Verfahren* zur Bestimmung einer Nullstelle des Gradienten $\nabla f(x)$. (8.3) heißt auch die zugehörige *Newton-Gleichung*.

Ist d_N eine Abstiegsrichtung? Einsetzen von (8.3) ergibt

$$\nabla f(x)^{\mathrm{T}} d_N \;=\; -g(x)^{\mathrm{T}}\,[\nabla^2 f(x)]^{-1}\, g(x), \qquad g(x) \;:=\; \nabla f(x).$$

Ist die Hesse-Matrix $\nabla^2 f(x)$ also positiv definit, so gilt dies auch für die inverse Hesse-Matrix und wir erhalten $\nabla f(x)^{\mathrm{T}} d_N < 0$, falls $\nabla f(x) \neq 0$, d_N ist dann also eine Abstiegsrichtung. Wir fassen nochmals zusammen:

Satz (8.4)

Für $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$ führt das Prinzip der lokalen Approximation mit quadratischem Modell auf die Newton-Richtung d_N aus (8.3).
Ist $\nabla^2 f(x)$ positiv definit, so ist d_N eine Abstiegsrichtung.

Unter den obigen Voraussetzungen ($f \in \mathrm{C}^2$, $\nabla^2 f$ positiv definit) lässt sich das Newton-Verfahren also als ein Abstiegsverfahren mit Schrittweite $t = 1$ interpretieren.

Satz (8.5) (Lokale Konvergenz des Newton-Verfahrens)

Sei $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, $D \subset \mathbb{R}^n$ offen und konvex. $x^* \in D$ sei ein stationärer Punkt von f. Ferner gelte mit einer Lipschitz-Konstanten $L > 0$ für alle $x, y \in D$:

a) $\nabla^2 f(x)$ regulär,

b) $\|[\nabla^2 f(x)]^{-1}(\nabla^2 f(y) - \nabla^2 f(x))\| \leq L \|y - x\|$.
Unter diesen Voraussetzungen ist das Newton-Verfahren

$$\nabla^2 f(x^k)\, d^k = -\nabla f(x^k), \quad x^{k+1} := x^k + d^k,$$

für alle Startwerte $x^0 \in D$ mit $\|x^0 - x^*\| < 2/L =: r$ und $K_r(x^*) \subset D$ wohldefiniert. Alle Iterierten x^k liegen in $K_r(x^*)$. Sie konvergieren *quadratisch* gegen x^*, genauer

$$\|x^{k+1} - x^*\| \leq \frac{L}{2} \|x^k - x^*\|^2.$$

Schließlich ist x^* einziger stationärer Punkt von f in $K_r(x^*)$.

Beweis:

Mit $g(x) := \nabla f(x)$ und $H(x) := \nabla^2 f(x)$ definiere man $\varphi : [0, 1] \to \mathbb{R}^n$ für feste Punkte $x, y \in K_r(x^*)$ durch

$$\varphi(t) := H(x)^{-1}\, g(x + t\,(y - x)), \qquad t \in [0, 1].$$

Dann ist φ stetig differenzierbar auf $[0, 1]$ mit der Ableitung

$$\varphi'(t) = H(x)^{-1}\, H(x + t\,(y - x))\,(y - x).$$

Damit folgt weiter mit der Voraussetzung b)

$$\begin{aligned} \|\varphi'(t) - \varphi'(0)\| &= \|H^{-1}(x)\,\big[H(x + t\,(y - x)) - H(x)\big]\,(y - x)\| \\ &\leq L\, t\, \|y - x\|^2. \end{aligned}$$

Hieraus erhält man nun die folgende Abschätzung

$$\begin{aligned}
\|H(x)^{-1}(g(y)-g(x)-H(x)\,(y-x))\| &= \|\varphi(1)-\varphi(0)-\varphi'(0)\| \\
&= \|\int_0^1 \big(\varphi'(t)-\varphi'(0)\big)\,dt\| \quad \le \quad \int_0^1 \|\varphi'(t)-\varphi'(0)\|\,dt \\
&\le \int_0^1 L\,t\,\|y-x\|^2\,dt \qquad\qquad = \quad \frac{L}{2}\,\|y-x\|^2
\end{aligned}$$

und schließlich

$$\begin{aligned}
x^{k+1}-x^* &= x^k - H(x^k)^{-1}g(x^k) - x^* \\
&= -H(x^k)^{-1}\big(g(x^k)-g(x^*)\big) \;+\; (x^k-x^*) \\
&= H(x^k)^{-1}\big(g(x^*)-g(x^k)-H(x^k)(x^*-x^k)\big)
\end{aligned}$$

Damit ist also $\|x^{k+1}-x^*\| \;\le\; \dfrac{L}{2}\,\|x^k-x^*\|^2$.

Gilt somit für den Startvektor $\dfrac{L}{2}\,\|x^0-x^*\| < 1$, so folgt per vollständiger Induktion

$$\|x^k-x^*\| \;\le\; \big(\frac{L}{2}\,\|x^0-x^*\|\big)^k\,\|x^0-x^*\| \;\to\; 0 \qquad (k\to\infty)$$

und somit die (monotone) Konvergenz und schließlich auch die quadratische Konvergenz der Folge (x^k) gegen x^*.

Zur Eindeutigkeit: Wäre $\tilde{x}$ ein weiterer stationärer Punkt von f in $K_r(x^*)$, so würde folgen

$$\begin{aligned}
\|\tilde{x}-x^*\| &= \|H(x^*)^{-1}\big(g(\tilde{x})-g(x^*)-H(x^*)\,(\tilde{x}-x^*)\big)\| \\
&\le \frac{L}{2}\,\|\tilde{x}-x^*\|^2 \;<\; \|\tilde{x}-x^*\|,
\end{aligned}$$

Widerspruch!! □

Beispiel (8.6)

Zu minimieren ist die Funktion $f \in \mathrm{C}^2(\mathbb{R}^2,\mathbb{R})$, gegeben durch

$$f(x_1,x_2) \;=\; x_1^4 \;+\; x_1\,x_2 \;+\; (1+x_2)^2.$$

Wir finden

$$\nabla f \;=\; \big(4x_1^3+x_2,\;\; x_1+2(1+x_2)\big)^{\mathrm{T}}, \qquad \nabla^2 f \;=\; \begin{pmatrix} 12x_1^2 & 1 \\ 1 & 2 \end{pmatrix}.$$

und somit

a) Es gibt genau einen stationären Punkt von f

$$(x_1^*, x_2^*) \approx \big(0.69588,\ -1.3479\big)^{\mathrm{T}}$$

und dieser ist ein globales Minimum mit positiv definiter Hesse-Matrix.

b) Die Hesse-Matrix $\nabla^2 f$ ist genau für $|x_1| > 1/\sqrt{24} \approx 0.20412$ positiv definit und genau für $|x_1| = 1/\sqrt{24}$ singulär.

c) Für alle Startvektoren $x^0 = (0, y)$ lautet die erste Newton-Iterierte $x^1 = (-2, 0)$.

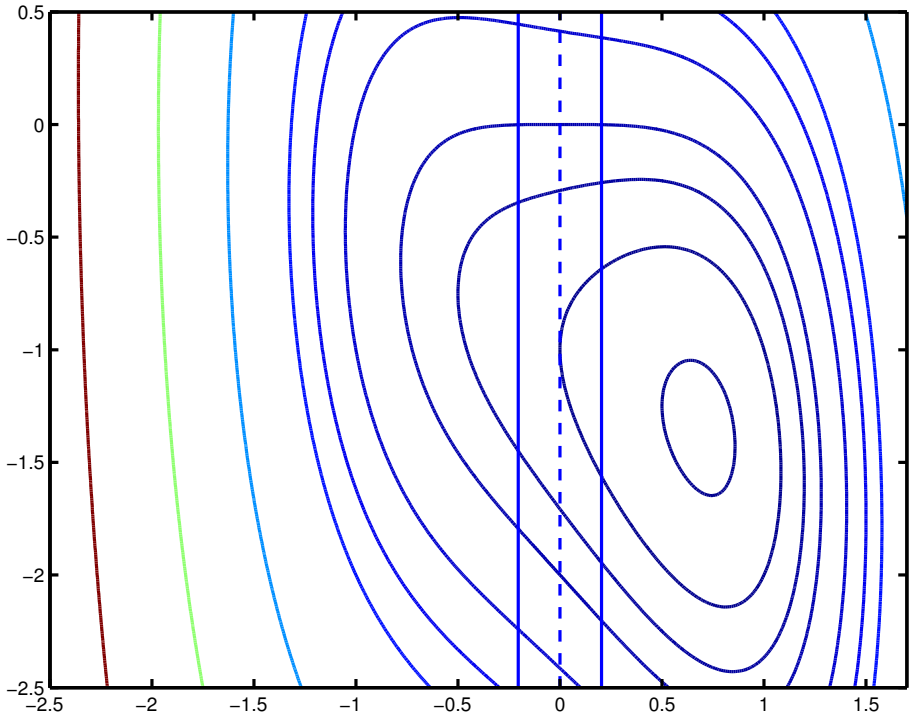


Abb. 8.1 Höhenlinien für Beispiel (8.6).

Startet man das Newton-Verfahren in einem Punkt des Streifens $|x_1| < 1/\sqrt{24}$, so liegt die nächste Iterierte links von diesem Streifen, genauer im Bereich $-\infty < x_1 < -1.86$. Damit ergibt sich für alle Startwerte in diesem Streifen oder links davon das folgende Verhalten: die Punkte wandern auf des gesuchte Minimum zu, landen im Streifen und werden dann wieder nach links geworfen. Man erhält also ein divergentes Verhalten des Newton-Verfahrens, während rechts vom Streifen schnelle Konvergenz vorliegt.

Das gedämpfte Newton-Verfahren.

Ein deutlich verbessertes globales Konvergenzverhalten des Newton-Verfahrens erhält man, wenn man mit einer Schrittweitenstrategie arbeitet. Natürlich ist das

nur in den Punkten sinnvoll, in denen die Newton-Richtung eine Abstiegsrichtung ist, bzw. in denen die Hesse-Matrix positiv definit ist.

Ferner hat man zur Sicherung der quadratischen Konvergenz darauf zu achten, dass tatsächlich die Schrittweite $t = 1$ gewählt wird, sofern diese Wahl zulässig ist.

Entscheidet man sich bei der Schrittweitenstrategie für die Armijo-Regel, so ergibt sich etwa der folgende Algorithmus (nach Geiger und Kanzow)

Algorithmus (8.7)

1.) Wähle Parameter $q > 0,\ p > 2,\ 0 < \alpha < 1,\ 0 < \sigma < 0.5,\ \mathrm{TOL} > 0.$

Wähle Startvektor $x^0 \in \mathbb{R}^n,\ k = 0;$

2.) Falls $\|\nabla f(x^k)\| < \mathrm{TOL}$: STOP;

3.) Löse das lineare Gleichungssystem $\nabla^2 f(x^k)\, d^k = -\nabla f(x^k);$

Ist $\nabla^2 f(x^k)$ (numerisch) singulär, oder gilt $\nabla f(x^k)^{\mathrm{T}} d^k > -q\, \|d^k\|^p,$

so setze man $d^k := -\nabla f(x^k).$

4.) Bestimme eine Armijo-Schrittweite $t_k = \alpha^\ell,\ \ell = 0, 1, 2, \ldots$ minimal, mit

$f(x^k + t_k\, d^k) \le f(x^k) + \sigma\, t_k\, \nabla f(x^k)^{\mathrm{T}} d^k;$

5.) $x^{k+1} := x^k + t_k\, d^k;\ k := k + 1;$ gehe zu Punkt 2.

Schrittweite t = 1?

Für die quadratische bzw. superlineare Konvergenz des gedämpften Newton-Verfahrens ist es wesentlich, dass in der Konvergenzphase auch tatsächlich die Schrittweite $t = 1$ gewählt wird. Wir untersuchen die üblichen Schrittweitenstrategien auf diese Wahlmöglichkeit und verwenden dazu als Modell wieder eine quadratische Funktion

$$f(x) = x^{\mathrm{T}} A x, \qquad A \in \mathbb{R}^n \text{ positiv definit.}$$

Hierfür ist dann $\nabla f(x) = Ax$ und $\nabla^2 f(x) = A$. Die Newton-Richtung ist also $d_N = -x$.
Ein Schritt des gedämpften Newton-Verfahrens liefert also $x^{\mathrm{neu}} = (1 - t)\, x$, so dass folgt

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = |1 - t_k|. \tag{8.8}$$

Die Konvergenz ist also nach Definition genau dann *superlinear*, wenn $t_k \to 1$ für $k \to \infty$.

Zur Untersuchung der Schrittweitenstrategien notieren wir

$$\Phi(t) \;=\; f(x+td) \;=\; (1-t)^2\,\Phi(0), \qquad \Phi'(t) \;=\; -2\,(1-t)\,\Phi(0).$$

Damit können wir die in den Schrittweitenstrategien auftretenden Bedingungen wie folgt auswerten:

- $\Phi(t) \;\le\; \Phi_o(t) \quad \Leftrightarrow \;(1-t)^2\,\Phi(0) \;\le\; \Phi(0) \;+\; \sigma\,t\,(-2\,\Phi(0))$
  $\Leftrightarrow \;(1-t)^2 \;\le\; 1-2\,\sigma\,t \quad\Leftrightarrow\quad t \;\le 2\,(1-\sigma).$
- $\Phi(t) \;\ge\; \Phi_u(t) \quad \Leftrightarrow \;(1-t)^2\,\Phi(0) \;\ge\; \Phi(0) \;+\; (1-\sigma)\,t\,(-2\,\Phi(0))$
  $\Leftrightarrow \;t \;\ge\; 2\,\sigma.$
- $\Phi'(t) \;\ge\; \rho\,\Phi'(0) \quad \Leftrightarrow \;-2\,(1-t)\Phi(0) \;\ge\; -2\,\rho\,\Phi(0)$
  $\Leftrightarrow \;t \;\ge\; 1 \;-\; \rho.$
- $|\,\Phi'(t)\,| \;\le\; -\,\rho\,\Phi'(0) \quad \Leftrightarrow \;-2\,\rho\,\Phi(0) \;\le\; -2\,(1-t)\,\Phi(0) \;\le\; 2\,\rho\,\Phi(0)$
  $\Leftrightarrow \;1 \;-\; \rho \;\le\; t \;\le\; 1 \;+\; \rho.$

An diesen Bedingungen erkennt man, dass sowohl die Armijo-Bedingung wie die Goldstein-Bedingung, die Wolfe-Powell Bedingung und auch die verschärfte Wolfe-Powell Bedingung für $t = 1$ erfüllt werden, sofern die Parameter $\sigma \in]0, 0.5[$ und $\rho \in]\sigma, 1[$ gewählt werden. Man sollte also bei den Algorithmen zur Schrittweitenstrategie jeweils im ersten Schritt prüfen, ob die Schrittweite $t = 1$ zulässig ist.

Konvergenz des gedämpften Newton-Verfahrens.

Wir zitieren (ohne Beweise) zwei Sätze zur globalen Konvergenz des gedämpften Newton-Verfahrens.

Der erste Konvergenzsatz bezieht sich direkt auf den Algorithmus (8.7) und ist dem Buch von Geiger, Kanzow (1999) entnommen.

Satz (8.9) (Konvergenz I)

Sei $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, $(x^k)_{k\in\mathbb{N}_0}$ eine durch das gedämpfte Newton-Verfahren (8.7) erzeugte Folge und x^* ein Häufungspunkt von (x^k), in dem die Hesse-Matrix $\nabla^2 f(x^*)$ positiv definit ist. Dann gelten die folgenden Aussagen:
x^* ist ein striktes lokales Minimum von f und die gesamte Folge konvergiert: $x^k \to x^*,\ k \to \infty$. Für hinreichend großes k ist d^k stets Lösung der Newton-Gleichung (8.3) und die Schrittweite ist $t_k = 1$. Ferner konvergiert (x^k) *superlinear* gegen x^* und sogar *quadratisch*, falls die Hesse-Matrix $\nabla^2 f$ lokal Lipschitz-stetig ist (vgl.die Voraussetzung b) in (8.5)).

Wir zitieren noch einen weiteren Konvergenzsatz, bei dem die Abstiegsrichtung d^k lediglich näherungsweise die Newton-Gleichung erfüllen muss. Der Satz ist dem Lehrbuch von Dennis und Schnabel (1983; Theorem 6.3.4) entnommen.

Satz (8.10) (Konvergenz II)

Sei $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, $(x^k)_{k\in\mathbb{N}_0}$ eine durch ein Abstiegsverfahren (5.4) erzeugte Folge, wobei die Schrittweiten t_k der Wolfe-Powell Bedingung (7.10) mit (festem) $\sigma \in]0, 0.5[$ und $\sigma < \rho < 1$ und die Abstiegsrichtungen d^k der folgenden Bedingung genügen

$$\frac{\|\nabla f(x^k) \; + \; \nabla^2 f(x^k)\, d^k\|}{\|d^k\|} \;\to\; 0, \qquad k \to \infty.$$

Ferner konvergiere die Folge (x^k) gegen einen Punkt x^*, in dem die Hesse-Matrix $\nabla^2 f(x^*)$ positiv definit und lokal Lipschitz-stetig ist. Dann gelten die folgenden Aussagen:
x^* ist ein striktes lokales Minimum von f. Für hinreichend großes k ist die Schrittweite $t_k = 1$ stets zulässig. Wird dann auch $t_k = 1$ gewählt, so konvergiert (x^k) *superlinear* gegen x^*.

Modifikationen der Newton-Richtung.

(a) Ansatz von Goldstein und Price: Man testet mittels Cholesky-Zerlegung, ob die Hesse-Matrix $\nabla^2 f(x)$ positiv definit, oder, ob die Newton-Richtung d_N eine Winkelbedingung (5.10) mit vorgegebener Konstanten $c > 0$ erfüllt:

$$-\frac{\nabla f(x)^{\mathrm{T}} d_N}{\|\nabla f(x)\|\,\|d_N\|} \;\geq\; c.$$

Ist dies nicht der Fall, so wähle man die negative Gradientenrichtung $d = -\nabla f(x)$.

Ein Nachteil diese Ansatzes ist es, dass häufige Gradientenwahl die Konvergenz erheblich verlangsamen kann.

(b) Ansatz von Levenberg und Marquardt (1944/1963): Ist die Hesse-Matrix nicht positiv definit, so bestimme man ein möglichst kleines $\lambda > 0$, so dass $\nabla^2 f(x) + \lambda\, I_n$ positiv definit ist und berechne eine Abstiegsrichtung d aus dem linearen Gleichungssystem

$$\big(\nabla^2 f(x) + \lambda\, I_n\big)\, d \;=\; -\nabla f(x). \tag{8.11}$$

Für $\lambda \downarrow 0$ geht (8.11) in die Newton-Gleichung über, für $\lambda \to \infty$ weist d in die Richtung des negativen Gradienten. Damit stellt das Levenberg Marquardt Verfahren einen Kompromiss zwischen Gradienten- und Newton-Verfahren dar.

Cholesky-Zerlegung.

Jede symmetrische und positiv definite Matrix $A \in \mathbb{R}^{(n.n)}$ besitzt eine so genannte (rationale) Cholesky-Zerlegung $A = L\,D\,L^{\mathrm{T}}$, wobei L eine normierte untere Dreiecksmatrix und D eine Diagonalmatrix bezeichnet. Die Cholesky-Zerlegung lässt sich mittels $L\,R$-Zerlegung (Gauß - Elimination) und einem Aufwand von $n^3/6 + \mathrm{O}(n^2)$ wesentlichen Operationen berechnen.

Algorithmus (8.12)

für $k = 1, \ldots, n$
  für $i = 1, \ldots, k-1$
    $r_{ik} \;=\; a_{ki} \;-\; \sum_{j=1}^{i-1} \ell_{ij}\, r_{jk};$
  end i;
  $d_k \;=\; a_{kk};$
  für $i = 1, \ldots, k-1$
    $t \;=\; r_{ik};$
    $\ell_{ki} \;=\; r_{ik}/d_i;$
    $d_k \;=\; d_k \;-\; \ell_{ki}\, t;$
  end i;
end k;

Der obige Algorithmus lässt sich auf beliebige symmetrische Matrizen anwenden. Er ist dann allerdings ev. nicht durchführbar, falls eins der Pivotelement d_i verschwindet. Der Algorithmus kann damit als Test auf Positive Definitheit verwendet werden. Ist er durchführbar, so gilt

$$A \text{ positiv definit } \;\Leftrightarrow\; \forall\, k \;=\; 1, \ldots, n: \;\; d_k \;>\; 0.$$

Man kann den obigen Algorithmus zur Cholesky-Zerlegung so modifizieren, dass für symmetrische, aber nicht notwendig positiv definite Hesse-Matrizen $\nabla^2 f(x)$ die Cholesky-Zerlegung einer "verschobenen" Matrix entsteht:

$$\nabla^2 f(x) \;+\; \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \;=\; L\,D\,L^{\mathrm{T}}.$$

Die $\lambda_i \geq 0$ werden im Lauf des Cholesky-Verfahrens so bestimmt, dass sie möglichst klein sind und die verschobene Matrix positiv definit ist. Dabei entsteht nur geringer Mehraufwand gegenüber einer einfachen Cholesky-Zerlegung. Man vergleiche hierzu das Lehrbuch von Gill, Murray und Wright.

Abhilfe an stationären Punkten.

Die bisher betrachteten Abstiegsverfahren brechen ab, falls eine Iterierte $x = x^k$ ein stationärer Punkt ist. Es gibt dann keine Richtung d, die der hinreichenden

Abstiegsbedingung $\nabla f(x)^{\mathrm{T}} d < 0$ genügt. Für $f \in \mathrm{C}^2$ gilt jedoch dann nach dem Taylorschen Satz:

$$f(x+td) \;=\; f(x) \;+\; \frac{1}{2}\, t^2\, d^{\mathrm{T}}\, \nabla^2 f(x)\, d \;+\; \mathrm{o}(t^2).$$

Ist also die Hesse-Matrix nicht *positiv semidefinit*, so gibt es einen Vektor $d \in \mathbb{R}^n$ mit $d^{\mathrm{T}}\, \nabla^2 f(x)\, d \,<\, 0$. Nach Obigem ist ein solcher Vektor d dann eine Abstiegsrichtung von f, d.h.

$$\exists\, t_0 > 0 :\;\; \forall\, t \in]0, t_0[:\;\; f(x+td) \;<\; f(x).$$

Man nennt d dann *eine Richtung negativer Krümmung.*

Zur Berechnung einer Richtung negativer Krümmung kann man nach **Fiacco, Mc Cormick (1968)** folgendermaßen vorgehen:

Man berechne (sofern möglich) die Cholesky-Zerlegung der Hesse-Matrix $\nabla^2 f(x) = L\, D\, L^{\mathrm{T}}$. Nach Annahme ist $\nabla^2 f(x)$ nicht positiv semidefinit. Dann gibt es Indizes $j \in \{1, \ldots, n\}$ mit $d_j < 0$. Man berechne nun d als Lösung des linearen Gleichungssystems

$$L^{\mathrm{T}}\, d \;=\; a, \qquad a_j \;:=\; \begin{cases} 1, & \text{falls } d_j < 0, \\ 0, & \text{sonst.} \end{cases}$$

Hiermit folgt dann

$$d^{\mathrm{T}} \nabla^2 f(x)\, d \;=\; d^{\mathrm{T}} L\, D\, L^{\mathrm{T}}\, d \;=\; a^{\mathrm{T}}\, D\, a \;=\; \sum_{d_j < 0} d_j \;<\; 0,$$

d.h. d ist tatsächlich eine Richtung negativer Krümmung.

Es sei ausdrücklich angemerkt, dass selbst eine reguläre, symmetrische Matrix keine Cholesky-Zerlegung besitzen muss. Man kann in diesem Fall versuchen durch geeignete Zeilen- und Spaltenvertauschungen eine zerlegbare Matrix zu erhalten:

$$P^{\mathrm{T}} \nabla^2\, f(x)\, P \;=\; L\, D\, L^{\mathrm{T}},$$

P: Permutationsmatrix.

Aufgabe (8.13)

Wenden Sie das Newton-Verfahren einmal ohne Schrittweitenbestimmung (ungedämpft) und einmal mit einer effizienten Schrittweitenbestimmung (z.B. einer Goldstein Schrittweite) an zur Minimierung der Funktion

$$f(x) \;=\; \frac{11}{546}\, x^6 \;-\; \frac{38}{364}\, x^4 \;+\; \frac{1}{2}\, x^2.$$

Startnäherung: $x_0 \;=\; 1.01$.

Zeigen Sie zur Interpretation der Ergebnisse, dass für kleine $|h|$ gilt:

$$x_k \;=\; \pm(1+h) \quad \Rightarrow \quad x_{k+1} \;=\; \mp\left(1 + \frac{h}{2}\right) \;+\; 0(h^2).$$

9. Quasi–Newton–Verfahren

Beim Newton–Verfahren besteht ein großer Anteil des numerischen Aufwandes darin, die Hesse-Matrix in jedem Schritt neu zu berechnen. Zudem ist es bei komplizierten praktischen Aufgabe ein Problem, analytische Ausdrücke für die Hesse-Matrix zu gewinnen und man ist hierzu meist auf die Verwendung von Programmpaketen wie MATHEMATICA ober MAPLE angewiesen, deren Ergebnisse aber häufig auch nur mit großem Aufwand numerisch ausgewertet werden können.

Es ist daher naheliegend, statt der (exakten) Hesse-Matrix $\nabla^2 f(x)$ eine Approximation $H \approx \nabla^2 f(x)$ zu verwenden. Man spricht dann von *Quasi-Newton-Verfahren* oder *Variable Metrik Verfahren*, *Update-Verfahren* oder *Skalierten Gradientenverfahren*.

Wesentlich ist dabei, dass die Approximationen H iterativ *aufdatiert* werden, d.h. man verwendet eine teilweise heuristische Vorschrift, die besagt, wie man aus einer Näherung H von $\nabla^2 f(x)$, $x = x^k$, und der Kenntnis der Daten $x^+ = x^{k+1}$, $f(x)$, $f(x^+)$, $\nabla f(x)$ und $\nabla f(x^+)$ zu einer neuen Approximation $H^+ \approx \nabla^2 f(x^+)$ gelangt. Genauer verwendet man *update–Formeln* der Form

$$\begin{aligned} H^+ &= \Phi(H;\, s,\, y), \\ s &:= x^+ - x, \\ y &:= \nabla f(x^+) - \nabla f(x). \end{aligned} \tag{9.1}$$

Der obige Ansatz geht historisch auf W.C. Davidon (1959) zurück und war seinerzeit revolutionär für die Entwicklung neuer numerischer Optimierungsverfahren.

Die Broyden-Approximation für nichtlineare Gleichungssysteme.

Erste update-Formeln für die Jacobi-Matrizen $Jg(x)$ zur Lösung nichtlinearer Gleichungssysteme $g(x) = 0$, mit $g \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R}^n)$, gehen auf C.G. Broyden (1965) zurück. Für das Newton-Verfahren

$$\begin{aligned} Jg(x)\, d_N &= -g(x), \\ x^+ &= x + t\, d_N \end{aligned} \tag{9.2}$$

soll $Jg(x^+)$ ersetzt werden durch eine Näherung H^+.

Aufgrund der Taylor-Entwicklung von g zum Entwicklungspunkt x^+ fordern wir

$$H^+ (x^+ - x) = g(x^+) - g(x). \tag{9.3}$$

Die Forderung (9.3) heißt *Quasi-Newton-Bedingung* oder *Sekantenbedingung*. Man beachte, dass in (9.3) alle Daten - bis auf das gesuchte H^+ - bekannt sind. (9.3) liefert also n (lineare) Gleichungen für n^2 Unbekannte.

Da keine weiteren Information über H^+ vorliegen, verlangt man, dass sich H und H^+ in Richtungen senkrecht zu $s := x^+ - x$ nicht unterscheiden, dass also gilt

$$\forall\, v \perp s := (x^+ - x): \quad (H^+ - H)\, v \;=\; 0,$$

wobei H die Jacobi-Matrix $Jg(x)$ bzw. eine Näherung dieser bezeichnet. Damit ist klar, dass H^+ die folgende Darstellung besitzen muss

$$H^+ \;=\; H \;+\; \frac{u\, s^{\mathrm{T}}}{s^{\mathrm{T}}\, s}\,.$$

Der unbekannte Vektor $u \in \mathbb{R}^n$ wird durch die Quasi-Newton Bedingung (9.3) festgelegt:

$$H^+ s \;=\; H\, s \,+\, u \;=\; g(x^+) \,-\, g(x).$$

Damit erhalten wir die *Broyden–Approximation*

$$\begin{aligned} H^+ \;&=\; H \;+\; \frac{(y \,-\, H\, s)\, s^{\mathrm{T}}}{s^{\mathrm{T}}\, s}, \\ s \;&:=\; x^+ - x, \\ y \;&:=\; g(x^+) \,-\, g(x). \end{aligned} \tag{9.4}$$

Die Broyden-Approximation (9.4) lässt sich auch durch die folgende Eigenschaft motivieren.

Satz (9.5)

Die Broyden-Approximation minimiert den Fehler $\|H^+ - H\|_2$ bei vorgegebener Matrix $H = Jg(x)$ für alle Matrizen H^+, die der Quasi-Newton-Bedingung $H^+ s = y$ genügen.

Bemerkungen (9.6)

a) Ersetzt man im globalisierten Newton-Verfahren (9.2) die Jacobi-Matrix durch die Broyden-Approximation, so konvergiert das Verfahren unter geeigneten Voraussetzungen an g und die Schrittweitenwahl superlinear gegen eine Nullstelle x^* von g, d.h.

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

b) Zur Lösung der Newton-Gleichung

$$H^+\, d_N \;=\; -g(x), \quad \text{mit} \quad H^+ \,=\, H \,+\, u\, v^{\mathrm{T}}$$

lässt sich die so genannte *Sherman-Morrison-Formel* verwenden:

$$(H \,+\, u\, v^{\mathrm{T}})^{-1} \;=\; H^{-1} \;-\; \frac{H^{-1}\, u\, v^{\mathrm{T}}\, H^{-1}}{1 \,+\, v^{\mathrm{T}}\, H^{-1}\, u}\,. \tag{9.7}$$

Ferner ist die Matrix H^+ nur dann singulär, wenn der Nenner in der Relation (9.7) verschwindet.

c) Es gibt weitere so genannte *Rang-1-update Formeln*. Von Broyden selbst wurde die folgende update-Formel für die *inverse* Jacobi-Matrix $B \approx Jg(x)^{-1}$ angegeben

$$B^+ \;=\; B \;+\; \frac{(s - By)\, y^{\mathrm{T}}}{y^{\mathrm{T}} y}\,. \tag{9.8}$$

Auch für diese update-Formel lässt sich die superlineare Konvergenz des entsprechenden Quasi-Newton-Verfahrens beweisen. Allerdings hat sich (9.8) in der Praxis als weniger erfolgreich herausgestellt (*Broyden's bad update*).

Update-Formeln für Hesse-Matrizen.

Wir beginnen mit einem Modell-Algorithmus für das Quasi-Newton-Verfahren zur Minimierung einer Funktion $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$:

Modell-Algorithmus (9.9)

1.) Startdaten: $x^0 \in \mathbb{R}^n$, $g^0 := \nabla f(x^0)$, $H_0 \in \mathbb{R}^{(n,n)}$ symmetrisch und positiv definit, $k := 0$, TOL: Genauigkeitsschranke;

2.) Falls $\|g^k\| < \mathrm{TOL}$ oder $k > k_{\max}$: Stop;

3.) Löse das lin. Gl.system $H_k\, d^k \;=\; -g^k$ (Abstiegsrichtung);

4.) Schrittweitenbestimmung (Armijo, Goldstein,...); $x^{k+1} \;:=\; x^k + t_k\, d^k$;

5.) $s^k \;:=\; x^{k+1} - x^k$; $\quad g^{k+1} \;:=\; \nabla f(x^{k+1})$; $\quad y^k \;:=\; g^{k+1} - g^k$;

6.) $H_{k+1} \;:=\; \Phi(H_k; s^k, y^k)$; $\quad$ (update-Formel)

7.) $k \;:=\; k+1$; Gehe zu 2.);

Verwendet man wie oben update-Formeln für die Hesse-Matrix selbst $H \approx \nabla^2 f(x)$, so spricht man von *direkten Quasi-Newton-Verfahren*. Alternativ kann man auch mit update-Formeln für die inverse Hesse-Matrix $B \approx [\nabla^2 f(x)]^{-1}$ arbeiten. Anstelle der Lösung des linearen Gleichungssystems in Schritt 3.) hat man dann nur eine Matrix-Multiplikation $d^k := -B_k\, g^k$ auszuführen. Man spricht dann von einem *inversen Quasi-Newton-Verfahren*. Die *Quasi-Newton-Bedingung* bleibt dabei sinngemäß erhalten

$$H^+ s \;=\; y, \qquad \text{oder} \qquad B^+ y \;=\; s. \tag{9.10}$$

Diese Bedingung ist eine wesentliche Voraussetzung zum Nachweis der superlinearen Konvergenz des Verfahrens.

Für die Approximation von Hesse-Matrizen lässt sich die Broydensche update-Formel (9.4) nicht verwenden. Diese erhält nämlich weder die Symmetrie noch die positive Definitheit der Approximationen.

Rang - 1 - Ansatz.

Wir versuchen diese Forderungen mit einem allgemeinen Ansatz einer Rang - 1 - Korrektur der Hesse-Matrix zu erfüllen. Die geforderte Symmetrie legt den folgenden Ansatz nahe $H^+ = H + \gamma\, u\, u^{\mathrm{T}},\ u \in \mathbb{R}^n$. Diesen setzen wir in die Quasi-Newton-Gleichung ein und erhalten

$$\begin{aligned} H\, s \;+\; \gamma\, u\, \left(u^{\mathrm{T}}\, s\right) \;&=\; y \\ \Rightarrow \qquad u \;=\; y \;-\; H\, s, \qquad \gamma \;&=\; \frac{1}{u^{\mathrm{T}}\, s}. \end{aligned}$$

Eine *symmetrische Rang-1-Approximation*, die zugleich die Quasi-Newton-Bedingung erfüllt, ist damit gegeben durch

$$H^+ \;=\; H \;+\; \frac{(y - H\, s)\,(y - H\, s)^{\mathrm{T}}}{(y - H\, s)^{\mathrm{T}}\, s}\,. \tag{9.11}$$

Zwar erhält diese Rang- 1-Formel die Symmetrie, aber leider i. Allg. nicht die positive Definitheit. Überdies kann der Nenner in (9.11) eventuell verschwinden. Damit ist klar, dass man zur Approximation von Hesse-Matrizen einen weitergehenden Ansatz benötigt.

Rang - 2 - Ansatz.

Wir verwenden nun einen symmetrischen Rang-2-Ansatz

$$H^+ \;=\; H \;+\; \gamma\, u\, u^{\mathrm{T}} \;+\; \delta\, v\, v^{\mathrm{T}}, \qquad u, v \,\in\, \mathbb{R}^n.$$

Die Quasi-Newton-Bedingung ergibt

$$H\, s \;+\; \gamma\, (u^{\mathrm{T}} s)\, u \;+\; \delta\, (v^{\mathrm{T}} s)\, v \;=\; y,$$

d.h. $(y - H\, s)$ ist Linearkombination von u und v. Setzt man nun $u := y$ und $v := H\, s$ und nimmt an, dass diese Vektoren linear unabhängig sind, so folgt

$$\gamma \;=\; \frac{1}{u^{\mathrm{T}} s} \;=\; \frac{1}{y^{\mathrm{T}} s}, \qquad \delta \;=\; \frac{-1}{v^{\mathrm{T}} s} \;=\; \frac{-1}{s^{\mathrm{T}} H\, s}\,.$$

Damit haben wir die so genannte **BFGS-update-Formel**, benannt nach Broyden, Fletcher, Goldfarb und Shanno ($\approx$ 1970), hergeleitet:

$$H^+ \;=\; H \;+\; \frac{y\, y^{\mathrm{T}}}{y^{\mathrm{T}} s} \;-\; \frac{(H\, s)\,(H\, s)^{\mathrm{T}}}{s^{\mathrm{T}} H\, s}\,. \tag{9.12}$$

Ist (9.12) wohldefiniert?

Bemerkung (9.13)

Verlangt man Symmetrie und positive Definitheit von H^+ (bzw. B^+) bei vergegebener positiv definiten Matrix H (bzw. B), so kann die Quasi-Newton-Bedingung $H^+ s = y$ bzw. $B^+ y = s$ nur erfüllt werden, wenn die folgende notwendige Bedingung gilt

$$y^{\mathrm{T}} s \;>\; 0. \tag{9.14}$$

Unter dieser Voraussetzung ist also die BFGS - Approximation (9.12) wohldefiniert.

Satz (9.15)

Ist H symmetrisch und positiv definit und gilt die notwendige Bedingung $s^{\mathrm{T}} y > 0$, so ist auch die BFGS-Approximation H^+ nach (9.12) auch symmetrisch und positiv definit.

Beweis: Durch direktes Ausrechnen von $z^{\mathrm{T}} H^+ z$ und Abschätzungen von $z^{\mathrm{T}} H\, s$ mittels Cauchy-Schwarzscher Ungleichung bezüglich des Skalarproduktes $\langle u, v\rangle_H := u^{\mathrm{T}} H\, v$. Die Einzelheiten sind in einer Übungsaufgabe auszuführen. □

Wählt man den obigen symmetrischen Rang - 2 - Ansatz analog für die Approximation B^+ der inversen Hesse-Matrix, so erhält man aus der Quasi-Newton-Gleichung die folgende update-Formel

$$B^+ \;=\; B \;+\; \frac{s\, s^{\mathrm{T}}}{y^{\mathrm{T}} s} \;-\; \frac{(B\, y)\,(B\, y)^{\mathrm{T}}}{y^{\mathrm{T}} B\, y}\,. \tag{9.16}$$

(9.16) heißt **DFP-update Formel** nach Davidon, Fletcher und Powell. Wie immer sind hierbei $s := x^+ - x$ und $y := \nabla f(x^+) - \nabla f(x)$.

Alternativ lassen sich Rang-2-update Formeln durch Anwendung der Sherman-Morrison Formel, vgl. (9.7), invertieren. Allerdings ist hierzu ein erheblicher Umformungsaufwand nötig. Wendet man diese Methode auf die beiden update Formeln (9.12) (direkte BFGS-Formel) und (9.16) (inverse DFP-Formel) an, so erhält man nach einiger Rechnung

a) die *inverse BFGS-update Formel*:

$$B^+ \;=\; B \;+\; \frac{(s - By)\, s^{\mathrm{T}} + s\,(s - By)^{\mathrm{T}}}{y^{\mathrm{T}} s} \;-\; \frac{(s - By)^{\mathrm{T}}\, y}{(y^{\mathrm{T}} s)^2}\, s\, s^{\mathrm{T}}\,, \tag{9.17}$$

sowie

b) die *direkte DFP-update Formel*:

$$H^+ \;=\; H \;+\; \frac{(y - Hs)\, y^{\mathrm{T}} + y\,(y - Hs)^{\mathrm{T}}}{y^{\mathrm{T}} s} \;-\; \frac{(y - Hs)^{\mathrm{T}}\, s}{(y^{\mathrm{T}} s)^2}\, y\, y^{\mathrm{T}}\,. \tag{9.18}$$

Anmerkungen (9.19)

a) Es gibt eine Vielzahl weiterer Update-Formeln (Stichwort: Broyden Familie). Nach den bisherigen numerischen Erfahrungen haben sich die beiden BFGS-Formeln als besonders effizient erwiesen.

In Geiger, Kanzow wird auch die so genannte *Powell-symmetric-Broyden-Formel* hergeleitet

$$H^+ \;=\; H \;+\; \frac{(y-Hs)s^{\mathrm{T}} + s(y-Hs)^{\mathrm{T}}}{s^{\mathrm{T}}s} \;-\; \frac{(y-Hs)^{\mathrm{T}}\,s}{(s^{\mathrm{T}}s)^2}\, s\, s^{\mathrm{T}}\,, \tag{9.20}$$

die den Fehler $\|H^+ - H\|_F$ unter allen symmetrischen Matrizen, die die Quasi-Newton-Gleichung erfüllen, minimiert; vgl. auch Satz (9.5). Hierbei bezeichnet $\|A\|_F \;=\; (\sum_{i,j} a_{ij}^2)^{0.5}$ die Frobenius-Norm.

b) Auf den ersten Blick scheint es naheliegend zu sein, mit inversen update-Formeln zu arbeiten, da dann die Lösung der linearen Gleichungssysteme entfällt. Der Nachteil ist jedoch, dass man dann kaum eine Kontrolle darüber hat, ob H^+ noch numerisch ausreichend positiv definit ist.

c) Arbeitet man mit der direkten BFGS-update-Formel, so kann man nach Dennis, Schnabel auch die Cholesky-Zerlegungen von H aufdatieren. Ist nämlich $H = L\,L^{\mathrm{T}}$ mit einer regulären unteren Dreiecksmatrix L, so folgt für die nach (9.12) aufdatierte Matrix

$$H^+ \;=\; J\,J^{\mathrm{T}}, \quad J \;=\; L \;+\; \frac{(y-Lw)w^{\mathrm{T}}}{w^{\mathrm{T}}w}\,, \quad w \;=\; \sqrt{\frac{y^{\mathrm{T}}s}{s^{\mathrm{T}}H\,s}}\; L^{\mathrm{T}}s.$$

Man berechnet hiernach die i. Allg. vollbesetzte Matrix J, sowie eine $Q\,R$-Zerlegung von J^{T} (dies kann effizient mit Hilfe von Givens-Rotationen erfolgen). Damit gilt dann

$$H^+ \;=\; J\,J^{\mathrm{T}} \;=\; R^{\mathrm{T}}Q^{\mathrm{T}}Q\,R \;=\; R^{\mathrm{T}}R.$$

Damit ist mit $L^+ := R^{\mathrm{T}}$ eine neue Cholesky-Zerlegung von H^+ bestimmt.

Wir zitieren zwei Konvergenzsätze zum Quasi-Newton-Verfahren. Der erste ist ein lokaler Konvergenzsatz für das inverse BFGS-Verfahren ohne Schrittweitenbestimmung (d.h. alle $t_k = 1$). Der zweite Satz ist eine globale Konvergenzaussage für das direkte BFGS-Verfahren, wobei die Schrittweitenwahl den Wolfe-Powell-Bedingungen genügen.

Satz (9.21) (Lokaler Konvergenzsatz)

Sei $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, $\nabla^2 f$ lokal Lipschitz-stetig und $x^* \in \mathbb{R}^n$ ein striktes lokales Minimum von f mit positiv definiter Hesse-Matrix $\nabla^2 f(x^*)$. Dann existieren $\delta, \varepsilon > 0$, so dass das inverse BFGS-Verfahren (mit Schrittweite $t_k = 1$) für alle Startvektoren $x^0 \in K_\varepsilon(x^*)$ und alle Startmatrizen B_0 mit $\|B_0 - \nabla^2 f(x^*)^{-1}\|_F < \delta$ wohldefiniert ist und eine Folge (x^k) erzeugt, die superlinear gegen x^* konvergiert.

Beweis: Geiger, Kanzow, Satz 11.33.

Satz (9.22) (Globaler Konvergenzsatz)

Sei $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, $x^0 \in \mathbb{R}^n$ ein beliebiger Startvektor mit konvexer Levelmenge $L_f(x^0)$. Ferner sei f gleichmäßig konvex auf $L_f(x^0)$, $\nabla^2 f$ lokal Lipschitz-stetig und $x^* \in \mathbb{R}^n$ ein striktes globales Minimum von f mit positiv definiter Hesse-Matrix $\nabla^2 f(x^*)$.
Mit beliebiger symmetrischer und positiv definiter Startmatrix H_0 ist dann das direkte BFGS-Verfahren mit Schrittweitenbestimmung nach Wolfe-Powell durchführbar und die erzeugte Folge x^k konvergiert gegen x^*. Die Konvergenz ist superlinear, falls $\lim_{k\to\infty} t_k = 1$ gilt.

Beweis: Werner: Numerische Mathematik 2, Abschnitte 7.3.3 und 7.3.4.

Aufgabe (9.23)

Für die Broyden–Rang 1–Approximation H_{k+1} der Jacobi–Matrix $Jg(x^{k+1})$ einer Funktion $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$:

$$\begin{aligned} H_{k+1} &:= H_k + \frac{(y^k - H_k s^k)(s^k)^{\mathrm{T}}}{(s^k)^{\mathrm{T}} s^k}, \\ s^k &:= x^{k+1} - x^k, \qquad y^k := g(x^{k+1}) - g(x^k) \end{aligned}$$

ist zu zeigen:

H_{k+1} minimiert den Abstand $\|H - H_k\|_2$ für alle Matrizen $H \in \mathbb{R}^{(n,n)}$, die die Quasi–Newton Bedingung $Hs^k = y^k$ erfüllen.

Aufgabe (9.24)

Gegeben sei eine symmetrische, positiv definite Matrix $H \in \mathbb{R}^{(n,n)}$.

Zeigen Sie, dass die durch die BFGS–Formel (9.12) bestimmte Matrix

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

genau dann positiv definit ist, wenn $y^{\mathrm{T}} s > 0$ ist. Zeigen Sie weiter:

Im Fall $\nabla f(x_0) \neq 0$ ist diese Bedingung erfüllt, wenn die verwendete Schrittweite den Wolfe-Powell Bedingungen genügt.

Hinweis: Man benutze die Cauchy-Schwarzsche Ungleichung für das Skalarprodukt

$$\langle z, s \rangle := z^{\mathrm{T}} Hs.$$

10. Das Verfahren der konjugierten Gradienten

Wir schließen an die früheren Anmerkungen über eine geeignete Skalierung des Gradientenverfahrens an. Unser Ziel ist die Konstruktion eines Verfahrens, das
a) für quadratische Optimierungsaufgaben in endlich vielen Schritten das (exakte) Minimum findet, und
b) die Abspeicherung und auch das Aufdatieren einer (n,n)-Matrix in jedem Iterationsschritt vermeidet.

CG-Verfahren für quadratische Zielfunktionen.

Das folgende Verfahren geht in seiner Grundform auf **Hestenes und Stiefel** (Journal of Research of the National Bureau of Standards, Bd. 49, 1952) zurück.

Gegeben sei eine quadratische Funktion

$$f(x) \;=\; \frac{1}{2}\, x^{\mathrm{T}} A\, x \;+\; b^{\mathrm{T}} x \;+\; c \tag{10.1}$$

mit $A \in \mathbb{R}^{(n,n)}$ symmetrisch und positiv definit. Wir definieren

Definition (10.2)

Vektoren $d^1, \ldots, d^n \in \mathbb{R}^n \setminus \{0\}$ heißen *konjugiert* bzgl. A - oder *A–konjugiert*, falls

$$\forall\, i \neq j \in \{1, \ldots, n\}: \quad (d^i)^{\mathrm{T}} A\, d^j \;=\; 0.$$

Bemerkungen (10.3)

a) A–konjugierte Vektoren $(d^1, \ldots, d^n)$ sind also orthogonale Vektoren bezüglich des Skalarproduktes $\langle u, v\rangle_A := u^{\mathrm{T}} A\, v$. Insbesondere sind sie daher linear unabhängig und sie bilden eine Basis des $\mathbb{R}^n$.

b) Aus einer beliebigen Basis $(g^1, \ldots, g^n)$ des $\mathbb{R}^n$ erhält man mit Hilfe des Gram-Schmidtschen Orthogonalisierungsverfahrens eine A–konjugierte Basis:

$$\begin{aligned} &d^1 \;:=\; -\, g^1; \\ &\text{für } k \;=\; 2, \ldots, n \\ &\qquad d^k \;:=\; -\, g^k \;+\; \sum_{j=1}^{k-1} \frac{\langle d^j, g^k\rangle_A}{\langle d^j, d^j\rangle_A}\, d^j\,; \\ &\text{end } k \end{aligned} \tag{10.4}$$

c) Aufgrund der quadratischen Zielfunktion kann man umformen:

$$\begin{aligned} f(x+d) &= \frac{1}{2}\,(x+d)^{\mathrm{T}}\,A\,(x+d) \;+\; b^{\mathrm{T}}\,(x+d) \;+\; c \\ &= f(x) \;+\; \frac{1}{2}\,d^{\mathrm{T}}A\,d \;+\; (A\,x+b)^{\mathrm{T}}\,d \\ &= f(x) \;+\; \frac{1}{2}\,d^{\mathrm{T}}A\,d \;+\; g(x)^{\mathrm{T}}\,d. \end{aligned}$$

Stellt man nun den Vektor d bezüglich einer A–konjugierten Basis $(d^1,\ldots,d^n)$ dar, $d=\sum\, t_k\,d^k$, so ergibt sich

$$\begin{aligned} f(x \;+\; \sum_k\, t_k\,d^k) &= f(x) \;+\; \sum_{k=1}^{n}\left[\frac{(d^k)^{\mathrm{T}}A\,d^k}{2}\,t_k^2 \;+\; (g(x)^{\mathrm{T}}d^k)\,t_k\right] \\ &=: f(x) \;+\; \sum_{k=1}^{n} F_k(t_k). \end{aligned} \tag{10.5}$$

Die zu minimierende Funktion f ist also bezüglich der Koordinaten t_k entkoppelt. Jede der Funktionen F_k ist dabei eine quadratische Funktion in t_k mit positivem Koeffizienten von t_k^2. Nach n eindimensionalen Minimierungen bzgl. t_k, $k=1,\ldots,n$, erhält man also das gesuchte Minimum x^*.

Damit erhalten wir den folgenden Modellalgorithmus:

Modellalgorithmus (10.6)

Start: $x^1\in\mathbb{R}^n,\; g^1 := \nabla f(x^1),\; d^1 := -g^1;$

Für $k \;=\; 1,\,2,\,\ldots,\,n$

Falls $\|g^k\| \;\le\; \mathrm{TOL}:$ Stop!

Für $k>1:$ Bestimme $d^k\neq 0$ mit $\forall\, j<k:\;\; (d^k)^{\mathrm{T}}A\,d^j=0;$

$t_k \;:=\; \mathrm{argmin}\, f(x^k \;+\; t\,d^k);$

$x^{k+1} \;:=\; x^k \;+\; t_k\,d^k;\quad g^{k+1} \;:=\; \nabla f(x^{k+1});$

end k.

Wir stellen nun die Eigenschaften zusammen, mit deren Hilfe sich der Modellalgorithmus konkretisieren lässt. Dabei beizeichnet k den Iterationsindex.

Schritt 1:

Aus $t_k := \mathrm{argmin} f(x^k+t\,d^k)$ ergibt sich bei quadratischer Zielfunktion (10.1):

$$\nabla f(x^k+t_k d^k)^{\mathrm{T}}d^k \;=\; 0.$$

Mit der Abkürzung $g^{k+1} := \nabla f(x^k+t_k d^k) \;=\; A(x^k+t_k d^k)+b \;=\; g^k+t_k\,A\,d^k$ folgt hieraus

$$\forall\, k: \quad (g^{k+1})^{\mathrm{T}} d^k \;=\; 0, \qquad g^{k+1} - g^k \;=\; t_k\, A\, d^k \tag{10.7}$$

und damit

$$\forall\, k: \quad t_k \;=\; -\,\frac{(g^k)^{\mathrm{T}} d^k}{(d^k)^{\mathrm{T}} A\, d^k}\,. \tag{10.8}$$

Schritt 2:

Aus (10.7) und der A-Konjugiertheit der d^j folgt für $j < k$

$$(g^{k+1} - g^k)^{\mathrm{T}} d^j \;=\; t_k\, (d^k)^{\mathrm{T}} A\, d^j \;=\; 0$$

und hiermit und mit (10.7) auch

$$(g^{k+1})^{\mathrm{T}} d^j \;=\; (g^{j+1})^{\mathrm{T}} d^j \;+\; (g^{j+2} - g^{j+1})^{\mathrm{T}} d^j \;+\; \ldots \;+\; (g^{k+1} - g^k)^{\mathrm{T}} d^j \;=\; 0.$$

Als Ergebnis halten wir fest

$$\forall\, j \le k: \quad (g^{k+1})^{\mathrm{T}} d^j \;=\; 0. \tag{10.9}$$

Schritt 3:

Bricht der Modellalgorithmus im k-ten Schritt nicht ab (bei TOL = 0), d.h. gilt $g^{k+1} \neq 0$, so gilt mit (10.9):

$$g^{k+1} \;\perp\; \mathrm{Spann}(d^1, \ldots, d^k) \tag{10.10}$$

und das Gram-Schmidtsche Orthogonalisierungverfahren lässt sich mit $(-g^{k+1})$ als neuem Basisvektor durchführen.

Mit (10.4) ergibt sich die Relation

$$d^{k+1} \;=\; -\,g^{k+1} \;+\; \sum_{j=1}^{k} \frac{(g^{k+1})^{\mathrm{T}} A\, d^j}{(d^j)^{\mathrm{T}} A\, d^j}\, d^j\,. \tag{10.11}$$

Multipliziert man diese Gleichung mit $(g^{k+1})^{\mathrm{T}}$, so erhält man mit (10.9)

$$(g^{k+1})^{\mathrm{T}} d^{k+1} \;=\; -\,\|g^{k+1}\|^2 \;<\; 0. \tag{10.12}$$

Insbesondere ist d^{k+1} also eine Abstiegsrichtung und mit (10.8) ergibt sich $t_{k+1} > 0$.

Schritt 4:

Wir zeigen nun, dass in der Relation (10.11) alle Summanden $j < k$ verschwinden. Nach (10.11) ist $g^{k+1} \in \mathrm{Spann}(d^1, \ldots, d^{k+1})$, aber mit (10.10) zugleich $g^{k+1} \perp \mathrm{Spann}(d^1, \ldots, d^k)$. Damit folgt

$$\forall\, j \le k: \;\; (g^{k+1})^{\mathrm{T}} g^j \;=\; 0 \tag{10.13}$$

und somit auch für $j < k$ mitteles (10.7):

$$(g^{k+1})^{\mathrm{T}} A\, d^j \;=\; \frac{1}{t_j}\,(g^{k+1})^{\mathrm{T}}(g^{j+1} - g^j) \;=\; 0.$$

Damit reduziert sich (10.11) auf zwei Summanden

$$d^{k+1} \;=\; -g^{k+1} \;+\; \frac{(g^{k+1})^{\mathrm{T}} A\, d^k}{(d^k)^{\mathrm{T}} A\, d^k}\, d^k\,. \tag{10.14}$$

Aus (10.7) und (10.13) folgt weiterhin

$$(g^{k+1})^{\mathrm{T}} A\, d^k \;=\; \frac{1}{t_k}\left(\|g^{k+1}\|^2 \;-\; (g^{k+1})^{\mathrm{T}} g^k\right) \;=\; \frac{1}{t_k}\,\|g^{k+1}\|^2$$

sowie aus (10.8) und (10.12)

$$(d^k)^{\mathrm{T}} A\, d^k \;=\; -\frac{1}{t_k}\,(g^k)^{\mathrm{T}} d^k \;=\; \frac{1}{t_k}\,\|g^k\|^2\,.$$

Diese Relationen in (10.14) eingesetzt ergibt

$$d^{k+1} \;=\; -g^{k+1} \;+\; \frac{(g^{k+1})^{\mathrm{T}} g^{k+1}}{(g^k)^{\mathrm{T}} g^k}\, d^k\,. \tag{10.15}$$

Bemerkung (10.16)

Wir fassen nochmals die bisher gezeigten Orthogonalitätseigenschaften zusammen. Dabei beziehen wir uns auf die Definition der A-Konjugiertheit sowie auf (10.9), (10.12) und (10.13). Für $j < k$ gelten:

$$(d^k)^{\mathrm{T}} A\, d^j \;=\; 0 \quad \text{nach (10.2)}, \qquad (g^k)^{\mathrm{T}} g^j \;=\; 0 \quad \text{nach (10.13)},$$

$$(g^k)^{\mathrm{T}} d^j \;=\; 0 \quad \text{nach (10.9)}, \qquad (g^k)^{\mathrm{T}} d^k \;=\; -\,\|g^k\|^2 \quad \text{nach (10.12)}.$$

Algorithmus (10.17) (CG–Verfahren)

1.) Startdaten: $x^1 \in \mathbb{R}^n$, $g^1 := \nabla f(x^1)$, $d^1 := -g^1$, $k := 1$, $\mathrm{TOL} > 0$;

2.) $t_k := \|g^k\|^2/((d^k)^{\mathrm{T}} A\, d^k)$, $x^{k+1} := x^k + t_k d^k$, $g^{k+1} := g^k + t_k\, A\, d^k$;

3.) Falls $\|g^{k+1}\| \le \mathrm{TOL}$ oder $k > k_{\max}$: Stop;

4.) $s_k := \|g^{k+1}\|^2/\|g^k\|^2$, $d^{k+1} := -g^{k+1} + s_k\, d^k$;

5.) $k := k + 1$; Gehe zu 2.);

Anmerkungen zu (10.17)

a) In jedem Iterationsschritt des CG–Verfahrens ist lediglich ein Produkt "Matrix $*$ Vektor" zu berechnen (Aufwand $O(n^2)$). Die anderen Operationen (zwei Skalarprodukte, drei Vektoradditionen) sind weniger aufwendig ($O(n)$).

b) Die A-konjugierten Vektoren d^k lassen sich nach Hestenes (1980) folgendermaßen charakterisieren: d^k ist - bis auf ein skalares Vielfaches - Lösung der Optimierungsaufgabe:

$$\begin{aligned} &\text{Minimiere} \quad (g^k)^{\mathrm{T}} d \\ &\text{Nebenbed.:} \quad d^{\mathrm{T}} d = 1, \quad d^{\mathrm{T}} A\, d^j = 0 \quad (j = 1, \ldots, k-1). \end{aligned}$$

c) Besitzt A genau m verschiedene Eigenwerte, so liefert das CG–Verfahren das gesuchte Minimum nach höchstens m Iterationen.

d) Für den Approximationsfehler des CG–Verfahrens lässt sich zeigen (Luenberger, 1973)

$$\|x^{k+1} - x^*\|_A \;\leq\; 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x^1 - x^*\|_A, \tag{10.18}$$

wobei $\kappa := \mathrm{cond}_2(A)$ die spektrale Konditionszahl bezeichnet, $\mathrm{cond}_2(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$.

Man vergleiche dieses Ergebnis mit der Aussage in (6.9) über die Konvergenzgeschwindigkeit des Gradientenverfahrens.

e) Das CG-Verfahren wird häufig zur iterativen Lösung linearer Gleichungssysteme $g(x) = A\,x + b = 0$ verwendet, insbesondere für große Dimensionszahlen n und dünn besetzte Matrizen A.

f) Ist A nicht symmetrisch aber regulär, so kann man zur Lösung eines linearen Gleichungssystems $Ax = b$ zur Normalgleichung $A^{\mathrm{T}}Ax = A^{\mathrm{T}}b$ übergehen (vgl. lineare Ausgleichsprobleme) und hierauf das CG-Verfahren anwenden. Methoden dieser Art heißen auch CGS-Verfahren (conjugate gradient squared methods). Hierzu gibt es wichtige Resultate von Stoer, Sonnevelt, Fletcher und anderen.

Präkonditionierung.

Man erkennt an der Relation (10.18), dass das CG-Verfahren um so schneller konvergiert, je kleiner die Konditionszahl $\kappa = \mathrm{cond}_2(A)$ ist. Es ist daher naheliegend, diese Konditionszahl durch eine geeignete Skalierung der Zielfunktion zu verkleinern.

Verwendet man eine lineare Transformation $x =: S\,z$ der Variablen x auf die Variable z - dabei sei $S \in \mathbb{R}^{(n,n)}$ regulär - und wendet man das normale CG-Verfahren auf die transformierte Zielfunktion

$$\tilde{f}(z) \;:=\; f(x) \;=\; \frac{1}{2} z^{\mathrm{T}} S^{\mathrm{T}} A S z \;+\; (S^{\mathrm{T}} b)^{\mathrm{T}} z \;+\; c$$

an, so ergibt sich nach der Rücktransformation

$$x^k \;=\; S\,z^k, \quad \tilde{g}^k \;=\; S^{\mathrm{T}} g^k, \quad d^k \;=\; S\,\tilde{d}^k$$

das folgende präkonditionierte CG-Verfahren (PCG-Verfahren):

Algorithmus (10.19) (PCG–Verfahren)

1.) Startdaten: $x^1 \in \mathbb{R}^n, \quad g^1 := \nabla f(x^1), \quad k := 1, \quad \mathrm{TOL} > 0;$

2.) Wähle reguläre Transformationsmatrix $S, \quad B \;:=\; S\,S^{\mathrm{T}}, \quad d^1 := -B\,g^1;$

3.) $t_k := \dfrac{(g^k)^{\mathrm{T}} B\,g^k}{(d^k)^{\mathrm{T}} A\,d^k}, \quad x^{k+1} := x^k + t_k\,d^k, \quad g^{k+1} := g^k + t_k\,A\,d^k;$

4.) Falls $\|g^{k+1}\| \le \mathrm{TOL}$ oder $k > k_{\max}$: Stopp;

5.) $s_k := \dfrac{(g^{k+1})^{\mathrm{T}} B\,g^{k+1}}{(g^k)^{\mathrm{T}} B\,g^k}, \quad d^{k+1} := -B\,g^{k+1} + s_k\,d^k;$

6.) $k \;:=\; k+1;$ Gehe zu 3.);

Im eigentlichen Algorithmus tritt nur die symmetrische und positiv definite Matrix $B = S\,S^{\mathrm{T}}$ auf. Diese Matrix ist so zu wählen, dass die transformierte Hesse-Matrix $S^{\mathrm{T}}AS$ von $\tilde{f}(z)$ bzw. die zu ihr ähnliche Matrix $B\,A$ eine möglichst kleine Kondition hat, wobei andererseits darauf zu achten ist, dass die Multiplikation By einfach zu berechnen bleibt. (Die naheliegende Wahl $B = A^{-1}$ ist also sinnlos!)

Die folgenden Standardansätze liefern häufig günstige Wahlen für B:

a) *Diagonale Vorkonditionierung*: Setze $B = D^{-1}$, wobei $D = \mathrm{diag}(a_{11}, \ldots, a_{nn})$ die Diagonale von A bezeichnet.

b) Man berechne die Cholesky-Zerlegung $A = L\,L^{\mathrm{T}}$ der Matrix A und approximiere L durch Weglassen kleiner Elemente. Mit dieser unteren Dreiecksmatrix $\tilde{L}$ setze man $B = \tilde{L}^{-\mathrm{T}}\tilde{L}^{-1}$. Diese Methode ist vorwiegend für dünn besetzte, diagonaldominante Matrizen A anwendbar.

CG-Verfahren für beliebige Zielfunktionen.

Für beliebige nichtquadratische Zielfunktionen $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$ lässt sich der CG-Algorithmus einfach formal übertragen. Die Wahl der Schrittweite t_k wird dabei nach Fletcher und Reeves (1964) ersetzt durch eine Schrittweitenstrategie, die den strengen Wolfe-Powell Bedingungen genügt. Damit ergibt sich der folgende Algorithmus für das so genannte Fletcher-Reeves-Verfahren.

Algorithmus (10.20) (Fletcher–Reeves–Verfahren)

1.) Startdaten: $x^1 \in \mathbb{R}^n$, $g^1 := \nabla f(x^1)$, $d^1 := -g^1$, $0 < \sigma < \rho < 0.5$, $k := 1$, TOL > 0;

2.) Falls $\|g^k\| \leq$ TOL oder $k > k_{\max}$: Stop;

3.) Bestimme Schrittweite $t_k > 0$ mit
$f(x^k + t_k\, d^k) \;\leq\; f(x^k) + \sigma\, t_k\, (g^k)^{\mathrm{T}} d^k, \quad |\nabla f(x^k + t_k d^k)^{\mathrm{T}} d^k| \;\leq\; -\rho\, (g^k)^{\mathrm{T}} d^k$;

4.) $x^{k+1} := x^k + t_k d^k, \quad g^{k+1} := \nabla f(x^{k+1})$,
$s_k := \|g^{k+1}\|^2 / \|g^k\|^2, \quad d^{k+1} := -g^{k+1} + s_k\, d^k$;

5.) $k := k + 1$; Gehe zu 2.);

Wir zitieren ohne Beweis den folgenden Konvergenzsatz aus Geiger, Kanzow:

Satz (10.21) (Konvergenz)

Ist $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$ nach unten beschränkt, $x^1 \in \mathbb{R}^n$ und ∇f Lipschitz-stetig auf $L_f(x^1)$, sowie $0 < \sigma < \rho < 0.5$, so gelten
a) Das Fletcher–Reeves–Verfahren (10.20) mit TOL $= 0$ ist wohldefiniert, die d^k sind Abstiegsrichtungen.
b) Für die gemäß (10.20) erzeugte Folge (x^k) gilt $\liminf\limits_{k\to\infty} \|\nabla f(x^k)\| \;=\; 0$.

Bemerkung (10.22)

Es gibt weitere Relationen zur Berechnung der Schrittweiten s_k, die sämtlich für quadratische Zielfunktionen äquivalent sind, jedoch für nichtquadratische Zielfunktionen zu verschiedenen Optimierungsverfahren führen:

$$
\begin{aligned}
s_k &= -\frac{(g^{k+1})^{\mathrm{T}} g^{k+1}}{(g^k)^{\mathrm{T}} d^k} && \text{(Myers)} \\
s_k &= \frac{(g^{k+1} - g^k)^{\mathrm{T}} g^{k+1}}{(g^k)^{\mathrm{T}} g^k} && \text{(Polak, Ribiere)} \\
s_k &= \frac{(g^{k+1} - g^k)^{\mathrm{T}} g^{k+1}}{(g^{k+1} - g^k)^{\mathrm{T}} d^k} && \text{(Hestenes, Stiefel)}
\end{aligned}
$$

Aufgabe (10.23)

Wenden Sie das CG–Verfahren (10.17) auf die Minimierung der folgenden Funktion an (Handrechnung)

$$n \;=\; 2, \quad f(x) \;=\; x_1^2 \;+\; 10 x_2^2$$

Startvektor : $x^1 \;=\; (1, 0.1)^{\mathrm{T}}$. Überzeugen Sie sich, dass das Verfahren in nur zwei Iterationen das Minimum von f liefert.

Wieviele Iterationen hätte das Gradientenverfahren (mit exakter Schrittweite) benötigt, um gemäß (6.9) eine Genauigkeit $f(x^{k+1}) \leq 10^{-5}$ zu garantieren?

11. Das Trust–Region Verfahren

A. Das Verfahren und seine Konvergenz.

Wir kommen zurück auf das *Prinzip der lokalen Approximation*, vgl. Abschnitt 8, ersetzen f wieder durch ein quadratisches Modell

$$q(z) \;:=\; f(x) \,+\, \nabla f(x)^{\mathrm{T}}(z-x) \,+\, \frac{1}{2}\,(z-x)^{\mathrm{T}}\,\nabla^2 f(x)\,(z-x), \tag{11.1}$$

minimieren nun aber $q(z)$ nicht über ganz $\mathbb{R}^n$ - dies würde ja auch die positive Definitheit von $\nabla^2 f(x)$ voraussetzen, - sondern über einen *Vertrauensbereich (trust region)*

$$\Omega(\Delta) \;:=\; \{z \in \mathbb{R}^n :\; \|z-x\|_2 \,\leq\, \Delta\} \;=\; \overline{K}_\Delta(x), \quad \Delta > 0. \tag{11.2}$$

Hierdurch wird berücksichtigt, dass $q(z)$ i. Allg. nur in der Nähe des Entwicklungspunktes x eine gute Approximation von f darstellt.

Das Grundprinzip der *Trust–Region–Verfahren* besteht nun darin, die Approximationsgüte von q zu testen und damit den Radius Δ des Vertrauensbereiches adaptiv anzupassen. Dies leistet der folgende Modellalgorithmus.

Modellalgorithmus (11.3)

Start: $x^0 \in \mathbb{R}^n, \quad \Delta_0 > 0, \quad \mathrm{TOL} \geq 0;$

Für $k \,=\, 0,\, 1,\, \ldots$

$g^k \;:=\; \nabla f(x^k); \quad H_k \;:=\; \nabla^2 f(x^k);$

Falls $\|g^k\| \,\leq\, \mathrm{TOL}:$ Stopp!

Bestimme $d^k \in \mathbb{R}^n$ als Lösung des *quadratischen Hilfsproblems*

Minimiere $q_k(d) \;:=\; f(x^k) \,+\, (g^k)^{\mathrm{T}} d \,+\, \frac{1}{2}\, d^{\mathrm{T}}\, H_k\, d,$

unter der Nebenbedingung: $\|d\| \leq \Delta_k;$

$$r_k \;:=\; \frac{f(x^k) - f(x^k + d^k)}{q_k(0) - q_k(d^k)};$$

$$\Delta_{k+1} \;:=\; \begin{cases} \Delta_k/4, & \text{falls} \quad r_k < 1/4, \\ 2\,\Delta_k, & \text{falls} \quad r_k > 3/4 \;\text{ und }\; \|d^k\| = \Delta_k, \\ \Delta_k, & \text{sonst}; \end{cases}$$

Falls $r_k \geq 1/4:$ $\quad x^{k+1} \;:=\; x^k \,+\, d^k; \quad$ sonst $\quad x^{k+1} \;:=\; x^k;$

end k.

Bemerkung (11.4)

Der Nenner $\Delta q_k(d^k) := q_k(0) - q_k(d^k)$ bei der Berechnung von r_k ist stets positiv, der Algorithmus (11.3) ist also wohldefiniert.

Wäre nämlich $\Delta q_k = 0$, so wäre $d = 0$ ein inneres lokales Minimum von q_k, also

$$\nabla q_k(0) \;=\; g^k \,+\, H_k\, d \;=\; g^k \;=\; 0,$$

d.h. in der Abfrage zuvor wäre ein Abbruch erfolgt. □

Bevor wir auf die Lösung des quadratischen Hilfsproblems eingehen, geben wir einen Satz über die Konvergenz des obigen Trust–Region–Algorithmus (11.3) an. Die Beweisideen gehen auf Sorensen (1982) und Moré (1983) zurück. Der Satz ist dem Buch von Werner (1992) entnommen. Varianten findet man in Fletcher (1987) und Geiger, Kanzow (1999).

Satz (11.5) (Konvergenz)

Sei $f \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, $x^0 \in \mathbb{R}^n$ mit $L_f(x^0)$ kompakt. Das Verfahren (11.3) ist dann wohldefiniert und breche für TOL $= 0$ nicht ab. Für die erzeugte Folge x^k gelten dann

a) $\lim\limits_{k\to\infty} g^k \;=\; 0$, d.h. jeder Häufungspunkt von x^k ist stationär.

b) Es gibt einen Häufungspunkt x^* von x^k mit $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv semidefinit.

c) Ist $\nabla^2 f(x^*)$ sogar positiv definit, so konvergiert die gesamte Folge, $x^k \to x^*$. Es gilt $r_k \to 1$. Damit existiert ein $k_0 \in \mathbb{N}$, so dass alle Iterationen $k \geq k_0$ erfolgreich sind. Ist $\nabla^2 f$ lokal Lipschitz-stetig, so konvergiert x^k *quadratisch* gegen x^*.

Beweis – nur a):

(i) Wir zeigen zunächst die Abschätzung

$$\Delta q_k(d^k) \;:=\; q_k(0) - q_k(d^k) \;\geq\; \frac{1}{2}\,\|g^k\|\,\min\left(\Delta_k, \frac{\|g^k\|}{\|H_k\|}\right).$$

Ist $\Delta_k\,\|H_k\| \leq \|g^k\|$, so folgt mit dem zulässigen Punkt $d := -\Delta_k\, g^k/\|g^k\|$

$$\begin{aligned}
\Delta q_k(d^k) \;&\geq\; \Delta q_k(d) \\
&=\; -(g^k)^{\mathrm{T}} d \;-\; \frac{1}{2}\, d^{\mathrm{T}} H_k\, d \\
&=\; \Delta_k\,\|g^k\| \;-\; \frac{\Delta_k^2}{2}\,\frac{(g^k)^{\mathrm{T}} H_k\, g^k}{\|g^k\|^2} \\
&\geq\; \Delta_k\|g^k\| \;-\; \frac{\Delta_k^2}{2}\,\|H_k\| \;\geq\; \frac{1}{2}\,\Delta_k\|g^k\|.
\end{aligned}$$

Ist dagegen $\Delta_k \|H_k\| > \|g^k\|$, so ist $d := -g^k/\|H_k\|$ zulässig und es folgt

$$\begin{aligned}
\Delta q_k(d^k) &\geq \Delta q_k(d) \\
&= -(g^k)^{\mathrm{T}} d \; - \; \frac{1}{2}\, d^{\mathrm{T}} H_k\, d \\
&= \frac{\|g^k\|^2}{\|H_k\|} \; - \; \frac{1}{2}\, \frac{(g^k)^{\mathrm{T}} H_k\, g^k}{\|H^k\|^2} \\
&\geq \frac{\|g^k\|^2}{\|H_k\|} \; - \; \frac{1}{2}\, \frac{\|H_k\|\, \|g^k\|^2}{\|H^k\|^2} \;=\; \frac{1}{2}\, \frac{\|g^k\|^2}{\|H_k\|}.
\end{aligned}$$

(ii) Behauptung: $\liminf\limits_{k\to\infty} \|g^k\| \;=\; 0.$

Würde dies nicht gelten, so gäbe es $\varepsilon > 0$ und $K_0 \in \mathbb{N}$ mit $\forall k \geq K_0 : \; \|g^k\| > \varepsilon.$

Wir zeigen die Hilfsaussage: $$\sum_{k=0}^{\infty} \Delta_k \;<\; \infty. \qquad (*)$$

Gibt es nur endlich viele erfolgreiche Iterationsschritte, so folgt für hinreichend große $k \geq K_1$: $\Delta_{k+1} = \Delta_k/4$ und damit $\sum_{k\geq 0} \Delta_k < \infty$ (geometrische Reihe).

Gibt es andererseits eine Teilfolge $k = k_j,\; j \in \mathbb{N}_0$, aller erfolgreichen Iterationsschritte, so folgt mit (i) für $k = k_j \geq K_0$:

$$\begin{aligned}
f(x^k) \;-\; f(x^{k+1}) &\geq \frac{1}{4}\, \Delta q_k \\
&\geq \frac{1}{8}\, \|g^k\| \, \min\left(\Delta_k, \frac{\|g^k\|}{\|H_k\|}\right) \\
&\geq \frac{\varepsilon}{8} \, \min\left(\Delta_k, \frac{\varepsilon}{\|H_k\|}\right).
\end{aligned}$$

Damit ergibt sich aus der Monotonie der $f(x^k)$:

$$\begin{aligned}
\frac{\varepsilon}{8} \sum_{j=j_0}^{\infty} \min\left(\Delta_{k_j}, \frac{\varepsilon}{\|H_{k_j}\|}\right) &\leq \sum_{j=j_0}^{\infty} \left(f(x^{k_j}) - f(x^{k_j+1})\right) \\
&\leq \sum_{j=j_0}^{\infty} \left(f(x^{k_j}) - f(x^{k_{j+1}})\right) \;<\; \infty,
\end{aligned}$$

Die Konvergenz der Reihe folgt aus der Beschränktheit von f auf $L_f(x^0)$. Da $\nabla^2 f$ ebenfalls auf $L_f(x^0)$ beschränkt ist, $\|H_k\| \leq C$, folgt somit

$$\sum_{j=j_0}^{\infty} \min\left(\Delta_{k_j}, \frac{\varepsilon}{C}\right) \;<\; \infty.$$

Damit ist für hinreichend große j notwendigerweise $\Delta_{k_j} < \varepsilon/C$ und somit auch $\sum\limits_{j=0}^{\infty} \Delta_{k_j} \;<\; \infty.$

Für die nicht erfolgreichen Iterationsschritte zwischen k_j und k_{j+1} gilt $\Delta_{i+1} = \Delta_i/4$, $i = k_j + 1, \ldots, k_{j+1} - 1$. Damit folgt mittels geometrischer Reihe

$$\begin{aligned} \sum_{i=k_j+1}^{k_{j+1}-1} \Delta_i &\leq \frac{1/4}{1-1/4}\,\Delta_{k_j} = \frac{1}{3}\,\Delta_{k_j} \\ \Rightarrow \quad \sum_{k=0}^{\infty} \Delta_k &\leq (1+1/3) \sum_{j=0}^{\infty} \Delta_{k_j} < \infty. \end{aligned}$$

Damit ist die Hilfsaussage $(*)$ gezeigt!

Weiter schließen wir hiermit

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| \leq \sum_{k=0}^{\infty} \Delta_k < \infty.$$

Somit ist (x^k) eine Cauchy-Folge, also auch konvergent $x^k \to x^*$ $(k \to \infty)$.

Für hinreichend große Indizes $k \geq K_2$ liefert die Abschätzung (i) und der Mittelwertsatz $(0 < \Theta < 1)$

$$\begin{aligned} |r_k - 1| &= \Big|\frac{f(x^k) - f(x^k + d^k)}{f(x^k) - q_k(d^k)} - 1\Big| \\ &= \frac{|f(x^k + d^k) - q_k(d^k)|}{|\Delta q_k(d^k)|} \\ &\leq \frac{|(f(x^k + d^k) - f(x^k)) - (g^k)^{\mathrm{T}} d^k - (1/2)\,(d^k)^{\mathrm{T}} H_k d^k|}{(1/2)\|g^k\|\Delta_k} \\ &= \frac{|(\nabla f(x^k + \Theta d^k)^{\mathrm{T}} d^k - (g^k)^{\mathrm{T}} d^k - (1/2)\,(d^k)^{\mathrm{T}} H_k d^k|}{(1/2)\|g^k\|\Delta_k} \\ &\leq \frac{2\,\|\nabla f(x^k + \Theta d^k) - \nabla f(x^k)\| + C\,\Delta_k}{\|g^k\|} \\ &\leq \frac{1}{\varepsilon}\,\big(2\,\|\nabla f(x^k + \Theta d^k) - \nabla f(x^k)\| + C\,\Delta_k\big) \\ &\to 0 \qquad (k \to \infty). \end{aligned}$$

Damit folgt, dass $r_k \to 1$, $(k \to \infty)$ und somit für hinreichend große k aufgrund des Algorithmus $\Delta_{k+1} \geq \Delta_k$, im Widerspruch zu Hilfsaussage $(*)$.

(iii) Behauptung: $\lim\limits_{k\to\infty} g^k = 0.$

Wir führen den Beweis wieder indirekt und nehmen also an, dass es ein $\varepsilon > 0$ gibt sowie eine Teilfolge $k = k_j$ mit der Eigenschaft $\forall j : \ \|g^{k_j}\| \geq 2\,\varepsilon$.

Wegen (ii) existiert zu jeden $j \in \mathbb{N}$ ein Index $\ell_j > k_j$ mit

$$\|g^{\ell_j}\| < \varepsilon, \quad \forall\, k_j \leq \ell < \ell_j : \ \|g^{\ell}\| \geq \varepsilon. \qquad (**)$$

Ist für beliebiges j ein Iterationsschritt Nummer ℓ mit $k_j \le \ell < \ell_j$ erfolgreich, so hat man mit (i) und $(**)$ die Abschätzung

$$\begin{aligned} f(x^\ell) - f(x^{\ell+1}) &\ge \frac{1}{4}\,\Delta q_\ell \\ &\ge \frac{1}{8}\,\|g^\ell\|\,\min\{\Delta_\ell,\,\frac{\|g^\ell\|}{\|H_\ell\|}\} \\ &\ge \frac{\varepsilon}{8}\,\min\{\|x^{\ell+1}-x^\ell\|,\,\frac{\varepsilon}{C}\}. \end{aligned}$$

Da die Folge $(f(x^k))$ notwendigerweise konvergiert (monoton fallend und beschränkt), folgt für hinreichend große Indizes j

$$\frac{\varepsilon}{8}\,\|x^{\ell+1}-x^\ell\| \;\le\; f(x^\ell) - f(x^{\ell+1})\,.$$

Dies gilt erst recht für alle nicht erfolgreichen Iterationsschritte, da dann ja $x^{\ell+1} = x^\ell$ ist. Folglich haben wir

$$\begin{aligned} \frac{\varepsilon}{8}\,\|x^{\ell_j} - x^{k_j}\| &\le \frac{\varepsilon}{8}\sum_{\ell=k_j}^{\ell_j-1}\|x^{\ell+1}-x^\ell\| \\ &\le \sum_{\ell=k_j}^{\ell_j-1}\big(f(x^\ell)-f(x^{\ell+1})\big) \\ &= f(x^{k_j}) - f(x^{\ell_j}) \;\to\; 0 \quad (j\to\infty). \end{aligned}$$

Damit konvergiert $\|x^{\ell_j} - x^{k_j}\| \to 0, \quad j\to\infty$. Wegen der gleichmäßigen Stetigkeit von ∇f auf $L_f(x^0)$ (kompakt!) folgt hieraus aber auch $\|g^{\ell_j} - g^{k_j}\| \to 0, \quad j\to\infty$.

Andererseits folgt nach unserer Konstruktion:

$$\|g^{\ell_j} - g^{k_j}\| \;\ge\; \|g^{k_j}\| - \|g^{\ell_j}\| \;\ge\; 2\,\varepsilon - \varepsilon \;=\; \varepsilon \;>\; 0.$$

Dies ist ein Widerspruch, womit schließlich die Aussage (a) des Satzes gezeigt ist. □

Anmerkungen (11.6)

Es gibt etliche Varianten sowohl des obigen Konvergenzsatzes wie auch des Trust-Region-Verfahrens selbst.

a) Ersetzt man die (exakte) Hesse-Matrix H_k durch irgendeine beschränkte Folge symmetrischer Matrizen, so lässt sich immer noch $g^k \to 0$ zeigen.

b) Es kann sinnvoll sein, die quadratischen Hilfsprobleme nur näherungsweise zu lösen.

c) Man kann die Trust-Region-Kugeln mittels regulärer Matrizen D_k skalieren: $\|D_k\,d\| \le \Delta_k$.

d) Die im Algorithmus (11.3) gewählten Parameter $(1/4, 3/4, 2)$ können modifiziert werden.

B. Das quadratisches Hilfsproblem.

Die Brauchbarkeit des Trust-Region-Ansatzes hängt wesentlich davon ab, ob es gelingt, das *quadratische Hilfsproblem* effizient zu lösen.

Hilfsproblem (11.7)

Bestimme zu vorgegebenen $f \in \mathbb{R}$, $g \in \mathbb{R}^n$, $H \in \mathbb{R}^{(n,n)}$ (symmetrisch) und $\Delta > 0$ ein globales Minimum d^* der Funktion

$$q(d) \;:=\; f \,+\, g^{\mathrm{T}} d \,+\, \frac{1}{2}\, d^{\mathrm{T}} H\, d$$

unter der Nebenbedingung $\|d\|_2 \,\leq\, \Delta$.

Satz (11.8)

$d^* \in \mathbb{R}^n$ löst genau dann das quadratische Hilfsproblem (11.7), wenn es einen *Lagrange–Multiplikator* $\lambda \geq 0$ gibt mit den Eigenschaften

a) $(H \,+\, \lambda\, I_n)\, d^* \;=\; -g,$

b) $(H \,+\, \lambda\, I_n)$ ist positiv semidefinit,

c) $\|d^*\| \,\leq\, \Delta, \quad \lambda \cdot (\|d^*\|^2 - \Delta^2) \,=\, 0.$

Die Bedingung in a) heißt *notwendige Bedingung erster Ordnung*, die Bedingung in b) *notwendige Bedingung zweiter Ordnung*, die Bedingungen in c) heißen *Zulässigkeit* und *Komplementarität*.

Beweis:

$\Leftarrow$: Wir setzen $\quad q_\lambda(d) \;:=\; q(d) \,+\, \dfrac{\lambda}{2}\, d^{\mathrm{T}} d \;=\; f \,+\, g^{\mathrm{T}} d \,+\, \dfrac{1}{2}\, d^{\mathrm{T}} (H + \lambda I_n)\, d.$

Direktes Ausrechnen unter Verwendung von a) ergibt

$$q_\lambda(d) - q_\lambda(d^*) \;=\; \frac{1}{2}\, (d - d^*)^{\mathrm{T}} (H + \lambda I_n)\, (d - d^*).$$

Damit folgt aus b) $\;\forall\, d \in \mathbb{R}^n :\; q_\lambda(d) \geq q_\lambda(d^*).\;$ Einsetzen von q_λ ergibt

$$\forall\, d \in \mathbb{R}^n : \qquad q(d) \;\geq\; q(d^*) \,+\, \frac{\lambda}{2} \left((d^*)^{\mathrm{T}} d^* \,-\, d^{\mathrm{T}} d \right)$$

Da ferner $\lambda \geq 0$ ist folgt weiter

$$\forall\, d \in \mathbb{R}^n : \qquad \|d\| \,\leq\, \Delta \;\Rightarrow\; q(d) \,\geq\, q(d^*) \,+\, \frac{\lambda}{2} \left(\|d^*\|^2 - \Delta^2 \right).$$

Aufgrund der Voraussetzung c) verschwindet schließlich der zweite Summand. Damit ist d^* Lösung des quadratischen Hilfsproblems.

$\Rightarrow$: Gilt $\|d^*\| < \Delta$, so ist d^* zugleich ein lokales Minimum von q und die Bedingungen a) - c) folgen mit $\lambda = 0$ aus den notwendigen Bedingungen für unrestringierte Optimierungsaufgaben.

Im Fall $\|d^*\| = \Delta$ ist die Nebenbedingung

$$c(d) \;:=\; \frac{1}{2}\left(d^{\mathrm{T}}d \;-\; \Delta^2\right)$$

aktiv und d^* ist ein *regulärer Punkt* von c, d.h. $\nabla c(d^*) = d^* \neq 0$.

Aufgrund der (erst später zu zeigenden) notwendigen Bedingungen für ungleichungsrestringierte Optimierungaufgaben (Abschnitt 13) existiert ein *Lagrange-Multiplikator* $\lambda \geq 0$, so dass für die *Lagrange-Funktion*

$$L(d) \;:=\; q(d) \;+\; \lambda\, c(d) \;=\; f \;+\; g^{\mathrm{T}}d \;+\; \frac{1}{2}\, d^{\mathrm{T}}\big(H + \lambda I_2\big)d \;-\; \frac{\lambda}{2}\,\Delta^2$$

die folgenden notwendigen Bedingungen gelten:

(i) $\nabla L(d^*) \;=\; 0,$

(ii) $\forall\, y \in \mathbb{R}^n :\; \nabla c(d^*)^{\mathrm{T}} y \;=\; 0 \;\Rightarrow\; y^{\mathrm{T}} \nabla^2 L(d^*)\, y \;\geq\; 0.$

Aus (i) folgt unmittelbar die Bedingung a):

$$\nabla L(d^*) \;=\; \nabla q(d^*) + \lambda\, \nabla c(d^*) \;=\; (g + H\, d^*) + \lambda\, d^* \;=\; 0.$$

Aus (ii) folgt: $\forall\, y \in \mathbb{R}^n :\; (d^*)^{\mathrm{T}} y \;=\; 0 \;\Rightarrow\; y^{\mathrm{T}}(H + \lambda\, I_n)\, y \;\geq\; 0.$

Somit bleibt zu zeigen, dass hierbei auf die Prämisse $y \perp d^*$ verzichtet werden kann. Sei also $y \in \mathbb{R}^n$ mit $(d^*)^{\mathrm{T}} y \neq 0$. Wir setzen

$$\begin{aligned} d \;&:=\; d^* \;+\; \alpha\, y, \qquad\qquad \alpha \;:=\; -2\,\frac{y^{\mathrm{T}} d^*}{\|y\|^2} \;\neq\; 0, \\ &=\; \Big[I_n - 2\,\Big(\frac{y}{\|y\|}\Big)\,\Big(\frac{y}{\|y\|}\Big)^{\mathrm{T}}\Big]\, d^*\,. \end{aligned}$$

d entsteht also aus d^* durch Spiegelung an der Hyperebene $y^{\perp}$. Insbesondere folgt $\|d^*\| = \|d\|$.

Wie in ersten Teil des Beweises schließt man nun

$$\begin{aligned} 0 \;&\leq\; q(d) \;-\; q(d^*) \\ &=\; q_\lambda(d) \;-\; q_\lambda(d^*) \;+\; \frac{\lambda}{2}\big(\|d^*\|^2 \;-\; \|d\|^2\big) \\ &=\; \frac{1}{2}\,(d - d^*)^{\mathrm{T}}\big(H + \lambda I_n\big)(d - d^*) \\ &=\; \frac{\alpha^2}{2}\; y^{\mathrm{T}}(H + \lambda I_n)\, y. \end{aligned}$$

Damit ist gezeigt, dass $H + \lambda I_n$ positiv semidefinit ist. $\square$

Bemerkung (11.9)

Man kann zeigen, dass der Lagrange-Multiplikator λ in (11.8) eindeutig bestimmt ist. Dies folgt aus der Regularität von d^*.

C. Numerische Realisierung.

Hierzu gibt es zwei prinzipiell unterschiedliche Ansätze:

I. Anstelle des Trust-Region-Radius Δ wird der Lagrange-Multiplikator λ gesteuert. Dies lässt sich dadurch rechtfertigen, dass die Abbildung $\lambda \mapsto \|d^*\| = \Delta$ monoton fällt. Nach Fletcher lässt sich diese Idee etwa wie folgt realisieren:

Algorithmus (11.10)

Start: $x \in \mathbb{R}^n, \quad \lambda > 0, \quad \text{TOL} \geq 0;$

Für $k = 0, 1, \ldots$

$f := f(x); \quad g := \nabla f(x); \quad H := \nabla^2 f(x);$

Falls $\|g\| \leq \text{TOL}:$ Stop!

Solange $(H + \lambda I_n)$ nicht positiv definit : $\lambda = 4\,\lambda;$

Bestimme $d \in \mathbb{R}^n$ aus $(H + \lambda I_n)\, d = -g;$

$r := \dfrac{f - f(x+d)}{f - q(d)};$ wobei $q(d) := f + g^{\mathrm{T}} d + \dfrac{1}{2}\, d^{\mathrm{T}} H\, d;$

$$\lambda := \begin{cases} 4\,\lambda, & \text{falls} \quad r < 1/4, \\ \lambda/2, & \text{falls} \quad r > 3/4, \\ \lambda, & \text{sonst;} \end{cases}$$

Falls $r > 0:$ $x := x + d;$

end k.

Bemerkungen (11.11)

Der Test auf Positive Definitheit kann wie früher mittels Cholesky-Zerlegung durchgeführt werden.

In der Konvergenzphase des Algorithmus muss ferner auf $\lambda = 0$ umgeschaltet werden um die quadratische Konvergenz des Verfahrens zu sichern.

Alternativ könnte man eine verallgemeinerte Schrittweitenbestimmung für den so genannten *Levenberg-Marquardt-Pfad* $\lambda \mapsto f(x + d(\lambda)),\ \lambda \geq 0,$ mit

$$(H + \lambda I_n)\, d(\lambda) = -g \tag{11.12}$$

durchführen.

II. Wir bleiben beim Modellalgorithmus (11.3) und der Steuerung über den Radius Δ des Vertrauensbereichs. Nach Satz (11.8) ist $\lambda \geq 0$ so zu bestimmen, dass die symmetrische Matrix $(H + \lambda I_n)$ positiv (semi-)definit ist und die nichtlineare Gleichung

$$F(\lambda) \;:=\; \frac{1}{\Delta} \,-\, \frac{1}{\|d(\lambda)\|} \;=\; 0 \tag{11.13}$$

erfüllt ist. Dabei beschreibt $d(\lambda)$ die Levenberg-Marquardt-Trajektorie gemäß (11.12).

Es ist sinnvoll, die Bestimmung von λ über (11.13) vorzunehmen, da $\|d(\lambda)\|$ in $\lambda = -\lambda_i$, mit $\lambda_1 \leq \ldots \leq \lambda_n$ Eigenwerte von H, i. Allg. Singularitäten besitzt, also $\|d(\lambda)\|^{-1}$ Nullstellen. Zur Berechnung von λ lässt sich das Newton-Verfahren verwenden. Im Einzelnen geht man dazu folgendermaßen vor:

1.) Löse das lineare Gleichungssystem $(H + \lambda I_n)\, d(\lambda) \;=\; -g$ mittels Cholesky-Zerlegung der Koeffizientenmatrix: $H + \lambda I_n \;=:\; L\, D\, L^{\mathrm{T}}$, L: normierte untere Dreiecksmatrix, $D = \mathrm{diag}(\delta_1, \ldots, \delta_n)$.

2.) Durch Differentiation dieses linearen Gleichungssystems nach dem Parameter λ erhält man ein lineares Gleichungssystem für die Ableitung $d'(\lambda)$:

$$(H + \lambda I_n)\, d'(\lambda) \;=\; -d(\lambda).$$

3.) Setzt man nun $w(\lambda) := \|d(\lambda)\|$, so folgt für die Ableitung

$$w'(\lambda) \;=\; \frac{d(\lambda)^{\mathrm{T}} d'(\lambda)}{w(\lambda)} \;=\; -\,\frac{d(\lambda)^{\mathrm{T}} (H + \lambda I_n)^{-1} d(\lambda)}{w(\lambda)}.$$

und mit der obigen Cholesky-Zerlegung

$$d^{\mathrm{T}} (H + \lambda I_n)^{-1} d \;=\; d^{\mathrm{T}} L^{-\mathrm{T}} D^{-1} L^{-1} d \;=\; z^{\mathrm{T}} D^{-1} z \;=\; \sum_{i=1}^{n} \frac{z_i^2}{\delta_i}\,,$$

wobei z aus dem gestaffelten linearen Gleichungssystem $L\, z \;=\; d$ berechnet werden kann (Vorwärts-Rekursion).

4.) Das Newton-Verfahren für das nichtlineare Gleichungssystem $F(\lambda) = 0$ ergibt

$$\begin{aligned} & F(\lambda) \;=\; \frac{1}{\Delta} \,-\, \frac{1}{w(\lambda)}\,, \quad F'(\lambda) \;=\; \frac{w'(\lambda)}{w(\lambda)^2} \\ \Rightarrow \qquad & \lambda^+ \;=\; \lambda \,-\, \frac{F(\lambda)}{F'(\lambda)} \;=\; \lambda \,-\, \frac{w(\lambda)^2}{w'(\lambda)} \left(\frac{1}{\Delta} \,-\, \frac{1}{w(\lambda)} \right) \\ \Rightarrow \qquad & \lambda^+ \;=\; \lambda \,+\, \frac{w(\lambda)}{w'(\lambda)} \left(1 \,-\, \frac{w(\lambda)}{\Delta} \right) . \end{aligned}$$

Insgesamt ergibt sich somit der folgende Algorithmus zur Lösung des quadratischen Hilfsproblems (11.7).

Algorithmus (11.14) (nach Sorensen, Moré, 1983)

Start: $x \in \mathbb{R}^n, \quad \Delta > 0, \quad g := \nabla f(x); \quad H := \nabla^2 f(x);$

(1) $H = L\,D\,L^{\mathrm{T}}, \quad D = \mathrm{diag}(\delta_1, \ldots, \delta_n);$ (Cholesky-Zerlegung)

Falls $\forall\, i: \ \delta_i > 0$: Löse $L\,D\,L^{\mathrm{T}} d = -g,$

Falls $\|d\| \le \Delta$: $\lambda := 0,$ Stop!;

Wähle $\lambda > 0$;

(2) $H + \lambda\, I_n = L\,D\,L^{\mathrm{T}}, \quad D = \mathrm{diag}(\delta_1, \ldots, \delta_n);$ (Cholesky-Zerlegung)

Falls $\exists\, i: \ \delta_i \le 0$: $\lambda = 2\,\lambda,$ Gehe zu (2);

(3) Löse $(LDL^{\mathrm{T}})\, d = -g; \quad w := \|d\|;$ Löse $L\, z = d;$

$$w' = -\Big(\sum_{i=1}^{n} \frac{z_i^2}{\delta_i}\Big)/w; \quad \lambda := \lambda + \frac{w}{w'}\,\Big(1 - \frac{w}{\Delta}\Big);$$

Abbruch, falls λ hinreichend genau; Gehe zu (2).

Bemerkungen (11.15)

a) Im obigen Algorithmus sind zusätzliche technischen Maßnahmen nötig, um sicher zu stellen, dass $H + \lambda I_n$ numerisch hinreichend positiv definit ist.

b) Für die Anwendung in einem Trust-Region-Verfahren genügt es i. Allg., das quadratische Hilfsproblem mit nur geringer Genauigkeit zu lösen. Zumeist werden lediglich wenige Newton-Iterationen durchgeführt.

c) Es gibt eine Reihe weiterer Ansätze, eine Lösung des quadratischen Hilfsproblems mittels einfacher numerischer Ansätze zu approximieren. Man spricht dann von *inexakten Trust-Region-Verfahren*. Stichworte hierzu sind

- Minimierung in Gradientenrichtung,
- Minimierung auf Teilräumen,
- CG-Ansätze zur Minimierung von $q(d)$,
- Kopplung mit Quasi-Newton Verfahren.

12. Gleichungsrestringierte Optimierungsaufgaben

In diesem und im nächsten Abschnitt untersuchen wir notwendige und hinreichende Bedingungen für allgemeine, restringierte Optimierungaufgaben in Standardform

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n \\ &\text{unter den Nebenbedingungen} \\ &g_i(x) \;\le\; 0, \quad i \in I \;:=\; \{1,\ldots,m\}, \\ &h_j(x) \;=\; 0, \quad j \in E \;:=\; \{1,\ldots,p\}. \end{aligned} \tag{12.1}$$

Wir setzen dabei wieder voraus, dass die beteiligten Funktionen f, g_i und h_j auf ganz $\mathbb{R}^n$ definiert und hinreichend oft stetig differenzierbar, zumindest aber C^1–Funktionen sind. Ferner setzen wir voraus, dass der zulässige Bereich

$$X \;:=\; \{x \in \mathbb{R}^n : \; g(x) \le 0 \;\wedge h(x) = 0\} \tag{12.2}$$

nichtleer ist, wobei wir zusammenfassen $g := (g_1,\ldots,g_m)^{\mathrm{T}}$ und $h := (h_1,\ldots,h_p)^{\mathrm{T}}$.

Wir beginnen in diesem Abschnitt mit der Untersuchung gleichungsrestringierter Optimierungsaufgaben, d.h. in (12.1) gelte $I = \emptyset$ und $E \neq \emptyset$.

Definition (12.3)

Ein zulässiger Punkt $x \in X$ heißt ein **regulärer Punkt** von X, falls die Gradienten $\nabla h_1(x),\ldots,\nabla h_p(x)$ linear unabhängig sind.

Beispiel (12.4)

Betrachtet man im $\mathbb{R}^3$ die beiden Gleichungsnebenbedingungen $h_1(x) := x_1^2 + x_2^2 + x_3^2 - 1 = 0$ und $h_2(x) := x_3 - 0.5 = 0$, so ist die Menge X der zulässigen Punkte nicht leer und in jeden Punkt $x \in X$ sind die Gradienten $\nabla h_1 = 2\,x$ und $\nabla h_2 = (0,0,1)^{\mathrm{T}}$ linear unabhängig. In diesem Beispiel sind also alle zulässigen Punkte regulär.

Ändert man dagegen die zweite Gleichungsnebenbedingung ab zu $h_2(x) := x_3 - 1 = 0$, so gibt es nur noch einen zulässigen Punkt, nämlich $x = (0,0,1)^{\mathrm{T}}$ und in diesem Punkt sind die beiden Gradienten ∇h_1 und ∇h_2 linear abhängig.

Bemerkung (12.5)

Sind alle Punkte der zulässigen Menge X regulär, so bildet X eine $(n-p)$–dimensionale *Untermannigfaltigkeit* des $\mathbb{R}^n$, vgl. z.B. Königsberger: Analysis II, Abschnitt 3.5.

Sei nun $x^0 \in X$ ein regulärer Punkt. Wir betrachten C^1-Wege $x :]-\varepsilon, \varepsilon[\to X,\ \varepsilon > 0$, mit der Eigenschaft $x(0) = x^0$. Durchläuft x alle C^1-Wege dieser Art, so durchlaufen die Geschwindigkeitsvektoren $x'(0)$ offensichtlich alle *Tangentialvektoren* an X im Punkt x^0.

Durch Differentiation der Identitäten $h_j(x(t)) = 0,\ j = 1, \ldots, p,$ finden wir

$$\forall\, j \in E :\ \forall\, t \in]-\varepsilon, \varepsilon[: \quad \nabla h_j(x(t))^{\mathrm{T}}\, x'(t) \ = \ 0$$

und speziell für $t = 0$:

$$\forall\, j \in E : \quad \nabla h_j(x^0)^{\mathrm{T}}\, x'(0) \ = \ 0.$$

Damit gelangen wir zu der folgenden Definition:

Definition (12.6)

Ist $x^0 \in X$ ein zulässiger Punkt, so heißt

$$T_X(x^0) \ := \ \{y \in \mathbb{R}^n :\ \forall\, j \in E :\ \nabla h_j(x^0)^{\mathrm{T}} y \ = \ 0\} \ = \ \operatorname{Kern} h'(x^0)$$

der *Tangentialraum* von X in x^0.

Ist x^0 regulär, so ist $T_X(x^0)$ als Kern der Jacobi-Matrix $h'(x^0)$ ein $(n-p)$-dimensionaler linearer Teilraum von $\mathbb{R}^n$.

Satz (12.7)

Ist $x^0 \in X$ ein regulärer Punkt, so lässt sich umgekehrt auch zu jedem Tangentialvektor $y \in T_X(x^0)$ ein C^1-Weg $x :]-\varepsilon, \varepsilon[\to X$ finden mit $x(0) = x^0$ und $x'(0) = y$.

Beweis: (Satz über implizite Funktionen)

Setze für $t \in \mathbb{R},\ v \in \mathbb{R}^p$:

$$H(t, v) \ := \ h(x^0 + ty + h'(x^0)^{\mathrm{T}}\, v) \ \in \ \mathbb{R}^p.$$

Dann gelten

$$\text{(i)} \quad H(0,0) \ = \ h(x^0) \ = \ 0,$$

$$\text{(ii)} \quad \frac{\partial H}{\partial v}(0,0) \ = \ h'(x^0)\, h'(x^0)^{\mathrm{T}} \ \in \mathbb{R}^{(p,p)} \quad \text{regulär!!}$$

Denn: $\quad h'(x^0)\, h'(x^0)^{\mathrm{T}}\, a \ = \ 0 \quad \Rightarrow \quad \|h'(x^0)^{\mathrm{T}}\, a\| \ = \ 0$

$$\Rightarrow \ h'(x^0)^{\mathrm{T}}\, a \ = \ 0 \quad \Rightarrow \quad \sum_{j=0}^{p} a_j\, \nabla h_j(x^0) \ = \ 0 \quad \Rightarrow \quad a \ = \ 0.$$

Mit (i) und (ii) sind die Voraussetzungen des Satzes über implizite Funktionen für H erfüllt und es folgt

$$\exists\, \varepsilon > 0,\ \ v \in \mathrm{C}^1(]-\varepsilon, \varepsilon[, \mathbb{R}^p): \quad v(0) = 0 \ \wedge\ \forall\, t:\ H(t, v(t)) \ =\ 0.$$

Setzt man nun $x(t) := x^0 + t\,y + h'(x^0)^{\mathrm{T}}\, v(t),\ \ |t| < \varepsilon,$ so ist x ein C^1-Weg mit

$$x(0) = x^0, \quad \forall\, |t| < \varepsilon: \quad h(x(t)) \ =\ H(t, v(t)) \ =\ 0.$$

Differentiation dieser Identität nach t ergibt

$$h'(x^0 + ty + h'(x^0)^{\mathrm{T}} v(t)) \ \cdot\ \big(y + h'(x^0)^{\mathrm{T}}\, v'(t)\big) \ \ =\ \ 0.$$

Für $t = 0$ folgt

$$h'(x^0)\, y \ +\ h'(x^0)\, h'(x^0)^{\mathrm{T}}\, v'(0) \ \ =\ \ 0.$$

Die Matrix $h'(x^0)\, h'(x^0)^{\mathrm{T}} \in \mathbb{R}^{(p,p)}$ ist regulär; ferner ist $h'(x^0)\, y = 0$, da nach Voraussetzung $y \in T_X(x^0)$.

Damit ist $v'(0) = 0$ und somit auch $x'(0) \ =\ y + h'(x^0)^{\mathrm{T}} v'(0) \ =\ y.$

□

Satz (12.8) (Notwendige Bedingungen erster Ordnung)

Ist $x^* \in X$ ein lokales Minimum von f auf X und ist x^* ein regulärer Punkt, so gibt es eindeutig bestimmte **Lagrange–Multiplikatoren** $\mu \in \mathbb{R}^p,$ so dass für die **Lagrange-Funktion**

$$L(x, \mu) \ \ :=\ \ f(x) \ +\ \mu^{\mathrm{T}} h(x)$$

gilt $\qquad \nabla_x L(x^*, \mu) \ \ =\ \ 0.$

Beweis:

Zu $y \in T_X(x^*)$ sei $x :]-\varepsilon, \varepsilon[\to X$ ein C^1–Weg mit $x(0) = x^*$ und $x'(0) = y.$

Wegen $f(x^*) \ \le\ f(x(t)), \quad \forall\, t:\ |t| < \delta \le \varepsilon,\ \delta > 0,$ folgt

$$\frac{d}{dt}\, f(x(t))|_{t=0} \ \ =\ \ \nabla f(x^*)^{\mathrm{T}}\, y \ \ =\ \ 0.$$

Damit gilt $\nabla f(x^*) \ \in T_X(x^*)^{\perp} \ =\ \mathrm{Spann}\{\nabla h_j(x^*) :\ \ j \in E\}.$

Die Eindeutigkeit der Lagrange-Multiplikatoren folgt aus der linearen Unabhängigkeit der $\nabla h_j(x^*),\ \ j \in E.$

□

Satz (12.9) (Notwendige Bedingungen zweiter Ordnung)

Sind $f,\ h_j \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$, so gilt unter den Voraussetzungen des Satzes (12.8)

$$\forall\, y \in T_X(x^*): \quad y^{\mathrm{T}}\, \nabla^2_{x\,x} L(x^*, \mu)\, y \ \geq\ 0,$$

d.h. die Hesse–Matrix der Lagrange–Funktion ist positiv semidefinit auf dem Tangentialraum $T_X(x^*)$.

Beweis:

Zu $y \in T_X(x^*)$ sei $x :]-\varepsilon, \varepsilon[\to X$ C^1–Weg in X gemäß (12.7) mit $x(0) = x^*$ und $x'(0) = y$. Aufgrund der Differenzierbarkeitsvoraussetzung ist x sogar ein C^2–Weg, vgl. den Satz über implizite Funktionen.

Nach ev. Verkleinerung von $\varepsilon > 0$ gilt $f(x^*) \leq f(x(t)),\ \forall t \in]-\varepsilon, \varepsilon[$. Damit folgt $d^2/dt^2\, f(x(t))|_{t=0} \geq 0$ und somit

$$\frac{d^2}{dt^2} f(x(t))|_{t=0} \ =\ x'(0)^{\mathrm{T}} \nabla^2 f(x^*)\, x'(0) \ +\ \nabla f(x^*)^{\mathrm{T}} x''(0) \ \geq\ 0. \qquad (*)$$

Aus $\sum_{j\in E} \mu_j\, h_j(x(t)) = 0$ folgt durch zweimaliges Differenzieren

$$x'(0)^{\mathrm{T}} \big(\sum_j \mu_j\, \nabla^2 h_j(x^*) \big)\, x'(0) \ +\ \big(\sum_j \mu_j\, \nabla h_j(x^*) \big)^{\mathrm{T}}\, x''(0) \ =\ 0. \qquad (**)$$

Addition von $(*)$ und $(**)$ und Anwendung von (12.8) ergibt

$$\begin{aligned} & x'(0)^{\mathrm{T}}\, \nabla^2_{x\,x} L(x^*, \mu)\, x'(0) \ +\ \nabla_x L(x^*, \mu)^{\mathrm{T}}\, x''(0) \ \geq\ 0 \\ \Rightarrow \qquad & y^{\mathrm{T}}\, \nabla^2_{x\,x} L(x^*, \mu)\, y \ \geq\ 0. \end{aligned}$$

□

Satz (12.10) (Hinreichende Bedingung)

Die Funktionen f und $h_j,\ j = 1, \ldots, p,$ seien zweifach stetig differenzierbar, $x^* \in X$ und $\mu \in \mathbb{R}^p$. Gelten dann für die Lagrange–Funktion

$$L(x, \mu) \ =\ f(x) \ +\ \sum_{j\in E} \mu_j\, h_j(x)$$

die Bedingungen

(i) $\quad \nabla_x L(x^*, \mu) \ =\ 0,$

(ii) $\quad \forall\, y \in T_X(x^*) \setminus \{0\}: \quad y^{\mathrm{T}}\, \nabla^2_{x\,x}\, L(x^*, \mu)\, y \ >\ 0,$

so ist x^* ein striktes lokales Minimum von f auf X.

Beweis: (indirekt)

Nehmen wir an, x^* sei kein striktes lokales Minimum. Dann gibt es eine Folge $(x^k) \in X^{\mathbb{N}}$ mit $\quad \forall\, k:\ x^k \neq x^*, \quad x^k \to x^* \ \ (k \to \infty) \quad$ und $\quad \forall\, k:\ f(x^k) \leq f(x^*).$

Durch Übergang zu einer Teilfolge kann man erreichen, dass zusätzlich gilt

$$\lim_{k\to\infty} \ \frac{x^k - x^*}{\|x^k - x^*\|} \ = \ y, \qquad \|y\| \ = \ 1.$$

Aufgrund der Differenzierbarkeit der Funktionen h_j in x^* folgt nun

$$\lim_{k\to\infty} \ \frac{h_j(x^k) \ - \ h_j(x^*) \ - \ \nabla h_j(x^*)^{\mathrm{T}}(x^k - x^*)}{\|x^k - x^*\|} \ = \ 0.$$

Da $x^k, x^* \in X,$ also $h_j(x^k) = h_j(x^*) = 0,$ folgt hieraus $\forall\, j \in E: \quad \nabla h_j(x^*)^{\mathrm{T}} y = 0,$ d.h. $y \in T_X(x^*) \setminus \{0\}$.

Der Taylorsche Satz liefert nun mit einer Zwischenstelle $\xi^k = x^* + \Theta(x^k - x^*),$ $0 < \Theta < 1,$ und der Voraussetzung (i):

$$\begin{aligned} f(x^*) \ \geq \ f(x^k) \ &= \ L(x^k, \mu) \\ &= \ L(x^*, \mu) \ + \ \nabla_x L(x^*, \mu)^{\mathrm{T}}(x^k - x^*) \ + \ \frac{1}{2}\,(x^k - x^*)^{\mathrm{T}} \nabla^2_{x\,x} L(\xi^k, \mu)\,(x^k - x^*) \\ &= \ f(x^*) \ + \ \frac{1}{2}\,(x^k - x^*)^{\mathrm{T}} \nabla^2_{x\,x} L(\xi^k, \mu)\,(x^k - x^*) \end{aligned}$$

Division durch $\|x^k - x^*\|^2$ und Grenzübergang $k \to \infty$ liefert somit

$$y^{\mathrm{T}}\, \nabla^2_{x\,x} L(x^*, \lambda)\, y \ \leq \ 0,$$

im Widerspruch zur Voraussetzung (ii)! $\square$

Bemerkungen (12.11)

a) Für die hinreichenden Bedingungen (12.10) wird die Regularität des zulässigen Punktes $x^* \in X$ nicht benötigt! Man beachte jedoch, dass nur für reguläre, zulässige Punkte die Eindeutigkeit der Lagrange-Multiplikatoren gesichert ist, vgl. (12.8).

b) Ist $(z^1, \ldots, z^{(n-p)})$ eine Basis des Tangentialraumes $T_X(x^*),$ so sind die notwendigen bzw. hinreichenden Bedingungen zweiter Ordnung äquivalent zu

$$Z^{\mathrm{T}}\, \nabla^2_{x\,x} L(x^*, \lambda)\, Z \ \in \ \mathbb{R}^{(n-p,n-p)} \quad \text{pos. semidefinit bzw. pos. definit.} \qquad (12.12)$$

Dabei ist $Z := (z^1, \ldots, z^{(n-p)}) \in \mathbb{R}^{(n,n-p)}.$ Die obige Matrix $Z^{\mathrm{T}}\, \nabla^2_{x\,x} L(x^*, \lambda)\, Z$ heißt **reduzierte Hesse–Matrix**.

c) Interpretation der Lagrange-Multiplikatoren:

Sei x^* ein reguläres lokales Minimum des gleichungsrestringierten Optimierungsproblems (12.1) mit $I = \emptyset$ und sei auch die hinreichende Bedingung zweiter Ordnung erfüllt.

Variiert man nun die Nebenbedingungen des Optimierungsproblems im folgenden Sinn:

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n \\ &\text{unter den Nebenbedingungen} \\ &h_j(x) \;=\; \varepsilon_j, \quad j \in E := \{1, \ldots, p\}, \end{aligned}$$

so ist auch dieses Optimierungsproblem für betragsmäßig kleine ε_j lösbar und es gibt eine von $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)^{\mathrm{T}}$ abhängige Lösung $x^*(\varepsilon)$ (striktes lokales Minimum des variierten Optimieruingsproblems), die in einer Umgebung von $x^* = x^*(0)$ liegt. $x^*(\varepsilon)$ hängt stetig differenzierbar von ε ab und es gilt

$$\mu_j \;=\; -\frac{\partial}{\partial \varepsilon_j} f(x^*(\varepsilon))\big|_{\varepsilon=0}, \qquad j \in E. \tag{12.13}$$

Der Lagrange–Multiplikator μ_j gibt also an, wie sich der optimale Zielfunktionswert in erster Näherung ändert, wenn man die Nebenbedingung $h_j(x) = 0$ in obigem Sinne variiert.

Beispiel (12.14)

Minimiere $\qquad f(x_1, x_2) \;:=\; x_1 \,+\, x_2$

unter der Nebenbedingung $\qquad h(x_1, x_2) \;:=\; x_1^2 \,+\, x_2^2 \,-\, 2 \;=\; 0.$

Mit der Lagrange–Funktion $\quad L(x, \mu) \;=\; x_1 \,+\, x_2 \,+\, \mu\,(x_1^2 \,+\, x_2^2 \,-\, 2)$ $\quad$ ergeben sich die folgenden notwendigen Bedingungen

$$\begin{aligned} L_{x_1} \;&=\; 1 \,+\, 2\,\mu\, x_1 \;\;\;\;\;\; &=\; 0, \\ L_{x_2} \;&=\; 1 \,+\, 2\,\mu\, x_2 \;\;\;\;\;\; &=\; 0, \\ L_{\mu} \;&=\; x_1^2 \,+\, x_2^2 \,-\, 2 \;\;\; &=\; 0. \end{aligned}$$

Aus der letzten Gleichung folgt

$$2\,\mu^2 \;=\; (\mu\, x_1)^2 \,+\, (\mu\, x_2)^2 \;=\; (-\frac{1}{2})^2 \,+\, (-\frac{1}{2})^2 \;=\; \frac{1}{2}.$$

Damit ist $\mu \;=\; \pm\, 1/2$ und $x \;=\; \mp (1, 1)^{\mathrm{T}}$.

Mit $\nabla h \;=\; 2\,x$ ergibt sich der Tangentialraum zu $T_X(x) \;=\; \{y \in \mathbb{R}^2 :\; x^{\mathrm{T}}\, y \;=\; 0\}$.

Ferner ist

$$\nabla_x^2 L \;=\; \begin{pmatrix} 2\,\mu & 0 \\ 0 & 2\,\mu \end{pmatrix} \;=\; \begin{pmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{pmatrix}.$$

Die Hesse–Matrix der Lagrange–Funktion ist also nur für das obere Vorzeichen auf dem Tangentialraum $T_X(x)$ positiv semidefinit, dann aber sogar positiv definit.

Damit lautet die eindeutig bestimmte Lösung der Optimierungsaufgabe

$$x^* \;=\; \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \qquad \mu \;=\; \frac{1}{2}$$

und x^* ist ein striktes lokales Minimum. x^* ist sogar ein striktes globales Minimum (warum?).

Beispiel (12.15)

Minimiere $\qquad f(x_1, x_2) \;:=\; x_1 \,+\, x_2^2$

unter der Nebenbedingung $\qquad h(x_1, x_2) \;:=\; x_1^2 \;=\; 0.$

Diese Optimierungsaufgabe hat offensichtlich die eindeutig bestimmte Lösung $x^* = 0$ und diese ist ein striktes globales Minimum.

Auf der anderen Seite gilt für *alle* $\mu \in \mathbb{R}$

$$\nabla_x L(x^*, \mu) \;=\; \nabla f(x^*) \,+\, \mu\, \nabla h(x^*) \;=\; \begin{pmatrix} 1 \\ 0 \end{pmatrix} \,+\, \mu \begin{pmatrix} 0 \\ 0 \end{pmatrix} \;\neq\; 0\,.$$

Die Regularitätsbedingung ist also *nicht* erfüllt, ebenso wenig die notwendigen bzw. hinreichenden Bedingungen!

Aufgabe (12.16)

Bestimmen Sie das (globale) Maximum der Funktion $f(x_1, \ldots, x_n) := \prod_{j=1}^{n} x_j^2$ unter der Nebenbedingung $h(x_1, \ldots, x_n) := \sum_{j=1}^{n} x_j^2 - 1 = 0$ und leiten Sie hieraus die folgende Abschätzung zwischen geometrischem und arithmetischem Mittel her

$$\forall\; a_1, \ldots, a_n \;>\; 0 \quad \left(\prod_{j=1}^{n} a_j \right)^{1/n} \;\leq\; \frac{1}{n} \sum_{j=1}^{n} a_j.$$

13. Allgemeine restringierte Optimierung

Wir kehren zu allgemeinen restringierten Optimierungaufgaben mit Gleichungs- und Ungleichungsrestriktionen zurück. Die Aufgabe lautet wie früher:

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n \\ &\text{unter den Nebenbedingungen} \\ &g_i(x) \;\le\; 0, \quad i \in I \;:=\; \{1,\ldots,m\}, \\ &h_j(x) \;=\; 0, \quad j \in E \;:=\; \{1,\ldots,p\}. \end{aligned} \tag{13.1}$$

Die beteiligten Funktionen f, g_i und h_j mögen wieder auf ganz $\mathbb{R}^n$ definiert und hinreichend oft stetig differenzierbar sein. Der zulässige Bereich ist

$$X \;=\; \{x \in \mathbb{R}^n : \; g(x) \le 0 \;\wedge h(x) = 0\} \tag{13.2}$$

als nichtleer vorausgesetzt mit $g := (g_1,\ldots,g_m)^{\mathrm{T}}$, $h := (h_1,\ldots,h_p)^{\mathrm{T}}$ und $m > 0$.

Für einen zulässigen Punkt $x \in X$ bezeichnen wir mit

$$I(x) \;:=\; \{i \in I : \quad g_i(x) \;=\; 0\} \tag{13.3}$$

die *Menge der aktiven Indizes*. Ist $i \in I \setminus I(x)$, so ist die Nebenbedingung Nummer i mit $g_i(x) < 0$ erfüllt. Damit ist diese Nebenbedingung auch in einer ganzen Umgebung von x (bzgl. $\mathbb{R}^n$) erfüllt. Für $i \in I(x)$ ist dagegen die Nebenbedingung $g_i \le 0$ i. Allg. in jeder Umgebung von x verletzt.

Wir beginnen wieder mit der Untersuchung der Tangentialvektoren an die zulässige Menge.

Definition (13.4)

a) Ein Vektor $y \in \mathbb{R}^n$ heißt im einem Punkt $x \in X$ **tangential** an X, oder ein **Tangentialvektor** an X, falls es Folgen $(t_k) \in (\mathbb{R} \setminus \{0\})^{\mathbb{N}}$ und $(x^k) \in X^{\mathbb{N}}$ gibt mit den Eigenschaften

$$t_k \;\downarrow\; 0 \quad (k \to \infty), \qquad \lim_{k\to\infty} \frac{x^k - x}{t_k} \;=\; y.$$

b) Die Menge der an X tangentialen Vektoren in einem zulässigen Punkt $x \in X$ wird wie früher mit $T_X(x)$ bezeichnet.

Bemerkungen (13.5)

a) $T_X(x)$ ist stets ein *Kegel* mit Spitze 0 im $\mathbb{R}^n$, d.h.

$$y \in T_X(x) \;\Rightarrow\; \forall\, t \geq 0: \quad t\,y \in T_X(x).$$

$T_X(x)$ heißt **Tangentialkegel** von X in x. $T_X(x)$ ist stets abgeschlossen (Übungsaufgabe), i. Allg. jedoch kein linearer Raum.

b) Im Fall $I = \emptyset$ (d.h. es gibt nur Gleichungsrestriktionen) stimmt der Tangentialkegel $T_X(x)$ in einem regulären Punkt $x \in X$ mit den in (12.6) definierten Tangentialraum überein (Übungsaufgabe).

c) Ist $x \in X$ ein innerer Punkt von X (dies setzt i. Allg. $E = \emptyset$ voraus), so gilt offenbar $T_X(x) = \mathbb{R}^n$.

Satz (13.6) (Notwendige Bedingung erster Ordnung)

Ist x^* ein lokales Minimum von $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$ unter der Nebenbedingung $x \in X$, so gilt

$$\forall\, y \in T_X(x^*): \qquad \nabla f(x^*)^{\mathrm{T}}\, y \geq 0.$$

Beweis: Zu $y \in T_X(x^*)$ seien $(x^k) \in X^{\mathbb{N}}$ und $(t_k) \in (\mathbb{R} \setminus \{0\})^{\mathbb{N}}$ gemäß (13.4) gegeben. Dann folgt mittels Taylor–Entwicklung für hinreichend große Indizes k:

$$0 \;\leq\; f(x^k) \,-\, f(x^*) \;=\; \nabla f(\xi^k)^{\mathrm{T}}(x^k \,-\, x^*)$$

mit $\xi^k := x^k + \Theta_k\,(x^k - x^*)$, $0 < \Theta_k < 1$. Nach (13.4) konvergiert $x^k \to x^*$.

Division durch $t_k > 0$ und Grenzübergang $k \to \infty$ liefert die Behauptung $0 \leq \nabla f(x^*)^{\mathrm{T}}\, y$. □

Definition (13.7)

Für $x \in X$ definieren wir neben dem Tangentialkegel $T_X(x)$ in Analogie zu (12.6) den **linearisierten Tangentialkegel** gemäß

$$T_X^0(x) := \{y \in \mathbb{R}^n : \; \nabla g_i(x)^{\mathrm{T}} y \leq 0 \;\; (\forall\, i \in I(x)) \;\wedge\; \nabla h_j(x)^{\mathrm{T}} y = 0 \;\; (\forall\, j \in E)\}.$$

Dabei bezeichnet $I(x)$ die in (13.3) definierte Menge der aktiven Indizes.

Interpretation: Eine zulässige Richtung y im Sinne der notwendigen Bedingungen erster Ordnung ist eine solche, für die $x + ty$ die Nebenbedingungen für kleine $t > 0$ in erster Ordnung nicht verletzt. Für Indizes $i \in I \setminus I(x)$, d.h. $g_i(x) < 0$ sind daher alle Richtungen zulässig, während für Indizes $i \in I(x)$, d.h. $g_i(x) = 0$ nach dem Taylorschen Satz

$$0 \;\geq\; g_i(x + t\,y) \;=\; 0 \;+\; t\,\left(\nabla g_i(x)^{\mathrm{T}} y\right) \;+\; \mathrm{o}(t)$$

nur solche y zulässig sein können, für die $\nabla g_i(x)^{\mathrm{T}} y \leq 0$ gilt.

Satz (13.8) $\forall\ x \in X: \quad T_X(x) \ \subset\ T_X^0(x).$

Beweis:

Zu $y \in T_X(x)$ seien die Folgen $(t_k) \in (\mathbb{R} \setminus \{0\})^{\mathbb{N}}$ und $(x^k) \in X^{\mathbb{N}}$ wie in Definition (13.4) gegeben. Für $i \in I(x)$ gilt dann mit dem Taylorschen Satz

$$0 \ \geq\ g_i(x^k) \ =\ g_i(x) \ +\ \nabla g_i(\xi^k)^{\mathrm{T}}(x^k - x), \qquad g_i(x) \ =\ 0.$$

Division durch $t_k > 0$ und Grenzübergang $k \to \infty$ liefert $0 \geq \nabla g_i(x)^{\mathrm{T}} y$.

Analog sind die Gleichungsnebenbedingungen $h_j(x) = 0,\ j \in E,$ zu behandeln:

$$0 \ =\ h_j(x^k) \ =\ h_j(x) + \nabla h_j(\xi^k)^{\mathrm{T}}(x^k - x), \qquad h_j(x) \ =\ 0. \qquad \square$$

Satz (13.9) (KKT – Bedingungen[1])

Für ein lokales Minimum x^* der restringierten Optimierungsaufgabe (13.1) gelte die so genannte **Abadie–Constraint Qualification**[2] (ACQ)

$$T_X(x^*) \ =\ T_X^0(x^*). \tag{13.10}$$

Dann existieren *Lagrange–Multiplikatoren* $\lambda \in \mathbb{R}^m$ und $\mu \in \mathbb{R}^p$, so dass mit der *Lagrange–Funktion*

$$L(x, \lambda, \mu) \ :=\ f(x) \ +\ \lambda^{\mathrm{T}}\, g(x) \ +\ \mu^{\mathrm{T}}\, h(x) \tag{13.11}$$

die folgenden notwendigen Bedingungen gelten

(i)	$\nabla_x L(x^*, \lambda, \mu) \ =\ 0,$	(notw. Bed. erster Ordnung)
(ii)	$g(x^*) \ \leq\ 0, \qquad h(x^*) \ =\ 0,$	(Zulässigkeit)
(iii)	$\lambda \ \geq\ 0,$	(Vorzeichenbedingung)
(iv)	$\lambda^{\mathrm{T}}\, g(x^*) \ =\ 0$	(Komplementarität)

Punkte (x^*, λ, μ), die diese KKT-Bedingungen erfüllen, werden auch *KKT-Punkte* genannt.

Geometrische Interpretation (13.12)

Wir geben eine physikalisch-geometrische Interpretation der KKT-Bedingungen für Ungleichungsrestriktionen an. Hierzu beschreibe x die Lage eines Massenpunktes (Ortsvektor), der sich in einem Kraftfeld $-\nabla f(x)$ (z.B. die Erdanziehungskraft) bewegt.

[1] benannt nach William Karush (1939) und Harold W. Kuhn, Albert W. Tucker (1951)

[2] J. Abadie: On the Kuhn-Tucker Theorem, In: Nonlinear Programming. North-Holland, pp. 21-36 (1967)

Die Zielfunktion $f(x)$ beschreibt also die potentielle Energie dieses Kraftfeldes. Die Ungleichungsnebenbedingungen $g_i(x) \leq 0$ werden als Zwangsbedingungen (Wände) interpretiert.

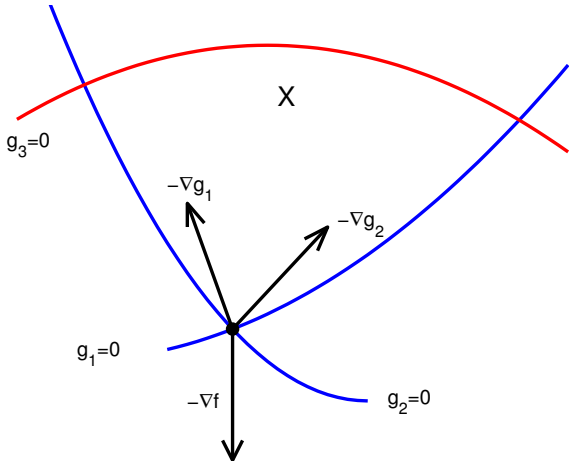


Abb. 13.1 KKT–Bedingungen.

Befindet sich der Massenpunkt nun an einer Stelle x^* in Ruhe, so nimmt die potentielle Energie dort ein lokales Minimum an. Auf den Massenpunkt wirkt dann die Kraft $-\nabla f(x^*)$, sowie Zwangskräfte, die von den aktiven Wänden ausgeübt werden. Diese Kräfte müssen auf den entsprechenden Wänden senkrecht stehen, damit haben sie notwendigerweise die Form $-\lambda_i \nabla g_i(x^*)$, da der Gradient $\nabla g_i(x)$ ja auf der Äquipotentialfläche $g_i(x) = 0$ senkrecht steht. Da dieser in die Richtung wachsender g_i-Werte weist, gilt $\lambda_i \geq 0$.

Aus dem Kräftegleichgewicht folgt schließlich

$$-\nabla f(x) \; + \sum_{i \in I(x^*)} \lambda_i \; (-\nabla g_i(x^*)) \; = \; 0.$$

Die entspricht gerade der notwendigen Bedingung erster Ordnung.

Ein Standardweg zum Beweis der KKT-Bedingungen verwendet das so genannte Lemma von Farkas, welches selbst wieder relativ leicht mittels des folgenden Trennungssatzes hergeleitet werden kann.

Satz (13.13) (Trennungssatz)

Sei $K \subset \mathbb{R}^m$ eine nichtleere, konvexe Teilmenge und $b \in \mathbb{R}^m \setminus \overline{K}$.
Dann existiert $y \in \mathbb{R}^m \setminus \{0\}$ mit

$$y^{\mathrm{T}} b \; < \; \inf \{y^{\mathrm{T}} x : \; x \in K\}.$$

Folgerung: Wählt man $\alpha \in \mathbb{R}$ mit $y^{\mathrm{T}} b < \alpha < \inf\{y^{\mathrm{T}} x : x \in K\}$, so *trennt* die Hyperebene $H = \{x : \; y^{\mathrm{T}} x = \alpha\}$ die konvexe Menge K und den Punkt b.

Beweis:

Sei $\delta := \operatorname{dist}(K,b) = \inf\{\|x-b\| : x \in K\}$. Nach Voraussetzung ($b \notin \overline{K}$) ist dann $\delta > 0$. Ein Kompaktheitsschluss zeigt weiter

$$\exists\, x^* \in \partial K : \quad \|x^* - b\| = \delta.$$

Mit K ist auch $\overline{K}$ konvex. Für $x \in K$ ist daher $x^* + t\,(x - x^*) \in \overline{K}$, $0 \le t \le 1$. Damit folgt für alle $t \in [0,1]$:

$$\begin{aligned} & \|(x^* + t\,(x-x^*)) - b\|^2 \ge \|x^* - b\|^2 \\ \Rightarrow \quad & \|(x^* - b) + t\,(x-x^*)\|^2 \ge \|x^* - b\|^2 \\ \Rightarrow \quad & 2\,t\,(x^*-b)^{\mathrm{T}}(x-x^*) + t^2\,\|x-x^*\|^2 \ge 0 \\ \Rightarrow \quad & (x^*-b)^{\mathrm{T}}(x-x^*) \ge 0. \end{aligned}$$

Mit $y := x^* - b$ folgt somit

$$\forall\, x \in K : \quad y^{\mathrm{T}}x \ge y^{\mathrm{T}}x^* = y^{\mathrm{T}}b + y^{\mathrm{T}}(x^*-b) = y^{\mathrm{T}}b + \delta^2.$$

und damit auch

$$y^{\mathrm{T}}b \le \inf\{y^{\mathrm{T}}x : x \in K\} - \delta^2 < \inf\{y^{\mathrm{T}}x : x \in K\}. \qquad \square$$

Satz (13.14) (Lemma von Farkas [3])

Seien $A \in \mathbb{R}^{(m,n)}$ und $b \in \mathbb{R}^m$. Dann gilt *genau eine* der beiden folgenden Aussagen

(a) $\exists\, z \in \mathbb{R}^n : \quad A\,z = b \ \wedge\ z \ge 0,$

(b) $\exists\, y \in \mathbb{R}^m : \quad A^{\mathrm{T}}\,y \ge 0 \ \wedge\ b^{\mathrm{T}}y < 0.$

Beweis:

Zunächst sehen wir, dass beide Aussagen nicht zugleich gelten können. Dann wäre nämlich

$$0 \le z^{\mathrm{T}}(A^{\mathrm{T}}y) = (y^{\mathrm{T}}A)\,z = y^{\mathrm{T}}(A\,z) = y^{\mathrm{T}}b = b^{\mathrm{T}}y < 0.$$

Widerspruch!

Gilt die Aussage (a) nicht, so ist $b \notin K := \{A\,z : z \ge 0\}$. Man sieht unmittelbar, dass K ein nichtleerer, abgeschlossener und konvexer Kegel in $\mathbb{R}^m$ ist. Nach dem Trennungssatz (13.13) folgt somit

$$\exists\, y \in \mathbb{R}^m \setminus \{0\} : \quad y^{\mathrm{T}}b < \inf\{y^{\mathrm{T}}x : x \in K\}. \qquad (*)$$

[3] Julius Farkas (1847–1930) Ungarischer Physiker u. Mathematiker
Über die Theorie der einfachen Ungleichungen; Journal für reine und angewandte Mathematik, Bd.124, 1-27, 1902.

Nun ist K ein Kegel. Daher liegen mit x auch alle nichtnegativen Vielfachen $t\,x$, $t \geq 0$, in K. Somit kann das obige Infimum in $(*)$ aber nur die Werte $-\infty$ oder 0 annehmen, je nachdem, ob es ein $x \in K$ gibt mit $y^{\mathrm{T}}x < 0$, oder nicht. Wegen $(*)$ ist der erste Fall jedoch ausgeschlossen, damit folgt $\inf\{y^{\mathrm{T}}x : \ x \in K\} = 0$.

Somit gelten die Ungleichungen

$$\forall\, x \in K : \quad y^{\mathrm{T}}b < 0 \leq y^{\mathrm{T}}x.$$

Setzen wir hierin nun für $x = A\,e_i =: a^i \in K$ (i-ter Spaltenvektor von A) ein, $i = 1, \ldots, n$, so ergibt sich gerade die Aussage (b): $y^{\mathrm{T}}b < 0$ und $A^{\mathrm{T}}y \geq 0$. □

Wir kommen nun zum Beweis der KKT-Bedingungen.

Beweis zu (13.9)

Da x^* ein lokales Minimum von f unter der Nebenbedingung $x \in X$ ist, folgt nach Satz (13.6) $\forall\, y \in T_X(x^*) : \ \nabla f(x^*)^{\mathrm{T}}y \geq 0$. Aufgrund der vorausgesetzten Abadie Constraint Qualification (ACQ) gilt also auch

$$\forall\, y \in T_X^0(x^*) : \quad \nabla f(x^*)^{\mathrm{T}}y \geq 0.$$

Wir definieren nun eine $(k + 2\,p, n)$–Matrix A^{T}, wobei $k := \#I(x^*)$, gemäß

$$A^{\mathrm{T}} := \begin{pmatrix} -\nabla g_i(x^*)^{\mathrm{T}}{}_{i \in I(x^*)} \\ -\nabla h_j(x^*)^{\mathrm{T}}{}_{j \in E} \\ \nabla h_j(x^*)^{\mathrm{T}}{}_{j \in E} \end{pmatrix}$$

Damit lässt sich der linearisierte Tangentialkegel, vgl. (13.7), wie folgt beschreiben $T_X^0(x^*) = \{y \in \mathbb{R}^n : \ A^{\mathrm{T}}y \geq 0\}$ und wir finden

$$\forall\, y \in \mathbb{R}^n : \quad A^{\mathrm{T}}y \geq 0 \quad \Rightarrow \quad \nabla f(x^*)^{\mathrm{T}}y \geq 0.$$

Die Aussage (b) des Farkas' Lemmas ist also falsch und es gilt somit die Aussage (a):

$$\exists\, z \in \mathbb{R}^{(k+2\,p)} : \quad A\,z = \nabla f(x^*) \ \wedge \ z \geq 0.$$

Zerlegt man den Vektor z analog zur Matrix A^{T}, also $z^{\mathrm{T}} =: (\lambda^{\mathrm{T}}, (\mu^+)^{\mathrm{T}}, (\mu^-)^{\mathrm{T}})$, so erhalten wir

$$\nabla f(x^*) = -\sum_{i \in I(x^*)} \lambda_i\, \nabla g_i(x^*) - \sum_{j \in E} (\mu_j^+ - \mu_j^-)\, \nabla h_j(x^*); \qquad \lambda_i,\, \mu_j^+,\, \mu_j^- \geq 0.$$

Erweitert man den Vektor λ zu $\lambda \in \mathbb{R}^m$ vermöge $\lambda_i := 0$ für $i \in I \setminus I(x^*)$ und setzt man $\mu_j := \mu_j^+ - \mu_j^-$, so erhält man die Existenz von $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^p$ mit

$$\begin{aligned} &\nabla f(x^*) + \sum_{i \in I} \lambda_i\, \nabla g_i(x^*) + \sum_{j \in E} \mu_j\, \nabla h_j(x^*) = 0, \\ &\lambda \geq 0, \quad \forall\, i \in I : \ \lambda_i\, g_i(x^*) = 0. \end{aligned}$$

□

In Satz (13.9) werden die KKT-Bedingungen unter der relativ schwachen Voraussetzung der Abdie CQ hergeleitet. Diese Voraussetzung lässt sich in den Anwendungen jedoch nur schwer überprüfen. Wir geben daher einige stärkere Voraussetzungen an, aus denen die ACQ folgt, die sich aber einfacher überprüfen lassen.

Satz (13.15) (Mangasarian-Fromovitz-CQ[4])

Erfüllt ein Punkt $x \in X$ die so genannte *Mangasarian-Fromovitz-Constraint-Qualification* (MFCQ)

(i) $\nabla h_j(x),\ j \in E,$ linear unabhängig,

(ii) $\exists\, d \in \mathbb{R}^n :\ \nabla g_i(x)^{\mathrm{T}} d < 0\ (i \in I(x))\ \wedge\ \nabla h_j(x)^{\mathrm{T}} d = 0\ (j \in E),$

so erfüllt er auch die ACQ.

Beweis:

Sei $y \in T_X^0(x)$, also

$$\forall\, i \in I(x) :\ \nabla g_i(x)^{\mathrm{T}} y \ \leq\ 0, \qquad \forall\, j \in E :\ \nabla h_j(x)^{\mathrm{T}} y \ =\ 0.$$

Wir zeigen, dass auch $y \in T_X(x)$ gilt. Dazu setzen wir $y(\tau) := y + \tau\, d,$ wobei d gemäß (ii) gewählt sei. Für ein festes $\tau > 0$ folgt dann

$$\nabla g_i(x)^{\mathrm{T}} y(\tau) \ <\ 0\ (i \in I(x)), \quad \nabla h_j(x)^{\mathrm{T}} y(\tau) \ =\ 0\ (j \in E).$$

Satz (12.7) liefert nun (bei festem $\tau > 0$) die Existenz eines C^1–Weges $x_\tau :]-\varepsilon, \varepsilon[\to \mathbb{R}^n,\ \varepsilon = \varepsilon_\tau > 0,$ mit

$$h(x_\tau(t)) \ =\ 0\ (\forall\, |t| < \varepsilon), \quad x_\tau(0) \ =\ x, \quad x_\tau'(0) \ =\ y(\tau).$$

Wegen $\nabla g_i(x)^{\mathrm{T}} x_\tau'(0) \ <\ 0\ (i \in I(x))$ ist $x(t)$ für hinreichend kleine $0 \leq t < \varepsilon$ zulässig, also $x_\tau(t) \in X$.

Ist $t_k \in]0, \varepsilon[$ nun eine Folge, die streng monoton gegen 0 konvergiert, so setzt man $x^k := x_\tau(t_k) \in X$. Für diese Folge gilt dann

$$\lim_{k\to\infty} \frac{x^k - x}{t_k} \ =\ \lim_{k\to\infty} \frac{x_\tau(t_k) - x_\tau(0)}{t_k} \ =\ x_\tau'(0) \ =\ y(\tau).$$

Damit ist $y(\tau) \ =\ y + \tau d \ \in\ T_X(x)$. Da dies für alle $\tau > 0$ gilt, folgt aus der Abgeschlossenheit von $T_X(x),$ dass auch $y \in T_X(x)$ liegt. □

[4] O.L.Mangasarian, S. Fromovitz: The Fritz John Necessary Optimality Conditions in the Presence of Equality and Inequality Constraints, J. Math. Anal. a. Appl., 17, 37-47, 1967

Satz (13.16) (Linear Independence CQ)

Erfüllt ein Punkt $x \in X$ die so genannte *Linear-Independence-Constraint-Qualification* (LICQ)

$$\nabla g_i(x),\ i \in I(x), \quad \text{und} \quad \nabla h_j(x),\ j \in E, \quad \text{linear unabhängig,}$$

so erfüllt er auch die MFCQ und damit auch die ACQ .

Beweis:

Aufgrund der linearen Unabhängigkeit der Vektoren $\nabla g_i(x),\ i \in I(x),$ und $\nabla h_j(x),\ j \in E,$ hat die Matrix

$$A := \begin{pmatrix} \nabla g_i(x)^{\mathrm{T}}_{i \in I(x)} \\ \\ \nabla h_j(x)^{\mathrm{T}}_{j \in E} \end{pmatrix} \in \mathbb{R}^{(\ell,n)}, \qquad \ell \leq n,$$

maximalen Rang ℓ. Damit besitzt das lineare Gleichungssystem $A\,d = b$ für jede vorgegebene rechte Seite $b \in \mathbb{R}^\ell$ (wenigstens) eine Lösung $d \in \mathbb{R}^n$. Setzen wir also

$$b := \begin{pmatrix} -1 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^\ell,$$

so folgt die Existenz eines Vektors $d \in \mathbb{R}^n$ mit $\nabla g_i(x)^{\mathrm{T}} d = -1 < 0,\ i \in I(x),$ und $\nabla h_j(x)^{\mathrm{T}} d = 0,\ j \in E.$ Damit erfüllt x die MFQC. □

Bemerkung (13.17)

Ist $x^* \in X$ ein lokales Minimum der restringierten Optimierungsaufgabe (13.1) und erfüllt x^* die LICQ Bedingung, so sind die Lagrange-Multiplikatoren λ und μ aus den zugehörigen KKT-Bedingungen *eindeutig bestimmt.*

Beispiele (13.18)

a) Wir betrachten die folgenden Ungleichungsnebenbedingungen in $\mathbb{R}^2$

$$\begin{aligned} g_1 &:= x_2 - x_1^2 \leq 0, \\ g_2 &:= -x_2 \leq 0. \end{aligned}$$

Der Punkt $x^* = 0$ ist zulässig, beide Nebenbedingungen sind aktiv. Ferner gilt

$$\nabla g_1(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \nabla g_2(0) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

Daher gibt es keinen Vektor $d \in \mathbb{R}^2$ mit $\nabla g_i(x^*)^{\mathrm{T}} d < 0$, $i = 1, 2$. Andererseits gilt $T_X(x^*) = \{y : \ y_2 = 0\} = T_X^0(x^*)$.

Die ACQ ist also erfüllt, die MFCQ jedoch nicht!

b) Wir betrachten die folgenden drei Ungleichungsnebenbedingungen in $\mathbb{R}^2$

$$\begin{aligned} g_1 &:= (x_1 - 1)^2 + (x_2 - 1)^2 - 2 \leq 0, \\ g_2 &:= (x_1 - 1)^2 + (x_2 + 1)^2 - 2 \leq 0, \\ g_3 &:= -x_1 \leq 0 \end{aligned}$$

Der Punkt $x^* = 0$ ist zulässig, alle drei Nebenbedingungen sind aktiv. Ferner gilt

$$\nabla g_1(0) = \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \quad \nabla g_2(0) = \begin{pmatrix} -2 \\ 2 \end{pmatrix}, \quad \nabla g_3(0) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Damit ist die MFCQ mit $d = (1,\ 0)^{\mathrm{T}}$ erfüllt, andererseits ist die LICQ Bedingung offensichtlich nicht erfüllt!

Bevor wir im Folgenden die notwendigen und hinreichenden Bedingungen zweiter Ordnung herleiten, betrachten wir die Spezialfälle der linearen und konvexen Optimierungsaufgaben.

Lineare Optimierungsaufgaben (lineare Programme).

Für lineare Restriktionen ist die ACQ stets erfüllt, daher sind die KKT-Bedingungen ohne zusätzliche Regularitätsannahmen notwendig.

Satz (13.19) (KKT-Bedingungen für lineare Restriktionen)

Ist $x^* \in \mathbb{R}^n$ ein lokales Minimum von $f \in \mathrm{C}^1(\mathbb{R}^n, \mathbb{R})$ unter linearen Nebenbedingungen

$$\begin{aligned} g_i(x) &:= (a^i)^{\mathrm{T}} x - p_i \leq 0, \qquad i \in I = \{1, \ldots, m\}, \\ h_j(x) &:= (b^j)^{\mathrm{T}} x - q_j = 0, \qquad j \in E = \{1, \ldots, p\}, \end{aligned}$$

so existieren Lagrange-Multiplikatoren $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^p$, so dass die folgenden KKT-Bedingungen erfüllt sind

(**i**) $\quad \nabla f(x^*) + \sum\limits_{i \in I} \lambda_i \, a^i + \sum\limits_{j \in E} \mu_j \, b^j = 0,$

(**ii**) $\quad (a^i)^{\mathrm{T}} x^* \leq p_i \quad (i \in I), \qquad (b^j)^{\mathrm{T}} x^* = q_j \quad (j \in E),$

(**iii**) $\quad \lambda_i \geq 0, \qquad \lambda_i \, ((a^i)^{\mathrm{T}} x^* - p_i) = 0 \quad (i \in I).$

Beweis:

Es genügt zu zeigen, dass die ACQ , $T_X^0(x^*) \subset T_X(x^*)$, erfüllt ist.

Sei also $y \in T_X^0(x^*)$, d.h. $(a^i)^{\mathrm{T}} y \leq 0$, $i \in I(x^*)$, und $(b^j)^{\mathrm{T}} y = 0$, $j \in E$. Weiter sei $t_k > 0$ eine Folge mit $t_k \downarrow 0$ $(k \to \infty)$.

Mit $x^k := x^* + t_k\, y$ folgt dann für hinreichend große Indizes k

$$\begin{aligned}
(a^i)^{\mathrm{T}}\, x^k &= (a^i)^{\mathrm{T}} x^* + t_k\, (a^i)^{\mathrm{T}}\, y \;\leq\; p_i, \qquad \forall\, i \in I(x^*),\\
(a^i)^{\mathrm{T}}\, x^k &= (a^i)^{\mathrm{T}} x^* + t_k\, (a^i)^{\mathrm{T}}\, y \;<\; p_i, \qquad \forall\, i \in I \setminus I(x^*),\\
(b^j)^{\mathrm{T}}\, x^k &= (b^j)^{\mathrm{T}} x^* + t_k\, (b^j)^{\mathrm{T}}\, y \;=\; q_j, \qquad \forall\, j \in E.
\end{aligned}$$

Damit ist $x^k \in X$ für hinreichend große k und es gilt $\lim\limits_{k\to\infty} \dfrac{x^k - x^*}{t_k} = y$, also $y \in T_X(x^*)$. □

Dualität.

Jedem linearen Optimierungsproblem wird ein neues, so genanntes **duales Optimierungsproblem** zugeordnet. In der symmetrischen Form eines linearen Programms lautet diese Zuordnung

Primales Problem (P) (13.20)

Minimiere	$J_P(x) = c^{\mathrm{T}} x, \qquad x \in \mathbb{R}^n,$
Nebenbedingungen	$A\, x \geq b \in \mathbb{R}^m, \quad x \geq 0.$

Duales Problem (D) (13.21)

Maximiere	$J_D(y) = b^{\mathrm{T}} y, \qquad y \in \mathbb{R}^m,$
Nebenbedingungen	$A^{\mathrm{T}}\, y \leq c \in \mathbb{R}^n, \quad y \geq 0.$

Bemerkungen (13.22)

a) Durch die Einführung von **Schlupfvariablen** lässt sich das Problem (P) auf die Standardform eines linearen Programms bringen

$$\text{Minim.} \quad J = c^{\mathrm{T}} x, \quad \text{NB.} \quad Ax = b, \quad x \geq 0. \tag{13.23}$$

Umgekehrt lässt sich natürlich auch jedes lineare Programm in Standardform durch Ersetzen der Gleichungsnebenbedingungen durch jeweils zwei Ungleichungsnebenbedingungen in die symmetrische Form bringen.

Für ein primales Problem in Standardform (13.23) lautet das zugehörige duale Problem

$$\text{Maxim.} \quad J_D(y) \;=\; b^{\mathrm{T}}y, \quad \text{NB.} \quad A^{\mathrm{T}}y \;\leq\; c.$$

b) Dualität ist **selbstinvers**, d.h. schreibt man das duale Problem (D) in die Form (P) um:

$$\text{Minim.} \quad J \;=\; (-b)^{\mathrm{T}}y, \quad \text{NB.} \quad (-A^{\mathrm{T}})y \;\geq\; (-c), \quad y \;\geq\; 0,$$

so ist das hierzu duale Problem wiederum zum Ausgangsproblem (P) äquivalent.

Satz (13.24) (Schwacher Dualitätssatz)

Ist $x \in \mathbb{R}^n$ zulässig für (P) und $y \in \mathbb{R}^m$ zulässig für (D), so gilt

$$J_P(x) \;\geq\; J_D(y).$$

Insbesondere gilt damit für die Optimalwerte

$$v(P) \;:=\; \inf\{c^{\mathrm{T}}x :\; Ax \geq b \wedge x \geq 0\} \;\geq\; v(D) \;:=\; \sup\{b^{\mathrm{T}}y :\; A^{\mathrm{T}}y \leq c \wedge y \geq 0\}.$$

Beweis:

Aus der Zulässigkeit folgt sofort

$$c^{\mathrm{T}}x \;=\; x^{\mathrm{T}}c \;\geq\; x^{\mathrm{T}}(A^{\mathrm{T}}y) \;=\; (Ax)^{\mathrm{T}}y \;\geq\; b^{\mathrm{T}}y.$$

□

Satz (13.25) (Existenz)

Ist die Menge X_P der zulässigen Punkte eines linearen Programms (P) nichtleer und ist der Optimalwert des Problems endlich, $v(P) > -\infty$, so hat (P) eine Lösung x^*.

Beweis:

Man transformiere das Problem in die Standardform (13.23) eines linearen Programms und führe $x_0 := c^{\mathrm{T}}x$ als neue Variable des Problems ein. Dann lautet die Optimierungsaufgabe

$$\begin{array}{ll} \text{Minimiere} & \tilde{J} \;=\; x_0, \qquad (x_0, x)^{\mathrm{T}} \;\in\; \mathbb{R}^{n+1}, \\ \text{Nebenbedingungen} & \tilde{A}\,x \;:=\; \begin{pmatrix} c^{\mathrm{T}} \\ A \end{pmatrix} x \;=\; \begin{pmatrix} x_0 \\ b \end{pmatrix} \;\in\; \mathbb{R}^{m+1}, \quad x \;\geq\; 0. \end{array}$$

Bezeichne nun $\tilde{a}^1, \ldots, \tilde{a}^n$ die Spaltenvektoren der erweiterten Matrix $\tilde{A}$, so ist

$$K \;:=\; \{z = \sum_{i=1}^{n} x_i \tilde{a}^i \;\in\; \mathbb{R}^{m+1} :\; x_i \;\geq\; 0,\; i = 1, \ldots, n\}$$

der von diesen erzeugte konvexe Kegel. Der Kern des Beweises (den wir hier auslassen wollen) besteht nun darin zu zeigen, dass dieser Kegel (da *endlich erzeugt*) sowohl *konvex* als auch *abgeschlossen* ist.

Das lineare Programm ist nun äquivalent zu

$$\text{Minimiere } \tilde{J} \;=\; x_0, \qquad \text{NB.} \quad \begin{pmatrix} x_0 \\ b \end{pmatrix} \in K.$$

Nun ist die Menge

$$K \;\cap\; \left\{ \begin{pmatrix} x_0 \\ b \end{pmatrix} \in \mathbb{R}^{m+1} : \; v(P) \;\le\; x_0 \;\le\; v(P)+1 \right\}$$

aufgrund der Voraussetzung nichtleer, beschränkt und auch abgeschlossen, also kompakt. Die stetige Funktion $f(x_0, x) := x_0$ nimmt daher auf dieser Menge ein Minimum an. □

Satz (13.26) (Starker Dualitätssatz)

Gilt für das lineare Programm (P): $X \neq \emptyset$ und $v(P) > -\infty$, so ist auch die zulässige Menge X_D des zugehörigen dualen linearen Programms (D) nichtleer und es gilt $v(P) = v(D)$, d.h. es gibt keine Dualitätslücke.

Beweis:

Die Nebenbedingung des Problems (P) haben die Form $g_1(x) := b - A\,x \le 0$ und $g_2(x) := -x \le 0$.

Nach (13.25) besitzt (P) eine Lösung x^*. Wir wenden die KKT-Bedingungen gemäß (13.19) an. Es gibt daher Lagrange-Parameter $\lambda \in \mathbb{R}^m$ und $\mu \in \mathbb{R}^n$, so dass gelten

$$\textbf{(i)} \qquad c \;-\; A^{\mathrm{T}}\lambda \;-\; \mu \;=\; 0,$$

$$\textbf{(ii)} \qquad b \;-\; A\,x \;\le\; 0, \qquad x \;\ge\; 0,$$

$$\textbf{(iii)} \qquad \lambda \;\ge\; 0, \quad \mu \;\ge\; 0, \qquad \lambda^{\mathrm{T}}(b - Ax) \;=\; 0, \quad \mu^{\mathrm{T}}x \;=\; 0.$$

Vergleicht man diese Bedingungen nun mit dem dualen linearen Programm (13.21), so findet man zunächst, dass $y := \lambda \in \mathbb{R}^m$ zulässig ist für (D)

$$A^{\mathrm{T}}\lambda \;=\; c - \mu \;\le\; c, \qquad \lambda \;\ge\; 0.$$

Ferner ist auch

$$J_D(\lambda) \;=\; b^{\mathrm{T}}\lambda \;=\; \lambda^{\mathrm{T}}\,A\,x^* \;=\; (x^*)^{\mathrm{T}}(c-\mu) \;=\; c^{\mathrm{T}}x^* \;=\; J_P(x^*).$$

Nach dem schwachen Dualitätssatz (13.24) ist daher $y^* = \lambda$ eine Lösung von (D) und es gilt $J_D(y^*) = J_P(x^*)$. □

Konvexe Optimierungsaufgaben.

Wir betrachten im Folgenden noch kurz *konvexe Optimierungsaufgaben*. Dies sind Aufgaben der Form

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n \\ &\text{unter den Nebenbedingungen} \\ &g_i(x) \;\leq\; 0, \quad i \in I := \{1, \ldots, m\}, \\ &h_j(x) := (b^j)^{\mathrm{T}} x - q_j \;=\; 0, \quad j \in E := \{1, \ldots, p\}, \end{aligned} \tag{13.27}$$

wobei die Funktionen f und g_i als *konvex* vorausgesetzt werden.

Definition (13.28)

Wir sagen, die konvexe Optimierungsaufgabe (13.27) erfüllt die **Slater-Bedingung**, falls es einen zulässigen Punkt $x \in X$ gibt, für den keine der Ungleichungsnebenbedingungen aktiv ist:

$$\exists\, x \in \mathbb{R}^n : \;\; g_i(x) \;<\; 0 \;\; (\forall\, i \in I) \;\; \wedge \;\; (b^j)^{\mathrm{T}} x \;=\; q_j \;\; (\forall\, j \in E).$$

Satz (13.29)

Erfüllt das konvexe Optimierungsproblem (13.27) die Slater-Bedingung, so erfüllt auch jedes lokale Minimum x^* von (13.27) die Abadie CQ.

Beweis:

Sei $x \in \mathbb{R}^n$ wie in der Slater-Bedingung gegeben und $d := x - x^*$. Aufgrund der Konvexität des Problems gelten dann (vgl. (3.5))

$$\begin{aligned} \nabla g_i(x^*)^{\mathrm{T}} d \;&\leq\; g_i(x) - g_i(x^*) \;=\; g_i(x) \;<\; 0 \quad (i \in I(x^*)) \\ \nabla h_j(x^*)^{\mathrm{T}} d \;&=\; h_j(x) - h_j(x^*) \;=\; 0 \quad (j \in E) \end{aligned}$$

Für $y \in T_X^0(x^*)$ (d.h. $\nabla g_i(x^*)^{\mathrm{T}} y \leq 0$ und $\nabla h_j(x^*)^{\mathrm{T}} y = 0$) und $t > 0$ setze man $y(t) := y + t\, d$. Für alle $t > 0$ erfüllt $y(t)$ damit die gleichen Bedingungen wie d selbst, nämlich

$$\nabla g_i(x^*)^{\mathrm{T}} y(t) \;<\; 0 \quad (i \in I(x^*)), \qquad \nabla h_j(x^*)^{\mathrm{T}} y(t) \;=\; 0 \quad (j \in E).$$

Man sieht nun allgemein mittels Taylor-Entwicklung (Übungsaufgabe), dass

$$T_{strikt} := \{z \in \mathbb{R}^n : \; \nabla g_i(x^*)^{\mathrm{T}} z < 0 \,(i \in I(x^*)) \;\wedge\; (b^j)^{\mathrm{T}} z = 0 \,(j \in E)\} \;\subset\; T_X(x^*).$$

Damit sind alle $y(t) \in T_X(x^*),\ t > 0$. Da $T_X(x^*)$ abgeschlossen ist, folgt für $t \downarrow 0$, dass auch $y \in T_X(x^*)$. □

Unter der vorausgesetzten Slater-Bedingungen sind somit die KKT-Bedingungen *notwendig* für ein lokales (und damit auch globales) Minimum des konvexen Optimierungsproblems.

Man kann nun leicht sehen, dass auch umgekehrt, die KKT-Bedingungen *hinreichend* sind für ein lokales (globales) Minimum, und zwar ohne weiterw Voraussetzungen.

Satz (13.30)

Ist (x^*, λ, μ) ein KKT-Punkt für das konvexe Optimierungsproblem (13.27), so ist x^* ein globales Minimum.

Beweis:

Für einen zulässigen Punkt $x \in X$ gilt mit der Konvexität und den KKT-Bedingungen

$$\begin{aligned} f(x) &\geq f(x^*) + \nabla f(x^*)^{\mathrm{T}}(x - x^*) \\ &= f(x^*) - \sum_{i \in I} \lambda_i \nabla g_i(x^*)^{\mathrm{T}}(x - x^*) - \sum_{j \in E} \mu_j (b^j)^{\mathrm{T}}(x - x^*) \\ &= f(x^*) - \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*)^{\mathrm{T}}(x - x^*) \\ &\geq f(x^*) - \sum_{i \in I(x^*)} \lambda_i (g_i(x) - g_i(x^*)) \\ &\geq f(x^*). \end{aligned}$$

□

Notwendige u. hinreichende Bedingungen zweiter Ordnung.

Im Folgenden setzen wir generell voraus, dass die auftretenden Funktionen f, g_i und h_j zweifach stetig differenzierbar sind. Für zulässige Punkte $x \in X$ definieren wir neben dem **Tangentialkegel**

$$T_X(x) := \{y \in \mathbb{R}^n : \ \exists (x^k) \in X^{\mathbb{N}},\, t_k > 0 : \ t_k \downarrow 0 \ \wedge \ \frac{x^k - x}{t_k} \to y\}$$

und dem **linearisierten Tangentialkegel**

$$T_X^0(x) := \{y \in \mathbb{R}^n : \ \nabla g_i(x)^{\mathrm{T}} y \leq 0 \ (\forall i \in I(x)) \ \wedge \ \nabla h_j(x)^{\mathrm{T}} y = 0 \ (\forall j \in E)\}$$

den so genannten **reduzierten Tangentialkegel**, der nunmehr nicht nur von dem zulässigen Punkt $x \in X$, sondern auch von einem Lagrange-Multiplikator $\lambda \in \mathbb{R}^m$,

$\lambda \geq 0$, abhängt. Der reduzierte Tangentialkegel schränkt die zugelassenen Variationen gegenüber dem linearisierten Tangentialkegel ein und behandelt dabei die aktiven Ungleichungsrestriktionen, für die *strikte Komplementarität* gilt, analog zu den Gleichungsnebenbedingungen.

$$T_X^1(x,\lambda) \;:=\; \Big\{ y \in \mathbb{R}^n : \quad \begin{array}{ll} \nabla g_i(x)^{\mathrm{T}} y \;=\; 0, & i \in I(x),\ \lambda_i > 0, \\ \nabla g_i(x)^{\mathrm{T}} y \;\leq\; 0, & i \in I(x),\ \lambda_i = 0, \\ \nabla h_j(x)^{\mathrm{T}} y \;=\; 0, & j \in E \end{array} \Big\}. \tag{13.31}$$

Wir stellen fest, dass offensichtlich $T_X^1(x,\lambda) \subset T_X^0(x)$ gilt.

Andererseits gilt für den so genannten **Tangentialraum der aktiven Nebenbedingungen**

$$T_X^a(x) \;:=\; \{ y : \; \nabla g_i(x)^{\mathrm{T}} y = 0 \,(i \in I(x)) \;\wedge\; \nabla h_j(x)^{\mathrm{T}} y = 0 \,(j \in E) \} \tag{13.32}$$

die umgekehrte Inklusion $T_X^1(x,\lambda) \;\supset\; T_X^a(x)$.

Satz (13.33) (Notwendige Bedingung zweiter Ordnung)

Ist $x^* \in X$ ein lokales Minimum der restrigierten Optimierungsaufgabe (13.1) mit $f,\ g_i,\ h_j \in \mathrm{C}^2$ und gilt die LICQ, so folgt mit den zugehörigen (eindeutig bestimmten) Lagrange-Multiplikatoren $\lambda \in \mathbb{R}^m,\ \mu \in \mathbb{R}^p$

$$\forall\, y \in T_X^1(x^*,\lambda) : \quad y^{\mathrm{T}}\, \nabla_{xx}^2 L(x^*,\lambda,\mu)\, y \;\geq\; 0.$$

Beweis:

Zu $y \in T_X^1(x^*,\lambda)$ definieren wir die folgenden Indexmengen:

$$\begin{aligned} I_> \;&:=\; \{ i \in I(x^*) : \quad \lambda_i > 0 \}, \\ I_0 \;&:=\; \{ i \in I(x^*) : \quad \lambda_i = 0 \}, \\ I_{0,<} \;&:=\; \{ i \in I_0 : \quad \nabla g_i(x^*)^{\mathrm{T}} y < 0 \}, \\ I_{0,0} \;&:=\; \{ i \in I_0 : \quad \nabla g_i(x^*)^{\mathrm{T}} y = 0 \}. \end{aligned}$$

Da $y \in T_X^1(x^*,\lambda)$, folgt $\nabla g_i(x)^{\mathrm{T}} y = 0$ für alle Indizes $i \in I_> \cup I_{0,0}$. Ferner sind die Gradienten $\nabla g_i(x^*),\ i \in I_> \cup I_{0,0}$ und $\nabla h_j(x^*),\ j \in E$ aufgrund der vorausgesetzten LICQ linear unabhängig.

Nach Satz (12.7) existiert nun ein C^2–Weg $x :]-\varepsilon, \varepsilon[\to \mathbb{R}^n$ mit den Eigenschaften

$$x(0) \;=\; x^*, \quad x'(0) \;=\; y, \quad g_i(x(t)) \;\equiv\; 0 \,(i \in I_> \cup I_{0,0}), \quad h_j(x(t)) \;\equiv\; 0 \,(j \in E).$$

Nach Konstruktion gilt ferner $g_i(x^*) < 0,\ i \in I \setminus I(x^*)$, sowie $g_i(x^*) = 0$ und $\nabla g_i(x^*)^{\mathrm{T}} y \;<\; 0$ für $i \in I(x^*) \setminus (I_> \cup I_{0,0})$. Mit dem Taylorschen Satz folgt nach ev. Verkleinerung von ε, dass $x(t) \in X$ zulässig ist für alle $0 \leq t < \varepsilon$.

Wir betrachten nun $\varphi(t) := L(x(t), \lambda, \mu)$. Aufgrund der Komplementarität und der obigen Konstruktion folgt

$$\varphi(t) \;=\; f(x(t)) \;+\; \sum_{i \in I} \lambda_i \, g_i(x(t)) \;+\; \sum_{j \in E} \mu_j \, h_j(x(t)) \;=\; f(x(t)).$$

Damit hat φ in $t = 0$ ein Minimum bezogen auf das Intervall $[0, \varepsilon[$. Man findet nun

$$\varphi'(t) \;=\; \nabla_x L(x(t), \lambda, \mu)^{\mathrm{T}} x'(t) \quad \Rightarrow \quad \varphi'(0) \;=\; 0$$

sowie

$$\varphi''(t) \;=\; \nabla_x L(x(t), \lambda, \mu)^{\mathrm{T}} x''(t) \;+\; x'(t)^{\mathrm{T}} \nabla^2_{xx} L(x(t), \lambda, \mu) x'(t)$$

$$\Rightarrow \qquad \varphi''(0) \;=\; y^{\mathrm{T}} \nabla^2_{xx} L(x^*, \lambda, \mu)\, y \;\geq\; 0. \qquad \square$$

Bemerkung (13.34)

Aus obigem Satz folgt insbesondere, dass in einem lokalen Minimum unter der obigen Voraussetzung auch die folgende notwendige Bedingung erfüllt ist

$$\forall\, y \in T^a_X(x^*): \quad y^{\mathrm{T}}\, \nabla^2_{xx} L(x^*, \lambda, \mu)\, y \;\geq\; 0.$$

Man beachte jedoch, dass diese Bedingung i. Allg. schwächer ist, als die in Satz (13.33) formulierte.

Beispiel (13.35)

$$\begin{aligned} &\text{Minimiere} \quad f(x_1, x_2, x_3) \;=\; -x_1^2 - x_2^2 + x_3^2 \\ &\text{unter den Nebenbedingungen} \\ &g_1(x) \;:=\; -x_1 \;\leq\; 0, \\ &g_2(x) \;:=\; -x_2 \;\leq\; 0. \end{aligned}$$

Die Lagrange-Funktion lautet $L = -x_1^2 - x_2^2 + x_3^2 - \lambda_1 x_1 - \lambda_2 x_2$. Die benötigten Gradienten sind

$$\nabla g_1 \;=\; \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \qquad \nabla g_2 \;=\; \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \qquad \nabla_x L \;=\; \begin{pmatrix} -2x_1 - \lambda_1 \\ -2x_2 - \lambda_2 \\ 2x_3 \end{pmatrix}.$$

Sei $x^* = 0,\ \lambda = (0,\, 0)^{\mathrm{T}}$. Damit ist $I(x^*) = \{1,\, 2\},\ \nabla^2_{xx} L \;=\; \mathrm{diag}(-2, -2, 2)$.

Für den Tangentialraum der aktiven Nebenbedingungen ergibt sich

$$T^a_X(x^*) \;=\; \{y:\ \nabla g_1^{\mathrm{T}} y \;=\; \nabla g_2^{\mathrm{T}} y \;=\} \;=\; \mathbb{R}\, e_3$$

Damit ist $\nabla^2_{xx} L$ positiv definit auf T^a_X, die notwendige Bedingung (13.34) ist also erfüllt!

Betrachtet man dagegen den reduzierten Tangentialkegel,

$$T^1_X(x^*) \;=\; \{y:\ \nabla g_1^{\mathrm{T}} y \le 0 \ \wedge\ \nabla g_2^{\mathrm{T}} y \le 0\} \;=\; \{y:\ y_1 \ge 0 \ \wedge\ y_2 \ge 0\},$$

so stellt man fest, dass die notwendigen Bedingung (13.33) *nicht* erfüllt ist! - Natürlich ist $x^* = 0$ kein lokales Minimum von f auf X.

Satz (13.36) (Hinreichende Bedingung)

Ist (x^*, λ, μ) ein KKT-Punkt der restringierten Optimierungsaufgabe (13.1) und gilt

$$\forall\, y \in T^1_X(x^*, \lambda) \setminus \{0\}: \quad y^{\mathrm{T}}\, \nabla^2_{xx}\, L(x^*, \lambda, \mu)\, y \;>\; 0,$$

so ist x^* ein striktes lokales Minimum von f auf X.

Beweis:

Wir führen den Beweis indirekt und nehmen an, dass x^* kein striktes lokales Minimum ist, also

$$\exists x^k \in X,\ x^k \ne x^*:\ x^k \to x^*\ (k \to \infty)\ \wedge\ f(x^k) \le f(x^*).$$

Durch Übergang auf eine Teilfolge erreicht man

$$\lim_{k\to\infty} \frac{x^k - x^*}{\|x^k - x^*\|} \;=\; \lim_{k\to\infty} y^k \;=:\; y, \qquad \|y\| \;=\; 1.$$

Da g_i und h_j C^1–Funktionen sind folgt weiter

$$\lim_{k\to\infty} \frac{g_i(x^k) - g_i(x^*) - \nabla g_i(x^*)^{\mathrm{T}}(x^k - x^*)}{\|x^k - x^*\|} \;=\; 0$$

$$\Rightarrow \qquad \forall\, i \in I(x^*): \quad \nabla g_i(x^*)^{\mathrm{T}} y \;\le\; 0.$$

Genauso zeigt man

$$\forall\, j \in E: \quad \nabla h_j(x^*)^{\mathrm{T}} y \;=\; 0$$

und schließlich wegen $f(x^k) \le f(x^*)$ bei gleicher Herleitung $\nabla f(x^*)^{\mathrm{T}} y \le 0$.

1. Fall: $\forall\, i \in I_>:\ \nabla g_i(x^*)^{\mathrm{T}} y \;=\; 0.$

In diesem Fall ist $y \in T^1_X(x^*, \lambda)$, vgl. (13.31). Mittels Taylor-Entwicklung folgt

$$\begin{aligned} f(x^*) \;\ge\; f(x^k) \;&\ge\; f(x^k) \;+\; \sum_i \lambda_i g_i(x^k) \;+\; \sum_j \mu_j h_j(x^k) \;=\; L(x^k, \lambda, \mu) \\ &=\; L(x^*, \lambda, \mu) \;+\; \nabla_x L^{\mathrm{T}}(x^k - x^*) \;+\; \frac{1}{2}\,(x^k - x^*)^{\mathrm{T}} \nabla^2_{xx} L(\xi^k, \lambda, \mu)(x^k - x^*) \\ &=\; f(x^*) \;+\; \frac{1}{2}\,(x^k - x^*)^{\mathrm{T}} \nabla^2_{xx} L(\xi^k, \lambda, \mu)(x^k - x^*) \end{aligned}$$

wobei wieder $\xi^k = x^k + \Theta_k(x^k - x^*),\ 0 < \Theta_k < 1$.

Damit folgt $0 \geq (x^k - x^*)^{\mathrm{T}} \nabla^2_{xx} L(\xi^k, \lambda, \mu)(x^k - x^*)$. Division durch $\|x^k - x^*\|^2$ und Limesbildung $k \to \infty$ ergibt

$$0 \geq y^{\mathrm{T}} \nabla^2_{xx} L(x^*, \lambda, \mu)\, y$$

im Widerspruch zur Voraussetzung.

2. Fall: $\exists\, i_0 \in I_> :\ \nabla g_{i_0}(x^*)^{\mathrm{T}} y < 0.$

Aus $\nabla_x L(x^*, \lambda, \mu) = 0,\ \lambda_i = 0\ (i \in I_0)$ und $\nabla g_i(x^*)^{\mathrm{T}} y \leq 0\ (i \in I_>)$ folgt

$$\begin{aligned} 0 &\geq \nabla f(x^*)^{\mathrm{T}} y \\ &= -\sum_{i \in I} \lambda_i \nabla g_i(x^*)^{\mathrm{T}} y - \sum_{j \in E} \mu_j \nabla h_j(x^*)^{\mathrm{T}} y \\ &= -\sum_{i \in I_>} \lambda_i \nabla g_i(x^*)^{\mathrm{T}} y \\ &\geq -\lambda_{i_0} \nabla g_{i_0}(x^*)^{\mathrm{T}} y > 0. \qquad \text{Widerspruch!} \end{aligned}$$

□

Aufgabe (13.37)

Gegeben sei eine konvexe Optimierungsaufgabe

$$\text{Minim. } f(x),\ x \in \mathbb{R}^n, \quad \text{NB. } g(x) \leq 0, \quad h(x) = Bx - q = 0,$$

mit $f, g_i : \mathbb{R}^n \to \mathbb{R}$ konvex, $i \in I = \{1, \ldots, m\}$, und $B \in \mathbb{R}^{(p,n)},\ B = (b^1, \ldots, b^p)^{\mathrm{T}}$, $E = \{1, \ldots, p\}$.

Für einen zulässigen Punkt x dieser Optimierungsaufgabe wird der strikte Tangentialkegel definiert durch

$$T_{\text{strikt}}(x) := \{d \in \mathbb{R}^n :\ \nabla g_i(x)^{\mathrm{T}} d < 0\ (i \in I(x)),\ \nabla h_j(x)^{\mathrm{T}} d = 0\ (j \in E)\}$$

Zeigen Sie: $T_{\text{strikt}}(x) \subset T_X(x)$. (Diese Aussage vervollständigt den Beweis von (13.29).)

Aufgabe (13.38)

Untersuchen Sie die folgende Optimierungsaufgabe auf lokale Minima:

$$\begin{aligned} \text{Minimiere} \quad f(x_1, x_2) &:= -0.1\,(x_1 - 4)^2 + x_2^2 \\ \text{NB.:} \qquad\quad g(x_1, x_2) &:= 1 - x_1^2 - x_2^2 \leq 0. \end{aligned}$$

Überprüfen Sie insbesondere die KKT–Bedingungen sowie die notwendigen und hinreichenden Bedingungen zweiter Ordnung.

14. Quadratische Programme, Strategie der aktiven Menge

A. Gleichungsrestriktionen.

Wir betrachten in diesem Abschnitt quadratische Optimierungsprobleme (*Quadratische Programme*), zunächst nur mit Gleichungsrestriktionen. Die Problemstellung lautet:

$$\begin{aligned} &\text{Minimiere} \quad f(x) \;=\; \frac{1}{2}\, x^{\mathrm{T}} Q\, x \;+\; c^{\mathrm{T}} x \;+\; \gamma, \quad x \in \mathbb{R}^n, \\ &\text{unter den Nebenbedingungen} \\ &h_j(x) \;:=\; (b^j)^{\mathrm{T}} x \;-\; q_j \;=\; 0, \quad j \in E \;:=\; \{1, \ldots, p\}. \end{aligned} \tag{14.1}$$

Dabei sei $Q \in \mathbb{R}^{(n,n)}$ eine *symmetrische* Matrix, $c, b^j \in \mathbb{R}^n$ und $\gamma, q_j \in \mathbb{R}$.

Die Nebenbedingungen lassen sich zusammenfassen zu

$$\begin{aligned} &h(x) \;:=\; B\, x \;-\; q \;=\; 0, \\ &B \;:=\; \begin{pmatrix} (b^1)^{\mathrm{T}} \\ \vdots \\ (b^p)^{\mathrm{T}} \end{pmatrix} \;\in\; \mathbb{R}^{(p,n)}, \quad q \;:=\; \begin{pmatrix} q_1 \\ \vdots \\ q_p \end{pmatrix} \;\in\; \mathbb{R}^p. \end{aligned} \tag{14.2}$$

Ist $x = x^*$ ein lokales Minimum von (14.1), so ergeben sich mit Satz (13.19) die folgenden notwendigen Bedingungen: Es existieren Lagrange-Multiplikatoren $\mu \in \mathbb{R}^p$ mit

$$\begin{aligned} Q\, x \;+\; c \;+\; \sum_{j=1}^{p} \mu_j\, b^j \;&=\; 0, \\ (b^j)^{\mathrm{T}} x \;-\; q_j \;&=\; 0 \quad (j \in E). \end{aligned}$$

Dieses Gleichungssystem lässt sich in Matrixschreibweise wie folgt darstellen

$$\begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ \mu \end{pmatrix} \;=\; \begin{pmatrix} -c \\ q \end{pmatrix} \tag{14.3}$$

Bemerkung (14.4)

Ist die Matrix Q positiv semidefinit, f also konvex auf $\mathbb{R}^n$, so liefert umgekehrt jede Lösung des Gleichungssystems (14.3) ein globales Minimum der quadratischen Optimierungsaufgabe (14.1); vgl. (13.30).

Wir formen das Gleichungssystem (14.3) weiter um und setzen $x = x^0 + \Delta x$, wobei $x^0 \in \mathbb{R}^n$ ein beliebiger zulässiger Punkt sei, also $B\, x^0 - q = 0$ gelte.

$$(14.3) \quad \Leftrightarrow \quad \begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} x^0 + \Delta x \\ \mu \end{pmatrix} = \begin{pmatrix} -c \\ q \end{pmatrix}$$

$$\Leftrightarrow \quad \begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \mu \end{pmatrix} = \begin{pmatrix} -c \\ q \end{pmatrix} - \begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} x^0 \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} -c - Q\,x^0 \\ q - B\,x^0 \end{pmatrix} = \begin{pmatrix} -\nabla f(x^0) \\ 0 \end{pmatrix}$$

Damit ist der folgende Satz gezeigt

Satz (14.5)

Ist $x^0 \in \mathbb{R}^n$ ein zulässiger Punkt für das quadratische Optimierungsproblem (14.1), so genügen $x = x^0 + \Delta x$ und $\mu \in \mathbb{R}^p$ genau dann den KKT-Bedingungen, wenn $(\Delta x, \mu)$ die folgende, so genannte **Lagrange-Newton-Gleichung** löst

$$\begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \mu \end{pmatrix} = \begin{pmatrix} -\nabla f(x^0) \\ 0 \end{pmatrix}. \tag{14.6}$$

Satz (14.7)

Hat die Matrix B maximalen Rang, d.h. sind die $b^1, \ldots b^p$ linear unabhängig, und ist die Hessematrix $Q = \nabla^2 f$ auf dem Tangentialraum $T_X = \mathrm{Kern}(B)$ positiv definit, so ist die Koeffizientenmatrix in der Lagrange-Newton-Gleichung (14.6) regulär.

Beweis: Wir zeigen, dass der Kern der obigen Koeffizientenmatrix nur aus dem Nullvektor besteht. Gelte also

$$\begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Wäre $u = 0$, so folgt aus der ersten Gleichung und der vorausgesetzten Regularität ($\mathrm{Rang}(B) = p$), dass auch $v = 0$ ist.

Nehmen wir an, dass $u \neq 0$ gelte. Aus der zweiten Gleichung ergibt sich dann $u \in T_X \setminus \{0\}$. Multipiziert man das obige lineare Gleichungssystem mit $(u^{\mathrm{T}}, v^{\mathrm{T}})$ von links, so ergibt sich

$$0 \;=\; (u^{\mathrm{T}}, v^{\mathrm{T}}) \begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \;=\; u^{\mathrm{T}}Qu \,+\, (Bu)^{\mathrm{T}}v + v^{\mathrm{T}}Bu \;=\; u^{\mathrm{T}}Qu,$$

im Widerspruch zur vorausgesetzten positiven Definitheit von Q auf dem Tangentialraum T_X. □

B. Allgemeine quadratische Programme.

Wir erweitern die Problemstellung nun durch lineare Ungleichungsnebenbedingungen, betrachten also lineare Programme der Form

$$\begin{aligned}
&\text{Minimiere} \quad f(x) \;=\; \frac{1}{2}\, x^{\mathrm{T}} Q\, x \;+\; c^{\mathrm{T}} x \;+\; \gamma, \quad x \in \mathbb{R}^n, \\
&\text{unter den Nebenbedingungen} \\
&g_i(x) \;:=\; (a^i)^{\mathrm{T}} x \;-\; p_i \;\leq\; 0, \quad i \in I \;:=\; \{1, \ldots, m\}, \\
&h_j(x) \;:=\; (b^j)^{\mathrm{T}} x \;-\; q_j \;=\; 0, \quad j \in E \;:=\; \{1, \ldots, p\}.
\end{aligned} \tag{14.8}$$

Zur numerischen Lösung soll Satz (14.5) für nur gleichungsrestringierte lineare Programme verwendet werden. Dies wäre möglich, falls die Menge $I(x^*)$ der aktiven Indizes einer Lösung x^* von (14.8) bekannt wäre. Man könnte dann anstelle von (14.8) das folgende Ersatzproblem lösen:

Ersatzproblem (14.9)

$$\begin{aligned}
&\text{Minimiere} \quad f(x) \;=\; \frac{1}{2}\, x^{\mathrm{T}} Q\, x \;+\; c^{\mathrm{T}} x \;+\; \gamma, \quad x \in \mathbb{R}^n, \\
&\text{unter den Nebenbedingungen} \\
&g_i(x) \;:=\; (a^i)^{\mathrm{T}} x \;-\; p_i \;=\; 0, \quad i \in I(x^*), \\
&h_j(x) \;:=\; (b^j)^{\mathrm{T}} x \;-\; q_j \;=\; 0, \quad j \in E \;:=\; \{1, \ldots, p\}.
\end{aligned}$$

Die Grundidee der **Strategie der aktiven Menge** besteht nun darin, die unbekannte Indexmenge $I(x^*)$ durch eine Schätzung $I_a \subset I = \{1, \ldots, m\}$ zu ersetzen. Diese Schätzung I_a wird dann während des Lösungsprozesses angepasst.

Bezeichne also $x = x^k$ eine Näherung der gesuchten Lösung x^* und k einen Iterationsindex. Ferner sei $I_a = I_a^k \subset I$ eine Näherung der zugehörigen Menge aktiver Indizes $I(x^k)$, kurz die **aktuelle aktive Menge**. Wir nehmen an, dass x zulässig ist für das Ausgangsproblem (14.8) und die aktuellen Gleichungsnebenbedingungen $g_i(x) = 0,\ i \in I_a,$ und $h_j(x) = 0,\ j \in E$ erfüllt sind.

Anstelle von (14.9) lösen wir dann das Hilfsproblem

$$\begin{aligned}
&\text{Minimiere} \quad f(x + \Delta x), \quad \Delta x \in \mathbb{R}^n, \\
&\text{unter den Nebenbedingungen} \\
&(a^i)^{\mathrm{T}} \Delta x \;=\; 0, \quad i \in I_a, \\
&(b^j)^{\mathrm{T}} \Delta x \;=\; 0, \quad j \in E.
\end{aligned} \tag{14.10}$$

Die Lösung erfolgt dabei über die Lagrange-Newton-Gleichung, vgl. (14.6)

$$\begin{pmatrix} Q & A_k^{\mathrm{T}} & B^{\mathrm{T}} \\ A_k & 0 & 0 \\ B & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x^k \\ \lambda^k \\ \mu^k \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) \\ 0 \\ 0 \end{pmatrix}. \tag{14.11}$$

Dabei ist die Matrix A_k durch die aktuelle aktive Menge bestimmt: $A_k = ((a^i)^{\mathrm{T}})_{i \in I_a^k}$.

Nach Satz (14.7) ist dieses lineare Gleichungssystem eindeutig lösbar, falls die Vektoren a^i, $i \in I_a^k$, und $b^j, j \in E$, linear unabhängig sind und Q auf dem zugehörigen aktuellen Tangentialraum Kern$\begin{pmatrix} A_k \\ B \end{pmatrix}$ positiv definit ist.

Die Strategie zur Anpassung der aktiven Menge verwendet nun die Lösung $\Delta x = \Delta x^k$, $\lambda = \lambda^k$ und $\mu = \mu^k$ von (14.11):

Fall A. Liefert (14.11) die Lösung $\Delta x = 0$, so sind die KKT-Bedingungen für das Hilfsproblem

$$\nabla f(x) \; + \; \sum_{i \in I_a} \lambda_i \, a^i \; + \; \sum_{j \in E} \mu_j \, b^j \; = \; 0$$

erfüllt.

(i) Falls nun $\lambda_i \geq 0$ für alle $i \in I_a$, so wird x als Lösung von (14.8) akzeptiert. Setzt man nämlich $\lambda_i := 0$ für alle $i \in I \setminus I_a$, so sind die KKT-Bedingungen für das Ausgangsproblem (14.8) erfüllt.

(ii) *Deaktivierungsschritt:* Gilt dagegen $\lambda_i < 0$ für einen Index $i \in I_a$, so sind die KKT-Bedingungen nicht erfüllt (Verletzung der Vorzeichenbedingung). Andererseits ist x ja ein Minimum von f unter den Gleichungsnebenbedingungen mit den Indizes $I_a \cup E$. Die aktuellen aktiven Nebenbedingungen müssen daher abgeschwächt werden.

Man bestimmt daher einen Index $\ell \in I_a$ mit $\lambda_\ell = \min\{\lambda_i : \, i \in I_a\} < 0$ und setzt $I_a^{k+1} := I_a \setminus \{\ell\}$ sowie $x^{k+1} := x^k$ (zulässig).

Fall B. Verschwindet dagegen die Lösung von (14.11) nicht, so ist Δx eine Abstiegsrichtung von f in $x = x^k$, die die aktuellen Nebenbedingungen mit den Indizes $I_a \cup E$ nicht verletzt, d.h. es gilt gemäß der letzen beiden Gleichungen von (14.11)

$$(a^i)^{\mathrm{T}} \Delta x \; = \; 0, \quad (b^j)^{\mathrm{T}} \Delta x \; = \; 0 \quad (i \in I_a, \, j \in E).$$

(i) Ist $x + \Delta x$ sogar zulässig für das Ausgangsproblem, d.h. gilt

$$(a^i)^{\mathrm{T}}(x + \Delta x) \; - \; p_i \; \leq \; 0 \quad (\forall \, i \in I \setminus I_a),$$

so setzt man $x^{k+1} := x + \Delta x$ und $I_a^{k+1} := I_a$.

(ii) Ist dagegen $x + \Delta x$ nicht zulässig, so bestimmt man eine möglichst große Schrittweite $t \geq 0$, so dass $x + t\,\Delta x$ zulässig ist, dass also gelten

$$\forall\, i \in I \setminus I_a : \quad (a^i)^{\mathrm{T}}(x + t\,\Delta x) \;-\; p_i \;\leq\; 0.$$

Da diese Bedingung für $t = 0$ nach Voraussetzung erfüllt ist, für $t = 1$ jedoch verletzt ist, muss es einen Index $i \in I \setminus I_a$ geben mit $(a^i)^{\mathrm{T}}\Delta x > 0$.

Daher ist die folgende Wahl der Schrittweite t wohldefiniert

$$t \;:=\; \min\,\{\, \frac{p_i - (a^i)^{\mathrm{T}}x}{(a^i)^{\mathrm{T}}\Delta x} \;:\; i \in I \setminus I_a,\; (a^i)^{\mathrm{T}}\Delta x \;>\; 0 \,\}. \tag{14.12}$$

Gleichzeitig bestimme man einen Index $\ell \in I \setminus I_a$ für den das obige Minimum angenommen wird. Der Index ℓ muss nicht eindeutig bestimmt sein. Nach Konstruktion ist $t \geq 0$, allerdings kann $t = 0$ nicht ausgeschlossen werden. Ferner ist nach Konstruktion $x^{k+1} := x + t\,\Delta x$ zulässig für das Ausgangsproblem und wird als neue Iterierte gewählt. Die Nebenbedingung g_ℓ wird in der neuen Iterierten aktiv, daher setzen wir $I_a^{k+1} := I_a \cup \{\ell\}$.

Zusammengefasst ergibt sich nun der folgende Modellalgorithmus zur Strategie der aktiven Menge:

Algorithmus (14.13)

1.) Start: $x = x^0 \in X$ zulässig!; $k := 0$; $I_a = I_a^0 := \{i \in I : (a^i)^{\mathrm{T}}x = p_i\}$;

2.) Bestimme $(\Delta x, \lambda, \mu)$ aus der Lagrange-Newton-Gleichung (14.11) und ergänze λ durch $\lambda_i := 0$, $i \notin I_a$.

3.) Ist $\Delta x = 0$ und $\lambda \geq 0$: Stop!

4.) Ist $\Delta x = 0$ und $\lambda_\ell := \min\{\lambda_i : i \in I_a\} < 0$:
$x := x^{k+1} := x^k$; $I_a := I_a^{k+1} := I_a \setminus \{\ell\}$; $k := k + 1$; gehe zu 2.)

5.) Falls $\forall i \in I \setminus I_a$: $(a^i)^{\mathrm{T}}(x + \Delta x) - p_i \leq 0$:
$x := x^{k+1} := x + \Delta x$; $I_a^{k+1} := I_a$; $k := k + 1$; gehe zu 2.)

6.) Bestimme $\ell \in I \setminus I_a$ mit

$$t \;:=\; \frac{p_\ell - (a^\ell)^{\mathrm{T}}x}{(a^\ell)^{\mathrm{T}}\Delta x} \;=\; \min\,\{\, \frac{p_i - (a^i)^{\mathrm{T}}x}{(a^i)^{\mathrm{T}}\Delta x} \;:\; i \in I \setminus I_a,\; (a^i)^{\mathrm{T}}\Delta x \;>\; 0 \,\}$$

und setze: $x := x^{k+1} := x + t\,\Delta x$; $I_a := I_a^{k+1} := I_a \cup \{\ell\}$, $k := k + 1$; gehe zu 2.)

Bemerkungen (14.14)

a) Ist die Matrix Q positiv definit und sind die Vektoren a^i, $i \in I_a^0$ und b^j, $j \in E$ linear unabhängig, so ist der Algorithmus (14.13) wohldefiniert, d.h. solange kein Abbruch erfolgt sind auch die Vektoren a^i, $i \in I_a^k$ und b^j, $j \in E$ linear unabhängig und die Lagrange-Newton-Gleichungen sind eindeutig lösbar. Der Algorithmus liefert dann nach endlich vielen Iterationen die eindeutig bestimmte Lösung von (14.8).

b) Zur Bestimmung eines zulässigen Startvektors $x^0 \in X$ kann man analog zur ersten Phase des Simplex-Algorithmus vorgehen. Dazu löst man das folgende **Hilfsproblem**:

$$\begin{aligned} &\text{Minimiere} \quad \widetilde{f}(x,y,z) \;=\; \frac{1}{2}\sum_{i\in I} y_i^2 \;+\; \frac{1}{2}\sum_{j\in E} z_j^2, \\ &\text{unter den Nebenbedingungen} \\ &(a^i)^{\mathrm{T}} x \;+\; y_i \;-\; p_i \;\le\; 0, \quad i \in I, \\ &(b^j)^{\mathrm{T}} x \;+\; z_j \;-\; q_j \;=\; 0, \quad j \in E. \end{aligned} \tag{14.15}$$

Für dieses quadratische Optimierungsproblem ist $(x^0, y^0, z^0) := (0, p, q)$ ein zulässiger Startvektor mit der aktiven Menge $I_a^0 := I$.

Besitzt das Ausgangsproblem einen zulässigen Punkt x, so hat das Hilfsproblem die - im Allg. nicht eindeutig bestimmte - Lösung $(x, 0, 0)$.

Beispiel (14.16)

$$\begin{aligned} &\text{Minimiere} \quad f(x_1,x_2) \;=\; \frac{1}{2}(x_1^2 + x_2^2) \\ &\text{unter den Nebenbedingungen} \\ &x_1 \;\le\; 2,\;\; x_2 \;\le\; 1,\;\; x_1 + 3\,x_2 \;\ge\; 2. \end{aligned}$$

Die Daten des Problems lauten also

$$\begin{aligned} &Q \;=\; I_2, \quad c \;=\; 0, \quad \gamma \;=\; 0, \quad \nabla f(x) \;=\; x, \\ &g_1 \;=\; (1,0)\,x \;-\; 2 \;\le\; 0, \\ &g_2 \;=\; (0,1)\,x \;-\; 1 \;\le\; 0, \\ &g_3 \;=\; (-1,-3)\,x \;+\; 2 \;\le\; 0. \end{aligned}$$

Start: Wir starten mit dem zulässigen Punkt $x^0 = (2,1)^{\mathrm{T}}$, $I_a = \{1,2\}$, $f(x^0) = 5/2$. Die Lagrange-Newton-Gleichung lautet

$$\left(\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right) \left(\begin{array}{c} \Delta x_1 \\ \Delta x_2 \\ \lambda_1 \\ \lambda_2 \end{array}\right) \;=\; \left(\begin{array}{c} -2 \\ -1 \\ 0 \\ 0 \end{array}\right).$$

Man erhält $\Delta x_1 = \Delta x_2 = 0,\ \lambda_1 = -2,\ \lambda_2 = -1$. Damit folgt nach Punkt 4.): $\Delta x = 0,\ \ell = 1$.

1. Iteration: $x^1 = (2,1)^{\mathrm{T}}\ \ I_a = \{2\},\ \ f(x^1) = 5/2$. Die Lagrange-Newton-Gleichung lautet

$$\left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 1 \\ \hline 0 & 1 & 0 \end{array}\right) \left(\begin{array}{c} \Delta x_1 \\ \Delta x_2 \\ \lambda_2 \end{array}\right) = \left(\begin{array}{c} -2 \\ -1 \\ 0 \end{array}\right).$$

Damit ist $\Delta x_1 = -2,\ \Delta x_2 = 0,\ \lambda_2 = -1$. Punkt 5.) des Algorithmus wird aktiv, denn $x + \Delta x = (0,1)^{\mathrm{T}}$ ist zulässig!

2. Iteration: $x^2 = (0,1)^{\mathrm{T}}\ \ I_a = \{2\},\ \ f(x^2) = 1/2$. Die Lagrange-Newton-Gleichung lautet

$$\left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 1 \\ \hline 0 & 1 & 0 \end{array}\right) \left(\begin{array}{c} \Delta x_1 \\ \Delta x_2 \\ \lambda_2 \end{array}\right) = \left(\begin{array}{c} 0 \\ -1 \\ 0 \end{array}\right).$$

Damit ist $\Delta x_1 = \Delta x_2 = 0,\ \lambda_2 = -1$. Im Algorithmus wird wieder Punkt 4.) aktiv mit $\Delta x = 0,\ \ell = 2$.

3. Iteration: $x^3 = (0,1)^{\mathrm{T}}\ \ I_a = \emptyset,\ \ f(x^3) = 1/2$. Die Lagrange-Newton-Gleichung lautet

$$\left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right) \left(\begin{array}{c} \Delta x_1 \\ \Delta x_2 \end{array}\right) = \left(\begin{array}{c} 0 \\ -1 \end{array}\right).$$

Somit ist $\Delta x_1 = 0,\ \Delta x_2 = -1$. Im Algorithmus wird Punkt 6.) aktiv, da $x + \Delta x = 0$ nicht zulässig ist. Die Auswertung der Quotienten ergibt $t = 1/3$ und $\ell = 3$. Damit wird

4. Iteration: $x^4 = (0,2/3)^{\mathrm{T}}\ \ I_a = \{3\},\ \ f(x^4) = 2/9$. Die Lagrange-Newton-Gleichung lautet

$$\left(\begin{array}{cc|c} 1 & 0 & -1 \\ 0 & 1 & -3 \\ \hline -1 & -3 & 0 \end{array}\right) \left(\begin{array}{c} \Delta x_1 \\ \Delta x_2 \\ \lambda_3 \end{array}\right) = \left(\begin{array}{c} 0 \\ -2/3 \\ 0 \end{array}\right).$$

Es ergibt sich die Lösung $\Delta x_1 = 1/5,\ \Delta x_2 = -1/15$ und $\lambda_3 = 1/5$. Man stellt fest, dass $\Delta x \neq 0$ und $x + \Delta x = (1/5, 3/5)^{\mathrm{T}}$ zulässig ist.

5. Iteration: $x^5 = (1/5, 3/5)^{\mathrm{T}}\ \ I_a = \{3\},\ \ f(x^5) = 1/5$. Die Lagrange-Newton-Gleichung lautet

$$\left(\begin{array}{cc|c} 1 & 0 & -1 \\ 0 & 1 & -3 \\ \hline -1 & -3 & 0 \end{array}\right) \left(\begin{array}{c} \Delta x_1 \\ \Delta x_2 \\ \lambda_3 \end{array}\right) = \left(\begin{array}{c} -1/5 \\ -3/5 \\ 0 \end{array}\right).$$

Die Lösung ist $\Delta x = 0$ und $\lambda_3 = 1/5 > 0$. Damit ist Punkt 3.) des Verfahrens erfüllt (**Lösung gefunden!**) und x^5 ist das eindeutig bestimmte globale Minimum des Optimierungsproblems.

15. Lagrange-Newton-Iteration, SQP Verfahren

A. Die Lagrange-Newton-Iteration.

Wir betrachten eine allgemeine gleichungsrestringierte Optimierungsaufgabe:

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n, \\ &\text{Nebenbedingungen:} \quad h_j(x) \;=\; 0, \quad j \in E := \{1, \ldots, p\}. \end{aligned} \tag{15.1}$$

Wir setzen dabei wieder voraus, dass $f,\ h_j \in \mathrm{C}^2(\mathbb{R}^n, \mathbb{R})$ gilt.

Unter der **Lagrange-Newton-Iteration** versteht man die direkte Lösung der KKT-Bedingungen zu Problem (15.1) – aufgefasst als ein im Allg. nichtlineares Gleichungssystem – mit dem Newton-Verfahren.

Mit der Lagrange-Funktion

$$L(x, \lambda) \;:=\; f(x) \;+\; \sum_{j \in E} \lambda_j \, h_j(x) \;=\; f(x) \;+\; \lambda^{\mathrm{T}} h(x) \tag{15.2}$$

lauten die KKT-Bedingungen

$$\nabla L(x, \lambda) \;=\; \begin{pmatrix} \nabla f(x) \;+\; \sum_{j \in E} \lambda_j \, \nabla h_j(x) \\ h(x) \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{15.3}$$

Man beachte, dass die Gradientenbildung der Lagrange-Funktion bezüglich *beider* Variablen (x, λ) erfolgt.

(15.3) ist ein im Allg. nichtlineares Gleichungssystem ($n + p$ Gleichungen) für die Unbekannten (x, λ). Das Newton-Verfahren hierzu lautet

$$\nabla^2 L(x, \lambda) \begin{pmatrix} \Delta x \\ \Delta \lambda \end{pmatrix} \;=\; -\nabla L(x, \lambda), \qquad \begin{pmatrix} x^+ \\ \lambda^+ \end{pmatrix} \;:=\; \begin{pmatrix} x \;+\; \Delta x \\ \lambda \;+\; \Delta \lambda \end{pmatrix}, \tag{15.4}$$

wobei die Hesse-Matrix die folgende Struktur besitzt

$$\begin{aligned} \nabla^2 L(x, \lambda) \;&=\; \begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \;\in\; \mathbb{R}^{(n+p, n+p)} \\ Q \;&=\; \nabla^2 f(x) \;+\; \textstyle\sum_{j \in E} \lambda_j \nabla^2 h_j(x) \;=\; \nabla^2_{x\,x} L(x, \lambda), \\ B \;&=\; h'(x) \;=\; \begin{pmatrix} \nabla h_1(x)^{\mathrm{T}} \\ \vdots \\ \nabla h_p(x)^{\mathrm{T}} \end{pmatrix}. \end{aligned} \tag{15.5}$$

Man beachte hierzu

$$\frac{\partial}{\partial x_k}\left[\nabla f(x) + \sum_j \lambda_j \nabla h_j(x)\right]_i = \frac{\partial^2 L}{\partial x_k \, \partial x_i}(x,\lambda)$$

$$\frac{\partial}{\partial \lambda_k}\left[\nabla f(x) + \sum_j \lambda_j \nabla h_j(x)\right]_i = \frac{\partial h_k}{\partial x_i}(x)$$

Unter Verwendung von (15.3) und (15.5) lässt sich die Newton-Gleichung weiter umformen:

$$\begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x) - B^{\mathrm{T}}\lambda \\ -h(x) \end{pmatrix}$$

oder mit $\lambda^+ = \lambda + \Delta\lambda$:

$$\begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \lambda^+ \end{pmatrix} = - \begin{pmatrix} \nabla f(x) \\ h(x) \end{pmatrix} \tag{15.6}$$

Damit haben wir den folgenden Algorithmus hergeleitet

Algorithmus (15.7) (Lagrange-Newton-Verfahren)

1.) Wähle $(x^0, \lambda^0)^{\mathrm{T}} \in \mathbb{R}^{(n+p)}$, $\mathrm{TOL} > 0$, $k := 0$;

2.) Berechne $\nabla f(x^k)$, $B := h'(x^k)$ und $F^k := \begin{pmatrix} \nabla f(x^k) + B^{\mathrm{T}}\lambda^k \\ h(x^k) \end{pmatrix}$

Falls $\|F^k\| < \mathrm{TOL}$: Stop!

Berechne $Q := \nabla^2 f(x^k) + \sum_{j\in E} \lambda_j^k \, \nabla^2 h_j(x^k)$.

3.) Löse das lineare Gleichungssystem

$$\begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \lambda^{k+1} \end{pmatrix} = - \begin{pmatrix} \nabla f(x^k) \\ h(x^k) \end{pmatrix}$$

4.) Setze $x^{k+1} := x^k + \Delta x$, $k := k+1$; gehe zu 2.)

Satz (15.8)

Hat $B = h'(x) \in \mathbb{R}^{(p,n)}$ für ein $x \in \mathbb{R}^n$ den Rang p und ist $Q = \nabla^2_{x\,x} L(x,\lambda)$ für ein $\lambda \in \mathbb{R}^p$ positiv definit auf $\mathrm{Kern}(B)$, so ist die Koeffizientenmatrix des linearen Gleichungssystems (15.6) $\nabla^2 L(x,\lambda)$ regulär.

Beweis: Vgl. Satz (14.7).

Bemerkung (15.9)

Die Voraussetzungen des obigen Satzes sind für einen regulären KKT-Punkt (x^*, λ), der den hinreichenden Bedingungen (12.10) genügt, erfüllt.

Satz (15.10)

Sind f, h_j sogar C^3-Funktionen und ist (x^*, λ^*) ein regulärer KKT-Punkt, der den hinreichenden Bedingungen (12.10) genügt, so konvergiert die Lagrange-Newton-Iteration für alle Startwerte (x^0, λ^0), für die $\|x^0 - x^*\|$ hinreichend klein und $\nabla^2 L(x^0, \lambda^0)$ regulär ist. Die Konvergenz ist quadratisch.

Beweis:

Mit $\varepsilon := x - x^*$ und $\delta := \lambda - \lambda^*$ gelten die folgenden Taylor-Entwicklungen

$$\begin{aligned} h(x^*) &= h(x) - B\,\varepsilon + O(\|\varepsilon\|^2), \\ \nabla f(x^*) &= \nabla f(x) - \nabla^2 f(x)\,\varepsilon + O(\|\varepsilon\|^2), \\ \nabla h_j(x^*) &= \nabla h_j(x) - \nabla^2 h_j(x)\,\varepsilon + O(\|\varepsilon\|^2) \quad (j \in E). \end{aligned}$$

Aus der Lagrange-Newton-Gleichung

$$\begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = - \begin{pmatrix} \nabla f(x) + \sum_j \lambda_j \nabla h_j(x) \\ h(x) \end{pmatrix}$$

folgt dann mit $\Delta x = \varepsilon^+ - \varepsilon$ und $\Delta\lambda = \delta^+ - \delta$:

$$\begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} \varepsilon^+ \\ \delta^+ \end{pmatrix} = \begin{pmatrix} Q\,\varepsilon + B^{\mathrm{T}}\delta - \nabla f(x) - \sum_j \lambda_j \nabla h_j(x) \\ B\,\varepsilon - h(x) \end{pmatrix}$$

Die rechte Seite wird nun mittels der obigen Taylor-Entwicklungen umgeformt

$$\begin{aligned} & Q\,\varepsilon + B^{\mathrm{T}}\delta - \nabla f(x) - \textstyle\sum_j \lambda_j \nabla h_j(x) \\ = \quad & \nabla^2 f(x)\,\varepsilon + \textstyle\sum_j \lambda_j \nabla^2 h_j(x)\,\varepsilon + \sum_j \delta_j \nabla h_j(x) - \nabla f(x) - \sum_j \lambda_j \nabla h_j(x) \end{aligned}$$

$$
\begin{aligned}
&= \big(\nabla^2 f(x)\,\varepsilon \;-\; \nabla f(x) \;+\; \nabla f(x^*)\big) \;+\; \big(\textstyle\sum_j \lambda_j^* \nabla h_j(x^*) \;+\; \sum_j (\delta_j - \lambda_j)\nabla h_j(x)\big) \\
&\quad + \;\textstyle\sum_j \lambda_j \nabla^2 h_j(x)\,\varepsilon \\
&= \big(\nabla^2 f(x)\,\varepsilon \;-\; \nabla f(x) \;+\; \nabla f(x^*)\big) \\
&\quad + \;\textstyle\sum_j \lambda_j^* \big(\nabla h_j(x^*) \;-\; \nabla h_j(x) \;+\; \nabla^2 h_j(x)\,\varepsilon\big) \\
&\quad + \;\textstyle\sum_j \delta_j \nabla^2 h_j(x)\,\varepsilon \\
&= O(\|\varepsilon\|^2) \;+\; O(\|\varepsilon\|^2) \;+\; O(\|\delta\|\,\|\varepsilon\|).
\end{aligned}
$$

Analog gilt für die zweite Komponente obiger Gleichung

$$
B\,\varepsilon \;-\; h(x) \;=\; B\,\varepsilon \;-\; h(x) \;+\; h(x^*) \;=\; O(\|\varepsilon\|^2).
$$

Ist nun die Koeffizientenmatrix der Lagrange-Newton-Gleichung für festes λ und hinreichend kleinen kleinem Anfangsfehler $\|x - x^*\| \;=\; \|\varepsilon\|$ regulär und deren Inverse gleichmäßig beschränkt, so folgt

$$
\left\| \begin{pmatrix} \varepsilon^+ \\ \delta^+ \end{pmatrix} \right\| \;=\; O(\|\varepsilon\|^2) \;+\; O(\|\varepsilon\|\,\|\delta\|).
$$

Damit lässt sich auch für große Anfangsfehler $\|\delta\|$ in λ der Anfangsfehler $\|\varepsilon\|$ in x so klein wählen, dass (x^+, λ^+) im Einzugsbereich des Newton-Verfahrens liegt. □

B. Das SQP-Verfahren.

Im Fall einer unrestringierten Optimierungsaufgabe, Minimiere $f(x),\ x \in \mathbb{R}^n,$ ist die Newton-Korrektur $\Delta x = -\,[\nabla^2 f(x)]^{-1} \nabla f(x)$ zugleich Lösung der quadratischen Optimierungsaufgabe

$$
\text{Minimiere} \quad q(\Delta x) \;=\; \frac{1}{2}\,\Delta x^{\mathrm{T}} Q\,\Delta x \;+\; g^{\mathrm{T}} \Delta x \;+\; \gamma, \quad \Delta x \in \mathbb{R}^n,
$$

wobei $Q := \nabla^2 f(x)$ (als positiv definit vorausgesetzt), $g := \nabla f(x)$ und $\gamma := f(x)$.

Eine analoge Eigenschaft lässt sich für die Lagrange-Newton Iteration feststellen. Die Lagrange-Newton Korrektur Δx ist nach (15.6) gegeben durch

$$
\begin{pmatrix} Q & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \lambda^+ \end{pmatrix} \;=\; -\begin{pmatrix} \nabla f(x) \\ h(x) \end{pmatrix}, \tag{15.11}
$$

wobei $Q \;:=\; \nabla^2 f(x) \;+\; \sum_{j \in E} \lambda_j\, \nabla^2 h_j(x)$ und $B = h'(x)$.

Hat B maximalen Rang und ist die Matrix Q positiv definit auf Kern(B), so ist dieses lineare Gleichungssystem eindeutig lösbar und nach Abschnitt 14 ist Δx zugleich Lösung des folgenden **quadratischen Teilproblems**, vgl. (14.1) und (14.3)

$$\begin{array}{ll} \text{Minimiere} & q(\Delta x) \;=\; \frac{1}{2}\,\Delta x^{\mathrm{T}} Q\,\Delta x \;+\; \nabla f(x)^{\mathrm{T}}\Delta x \;+\; f(x), \quad \Delta x \in \mathbb{R}^n, \\ \text{Nebenbedingungen:} & B\,\Delta x \;+\; h(x) \;=\; 0. \end{array} \tag{15.12}$$

Bemerkung (15.13)

a) Ist (x^*, λ^*) ein regulärer KKT-Punkt, der die hinreichenden Bedingungen (12.10) erfüllt, so sind die obigen Voraussetzungen für alle (x, λ) in einer Umgebung von (x^*, λ^*) erfüllt und das quadratische Hilfsproblem ist in dieser Umgebung lösbar (Übungsaufgabe).

b) Der Lagrange-Parameter λ^+ aus (15.11) stimmt überein mit dem Lagrange-Parameter des quadratischen Hilfsproblems. Die Lagrange-Funktion dieses Problems lautet nämlich

$$L_q(\Delta x, \mu) \;=\; \frac{1}{2}\,\Delta x^{\mathrm{T}} Q\,\Delta x \;+\; \nabla f(x)^{\mathrm{T}}\Delta x \;+\; f(x) \;+\; \mu^{\mathrm{T}}(B\,\Delta x + h(x)),$$

woraus sich die folgenden notwendigen Bedingungen ergeben

$$\nabla L_q \;=\; \begin{pmatrix} Q\,\Delta x \;+\; \nabla f(x) \;+\; B^{\mathrm{T}}\mu \\ B\,\Delta x \;+\; h(x) \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Dies entspricht aber genau dem linearen Gleichungssystem (15.11) mit $\mu = \lambda^+$.

c) Im Fall $\mathrm{Rang}(B) = p$ ist die Existenz (und Eindeutigkeit) eines Lagrange-Parameters λ^+ gesichert. Das lineare (Teil-)Gleichungssystem (bei bekanntem Δx)

$$B^{\mathrm{T}}\,\lambda^+ \;=\; -\,(\nabla f(x) \;+\; Q\,\Delta x)$$

ist jedoch überbestimmt. Es empfiehlt sich daher, λ^+ als Lösung des linearen Ausgleichsproblems

$$\text{Minimiere} \quad \{\|B^{\mathrm{T}}\,\lambda^+ \;+\; \nabla f(x) \;+\; Q\,\Delta x\|_2 : \quad \lambda^+ \in \mathbb{R}^p\}$$

zu ermitteln (d.h. mittels QR-Zerlegung von B^{T} und hierbei die Kleinheit des Residuums zu überprüfen.

Wir übertragen die Grundidee des SQP Verfahrens nun auf **allgemeine restringierte Optimierungsaufgaben** des Standardtyps

$$\begin{array}{lll} \text{Minimiere} & f(x), & x \in \mathbb{R}^n, \\ \text{Nebenbedingungen:} & g_i(x) \;\le\; 0, & i \in I := \{1, \ldots, m\} \\ & h_j(x) \;=\; 0, & j \in E := \{1, \ldots, p\}. \end{array} \tag{15.14}$$

Hierbei seien alle auftretenden Funktionen $f,\ g_i,\ h_j$ zweifach stetig differenzierbar.

Zu vorgegebenem $x \in \mathbb{R}^n,\ \lambda \in \mathbb{R}^m$ und $\mu \in \mathbb{R}^p$ definiert man

$$Q \;=\; Q(x,\lambda,\mu) \;:=\; \nabla^2 f(x) \;+\; \sum_{i\in I} \lambda_i \,\nabla^2 g_i(x) \;+\; \sum_{j\in E} \mu_j \,\nabla^2 h_j(x) \tag{15.15}$$

und hiermit das **quadratische Hilfsproblem**

$$\begin{aligned} &\text{Minimiere} \quad q(\Delta x) \;=\; \frac{1}{2}\,\Delta x^{\mathrm{T}} Q\,\Delta x \;+\; \nabla f(x)^{\mathrm{T}}\Delta x \;+\; f(x), \quad \Delta x \in \mathbb{R}^n, \\ &\text{Nebenbed.:} \quad \nabla g_i(x)^{\mathrm{T}}\,\Delta x \;+\; g_i(x) \;\le\; 0, \qquad i \in I, \\ &\qquad\qquad \nabla h_j(x)^{\mathrm{T}}\,\Delta x \;+\; h_j(x) \;=\; 0, \quad j \in E. \end{aligned} \tag{15.16}$$

Damit ergibt sich der folgende Modell-Algorithmus für ein so genanntes SQP-Verfahren (*sequential quadratic programming*)

Algorithmus (15.17) (SQP-Verfahren)

1.) Wähle $(x^0,\lambda^0,\mu^0)^{\mathrm{T}} \in \mathbb{R}^{(n+m+p)},\ \mathrm{TOL} > 0,\ k := 0;$

2.) Berechne $\nabla f(x^k),\ A := g'(x^k),\ B := h'(x^k)$ und

$$F^k \;:=\; \begin{pmatrix} \nabla f(x^k) \;+\; A^{\mathrm{T}}\lambda^k \;+\; B^{\mathrm{T}}\mu^k \\ h(x) \end{pmatrix}$$

Falls $\|F^k\| < \mathrm{TOL},\ \lambda^k \ge 0,\ g(x^k) \le 0$ und $|(\lambda^k)^{\mathrm{T}} g(x^k)| \le \mathrm{TOL}$: Stop!

Berechne $Q := \nabla^2 f(x^k) + \sum_{i\in I} \lambda_i^k \,\nabla^2 g_i(x^k) + \sum_{j\in E} \mu_j^k \,\nabla^2 h_j(x^k).$

3.) Löse das quadratische Hilfsproblem (15.16) mit $x := x^k$.
Hierbei bestimme man auch die aktive Menge $I_{\mathrm{lin}}(\Delta x)$ und die Lagrange-Multiplikatoren $\lambda_{\mathrm{lin}},\ \mu_{\mathrm{lin}}$ – beispielsweise aus dem überbestimmten linearen Gleichungssystem

$$\begin{pmatrix} \nabla g_i(x)^{\mathrm{T}}_{i\in I_{\mathrm{lin}}(\Delta x)} \\ \nabla h_j(x)^{\mathrm{T}}_{j\in E} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \lambda_{\mathrm{lin}} \\ \mu_{\mathrm{lin}} \end{pmatrix} \;=\; -\,(Q\,\Delta x \;+\; \nabla f(x))\ .$$

4.) Setze $x^{k+1} := x^k + \Delta x,\ \lambda_i^{k+1} := \lambda_{i,\mathrm{lin}},$ falls $i \in I_{\mathrm{lin}}(\Delta x)$ und $\lambda_i^{k+1} := 0,$ sonst, $\mu^{k+1} := \mu_{\mathrm{lin}},\ k := k+1;$ gehe zu 2.)

Bemerkungen (15.18)

a) Es ist klar, dass man die Konvergenz des obigen SQP-Verfahrens nur unter starken Voraussetzungen wird garantieren können. So müssen sämtliche quadratischen Hilfsprobleme (eindeutig) lösbar sein und die aktive Indexmenge muss in der Konvergenzphase festliegen. Andererseits erreicht man mit diesem Ansatz die schnelle, quadratische Konvergenz des Verfahrens.

b) Liefert das quadratische Hilfsproblem die Lösung $\Delta x = 0$, so ist die neue Näherung $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ bereits ein KKT-Punkt des Ausgangsproblems. Die Zulässigkeit folgt aus den Nebenbedingungen des quadratischen Hilfsproblems, die KKT-Bedingungen ebenfalls aus denen des quadratischen Hilfsproblems.

Satz (15.19) (Lokale Konvergenz)

Sei (x^*, λ^*, μ^*) ein KKT-Punkt des Ausgangsproblems (15.14), der den folgenden Voraussetzungen genügt

(i) Strikte Komplementarität:

$$\forall\, i \in I: \quad g_i(x^*) \,-\, \lambda_i^* \,<\, 0,$$

(ii) LICQ-Bedingung:

$$\nabla g_i(x^*),\ i \in I(x^*), \quad \nabla h_j(x^*),\ j \in E \quad \text{linear unabhängig},$$

(iii) Hinreichende Bedingung: (Beachte (i))

$$\forall\, y \in T_X^a(x^*) \setminus \{0\}: \quad y^{\mathrm{T}}\, \nabla_{x\,x}^2 L(x^*, \lambda^*, \mu^*)\, y \;>\; 0.$$

Dann ist das obige SQP-Verfahren für alle Startwerte (x^0, λ^0, μ^0) aus einer geeigneten Umgebung von (x^*, λ^*, μ^*) wohldefiniert und liefert eine gegen (x^*, λ^*, μ^*) konvergente Folge. Sind alle zweiten Ableitungen lokal Lipschitz-stetig, so ist die Konvergenz quadratisch, vgl. Geiger, Kanzow, Band 2, Satz 5.31.

C. Globalisierung des SQP-Verfahrens.

Das SQP-Verfahren ist – wie auch das Newton-Verfahren – im Allg. nur lokal konvergent, wobei der *Einzugsbereich* für höhere Dimensionszahlen oder für komplexe Probleme durchaus klein ausfallen kann. Maßnahmen zur Globalisierung des Verfahrens (Schrittweitenstrategien, trust region Ansatz, ...) sind daher für die praktischen Anwendungen wesentlich.

Zur Anwendung von Schrittweitenstrategien ist zunächst die Definition einer geeigneten Abstiegsfunktion (Testfunktion) notwendig. Man beachte, dass die Iterierten des SQP-Verfahrens im Allg. *nicht zulässig* sind, ein Abstieg von f also kein geeignetes Maß darstellt.

Statt dessen betrachtet man die so genannte **exakte ℓ_1– Penalty Funktion**

$$P_1(x;\alpha) \;:=\; f(x) \;+\; \alpha \sum_{i\in I} \max(0, g_i(x)) \;+\; \alpha \sum_{j\in E} |h_j(x)| \tag{15.20}$$

Hierbei ist $\alpha > 0$ ein fester **Penalty Parameter**. Die Schrittweite des SQP-Verfahrens wird dann so bestimmt, dass die exakte ℓ_1-Penalty Funktion in jeder Iteration ausreichend fällt.

Nun ist $P_1(\cdot,\alpha)$ zwar im Allg. nicht differenzierbar, besitzt aber gleichwohl Richtungsableitungen

$$P_1'(x,y;\alpha) \;:=\; \lim_{t\downarrow 0} \frac{P_1(x+t\,y;\alpha) \;-\; P_1(x;\alpha)}{t}$$

für alle $y \in \mathbb{R}^n$.

Für den vorliegenden Fall der exakten ℓ_1-Penalty Funktion (15.20) ergibt sich

$$\begin{aligned} P_1'(x,y;\alpha) \;&=\; \nabla f(x)^{\mathrm{T}} y \;+\; \alpha \left\{ \sum_{g_i>0} \nabla g_i(x)^{\mathrm{T}} y \;+\; \sum_{g_i=0} \max(0, \nabla g_i(x)^{\mathrm{T}} y) \right\} \\ &+\; \alpha \left\{ \sum_{h_j>0} \nabla h_j(x)^{\mathrm{T}} y \;-\; \sum_{h_j<0} \nabla h_j(x)^{\mathrm{T}} y \;+\; \sum_{h_j=0} |\nabla h_j(x)^{\mathrm{T}} y| \right\} \end{aligned} \tag{15.21}$$

Vergleiche hierzu 5.32 – 5.34 in Geiger, Kanzow, Band 2.

Hiermit zeigen wir nun, dass die Lösung Δx des quadratischen Hilfsproblems stets eine Abstiegsrichtung von $P_1(\cdot,\alpha)$ ist.

Satz (15.22)

Für das quadratische Hilfsproblem (15.16) mit symmetrischer und positiv definiter Matrix $Q \in \mathbb{R}^{(n,n)}$ sei $\Delta x \neq 0$ eine Lösung mit zugehörigen Lagrange-Multiplikatoren λ, μ.
Für Penalty Parameter $\alpha \geq \max\{\lambda_1, \ldots, \lambda_m, |\mu_1|, \ldots, |\mu_p|\}$ gilt

$$P_1'(x, \Delta x;\alpha) \;\leq\; -\,\Delta x^{\mathrm{T}}\, Q\, \Delta x \;<\; 0,$$

d.h. Δx ist eine Abstiegsrichtung der exakten ℓ_1-Penalty Funktion.

Beweis:

Wir formen

$$\begin{aligned} P_1'(x, \Delta x;\alpha) \;&=\; \nabla f(x)^{\mathrm{T}} \Delta x \;+\; \alpha \left\{ \sum_{g_i>0} \nabla g_i(x)^{\mathrm{T}} \Delta x \;+\; \sum_{g_i=0} \max(0, \nabla g_i(x)^{\mathrm{T}} \Delta x) \right\} \\ &+\; \alpha \left\{ \sum_{h_j>0} \nabla h_j(x)^{\mathrm{T}} \Delta x \;-\; \sum_{h_j<0} \nabla h_j(x)^{\mathrm{T}} \Delta x \;+\; \sum_{h_j=0} |\nabla h_j(x)^{\mathrm{T}} \Delta x| \right\} \end{aligned}$$

mit Hilfe der KKT-Bedingungen für das quadratische Hilfsproblem um.

Aus der Bedingung erster Ordnung $\nabla f \; + \; Q\,\Delta x \; + \; \sum\limits_{i\in I} \lambda_i \nabla g_i \; + \; \sum\limits_{j\in E} \mu_j \nabla h_j \;\; = \;\; 0$ folgt

$$\nabla f^{\mathrm{T}} \Delta x \;\; = \;\; -\,\Delta x^{\mathrm{T}}\, Q\, \Delta x \; - \; \sum_{i\in I} \lambda_i \, \nabla g_i^{\mathrm{T}} \Delta x \; - \; \sum_{j\in E} \mu_j \, \nabla h_j^{\mathrm{T}} \Delta x.$$

Die Zulässigkeit bezüglich der Ungleichungsnebenbedingungen $\nabla g_i^{\mathrm{T}} \Delta x + g_i \;\; \leq \;\; 0$ liefert

$$\begin{aligned} \max\{0, \nabla g_i^{\mathrm{T}} \Delta x\} \;\; &= \;\; 0, \quad \text{für} \quad g_i \; = \; 0, \\ \nabla g_i^{\mathrm{T}} \Delta x \;\; &\leq \;\; -g_i, \qquad\quad \text{für} \quad g_i \; > \; 0, \end{aligned}$$

und bezüglich der Gleichungsnebenbedingungen $\nabla h_j^{\mathrm{T}} \Delta x \; + \; h_j \; = \; 0$:

$$\nabla h_j^{\mathrm{T}} \, \Delta x \;\; = \;\; -\,h_j, \qquad j \in E.$$

Schließlich ergibt die Komplementarität

$$\lambda_i \; \left(\nabla g_i^{\mathrm{T}} \, \Delta x \right) \;\; = \;\; -\,\lambda_i \, g_i, \qquad i \in I.$$

Setzt man diese Ergebnisse nun in die obigen Terme für die Richtungsableitung der ℓ_1-Penalty Funktion ein, so erhält man

$$\begin{aligned} P_1'(x, \Delta x; \alpha) \;\; &= \;\; -\,\Delta x^{\mathrm{T}}\, Q\, \Delta x \; - \; \sum_{i\in I} \lambda_i \, \nabla g_i^{\mathrm{T}} \Delta x \; - \; \sum_{j\in E} \mu_j \, \nabla h_j^{\mathrm{T}} \Delta x \\ &\quad + \; \alpha \; \left\{ \sum_{g_i>0} \nabla g_i^{\mathrm{T}} \Delta x \; + \; \sum_{g_i=0} \max(0, \nabla g_i^{\mathrm{T}} \Delta x) \right\} \\ &\quad + \; \alpha \; \left\{ \sum_{h_j>0} \nabla h_j^{\mathrm{T}} \Delta x \; - \; \sum_{h_j<0} \nabla h_j^{\mathrm{T}} \Delta x \; + \; \sum_{h_j=0} |\nabla h_j^{\mathrm{T}} \Delta x| \right\} \\ &\leq \;\; -\,\Delta x^{\mathrm{T}}\, Q\, \Delta x \; + \; \sum_{i\in I} \lambda_i \, g_i \; + \; \sum_{j\in E} \mu_j \, h_j \\ &\quad + \; \alpha \; \left\{ - \sum_{g_i>0} g_i \right\} \; + \; \alpha \; \left\{ - \sum_{h_j>0} h_j \; + \; \sum_{h_j<0} h_j \right\} \\ &= \;\; -\,\Delta x^{\mathrm{T}}\, Q\, \Delta x \; + \; \sum_{g_i>0} (\lambda_i - \alpha) \, g_i \; + \; \sum_{g_i<0} \lambda_i \, g_i \\ &\quad + \; \sum_{h_j>0} (\mu_j - \alpha) \, h_j \; + \; \sum_{h_j<0} (\mu_j + \alpha) \, h_j \\ &\leq \;\; -\,\Delta x^{\mathrm{T}}\, Q\, \Delta x \;\; < \;\; 0. \qquad \square \end{aligned}$$

Aufgrund des obigen Satzes lässt sich dass SQP-Verfahren nun folgendermaßen mit einer Schrittweitensteuerung erweitern.

Algorithmus (15.23) (Globalisiertes SQP-Verfahren)

1.) Start: $(x, \lambda, \mu)^{\mathrm{T}} = (x^0, \lambda^0, \mu^0)^{\mathrm{T}} \in \mathbb{R}^{(n+m+p)}$, $\alpha > 0$, $\sigma \in]0, 1[$, $Q = Q_0 \in \mathbb{R}^{(n,n)}$ symmetrisch, $k := 0$, $\mathrm{TOL} > 0$;

2.) Berechne $\nabla f(x)$, $\nabla g_i(x)$, $\nabla h_j(x)$, $i \in I$, $j \in E$.
Ist (x, λ, μ) innerhalb der vorgegebenen Genauigkeit ein KKT-Punkt des Ausgangsproblems: Stop!

3.) Berechnen eine Lösung Δx des quadratischen Hilfsproblem (15.16) sowie die zugehörige aktive Menge und die Lagrange-Multiplikatoren λ^{k+1} und μ^{k+1}. Ist $\|\Delta x\| \leq \mathrm{TOL}$: Stop!

4.) Bestimme eine Schrittweite $t = 0.5^{\nu}$, $\nu \in \mathbb{N}_0$ minimal, mit

$$P_1(x + t\,\Delta x; \alpha) \;\leq\; P_1(x; \alpha) \;+\; \sigma\, t\, P_1'(x, \Delta x; \alpha)$$

(Armijo–Schrittweite)

5.) Setze $x = x^{k+1} := x^k + t\,\Delta x$. Bestimme Q_{k+1} durch eine geeignete Update-Formel; $k := k + 1$; gehe zu 2.)

Bemerkungen (15.24)

a) Globale Konvergenzaussagen für globalisierte SQP-Verfahren findet man beispielsweise bei Han[5].

b) Im Allg. wird man den Penalty-Parameter α in jedem Iterationschritt anpassen müssen. Dies könnte beispielsweise folgendermaßen erfolgen

$$\alpha_{k+1} \;=\; \max\{\alpha_k, \max\{\lambda_1^{k+1}, \ldots, \lambda_m^{k+1}, |\mu_1^{k+1}|, \ldots, |\mu_p^{k+1}|\} + \gamma\},$$

wobei $\gamma > 0$ vorgegeben sei. Man hat jedoch darauf zu achten, dass die Folge (α_k) beschränkt bleibt.

c) Von M.J.D. Powell (1978) stammt eine modifizierte BFGS-Update Formel für die Hesse-Matrizen Q_k, die deren positive Definitheit garantiert und für die Powell die superlineare Konvergenz des globalisierten SQP-Verfahrens zeigen konnte, selbst bei Verwendung der Armijo-Schrittweite und einer indefiniten Hesse-Matrix $\nabla^2_{x\,x} L$ im Lösungspunkt.

d) Weitere Modifikationen sind notwendig, um im Zusammenhang mit der Schrittweitensteuerung die quadratische Konvergenz des Verfahrens zu sichern; Stichworte:

[5] S.P. Han: A globally convergent method for nonlinear programming. Journal of Optimization Theory and Applications, 22, 297-309, 1977.

Maratos-Effekt (d.h. das Verfahren arbeitet auch im Einzugsbereich einer Lösung mit Schrittweiten $t < 1$), nichtmonotone Schrittweitensteuerung. Für Einzelheiten sei auf die Abschnitte 5.5.6 – 5.5.8 im Buch von Geiger und Kanzow, Band 2, verwiesen.

e) Abschließend sei bemerkt, dass SQP-Verfahren in verschiedenen Varianten aktuell zu den effizientesten Verfahren für allgemeine restringierte Optimierungsaufgaben gehören. Die gängigen numerischen Programmpakete IMSL, NAG und auch MATLAB bieten jeweils ausgefeilte SQP-Verfahren an.

16. Reduktionsmethoden

A. Probleme mit linearen Gleichungsnebenbedingungen.

Wir betrachten eine Optimierungsaufgabe mit linearen Gleichungsrestriktionen:

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n, \\ &\text{Nebenbedingungen:} \quad h(x) \;=\; B\,x \;-\; q \;=\; 0 \in \mathbb{R}^p. \end{aligned} \tag{16.1}$$

Die Matrix $B \in \mathbb{R}^{(p,n)}$ habe hierbei maximalen Rang $p < n$. Insbesondere ist damit jeder zulässige Punkt regulär und die zulässige Menge X ist ein $(n-p)$ dimensionaler affin-linearer Teilraum des $\mathbb{R}^n$.

Für X hat man damit die folgende *Parametrisierung*

$$X \;=\; \{x \in \mathbb{R}^n : \; B\,x \;=\; q\} \;=\; \hat{x} \;+\; \text{Kern}(B). \tag{16.2}$$

Dabei bezeichne $\hat{x} \in X$ einen beliebigen zulässigen Punkt. Ferner ist $\text{Kern}(B) = T_X(\hat{x})$.

Sei nun $(z^1, \ldots, z^{n-p})$ eine Basis von $\text{Kern}(B)$ und $Z := (z^1, \ldots, z^{n-p}) \in \mathbb{R}^{(n,n-p)}$. Damit hat man die folgende Darstellung für die zulässige Menge

$$X \;=\; \{x \;=\; \hat{x} + Z\,u : \; u \in \mathbb{R}^{n-p}\} \tag{16.3}$$

und die Optimierungsaufgabe (16.1) ist äquivalent zu der folgenden unrestringierten Optimierungsaufgabe

$$\text{Minimiere} \quad \widetilde{f}(u) \;:=\; f(\hat{x} + Z\,u), \quad u \in \mathbb{R}^{n-p}. \tag{16.4}$$

(16.4) heißt die **reduzierte Optimierungsaufgabe** zu (16.1).

Zur numerischen Lösung von (16.4) werden noch Gradient und Hesse-Matrix der Zielfunktion benötigt:

$$\begin{aligned} \nabla \widetilde{f}(u) \;&=\; Z^{\mathrm{T}}\,\nabla f(\hat{x} + Z\,u) \;\in\; \mathbb{R}^{n-p} \\ \nabla^2 \widetilde{f}(u) \;&=\; Z^{\mathrm{T}}\,\nabla^2 f(\hat{x} + Z\,u)\,Z \;\in\; \mathbb{R}^{(n-p,n-p)} \end{aligned} \tag{16.5}$$

Man spricht hierbei vom **reduzierten Gradienten** und von der **reduzierten Hesse-Matrix**.

Wir beschreiben im Folgenden drei Verfahren zur Berechnung von $\hat{x}$ und Z.

I. Eliminationsmethode.

Man zerlegt die Matrix B unter ev. Vertauschen von Variablen (Permutationsvektor!) wie folgt

$$B \;=\; (B_1|B_2); \quad B_1 \in \mathbb{R}^{(p,p)} \text{ regulär}, \quad B_2 \in \mathbb{R}^{(p,n-p)}.$$

Die entsprechende Zerlegung von $x = (x_1, u)^{\mathrm{T}}$ mit $u \in \mathbb{R}^{(n-p)}$ liefert dann

$$\begin{aligned} B\,x \;=\; q \;&\Leftrightarrow\; B_1\,x_1 \,+\, B_2\,u \;=\; q \\ &\Leftrightarrow\; x_1 \;=\; -B_1^{-1}\,B_2\,u \,+\, B_1^{-1}\,q \\ &\Leftrightarrow\; x \;=\; \begin{pmatrix} B_1^{-1} \\ 0 \end{pmatrix} q \,+\, \begin{pmatrix} -B_1^{-1}\,B_2 \\ I_{n-p} \end{pmatrix} u \end{aligned}$$

Setzt man also

$$\hat{x} \;:=\; Y\,q, \qquad Y \;:=\; \begin{pmatrix} B_1^{-1} \\ 0 \end{pmatrix}, \qquad Z \;:=\; \begin{pmatrix} -B_1^{-1}\,B_2 \\ I_{n-p} \end{pmatrix}, \tag{16.6}$$

so erhält man gerade die Darstellung (16.3) für die zulässige Menge.

Bemerkungen (16.7)

a) Für die in (16.6) gegebenen Matrizen $Y \in \mathbb{R}^{(n,p)}$, $Z \in \mathbb{R}^{(n,n-p)}$ gelten offensichtlich

$$B\,Y \;=\; I_p\,, \qquad B\,Z \;=\; 0\,, \qquad \mathrm{Rang}(Z) \;=\; n-p. \tag{16.8}$$

b) Umgekehrt folgt aus (16.8), dass $\hat{x} := Y\,q$ ein zulässiger Punkt ist und die Spaltenvektoren von Z eine Basis von $\mathrm{Kern}(B)$ bilden.

c) Numerisch kann man die gewünschte Zerlegung mittels Gauß-Elimination aus dem linearen Gleichungssystem $B\,x = q$ erhalten, wobei ev. Zeilen- und Spaltenvertauschungen ausgeführt werden müssen. Mit einer regulären oberen Dreiecksmatrix $R_1 \in \mathbb{R}^{(p,p)}$ wird dabei:

$$\begin{aligned} B\,x \;=\; q \;&\rightarrow\; (R_1|\widetilde{B}_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;=\; \widetilde{q} \\ &\rightarrow\; \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;=\; \begin{pmatrix} R_1^{-1} \\ 0 \end{pmatrix} \widetilde{q} \,+\, \begin{pmatrix} -\,R_1^{-1}\widetilde{B}_2 \\ I_{n-p} \end{pmatrix} x_2. \end{aligned}$$

II. Orthogonalzerlegungsmethode.

Numerisch stabiler als Gauß-Elimination ist die Orthogonalzerlegung (QR-Zerlegung) der Matrix B^{T}:

$$B^{\mathrm{T}} \;=\; Q \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad Q \in \mathbb{R}^{(n,n)}, \quad R \in \mathbb{R}^{(p,p)}, \tag{16.9}$$

wobei Q eine orthogonale Matrix und R eine reguläre obere Dreiecksmatrix ist.

Zerlegt man die Matrix Q analog zur Zerlegung von R

$$B^{\mathrm{T}} \;=\; Q \begin{pmatrix} R \\ 0 \end{pmatrix} \;=\; (Q_1 | Q_2) \begin{pmatrix} R \\ 0 \end{pmatrix} \;=\; Q_1\, R,$$

so lässt sich die allgemeine Lösung von $B\,x = q$ wie folgt darstellen

$$\begin{aligned} & B\,x \;=\; q \;\Leftrightarrow\; (R^{\mathrm{T}} | \,0)\,(Q^{\mathrm{T}} x) \;=\; q, \quad \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} := Q^{\mathrm{T}}\, x \\ & \Leftrightarrow\; R^{\mathrm{T}} u_1 \;=\; q, \quad x \;=\; Q \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad u_2 \text{ beliebig} \\ & \Leftrightarrow\; x \;=\; Q_1\, u_1 \;+\; Q_2\, u_2 \;=\; (Q_1\, R^{-\mathrm{T}})\, q \;+\; Q_2\, u_2. \end{aligned}$$

Setzt man also $Y := Q_1\, R^{-\mathrm{T}}$ und $Z := Q_2$, so erhält man die gewünschte Parametrisierung der zulässigen Menge:

$$B\,x \;=\; q \;\Leftrightarrow\; x \;=\; Y\, q \;+\; Z\, u_2, \quad u_2 \in \mathbb{R}^{n-p}. \tag{16.10}$$

Bemerkungen (16.11)

a) Die Matrizen $Y := Q_1\, R^{-\mathrm{T}} \in \mathbb{R}^{(n,p)}$ und $Z := Q_2 \in \mathbb{R}^{(n,n-p)}$ erfüllen die Eigenschaften (16.8)

$$\begin{aligned} B\,Y \;&=\; R^{\mathrm{T}}\, Q_1^{\mathrm{T}}\, Q_1\, R^{-\mathrm{T}} \;=\; R^{\mathrm{T}}\, R^{-\mathrm{T}} \;=\; I_p\,, \\ B\,Z \;&=\; R^{\mathrm{T}}\, Q_1^{\mathrm{T}}\, Q_2 \;=\; R^{\mathrm{T}}\, 0 \;=\; 0\,, \\ Z \;&=\; Q_2 \quad \text{hat maximalen Rang.} \end{aligned}$$

b) Der spezielle zulässige Punkt $\hat{x} \;=\; Y\, q \;=\; Q_1\,(R^{-\mathrm{T}}\, q)$ ist eine Linearkombination der Spaltenvektoren von Q_1, die sich numerisch mittels Rückwärtsrekursion berechnen läßt.

c) $x_h := Z\, u_2 \;=\; Q_2\, u_2$ bildet den Kern der Matrix B. Die Spaltenvektoren von Q_2 bilden also eine ONB von Kern(B) und die spezielle Lösung $\hat{x}$ steht senkrecht auf diesem Raum.

III. Ein allgemeiner Ansatz.

Man erweitert die Matrix B durch Hinzunahme beliebiger weiterer Zeilen, so dass eine reguläre (quadratische) Matrix entsteht. Damit wird aus dem unterbestimmten linearen Gleichungssystem $B\,x = q$ ein eindeutig lösbares (erweitertes) Gleichungssystem

$$\begin{pmatrix} B \\ V \end{pmatrix} x \;=\; \begin{pmatrix} q \\ u \end{pmatrix}, \qquad u \in \mathbb{R}^{n-p}$$

Dabei ist u beliebig vorzugeben, u spielt die Rolle der Parametrisierung der zulässigen Menge. Es folgt

$$x \;=\; \begin{pmatrix} B \\ V \end{pmatrix}^{-1} \begin{pmatrix} q \\ u \end{pmatrix} \;=:\; (\,Y \,|\, Z\,) \begin{pmatrix} q \\ u \end{pmatrix}$$

Hierdurch sind $Y \in \mathbb{R}^{(n,p)}$ und $Z \in \mathbb{R}^{(n,n-p)}$ wohldefiniert (als Inverse zu (B/V)) und erfüllen die allgemeine Vorassetzung (16.8).

Die früheren Methoden erweisen sich als Spezialfälle, nämlich:

- Eliminationsmethode: $V \;=\; (\,0\,|\,I_{n-p}\,)$,
- Orthogonalzerlegung: $V \;=\; Q_2^{\mathrm{T}}$.

Das reduzierte Problem.

Ist ein spezieller zulässiger Punkt $\hat{x} \in X$ sowie eine Basis von Kern(B) bekannt, so verbleibt die Lösung des (unrestringierten) reduzierten Problems

$$\text{Minimiere} \quad \widetilde{f}(u) \;:=\; f(\hat{x} \,+\, Z\,u), \quad u \in \mathbb{R}^p. \tag{16.12}$$

Es ist hierbei aus numerischen Gründen zu empfehlen, für $\hat{x}$ jeweils die aktuelle Näherung x^k zu wählen. Die zugehörige Näherung für u ist dann $u^k = 0$.

Das Newton-Verfahren ergibt sich durch Minimierung des quadratischen Modells von $\widetilde{f}$ $(u^k = 0,\ u = \Delta u)$

$$\begin{aligned} q(u) \;&=\; \widetilde{f}(0) \,+\, \nabla\widetilde{f}(0)^{\mathrm{T}}u \,+\, \frac{1}{2}\,u^{\mathrm{T}}\nabla^2\widetilde{f}(0)\,u \\ &=\; f^k \,+\, (g^k)^{\mathrm{T}}Z\,u \,+\, \frac{1}{2}\,u^{\mathrm{T}}\,(Z^{\mathrm{T}}G^k\,Z)\,u \end{aligned}$$

mit $f^k := f(x^k)$, $g^k \;:=\; \nabla f(x^k)$ und $G^k \;:=\; \nabla^2 f(x^k)$.

Für die Newton-Richtung $s^k := \Delta u^k = u^{k+1}$ erhält man damit das lineare Gleichungssystem

$$(Z^{\mathrm{T}}\,G^k\,Z)\,s^k \;=\; -Z^{\mathrm{T}}\,g^k. \tag{16.13}$$

Die Rücktransformation in den x-Bereich ergibt dann die Suchrichtung $d^k \;=\; Z\,s^k$.

Ein **reduziertes Quasi-Newton-Verfahren** erhält man durch die Ersetzung der Hesse-Matrix von $\widetilde{f}$ durch eine Approximation $H_k \approx \nabla^2 \widetilde{f}(u^k)$, die mit einer Quasi-Newton Formel, z.B. dem BFGS-update (9.12) aufdatiert wird. (16.13) wird dann ersetzt durch

$$H_k \, s^k \;=\; -\, Z^{\mathrm{T}} \, g^k. \tag{16.14}$$

Für die Vektoren s und y der BFGS-update-Formel

$$H_{k+1} \;=\; H_k \;+\; \frac{y\, y^{\mathrm{T}}}{y^{\mathrm{T}} s} \;-\; \frac{(H_k\, s)\,(H_k\, s)^{\mathrm{T}}}{s^{\mathrm{T}} H_k\, s}\,. \tag{16.15}$$

ergibt sich

$$\begin{aligned} s \;&:=\; u^{k+1} - u^k \;=\; u^{k+1} \;=\; s^k \\ y \;&:=\; \nabla \widetilde{f}^{k+1} \;-\; \nabla \widetilde{f}^{k} \;=\; Z^{\mathrm{T}}\,(g^{k+1} - g^k). \end{aligned} \tag{16.16}$$

Die Lagrange-Multiplikatoren.

Häufig wird gewünscht, neben der Lösung x^* des Ausgangsproblems (16.1) auch die zugehörigen Lagrange-Multiplikatoren $\mu \in \mathbb{R}^p$ zu ermitteln. Nimmt man an, dass x^* ein *reguläres* lokales Minimum ist, so ist μ bekanntlich eindeutig bestimmt und mit der Lagrange-Funktion $L(x,\mu) \;=\; f(x) \;+\; \mu^{\mathrm{T}}(B\,x - q)$ hat man das folgende überbestimmte lineare Gleichungssystem für μ

$$\nabla_x L \;=\; \nabla f(x) \;+\; B^{\mathrm{T}} \mu \;=\; 0. \tag{16.17}$$

Sind nun $Y \in \mathbb{R}^{(n,p)}$ und $Z \in \mathbb{R}^{(n,n-p)}$ gegeben mit den Eigenschaften (16.8), so folgt durch Multiplikation von (16.17) mit Y^{T} von links

$$Y^{\mathrm{T}} \nabla f(x) \;=\; -Y^{\mathrm{T}} B^{\mathrm{T}} \mu \;=\; -\,(B\,Y)^{\mathrm{T}}\,\mu \;=\; -\,\mu,$$

also

$$\mu \;=\; -\,Y^{\mathrm{T}} \nabla f(x). \tag{16.18}$$

Es ist hierbei zu beachten, dass das Gleichungssysten (16.17) nur im Lösungspunkt x^* lösbar sein muss, während die Relation (16.18) auch während der Iteration als eine Näherung für den Lagrange-Multiplikator μ ausgewertet werden kann.

Speziell für die **Eliminationsmethode** ergibt sich, vgl. (16.6),

$$B\,x \;=\; q \quad\to\quad (B_1 \,|\, B_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;=\; q, \qquad B_1 \;=\; L_1\, R_1$$

$$\Rightarrow \quad Y \;=\; \begin{pmatrix} B_1^{-1} \\ 0 \end{pmatrix} \;=\; \begin{pmatrix} R_1^{-1}\, L_1^{-1} \\ 0 \end{pmatrix}, \qquad Z \;:=\; \begin{pmatrix} -B_1^{-1}\, B_2 \\ I_{n-p} \end{pmatrix}$$

$$\Rightarrow \quad \mu \;=\; -\;\left(L_1^{-\mathrm{T}}\, R_1^{-\mathrm{T}} \,|\, 0 \right) \begin{pmatrix} \nabla f_1(x) \\ \nabla f_2(x) \end{pmatrix}.$$

Damit folgt

$$\mu \;=\; -\,L_1^{-\mathrm{T}}\,R_1^{-\mathrm{T}}\,\nabla f_1(x). \tag{16.19}$$

Die Berechnung von μ kann also relativ einfach mittels Vorwärts-/Rückwärtssubstitution erfolgen.

Für die **Orthogonalzerlegungsmethode** ergibt sich mit $Y := Q_1\,R^{-\mathrm{T}}$

$$\mu \;=\; -\,Y^{-\mathrm{T}}\,\nabla f(x) \;=\; -\,R^{-1}\,Q_1^{\mathrm{T}}\,\nabla f(x). \tag{16.20}$$

Der Lagrange-Multiplikator μ aus (16.20) minimiert das Residuum der Ausgangsgleichung (16.17):

$$\begin{aligned} r \;&:=\; B^{\mathrm{T}}\,\mu \,+\, \nabla f(x) \;=\; Q \begin{pmatrix} R \\ 0 \end{pmatrix} \mu \,+\, \nabla f(x) \\ &=\; Q\,\Big\{ \begin{pmatrix} R \\ 0 \end{pmatrix} \mu \,+\, Q^{\mathrm{T}}\,\nabla f(x) \Big\} \\ &=\; Q \begin{pmatrix} R\,\mu \,+\, Q_1^{\mathrm{T}}\,\nabla f(x) \\ Q_2^{\mathrm{T}}\,\nabla f(x) \end{pmatrix}. \end{aligned}$$

Damit wird $\|r\|_2$ gerade für μ gemäß (16.20) minimal!

Beispiel (16.21)

$$B^{\mathrm{T}} \;=\; \begin{pmatrix} 1 & 1 \\ 1+\varepsilon & 1 \\ 0 & -1 \end{pmatrix}, \quad \nabla f(x^*) \;=\; \begin{pmatrix} 2 \\ 2+\varepsilon \\ -1 \end{pmatrix}, \quad \nabla f(x) \;=\; \begin{pmatrix} 2+\varepsilon \\ 2 \\ -1+\varepsilon \end{pmatrix},$$

$$(16.17) \quad \text{für} \quad x^*: \quad \begin{pmatrix} 1 & 1 \\ 1+\varepsilon & 1 \\ 0 & -1 \end{pmatrix} \mu^* \;=\; \begin{pmatrix} 2 \\ 2+\varepsilon \\ -1 \end{pmatrix}, \quad \mu^* \;=\; \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$(16.17) \quad \text{für} \quad x: \quad \begin{pmatrix} 1 & 1 \\ 1+\varepsilon & 1 \\ 0 & -1 \end{pmatrix} \mu \;=\; \begin{pmatrix} 2+\varepsilon \\ 2 \\ -1+\varepsilon \end{pmatrix}.$$

Wählt man hierbei die ersten beiden Gleichungen zur Bestimmung von μ, so erhält man $\mu = (-1,\, 3+\varepsilon)^{\mathrm{T}}$, also offenbar einen Wert, der nicht in der Nähe von μ^* liegt.

Fazit: Die Orthogonalisierungsmethode liefert im Allg. wesentlich zuverlässigere Näherungen für die Lagrange-Multiplikatoren.

B. Quadratische Optimierungsaufgaben.

Wir betrachten die obigen Reduktionsmethoden für den Spezialfall einer quadratischen Optimierungsaufgabe

$$\begin{aligned} &\text{Minimiere} \quad f(x) \;=\; \frac{1}{2}\, x^{\mathrm{T}} W\, x \;+\; g^{\mathrm{T}} x, \quad x \in \mathbb{R}^n, \\ &\text{Nebenbedingungen:} \quad h(x) \;=\; B\, x \;-\; q \;=\; 0, \end{aligned} \tag{16.22}$$

mit symmetrischer Matrix $W \in \mathbb{R}^{(n,n)}$ und $\mathrm{Rang}(B) = m$. Beachte die geänderte Notation gegenüber Abschnitt 15.

Mittels QR–Zerlegung von B^{T} werden, wie in Abschnitt A II beschrieben, Matrizen $Y \in \mathbb{R}^{(n,p)}$ und $Z \in \mathbb{R}^{(n,n-p)}$ bestimmt, die den Bedingungen (16.8) genügen.

Die reduzierte Zielfunktion lautet damit

$$\begin{aligned} \widetilde{f}(u) \;&:=\; f(Y\, q \;+\; Z\, u) \\ &=\; \frac{1}{2}\,(Y\, q + Z\, u)^{\mathrm{T}} W\,(Y\, q + Z\, u) \;+\; g^{\mathrm{T}}\,(Y\, q + Z\, u) \\ &=\; \frac{1}{2}\, u^{\mathrm{T}}\,(Z^{\mathrm{T}} W\, Z)\, u \;+\; (g \;+\; W\, Y\, q)^{\mathrm{T}}\, Z\, u \;+\; (g + \frac{1}{2}\, W\, Y\, q)^{\mathrm{T}}\, Y\, q. \end{aligned}$$

Sie ist natürlich ebenfalls quadratisch in u und besitzt – sofern die reduzierte Hesse-Matrix $Z^{\mathrm{T}} W\, Z$ positiv definit ist – ein eindeutig bestimmtes globales Minimum, welches durch das folgende lineare Gleichungssystem bestimmt werden kann

$$(Z^{\mathrm{T}} W\, Z)\, u^* \;=\; -\, Z^{\mathrm{T}}\,(g \;+\; W\, Y\, q). \tag{16.23}$$

Wir fassen die bisherigen Ergebnisse im folgenden Algorithmus zur Lösung des quadratischen Optimierungsproblems (16.22) zusammen.

Algorithmus (16.24)

1.) Berechne die QR–Zerlegung

$$B^{\mathrm{T}} \;=\; Q \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad Q \;=\; (\, Q_1 \,|\, Q_2\,) \quad \text{orthogonal}, \quad Q_1 \in \mathbb{R}^{(n,p)}.$$

Falls R (obere Dreiecksmatrix) singulär: Abbruch!

2.)

$$\begin{aligned} W_r \;&:=\; Q_2^{\mathrm{T}}\, W\, Q_2 \;\in\; \mathbb{R}^{(n-p,n-p)}, \\ q_r \;&:=\; Q_1\, R^{-\mathrm{T}}\, q \quad (=\; Y\, q), \\ g_r \;&:=\; Q_2^{\mathrm{T}}\,(W\, q_r \;+\; g). \end{aligned}$$

Bestimme die Cholesky-Zerlegung von W_r.

Falls W_r nicht positiv definit: Abbruch!.

Löse $W_r\, u \;=\; -g_r$.

3.)

$$\begin{aligned} x &:= Q_2\, u \,+\, q_r, \\ \mu &:= -\,R^{-1}\, Q_1^{\mathrm{T}}\,(W\, x \,+\, g). \end{aligned}$$

C. Ein reduziertes Quasi-Newton-Verfahren.

Wir kehren zur Lagrange-Newton-Iteration für allgemeine gleichungsrestringierte Optimierungsaufgaben zurück.

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n, \\ &\text{Nebenbedingungen:} \quad h(x) \;=\; 0 \;\in \mathbb{R}^p. \end{aligned} \tag{16.25}$$

Ist $(x,\mu) \;=\; (x^k, \mu^k)$ eine aktuelle Näherung für einen KKT-Punkt von (16.24), so lautet die Lagrange-Newton-Gleichung zur Nullstellenbestimmung von $\nabla L = 0$ nach Früherem, vgl (15.6),

$$\begin{pmatrix} W & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\mu \end{pmatrix} \;=\; - \begin{pmatrix} \nabla f(x) \,+\, B^{\mathrm{T}}\,\mu \\ h(x) \end{pmatrix}. \tag{16.26}$$

Dabei ist

$$\begin{aligned} W \;&=\; \nabla^2_{x\,x} L \;=\; \nabla^2 f(x) \,+\, \sum_{j\in E} \mu_j\, \nabla^2 h_j(x), \\ B \;&=\; h'(x) \;=\; \begin{pmatrix} \nabla h_1(x)^{\mathrm{T}} \\ \vdots \\ \nabla h_p(x)^{\mathrm{T}} \end{pmatrix}. \end{aligned}$$

Gemäß der Orthogonalisierungsmethode bilden wir die QR-Zerlegung von B^{T}

$$B^{\mathrm{T}} \;=\; Q \begin{pmatrix} R \\ 0 \end{pmatrix} \;=\; (\, Q_1 \,|\, Q_2 \,) \begin{pmatrix} R \\ 0 \end{pmatrix} \;=\; Q_1\, R$$

mit einer orthogonalen (n,n)-Matrix Q und einer oberen (p,p)-Dreiecksmatrix R.

Erweitert man Q gemäß der Partitionierung in (16.26) zu

$$\widetilde{Q} \;:=\; \left(\begin{array}{c|c} Q & 0 \\ \hline 0 & I_p \end{array}\right),$$

so ist offenbar auch $\widetilde{Q}$ orthogonal. Wir setzen weiterhin

$$\widetilde{Q}^{\mathrm{T}} \begin{pmatrix} \Delta x \\ \Delta\mu \end{pmatrix} = \begin{pmatrix} Q_1^{\mathrm{T}}\Delta x \\ Q_2^{\mathrm{T}}\Delta x \\ \Delta\mu \end{pmatrix} =: \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\mu \end{pmatrix}, \tag{16.27}$$

$$\begin{aligned} \widetilde{Q}^{\mathrm{T}} \begin{pmatrix} \nabla_x L(x,\mu) \\ h(x) \end{pmatrix} &= \begin{pmatrix} Q_1^{\mathrm{T}}\,\nabla_x L \\ Q_2^{\mathrm{T}}\,\nabla_x L \\ h \end{pmatrix} \\ &= \begin{pmatrix} Q_1^{\mathrm{T}}\big[\nabla f \,+\, (Q_1|Q_2) \begin{pmatrix} R \\ 0 \end{pmatrix} \mu\big] \\ Q_2^{\mathrm{T}}\big[\nabla f \,+\, (Q_1|Q_2) \begin{pmatrix} R \\ 0 \end{pmatrix} \mu\big] \\ h \end{pmatrix} \\ &= \begin{pmatrix} Q_1^{\mathrm{T}}\nabla f(x) \,+\, R\,\mu \\ Q_2^{\mathrm{T}}\nabla f(x) \\ h(x) \end{pmatrix}. \end{aligned} \tag{16.28}$$

Die Lagrange-Newton-Gleichung (16.26) wird wie folgt erweitert

$$\Big[\widetilde{Q}^{\mathrm{T}} \begin{pmatrix} W & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \widetilde{Q}\Big]\, \widetilde{Q}^{\mathrm{T}} \begin{pmatrix} \Delta x \\ \Delta\mu \end{pmatrix} = -\,\widetilde{Q}^{\mathrm{T}} \begin{pmatrix} \nabla_x L \\ h(x) \end{pmatrix}.$$

Multipliziert man hier die eckige Klammer aus, verwendet dabei $B^{\mathrm{T}} = Q_1\,R$ und die Orthogonalität von Q, und setzt für die restlichen Terme (16.27) und (16.28) ein, so erhält man das folgende, zu (16.26) äquivalente lineare Gleichungssystem

$$\begin{pmatrix} Q_1^{\mathrm{T}}W\,Q_1 & Q_1^{\mathrm{T}}W\,Q_2 & R \\ Q_2^{\mathrm{T}}W\,Q_1 & Q_2^{\mathrm{T}}W\,Q_2 & 0 \\ R^{\mathrm{T}} & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\mu \end{pmatrix} = -\begin{pmatrix} Q_1^{\mathrm{T}}\nabla f(x) \,+\, R\,\mu \\ Q_2^{\mathrm{T}}\nabla f(x) \\ h(x) \end{pmatrix}. \tag{16.29}$$

Die Koeffizientenmatrix in (16.29) hat nunmehr Block-Dreiecksgestalt und ist damit relativ einfach zu lösen. Ist dies erfolgt, so erhält man Δx gemäß, vgl. (16.27)

$$\Delta x \;=\; Q\,Q^{\mathrm{T}}\,\Delta x \;=\; Q_1\,Q_1^{\mathrm{T}}\Delta x \,+\, Q_2\,Q_2^{\mathrm{T}}\,\Delta x \;=\; Q_1\,\Delta x_1 \,+\, Q_2\,\Delta x_2. \tag{16.30}$$

Insgesamt haben wir nun den folgenden Algorithmus für die **reduzierte Lagrange-Newton-Iteration** hergeleitet (ein Iterationsschritt)

Algorithmus (16.31)

1.) Berechne die QR–Zerlegung von B^{T}

$$B^{\mathrm{T}} \;=\; Q \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad Q \;=\; (\, Q_1 \,|\, Q_2 \,) \quad \text{orthogonal}, \quad Q_1 \in \mathbb{R}^{(n,p)}.$$

Falls R (obere Dreiecksmatrix) singulär: Abbruch!

2.) Berechne Δx_1 aus $\quad R^{\mathrm{T}}\, \Delta x_1 \;=\; -\, h(x).$

3.) Berechne Δx_2 aus

$$(Q_2^{\mathrm{T}}\, W\, Q_2)\, \Delta x_2 \;=\; -\,[\, Q_2^{\mathrm{T}}\, \nabla f(x) \;+\; Q_2^{\mathrm{T}}\, W\, Q_1\, \Delta x_1\,]$$

4.) Setze $\Delta x \;:=\; Q_1\, \Delta x_1 \;+\; Q_2\, \Delta x_2.$

5.) Berechne μ^+ aus

$$R\, \mu^+ \;=\; -\,[\, Q_1^{\mathrm{T}}\, \nabla f(x) \;+\; Q_1^{\mathrm{T}}\, W\, \Delta x\,].$$

Bemerkung (16.32)

Aus der obigen Umformung sieht man, vgl. (16.29), dass die Koeffizientenmatrix der Lagrange-Newton-Iteration genau dann regulär ist, wenn R und $Q_2^{\mathrm{T}} W Q_2$ regulär sind.

Will man im Algorithmus (16.31) die Verwendung der zweiten Ableitungen $W = \nabla^2_{x\,x} L(x,\mu)$ vermeiden, so kann man die reduzierte Hesse-Matrix $Q_2^{\mathrm{T}}\, W\, Q_2$ durch ein BFGS-Update ersetzen

$$\begin{aligned} H_+ \;&:=\; H \;+\; \frac{y\, y^{\mathrm{T}}}{y^{\mathrm{T}} s} \;-\; \frac{(H\, s)\, (H\, s)^{\mathrm{T}}}{s^{\mathrm{T}} H\, s} \\ s \;&:=\; Q_2^{\mathrm{T}}\, (x^+ - x) \\ y \;&:=\; Q_2^{\mathrm{T}}\, (\nabla_x L(x^+, \mu^+) \;-\; \nabla_x L(x,\mu))\,. \end{aligned} \tag{16.33}$$

Zur Realisierung des Algorithmus (16.31) ohne Verwendung der zweiten Ableitungen lässt man ferner den zweiten Summanden $Q_2\, W\, Q_1\, \Delta x_1$ in der rechten Seite von Schritt 3.) fort. Ferner wird die Auswertung von $\Delta\mu$ in Schritt 5.) ersetzt durch die Berechnung von μ^+ aus der folgenden Relation, vgl. (16.20)

$$R^+\, \mu^+ \;=\; -\,(Q_1^+)^{\mathrm{T}}\, \nabla f(x^+). \tag{16.34}$$

Insgesamt rehält man nun den folgenden Algorithmus für ein **reduziertes Quasi-Newton-Verfahren**.

Algorithmus (16.35)

1.) Start: $(x,\mu)^{\mathrm{T}} = (x^0,\mu^0)^{\mathrm{T}} \in \mathbb{R}^{(n+p)}$,
$H = H_0 \in \mathbb{R}^{(n-p,n-p)}$ symmetrisch und positiv definit, $k := 0$;

2.) Berechne die QR–Zerlegung von $B^{\mathrm{T}} := h'(x)^{\mathrm{T}}$:

$$B^{\mathrm{T}} = Q \begin{pmatrix} R \\ 0 \end{pmatrix}, \quad Q = (\, Q_1 \,|\, Q_2 \,) \quad \text{orthogonal}, \quad Q_1 \in \mathbb{R}^{(n,p)};$$

Falls R (obere Dreiecksmatrix) singulär: Abbruch!

3.) Berechne Δx_1 aus $R^{\mathrm{T}}\,\Delta x_1 = -\,h(x)$;

Berechne Δx_2 aus $H\,\Delta x_2 = -\,Q_2^{\mathrm{T}}\,\nabla f(x)$;

Setze $\Delta x := Q_1\,\Delta x_1 + Q_2\,\Delta x_2, \quad x^+ := x + \Delta x$;

4.) Berechne die QR–Zerlegung von $(B^+)^{\mathrm{T}} := h'(x^+)^{\mathrm{T}}$:

$$(B^+)^{\mathrm{T}} = \left(\, Q_1^+ \,|\, Q_2^+ \,\right) \begin{pmatrix} R^+ \\ 0 \end{pmatrix}, \quad Q_1^+ \in \mathbb{R}^{(n,p)};$$

Falls R^+ singulär: Abbruch!

5.) Berechne μ^+ aus $R^+\,\mu^+ = -\,(Q_1^+)^{\mathrm{T}}\,\nabla f(x^+)$;

6.) Update der reduzierten Hesse-Matrix

$$\begin{aligned} s &:= (Q_2^+)^{\mathrm{T}}\,\Delta x \\ y &:= (Q_2^+)^{\mathrm{T}}\,(\nabla_x L(x^+,\mu^+) - \nabla_x L(x,\mu)) \\ H &:= H + \frac{y\,y^{\mathrm{T}}}{y^{\mathrm{T}}s} - \frac{(H\,s)\,(H\,s)^{\mathrm{T}}}{s^{\mathrm{T}}H\,s}; \end{aligned}$$

7.) $x := x^{k+1} := x^+; \quad \mu := \mu^+; \quad Q_i := Q_i^+,\ i = 1,2; \quad R := R^+,$

$k := k+1, \quad$ Gehe zu 3.);

Der obige Algorithmus lässt sich analog zum SQP Verfahren mit Hilfe der exakten ℓ_1-Penalty-Funktion zu einem globalen Verfahren erweitern.

17. Penalty- und Barriere-Methoden

Penalty- und Barriere–Methoden gehören zu den ältesten Ansätzen zur Lösung allgemeiner restringierter Optimierungsaufgaben.

Die grundlegende Idee besteht darin, die Verletzung der Nebenbedingungen durch Strafterme (penalty) in der Zielfunktion zu berücksichtigen.

Gegeben sei eine gleichungsrestringierte Optimierungsaufgabe der Form

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n, \\ &\text{Nebenbedingungen:} \quad h_j(x) \;=\; 0, \quad j \in E \;=\; \{1, \dots p\}. \end{aligned} \tag{17.1}$$

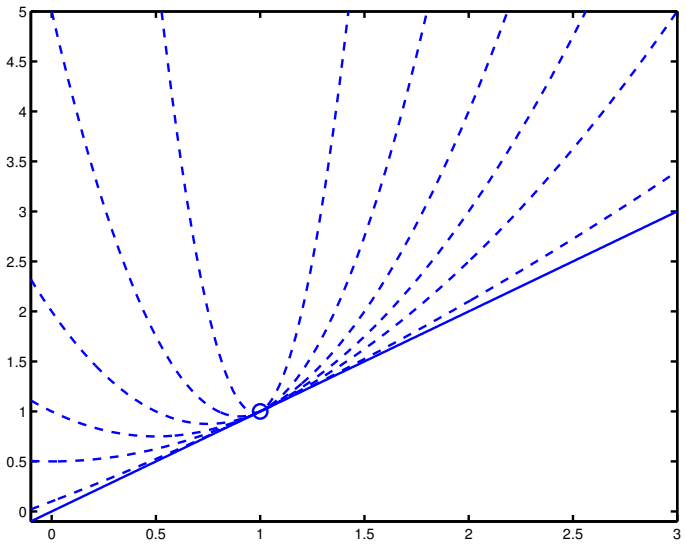
Nach **Courant (1943)** betrachtet man anstelle von (17.1) die folgende unrestringierte Optimierungsaufgabe:

$$\text{Minimiere} \quad P(x;\gamma) \;:=\; f(x) \;+\; \frac{\gamma}{2}\,\|h(x)\|^2, \quad x \in \mathbb{R}^n. \tag{17.2}$$

Dabei ist $\gamma > 0$ ein fester **Penalty–Parameter**.

Beispiel (17.3)

$$\begin{aligned} &\text{Minimiere} \quad f(x) \;:=\; x, \quad x \in \mathbb{R}, \\ &\text{Nebenbedingung:} \quad h(x) \;=\; x - 1 \;=\; 0. \end{aligned}$$



Die obige Abbildung zeigt $P(x;\gamma)$ für verschiedene Werte des Penalty-Parameters $\gamma > 0$. Die Lösung des Hilfsproblems (17.2) lautet hierfür $x_{\min} = 1 - 1/\gamma$.

Penalty–Verfahren (17.4)

Zu einer Folge von Penalty-Parametern $0 < \gamma_1 < \gamma_2 < \ldots$ mit $\lim_{k\to\infty} \gamma_k = \infty$ bestimme man jeweils ein Minimum x^k von $P(x;\gamma_k)$. Man breche das Verfahren ab, falls $\|h(x^k)\|$ hinreichend klein ist.

In jedem Iterationsschritt des obigen Verfahrens ist also eine unrestringierte Optimierungsaufgabe zu lösen. Das Penalty–Verfahren wird daher auch **SUMT (Sequential Unrestricted Minimization Technique)** genannt.

Beispiel (17.5)

$$\begin{aligned} &\text{Minimiere} \quad f(x_1,x_2) := -x_1 - x_2, \quad x = (x_1,x_2)^{\mathrm{T}} \in \mathbb{R}^2, \\ &\text{Nebenbedingung:} \quad h(x_1,x_2) = 1 - x_1^2 - x_2^2 = 0. \end{aligned}$$

Die Anwendung von (17.4) ergibt die folgende Tabelle:

γ_k	$x_1^k = x_2^k$	$h(x^k)$	$P(x^k;\gamma_k)$
10^0	0.88465	-0.5652×10^0	-1.6096
10^1	0.73089	-0.6841×10^{-1}	-1.4384
10^2	0.70959	-0.7046×10^{-2}	-1.4167
10^3	0.70736	-0.7069×10^{-3}	-1.4145
10^4	0.70713	-0.7071×10^{-4}	-1.4142
10^5	0.70711	-0.7071×10^{-5}	-1.4142
10^6	0.70711	-0.7071×10^{-6}	-1.4142

Satz (17.6) (Konvergenzsatz)

Sei $X := h^{-1}(0) \neq \emptyset$, $\gamma_k > 0$ sei strikt monoton wachsend mit $\lim_{k\to\infty} \gamma_k = \infty$. Ferner seien f, $h_j : \mathbb{R}^n \to \mathbb{R}$ stetig, f sei auf X nach unten beschränkt und das Hilfsproblem (17.2) sei für alle $\gamma > 0$ eindeutig lösbar (globales Minimum). Dann gelten:

a) Die Folge $P(x^k;\gamma_k)$ wächst monoton.

b) Die Folge $\|h(x^k)\|$ fällt monoton.

c) Die Folge $f(x^k)$ wächst monoton.

d) $h(x^k) \to 0$ $(k \to \infty)$.

e) Jeder Häufungspunkt der Folge (x^k) löst die restringierte Optimierungsaufgabe (17.1).

Beweis:

Aus der vorausgesetzten Monotonie der γ_k erhält man

$$P(x^k;\gamma_k) \;\le\; P(x^{k+1};\gamma_k) \;\le\; P(x^{k+1};\gamma_{k+1}) \;\le\; P(x^k;\gamma_{k+1}). \qquad (*)$$

Hieraus folgt insbesondere die Behauptung a). Die erste und die letzte Ungleichung in $(*)$ lauten explizit:

$$\begin{aligned} f(x^k) \;+\; \frac{\gamma_k}{2}\,\|h(x^k)\|^2 \;&\le\; f(x^{k+1}) \;+\; \frac{\gamma_k}{2}\,\|h(x^{k+1})\|^2 \\ f(x^{k+1}) \;+\; \frac{\gamma_{k+1}}{2}\,\|h(x^{k+1})\|^2 \;&\le\; f(x^k) \;+\; \frac{\gamma_{k+1}}{2}\,\|h(x^k)\|^2 \end{aligned}$$

Addition dieser beiden Ungleichungen ergibt

$$\frac{\gamma_k}{2}\,\|h(x^k)\|^2 \;+\; \frac{\gamma_{k+1}}{2}\,\|h(x^{k+1})\|^2 \;\le\; \frac{\gamma_k}{2}\,\|h(x^{k+1})\|^2 \;+\; \frac{\gamma_{k+1}}{2}\,\|h(x^k)\|^2$$

und damit

$$\frac{1}{2}\,(\gamma_k - \gamma_{k+1}) \cdot \big(\|h(x^k)\|^2 - \|h(x^{k+1})\|^2\big) \;\le\; 0,$$

woraus sich wegen der strikten Monotonie der γ_k die Behauptung b) ergibt.

Aus der ersten Ungleichung von $(*)$ folgt damit nun auch $f(x^k) \le f(x^{k+1})$, also die Behauptung c).

Schließlich gilt aufgrund der Beschränktheit von f

$$P(x^k;\gamma_k) \;=\; \min\{f(x) + \frac{\gamma_k}{2}\,\|h(x)\|^2 :\; x \in \mathbb{R}^n\} \;\le\; \inf\{f(x) :\; x \in X\} \;=:\; f_{\min} \quad (**)$$

Damit ist also

$$f(x^k) + \frac{\gamma_k}{2}\,\|h(x^k)\|^2 \;\le\; f_{\min}$$

und somit

$$\gamma_k\,\|h(x^k)\|^2 \;\le\; 2\,(f_{\min} - f(x^k)) \;\le\; 2\,(f_{\min} - f(x^1)).$$

Da $\lim_{k\to\infty} \gamma_k = \infty$ vorausgesetzt wurde, ergibt sich hieraus: $h(x^k) \to 0$ $(k \to \infty)$. Damit ist auch die Behauptung d) gezeigt.

Ist nun x^* ein Häufungspunkt von (x^k), so gilt wegen d) $h(x^*) = 0$, also $x^* \in X$ und daher $f(x^*) \ge f_{\min}$.

Andererseits gilt für die Teilfolge $x^{k_j} \to x^*$ nach $(**)$

$$f(x^*) \;=\; \lim_{j\to\infty} \Big[f(x^{k_j}) + \frac{\gamma_{k_j}}{2}\,\|h(x^{k_j})\|^2\Big] \;\le\; f_{\min}.$$

Damit ist also $f(x^*) = f_{\min}$, d.h. x^* löst die restringierte Optimierungsaufgabe (17.1). □

Bemerkung (17.7)

Es sei angemerkt, dass Satz (17.6) ohne Differenzierbarkeits- oder Regularitätsvoraussetzungen auskommt. Es wird lediglich die Stetigkeit der Funktionen f und h vorausgesetzt.

Lagrange–Multiplikatoren.

Sind die Funktionen f und h hinreichend glatt, so folgt für das Hilfsproblem (17.2)

$$\nabla P(x^k;\gamma_k) \;=\; \nabla f(x^k) \,+\, \gamma_k \,\sum_j h_j(x^k)\cdot\nabla h_j(x^k) \;=\; 0.$$

Der Vergleich mit der notwendigen Bedingung erster Ordnung (12.8)

$$\nabla L(x^*;\mu) \;=\; \nabla f(x^*) \,+\, \sum_j \mu_j\cdot\nabla h_j(x^*) \;=\; 0$$

legt es nahe, die Größen

$$\mu_j^k \;:=\; \gamma_k\cdot h_j(x^k), \quad j=1,\ldots,p, \tag{17.8}$$

als Näherungen für die Lagrange–Multiplikatoren zu verwenden.

Für das Beispiel (17.5) ergibt sich mit $\gamma_6 = 10^{-6}$ die Näherung $\mu^6 = \gamma_6\cdot h(x^6) = -0.7071$. Diese stimmt für die angegebenen Dezimalstellen mit dem zugehörigen Lagrange-Multiplikator $\mu = -1/\sqrt{2}$ überein.

Satz (17.9) (Lagrange–Multiplikatoren)

Sind f und h stetig differenzierbar, gelten die Voraussetzungen von Satz (17.6) mit $\lim_{k\to\infty} x^k = x^*$ und ist x^* ein regulärer Punkt, so konvergieren die Näherungen μ^k gegen den (eindeutig bestimmten) Lagrange–Multiplikator: $\lim_{k\to\infty}\mu^k = \mu$.

Beweis

Mit der (p,n)–Matrix $B_k := h'(x^k)$ folgt mit Hilfe der notwendigen Bedingung $\nabla P(x^k;\gamma_k) = 0$:

$$B_k^{\mathrm{T}}\mu^k \;=\; -\nabla f(x^k).$$

Multilikation mit B_k ergibt

$$(B_k\,B_k^{\mathrm{T}})\,\mu^k \;=\; -B_k\,\nabla f(x^k).$$

Die Matrix $B := h'(x^*)$ hat nun wegen der vorausgesetzten Regularität von x^* maximalen Rang. Aus Stetigkeitsgründen gilt damit für hinreichend große Indizes

k auch $\text{Rang}\, B_k = \text{Rang}\, B_k^{\mathrm{T}} = p$. Für hinreichend große k ist $B_k\, B_k^{\mathrm{T}}$ also regulär. Damit folgt

$$\begin{aligned} \mu^k &= -(B_k\, B_k^{\mathrm{T}})^{-1}\, B_k\, \nabla f(x^k) \\ &\to -(B\, B^{\mathrm{T}})^{-1}\, B\, \nabla f(x^*) \qquad (k \to \infty). \end{aligned}$$

Genauso folgt aber aus der notwendigen Bedingung erster Ordnung (12.8)

$$\nabla L(x^*; \mu) \;=\; \nabla f(x^*) \;+\; \sum_j \mu_j \cdot \nabla h_j(x^*) \;=\; 0$$

die Relation $B^{\mathrm{T}}\mu = -\nabla f(x^*)$ und damit

$$\mu \;=\; -(B\, B^{\mathrm{T}})^{-1}\, B\, \nabla f(x^*)$$

□

Bemerkung (17.10)

Ohne Beweis seien die folgenden Fehlerabschätzungen angegeben:

Sind f und h C^2–Funktionen und ist $x^* \in X$ ein regulärer Punkt, der die hinreichenden Bedingungen erfüllt, vgl. (12.10), so gelten

$$\begin{aligned} f(x^*) &= f(x^k) \;+\; \frac{1}{2}\,\gamma_k\, \|h(x^k)\|^2 \;+\; o(1/\gamma_k), \\ x^k \;-\; x^* &= \frac{1}{\gamma_k}\, c \;+\; o(1/\gamma_k), \qquad c \in \mathbb{R}^n. \end{aligned}$$

Grenzen der Penalty-Verfahren.

Um den Nebenbedingungen Geltung zu verschaffen, muss der Penalty–Parameter γ_k im Allg. sehr groß sein. Dadurch werden die Hilfsprobleme (17.2) allerdings für größere k zunehmend schlechter konditioniert. Dies kann man folgendermaßen sehen:

Für die Hesse–Matrix der Penalty–Funktion $P(x;\gamma)$ ergibt sich bei hinreichender Glattheit der Funktionen f und h:

$$\nabla^2 P(x;\gamma) \;=\; \nabla^2 f(x) \;+\; \gamma\, \Big[\sum_{j\in E} h_j(x)\, \nabla^2 h_j(x) \;+\; B^{\mathrm{T}} B\Big] \;=\; W \;+\; \gamma\, (B^{\mathrm{T}} B)$$

mit $B := h'(x)$ und $W := W(x) := \nabla^2 f(x) \;+\; \sum_{j\in E} (\gamma h_j(x))\, \nabla^2 h_j(x)$.

Für hinreichend große Indizes k und $x = x^k,\ \gamma = \gamma_k$ hat $B = B_k$ maximalen Rang $p < n$ und es gilt $W = W(x_k) \approx W^* = W(x^*)$. Damit hat man

$$\nabla^2 P(x^k; \gamma_k) \;\approx\; W^* \;+\; \gamma_k\, (B_k^{\mathrm{T}} B_k).$$

Die Matrix $B_k^{\mathrm{T}} B_k \in \mathbb{R}^n$ hat ebenfalls den Rang $p < n$. Wegen $\gamma_k \to \infty$ konvergieren somit genau p Eigenwerte der Hesse–Matrix $\nabla^2 P(x^k; \gamma_k)$ dem Betrage nach gegen ∞, während die anderen Eigenwerte beschränkt bleiben. Daher folgt:

$$\text{cond}_2 \nabla^2 P(x^k; \gamma_k) \;\to\; \infty \qquad (k \to \infty).$$

Hiermit ist die schlechte Kondition der Hilfsprobleme (17.2) für größere Werte von γ_k belegt. Ferner ist damit auch die Verwendung von **Homotopiemethoden**, wie in (17.4) beschrieben, zu begründen. Man startet mit einem moderaten Wert γ_1 und vergrößert γ_k dann schrittweise, wobei jeweils die Lösung x^k als Startwert zur numerischen Bestimmung der Lösung für γ_{k+1} verwendet wird. Dies macht das Verfahren insgesamt relativ aufwendig.

Ungleichungsrestriktionen.

Für die allgemeine Optimierungsaufgabe

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n \\ &\text{unter den Nebenbedingungen} \\ &g_i(x) \;\leq\; 0, \quad i \in I := \{1, \ldots, m\}, \\ &h_j(x) \;=\; 0, \quad j \in E := \{1, \ldots, p\}, \end{aligned} \tag{17.11}$$

lässt sich beispielsweise die folgende (klassische) Penalty–Funktion verwenden:

$$P(x;\gamma) \;=\; f(x) \;+\; \frac{\gamma}{2}\Big\{\sum_{i\in I} \max(0, g_i(x))^2 \;+\; \sum_{j\in E} h_j(x)^2\Big\} \tag{17.12}$$

Diese ist allerdings selbst für glatte Funktionen f, g und h im Allg. lediglich stetig differenzierbar. Es gilt zwar auch hierfür der Konvergenzsatz (17.6) analog, allerdings hat man aufgrund der fehlenden Glattheit mit numerischen Schwierigkeiten bei der Minimierung der Penalty–Funktion zu rechnen.

Barriere–Methoden, Innere–Punkte–Verfahren.

Wir betrachten o.B.d.A. eine Optimierungsaufgabe für die nur Ungleichungsrestriktionen gegeben sind:

$$\begin{aligned} &\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n \\ &\text{unter den Nebenbedingungen} \\ &g_i(x) \;\leq\; 0, \quad i \in I := \{1, \ldots, m\}, \end{aligned} \tag{17.13}$$

Barriere–Methoden sind dadurch gekennzeichnet, dass anstelle eines Penalty–Terms, der die Verletzung der Nebenbedingungen bestraft, versucht wird, durch einen Barriere–Term in der Zielfunktion das Verlassen des zulässigen Bereichs gänzlich zu verhindern. Ein Barriere–Verfahren arbeitet daher – im Gegensatz zu den Penalty–Methoden – nur mit zulässigen Punkten. Daher kommt auch der Name: *Innere–Punkte–Verfahren.*

Als gebräuchliche Barriere–Funktionen sind zu nennen:

a) Logarithmische Barriere-Funktion (Frisch, 1955)

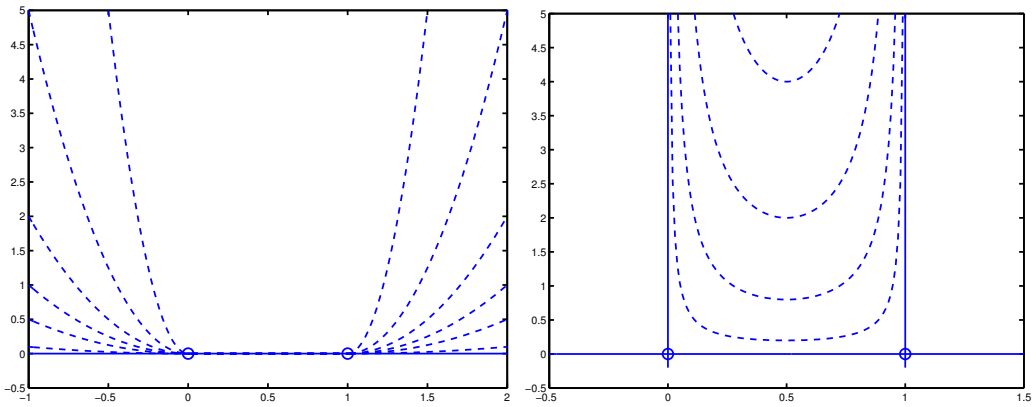
$$B_1(x;\gamma) \;=\; f(x) \;-\; \frac{1}{\gamma}\sum_{i\in I}\ln[-g_i(x)], \tag{17.14}$$

b) Inverse Barriere-Funktion (Carroll, 1961)

$$B_2(x;\gamma) \;=\; f(x) \;-\; \frac{1}{\gamma}\sum_{i\in I}\frac{1}{g_i(x)}. \tag{17.15}$$

Sinnvoll anwenden lassen sich solche Barriere-Verfahren nur auf Optimierungsaufgaben mit in gewissem Sinn *robusten* zulässigen Mengen X. Das soll heißen, dass jeder zulässige Punkt auch ein Häufungspunkt innerer Punkte der zulässigen Menge ist.

Das folgende Bild zeigt das qualitative Verhalten des Penalty-Terms und des Barriere-Terms für eine unabhängige Variable ($n = 1$) und den Restriktionen $g_1(x) := -x \le 0$ und $g_2(x) := x - 1 \le 0$.



Man beachte, dass der Barriere-Term nur für innere Punkte des zulässigen Bereichs, hier $X = [0, 1]$, definiert ist.

Analog zur Homotopietechnik bei den Penalty–Methoden, vgl. (17.4), wird wieder zu einer Folge

$$0 \;<\; \gamma_1 \;<\; \gamma_2 \;<\; \gamma_3 \;\ldots, \qquad \lim_{k\to\infty}\gamma_k \;=\; \infty$$

jeweils die unrestringierte Optimierungsaufgabe

$$x^k \;:=\; \operatorname{argmin} B(x;\gamma_k) \tag{17.16}$$

gelöst, wobei mit einem zulässigen Punkt $x^1 \in \operatorname{int}(X)$ begonnen wird.

Analog zum Konvergenzsatz (17.6) lässt sich für stetige Funktionen f und g zeigen, dass jeder Häufungspunkt der Folge x^k das Ausgangsproblem (17.13) löst, wobei natürlich die Lösbarkeit der Hilfsprobleme (17.16) vorausgesetzt werden muss.

Auch bei dieser Vorgehensweise werden die unrestringierten Teilprobleme (17.16) mit größer werdendem Barriere-Parameter γ_k zunehmend schlecht konditioniert. Die Singularitäten von $B(x;\gamma_k)$ am Rand des zulässigen Bereichs erzwingt ferner besondere Vorkehrungen bei der Schrittweitenbestimmung zur Lösung der unrestringierten Optimierungsprobleme.

Lineare Optimierung.

Wir betrachten eine lineare Optimierungsaufgabe in Standardform

$$\begin{aligned} &\text{Minimiere} \quad f(x) \;=\; c^{\mathrm{T}}x, \quad x \in \mathbb{R}^n \\ &\text{unter den Nebenbedingungen} \\ &A\,x \;=\; b, \quad x \;\geq\; 0. \end{aligned} \tag{17.17}$$

Behandelt man lediglich die Ungleichungsnebenbedingungen $x_i \geq 0$ mit der logarithmischen Barriere–Funktion, so ergibt sich das Ersatzproblem $(\gamma > 0)$:

$$\begin{aligned} &\text{Minimiere} \quad B(x;\gamma) \;=\; c^{\mathrm{T}}x \;-\; \frac{1}{\gamma}\,\sum_{i=1}^{n} \ln[x_i] \\ &\text{unter der Nebenbedingung} \quad A\,x \;=\; b. \end{aligned} \tag{17.18}$$

Eine unrestringierte Optimierungsaufgabe erhält man hieraus durch Anwendung von Reduktionsmethoden.

Wir stellen nun die KKT–Bedingungen für die restringierte Optimierungsaufgabe (17.18) auf. Die zugehörige Lagrange-Funktion lautet

$$L(x,\mu;\gamma) \;=\; c^{\mathrm{T}}x \;-\; \frac{1}{\gamma}\,\sum_{i=1}^{n} \ln[x_i] \;+\; \mu^{\mathrm{T}}(b - A\,x).$$

Damit ist nach (13.9) eine notwendige Bedingung gegeben durch

$$\frac{\partial L}{\partial x_i} \;=\; c_i \;-\; \frac{1}{\gamma\,x_i} \;-\; (A^{\mathrm{T}}\mu)_i \;=\; 0.$$

Setzt man nun $\lambda_i \;:=\; \lambda_i(\gamma) \;:=\; 1/(\gamma\,x_i)$, so folgt mit (13.9)

$$\begin{aligned} &A^{\mathrm{T}}\mu \;+\; \lambda \;=\; c, \\ &A\,x \;=\; b, \\ &x \;>\; 0, \quad \lambda \;>\; 0, \\ &\forall\, i = 1,\ldots,n: \quad x_i \cdot \lambda_i \;=\; \frac{1}{\gamma} \;=:\; \tau \;>\; 0. \end{aligned} \tag{17.19}$$

Diese Bedingungen können wir nun interpretieren als die mit dem Parameter $\tau > 0$ gestörten KKT–Bedingungen des Ausgangsproblems (17.17). Diese lauten nämlich (ebenfalls mit (13.9))

$$\begin{aligned} A^{\mathrm{T}}\mu + \lambda &= c, \\ A\,x &= b, \\ x \geq 0, \quad \lambda &\geq 0, \\ \forall\, i = 1, \ldots, n: \quad x_i \cdot \lambda_i &= 0. \end{aligned} \tag{17.20}$$

Hat (17.19) für alle $\tau > 0$ eine Lösung $(x(\tau), \mu(\tau), \lambda(\tau))$, so heißt die Abbildung

$$\tau \;\mapsto\; (x(\tau), \mu(\tau), \lambda(\tau))$$

der **zentrale Pfad** der linearen Optimierungsaufgabe (17.17).

Der logarithmische Barriere–Ansatz beschreibt also eine Homotopie von $\tau_1 = 1/\lambda_1 > 0$ zu $\tau_\infty = 0$ längs des zentralen Pfades.

Exakte Penalty–Funktionen.

Die klassische Penalty–Funktion $P(x;\gamma)$, vgl. (17.12), für eine allgemeine restringierte Optimierungsaufgabe hat den Nachteil, dass zur Approximation der Lösung x^* ein Homotopieverfahren mit $\gamma_k \to \infty$ benötigt wird. Für große Werte von γ_k bereitet dann die schlechte Kondition der Penalty–Hilfsprobleme Schwierigkeiten.

Verzichtet man dagegen auf die C^1–Eigenschaft der Penalty–Funktion, so lassen sich zumindest für konvexe Optimierungsaufgaben Penalty–Funktionen finden, deren Minimum bereits für hinreichend große, aber *endliche* γ_k mit dem Minimum des Ausgangsproblems übereinstimmt. Solche Penalty–Funktionen heißen **exakt**.

Eine Klasse exakter Penalty–Funktionen ist durch die so genannten ℓ_q–**Penalty–Funktionen** gegeben:

$$P_q(x;\gamma) \;=\; f(x) + \gamma \left\| \begin{pmatrix} h(x) \\ g(x)_+ \end{pmatrix} \right\|_q. \tag{17.21}$$

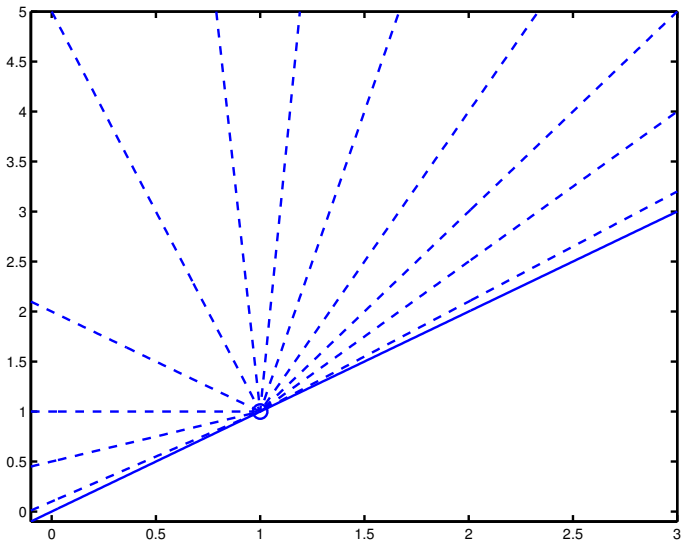
Dabei ist $g(x)_+ := (\max(0, g_1(x)), \ldots, \max(0, g_m(x)))^{\mathrm{T}}$ und $\|\cdot\|_q$ bezeichnet die übliche q–Norm, $1 \leq q \leq \infty$.

Beispiel (17.22)

Fur das bereits in (17.3) betrachtete Beispiel

$$\begin{aligned} &\text{Minimiere} \quad f(x) := x, \quad x \in \mathbb{R}, \\ &\text{Nebenbedingung:} \quad h(x) = x - 1 = 0. \end{aligned}$$

ergibt sich für $q = 1$ und $q = 2$ die folgende Darstellung der exakten Penalty–Funktion



Satz (17.24) (Exaktheit der ℓ_1–Penalty–Funktion)

Seien f und g stetig differenzierbare, konvexe Funktionen und sei h affin–linear. Ferner sei (x^*, λ^*, μ^*) ein KKT–Punkt der *konvexen* Optimierungsaufgabe (17. 11). Dann gilt:
Es gibt ein $\gamma^* > 0$, so dass x^* für alle $\gamma \geq \gamma^*$ ein Minimum der ℓ_1–Penalty–Funktion $P_1(x, \gamma)$ ist.

Beweis

Die Lagrange–Funktion der Optimierungsaufgabe lautet

$$L(x, \lambda, \mu) \;=\; f(x) \;+\; \sum_{i \in I} \lambda_i \, g_i(x) \;+\; \sum_{j \in E} \mu_j \, h_j(x).$$

Damit hat man die folgenden KKT–Bedingungen für (x^*, λ^*, μ^*):

$$\begin{aligned} \nabla_x L(x^*, \lambda^*, \mu^*) \;&=\; 0, \\ h(x^*) \;&=\; 0, \\ \lambda_i^* \;\geq\; 0, \quad g_i(x^*) \;&\leq\; 0, \\ \lambda_i^* \cdot g_i(x^*) \;&=\; 0, \quad i = 1, \ldots, m. \end{aligned}$$

Da die Lagrange–Funktion $L(\cdot\,, \lambda^*, \mu^*)$ als Funktion von x nach Voraussetzung konvex ist und x^* stationärer Punkt dieser Funktion ist, folgt nach (3.4) b)

$$\forall\, x \in \mathbb{R}^n : \quad L(x^*, \lambda^*, \mu^*) \;\leq\; L(x, \lambda^*, \mu^*).$$

Setzt man nun $\gamma^* := \| \begin{pmatrix} \lambda^* \\ \mu^* \end{pmatrix} \|_\infty$, so folgt für $\gamma \geq \gamma^*$ und $x \in \mathbb{R}^n$

$$
\begin{aligned}
P_1(x^*,\gamma) &= f(x^*) \,+\, \gamma \Big\{ \sum_{i\in I} \max(0, g_i(x^*)) \,+\, \sum_{j\in E} |h_j(x^*)| \Big\} \\
&= f(x^*) \\
&= f(x^*) \,+\, \sum_{i\in I} \lambda_i^*\, g_i(x^*) \,+\, \sum_{j\in E} \mu_j^*\, h_j(x^*) \\
&\le f(x) \,+\, \sum_{i\in I} \lambda_i^*\, g_i(x) \,+\, \sum_{j\in E} \mu_j^*\, h_j(x) \\
&\le f(x) \,+\, \sum_{i\in I} \lambda_i^*\, \max(0, g_i(x)) \,+\, \sum_{j\in E} |\mu_j^*|\, |h_j(x)| \\
&\le f(x) \,+\, \gamma \Big\{ \sum_{i\in I} \max(0, g_i(x)) \,+\, \sum_{j\in E} |h_j(x)| \Big\} \\
&= P_1(x;\gamma).
\end{aligned}
$$

Damit ist gezeigt, dass x^* ein globales Minimum von $P_1(x;\gamma)$ ist. □

Multiplier–Penalty–Funktionen.

Hiermit sollen glatte und exakte Penalty–Funktionen konstruiert werden. Dies gelingt allerdings nur dadurch, dass man die Penalty–Funktion in Abhängigkeit von den Lagrange–Multiplikatoren der Optimierungsaufgabe konstruiert.

Wir betrachten zunächst wieder eine gleichungsrestringierte Optimierungsaufgabe:

$$
\begin{aligned}
&\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n \\
&\text{unter den Nebenbedingungen} \\
&h_j(x) \;=\; 0, \quad j \in E \,:=\, \{1,\ldots,p\}.
\end{aligned}
\tag{17.25}
$$

Wir stellen hierzu die klassische Penalty–Funktion auf, verschieben allerdings die Penalty–Terme geeignet:

$$
\widetilde{P}(x;\sigma,\gamma) \;:=\; f(x) \,+\, \frac{\gamma}{2} \sum_{j=1}^{p} \big(h_j(x) + \sigma_j\big)^2, \quad x \in \mathbb{R}^n.
$$

Multipliziert man dies aus und lässt man den von x unabhängigen Summanden weg, so ergibt sich

$$
P(x;\sigma,\gamma) \;:=\; f(x) \,+\, \sum_{j=1}^{p} \big(\gamma\,\sigma_j\big)\, h_j(x) \,+\, \frac{\gamma}{2}\, \|h(x)\|^2, \quad x \in \mathbb{R}^n.
$$

Mit der Definition $\mu_j := \gamma\,\sigma_j$ entsprechen die ersten beiden Summanden der obigen modifiziertem Penalty–Funktion gerade der Lagrange–Funktion der Optimierungsaufgabe. Daher setzt man

$$
L_\gamma(x;\mu,\gamma) \;:=\; f(x) \,+\, \mu^{\mathrm{T}} h(x) \,+\, \frac{\gamma}{2}\, \|h(x)\|^2, \quad x \in \mathbb{R}^n. \tag{17.26}
$$

Die Funktion L_γ heißt *erweiterte Lagrange–Funktion* oder *Multiplier–Penalty–Funktion* der gleichungsrestringierten Optimierungsaufgabe (17.25).

Wir sind daran interessiert, dass die Minimierung der Multiplier–Penalty–Funktion $L_\gamma(x;\mu,\gamma)$ bereits für *endliches*, hinreichend großes γ die exakte Lösung x^* des Ausgangsproblems liefert. Da aber x^* unter Konvexitätsvoraussetzungen die exakte Lagrange–Funktion $L(x;\mu^*)$ minimiert, erscheint es sinnvoll zu sein, auch in (17.26) den exakten Lagrange–Multiplikator $\mu = \mu^*$ einzusetzen.

Wir zeigen nun, dass $L_\gamma(x;\mu^*,\gamma)$ tatsächlich eine exakte Penalty–Funktion ist. Hierzu brauchen wir zunächst den folgenden Hilfssatz:

Lemma (17.27)

Ist $Q \in \mathbb{R}^{(n,n)}$ symmetrisch und positiv semidefinit und ist $W \in \mathbb{R}^{(n,n)}$ symmetrisch und positiv definit auf dem Kern der Matrix Q, so existiert ein $\gamma^* > 0$ mit

$$\forall\ \gamma\ \geq\ \gamma^* :\quad (W + \gamma\, Q)\ \text{positiv definit auf } \mathbb{R}^n.$$

Beweis (indirekt)

Wir nehmen an, es gibt eine Folge $\gamma_k > 0$ mit $\lim_{k\to\infty} \gamma_k\ =\ \infty$, sowie

$$\forall\, k :\ \exists\ x^k \in \mathbb{R}^n \setminus \{0\} :\quad (x^k)^{\mathrm{T}}(W + \gamma_k\, Q)\, x^k\ \leq\ 0. \tag{1}$$

O.B.d.A. kann die Folge (x^k) normiert werden zu $\|x^k\| = 1$. Ferner lässt sich durch Übergang auf eine Teilfolge erreichen, dass die Folge (x^k) gegen einen Einheitsvektor konvergiert: $x^k \to x^* \neq 0\ (k \to \infty)$.

Die Beziehung (1) lautet ausmultipliziert:

$$\forall\, k :\ \Big((x^k)^{\mathrm{T}} W\, x^k\Big)\ +\ \gamma_k \Big((x^k)^{\mathrm{T}} Q\, x^k\Big)\ \leq\ 0. \tag{2}$$

Der erste Summand konvergiert gegen den festen Wert $(x^*)^{\mathrm{T}} W\, x^*$.

Da weiter $\gamma_k \to \infty$ und nach Voraussetzung $(x^k)^{\mathrm{T}} Q\, x^k \geq 0$ gilt, folgt notwendigerweise $(x^*)^{\mathrm{T}} Q\, x^* = 0$.

Da Q als positiv semidefinit vorausgesetzt wurde, ist x^* also ein globales Minimum der quadratischen Funktion $q(x)\ :=\ \frac{1}{2}\, x^{\mathrm{T}} Q x$ und es folgt $\nabla q(x^*) = Q\, x^* = 0$.

Der Einheitsvektor $x^* \neq 0$ liegt also im Kern der Matrix Q. Somit gilt nach Voraussetzung $(x^*)^{\mathrm{T}} W\, x^*\ >\ 0$, was der Beziehung (2) widerspricht! □

Satz (17.28) (Exaktheit von $L_\gamma(x;\mu^*,\gamma)$)

Sei (x^*,μ^*) ein KKT–Punkt der Optimierungsaufgabe (17.25), der die hinreichenden Bedingungen zweiter Ordnung erfüllt, vgl. (12.10).
Dann existiert ein $\gamma^* > 0$, so dass x^* für jedes $\gamma \geq \gamma^*$ ein lokales Minimum der Multiplier–Penalty–Funktion $L_\gamma(\cdot\,;\mu^*,\gamma)$ ist.

Beweis

Wir überprüfen die hinreichenden Bedingungen zweiter Ordnung für die Minimierung von $x \mapsto L_\gamma(x;\mu^*,\gamma)$.

Zunächst gilt mit der Zulässigkeit von x^* und den KKT–Bedingungen:

$$\nabla_x L_\gamma(x^*;\mu^*,\gamma) \;=\; \nabla_x L(x^*,\mu^*) \;+\; \gamma \sum_{j=1}^{p} h_j(x^*)\,\nabla h_j(x^*) \;=\; 0.$$

Weiter folgt für die Hesse–Matrix

$$\begin{aligned}\nabla^2_{xx} L_\gamma(x^*;\mu^*,\gamma) &= \nabla^2_{xx} L(x^*,\mu^*) \;+\; \gamma \sum_{j=1}^{p} \Big(h_j(x^*)\,\nabla^2 h_j(x^*) + \nabla h_j(x^*)\,\nabla h_j(x^*)^{\mathrm{T}} \Big) \\ &= W \;+\; \gamma\, B^{\mathrm{T}} B,\end{aligned}$$

mit $W := \nabla^2_{xx} L(x^*,\mu^*)$ und $B := h'(x^*) \in \mathbb{R}^{(p,n)}$.

Die Matrix $Q := B^{\mathrm{T}}B \in \mathbb{R}^{(n,n)}$ ist offenbar symmetrisch und positiv semidefinit. Nun gilt weiter: $\mathrm{Kern}(B) = \mathrm{Kern}(Q)$. Damit ist W aufgrund der KKT–Bedingungen auch positiv definit auf $\mathrm{Kern}(Q)$.

Die Voraussetzungen von Lemma (17.27) sind also erfüllt, d.h. für hinreichend große $\gamma \geq \gamma^*$ ist die Hesse–Matrix $\nabla^2_{xx} L_\gamma(x^*;\mu^*,\gamma)$ positiv definit.

Damit ist gezeigt, dass x^* für solche γ ein striktes lokales Minimum von L_γ ist. □

Um die Exaktheit des Multiplier–Penalty Ansatzes ausnutzen zu können benötigt man jedoch eine hinreichend gute Approximation des Lagrange–Multiplikators μ^*.

Dazu sei $x^{k+1} := \mathrm{argmin}\big\{ L_\gamma(x;\mu^k,\gamma) \;:\; x \in \mathbb{R}^n \big\}$. Dann gilt notwendigerweise

$$\begin{aligned} 0 &= \nabla_x L_\gamma(x^{k+1};\mu^k,\gamma) \\ &= \nabla f(x^{k+1}) \;+\; \sum_{j=1}^{p} (\mu_j^k + \gamma\, h_j(x^{k+1}))\,\nabla h_j(x^{k+1}).\end{aligned}$$

Vergleicht man dies mit der KKT–Bedingung $0 = \nabla f(x^*) + \sum_{j=1}^{p} \mu_j^* \nabla h_j(x^{k+1})$, so wird die folgende Aufdatierung für die Näherungen der Lagrange–Multiplikatoren nahegelegt:

$$\mu^{k+1} \;:=\; \mu^k \;+\; \gamma\, h(x^{k+1}). \tag{17.29}$$

(17.29) heißt auch die **Aufdatierungsformel von Hestenes und Powell**.

Insgesamt ergibt sich damit der folgende Algorithmus für das Multiplier–Penalty–Verfahren:

Algorithmus (17.30)

1.) Wähle $(x^0, \mu^0)^{\mathrm{T}} \in \mathbb{R}^{(n+p)}$, $\gamma_0 > 0$, $c \in]0,1[$, $\mathrm{TOL} > 0$, $k := 0$.

2.) Falls (x^k, μ^k) ein KKT–Punkt ist: Stop!

3.) Berechne: $x^{k+1} := \operatorname{argmin}\big\{ L_\gamma(x; \mu^k, \gamma) \;:\; x \in \mathbb{R}^n \big\}$.

4.) Setze $\mu^{k+1} := \mu^k + \gamma_k\, h(x^{k+1})$.

5.) Setze $\gamma_{k+1} := \gamma_k$,
falls $\|h(x^{k+1})\| \geq c\, \|h(x^k)\|$: $\gamma_{k+1} := 10 \cdot \gamma_k$.

6.) Setze $k := k+1$; gehe zu 2.) .