Manuskript zur Vorlesung
Nichtlineare Optimierung
SS 1993

Uwe Zimmermann
Abteilung Mathematische Optimierung
Institut für Angewandte Mathematik
Technische Universität Braunschweig

3. Juli 1996

Inhaltsverzeichnis

Kapitel 1

Einführung und Übersicht

Wir haben in der Einführungsvorlesung bereits nichtlineare Optimierungsprobleme ohne Restriktionen kennengelernt. In dieser Vorlesung behandeln wir nichtlineare Optimierungsaufgaben der Form

$$\min \left\{ f(x) \;\middle|\; x \in P \right\}, \tag{1.1}$$

wobei $P \subseteq \mathbf{R}^n$ eine abgeschlossene Menge ist. Die Funktion $f\colon P \to \mathbf{R}$ heißt **Zielfunktion**. Natürlich sind auch die aus der Einführungsvorlesung bekannten Linearen Optimierungsprobleme (LPs) ein Spezielfall von (1.1), hier ist f linear und P ein Polyeder.

Aus dem großen Gebiet nichtlinearer Optimierungsaufgaben stellen wir zunächst einige spezielle Problemtypen durch Beispiele vor.

Im zweiten Kapitel werden wir Verfahren zur nichtlinearen Optimierung ohne Restriktionen kennenlernen. Das dritte Kapitel führt schließlich in die Optimierung mit Nebenbedingungen ein.

1.1 Beispiele

Ausgleichsrechnung

(aus: Krabs: Nichtlineare Optimierung für Ingenieure). Gegeben sei folgende Abhängigkeit:

$$s(t) = \underbrace{\frac{1}{2} g \sin(\alpha)}_{x_1} t^2 + \underbrace{v_0}_{x_2} t + \underbrace{s_0}_{x_3}$$

Die x_i sind Unbekannte; sie sollen durch mehrmalige (fehlerbehaftete) Messung der Größen s und t bestimmt werden. Tritt bei einer Messung der Fehler ε_i auf, der als Fehler von s_i interpretiert wird (d. h. die Messung von t_i wird als exakt angenommen), so erhält man bei n Messungen ein System von Gleichungen der Form

$$s_i = x_1 t_i^2 + x_2 t_i + x_3 + \varepsilon_i \quad (1 \leqslant i \leqslant n).$$

Unter der Annahme, daß die Meßfehler klein sind, erhält man die folgende Optimierungsaufgabe: $\min \sum_{i=1}^{n} \varepsilon_i^2$ unter $x \in \mathbf{R}^3$.

(Minimierung einer quadratischen Zielfunktion ohne Restriktionen.)

Angebotsauswertung

Eine Fabrik benötigt eine Menge a eines bestimmten Rohstoffs, der von verschiedenen Zulieferern i angeboten wird. Sei $f_i(x_i)$ der Preis bei Abnahme der Menge x_i von Zulieferer i. Die f_i seien z. B. stückweise lineare Funktionen. Sei M_i die maximale Liefermenge von Anbieter i. Dann lautet die Aufgabe: min $\sum_i f_i(x_i)$ unter $1^T x = a$, $0 \leqslant x \leqslant M$.
(Minimierung einer separablen Zielfunktion mit linearen Restriktionen.)

Standortplanung

Ein Konzern möchte n neue Lager der Kapazität a_i $(1 \leqslant i \leqslant n)$ errichten. Gegeben sind die Standorte der Abnehmer (α_j, β_j), sowie deren Bedarf b_j. In einigen Gebieten darf nicht gebaut werden (Umgebungen von (γ_k, δ_k)), etwa aufgrund von Umweltschutzbestimmungen bei Lagerung von Gefahrgut.
Gesucht sind die Standorte der Lager (x_i, y_i), die den Transportaufwand minimieren, sowie ein Lieferplan z (Transportmengen z_{ij} von Lager i zu Abnehmer j).

Benennung der Variablen:
Entfernungen Lager i – Abnehmer j: d_{ij}
Entfernungen Lager i – Verbotszone k: Δ_{ik}

Die Aufgabe lautet dann: min $\sum_{i,j} d_{ij} z_{ij}$ unter
$\sum_j z_{ij} \leqslant a_i$ für alle Lager i,
$\sum_i z_{ij} = b_j$ für alle Abnehmer j,
$\Delta_{ik} \geqslant \varepsilon_k$ für alle Bauverbote k,
$z \geqslant 0$.
Diese Aufgabe kann für verschiedene Entfernungsmaße gestellt werden, etwa für die Rechteckdistanz (Manhattandistanz) $d_{ij} = |x_i - \alpha_j| + |y_i - \beta_j|$ oder für die euklidische Distanz $\Delta_{ik} = \sqrt{(x_i - \gamma_k)^2 + (y_i - \delta_k)^2}$.
Wird in beiden Fällen die euklidische Distanz gewählt, so erhält man eine Optimierungsaufgabe mit quadratischer Zielfunktion und quadratischen Beschränkungen.

Chemische Gleichgewichte

Eine Anzahl m von chemischen Elementen wird gemischt. Dabei können n verschiedene Verbindungen entstehen. Es seien b_i $(1 \leqslant i \leqslant m)$ die Atomgewichte der Elemente. Es sei a_{ij} $(1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n)$ die Anzahl der Elemente vom Typ i in der Verbindung j.
Laut Massenerhaltungsgesetz gilt $\sum_j a_{ij} x_j = b_i$ für alle Elemente, wobei x_j die (unbekannte) Anzahl der Mole der Verbindung j ist.
Im Gleichgewichtszustand wird die freie Energie E (Gibbssche Enthalpie)

$$E = \sum_j c_j x_j + \sum_j x_j \ln\left(\frac{x_j}{\sum_k x_k}\right)$$

des Gemisches minimiert. Die $c_j = c_j(T, P)$ sind tabellierte Konstanten.

Aufgabe: $\min E(x)$ unter $Ax = b, x \geqslant 0$.
(Minimierung einer nichtlinearen Zielfunktion mit linearen Restriktionen.)

Diskrete Steuerung

Zustand y_0 wird durch sukzessive Steuerung mit Steuervektoren $u_1, \dots, u_n$ in die Zustände $y_1, \dots, y_n$ überführt.
Übergangsbeschreibung: $y_k := y_{k-1} + \Phi(y_{k-1}, u_k)$
Beschränkungen (B): $y_k \in Y_k, u_k \in U_k$ für $1 \leqslant k \leqslant n$ $\Psi(y_0, \dots, y_n, u_1, \dots, u_n) \in D$
Aufgabe: $\min \alpha(y, u)$ unter (B)
Beispiel: Flugzeitminimierung einer unbemannten Raumkapsel: Der Weg von A nach B ist in 10 Intervalle gleicher Länge L aufgeteilt. Sei y_i die Geschwindigkeit am Ende des i-ten Intervalls ($y_0 = y_{10} = 0$). Sei u_i die konstante Beschleunigung (Schub) im Intervall i. $-c \leqslant u_i \leqslant c$ gibt den maximalen Schub an. Sei t_i die Flugzeit füt Intervall i und m die Masse der Raumkapsel.
Übergangsbeschreibung: $y_i = y_{i-1} + \frac{u_i}{m} t_i$ $L = y_{i-1} t_i + \frac{u_i}{2m} t_i^2$. Diese quadratische Gleichung (für t_i) hat für $u_i = 0$ die Lösung $t_i = \frac{L}{y_{i-1}}$ und für $u_i \neq 0$ die positive Lösung

$$t_i = \frac{m}{u_i} \left(-y_{i-1} + \sqrt{y_{i-1}^2 + 2\frac{L}{m} u_i} \right)$$

In beiden Fällen erhält man

$$y_i = \sqrt{y_{i-1}^2 + 2\frac{L}{m} u_i}$$

Aufgabe: $\min \sum_i \frac{1}{u_i} \left(-y_{i-1} + \sqrt{y_{i-1}^2 + 2\frac{L}{m} u_i} \right)$ unter

- $y_i - \sqrt{y_{i-1}^2 + 2\frac{L}{m} u_i} = 0$ für alle Intervalle
- $y_0 = y_{10} = 0$
- $-c1 \leqslant u \leqslant c1$

Entwurf eines Kurbelgetriebes [7]

Bestimme a, b, c, d so, daß $\Psi(\varphi) \approx \Psi_0(\varphi)$ bei Kurbelbetrieb von a.
$\Psi(\varphi) = \alpha(\varphi) + \beta(\varphi)$
$\alpha(\varphi) = \arctan(\frac{a \sin \varphi}{d - a \cos \varphi})$
$\beta(\varphi) = \arccos(\frac{c^2 + r^2 - b^2}{2cr})$
$r = \sqrt{a^2 + d^2 - 2ad \cos \varphi}$
Beschränkungen: (R)

- $l_{\min} > 0$ (technische Vorgabe)
- $a, b \geqslant l_{\min}$

- $d \geqslant a + l_{\min}$
- $b + c \geqslant a + d$
- $(b - c)^2 \leqslant (d - a)^2$
- $b^2 \leqslant (d - a)^2 + c^2$
- $c^2 \leqslant (d - a)^2 + b^2$
- $c \geqslant a + l_{\min}$ (Schwingbewegung; keine Drehung)

Diskretisierung: $\varphi_i = ih$ für $0 \leqslant i \leqslant N$: $h = \frac{2\pi}{N+1}$
min $\sum_i (\Psi(\varphi_i) - \Psi_0(\varphi_i))^2$ unter (R). (quadratische Beschränkungen, Zielfunktion nichtlinear)

Graphische Lösung von Optimierungsproblemen im $\mathbf{R}^2$

Es sei die Aufgabe

$$\begin{array}{llrcll} \min & (x_1-2)^2+(x_2-1)^2 & & & & \\ \text{unter} & & x_1^2+x_2^2-2 & \leqslant & 0 & g_1 \\ & & x_1-1 & \leqslant & 0 & g_2 \\ & & -x_1 & \leqslant & 0 & g_3 \end{array}$$

gegeben.
Lösung:

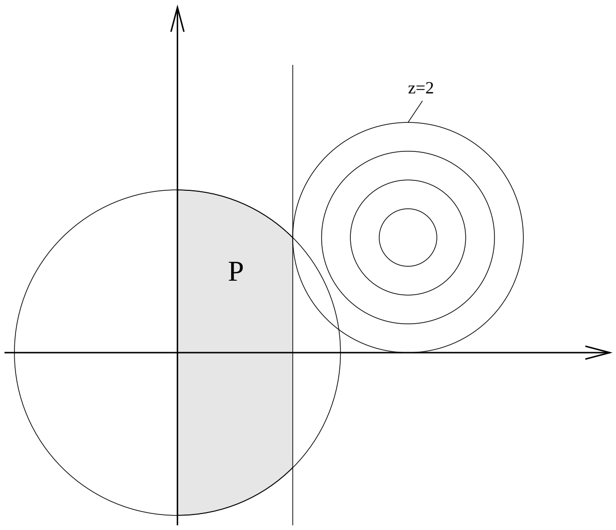

Abbildung 1: Graphische Lösung eines nichtlinearen Optimierungsproblems.

Die Lösung wird durch den Schnittpunkt des kleinsten Kreises um $(2, 1)$ mit P gegeben, es ist der Punkt $(1, 1)$. Der Zielfunktionswert beträgt 2.

Vorsicht: Nichtlineare Probleme sind pathologisch schwer, wie z. B. die folgende Formulierung zeigt: Es sei

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

Eine Lösung der Aufgabe

$$\min\{f(x)|x_1, x_2, x_3 \geqslant 1, x_4 \geqslant 3\}$$

löst das Fermatsche Problem „Existieren ganze $x, y, z \geqslant 1$ und $n \geqslant 3$ mit $x^n + y^n = z^n$?"

1.2 Allgemeine Aufgabenstellung

Wir werden die allgemeine Aufgabenstellung (1.1) nun etwas einschränken, indem wir die Darstellung von P präzisieren. Gegeben seien $D \subseteq \mathbf{R}^n$, eine Zielfunktion $f\colon D \to \mathbf{R}$, sowie Restriktionsfunktionen $g\colon D \to \mathbf{R}^m$, $h\colon D \to \mathbf{R}^p$.
Die allgemeine nichtlineare Optimierungsaufgabe lautet dann

$$\min\left\{f(x) \,\middle|\, x \in P\right\} \text{ mit } P := \left\{x \in D \,\middle|\, g(x) \leqslant 0,\ h(x) = 0\right\}.$$

Die Menge P heißt **Menge der zulässigen Punkte**.
Wir fordern für die in dieser Vorlesung behandelten Probleme die folgendenden Mindestvoraussetzungen:

1. $P \neq \emptyset$
2. D offen
3. f, g, h sind stetig differenzierbar auf D

Wir betrachten keine der praktisch interessanten Beispiele für nichtdifferenzierbare Probleme mit lipschitz-stetigen Funktionen. Außerdem betrachten wir nur Probleme mit endlich vielen Restriktionen.

1.3 Bezeichnungen

Es seien $\mathbf{R}$ die reellen Zahlen, $R_{\geqslant}$ die nichtnegativen reellen Zahlen und $\mathbf{R}_{>}$ die echt positiven reellen Zahlen.
Für $D \subseteq \mathbf{R}^n$ bezeichnen wir mit $C_i(D)$ die Menge der i-fach stetig differenzierbaren Funktionen von D nach $\mathbf{R}$. Ist $f\colon D \to \mathbf{R}$ stetig differenzierbar, so bezeichnen wir mit ∇f den Zeilenvektor $\left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)$ (Gradient von f). Ist f zweimal stetig differenzierbar, so sei

$$F := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

die Matrix der zweiten Ableitungen.

Ist $f\colon D \to \mathbf{R}$ und $x_0 \in D$, so bezeichnen wir mit $N(x_0)$ die **Niveaumenge** von x_0, wobei $N(x_0) := \{x \in D \mid f(x) \leqslant f(x_0)\}$.

Der Vektor $\mathbf{1}$ sei der Vektor aus lauter Einsen. Seine Dimension wird jeweils durch den Zusammenhang klar.

Die Funktion sgn heißt **Vorzeichenfunktion** und ist durch

$$\operatorname{sgn}(x) := \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

auf den reellen Zahlen definiert. Wir definieren die Anwendung der sgn-Funktion auf Vektoren und Matrizen komponentenweise, d. h. für $x \in \mathbf{R}^n$ sei $(\operatorname{sgn}(x))_i = \operatorname{sgn}(x_i)$. Die Erweiterung der Betragsfunktion auf Vektoren und Matrizen sei analog.

Wir benutzen zur Abschätzung von Resttermen die Landau-Symbole $O()$ und $o()$. Es bedeutet $f(x) = O(g(x))$, daß es eine Konstante $c \in \mathbf{R}_>$ gibt, so daß $f(x) \leqslant cg(x)$ für alle x gilt. Der Wert der Konstanten c ist hierbei nicht von Interesse. Wir schreiben $f(x) = o(g(x))$, wenn gilt

$$g(x) > 0 \text{ für } x \in \mathbf{R}_> \text{ und } \lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$$

gilt. Je nach Kontext kann der Grenzwert auch gegen null laufen, etwa bei Darstellung der Taylor-Entwicklung für zweifach stetig differenzierbare Funktionen:

$$f(x + td) = f(x) + t\nabla f(x)d + to(t).$$

Wir benutzen die Schreibweise $f(x) = \Theta(g(x))$, wenn sowohl $f(x) = O(g(x))$ wie auch $g(x) = O(f(x))$ gilt, d. h. wenn die Terme von gleicher Größenordnung sind.

Kapitel 2

Nichtlineare Optimierung ohne Restriktionen

In der Einführung wurden bereits Verfahren zur Minimierung nichtlinearer Funktionen ohne Nebenbedingungen im eindimensionalen Fall und einfache Verfahren für den mehrdimensionalen Fall vorgestellt. In diesem Kapitel werden wir generell Methoden für den mehrdimensionalen Fall ohne Restriktionen behandeln. Die aus der Einführung bekannten Verfahren werden genauer untersucht und wesentlich verfeinert, um auch Funktionen minimieren zu können, bei denen die einfachen Verfahren nur schlecht oder überhaupt nicht konvergieren.

2.1 Abstiegsverfahren

Für diesen Abschnitt sollen die folgenden Voraussetzungen stets erfüllt sein: Sei $f \in C_2(\mathbf{R}^n)$ nach unten beschränkt. Es sei ein $x_0 \in \mathbf{R}^n$ gegeben, für das $N(x_0)$ beschränkt ist. f besitze eine lokale Minimalstelle $x_* \in N(x_0)$.

Weil $\nabla f(x_*) = 0$, besitzen wir eine notwendiges Kriterium für Minimalstellen. Um f zu minimieren, suchen wir zunächst $x_* \in N(x_0)$ mit $\nabla f(x_*) = 0$. Dann testen wir, ob x_0 eine Minimalstelle ist.

2.1.1 Das Prinzip des hinreichenden Abstiegs

Definition 2.1.1. Ein Vektor $d \in \mathbf{R}^n$ heißt **Abstiegsrichtung** in x, falls $\nabla f(x)d < 0$ gilt. Gibt es positive von x und d unabhängige Konstanten $\alpha, \beta > 0$, so daß für ein $\lambda \in \mathbf{R}$ die Bedingungen

1. $f(x + \lambda d) - f(x) \leqslant \alpha\lambda\nabla f(x)d$ und
2. $\beta|\nabla f(x)d|\frac{1}{\|d\|^2} \leqslant \lambda$

erfüllt sind, so heißt λ **akzeptable Schrittweite** für (x, d).

Bemerkung 2.1.1. Die erste Forderung für eine akzeptable Schrittweite erzwingt einen kontrollierten minimalen Abstieg in f, die zweite verhindert zu kleine Schrittweiten.

Bemerkung 2.1.2. Es sei $N(x_0)$ beschränkt, also kompakt. Dann ist

$$M := \max\left\{\|F(x)\| \;\middle|\; x \in \operatorname{conv}(N(x_0))\right\}$$

wohldefiniert.

Lemma 2.1.1. Sei $\delta \in (0,1)$, $x \in N(x_0)$ und d eine Abstiegsrichtung in x. Dann gibt es ein $\tau = \tau(x,d,\delta)$ mit den Eigenschaften

1. $f(x+\lambda d) - f(x) < \delta\lambda\nabla f(x)d$ für alle $\lambda \in (0,\tau)$
2. $f(x+\tau d) - f(x) = \delta\tau\nabla f(x)d$
3. $\tau \geqslant 2\frac{(1-\delta)}{M}\frac{|\nabla f(x)d|}{\|d\|^2} =: \varrho$
4. $\nabla f(x+\lambda d)d < \delta\nabla f(x)d$ für alle $\lambda \in [0,\frac{\varrho}{2})$

Beweis. Die Taylorentwicklung von f bei x ergibt

$$f(x+\lambda d) - f(x) = \lambda\underbrace{\nabla f(x)d}_{<0} + \frac{1}{2}\lambda^2 d^T F(x+\theta\lambda d)d \text{ für ein } \theta \in (0,1). \tag{2.1}$$

Wir setzen $T := \left\{\tau' > 0 \;\middle|\; f(x+\lambda d) - f(x) < \delta\lambda\nabla f(x)d \text{ für alle } \lambda \in (0,\tau')\right\}$. Wegen (2.1) ist $T \neq \emptyset$ und $[x, x+\tau' d] \subset N(x_0)$ für alle $\tau' \in T$. Da $N(x_0)$ kompakt ist, ist T beschränkt. Wir zeigen nun, daß $\tau := \sup T$ alle vier Behauptungen erfüllt.

1. trivial.
2. Dies gilt, denn sonst wäre ein größeres τ möglich.
3. Mit $\lambda = \tau$ in (2.1) folgt

$$|\underbrace{f(x+\tau d) - f(x) - \tau\nabla f(x)d}_{=(\delta-1)\tau\nabla f(x)d}| \leqslant \frac{\tau^2}{2}M\|d\|^2,$$

   also gilt $\frac{2(1-\delta)}{M}\frac{|\nabla f(x)d|}{\|d\|^2} \leqslant \tau$, womit auch dieser Punkt erfüllt ist.

4. Sei $\lambda \in [0,\frac{\varrho}{2})$. Dann gilt

$$\begin{aligned}
\nabla f(x+\lambda d)d &= \nabla f(x)d + (\nabla f(x+\lambda d) - \nabla f(x))d \\
&\leqslant \nabla f(x)d + \lambda M d^T d \\
&< \nabla f(x)d + \frac{\varrho}{2}M\|d\|^2 \\
&= \nabla f(x)d + (\delta-1)\nabla f(x)d \\
&= \delta\nabla f(x)d.
\end{aligned}$$

$\square$

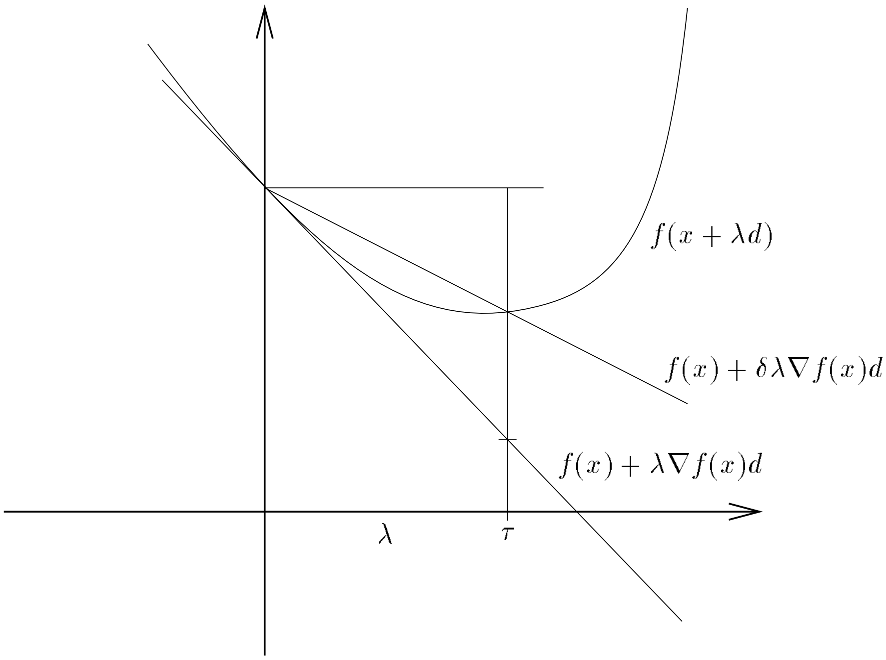

Abbildung 2: Skizze zu Lemma 2.1.1 für $\delta = \frac{1}{2}$.

Bemerkung 2.1.3. Lemma 2.1.1 sichert die Existenz von akzeptablen Schrittweiten. Wir wählen etwa $\alpha := \delta$ und $\beta := \frac{2(1-\delta)}{M}$. Dann ist τ wegen Lemma 2.1.1.1 und Lemma 2.1.1.3 eine akzeptable Schrittweite.

Da die Bestimmung von τ nach Lemma 2.1.1 aufwendig ist, werden wir nun ein schnelleres Verfahren zur Bestimmung einer akzeptablen Schrittweite kennenlernen.

2.1.2 Die Goldstein-Armijo-Suche

Die Goldstein-Armijo-Suche ist eine Vorschrift, die zu gegebener Abstiegsrichtung eine Schrittweite bestimmt. Es sei $\alpha \in (0,1)$, $x \in N(x_0)$ und d eine Abstiegsrichtung in x. Seien β', β'' zwei Konstanten, die $0 < \beta' \ll \beta''$ erfüllen. Mit diesen Eingabedaten lautet die Vorschrift der Goldstein-Armijo-Suche wie folgt:

Algorithmus 2.1.1. (Goldstein-Armijo-Suche)
Es seien $\alpha, \beta', \beta'' \in \mathbf{R}$ und $x, d \in \mathbf{R}^n$ gegeben.

1. Wähle einen Startwert $\lambda_0 \in [\beta', \beta''] \frac{|\nabla f(x) d|}{\|d\|^2}$.
2. Setze $i := \min \left\{ j \in \mathbf{N}_0 \;\middle|\; f(x+\lambda(j)d) - f(x) \leqslant \alpha\lambda(j)\nabla f(x) d \right\}$ mit $\lambda(j) := (\frac{1}{2})^j \lambda_0$.
3. Setze $\lambda := \lambda(i)$.

Satz 2.1.2. Die Goldstein-Armijo-Suche liefert eine akzeptable Schrittweite λ.

Beweis. Nach Lemma 2.1.1 ist λ wohldefiniert und erfüllt Bedingung 1 aus Definition 2.1.1. Für $i = 0$ ist $\lambda = \lambda_0$ und die zweite Bedingung mit $\beta \leqslant \beta'$ erfüllt.
Sei nun $i > 0$, d. h.

$$f(x + 2\lambda d) - f(x) > \alpha 2\lambda \nabla f(x)d. \tag{2.2}$$

Setze $\bar{\lambda} := \min\{\mu > 0 \mid \nabla f(x + \mu d)d = 0\}$, welches existiert, da $N(x)$ kompakt ist. Wegen $\nabla f(x + \mu d)d < 0$ für alle $\mu \in [0, \bar{\lambda})$ gilt $[x, x + \bar{\lambda}d] \subseteq N(x)$. Für $\lambda = \bar{\lambda}$ ist Lemma 2.1.1.4. (mit $\delta = \alpha$) nicht erfüllt, d. h. $\bar{\lambda} > \frac{\varrho}{2}$, also

$$\bar{\lambda} > \frac{(1-\alpha)}{M} \frac{|\nabla f(x)d|}{\|d\|^2}.$$

Fall 1: $x + 2\lambda d \notin N(x)$. Also gilt $2\lambda > \bar{\lambda} > \frac{1-\alpha}{M}\frac{|\nabla f(x)d|}{\|d\|^2}$, d. h. Definition 2.1.1.2. ist erfüllt, wenn wir $\beta \leqslant \frac{1-\alpha}{2M}$ wählen.

Fall 2: $x + 2\lambda d \in N(x)$. Nach dem Satz von Taylor gibt es $\theta \in (0, 1)$ mit

$$f(x + 2\lambda d) - f(x) = \nabla f(x + \theta 2\lambda d) 2d\lambda.$$

Nun folgt mit (2.2) $\nabla f(x + \theta 2\lambda d)d > \alpha \nabla f(x)d$.
Also gilt $(\nabla f(x + \theta 2\lambda d) - \nabla f(x))d > (1 - \alpha)|\nabla f(x)d|$. Die Taylorentwicklung von ∇f liefert nun

$$\nabla f(x + \theta 2\lambda d)d - \nabla f(x)d = 2\lambda\theta \underbrace{d^T F(x + 2\lambda\theta\theta' d)d}_{\leqslant M\|d\|^2}$$

für ein $\theta' \in (0, 1)$, woraus nach Auflösung nach λ

$$\lambda > \theta\lambda > \frac{1-\alpha}{2M}\frac{|\nabla f(x)d|}{\|d\|^2}$$

entsteht, d. h. Definition 2.1.1.2. ist für $\beta \leqslant \frac{1-\alpha}{2M}$ erfüllt.

Insgesamt haben wir nachgewiesen, daß die Goldstein-Armijo-Suche eine akzeptable Schrittweite liefert, wenn $\beta = \min\left\{\beta', \frac{1-\alpha}{2M}\right\}$ gesetzt wird. □

Bemerkung 2.1.4. Häufig setzt man $\alpha = \frac{1}{100}$. Für die Konstanten β' und β'' benutzt man Werte von etwa $\beta' = 10^{-4}$, $\beta'' = 10^4$. (Man findet etliche Überlegungen zur Wahl der Konstanten in [7], ebenso andere Ansätze zur Bestimmung akzeptabler Schrittweiten.)

Bemerkung 2.1.5. Bei exakter Minimierung von f in Richtung d gilt

$$\frac{\nabla f(x + \bar{\lambda}d)d}{\nabla f(x)d} = 0.$$

Die Güte einer mit der Goldstein-Armijo-Suche gewonnenen Schrittweite λ wird oft mit diesem Quotienten (mit $\lambda \equiv \bar{\lambda}$) bewertet.

2.1.3 Asymptotisch exakte Schrittweiten

Definition 2.1.2. Zu jedem $x \in U(x_*)$ sei eine Abstiegsrichtung d und eine Schrittweite λ gegeben. Dann heißt die Schrittweite λ **asymptotisch exakt von der Ordnung** k ($k \in \mathbf{N}$), falls

$$\frac{|\nabla f(x+\lambda d)d|}{|\nabla f(x)d|} = O\Big(\|\nabla f(x)\|^k\Big)$$

gilt. λ heißt **asymptotisch exakt**, falls für $x \to x_*$ mit $\nabla f(x_*) = 0$

$$\frac{|\nabla f(x+\lambda d)d|}{|\nabla f(x)d|} = o(1)$$

gilt.

Bemerkung 2.1.6. (Konvergenz von Abstiegsverfahren bei Verwendung asymptotisch exakter Schrittweiten)
Sei $x_{k+1} := x_k + \lambda_k d_k$ mit Abstiegsrichtungen d_k bei x_k für $k \in \mathbf{N}_0$ und akzeptablen Schrittweiten λ_k. Dann gilt

$$f(x_{k+1}) - f(x_k) \leqslant \alpha\lambda_k \nabla f(x_k)d_k \leqslant -\alpha\beta\frac{|\nabla f(x_k)d_k|^2}{\|d_k\|^2}. \tag{2.3}$$

Die Folge $\frac{|\nabla f(x_k)d_k|}{\|d_k\|}$ geht also gegen 0. Gesucht ist eine Folge mit $\nabla f(x_k) \to 0$.

2.1.4 Streng gradientenbezogene Richtungen

Definition 2.1.3. Zu jedem $x \in U(x_*)$ sei eine Abstiegsrichtung d gegeben. Dann heißt d heißt **streng gradientenbezogen**, falls es von x unabhängige Konstanten $\varrho_1, \varrho_2 > 0$ gibt mit

$$\begin{aligned} \varrho_1\|\nabla f(x)\| &\geqslant \|d\| \geqslant \frac{1}{\varrho_1}\|\nabla f(x)\| && (2.4)\\ \nabla f(x)d &\leqslant -\varrho_2\|\nabla f(x)\|\cdot\|d\|. && (2.5)\end{aligned}$$

Bemerkung 2.1.7. Für streng gradientenbezogene Abstiegsrichtungen kann man die Abschätzung (2.3) verbessern. Hier gilt

$$f(x_{k+1}) - f(x_k) \leqslant -\alpha\beta\varrho_2^2\|\nabla f(x_k)\|^2,$$

also folgt $\nabla f(x_k) \to 0$.

Satz 2.1.3. Es sei die Menge $\Omega := \{x \in N(x_0) \mid \nabla f(x) = 0\}$ von endlicher Mächtigkeit. Die in (2.3) betrachtete Folge sei durch streng gradientenbezogene Suchrichtungen erzeugt. Sind die zugehörigen Schrittweiten λ_k akzeptabel und beschränkt, so konvergiert die Folge (x_k) gegen ein $x_* \in \Omega$.

Beweis. Wegen Bemerkung 2.1.7 gilt $\nabla f(x_k) \to 0$. Da $N(x_0)$ kompakt ist, besitzt die Folge (x_k) Häufungspunkte. Ist $|\Omega| = 1$, so ist der Häufungspunkt der Grenzwert x_*. Wegen $\nabla f(x_k) \to 0$ folgt $\nabla f(x_*) = 0$, also $\Omega = \{x_*\}$. Damit sind wir fertig.

Ist $|\Omega| > 1$, so existiert $\sigma := \min\{\|x-y\| \mid x, y \in \Omega, x \neq y\} > 0$, da Ω endlich ist. Für genügend großes k_0 gilt

$$\operatorname{dist}(x_k, \Omega) \leqslant \frac{\sigma}{4} \text{ für alle } k \geqslant k_0. \tag{2.6}$$

Es sei $c := \sup\{\lambda_k \mid k \in \mathbf{N}_0\}$. Wegen $\nabla f(x_k) \to 0$ und weil d_k streng gradientenbezogen ist, gibt es ein $k_1 \in \mathbf{N}$ mit $\|x_k - x_{k+1}\| = \lambda_k \|d_k\| \leqslant c\varrho_1 \|\nabla f(x_k)\| \leqslant \frac{\sigma}{4}$ für alle $k \geqslant k_1$.

Sei $k \geqslant \max\{k_0, k_1\}$ und $\|x_k - x_*\| \leqslant \frac{\sigma}{4}$ für ein $x_* \in \Omega$. Dann folgt aus (2.6) $\|x_{k+1} - x_*\| \leqslant \frac{\sigma}{4}$ und somit auch $\|x_k - x_*\| \leqslant \frac{\sigma}{4}$ für alle $k \geqslant \max\{k_0, k_1\}$. Also besitzt (x_k) nur einen einzigen Häufungspunkt x_*, d. h. $(x_k) \to x_*$. □

Definition 2.1.4. Die Matrixfolge (A_k) symmetrischer Matrizen heißt **beschränkt**, falls es eine Konstante $R \in \mathbf{R}$ gibt, so daß

$$|y^T A_k y| \leqslant R y^T y \text{ für alle } k \text{ und alle } y \in \mathbf{R}^n$$

gilt. Sie heißt **gleichmäßig positiv definit**, falls es eine Konstante $r > 0$ gibt, so daß

$$r y^T y \leqslant y^T A_k y \text{ für alle } k \in \mathbf{N} \text{ und alle } y \in \mathbf{R}^n$$

gilt.

Lemma 2.1.4. Sei (A_k) eine beschränkte, gleichmäßig positiv definite Matrixfolge. Dann ist $d_k := -A_k^{-1} \nabla f(x_k)^T$ streng gradientenbezogen.

Beweis. Es sei $\|\cdot\|$ die euklidische Vektornorm. Es seien $r, R > 0$ nach Definition 2.1.4 gegeben. Dann gilt

$$\lambda_{\min}(A_k^{-1}) = (\lambda_{\max}(A_k))^{-1} \geqslant \frac{1}{R} \quad \text{und} \quad \lambda_{\max}(A_k^{-1}) = (\lambda_{\min}(A_k))^{-1} \leqslant \frac{1}{r} \text{ für alle } k \in \mathbf{N}.$$

Es folgt

$$\begin{aligned} \frac{1}{R}\|\nabla f(x)\| &\leqslant \lambda_{\min}(A_k^{-1}) \|\nabla f(x)\| \\ &\leqslant \|A_k^{-1} \nabla f(x)\| \\ &= \|d_k\| \\ &\leqslant \lambda_{\max}(A^{-1}) \|\nabla f(x)\| \\ &\leqslant \frac{1}{r}\|\nabla f(x)\|. \end{aligned}$$

Somit ist durch Wahl von $\varrho_1 := \max\left\{\frac{1}{r}, R\right\}$ Forderung (2.4) erfüllt.

Sei $\{v_1, \ldots, v_n\}$ eine Orthonormalbasis von Eigenvektoren von A_k^{-1} und seien $\{\lambda_1, \ldots, \lambda_n\}$ die Eigenwerte von A_k^{-1}. Wir stellen nun den Gradienten von f als Linearkombination der Eigenvektoren v_i dar. Man erhält mit dem Ansatz $\nabla f(x_k)^T = \sum_i c_i v_i$

$$\begin{aligned} -\nabla f(x_k) d_k &= \nabla f(x_k) A_k^{-1} \nabla f(x_k)^T \\ &= \Big(\sum_i c_i v_i\Big)^T A_k^{-1} \Big(\sum_j c_j v_j\Big) \\ &= \Big(\sum_i c_i v_i\Big)^T \Big(\sum_j c_j \lambda_j v_j\Big) \\ &= \sum_i c_i^2 \lambda_i. \end{aligned}$$

Andererseits gilt

$$\|\nabla f(x_k)\| \cdot \|A_k^{-1}\nabla f(x_k)\| = \sqrt{\sum_i c_i^2}\sqrt{\sum_i c_i^2\lambda_i^2}.$$

Wählen wir nun $\varrho_2 := \sqrt{\frac{r}{R}} \leqslant \sqrt{\frac{\lambda_{\min}(A_k^{-1})}{\lambda_{\max}(A_k^{-1})}}$, so gilt

$$\Big(- \nabla f(x_k)d_k\Big)^2 = \sum_{i,j} c_i^2 c_j^2 \lambda_i\lambda_j \geqslant \varrho_2^2 \sum_{i,j} c_i^2 c_j^2 \lambda_i^2 = \Big(\varrho_2 \|\nabla f(x_k)\| \cdot \|A_k^{-1}\nabla f(x_k)\|\Big)^2,$$

also ist auch Forderung (2.5) erfüllt. □

Lemma 2.1.5. Sei $x_{k+1} := x_k + \lambda_k d_k$ und $(x_k) \to x_*$, wobei die d_k streng gradientenbezogen und die λ_k beschränkt und asymptotisch exakt von der Ordnung 1 sind. Es sei $f \in C_3(\mathbf{R}^n)$, $\nabla f(x_*) = 0$ und $F(x_*)$ positiv definit.
Dann gilt

$$\lambda_k = \frac{-\nabla f(x_k)d_k}{d_k^T F(x_*)d_k}\Big(1 + O(\|\nabla f(x_k)\|)\Big).$$

Beweis. Da die λ_k asymptotisch exakt von der Ordnung 1 sind, und da d_k streng gradientenbezogen ist, gilt

$$\nabla f(x_{k+1})d_k = \nabla f(x_k)d_k\, O(\|\nabla f(x_k)\|) = O(\|d_k\|^3). \tag{2.7}$$

Andererseits ergibt die Taylorentwicklung von ∇f bei x_k wegen $f \in C_3(\mathbf{R}^n)$

$$\nabla f(x_{k+1})d_k = \nabla f(x_k)d_k + \lambda_k d_k^T F(x_k)d_k + O(\lambda_k^2\|d_k\|^3). \tag{2.8}$$

Vergleich von (2.7) und (2.8) ergibt

$$\nabla f(x_k)d_k + \lambda_k d_k^T F(x_k)d_k + O(\lambda_k^2\|d_k\|^3) = O(\lambda_k^2\|d_k\|^3),$$

wobei der $O()$-Term auf der linken Seite fortgelassen werden kann, da λ_k beschränkt ist.
Da $f \in C_2(\mathbf{R}^n)$, gibt es für alle $x \in U(x_*)$ ein $\theta \in (0,1)$ mit $\nabla f(x) = \nabla f(x) - \nabla f(x_*) = (x - x_*)^T F(x_* + \theta(x - x_*))$.
Wähle $U(x_*)$ klein genug, so daß $0 < \frac{1}{2}\lambda_{\min}(F(x_*)) \leqslant \lambda_{\min}(F(y))$ für alle $y \in U(x_*)$ gilt und $\lambda_{\max}(F(y)) \leqslant 2\lambda_{\max}(F(x_*))$. Dann gilt $\|\nabla f(x)\| = O(\|x - x_*\|)$ und $\|x - x_*\| = O(\|\nabla f(x)\|)$.
Da d_k streng gradientenbezogen ist, folgt $\|d_k\| = O(\|x - x_*\|)$ und $\|x - x_*\| = O(\|d_k\|)$. Also folgt durch Taylorentwicklung: $F(x_k) = F(x_*) + O(\|x - x_*\|) = F(x_*) + O(\|d_k\|)$. Wegen (2.8) gilt weiterhin

$$\nabla f(x_k)d_k + \lambda_k \underbrace{d_k^T F(x_*)d_k}_{\Theta(\|d_k\|^2)} = O(\|d_k\|^3)$$

und somit

$$\lambda_k = \frac{-\nabla f(x_k)d_k}{d_k^T F(x_*)d_k}\Big(1 + O(\|d_k\|)\Big).$$

□

Man könnte also unter den Voraussetzungen von Lemma 2.1.5 die asymptotische Form der Schrittweite angegeben, wenn $F(x_*)$ bekannt wäre.

Den nun folgenden Satz von Kantorovich werden wir für eine Abschätzung im Beweis von Satz 2.1.7 benötigen.

Satz 2.1.6. (Satz von Kantorovich) Sei $S \in \mathbf{R}^{n\times n}$ positiv definit. Sei ϱ_1 kleinster Eigenwert von S und ϱ_n größter Eigenwert von S. Dann gilt

$$\frac{(x^Tx)^2}{(x^TSx)(x^TS^{-1}x)} \geqslant \frac{4\varrho_1\varrho_n}{(\varrho_1+\varrho_n)^2} \text{ für alle } x \in \mathbf{R}^n,\ x \neq 0.$$

Beweis. S ist mit Hilfe einer orthogonalen Matrix diagonalisierbar. Sei $D = USU^T$, also $S = U^TDU$, $S^{-1} = U^TD^{-1}U$. Setze $y := \frac{1}{\|x\|_2}Ux$. Es gilt $\|y\| = 1$, weil U orthogonal ist. Nun gilt mit $D = \operatorname{diag}(\varrho_1,\dots,\varrho_n)$ (wobei $\varrho_1 \leqslant \varrho_2 \leqslant \cdots \leqslant \varrho_n$)

$$\begin{aligned}
\min_{x\neq 0}\frac{(x^Tx)^2}{(x^TSx)(x^TS^{-1}x)} &= \min_{x\neq 0}\frac{\|x\|_2^4}{(x^TU^TDUx)(x^TU^TD^{-1}Ux)}\\
&= \min_{\|y\|=1}\frac{1}{(y^TDy)(y^TD^{-1}y)}\\
&= \min_{\|y\|=1}\left[\left(\sum_{i=1}^n \varrho_i y_i^2\right)\cdot\left(\sum_{i=1}^n \frac{1}{\varrho_i}y_i^2\right)\right]^{-1}\\
&= \min_{\xi}\left[\left(\sum_{i=1}^n \varrho_i \xi_i\right)\cdot\left(\sum_{i=1}^n \frac{1}{\varrho_i}\xi_i\right)\right]^{-1} =: \min_{\xi} z(\xi)
\end{aligned}$$

mit der Substitution $\xi_i := y_i^2$ für $i = 1,\dots,n$. Wegen $1 = \|y\|_2 = \sum_{i=1}^n y_i^2$ folgt $\sum_{i=1}^n \xi_i = 1$ und $\xi_i \geqslant 0$.

Wir minimieren nun $z(\xi)$. Es sei $\varrho := \sum_{i=1}^n \xi_i\varrho_i$. Zu festem ϱ ist die Lösung von

$$\min_{\xi} z(\xi,\varrho) = \min\left\{\frac{1/\varrho}{\sum_{i=1}^n \frac{1}{\varrho_i}\xi_i}\ \middle|\ \varrho = \sum_{i=1}^n \varrho_i\xi_i,\ \xi_i \geqslant 0, \sum_{i=1}^n \xi_i = 1\right\}$$

äquivalent zur Lösung des linearen Optimierungsproblems

$$\max\left\{\sum_{i=1}^n \frac{1}{\varrho_i}\xi_i\ \middle|\ \sum_{i=1}^n \xi_i = 1,\ \sum_{i=1}^n \varrho_i\xi_i = \varrho,\ \xi \geqslant 0\right\}.$$

Dieses Optimierungsproblem ist für alle möglichen Werte von ϱ (also für $\varrho_1 \leqslant \varrho \leqslant \varrho_n$) zu lösen.

Behauptung: Durch $\xi_1 = \frac{\varrho_n-\varrho}{\varrho_n-\varrho_1}$, $\xi_n = 1-\xi_1$ und $\xi_i = 0$ für $i = 2,\dots,n-1$ wird eine Optimallösung gegeben. Zunächst rechnet man nach, daß die Bedingung $\varrho = \sum_{i=1}^n \xi_i\varrho_i$ erfüllt ist und $0 \leqslant \xi \leqslant 1$ gilt. Also ist die Lösung zulässig.

Falls $\varrho_1 = \varrho_n$, so sind alle ϱ_i gleich und die Lösung ist trivialerweise optimal.

Anderenfalls berechne $\tilde{c}_N$ zur zulässigen Basislösung ξ. Wegen $\varrho_1 < \varrho_n$ ist

$$A_B = \begin{pmatrix} 1 & 1 \\ \varrho_1 & \varrho_n \end{pmatrix}$$

regulär. Man berechnet nun

$$c_B^T = \left(\frac{1}{\varrho_1}, \frac{1}{\varrho_n}\right) \text{ und } A_B^{-1} = \begin{pmatrix} \varrho_n & -1 \\ -\varrho_1 & 1 \end{pmatrix} \cdot \frac{1}{\varrho_n - \varrho_1}.$$

Damit ergibt sich

$$c_B^T A_B^{-1} = (\varrho_1 + \varrho_n, -1)\frac{1}{\varrho_1 \varrho_n}$$

und schließlich

$$\begin{aligned} \tilde{c}_{N(i)} &= \frac{1}{\varrho_i} - \frac{1}{\varrho_1 \varrho_n}(\varrho_1 + \varrho_n - \varrho_i) \\ &= -\frac{1}{\varrho_1} - \frac{1}{\varrho_n} + \frac{1}{\varrho_i} + \frac{\varrho_i}{\varrho_1 \varrho_n} \\ &= \frac{1}{\varrho_1 \varrho_i \varrho_n}(\varrho_i^2 + \varrho_1 \varrho_n - \varrho_1 \varrho_i - \varrho_i \varrho_n) \\ &= \frac{1}{\varrho_1 \varrho_i \varrho_n} \underbrace{(\varrho_i - \varrho_1)}_{\geqslant 0} \underbrace{(\varrho_i - \varrho_n)}_{\leqslant 0} \\ &\leqslant 0 \end{aligned}$$

Da die reduzierten Kostenkoeffizienten nicht positiv sind, ist die angegebene Lösung optimal. Insbesondere ist die optimale Basis unabhängig vom Parameter ϱ.
Damit ergibt sich

$$\begin{aligned} \min_{x \neq 0} \frac{(x^T x)^2}{(x^T S x)(x^T S^{-1} x)} &= \min_{\varrho} \left[\varrho \cdot \left(\frac{\varrho_n - \varrho}{\varrho_n - \varrho_1} \cdot \frac{1}{\varrho_1} + \frac{\varrho - \varrho_1}{\varrho_n - \varrho_1} \cdot \frac{1}{\varrho_n} \right) \right]^{-1} \\ &= \min_{\varrho} \left[\frac{\varrho}{(\varrho_n - \varrho_1)\varrho_1 \varrho_n} \Big((\varrho_n - \varrho)\varrho_n + (\varrho - \varrho_1)\varrho_1 \Big) \right]^{-1} \\ &= \min_{\varrho} \left[\frac{\varrho}{(\varrho_n - \varrho_1)\varrho_1 \varrho_n} \Big((\varrho_n - \varrho_1)(\varrho_n + \varrho_1 - \varrho) \Big) \right]^{-1} \\ &= \min_{\varrho} \left[\frac{\varrho}{\varrho_1 \varrho_n} (\varrho_n + \varrho_1 - \varrho) \right]^{-1} \\ &= \frac{4 \varrho_1 \varrho_n}{(\varrho_1 + \varrho_n)^2}, \end{aligned}$$

denn die Minimierung in der letzten Formulierung ist ein quadratisches Problem, dessen Lösung offensichtlich bei $\varrho = \frac{1}{2}\varrho_1 + \frac{1}{2}\varrho_n$ angenommen wird. □

Satz 2.1.7. Sei $x_0 \in \mathbf{R}^n$ und $\Omega := \{x \in N(x_0) \mid \nabla f(x) = 0\}$ endlich. Sei (A_k) eine Folge beschränkter gleichmäßig positiv definiter Matrizen. Die Schrittweite λ_k der Folge $x_{k+1} := x_k + \lambda_k d_k$ $(k \in \mathbf{N}_0)$ mit $d_k := -A_k^{-1} \nabla f(x_k)^T$ sei akzeptabel und beschränkt.
Dann gibt es ein $x_* \in \Omega$ mit $(x_k) \to x_*$.
Falls außerdem

1. $f \in C_3(U(x_*))$, $F(x_*)$ positiv definit ist und

2. λ_k asymptotisch exakt von der Ordnung 1 ist,

so gilt $f(x_{k+1}) - f(x_*) \leqslant \Big[\Big(\frac{\mu_n - \mu_1}{\mu_n + \mu_1}\Big)^2 + \varepsilon_k \Big] (f(x_k) - f(x_*))$ mit $(\varepsilon_k) \to 0$, wobei

$$\begin{aligned} \mu_1 &:= \liminf \lambda_{\min}\Big(A_k^{-\frac{1}{2}} F(x_*) A_k^{-\frac{1}{2}} \Big) > 0 \text{ und} \\ \mu_n &:= \limsup \lambda_{\max}\Big(A_k^{-\frac{1}{2}} F(x_*) A_k^{-\frac{1}{2}} \Big) < \infty. \end{aligned}$$

Beweis. $(x_k) \to x_* \in \Omega$ folgt nach Lemma 2.1.4 und Satz 2.1.3. Die Taylorentwicklung von f um x_* ergibt

$$f(x_k) - f(x_*) = \frac{1}{2}(x_k - x_*)^T F(x_*)(x_k - x_*) + O(\|x_k - x_*\|^3), \tag{2.9}$$

weil der erste Term wegen $\nabla f(x_*) = 0$ entfällt. Die Taylorentwicklung von ∇f um x_*

$$\nabla f(x_k) - \nabla f(x_*) = (x_k - x_*)^T F(x_*) + O(\|x_k - x_*\|^2)$$

läßt sich wie folgt nach $(x_k - x_*)$ auflösen:

$$(x_k - x_*)^T = \nabla f(x_k)(F(x_*))^{-1} + O(\|x_k - x_*\|^2)$$

und in (2.9) einsetzen:

$$f(x_k) - f(x_*) = \frac{1}{2}\nabla f(x_k)(F(x_*))^{-1}\nabla f(x_k)^T + O(\|x_k - x_*\|^3). \tag{2.10}$$

Wir schreiben nun (2.9) analog für x_{k+1}:

$$f(x_{k+1}) - f(x_*) = \frac{1}{2}(x_k - x_* + \lambda_k d_k)^T F(x_*)(x_k - x_* + \lambda_k d_k) + O(\|x_k - x_*\|^3).$$

Dies ergibt nach Ausmultiplizieren und Einsetzen von (2.9)

$$f(x_{k+1}) - f(x_*) = f(x_k) - f(x_*) + \lambda_k d_k^T F(x_*)(x_k - x_*) + \tfrac{1}{2}\lambda_k^2 \underbrace{d_k^T F(x_*) d_k}_{=O(\|d_k\|^2)} + O(\|x_k - x_*\|^3).$$

Wie im Beweis von Lemma 2.1.5 kann man auch hier zeigen, daß in einer genügend kleinen Umgebung von x_* die Terme $\|\nabla f(x_k)\|$, $\|d_k\|$ und $\|x_k - x_*\|$ größenordnungsmäßig gleich sind und gegeneinander abgeschätzt werden können. Außerdem können wir $(x_k - x_*)$ gemäß (2.1.4) ersetzen, die Darstellung von λ_k aus Lemma 2.1.5 benutzen und erhalten damit

$$f(x_{k+1}) - f(x_*) = f(x_k) - f(x_*) - \frac{1}{2}\frac{(\nabla f(x_k) d_k)^2}{d_k^T F(x_*) d_k} + O(\|x_k - x_*\|^3).$$

Dies ergibt nach Ausklammern von $f(x_k) - f(x_*)$ und weil $f(x_k) - f(x_*) = \Theta(\|x_k - x_*\|^2)$ ist zusammen mit (2.10)

$$f(x_{k+1}) - f(x_*) = (f(x_k) - f(x_*))\Big[1 - \frac{(\nabla f(x_k) d_k)^2}{d_k^T F(x_*) d_k \nabla f(x_k) F(x_*)^{-1} \nabla f(x_k)^T} + O(\|x_k - x_*\|)\Big].$$

Der Bruch läßt sich mit Hilfe des Satzes von Kantorovich abschätzen: Man setzt dafür $S := A_k^{-\frac{1}{2}} F(x_*) A_k^{-\frac{1}{2}}$ und $s := A_k^{\frac{1}{2}} d_k$. Wegen $\nabla f(x_k) = -d_k^T A_k$ lautet der Bruch dann $\frac{(s^T s)^2}{(s^T S s)(s^T S^{-1} s)}$ und läßt sich durch $\frac{4\mu_1\mu_n}{(\mu_1+\mu_n)^2}$ nach unten abschätzen, wobei $\mu_1 = \lambda_{\min}(S), \mu_n = \lambda_{\max}(S)$. Damit haben wir

$$\begin{aligned} f(x_{k+1}) - f(x_*) &\leqslant (f(x_k) - f(x_*)) \left[1 - \frac{4\mu_1\mu_n}{(\mu_1+\mu_n)^2} + O(\|x_k - x_*\|)\right] \\ &= (f(x_k) - f(x_*)) \left[\frac{(\mu_n - \mu_1)^2}{(\mu_n+\mu_1)^2} + O(\|x_k - x_*\|)\right], \end{aligned}$$

wobei $\mu_n < \infty$, weil (A_k) gleichmäßig positiv definit ist und $\mu_1 > 0$, weil (A_k) beschränkt ist. □

Bemerkung 2.1.8. Falls $\mu_1 = \mu_n$, so folgt in Satz 2.1.7 superlineare Konvergenz. Sei etwa $\varrho_1 := \lambda_{\min}(F(x_*))$ und $\varrho_n := \lambda_{\max}(F(x_*))$. Lokal gilt dann wegen Satz 2.1.7 und zweifacher Anwendung von (2.9) für genügend große k:

$$\begin{aligned} \frac{\varrho_1}{4}\|x_{k+1} - x_*\|^2 &\leqslant f(x_{k+1}) - f(x_*) \\ &\leqslant \varepsilon_k (f(x_k) - f(x_*)) \\ &\leqslant \varepsilon_k \varrho_n \|x_k - x_*\|^2. \end{aligned}$$

Also gilt

$$\|x_{k+1} - x_*\| \leqslant \underbrace{\sqrt{\frac{4\varrho_n\varepsilon_k}{\varrho_1}}}_{=:\varepsilon'_k} \|x_k - x_*\|$$

mit $(\varepsilon'_k) \to 0$.
Insbesondere gilt $\mu_1 = \mu_n$, also superlineare Konvergenz, wenn $(A_k) \to F(x_*)$.
Schlechte Konvergenz erhalten wir dagegen für $\mu_1 \ll \mu_2$, z. B. für das Gradientenverfahren. (Siehe auch Vorlesungsskriptum „Einführung in die Mathematische Optimierung“, Satz 3.2 über die Konvergenz des Gradientenverfahrens für quadratische Funktionen, sowie die daran anschließende Bemerkung.)

Lemma 2.1.8. Sei $(x_k) \to x_*$ mit $\nabla f(x_*) = 0$, $F(x_*)$ positiv definit, $f \in C_3(U(x_*))$. Sei $d_k = -A_k^{-1} f(x_k)^T$ mit beschränkter gleichmäßig positiv definiter Matrixfolge (A_k). Setzt man

$$\lambda_k := \frac{-\nabla f(x_k) d_k}{2(f(x_k + d_k) - f(x_k) - \nabla f(x_k) d_k)},$$

so erfüllt für $0 < \alpha < \frac{1}{2}$, $\beta' := \frac{1}{2\lambda_{\max}(F(x_*))}$, $\beta'' := \frac{2}{\lambda_{\min}(F(x_*))}$ und hinreichend großes k die Schrittweite λ_k die Bedingungen der Goldstein-Armijo-Suche für $i = 0$ und λ_k ist asymptotisch exakt von der Ordnung 1.

Beweis. Mittels Taylorentwicklung zeigt man, daß der Nenner von λ_k gleich $d_k^T F(x_k) d_k + O(\|d_k\|^3) = d_k^T F(x_*) d_k + O(\|d_k\|^3)$ ist. Im folgenden Beweis ist zu beachten, daß aufgrund der

Voraussetzungen die Terme $\|d_k\|$, $\|\nabla f(x_k)\|$ und $\|x_k - x_*\|$ die gleiche Größenordnung haben und gegenseitig durch $O()$-Terme abgeschätzt werden können. Es gilt also

$$\begin{aligned}
\lambda_k &= \frac{-\nabla f(x_k)d_k}{\underbrace{d_k^T F(x_*)d_k}_{=\Theta(\|d_k\|^2)} + O(\|d_k\|^3)} \\
&= \frac{-\nabla f(x_k)d_k}{d_k^T F(x_*)d_k(1 + O(\|d_k\|))} \\
&= \frac{-\nabla f(x_k)d_k}{d_k^T F(x_*)d_k}\Big(1 + O(\|d_k\|)\Big) \\
&= \underbrace{\frac{-\nabla f(x_k)d_k}{d_k^T F(x_*)d_k}}_{=\Theta(1)}\Big(1 + O(\|\nabla f(x_k)\|)\Big).
\end{aligned}$$

Insbesondere ist λ_k beschränkt, da d_k streng gradientenbezogen ist. Mit $\varrho_1 = \lambda_{\min}(F(x_*))$ und $\varrho_n = \lambda_{\max}(F(x_*))$ gilt $\varrho_1\|d_k\|^2 \leqslant d_k^T F(x_*)d_k \leqslant \varrho_n\|d_k\|^2$. Für hinreichend großes k gilt $\frac{1}{2} \leqslant 1 + O(\|\nabla f(x_k)\|) \leqslant 2$, also

$$\frac{1}{2\varrho_n}\frac{-\nabla f(x_k)d_k}{\|d_k\|^2} \leqslant \lambda_k \leqslant \frac{2}{\varrho_1}\frac{-\nabla f(x_k)d_k}{\|d_k\|^2},$$

d. h. λ_k ist als Startschrittweite der Goldstein-Armijo-Suche wählbar.

Wir zeigen nun, daß λ_k asymptotisch exakt von der Ordnung 1 ist:

$$\begin{aligned}
\nabla f(x_k + \lambda_k d_k)d_k &= (\nabla f(x_k) + \lambda_k d_k^T F(x_k) + O(\|d_k\|^2))d_k \\
&= \nabla f(x_k)d_k + \lambda_k d_k^T F(x_k)d_k + O(\|d_k\|^3) \\
&= \nabla f(x_k)d_k O(\|\nabla f(x_k)\|) + O(\|\nabla f(x_k)\|^3) \\
&= \nabla f(x_k)d_k O(\|\nabla f(x_k)\|).
\end{aligned}$$

Die Schrittweite ist akzeptabel, denn

$$\begin{aligned}
f(x_k + \lambda_k d_k) - f(x_k) &= \lambda_k\Big(\nabla f(x_k)d_k + \frac{1}{2}\lambda_k d_k^T F(x_k)d_k + O(\|d_k\|^3\Big) \\
&= \lambda_k\Big(\frac{1}{2}\nabla f(x_k)d_k + O(\|d_k\|^3)\Big) \\
&= \lambda_k\Big(\alpha\nabla f(x_k)d_k + \underbrace{\Big(\tfrac{1}{2} - \alpha\Big)}_{>0}\underbrace{\nabla f(x_k)d_k}_{<0} + O(\|d_k\|^3)\Big).
\end{aligned}$$

Für genügend großes k gilt

$$f(x_k + \lambda_k d_k) - f(x_k) \leqslant \lambda_k\alpha\nabla f(x_k)d_k,$$

also erfüllt λ_k die Bedingung der Goldstein-Armijo-Suche bereits für $i = 0$. □

Bemerkung 2.1.9. (Zu Lemma 2.1.8)

1. Bei quadratischen Funktionen ist λ_k exakt, d. h. in jedem Schritt wird das Minimum von f in Richtung d_k bestimmt.

2. Im Beweis sieht man, daß jede Schrittweite der Form
$$\lambda_k = \frac{-\nabla f(x_k)d_k}{d_k^T F(x_*)d_k}\Big(1 + O(\|d_k\|)\Big)$$
asymptotisch exakt von der Ordnung 1 und für genügend großes k akzeptabel ist.

2.1.5 Regularisierung nach Levenberg-Marquardt

Wie wir in Lemma 2.1.4 und Lemma 2.1.8 gesehen haben, ist die Verwendung von Abstiegsrichtungen der Form $d_k := -A_k^{-1}\nabla f(x_k)$ interessant, weil gute Aussagen über das Konvergenzverhalten der Folge (x_k) gemacht werden können und weil die Schrittweite leicht bestimmbar ist, wenn die Matrixfolge (A_k) beschränkt und gleichmäßig positiv definit ist.

Häufig kann man zwar die Beschränkheit von (A_k) zeigen, jedoch nicht die gleichmäßige positive Definitheit. In diesem Fall kann man sich mit der **Regularisierung nach Levenberg-Marquardt** abhelfen, die den kleinsten Eigenwert von A_k gegebenenfalls auf einen Wert größer als Null anhebt, indem ein skalares Vielfaches der Einheitsmatrix zu A_k addiert wird.

Falls A' symmetrisch und beschränkt ist, etwa $\|A'\| \leqslant R'$ mit kleinstem Eigenwert λ', so wähle für ein festes $\delta > 0$ eine Näherung $\lambda \in [\lambda' - \delta, \lambda' + \delta]$ und setze $A := A' + \mu I$ für $\mu := \max\{0, 2\delta - \lambda\}$. Dann gilt $\delta y^T y \leqslant y^T A y \leqslant (2R' + 3\delta)y^T y$.

Beweis. Die Abschätzung $\lambda_{\min}(A) \geqslant \mu + \lambda' = \underbrace{\max\{\lambda' - \delta, \lambda' + \delta - \lambda\}}_{\geqslant 0} + \delta \geqslant \delta$ ist trivial.

Wir zeigen nun, daß $\lambda_{\max}(A) \leqslant \|A\| \leqslant R' + \mu$ gilt. Für $\mu = 0$ ist diese Behauptung richtig. Anderenfalls ist $\mu = 2\delta - \lambda > 0$, also gilt $\mu \leqslant 2\delta - \lambda \leqslant 2\delta - (\lambda' - \delta) = 3\delta - \lambda' \leqslant 3\delta + R'$, womit die Behauptung ebenfalls folgt. $\square$

Wir können also aus einer durch R' beschränkten Matrixfolge eine beschränkte, gleichmäßig positiv definite Matrixfolge mit Schranken δ und $2R' + 3\delta$ gewinnen.

Die Bestimmung eines Schätzwertes für den kleinsten Eigenwert von A' ist relativ aufwendig, doch man kann dies gleich mit der Berechnung der Abstiegsrichtung kombinieren.

Algorithmus 2.1.2. (Berechnung von d bei Regularisierung von A')

1. Bringe A' mit einer unitären Transformationsmatrix U auf Tridiagonalform, etwa durch Householder-Transformationen (siehe z. B. [9]).
   $T := UA'U^T$

2. Bestimme $\lambda \approx \lambda_{\min}(A') + [-\delta, \delta]$ mit Hilfe des Bisektionsverfahrens Sturmsche Kette. (siehe auch [8])
   Setze $\mu := \max\{0, 2\delta - \lambda\}$.

3. Bestimme die Cholesky-Zerlegung von $T + \mu I =: LL^T$.
   Diese ist aufgrund der Bandstruktur von T sehr schnell berechenbar.

4. Löse die Dreiecksysteme $Lz = U\nabla f(x)^T$ und $L^T y = z$.
   Setze $d := U^T y$. (Damit gilt $(A' + \mu I)d = \nabla f(x)^T$.)

2.2 Newton-Verfahren

Definition 2.2.1. Sei $D \subseteq \mathbf{R}^n$ nichtleer und konvex. $f\colon D \to \mathbf{R}$ heißt **gleichmäßig konvex** auf D, wenn es ein $\gamma > 0$ gibt, so daß für alle $x, y \in D$ und alle $\lambda \in [0,1]$ gilt:

$$\lambda f(x) + (1-\lambda) f(y) \geqslant f(\lambda x + (1-\lambda)y) + \gamma\lambda(1-\lambda)\|x-y\|^2.$$

Sei f gleichmäßig konvex. Wählt man $A_k := F(x_k)$, also $d_k = -F(x_k)^{-1}\nabla f(x_k)$ für $k \in \mathbf{N}_0$, so erhält man das **gedämpfte Newton-Verfahren**. (Das gedämpfte Newton-Verfahren unterscheidet sich vom Newton-Verfahren darin, daß Schrittweiten unterschiedlich von 1 zugelassen sind.)

Lemma 2.2.1. Es sei $D \subseteq \mathbf{R}^n$ offen und konvex und $f \in C_2(D)$.
Dann ist f genau dann gleichmäßig konvex auf D, wenn es ein $\gamma > 0$ gibt, so daß

$$\inf\{\lambda_{\min}(F(x)) \mid x \in D\} \geqslant \gamma$$

gilt.

Beweis. Sei $\gamma > 0$ gegeben mit $\inf\{\lambda_{\min}(F(x)) \mid x \in D\} \geqslant \gamma$. Seien $x, y \in D$. Dann gilt

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)(y-x) + \frac{1}{2}(y-x)^T F(x + \theta(y-x))(y-x) \text{ mit einem } \theta \in (0,1) \\ &\geqslant f(x) + \nabla f(x)(y-x) + \frac{1}{2}\gamma\|y-x\|^2, \end{aligned}$$

also $f(y) - f(x) \geqslant \nabla f(x)(y-x) + \frac{1}{2}\gamma\|y-x\|^2$.
Mit $z := \lambda x + (1-\lambda)y \in D$, also $x - z = (1-\lambda)(x-y)$ und $y - z = \lambda(y-x)$ folgt

$$f(y) - f(z) \geqslant \nabla f(z)(y-z) + \frac{1}{2}\gamma\lambda^2\|y-x\|^2$$

und

$$f(x) - f(z) \geqslant \nabla f(z)(x-z) + \frac{1}{2}\gamma(1-\lambda)^2\|y-x\|^2.$$

Wird die erste Ungleichung mit $1-\lambda$ und die zweite mit λ multipliziert und beide addiert, so erhält man

$$(1-\lambda)f(y) + \lambda f(x) - f(z) \geqslant \frac{1}{2}\gamma\lambda(1-\lambda)\|y-x\|^2,$$

was nach Rücksubstitution von z direkt die Definition von gleichmäßig konvexen Funktionen ergibt.
Sei umgekehrt f gleichmäßig konvex. Dann ist

$$\lambda f(x) + (1-\lambda)f(y) \geqslant f(\lambda x + (1-\lambda)y) + \lambda(1-\lambda)\gamma\|x-y\|^2,$$

also

$$f(x) - f(y) \geqslant \frac{1}{\lambda}\Big[f(\underbrace{\lambda x + (1-\lambda)y}_{=y+\lambda(x-y)}) - f(y)\Big] + (1-\lambda)\gamma\|x-y\|^2$$

Bei Grenzübergang $\lambda \to 0$ ergibt sich somit

$$f(x) - f(y) \geqslant \nabla f(y)(x-y) + \gamma\|x-y\|^2,$$

was natürlich gilt, wenn man x und y vertauscht. Werden beide Ungleichungen summiert, so erhält man

$$(\nabla f(x) - \nabla f(y))(x-y) \geqslant 2\gamma\|x-y\|^2.$$

Die linke Seite kann man auch als

$$(x-y)^T\left(\int_0^1 F(y+\theta(x-y))d\theta\right)(x-y)$$

schreiben. Division durch $\|x-y\|^2$ ergibt nun

$$\min\left\{\frac{(x-y)^T}{\|x-y\|}\int_0^1 F(y+\theta(x-y))d\theta\frac{(x-y)}{\|x-y\|}\;\middle|\; x,y\in D, x\neq y\right\} \geqslant 2\gamma$$

Insbesondere liefert jede Folge mit $x \to y$ nun nach Grenzübergang $\lambda_{\min}(F(y)) \geqslant 2\gamma$. □

Satz 2.2.2. Sei $f \in C_3(\mathbf{R}^n)$ gleichmäßig konvex, $x_0 \in \mathbf{R}^n$. Sei λ_k $(k \in \mathbf{N}_0)$ akzeptabel und asymptotisch exakt von der Ordnung 1. Dann konvergiert die Folge (x_k), definiert durch $x_{k+1} := x_k + \lambda_k d_k$ $(k \in \mathbf{N}_0)$ mit $d_k = -F(x_k)^{-1}\nabla f(x_k)$, quadratisch gegen das eindeutige Minimum x_* von f. Für hinreichend großes k ist die Schrittweite $\lambda_k' \equiv 1$ akzeptabel und asymptotisch exakt von der Ordnung 1.

Beweis. Da f gleichmäßig konvex, ist $N(x_k)$ kompakt und $(F(x_k))$ gleichmäßig positiv definit und beschränkt. Nach Lemma 2.1.4 ist d_k streng gradientenbezogen. λ_k akzeptabel impliziert dann $\nabla f(x_k) \to 0$, also erfüllt jeder Häufungspunkt x_* der Folge (x_k) die Bedingung $\nabla f(x_*) = 0$. Also ist x_* der einzige Häufungspunkt, d. h. es gilt $(x_k) \to x_*$. Also sind die Voraussetzungen von Lemma 2.1.5 erfüllt. Wie dort im Beweis gezeigt, gilt

$$F(x_k) = F(x_*) + O(\|x_k - x_*\|) = F(x_*) + O(\|\nabla f(x_k)\|).$$

Nach Lemma 2.1.5 gilt

$$\begin{aligned}
\lambda_k &= \frac{-\nabla f(x_k)d_k}{d_k^T F(x_k)d_k}\Big(1 + O(\|\nabla f(x_k)\|)\Big)\\
&= \frac{\nabla f(x_k)F(x_k)^{-1}\nabla f(x_k)^T}{\nabla f(x_k)F(x_k)^{-1}F(x_k)F(x_k)^{-1}\nabla f(x_k)^T}\Big(1 + O(\|\nabla f(x_k)\|)\Big)\\
&= 1 + O(\|\nabla f(x_k)\|),
\end{aligned}$$

also ist λ_k für genügend große k akzeptabel (siehe Bemerkung zu Lemma 2.1.8). Da auch die konstante Folge (λ_k'), definiert durch $\lambda_k' = 1$, diese Form hat, ist sie nach der gleichen

Bemerkung ebenfalls asymptotisch exakt von der Ordnung 1 und für genügend große k ist λ_k' akzeptabel.
Nun erhält man durch Taylor-Entwicklung von ∇f

$$\begin{aligned}\nabla f(x_{k+1}) &= \nabla f(x_k) + d_k^T \lambda_k F(x_k) + O(\|d_k\|^2)\\ &= \nabla f(x_k) + (1 + O(\|\nabla f(x_k)\|))d_k^T F(x_k) + O(\|d_k\|^2)\\ &= O(\|\nabla f(x_k)\|^2)\\ &= O(\|x_k - x_*\|^2).\end{aligned}$$

Andererseits gilt

$$\begin{aligned}\|\nabla f(x_{k+1})\| &= \|\nabla f(x_{k+1}) - \underbrace{\nabla f(x_*)}_{=0}\|\\ &= \|H(x_{k+1}, x_*)(x_{k+1} - x_*)\|\\ &\geqslant \frac{\|x_{k+1} - x_*\|}{\|H(x_{k+1}, x_*)^{-1}\|}\end{aligned}$$

mit $H(x,y) = \int_0^1 F(y+\tau(x-y))d\tau$. Da f gleichmäßig konvex ist, gilt $\|H(x,y)^{-1}\| \leqslant c$ für ein festes $c > 0$. Also folgt $\|x_{k+1} - x_*)\| = O(\|\nabla f(x_{k+1})\|)$. Zusammen mit der vorigen Abschätzung gilt also $\|x_{k+1} - x_*\| = O(\|x_k - x_*\|^2)$, womit die quadratische Konvergenz gezeigt ist. □

Bemerkung 2.2.1. Sei f gleichmäßig konvex. Dann gilt:

1. Das gedämpfte Newton-Verfahren ist ebenfalls quadratisch konvergent.
   Dies ist leicht einzusehen, weil nach Satz 2.2.2 die Schrittweiten für genügend großes k gegen 1 gehen, das gedämpfte Newton-Verfahren also in das Newton-Verfahren übergeht.

2. Bei Verwendung der Suchrichtungen $d_k := -F(x_0)^{-1}\nabla f(x_k)^T$ zwecks Aufwandverringerung (**vereinfachtes Newton-Verfahren**) hat man zwar weiterhin globale, aber nur lineare Konvergenz.

3. Aus Differenzenquotienten mit genügend kleiner Diskretisierung bestimmte A_k bleiben gleichmäßig postiv definit. Dabei hat die Matrix A_k die gleiche Struktur wie $F(x_k)$, d. h. dünnbesetzte $F(x_k)$ liefern auch dünnbesetzte A_k.

Das im folgenden angegebene Verfahren bestimmt A_k alle n Schritte neu und besitzt trotzdem superlineare Konvergenz.

Algorithmus 2.2.1. (Rekursive Definition der A_k)
Es sei $\tau_k := \min\{\tau, \|\nabla f(x_k)\|\}$ für festes, genügend kleines $\tau > 0$.
Falls $k \equiv 0 \pmod n$, dann
 setze $A_k = 0$,
 für $j = 1, \ldots, n$ berechne
  $h_j := \mathrm{diag}(\underbrace{0,\ldots,0}_{j-1}, \frac{1}{2}, \underbrace{1,\ldots,1}_{n-j})(\nabla f(x_k + \tau_k e_j) - \nabla f(x_k))^T \tau_k^{-1}$
  $A_k := A_k + h_j e_j^T + e_j h_j^T$
sonst
 setze $A_k = A_{k-1}$.

Satz 2.2.3. Die Voraussetzungen von Satz 2.1.7 seien erfüllt. Die Richtungen seien mit Hilfe der in Algorithmus 2.2.1 definierten A_k berechnet, mit hinreichend kleinem τ. Dann konvergiert (x_k) superlinear gegen das Minimum x_* von f. Dabei ist $\lambda_k = 1 + o(1)$ und $\lambda'_k \equiv 1$ ist asymptotisch exakt von der Ordnung 1.

Beweis. Superlineare Konvergenz folgt, da $(A_k) \to F(x_*)$. Nach Wahl der τ_k gilt $A_k = F(x_k) + o(1)$ für $k \equiv 0 \pmod n$, woraus man $\lambda_k = 1 + o(1)$ und $\lambda'_k \equiv 1$ asymptotisch exakt folgert. ($O(\|\nabla f(x_k)\|)$ muß durch $o(1)$ ersetzt werden.) □

Bemerkung 2.2.2. Ist f nicht gleichmäßig konvex, so ist die Anwendung der Regularisierungstechnik möglich, aber man hat global ähnlich schlechte Konvergenz wie beim Gradientenverfahren.

2.3 Quasi-Newton-Verfahren

Im folgenden sei $d_k := -A_k^{-1}\nabla f(x_k)^T$ gewählt, wobei A_k aus den Iterationspunkten $x_0, \ldots, x_k$ und den Gradienten $\nabla f(x_0), \ldots, \nabla f(x_k)$ berechnet sein soll. Die d_k sind also streng gradientenbezogen. f kann in der Nähe des Minimums durch eine quadratische Funktion approximiert werden:

$$\begin{aligned} f(x) &= c - b^T x + \frac{1}{2} x^T A x \\ \nabla f(x) &= (Ax - b)^T, \end{aligned}$$

wobei A symmetrisch und positiv definit ist. Hiermit gilt dann

$$A(x_{k+1} - x_k) = \nabla f(x_{k+1})^T - \nabla f(x_k)^T. \tag{2.11}$$

Sind die Richtungen $x_{k+1} - x_k$ linear unabhängig für $0 \leqslant k < n$, so kann A aus den x_k und $\nabla f(x_k)$ mit Hilfe der Formel (2.11) berechnet werden.
Die Idee der Quasi-Newton-Verfahren ist die folgende: Aktualisiere A_k stets so, daß in Schritt k gilt:

$$A_{k+1}(x_{j+1} - x_j) = \nabla f(x_{j+1})^T - \nabla f(x_j)^T \text{ für alle } j \in \{0, \ldots, k\}.$$

Für A_0 wird dabei eine beliebige symmetrische positiv definite Matrix benutzt, meist nimmt man hier $A_0 = I$.

Definition 2.3.1. Ein Verfahren, bei dem A_{k+1} aus A_k, s_k und y_k (mit $s_k := x_{k+1} - x_k$ und $y_k^T := \nabla f(x_{k+1}) - \nabla f(x_k)$) hervorgeht, und welches die **Sekantenrelation**

$$A_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1})^T - \nabla f(x_k)^T$$

erfüllt, heißt **Quasi-Newton-Verfahren**.

In der Praxis benutzt man meist additive Änderungen der Form $A_{k+1} = A_k + \Delta_k$, wobei Δ_k möglichst niedrigen Rang hat (Einfachheit der Updates).

Rang-1-Updates

Der Ansatz $A_{k+1} = A_k + \beta_k z_k z_k^T$ führt auf $\beta_k := \frac{1}{(y_k - A_k s_k)^T s_k}$ und $z_k = y_k - A_k s_k$ (Herleitung als Übungsaufgabe). Die gewonnene Update-Formel

$$A_{k+1} = A_k + \frac{1}{(y_k - A_k s_k)^T s_k}(y_k - A_k s_k)(y_k - A_k s_k)^T$$

wird gewöhnlich als **SR1-Formel** bezeichnet.

Beispiel 2.3.1. Wie dieses Beispiel zeigt, ist bei Anwendung der SR1-Formel nicht gewährleistet, daß A_{k+1} positiv definit ist: Sei etwa

$$f(x) := \frac{1}{2} x^T \begin{pmatrix} \frac{1}{10} & 0 \\ 0 & \frac{1}{100} \end{pmatrix} x, \ x_0 := \begin{pmatrix} 10 \\ 100 \end{pmatrix} \text{ und } A_0 := I.$$

Die optimale Schrittweite ist $\tau_0 = \frac{200}{11}$. Damit folgt $x_1 = \frac{1}{11} \begin{pmatrix} -90 \\ 900 \end{pmatrix}$.

Die sich hieraus ergebende Matrix $A_1 = \begin{pmatrix} \frac{19}{100} & \frac{-891}{1890} \\ \frac{-891}{1890} & \frac{9801}{18900} \end{pmatrix}$ ist indefinit.

Verzichtet man sogar auf die Symmetrieforderung für A_{k+1}, so sind weitere Updateverfahren möglich. Ein bekanntes Rang-1-Verfahren ist das **Update nach Broyden** (1965):

$$A_{k+1} = A_k + \frac{(y_k - A_k s_k) s_k^T}{s_k^T s_k}.$$

Rang-2-Updates

Noch älter als das Broyden-Verfahren ist ein sehr gutes Rang-2-Update-Verfahren, das **DFP-Verfahren**, welches 1959 von Davidon vorgeschlagen und 1963 von Fletcher und Powell untersucht und verbessert wurde. Das DFP-Update gewährleistet ebenfalls nicht, daß A_{k+1} positiv definit ist.

Die Verfahren mit der Updateformel

$$A_{k+1} = A_k - \frac{A_k s_k (A_k s_k)^T}{s_k^T A_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + \varphi A_k s_k v_k v_k^T \text{ mit } v_k := \frac{1}{y_k^T s_k} y_k - \frac{1}{s_k^T A_k s_k} A_k s_k \quad (2.12)$$

und konstantem $\varphi \in [0,1]$ heißen Verfahren der konvexen **Broyden-Klasse**. Für $\varphi = 1$ erhalten wir das DFP-Update.

Ein weiteres bekanntes Rang-2-Update-Verfahren, das BFGS-Verfahren, werden wir nun sehr genau untersuchen. Weil das BFGS-Verfahren bei geeigneter Schrittweitenwahl nur positiv definite A_k bestimmt, heißt das Verfahren in der Literatur auch das „positiv definite Sekantenupdateverfahren“.

2.3.1 Das BFGS-Verfahren

Da mit Rang-1-Updates die notwendigen Forderungen nicht erfüllbar sind, muß mindestens mit Rang-2-Updates gearbeitet werden. Das beste bis heute bekannte Quasi-Newton-Verfahren ist das BFGS-Verfahren, benannt nach Broyden, Fletcher, Goldfarb und Shanno, die es 1970 unabhängig voneinander gefunden haben. Das BFGS-Verfahren ist ein Rang-2-Update-Verfahren, das als

$$\begin{aligned} \tilde{A}_k &= A_k - \tfrac{A_k s_k (A_k s_k)^T}{s_k^T A_k s_k} \\ A_{k+1} &= \tilde{A}_k + \tfrac{y_k y_k^T}{y_k^T s_k} \end{aligned}$$

geschrieben werden kann und somit zur Broyden-Klasse gehört. (Man setze in (2.12) $\varphi = 0$.)[1]

Bemerkung 2.3.1. (zur Idee des BFGS-Updates)

1. Das erste Update ist so gewählt, daß $\tilde{A}_k$ positiv semidefinit ist.

   Beweis. Es gilt $\tilde{A}_k s_k = A_k s_k - \frac{A_k s_k (A_k s_k)^T}{s_k^T A_k s_k} s_k = A_k s_k - A_k s_k \frac{(A_k s_k)^T s_k}{s_k^T A_k s_k} = 0$. Also gilt insbesondere $s_k^T \tilde{A}_k s_k = 0$. Andererseits folgt aus der Cauchy-Schwarzschen Ungleichung mit dem durch $< x, y >_A := x^T A y$ definierten Skalarprodukt und $\|x\|_A := \sqrt{< x, x >_A}$

   $$\|x\|_{A_k}^2 \|s_k\|_{A_k}^2 \geqslant < x, s_k >_{A_k}^2,$$

   also

   $$x^T A_k x s_k^T A_k s_k \geqslant (x^T A_k s_k)^2,$$

   woraus $x^T A_k x - \frac{x^T A_k s_k s_k^T A_k x}{s_k^T A_k s_k} = x^T \tilde{A}_k x \geqslant 0$ folgt. Voraussetzung hierbei ist, daß A_k positiv definit ist, da $< \cdot, \cdot >_A$ sonst kein Skalarprodukt definiert. (Semidefinitheit reicht für die Cauchy-Schwarzsche Ungleichung aus, aber dann ist nicht mehr gewährleistet, daß der Nenner echt positiv ist.)

   Also ist $\tilde{A}_k$ symmetrisch und positiv semidefinit, aber wegen $\tilde{A}_k s_k = 0$ nicht positiv definit. □

2. Das zweite Update ergibt sich zwangsläufig, wenn man ein weiteres Rang-1-Update ansetzt und fordert, daß A_{k+1} die Sekantenrelation erfüllen soll. Der gleiche Ansatz wie bei der SR1-Formel liefert hier unter Ausnutzung der Eigenschaft $\tilde{A}_k s_k = 0$ die erhaltene Formel. Man benötigt hier die zusätzliche Voraussetzung, daß der Nenner $y_k^T s_k$ echt positiv ist, was bei gleichmäßig konvexen Funktionen aber stets gegeben ist. Wir zeigen nun, daß A_{k+1} positiv definit ist.

   Beweis. Unter erneuter Ausnutzung von $y_k^T s_k > 0$ weisen wir zunächst nach, daß $A_k^{-\frac{1}{2}} A_{k+1} A_k^{-\frac{1}{2}}$ positiv definit ist:

   $$A_k^{-\frac{1}{2}} A_{k+1} A_k^{-\frac{1}{2}} = I - \frac{A_k^{\frac{1}{2}} s_k s_k^T A_k^{\frac{1}{2}}}{s_k^T A_k^{\frac{1}{2}} A_k^{\frac{1}{2}} s_k} + \frac{A_k^{-\frac{1}{2}} y_k y_k^T A_k^{-\frac{1}{2}}}{y_k^T A_k^{-\frac{1}{2}} A_k^{\frac{1}{2}} s_k},$$

[1] Das BFGS-Verfahren ist allerdings nicht als Spezialfall von (2.12) anzusehen; vielmehr ist (2.12) als Verallgemeinerung der erfolgreichen Update-Verfahren DFP und BFGS vorgeschlagen worden.

weshalb man die Substitutionen $\hat{y}_k := A_k^{-\frac{1}{2}} y_k$ und $\hat{s}_k := A_k^{\frac{1}{2}} s_k$ durchführt. Dann ist zu zeigen, daß

$$A_k^{-\frac{1}{2}} A_{k+1} A_k^{-\frac{1}{2}} = I - \frac{\hat{s}_k \hat{s}_k^T}{\hat{s}_k^T \hat{s}_k} + \frac{\hat{y}_k \hat{y}_k^T}{\hat{y}_k^T \hat{s}_k}$$

positiv definit ist. Sei $z \neq 0$. Man zerlegt z nun in den Anteil parallel zu $\hat{s}_k$ und den Anteil senkrecht dazu, etwa $z = \lambda \hat{s}_k + z_\perp$. Ausmultiplizieren ergibt nun

$$z^T \left(A_k^{-\frac{1}{2}} A_{k+1} A_k^{-\frac{1}{2}} \right) z = z_\perp^T z_\perp + \frac{(z^T \hat{y}_k)^2}{\hat{y}_k^T \hat{s}_k} \geqslant 0.$$

Ist nun $z_\perp \neq 0$, so gilt „echt größer“, anderenfalls ist bei $z \neq 0$ dann $z = \lambda \hat{s}_k$, woraus wegen $\hat{y}_k^T \hat{s}_k = y_k^T s_k > 0$ ebenfalls „echt größer“ folgt. Also ist die betrachtete Matrix positiv definit.

Also gilt

$$u^T A_{k+1} u = u^T \left(A_k^{\frac{1}{2}} A_k^{-\frac{1}{2}} \right) A_{k+1} \left(A_k^{-\frac{1}{2}} A_k^{\frac{1}{2}} \right) u = z^T \left(A_k^{-\frac{1}{2}} A_{k+1} A_k^{-\frac{1}{2}} \right) z > 0$$

für $u \neq 0$, also $z \neq 0$. Deshalb ist A_{k+1} positiv definit. □

Satz 2.3.1. Es sei $f(x) = c - b^T x + \frac{1}{2} x^T A x$ mit symmetrischer positiv definiter Matrix A. Es sei $x_0 \in \mathbf{R}^n, A_0 \in \mathbf{R}^{n \times n}$ symmetrisch und positiv definit. Die Folgen (x_k) und (A_k) für $0 \leqslant k \leqslant N-1$ seien gegeben durch

$$\begin{aligned}
d_k &:= -A_k^{-1} \nabla f(x_k)^T, \\
\tau_k &:= \frac{-\nabla f(x_k) d_k}{d_k^T A d_k}, \\
x_{k+1} &:= x_k + \tau_k d_k, \\
s_k &:= x_{k+1} - x_k, \\
y_k &:= \nabla f(x_{k+1})^T - \nabla f(x_k)^T \text{ und} \\
A_{k+1} &:= A_k - \frac{A_k s_k (A_k s_k)^T}{s_k^T A_k s_k} + \frac{y_k y_k^T}{y_k^T s_k},
\end{aligned}$$

wobei $N := \min\{k \mid \nabla f(x_k) = 0\}$, d. h. N ist durch den BFGS-Algorithmus definiert.
Dann folgt:

1. $d_k \neq 0$ und A_{k+1} ist symmetrisch und positiv definit für $0 \leqslant k \leqslant N-1$. τ_k ist die optimale Schrittweite.

2. Es gilt $d_i^T A d_j = 0$ für $0 \leqslant i < j \leqslant k$ und $A_{k+1}^{-1} A d_j = d_j$ für $0 \leqslant j \leqslant k$ und $0 \leqslant k \leqslant N-1$.

3. Es gilt $N \leqslant n$ und $x_N = x_* := A^{-1} b$.

4. Falls $N = n$, so gilt $A_n = A$.

Beweis.

1. Da A_k positiv definit, ist $d_k \neq 0$. Weiterhin ist $\tau_k > 0$, da auch A positiv definit ist. Also ist $s_k \neq 0$ und wegen $\nabla f(x) = (Ax - b)^T$ ist $y_k^T s_k = s_k^T A s_k > 0$. Hiermit folgt nach der vorigen Bemerkung die positive Definitheit von A_{k+1}, denn die Voraussetzung $s_k^T y_k = s_k^T A s_k > 0$ ist gegeben.

   τ_k ist die optimale Schrittweite, da Taylorentwicklung von ∇f um x

   $$\nabla f(x + \tau d) = \nabla f(x) + \tau d^T \underbrace{F(x)}_{=A} + \frac{1}{2}\tau^2 0$$

   ergibt. Multiplikation mit d ergibt nun

   $$0 = \nabla f(x + \tau d)d = \nabla f(x)d + \tau d^T A d,$$

   woraus die Behauptung nach Auflösung nach τ folgt.

2. Beweis induktiv über k: Sei $g_k := \nabla f(x_k)^T$.

   Induktionsanfang: $k = 0$: Es ist zu zeigen, daß $A_1^{-1} A d_0 = d_0$, also $A d_0 = A_1 d_0$. Es gilt

   $$A_1 d_0 = \frac{1}{\tau_0} A_1 s_0 = \frac{1}{\tau_0} y_0 = \frac{1}{\tau_0} A s_0 = A d_0$$

   wegen $d_k = \frac{1}{\tau_k} s_k$, der Sekantenrelation $A_{k+1} s_k = y_k$, $y_k = A s_k$ und $\tau_k \neq 0$.

   Induktionsschritt: Die Behauptung gelte für $k - 1$. Die Äquivalenzen $d_k = \frac{s_k}{\tau_k}$ und $y_k = A s_k = \tau_k A d_k$ werden häufig benutzt.

   Es ist nach Teleskopsummation $g_k = g_{i+1} + \sum_{j=i+1}^{k-1} y_j = g_{i+1} + \sum_{j=i+1}^{k-1} \tau_j A d_j$ für $i < k$. Nutzt man nun $d_i^T g_{i+1} = 0$ (Exaktheit der Schrittweite) und die Induktionsvoraussetzung $d_i^T A d_j = 0$ für $j \leqslant k - 1$ aus, so folgt

   $$d_i^T g_k = \underbrace{d_i^T g_{i+1}}_{=0} + \sum_{j=i+1}^{k-1} \tau_j \underbrace{d_i^T A d_j}_{=0} = 0$$

   für $i < k$.

   Wegen $d_i^T A A_k^{-1} = d_i^T$ für $i < k$ ist

   $$d_i^T A d_k = -d_i^T A A_k^{-1} g_k = -d_i^T g_k = 0$$

   für $i < k$, womit die erste Behauptung für k gezeigt ist.

   Weiterhin gilt aufgrund von $A_k^{-1} A d_i = d_i$ für $0 \leqslant i \leqslant k - 1$

   $$y_k^T A_k^{-1} A d_i = \tau_k d_k^T A d_i = 0$$

   für $0 \leqslant i \leqslant k - 1$.

Nun folgt mit $s_k = \tau_k d_k$ und $y_k = As_k = \tau_k Ad_k$ und $d_k^T Ad_j = 0$ sowie $A_k d_j = Ad_j$ für $j \leqslant k-1$:

$$\begin{aligned} A_{k+1}d_j &= A_k d_j - A_k s_k \frac{s_k^T A_k d_j}{s_k^T A_k s_k} + y_k \frac{y_k^T d_j}{y_k^T s_k} \\ &= Ad_j - A_k d_k \frac{d_k^T Ad_j}{d_k^T A_k d_k} - \tau_k Ad_k \frac{d_k^T Ad_j}{d_k^T Ad_k} \\ &= Ad_j \end{aligned}$$

für $j \leqslant k-1$. Aus der Sekantenrelation folgt nun noch $A_{k+1}d_k = Ad_k$, denn $\tau_k Ad_k = As_k = y_k = A_{k+1}s_k = \tau_k A_{k+1}d_k$, womit die zweite Behauptung bewiesen ist.

3. Man definiert durch $< a, b >_A := a^T Ab$ ein Skalarprodukt. Wegen 2. bilden die d_j bzgl. $< \cdot, \cdot >_A$ ein Orthogonalsystem, sind also insbesondere voneinander linear unabhängig. Da der Raum die Dimension n hat, muß das Verfahren spätestens im n-ten Schritt abbrechen.

   Bei Abbruch gilt $\nabla f(x_N) = 0$ laut Definition von N. Da $x_* = A^{-1}b$ die einzige Nullstelle von ∇f ist, folgt $x_N = x_*$.

4. Sei $n = N$. Dann gilt nach 2. $A_n d_j = Ad_j$ für $j < n$, also $(A_n - A)d_j = 0$. Da d_0 bis d_{n-1} den gesamten Raum aufspannen, folgt $A_n - A = 0$, also $A_n = A$.

□

2.3.2 Cholesky-Update von A_k beim BFGS-Verfahren

Anstatt A_k in jeder Iteration neu zu invertieren (welches mit praktikablen Methoden den Aufwand $O(n^3)$ bedeutet), kann auch direkt die Inverse von A_k mit Aufwand $O(n^2)$ upgedated werden. Grundlage hierfür ist das folgende Lemma:

Lemma 2.3.2. (Sherman, Morrison und Woodbury)
Es sei $A \in \mathbf{R}^{n\times n}$ regulär und $u, v \in \mathbf{R}^n$. Dann ist $A + uv^T$ genau dann regulär, wenn

$$\varrho := 1 + v^T A^{-1} u \neq 0.$$

In diesem Fall gilt

$$(A + uv^T)^{-1} = A^{-1} - \frac{1}{\varrho} A^{-1} uv^T A^{-1}. \tag{2.13}$$

Beweis. (Übungsaufgabe) □

Durch zweifache Anwendung der Sherman-Morrison-Woodbury-Formel (2.13) erhält man mit $q_k := s_k^T y_k$ das Ergebnis

$$A_{k+1}^{-1} = A_i^{-1} - \frac{1}{q_k}\Big(s_k y_k^T A_k^{-1} + A_k^{-1} y_k s_k^T\Big) + \frac{s_k s_k^T}{q_k}\Big(1 + \frac{y_k^T A_k^{-1} y_k}{q_k}\Big).$$

(Die Herleitung dieser Formel ist Übungsaufgabe.)

Dieses Update-Verfahren weist jedoch nicht die beste numerische Stabilität auf. Eine Alternative ist, stets die Cholesky-Zerlegung von A_k zu speichern und das Update direkt auf die Zerlegung anzuwenden.

Da das BFGS-Update als aus zwei Rang-1-Updates bestehend aufgefaßt werden kann, kann auch das Update der Cholesky-Zerlegung in zwei einfache Schritte aufgeteilt werden. Die Reihenfolge der Schritte ist hier wesentlich. Da $\tilde{A}_k$ singulär ist, muß zunächst

$$\tilde{A}'_k = A_k + \frac{y_k y_k^T}{y_k^T s_k} =: \tilde{L}'_k \left(\tilde{L}'_k\right)^T$$

berechnet werden und danach

$$A_{k+1} = \tilde{A}'_k - \frac{A_k s_k (A_k s_k)^T}{s_k^T A_k s_k} = L_{k+1} L_{k+1}^T.$$

Der folgende Algorithmus bestimmt $\tilde{L}$ aus der Gleichung $\tilde{L}\tilde{L}^T = LL^T + aa^T$ zu gegebenen L und a.

Algorithmus 2.3.1. (Update einer Cholesky-Zerlegung, Goldfarb, 1976)

1. Löse das Dreiecksystem $Lb = a$.

2. Berechne Givens-Rotationen $U_n \cdots U_2$ der Bandbreite 1, so daß

   $$U_2 \cdots U_n b = \pm \|b\| e_1$$

   gilt.

3. Berechne $H := LU_n \cdots U_2$. H ist eine untere Hessenberg-Matrix. Weiterhin gilt

   $$LL^T + aa^T = L(I + bb^T)L^T = H(I + \|b\|^2 e_1 e_1^T)H^T.$$

4. Setze $D := \operatorname{diag}(\sqrt{1 + \|b\|^2}, 1, 1, \ldots, 1)$.
   Setze $\hat{H} := HD$, also $LL^T + aa^T = \hat{H}\hat{H}^T$.

5. Berechne Givens-Rotationen $W_2, \ldots, W_n$, so daß $\hat{H}W_2 \cdots W_n =: \tilde{L}$ eine untere Dreiecksmatrix ist.

Bemerkung 2.3.2. Der Aufwand des Cholesky-Updates nach Algorithmus 2.3.1 ist $O(n^2)$. Für Rang-1-Updates mit Subtraktion kann das Verfahren analog angewandt werden, aus Gründen numerischer Stabilität wird es jedoch häufig noch etwas modifiziert.

2.3.3 Die Restart-Version des BFGS-Verfahrens

Um zu vermeiden, daß die Konditionszahl von A_k zu schlecht wird, kann man nach je m Iterationen (oder bei Bedarf) einen **Restart** einfügen, d. h. man ersetzt in den Restart-Iterationen das BFGS-Update durch eine Neuzuweisung auf eine gut konditionierte Matrix, etwa das Vielfache einer Einheitsmatrix.

Das Verfahren wird wie folgt durchgeführt:

Algorithmus 2.3.2. (Restart-Version des BFGS-Verfahrens)
Es sei x_0 gegeben, A_0 symmetrisch und positiv definit, B_j ($j \in \mathbf{N}$) gleichmäßig positiv definit.
Für $k := 0, 1, 2, \ldots$
  bestimme $d_k := -A_k^{-1} \nabla f(x_k)^T$,
  bestimme eine Schrittweite λ_k, z. B. mit der Goldstein-Armijo-Suche,
  setze $x_{k+1} := x_k + \lambda_k d_k$,
  falls $k + 1 = jm$ für ein $j \in \mathbf{N}$,
    setze $A_{k+1} := B_j$,
  sonst
    führe das BFGS-Update durch.

Die B_j sollen gleichmäßig positiv definite Matrizen sein, z. B. $B_j = I$. Ist f gleichmäßig konvex, so ist die Skalierung

$$B_j = \frac{\|\nabla f(x+s) - \nabla f(x)\|^2}{(\nabla f(x+s) - \nabla f(x))s}$$

für $x = x_{k+1}$ und ein $s \neq 0$ sinnvoll.

Satz 2.3.3. Sei $f \in C_3(\mathbf{R}^n)$ gleichmäßig konvex. Sei $x_0 \in \mathbf{R}^n$, A_0 symmetrisch und positiv definit. Die Folgen der (x_k) und (A_k) seien durch das BFGS-Verfahren gegeben, wobei die Schrittweiten asymptotisch exakt von erster Ordnung und akzeptabel seien.
Dann gilt: (A_k) ist beschränkt und gleichmäßig positiv definit, und die Folge (x_k) konvergiert superlinear gegen die eindeutige Minimalstelle x_* von f.

Beweis. (Siehe [7], Satz 3.1.13; Seite 143–150.) □

Bemerkung 2.3.3. Für das BFGS-Verfahren konvergiert die Folge der (A_k) nicht notwendig gegen $F(x_*)$.

2.4 Minimierung von Quadratsummen

Eine wichtige Klasse nichtlinearer Optimierungsprobleme ohne Restriktionen werden durch die **Ausgleichsprobleme** gegeben. Ein Ausgleichsproblem hat die spezielle Form

$$\min \left\{ f(x) \;\middle|\; x \in \mathbf{R}^n \right\} \text{ mit } f(x) = \frac{1}{2} h(x)^T h(x), \tag{2.14}$$

wobei $h \colon \mathbf{R}^n \to \mathbf{R}^m$ zweimal stetig differenzierbar ist. Bei Ausgleichsproblemen gilt $m \geqslant n$, meist sogar $m \gg n$.

Beispiel 2.4.1. Die optimale Lösung eines überbestimmten Gleichungssystems ist ein Ausgleichsproblem: Statt $h(x) = 0$ löse $\min \left\{ \|h(x)\|^2 \;\middle|\; x \in \mathbf{R}^n \right\}$.
Ausgleichsprobleme treten auch bei der Regressionsrechnung auf: Sei $h_i(x) := g(x, t_i) - y_i$, etwa mit

$$g(x,t) := x_1 + \sum_{j=1}^{N} x_{2j} e^{x_{2j+1} t}.$$

Derartige Aufgaben verursachen große numerische Probleme, da die Exponentialfunktion in hohem Maße nichtlinear ist.

Wir führen folgende Bezeichnungen ein:

$$\begin{aligned} J &:= \nabla h = \left(\frac{\partial h_i}{\partial x_j}\right)_{ij} \\ H &:= (H_l) = \left(\left(\frac{\partial^2 h_l}{\partial x_i \partial x_j}\right)_{ij}\right)_l \end{aligned}$$

Damit gilt $\nabla f(x) = h(x)^T J(x)$ und $F(x) = J(x)^T J(x) + h(x)^T H(x)$.

Bemerkung 2.4.1. Die Dimension n ist in den Anwendungen meist relativ klein, aber $\mathrm{cond}_{\|.\|} F(x_*) \gg 1$. Deshalb ist der Aufwand zur Lösung des Systems $-F(x_k)d_k = \tilde{b}$ zur Bestimmung einer neuen Suchrichtung zwar relativ gering, aber wir haben aufgrund numerischer Instabilitäten sehr schlechte Konvergenz.

2.4.1 Lineare Ausgleichsrechnung

Definition 2.4.1. Ein Ausgleichsproblem der Form $h(x) := Ax - b$ mit einer $(m \times n)$-Matrix A und einem Vektor $b \in \mathbf{R}^m$ ist ein **lineares Ausgleichsproblem**.

Für lineare Ausgleichsprobleme gilt

$$J(x) = A,\ \nabla f(x) = (Ax - b)^T A \text{ und } F(x) = A^T A.$$

Ist $\mathrm{rang} A = n$, so ist $A^T A$ positiv definit. Dann ist $\nabla f(x_*) = 0$ äquivalent zu

$$A^T A x_* = A^T b. \tag{2.15}$$

Das System (2.15) heißt **Gaußsches Normalgleichungssystem**. Es ist numerisch ungünstig, dieses System direkt zu lösen, weil A meistens schlecht konditioniert ist, und weil deshalb wegen

$$\mathrm{cond}_{\|.\|}(A^T A) = (\mathrm{cond}_{\|.\|} A)^2$$

die Konditionszahl von (2.15) sogar extrem schlecht ist. Wir beschreiben deshalb einen anderen Ansatz zur Lösung von (2.14) mit linearem h.

Lösung linearer Ausgleichsprobleme durch QR-Zerlegung von A

Wir bestimmen eine Zerlegung $QA = \begin{pmatrix} R \\ 0 \end{pmatrix}$, wobei Q orthogonale $(m \times m)$-Matrix ist (also $Q^T Q = I$). R ist eine obere Dreiecksmatrix der Größe $n \times n$. Nun bestimmen wir $Qb := \begin{pmatrix} r \\ s \end{pmatrix}$ mit $r \in \mathbf{R}^n$, $s \in \mathbf{R}^{m-n}$. Wir haben hiermit eine alternative Darstellung von f gewonnen: Bei Verwendung der Frobeniusnorm gilt

$$\begin{aligned} f(x) &= \frac{1}{2}\|Ax - b\|^2 = \frac{1}{2}\|Q(Ax - b)\|^2 \\ &= \frac{1}{2}(\|Rx - r\|^2 + \|s\|^2). \end{aligned}$$

Das Minimum wird offensichtlich bei $Rx_* = r$ angenommen, denn dann ist $f(x_*) = \frac{1}{2}\|s\|^2$. x_* kann in diesem Fall besser bestimmt werden, weil in der Regel $\mathrm{cond}_{\|.\|} R = \mathrm{cond}_{\|.\|} A$ gilt. Dieses Verfahren ist besonders stabil, falls $\frac{\|s\|}{\|r\|} \ll 1$ (siehe Literaturhinweise in [7]).

2.4.2 Das Gauß-Newton-Verfahren

Das Gauß-Newton-Verfahren bestimmt einen neuen Iterationspunkt x_{k+1} zur Minimierung von f aus der quadratischen Approximation von f an der Stelle x_k, d. h. h wird linear approximiert: $h(x_k + \lambda d_k) \approx h(x_k) + \lambda J(x_k) d_k$. Dabei wird die Suchrichtung d_k durch

$$d_k := \operatorname{argmin}\left\{ \|h(x_k) + J(x_k)d\|^2 \;\middle|\; d \in \mathbf{R}^n \right\}$$

bestimmt.

Falls $\operatorname{rang} J(x_k) = n$ gilt, so läßt sich d_k aus einer linearen Ausgleichsaufgabe mit dem Normalgleichungssystem

$$J(x_k)^T J(x_k) d_k = -\underbrace{J(x_k)^T h(x_k)}_{=\nabla f(x_k)^T} \tag{2.16}$$

bestimmen. Dies führt auf das Minimum $x_* = x_k + d_k$ der quadratischen Approximation von f.

Definition 2.4.2. Ein Abstiegsverfahren zur Lösung eines Ausgleichsproblems (2.14), bei dem die Suchrichtung d_k durch die Lösung des linearen Ausgleichsproblems (2.16) bestimmt wird, heißt **Gauß-Newton-Verfahren**, wenn die Schrittweite $\lambda_k = 1$ benutzt wird, und **gedämpftes Gauß-Newton-Verfahren**, wenn die Schrittweite variabel ist (etwa mit Hilfe der Goldstein-Armijo-Suche bestimmt wird).

Bemerkung 2.4.2. Das Gauß-Newton-Verfahren erfordert die numerisch stabile Berechnung der Richtung d_k, also **nicht** durch Berechnung von $d_k = -(J(x_k)^T J(x_k))^{-1} \nabla f(x_k)^T$. Wird die Schrittweite konstant 1 gewählt (ungedämpftes Gauß-Newton-Verfahren), so ist die Konvergenz des Verfahrens nicht gewährleistet.

Das Konvergenzverhalten des Gauß-Newton-Verfahrens

Wir setzen im folgenden voraus, daß $F(x_*)$ positiv definit ist. Unter den weiteren üblichen Voraussetzungen wird das Konvergenzverhalten dann nach Satz 2.1.7 durch die positiven maximalen und minimalen Eigenwerte μ_1 und μ_n von

$$A(x)^{-\frac{1}{2}} F(x) A(x)^{-\frac{1}{2}} = I + \underbrace{A(x)^{-\frac{1}{2}} h(x)^T H(x) A(x)^{-\frac{1}{2}}}_{=:C(x)}$$

für $x \to x_*$ bestimmt, wobei $A(x) := J(x)^T J(x)$. Hieraus lassen sich die folgenden Aussagen ableiten:

1. Gilt $h(x_*) = 0$, so folgt $C(x_*) = 0$, also $\mu_1 = \mu_n = 1$, d. h. wir erhalten lokal quadratische Konvergenz wie beim Newton-Verfahren.

2. Anderenfalls ist die Konvergenz mindestens linear. Wir setzen

$$\lambda_1 := \lambda_{\min}\left(A(x_*)^{\frac{1}{2}} C(x_*) A(x_*)^{\frac{1}{2}} \right) \text{ und } \lambda_n := \lambda_{\max}\left(A(x_*)^{\frac{1}{2}} C(x_*) A(x_*)^{\frac{1}{2}} \right).$$

Damit gilt $\mu_1 = 1 + \lambda_1$ und $\mu_n = 1 + \mu_n$. Nach Satz 2.1.7 ist der Konvergenzfaktor für $f(x_k) - f(x_*)$

$$\left(\frac{\mu_n - \mu_1}{\mu_n + \mu_1}\right)^2 = \left(\frac{\lambda_n - \lambda_n}{2 + \lambda_1 + \lambda_n}\right)^2 < 1.$$

Weil $F(x_*)$ positiv definit ist, bekommen wir analog zu Bemerkung 2.1.8 lineare Konvergenz der Folge (x_k) gegen x_* mit dem Faktor $\frac{\lambda_n - \lambda_n}{2+\lambda_1+\lambda_n}$.
Je kleiner das Residuum $\|h(x_*)\|$, desto besser ist die Konvergenz.

Abschätzung der optimalen Schrittweite

Laut Bemerkung 2.1.9 ist die optimale Schrittweite für x nahe x_* von der Form

$$\bar{\lambda} = \frac{-\nabla f(x)d}{d^T F(x)d}\Big(1 + O(\|\nabla f(x)\|)\Big).$$

Wegen $A(x)d = -\nabla f(x)^T$ und $\nabla f(x) = h(x)^T J(x)$ folgt dann

$$\bar{\lambda} = \frac{d^T A(x)d}{d^T(A(x) + h(x)^T H(x))d}\Big(1 + O(\|h(x)\|)\Big) = 1 + O(\|h(x)\|).$$

Bestimmung einer günstigen Schrittweite

Die optimale Schrittweite weicht bei ungenauem Startvektor x_0 bzw. großem Residuum $h(x_0)$ weit von 1 ab. Notwendig ist daher eine gute lokale Approximation der optimalen Schrittweite. Die Forderung einer akzeptablen Schrittweite liefert nach Definition 2.1.1 wegen

$$\lambda(-\alpha\nabla f(x)d) \leqslant f(x) - \underbrace{f(x + \lambda d)}_{\geqslant 0} \leqslant f(x)$$

die obere Schranke

$$\lambda \leqslant \frac{f(x)}{-\alpha\nabla f(x)d} =: \lambda_0.$$

Wir führen nun eine quadratische Interpolation der Komponenten h_i durch:

$$q_i(\lambda) \approx h_i(x + \lambda d).$$

Zur Verfügung stehen für das Suchintervall $[a, b]$ (anfangs $[0, \lambda_0]$) die Daten $h_i(a)$, $h_i(b)$ und $h_i'(a)$. Daraus ergibt sich die Darstellung

$$q_i(\lambda) = h_i(a) + (\lambda - a)h_i'(a) + \frac{(\lambda - a)^2}{b - a}\left(\frac{h_i(b) - h_i(a)}{b - a} - h_i'(a)\right)$$

für q_i. Mit Hilfe der q_i approximieren wir f nun in Richtung d quadratisch durch

$$f(x + \lambda d) \approx Q(\lambda) := \frac{1}{2}q(\lambda)^T q(\lambda).$$

Die optimale Schrittweite der Approximation läßt sich nun durch Bestimmung der Nullstellen eines Polynoms dritten Grades aus

$$\lambda_* := \operatorname{argmin}\Big\{Q(\lambda) \;\Big|\; a \leqslant \lambda \leqslant b\Big\}$$

bestimmen.

2.4.3 Das Verfahren von Levenberg, Marquardt und Moré

Ein Nachteil des Gauß-Newton-Verfahrens ist, daß die Bestimmung der Suchrichtung d_k problematisch wird, wenn $J(x_k)$ nicht maximalen Rang hat, weil $J(x_k)^T J(x_k)$ dann nicht positiv definit ist. Levenberg (1944) und Marquardt (1963) schlugen deshalb die Anwendung der Regularisierungstechnik

$$A_k := J(x_k)^T J(x_k) + \mu_k I \text{ mit } \mu_k \geqslant 0$$

vor, die schon in Abschnitt 2.1.5 behandelt wurde. Das Verfahren wurde von Moré verfeinert (1977, 1983), der die Einheitsmatrix durch eine adaptiv bestimmte positiv definite Matrix ersetzte. Der Ansatz von Moré lautet

$$A_k := J(x_k)^T J(x_k) + \mu_k \operatorname{diag}(\left(\delta_j^{(k)}\right)^2) \text{ mit } \mu_k \geqslant 0,$$

wobei

$$\delta_j^{(k)} := \max\left\{\delta_j^{(k-1)}, \|J(x_k)e_j\|\right\} \text{ und } \delta_j^{(-1)} := 1 \tag{2.17}$$

zu setzen ist.

Lemma 2.4.1. Es sei $d(\mu) := -\left(J(x)^T J(x) + \mu I\right)^{-1} \nabla f(x)^T$. Dann gilt

1. $\|d(\mu)\|$ als Funktion von μ ist monoton fallend.
2. $\lim\limits_{\mu\to\infty} \|d(\mu)\| = 0$

Beweis. Da $J^T(x)J(x)$ positiv semidefinit ist, gilt $J(x)^T J(x) = S^T D S$ mit $D = \operatorname{diag}(\sigma_j)$, $\sigma_j \geqslant 0$ und S orthonormal. Es folgt

$$B := J(x)^T J(x) + \mu I = S^T \underbrace{(D + \mu I)}_{=:\tilde{D}} S$$

und somit $B^z = S^T \tilde{D}^z S$ für alle $z \in \mathbf{Z}$.
Wegen $Bd(\mu) = -\nabla f(x)^T$ gilt $d(\mu) = -B^{-1}\nabla f(x)^T$, also

$$\|d(\mu)\|^2 = \nabla f(x) B^{-2} \nabla f(x)^T = (S\nabla f(x)^T)^T \tilde{D}^{-2} (S\nabla f(x)^T).$$

Setzen wir nun $v := S\nabla f(x)^T$, so erhalten wir

$$\|d(\mu)\|^2 = \sum_{j=1}^{n} \frac{v_j^2}{(\sigma_j + \mu)^2},$$

was für $\mu \to \infty$ monoton gegen 0 konvergiert. □

Bemerkung 2.4.3. Ist μ_k groß genug, so ist A_k positiv definit. Außerdem wird ein Abstieg erzwungen, wenn $\|d_k\|$ klein genug ist.

Lemma 2.4.2. Für hinreichend große, aber beschränkte μ ist die Suchrichtung

$$d := -\Big(J(x)^T J(x) + \mu\Delta^2\Big)^{-1} \nabla f(x)^T \text{ mit } \Delta = \operatorname{diag}(\delta_j)$$

streng gradientenbezogen und die Schrittweite $\lambda = 1$ akzeptabel.

Beweis. Es sei $\mu > 0$. Die Taylorentwicklung von f an der Stelle x ergibt

$$f(x+d) - f(x) = \nabla f(x) d + \frac{1}{2} d^T F(x + \theta d) d \text{ für ein } \theta \in (0,1).$$

Sei $\|F\| \leqslant M$. Dann folgt

$$f(x+\lambda d) - f(x) \leqslant \nabla f(x) d + \frac{M}{2}\|d\|^2. \tag{2.18}$$

Nun gilt wegen $d = -\Big(J(x)^T J(x) + \mu\Delta^2\Big)^{-1} \nabla f(x)^T$ für $\mu > 0$

$$d = -\frac{1}{\mu}\Delta^{-1}\Big(\underbrace{\frac{1}{\mu}\Delta^{-1} J(x)^T J(x) \Delta^{-1}}_{\text{positiv semidefinit}} + I\Big)^{-1} \Delta^{-1} \nabla f(x)^T. \tag{2.19}$$

Alle Eigenwerte der zu invertierenden Matrix in (2.19) sind mindestens 1, also hat die Inverse eine Norm von höchstens 1. Da auch $\|\Delta^{-1}\| \leqslant 1$, folgt die Abschätzung $\|d\| \leqslant \frac{1}{\mu}\|\Delta^{-1}\nabla f(x)^T\|$. Wir multiplizieren nun (2.19) von links mit $-\nabla f(x)$ und erhalten

$$\frac{1}{\mu}\Big(1 - \frac{\|J(x)\Delta^{-1}\|^2}{\mu}\Big)\|\Delta^{-1}\nabla f(x)^T\|^2 \leqslant -\nabla f(x) d \leqslant \frac{1}{\mu}\|\Delta^{-1}\nabla f(x)^T\|^2,$$

woraus wegen $\|J(x)\Delta^{-1}\| \leqslant \sqrt{n}$ (nach Konstruktion von Δ)

$$\frac{1}{\mu}\Big(1 - \frac{n}{\mu}\Big)\|\Delta^{-1}\nabla f(x)^T\|^2 \leqslant -\nabla f(x) d \leqslant \frac{1}{\mu}\|\Delta^{-1}\nabla f(x)^T\|^2 \tag{2.20}$$

folgt. Mit (2.18) ergibt sich damit

$$\begin{aligned} f(x+d) - f(x) &\leqslant -\frac{1}{\mu}\Big(1 - \frac{n}{\mu}\Big)\|\Delta^{-1}\nabla f(x)^T\|^2 + \frac{M}{2\mu^2}\|\Delta^{-1}\nabla f(x)^T\|^2 \\ &= -\frac{1}{\mu}\Big(1 - \frac{n + \frac{M}{2}}{\mu}\Big)\|\Delta^{-1}\nabla f(x)^T\|^2. \end{aligned}$$

Sei nun $0 < \alpha < 1$ und $\mu \geqslant \frac{n+\frac{M}{2}}{1-\alpha}$. Dann folgt aus der rechten Ungleichung von (2.20)

$$f(x+d) - f(x) \leqslant -\frac{1}{\mu}\alpha\|\Delta^{-1}\nabla f(x)^T\|^2 \leqslant \alpha \nabla f(x) d,$$

d. h. $\lambda = 1$ ist akzeptabel für diese μ. □

Bemerkung 2.4.4. Im allgemeinen ist die im Beweis von Lemma 2.4.2 gefundene Schranke $\mu \geqslant \frac{n+\frac{M}{2}}{1-\alpha}$ zu groß für eine schnelle Konvergenz. Es ist günstiger, μ flexibel anzupassen, um gegebenenfalls größere Schrittweiten ausnutzen zu können.

Vertrauensbereich der Approximation

Die vorgenommene Näherung

$$h(x+d) \approx h(x) + J(x)d$$

ist nur dann hinreichend gut, wenn d klein bleibt, z. B. $\|\Delta d\| \leqslant \varrho$ für eine genügend kleine Konstante $\varrho > 0$. Wir wollen die Suchrichtung d aus

$$\min \frac{1}{2}\|h(x) + J(x)d\|^2$$

bestimmen. Die Lösung dieses Problems ergibt sich aus der Multiplikatorregel von Lagrange. Wir setzen $d_N := d(0)$, d. h. $J(x)^T d_N = -J(x)^T h(x)$. Gilt nun $\|\Delta d_N\| \leqslant \varrho$, so verwende $d := d_N$ als Suchrichtung. Anderenfalls berechne $d_R := d(\mu_R)$, welches gerade $\|\Delta d_R\| = \varrho$ erfüllt. Die Bestimmung von μ_R erfolgt durch Lösung eines passenden Nullstellenproblems. Weil dies die Berechnung etlicher $d(\mu)$ erfordert, ist die Verwendung einer Updatetechnik für die QR-Zerlegung der Matrix $\begin{pmatrix} J(x) \\ \sqrt{\mu}D \end{pmatrix}$ wesentlich für die Einsparung von Rechenzeit, insbesondere wenn $m \gg n$. Auf diese Updatetechnik werden wir später eingehen.

Bestimmung von ϱ

Die lineare Näherung von h gilt nur, wenn

$$f(x+d) \approx \frac{1}{2}\|h(x) + J(x)d\|^2.$$

Dies ist äquivalent zu

$$\kappa := \frac{\|h(x)\|^2 - \|h(x+d)\|^2}{\|h(x)\|^2 - \|h(x) + J(x)d\|^2} \approx 1, \tag{2.21}$$

d. h. κ kann als Gütekriterium verwendet werden. Die Auswertung der Formel (2.21) ist allerdings problematisch, da im Nenner etwa gleich große Zahlen subtrahiert werden, d. h. es kann zur Auslöschung und damit zu großen Rundungsfehlern kommen. Wir werden deshalb zunächst eine günstigere Darstellung für κ herleiten. Mit der Beziehung

$$(J(x)^T J(x) + \mu\Delta^2)d = -J(x)^T h(x)$$

folgt für den Nenner von (2.21) die Darstellung

$$\begin{aligned}
\|h(x)\|^2 - \|h(x) + J(x)d\|^2 &= -2d^T J(x)^T h(x) - d^T J(x)^T J(x)d \\
&= 2d^T \Big(J(x)^T J(x) + \mu\Delta^2\Big) d - d^T J(x)^T J(x)d \\
&= d^T J(x)^T J(x)d + 2\mu d^T \Delta^2 d.
\end{aligned}$$

Also erhalten wir κ durch

$$\kappa = \frac{1 - \left(\frac{\|h(x+d)\|}{\|h(x)\|}\right)^2}{\left(\frac{\|J(x)d\|}{\|h(x)\|}\right)^2 + 2\left(\frac{\sqrt{\mu}\|\Delta d\|}{\|h(x)\|}\right)^2}.$$

Die genaue Form des Updates von ϱ in Abhängigkeit von κ ist im folgenden Algorithmus angegeben.

Algorithmus 2.4.1. (Verfahren von Levenberg, Marquardt und Moré)
Gegeben seien die Verfahrensparameter $\theta \in (0,1)$ und $0 < \varepsilon \ll \frac{1}{4}$, sowie ein Startvektor $x_0 \in \mathbf{R}^n$.
Setze $k := 0$. Setze $\varrho_0 := \|x_0\| + \theta$.
Für $1 \leqslant i \leqslant n$
    setze $\delta_i^{(0)} := \max\{1, \|J(x_0)e_i\|\}$.
Für $k = 0, 1, \ldots$

1. Falls $\|d(0)\| \leqslant (1+\theta)\varrho_k$
        setze $\mu_k := 0$, $d_k := d(0)$,
    anderenfalls
        bestimme $\mu_k > 0$ mit $(1-\theta)\varrho_k \leqslant \|d(\mu_k)\| \leqslant (1+\theta)\varrho_k$,
        setze $d_k := d(\mu_k)$.
2. Berechne κ.
3. Falls $\kappa < \varepsilon$
        setze $x_{k+1} := x_k$,
    anderenfalls
        setze $x_{k+1} := x_k + d_k$,
        für $1 \leqslant i \leqslant n$
            setze $\delta_i := \max\{\delta_i, \|J(x_k)e_i\|\}$.
4. Falls $\kappa < \frac{1}{4}$
        wähle $\varrho_{k+1} \in [\frac{1}{10}\varrho_k, \frac{1}{2}\varrho_k]$.
    Falls $\kappa \in [\frac{1}{4}, \frac{3}{4}]$ und $\mu_k > 0$
        setze $\varrho_{k+1} := \varrho_k$.
    Falls ($\kappa \in [\frac{1}{4}, \frac{3}{4}]$ und $\mu_k = 0$) oder $\kappa > \frac{3}{4}$
        setze $\varrho_{k+1} := \min\{2\varrho_k, \varrho_0\}$.

Im Fall $\kappa < \frac{1}{4}$ (Schritt 4 in Algorithmus 2.4.1), wird ϱ_{k+1} mit Hilfe einer Interpolationsmethode gewählt. Die Interpolation liefert eine Schätzung des lokalen Minimums von f in Richtung d_k:

$$\varphi(\lambda) := f(x_k + \lambda d_k)$$

Die quadratische Interpolation zu $\varphi(0)$, $\varphi'(0)$ und $\varphi(1)$ hat das Minimum

$$\begin{aligned}
\lambda_{\min} &= \frac{1}{2}\frac{-\nabla f(x_k)d}{f(x_k+d_k) - f(x_k) - \nabla f(x_k)d_k} \\
&= \frac{1}{2}\frac{-d_k^T J(x_k)^T h(x_k)}{\frac{1}{2}\|h(x_k+d_k)\|^2 - \frac{1}{2}\|h(x_k)\|^2 - d_k^T J(x_k)^T h(x_k)} \\
&= \frac{\frac{1}{2}\gamma}{\gamma + \frac{1}{2}\left(\left(\frac{\|h(x_k+d_k)\|}{\|h(x_k)\|}\right)^2 - 1\right)} \quad \text{mit } \gamma := \frac{-\nabla f(x_k)d_k}{\|h(x_k)\|^2} > 0.
\end{aligned}$$

Setze nun

$$\varrho_{k+1} := \begin{cases} \max\left\{\lambda_{\min}, \frac{1}{10}\right\}\varrho_k & \text{falls } \|h(x_k+d_k)\| > \|h(x_k)\|, \\ \frac{1}{2}\varrho_k & \text{sonst.} \end{cases}$$

Im ersten Fall gilt $\lambda_{\min} \leqslant \frac{1}{2}$, deshalb ist $\varrho_{k+1} \in [\frac{1}{10}, \frac{1}{2}]\varrho_k$ erfüllt.

Der folgende Satz liefert eine Konvergenzaussage zum Verfahren von Levenberg, Marquardt und Moré:

Satz 2.4.3. Sei $N(x_0)$ kompakt, $f \in C_3(N(x_0))$. Dann gilt für die nach Algorithmus 2.4.1 berechnete Folge (x_k)

$$\nabla f(x_k) \to 0.$$

Beweis. (Siehe [7]; Beweis zu Satz 3.2.21, S. 207–209. Der Beweis ist sehr technisch.) □

Bemerkung 2.4.5. Erfüllt ein Häufungspunkt x_* von (x_k) die Bedingung $\operatorname{rang} J(x_*) = n$ und ist $F(x_*)$ positiv definit, so gilt $(x_k) \to x_*$.

Eine Updatetechnik zur Bestimmung der $d(\mu)$

Die Suchrichtung $d(\mu)$ wird durch das lineare Gleichungssystem

$$(J(x)^T J(x) + \mu\Delta^2)d(\mu) = -J(x)^T h(x)$$

definiert. Zur Lösung wird analog zu Abschnitt 2.4.1 eine QR-Zerlegung von $A = \begin{pmatrix} J(x) \\ \sqrt{\mu}\Delta \end{pmatrix}$ bestimmt. Zunächst wird die QR-Zerlegung von $J(x)$

$$QJ(x)P = \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix}$$

mit geeigneter Permutationsmatrix P berechnet. R ist eine obere Dreieckmatrix, S ist voll besetzt. Zu $J(x)$ ist

$$J(x)^- := P \begin{pmatrix} R^{-1} & 0 \\ 0 & 0 \end{pmatrix} Q$$

eine verallgemeinerte Inverse. $J(x)^-$ hat die folgende Eigenschaft:

$$\begin{aligned} J(x)J(x)^- &= Q^T \begin{pmatrix} RS \\ 0 \end{pmatrix} P^T P \begin{pmatrix} R^{-1} & 0 \\ 0 & 0 \end{pmatrix} Q \\ &= Q^T \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} Q \\ &= \left(J(x)J(x)^-\right)^T. \end{aligned}$$

Weiterhin gilt

$$J(x)J(x)^- J(x) = Q^T \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} QQ^T \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} P^T = Q^T \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} P^T = J(x)$$

und analog zeigt man, daß auch $J(x)^- J(x) J(x)^- = J(x)^-$ gilt.
Für die Gesamtmatrix A läßt sich die Zerlegung wie folgt fortsetzen:

$$\underbrace{\begin{pmatrix} Q & 0 \\ 0 & P^T \end{pmatrix}}_{=:\tilde{Q}} \begin{pmatrix} J(x) \\ \sqrt{\mu}\Delta \end{pmatrix} P = \tilde{Q}AP = \begin{pmatrix} R & S \\ 0 & 0 \\ \multicolumn{2}{c}{(\sqrt{\mu}P^T\Delta P)} \end{pmatrix}.$$

Weil $P^T \Delta P$ wieder eine Diagonalmatrix ist, hat $\tilde{Q}AP$ die folgende Form:

$$\tilde{Q}AP = \begin{pmatrix} \times & \times & \cdots & \times & \times \\ & \times & \cdots & \times & \times \\ & & \ddots & \times & \times \\ & & & \times & \times \\ & & & & \\ \times & & & & \\ & \times & & & \\ & & \times & & \\ & & & \times & \\ & & & & \times \end{pmatrix}.$$

Mit Givens-Rotationen erzeugt man nun oberhalb der Diagonalen neue Nichtnullelemente an den Positionen $(m+j,k)$ für $j = 1,\ldots,n$ und $k \geqslant j+1$. Iterativ wird die Diagonale nach oben verschoben, bis die gewünschte Form erreicht ist. Dies liefert sukzessive die Lösung der Aufgabe $\min \|Ad + (h(x)^T, 0)^T\|^2$.

Für ein neues μ muß nur jeweils ein Update der Zerlegung durchgeführt werden. Dazu sind die Daten R, S und $b := \tilde{Q}\begin{pmatrix} h(x) \\ 0 \end{pmatrix}$ zu speichern.

Kapitel 3

Nichtlineare Optimierung mit Restriktionen

3.1 Optimalitätsbedingungen

Für nichtlineare Optimierungsprobleme mit Restriktionen müssen wir zunächst Optimalitätsbedingungen herleiten. Wir beschäftigen uns in diesem Kapitel mit Problemen in folgender Form: Gegeben seien $D \subseteq \mathbf{R}^n$ offen, $f \in C_2(D)$, $g \in C_1(D, \mathbf{R}^m)$ und $h \in C_1(D, \mathbf{R}^p))$. Gesucht ist eine Lösung der Aufgabe

$$x_* = \operatorname{argmin}\Big\{ f(x) \;\Big|\; \underbrace{g(x) \leqslant 0,\ h(x) = 0,\ x \in D}_{P \neq \emptyset} \Big\}. \tag{3.1}$$

3.1.1 Tangentialkegel und die Guignard-Bedingung

Satz 3.1.1. Es sei x_* lokales Minimum der Aufgabe (3.1), (x_k) eine gegen x_* konvergente Folge mit $x_k \neq x_*$ für alle $k \in \mathbf{N}$ und z ein Häufungspunkt der Folge $\left(\frac{x_k - x_*}{\|x_k - x_*\|}\right)$.
Dann gilt $\nabla f(x_*) z \geqslant 0$.

Beweis. Es sei $\lambda_k := \|x_k - x_*\|$. Nach Voraussetzung gilt $(\lambda_k) \to 0$. Mit $z_k := \frac{1}{\lambda_k}(x_k - x_*)$ gilt $x_k = x_* + \lambda_k z_k$. Da $\|z_k\| = 1$, existiert eine gegen z konvergente Teilfolge $(z_{k'})$. O. B. d. A. sei $(z_{k'}) = (z_k)$. Es ist nach Taylorentwicklung

$$f(x_k) = f(x_*) + \lambda_k \nabla f(x_*) z_k + \frac{1}{2}\lambda_k^2 z_k^T F(x_* + \theta \lambda_k z_k) z_k \text{ für ein } \theta \in (0,1).$$

Weil x_* lokales Minimum ist, gilt $f(x_k) \geqslant f(x_*)$ in einer genügend kleinen Umgebung von x_*, also $0 \leqslant \nabla f(x_*) z_k + O(\lambda_k)$. Der Grenzübergang $x_k \to x_*$ liefert nun $\nabla f(x_*) z \geqslant 0$. □

Definition 3.1.1. Es sei $T \subseteq \mathbf{R}^n$ und $0 \in T$.
Dann heißt T **Kegel**, wenn für alle $x \in T$, $x \neq 0$ gilt

$$x \in T \Rightarrow \tau x \in T \text{ für alle } \tau > 0.$$

Es sei P eine Menge wie in (3.1) und $x \in P$. Es sei

$$T_P(x,\delta) := \Big\{\lambda(y-x) \;\Big|\; \lambda \geqslant 0,\ y \in P,\ \|y-x\| < \delta\Big\}.$$

Dann ist der **abgeschlossene Tangentialkegel** von P in x durch

$$T_P(x) := \bigcap_{k=1}^{\infty} \overline{T_P\Big(x, \frac{1}{k}\Big)}$$

definiert.

$T_P(x)$ enthält alle gemeinsamen Häufungspunkte der $T_P(x, \frac{1}{k})$.

Satz 3.1.2. Es gilt $z \in T_P(x)$ genau dann, wenn es eine Folge positiver Zahlen (λ_k) und eine gegen x konvergente Folge $(x_k) \subseteq P$ gibt mit $z = \lim\limits_{k\to\infty} (\lambda_k(x_k - x))$.

Beweis. Sei $z \in T_P(x)$. Dann ist z Häufungspunkt einer Folge $\lambda_k(x_k - x)$ mit $\lambda_k > 0$ und $\|x_k - x\| \leqslant \frac{1}{k}$. Wir betrachten O. B. d. A. eine konvergente Teilfolge. Für diese gilt $(x_k) \to x$ und $(\lambda_k(x_k - x)) \to z$.
Sei umgekehrt $(x_k) \to x$, $(\lambda_k(x_k - x)) \to z$ mit $\lambda_k > 0$ und $(x_k) \subseteq P$. Sei $p \in \mathbf{N}$. Für hinreichend große k, etwa $k \geqslant k_0(p)$, gilt $\|x_k - x\| < \frac{1}{p}$, also $\lambda_k(x_k - x) \in \overline{T_P(x, \frac{1}{p})}$. Da p beliebig gewählt werden kann, folgt $z \in T_P(x)$. ☐

Definition 3.1.2. Es sei $S \subseteq \mathbf{R}^n$. Dann heißt die Menge

$$S' := \Big\{y \in \mathbf{R}^n \;\Big|\; y^T x \geqslant 0 \text{ für alle } x \in S\Big\}$$

positiver Normalkegel von S.

Bemerkung 3.1.1. Es sei x_* lokales Minimum von (3.1). Dann gilt $\nabla f(x_*)z \geqslant 0$ für alle $z \in T_P(x)$. Also folgt $\nabla f(x_*)^T \in T_P(x_*)'$.

Wir möchten nun eine handliche Darstellung der Mengen $T_P(x_*)$ bzw. $T_P(x_*)'$ bekommen.

Definition 3.1.3. Die Menge

$$\mathcal{A}(x) := \Big\{i \;\Big|\; g_i(x) = 0\Big\}$$

heißt **Indexmenge der in x aktiven Restriktionen**. Der Kegel

$$L(x) := \Big\{z \;\Big|\; \nabla h(x)z = 0,\ \nabla g_i(x)z \leqslant 0 \text{ für alle } i \in \mathcal{A}(x)\Big\}$$

heißt **linearisierter Kegel zu h und g in x.**

Satz 3.1.3. Es sei $x \in P$. Dann gilt $T_P(x) \subseteq L(x)$.

Beweis. Es sei $z \in T_P(x)$. Also gibt es eine Folge $(\lambda_k) \subseteq \mathbf{R}_>$ und $(x_k) \subseteq P$ mit $(x_k) \to x$ und $(\lambda_k(x_k - x)) \to z$. Wir setzen $z_k := \lambda_k(x_k - x)$. Damit gilt für $i \in \{1, \ldots, p\}$

$$0 = h_i(x_k) = h_i(x) + \nabla h_i(x)(x_k - x) + o(\|x_k - x\|).$$

Multiplikation mit λ_k ergibt

$$0 = \nabla h_i(x) z_k + o(\|z_k\|)$$

Grenzübergang $k \to \infty$ liefert nun $\nabla h_i(x) z = 0$.
Sei $i \in \mathcal{A}(x)$. Dann gilt $0 \geqslant g_i(x_k) = g_i(x) + \nabla g_i(x)(x_k - x) + o(\|x_k - x\|)$. Wie zuvor zeigt man nun $\nabla g_i(x) z \leqslant 0$. Also gilt $z \in L(x)$. □

Bemerkung 3.1.2. Das folgenden Beispiel (Kuhn, Tucker) zeigt, daß der Fall $T_P(x) \subset L(x)$ eintreten kann. Es sei $g \colon \mathbf{R}^2 \to \mathbf{R}^3$ mit

$$\begin{aligned} g_1(x) = &\ x_2 - (1 - x_1)^3 && \leqslant 0 \\ g_2(x) = &\ -x_1 && \leqslant 0 \\ g_3(x) = &\ -x_2 && \leqslant 0. \end{aligned}$$

Für $x_* = (1, 0)^T$ gilt hier

$$T_P(x_*) = \left\{ (\alpha, 0)^T \;\middle|\; \alpha \leqslant 0 \right\} \neq L(x_*) = \left\{ (\alpha, 0)^T \;\middle|\; \alpha \in \mathbf{R} \right\}.$$

Satz 3.1.4. Es sei x_* lokales Minimum der Aufgabe (3.1). Falls außerdem die **Guignard-Bedingung** $L(x_*)' = T_P(x_*)'$ gilt, so existieren Vektoren $\lambda \in \mathbf{R}^m_{\geqslant}$ und $\mu \in \mathbf{R}^p$ mit

$$\nabla f(x_*) + \lambda^T \nabla g(x_*) + \mu^T \nabla h(x_*) = 0 \tag{3.2}$$

$$\lambda^T g(x_*) = 0. \tag{3.3}$$

Beweis. Nach einer Umformulierung von Satz 3.1.1 gilt: $\nabla f(x_*)^T \in T_P(x_*)'$. Es gilt also $\nabla f(x_*)^T \in L(x_*)'$, d. h. $\nabla f(x_*) z \geqslant 0$ für alle $z \in L(x_*)$.
Eine äquivalente Aussage ist: Das Gleichungssystem

$$\nabla f(x_*) z < 0, \quad \nabla h(x_*) z = 0, \quad -\nabla g_i(x_*) z \geqslant 0 \text{ für alle } i \in \mathcal{A}(x_*)$$

ist unlösbar. Also ist nach dem Lemma von Farkas (siehe Skriptum zur Vorlesung „Einführung in die Mathematische Optimierung“) das System

$$\sum_{i \in \mathcal{A}(x_*)} u_i(-\nabla g_i(x_*)) + v^T \nabla h(x_*) = \nabla f(x_*),$$

$$u_i \geqslant 0 \text{ für alle } i \in \mathcal{A}(x_*)$$

lösbar. Mit $\lambda_i := \begin{cases} u_i & i \in \mathcal{A}(x_*) \\ 0 & i \notin \mathcal{A}(x_*) \end{cases}$ und $\mu := -v$ folgen (3.2) und (3.3). □

Bemerkung 3.1.3. Hat die Aufgabe (3.1) nur Gleichheitsrestriktionen, so ist (3.2) als **Lagrangesche Multiplikatorregel** bekannt. Die allgemeine Form wurde von Karush und von Kuhn und Tucker gefunden. Die Bedingungen (3.2) und (3.3) heißen deshalb **Kuhn-Tucker-Bedingungen** oder auch „notwendige Bedingungen erster Ordnung“. Ein Punkt x_*, der diese Bedingungen erfüllt, heißt **Kuhn-Tucker-Punkt**.

Bemerkung 3.1.4. Die Guignard-Bedingung ist abhängig von der Darstellung von P durch g und h. Derartige Bedingungen heißen daher **Restriktionsqualifikationen**. Das System aus Bemerkung 3.1.2 kann durch Hinzunahme einer weiteren Restriktion so erweitert werden, daß die Guignard-Bedingung gilt, ohne daß P sich ändert:

Durch Hinzufügen der redundanten Restriktion

$$g_4(x) = x_1 - 1 \leqslant 0$$

wird $T_P = L$ erreicht.

Bemerkung 3.1.5. Die Guignard-Bedingung ist in der Praxis sehr schwer nachprüfbar. Deshalb sucht man einfachere Bedingungen, die die Guignard-Bedingung implizieren. Von Gould und Tolle wurde die folgende Äquivalenz gezeigt:

Es sei $x_* \in P$. Dann gilt die Guignard-Bedingung genau dann in x_*, wenn für jede Funktion $f \in C_1(D)$, die in x_* ein lokales Minimum besitzt, (3.2) in x_* erfüllbar ist.

Für eine gegebene Funktion f ist die Bedingung (3.2) zwar notwendig, aber nicht hinreichend für die Charakterisierung von x_* als Minimum.

Der folgende Satz gibt nachprüfbare Bedingungen an, die die Guignard-Bedingung implizieren.

Satz 3.1.5. (Stärkere Restriktionsbedingungen) Es sei $x_* \in P$. Dann impliziert jede der folgenden Bedingungen die Guignard-Bedingung:

1. (Abadie) Es gilt $L(x_*) = T_P(x_*)$.

2. (Mangasarian-Fromowitz) Es existiert ein $z \in \mathbf{R}^n$ mit

$$\begin{aligned}
&\nabla g_i(x_*)z < 0 \text{ für alle } i \in \mathcal{A}(x_*),\\
&\nabla h(x_*)z = 0,\\
&\nabla h(x_*) \text{ hat maximalen Rang.}
\end{aligned}$$

3. (Regularitätsbedingung) Die Matrix $\begin{pmatrix} \nabla h(x_*) \\ \nabla g_{\mathcal{A}(x_*)}(x_*) \end{pmatrix}$ hat maximalen Rang.

Beweis. Wir beweisen jeden Teil separat.

1. Die Behauptung gilt, denn $L(x_*) = T_P(x_*)$ impliziert $L(x_*)' = T_P(x_*)'$.

2. Es genügt zu zeigen, daß $L(x_*) \subseteq T_P(x_*)$ gilt, denn hieraus folgt dann die Abadie-Bedingung.

   Es sei $a \in L(x_*)$ und $z \in \mathbf{R}^n$ erfülle die Mangasarian-Fromowitz-Bedingung. Gesucht sind $(\alpha_k) \subseteq \mathbf{R}_>$, $(x_k) \subseteq P$ mit $(x_k) \to x_*$ und $\alpha_k(x_k - x_*) \to a$, denn bei Nachweis der Existenz solcher Folgen folgt nach Satz 3.1.2 $a \in T_P(x_*)$.

   O. B. d. A. sei $\nabla h(x_*) = [A \vdots B]$, wobei A eine reguläre $p \times p$-Matrix ist.

Sei $x_* = \begin{pmatrix} u_* \\ v_* \end{pmatrix}$ mit $u_* \in \mathbf{R}^p$. Nach dem Hauptsatz über implizite Funktionen existiert eine stetig differenzierbare Funktion $u\colon U_1(v_*) \to U_2(u_*)$ mit

$$h(x) = 0 \text{ für } x \in U_2(u_*) \times U_1(v_*) \text{ genau dann, wenn } x = \begin{pmatrix} u(v) \\ v \end{pmatrix}$$

ist. Dabei gilt aufgrund der Kettenregel der Differentiation

$$\frac{\partial u}{\partial v} = -\left(\frac{\partial h}{\partial u}(x)\right)^{-1} \frac{\partial h}{\partial v}(x) \text{ für } x = \begin{pmatrix} u(v) \\ v \end{pmatrix},\ v \in U_1(v_*).$$

Wir setzen nun

$$a = \begin{pmatrix} a' \\ a'' \end{pmatrix} \text{ und } z = \begin{pmatrix} z' \\ z'' \end{pmatrix} \text{ mit } a', z'' \in \mathbf{R}^p.$$

Für positive Nullfolgen $(\tilde{\alpha}_k)$ und (ε_k) sei $v_k := v_* + \tilde{\alpha}_k(a'' + \varepsilon_k z'')$. O. B. d. A. sei $v_k \in U_1(v_*)$ und $\varepsilon_k \leqslant 1$ bereits für alle $k \geqslant 1$. Dann gilt mit $x_k := \begin{pmatrix} u(v_k) \\ v_k \end{pmatrix}$ sofort $(x_k) \to \begin{pmatrix} u_* \\ v_* \end{pmatrix} = x_*$.

Nun ist noch zu zeigen, daß $g(x_k) \leqslant 0$ und $\frac{1}{\tilde{\alpha}_k}(x_k - x_*) \to a$. Die Taylorentwicklung von u um v_* liefert:

$$\begin{aligned} u(v_k) - u_* = u(v_k) - u(v_*) &= \tilde{\alpha}_k \frac{\partial u}{\partial v}(v_*)(a'' + \varepsilon_k z'') + \underbrace{\tilde{\alpha}_k r(\tilde{\alpha}_k, \varepsilon_k)}_{=O(\tilde{\alpha}^2)} \\ &= \tilde{\alpha}_k \left(-\frac{\partial h}{\partial u}(x)\right)^{-1} \frac{\partial h}{\partial v}(x)(a'' + \varepsilon_k z'') + \tilde{\alpha}_k r(\tilde{\alpha}_k, \varepsilon_k) \end{aligned}$$

mit $r(\tilde{\alpha}, \varepsilon_k) \to 0$ wenn $(\tilde{\alpha}_k) \to 0$. Außerdem gilt nach Zerlegung von a und z

$$0 = \nabla h(x_*)(a + \varepsilon_k z) = \left[\frac{\partial h}{\partial u}(x_*) \vdots \frac{\partial h}{\partial v}(x_*)\right] \begin{pmatrix} a' + \varepsilon_k z' \\ a'' + \varepsilon_k z'' \end{pmatrix}, \tag{3.4}$$

also

$$-\frac{\partial h}{\partial u}(x_*)(a' + \varepsilon_k z') = \frac{\partial h}{\partial v}(a'' + \varepsilon_k z'').$$

Damit ergibt sich

$$u(v_k) - u_* = \tilde{\alpha}_k \underbrace{\left(-\frac{\partial h}{\partial u}(x_*)\right)^{-1} \left(-\frac{\partial h}{\partial u}(x_*)\right)}_{=I} (a' + \varepsilon_k z') + \tilde{\alpha}_k r(\tilde{\alpha}_k, \varepsilon_k).$$

Also gilt

$$\frac{1}{\tilde{\alpha}_k}(x_k - x_*) = \begin{pmatrix} a' + \varepsilon_k z' + r(\tilde{\alpha}_k, \varepsilon_k) \\ a'' + \varepsilon_k z'' \end{pmatrix}$$

und somit folgt die Konvergenz gegen a.

Es bleibt zu zeigen, daß $g(x_k) \leqslant 0$ für geeignete Wahl von ε_k gilt. Dies ist in einer hinreichend kleinen Umgebung von x_* trivial für $i \notin \mathcal{A}(x_*)$. Sei also $i \in \mathcal{A}(x_*)$. Dann gilt

$$\begin{aligned}
g_i(x_k) &= g_i\begin{pmatrix} u(v_k) \\ v_k \end{pmatrix} \\
&= \underbrace{g_i(x_*)}_{=0} + \frac{\partial g_i}{\partial u}(x_*)(u(v_k) - u_*) + \frac{\partial g_i}{\partial v}(x_*)(v_k - v_*) + o(\|v_k - v_*\|) \\
&= \tilde{\alpha}_k \left(\frac{\partial g_i}{\partial u}(x_*)\right)\Big(a' + \varepsilon_k z' + r(\tilde{\alpha}_k, \varepsilon_k)\Big) + \tilde{\alpha}_k \left(\frac{\partial g_i}{\partial v}(x_*)\right)(a'' + \varepsilon_k z'') \\
&\quad + \tilde{\alpha}_k \varrho_i(\tilde{\alpha}_k, \varepsilon_k) \\
&= \tilde{\alpha}_k \nabla g_i(x_*)(a + \varepsilon_k z) + \tilde{\alpha}_k \left(\frac{\partial g_i}{\partial u}(x_*) r(\tilde{\alpha}_k, \varepsilon_k) + \varrho_i(\tilde{\alpha}_k, \varepsilon_k)\right)
\end{aligned}$$

mit $(\varrho_i) \to 0$ wenn $(\tilde{\alpha}_k) \to 0$. Sei

$$\eta_k := \max_{i \in \mathcal{A}(x_*)} \sup \left\{ \left| \left(\frac{\partial g_i}{\partial u}(x_*)\right) r(\tilde{\alpha}_j, \varepsilon_j') + \rho_i(\tilde{\alpha}_j, \varepsilon_j') \right| \;\middle|\; j \geqslant k,\ 0 \leqslant \varepsilon_j' \leqslant 1 \right\}.$$

Es gilt $(\eta_k) \to 0$ für $k \to \infty$. Setze nun $\gamma := \max\left\{\nabla g_i(x_*)z \;\middle|\; i \in \mathcal{A}(x_*)\right\} < 0$ und $\varepsilon_k := \min\left\{1, \frac{\eta_k}{-\gamma}\right\}$. Dann gilt $0 \leqslant \varepsilon_k \leqslant 1$ und $(\varepsilon_k) \to 0$. Damit gilt für $i \in \mathcal{A}(x_*)$

$$g_i(x_k) \leqslant \tilde{\alpha}_k \underbrace{\nabla g_i(x_*)a}_{\leqslant 0} + \tilde{\alpha}_k \varepsilon_k \underbrace{\nabla g_i(x_*)z}_{<0} + \tilde{\alpha}_k \eta_k \leqslant \tilde{\alpha}_k(\varepsilon_k \gamma + \eta_k) \leqslant 0.$$

3. Es sei $a \in L(x_*)$, $\mathcal{A}_1(x_*) := \left\{i \in \mathcal{A}(x_*) \;\middle|\; \nabla g_i(x_*)a = 0\right\}$ und $\mathcal{A}_2(x_*) := \mathcal{A}(x_*) \setminus \mathcal{A}_1(x_*)$. Wegen $\nabla h(x_*)a = 0$ kann man die Konstruktion von 2. wiederholen mit $z = a$, $\varepsilon_k = 0$ und
$$\begin{pmatrix} \nabla h(x_*) \\ g_{\mathcal{A}_1(x_*)}(x_*) \end{pmatrix} \text{ statt } \nabla h(x_*).$$

□

Bemerkung 3.1.6. (Erweiterung von Satz 3.1.5) Gibt es ein $z \in \mathbf{R}^n$ mit $\nabla h(x_*)z = 0$ und $\nabla g_i(x_*)z < 0$ für alle $i \in \mathcal{A}(x_*)$ und ist h affin linear, so gilt die Guignard-Bedingung ebenfalls, denn aus $h(x_*) = 0$ folgt, daß die linearen Gleichungen lösbar sind; durch Streichung der abhängigen Gleichungen entsteht somit ein reduziertes System, in dem die Rang-Bedingung erfüllt ist.

Somit erhält man die **Mangasarian-Fromowitz-Bedingung in modifizierter Form**.

Die folgende Sprechweise für die Regularitätsbedingung ist üblich:

Definition 3.1.4. Falls die Regularitätsbedingung in x_* erfüllt ist, so heißt x_* **regulärer Punkt** von P.

3.1.2 Restriktionsqualifikationen 2. Ordnung

Im folgenden seien f, g, h zweimal stetig differenzierbar auf D.

Definition 3.1.5. Es sei $x \in P$ und

$$L_0(x) := \left\{ z \,\middle|\, \nabla h(x)z = 0,\ \nabla g_i(x)z = 0 \text{ für alle } i \in \mathcal{A}(x) \right\}.$$

Dann erfüllen g und h die **Restriktionsqualifikation 2. Ordnung**, falls es für alle $z \in L_0(x)$ eine C_2-Kurve $\phi : [-1, 1] \to P$ mit

$$\phi(0) = x,\ \dot{\phi}(0) = z,\ \begin{pmatrix} h(\phi(t)) \\ g_{\mathcal{A}(x)}(\phi(t)) \end{pmatrix} \equiv 0$$

gibt.

Satz 3.1.6. Für reguläre Punkte x_* gilt die Restriktionsqualifikation 2. Ordnung.

Beweis. Die Behauptung folgt nach Anwendung des Hauptsatzes über implizite Funktionen auf die Gleichung

$$\begin{pmatrix} h \\ g_{\mathcal{A}(x_*)} \end{pmatrix} = 0.$$

□

Satz 3.1.7. (Notwendige Bedingungen 2. Ordnung) Sei x_* lokales Minimum von f auf P. Es gelte sowohl die Guignard-Bedingung wie auch die Restriktionsqualifikation 2. Ordnung. Dann gibt es Vektoren $\lambda \in \mathbf{R}^n_{\geqslant}$ und $\mu \in \mathbf{R}^p$ mit

1. $\nabla f(x_*) + \lambda^T \nabla g(x_*) + \mu^T \nabla h(x_*) = 0$,
2. $\lambda^T g(x_*) = 0$,
3. die Lagrange-Matrix $\mathcal{L}(x_*) := F(x_*) + \lambda^T G(x_*) + \mu^T H(x_*)$ ist positiv semidefinit auf $L_0(x_*)$.

Beweis. Nach Satz 3.1.4 existieren λ, μ mit 1. und 2. Wähle ein beliebiges $z \in L_0(x_*)$ und dazu eine passende C_2-Kurve $x(t)$ mit $x(0) = x_*$, die gemäß Restriktionsqualifikation existiert. Weil x_* lokales Minimum ist, folgt mit der Ketten- und Produktregel der Differentiation

$$0 \leqslant \frac{d^2}{dt^2} f(x(t))|_{t=0} = \frac{d}{dt}(\nabla f(x(t))\dot{x}(t))|_{t=0} = \dot{x}(0)^T F(x_*)\dot{x}(0) + \nabla f(x_*)\ddot{x}(0). \tag{3.5}$$

Da $h(x(t)) \equiv 0$ und $g_{\mathcal{A}(x_*)}(x(t)) \equiv 0$ folgt $\lambda^T g(x(t)) + \mu^T h(x(t)) \equiv 0$. Zweimaliges Differenzieren ergibt nun:

$$\begin{aligned} 0 &= \frac{d^2}{dt^2}\Big(\lambda^T g(x(t)) + \mu^T h(x(t))\Big)|_{t=0} \\ &= \frac{d}{dt}\Big(\lambda^T \nabla g(x(t))\dot{x}(t) + \mu^T \nabla h(x(t))\dot{x}(t)\Big)|_{t=0} \\ &= \dot{x}(0)^T \left[\lambda^T G(x_*) + \mu^T H(x_*)\right] \dot{x}(0) + \left[\lambda^T \nabla g(x_*) + \mu^T \nabla h(x_*)\right] \ddot{x}(0) \end{aligned}$$

Addition der Ungleichung (3.5) ergibt nun

$$\begin{aligned}
0 &\leqslant \dot{x}(0)^T \Big[\underbrace{F(x_*) + \lambda^T G(x_*) + \mu^T H(x_*)}_{=\mathcal{L}(x_*)} \Big] \dot{x}(0) + \Big[\underbrace{\nabla f(x_*) + \lambda^T \nabla g(x_*) + \mu^T \nabla h(x_*)}_{=0 \text{ nach Multiplikatorregel}} \Big] \ddot{x}(0) \\
&= \dot{x}(0)^T \mathcal{L}(x_*) \dot{x}(0) = \dot{z}^T \mathcal{L}(x_*) z.
\end{aligned}$$

□

Bemerkung 3.1.7. Die Bedingungen aus Satz 3.1.7 sind nicht hinreichend, selbst dann nicht, wenn $\mathcal{L}(x_*)$ positiv definit auf $L_0(x_*)$ ist.

Wir werden nun hinreichende Bedingungen formulieren.

Satz 3.1.8. (Hinreichende Bedingungen 2. Ordnung) Es sei $x_* \in P$ und die Multiplikatorregel sei für $\lambda \in \mathbf{R}^n_{\geqslant}$, $\mu \in \mathbf{R}^p$ erfüllt. Ist $\mathcal{L}(x_*)$ positiv definit auf

$$\hat{L}_0(x_*) := \Big\{ z \;\Big|\; \nabla h(x_*) z = 0,\ \nabla g_i(x_*) z = 0 \text{ für alle } \lambda_i > 0,\ \nabla g_{\mathcal{A}(x_*)} z \leqslant 0 \Big\},$$

so ist x_* strenges lokales Minimum von f auf P.

Beweis. (indirekt) Wir nehmen an, es gebe eine Folge $(x_k) \subseteq P$, $(x_k) \to x_*$ mit $f(x_k) \leqslant f(x_*)$. Es sei $x_k = x_* + \delta_k s_k$ mit $(\delta_k) \subseteq \mathbf{R}_>$, $(\delta_k) \to 0$, $s_k \in \mathbf{R}^n$ und $\|s_k\| = 1$. O. B. d. A. betrachten wir nur eine Teilfolge mit $(s_k) \to s_*$. Für hinreichend große k gilt

$$0 > \frac{1}{2} g_i(x_*) \geqslant g_i(x_k) \text{ für } i \notin \mathcal{A}(x_*)$$

und mittels Taylorentwicklung zeigt man

$$\begin{aligned}
&0 \geqslant f(x_k) - f(x_*) = \delta_k \nabla f(x_*) s_k + \frac{1}{2}\delta_k^2 s_k^T F(x_{1,k}) s_k \text{ mit } x_{1,k} \in [x_k, x_*], \\
&0 \geqslant g_i(x_k) = \underbrace{g_i(x_*)}_{=0} + \delta_k \nabla g_i(x_*) s_k + \frac{1}{2}\delta_k^2 s_k^T G_i(x_{2,k,i}) s_k \text{ mit } x_{2,k,i} \in [x_k, x_*] \text{ für } i \in \mathcal{A}(x_*), \\
&0 = h_i(x_k) = \underbrace{h_i(x_*)}_{=0} + \delta_k \nabla h_i(x_*) s_k + \frac{1}{2}\delta_k^2 s_k^T H_i(x_{3,k,i}) s_k \text{ mit } x_{3,k,i} \in [x_k, x_*] \text{ für } i \in \{1, \dots, p\}.
\end{aligned}$$

Division durch δ_k und nachfolgender Grenzübergang $k \to \infty$ ergibt die Beziehungen

$$\begin{aligned}
\nabla g_i(x_*) s_* &\leqslant 0 \text{ für } i \in \mathcal{A}(x_*), \\
\nabla h_i(x_*) s_* &= 0 \text{ für } i \in \{1, \dots, p\}, \\
\nabla f(x_*) s_* &\leqslant 0.
\end{aligned}$$

Damit gilt laut Multiplikatorregel

$$0 \leqslant -\nabla f(x_*) s_* = \mu^T \underbrace{\nabla h(x_*) s_*}_{=0} + \sum_{i=1}^m \lambda_i \underbrace{\nabla g_i(x_*) s_*}_{\leqslant 0} \leqslant 0,$$

es gilt also überall Gleichheit und insbesondere sind alle Terme $\lambda_i \nabla g_i(x_*) s_*$ null. Damit gilt $\nabla g_i(x_*) s_* = 0$ falls $\lambda_i > 0$. Folglich ist $s_* \in \hat{L}_0(x_*)$. Wir bilden nun

$$\begin{aligned} 0 &\geqslant f(x_k) - f(x_*) + \lambda^T g(x_k) + \mu^T h(x_k) \\ &= \delta_k \Big(\underbrace{\nabla f(x_*) + \lambda^T \nabla g(x_*) + \mu^T h(x_*)}_{=0 \text{ nach Multiplikatorregel}} \Big) s_k \\ &\quad + \frac{1}{2} \delta_k^2 s_k^T \Big(F(x_{1,k}) + \sum_{i=1}^m \lambda_i G_i(x_{2,k,i}) + \sum_{i=1}^p \mu_i H_i(x_{3,k,i}) \Big) s_k. \end{aligned}$$

Mittels Division durch δ_k^2 und nachfolgendem Grenzprozeß $k \to \infty$ ergibt sich hieraus der Widerspruch $s_*^T \mathcal{L}(x_*) s_* \leqslant 0$. □

Die Bedingung aus Satz 3.1.8 wird **hinreichende Bedingung 2. Ordnung** genannt.

3.1.3 Strikte Komplementarität

Definition 3.1.6. Falls zusätzlich zur Multiplikatorregel für x_*, λ und μ gilt: $\lambda > g(x_*)$[1], so sagt man, in x_* sei die Bedingung der **strikten Komplementarität** erfüllt.

Satz 3.1.9. Es sei $x_* \in P$ regulär. Dann gilt:

1. Ist x_* ein lokales Minimum von f auf P, so gibt es eindeutig bestimmte $\lambda \in \mathbf{R}^m_{\geqslant}$ und $\mu \in \mathbf{R}^p$, die die Kuhn-Tucker-Bedingungen in x_* erfüllen. Außerdem ist $\mathcal{L}(x_*)$ positiv semidefinit auf $L_0(x_*)$.

2. Gibt es umgekehrt $\lambda \in \mathbf{R}^m_{\geqslant}$ und $\mu \in \mathbf{R}^p$, die die Kuhn-Tucker-Bedingungen in x_* erfüllen, ist $\mathcal{L}(x_*)$ positiv definit auf $L_0(x_*)$ und ist in x_* mit λ die strikte Komplementaritätsbedingung erfüllt, so ist x_* strenges lokales Minimum von f auf P.

Beweis. Wegen Satz 3.1.5 ist die Guignard-Bedingung in x_* erfüllt. Nach Satz 3.1.6 gilt die Restriktionsqualifikation 2. Ordnung in x_*. Damit sind die Voraussetzungen von Satz 3.1.7 erfüllt, hieraus folgt dann die erste Behauptung.

Sei nun in x_* die strikte Komplementaritätsbedingung für λ und μ erfüllt. Also gilt $L_0(x_*) = \hat{L}_0(x_*)$ und damit ist $\mathcal{L}(x_*)$ positiv definit auf $\hat{L}_0(x_*)$. Nun folgt die zweite Behauptung aus Satz 3.1.8. □

Definition 3.1.7. Ein Vektor $x \in \mathbf{R}^n$ heißt **streng regulär**, wenn $x \in P$, x regulärer Punkt ist und wenn sowohl die Kuhn-Tucker-Bedingungen in x, die strikte Komplementarität und die hinreichende Bedingung 2. Ordnung gelten.

3.1.4 Sattelpunkte und die Lagrangefunktion

Definition 3.1.8. Die Funktion $\ell \colon \mathbf{R}^n \times \mathbf{R}^m_{\geqslant} \times \mathbf{R}^p \to \mathbf{R}$, definiert durch

$$\ell(x, \lambda, \mu) := f(x) + \lambda^T g(x) + \mu^T h(x),$$

[1] Eine andere (längere) Formulierung hierfür ist: Für $i = 1, \ldots, m$ gilt $\lambda_i > 0$ oder $g_i(x_*) < 0$.

heißt **Lagrange-Funktion**. Ein Tupel (x_*, λ_*, μ_*) mit

$$\ell(x_*, \lambda, \mu) \leqslant \ell(x_*, \lambda_*, \mu_*) \leqslant \ell(x, \lambda_*, \mu_*) \text{ für alle } x \in \mathbf{R}^n,\ \lambda \in \mathbf{R}^m_{\geqslant},\ \mu \in \mathbf{R}^p \tag{3.6}$$

heißt **Sattelpunkt** der Lagrangefunktion.

Bemerkung 3.1.8. Mit Hilfe der Lagrangefunktion lassen sich die notwendigen Extremalbedingungen erster und zweiter Ordnung wie folgt formulieren:

$$\begin{array}{ll}
x_* \in P & \text{primale Zulässigkeit} \\
\lambda_* \geqslant 0 & \text{duale Zulässigkeit} \\
\nabla_x \ell(x_*, \lambda_*, \mu_*) = 0 & \text{Multiplikatorregel (3.2)} \\
\lambda_*^T g(x_*) = 0 & \text{Komplementarität (3.3)} \\
\multicolumn{2}{l}{\mathcal{L}(x_*) = \nabla^2_{xx} \ell(x_*, \lambda_*, \mu_*) \text{ ist positiv semidefinit auf } L_0(x_*)}
\end{array}$$

Satz 3.1.10. Ist (x_*, λ_*, μ_*) ein Sattelpunkt von ℓ, so ist x_* eine globale Optimallösung mit $\lambda_*^T g(x_*) = 0$.

Beweis. Wir zeigen zunächst, daß $x_* \in P$ gilt: Weil (x_*, λ_*, μ_*) ein Sattelpunkt von ℓ ist, folgt nach (3.6)

$$\lambda^T g(x_*) + \mu^T h(x_*) \leqslant \lambda_*^T g(x_*) + \mu_*^T h(x_*) \text{ für alle } \lambda \in \mathbf{R}^m_{\geqslant},\ \mu \in \mathbf{R}^p,$$

denn gäbe es ein $h_i(x_*) \neq 0$, so wäre die Ungleichung bei Wahl von $\lambda = \lambda^*$ und geeigneter Wahl von μ verletzt (nur Komponente i von μ unterschiedlich von μ_* wählen). Also folgt $h(x_*) = 0$. Analog folgt $g(x_*) \leqslant 0$, mit $(\lambda_*)_i = 0$, wenn $g_i(x_*) < 0$. Demnach gilt die Komplementarität (3.3) für λ_* und x_*.
Nun zeigen wir, daß x_* globales Optimum auf P ist: Es gilt aufgrund von Zulässigkeit und Komplementarität von x_* und λ_*

$$f(x_*) + \underbrace{\lambda_*^T g(x_*)}_{=0} + \underbrace{\mu_*^T h(x_*)}_{=0} \leqslant f(x) + \lambda_*^T g(x) + \mu_*^T h(x) \text{ für alle } x \in \mathbf{R}^n.$$

Hieraus folgt für $x \in P$

$$f(x_*) \leqslant f(x) + \lambda_*^T \underbrace{g(x)}_{\leqslant 0} + \mu_*^T \underbrace{h(x)}_{=0} \leqslant f(x).$$

$\square$

Korollar 3.1.11. (Regularitätseigenschaften von Sattelpunkten) f, g und h seien C_1-Funktionen. Ist (x_*, λ_*, μ_*) Sattelpunkt von ℓ, so gilt die Multiplikatorregel und $L(x_*)$ enthält keine Abstiegsrichtung.

Beweis. Es sei $(\delta_k) \subseteq \mathbf{R}_>$ eine Nullfolge und $s \in \mathbf{R}^n$ beliebig. Weil

$$\ell(x_*, \lambda_*, \mu_*) \leqslant \ell(x_* + \delta_k s, \lambda_*, \mu_*)$$

gilt, ist die Ableitung von ℓ in Richtung s nichtnegativ. Da dies für alle $s \in \mathbf{R}^n$ gilt, muß $\nabla_x \ell(x_*, \lambda_*, \mu_*) = 0$ gelten. Damit ist die Multiplikatorbedingung (3.2) erfüllt.

Sei nun $s \in L(x_*)$. Dann gilt $\nabla g_i(x_*)s \leqslant 0$ für $i \in \mathcal{A}(x_*)$ und $\nabla h(x_*)s = 0$. Die Richtungsableitung von f nach s ist damit wegen $(\lambda_*)_i = 0$ für $i \notin \mathcal{A}(x_*)$

$$\nabla f(x_*)s = -\underbrace{\lambda_*^T \nabla g(x_*)s}_{\leqslant 0} - \mu_*^T \underbrace{\nabla h(x_*)s}_{=0} \geqslant 0.$$

□

Bemerkung 3.1.9. Die Bestimmung von Sattelpunkten kann in bestimmten Fällen zur Lösung nichtlinearer Optimierungsprobleme herangezogen werden. Bestimme etwa

$$\Phi(\lambda, \mu) = \min_{x \in \mathbf{R}^n} \ell(x, \lambda, \mu).$$

Falls dies einfach (in geschlossener Form) zu lösen ist, bestimme anschließend

$$\max_{\lambda \in \mathbf{R}^m_{\geqslant},\, \mu \in \mathbf{R}^p} \Phi(\lambda, \mu).$$

Vorsicht! Nicht immer wenn Minima der Optimierungsaufgabe existieren gibt es auch einen Sattelpunkt der Lagrangefunktion.

Min-Max-Aufgaben

Es sei $\Phi\colon D \times \mathcal{E} \subseteq \mathbf{R}^{n_1} \times \mathbf{R}^{n_2} \to \mathbf{R}$. Zu $x \in D$ und $y \in \mathcal{E}$ seien

$$\varphi(x) := \max_{v \in \mathcal{E}} \Phi(x, v) \text{ und } \Psi(y) := \min_{u \in D} \Phi(u, y)$$

wohldefiniert und es sei

$$\bar{x} := \operatorname{argmin}\Big\{\varphi(x) \;\Big|\; x \in D\Big\} \text{ und } \bar{y} := \operatorname{argmax}\Big\{\Psi(y) \;\Big|\; y \in \mathcal{E}\Big\}.$$

Dann gilt der folgende Satz:

Satz 3.1.12. (Min-Max-Probleme und Sattelpunkte)
Es gilt

$$\max_{y \in \mathcal{E}} \min_{x \in D} \Phi(x, y) = \Psi(\bar{y}) \leqslant \varphi(\bar{x}) = \min_{x \in D} \max_{y \in \mathcal{E}} \Phi(x, y).$$

Außerdem gilt

$$\Psi(\bar{y}) = \Phi(\bar{x}, \bar{y}) = \varphi(\bar{x}) \tag{3.7}$$

genau dann, wenn $(\bar{x}, \bar{y})$ Sattelpunkt von Φ ist.

Beweis. Es gilt $\Psi(y) \leqslant \Phi(x, y) \leqslant \varphi(x)$ für alle $x \in D$, $y \in \mathcal{E}$. Hieraus folgt sofort die erste Behauptung.
Ist $(\bar{x}, \bar{y})$ Sattelpunkt, d. h.

$$\Phi(\bar{x}, y) \leqslant \Phi(\bar{x}, \bar{y}) \leqslant \Phi(x, \bar{y}) \text{ für alle } x \in D,\ y \in \mathcal{E},$$

so gilt

$$\varphi(\bar{x}) = \max_{y \in \mathcal{E}} \Phi(\bar{x}, y) \leqslant \Phi(\bar{x}, \bar{y}) \leqslant \min_{x \in D} \Phi(x, \bar{y}) = \Psi(\bar{y}) \leqslant \varphi(\bar{x}),$$

also gilt in obiger Ungleichung Gleichheit, womit (3.7) folgt.
Gilt umgekehrt (3.7), so folgt für $x \in D$ und $y \in \mathcal{E}$:

$$\Phi(\bar{x}, y) \leqslant \varphi(\bar{x}) = \Phi(\bar{x}, \bar{y}) = \Psi(\bar{y}) \leqslant \Phi(x, \bar{y}),$$

also ist $(\bar{x}, \bar{y})$ Sattelpunkt. □

3.1.5 Konvexe Optimierungsaufgaben

Definition 3.1.9. Es sei $D \subseteq \mathbf{R}^n$ konvex und offen. Es seien f, g konvex und stetig differenzierbar, sowie h affin linear. Dann bezeichnet man die Aufgabe

$$\min \left\{ f(x) \;\middle|\; g(x) \leqslant 0,\ h(x) = 0 \right\} \tag{3.8}$$

als **konvexes Optimierungsproblem**. Insbesondere ist P hier konvex.

In diesem Abschnitt sei H nicht die zweite, sondern die erste Ableitung von h, es sei also $h(x) = Hx - h_0$.

Satz 3.1.13. (Hinreichende Bedingungen für konvexe Probleme) Es sei $x_* \in P$. Falls in x_* die Multiplikatorregel gilt, so ist x_* globales Minimum von f auf P.

Beweis. Es sei $x \in P$. λ und μ seien durch die Multiplikatorregel definiert. Offensichtlich gilt

$$f(x) + \underbrace{\lambda^T g(x)}_{\leqslant 0} + \mu^T \underbrace{h(x)}_{=0} \leqslant f(x).$$

Die Gradientenungleichung für konvexe Funktionen liefert

$$\begin{aligned} f(x) &\geqslant f(x_*) + \nabla f(x_*)(x - x_*), \\ g(x) &\geqslant g(x_*) + \nabla g(x_*)(x - x_*), \\ h(x) &= h(x_*) + \nabla h(x_*)(x - x_*). \end{aligned}$$

Die gewichtete Summe dieser (Un-)Gleichungen ergibt (zweite Ungleichung mit λ^T, dritte mit μ^T multipliziert)

$$f(x) \geqslant f(x_*) + \lambda^T \underbrace{g(x_*)}_{=0} + \mu^T \underbrace{h(x_*)}_{=0} + \underbrace{\left[\nabla f(x_*) + \lambda^T \nabla g(x_*) + \mu^T \nabla h(x_*) \right]}_{=0 \text{ nach Multiplikatorregel}} (x - x_*),$$

also $f(x) \geqslant f(x_*)$. □

Die Slater-Bedingung

Für konvexe Aufgaben ist die Slater-Bedingung von großer Bedeutung.

Definition 3.1.10. Die **Slater-Bedingung** lautet:

$$\text{Es gibt ein } x_0 \in D \text{ mit } g(x_0) < 0,\ h(x_0) = 0.$$

Bemerkung 3.1.10. Die Slater-Bedingung impliziert die modifizierte Form der Bedingung von Mangasarian und Fromowitz (siehe Bemerkung zu Satz 3.1.5) für alle $x \in P$.

Beweis. Die Slater-Bedingung sei mit $x_0 \in D$ erfüllt. Zu zeigen: Es gibt ein $z \in \mathbf{R}^n$ mit $\nabla h(x)z = 0$, $\nabla g_{\mathcal{A}(x)}(x)z < 0$. Wähle $z := x_0 - x$. Dann gilt

$$0 = h(x_0) = \underbrace{h(x)}_{=0} + \nabla h(x) \underbrace{(x_0 - x)}_{=z},$$

woraus $\nabla h(x)z = 0$ folgt.

$$0 > g_{\mathcal{A}(x)}(x_0) \geqslant \underbrace{g_{\mathcal{A}(x)}(x)}_{=0} + \nabla g_{\mathcal{A}(x)}(x) \underbrace{(x_0 - x)}_{=z}$$

ergibt $\nabla g_{\mathcal{A}(x)}(x)z < 0$. □

Satz 3.1.14. (Notwendige und hinreichende Bedingungen für konvexe Probleme) Sei $x_* \in P$. Die Slaterbedingung sei erfüllt. Dann ist x_* globales Minimum von f auf P genau dann, wenn in x_* die Multiplikatorregel gilt.

Beweis. Die Multiplikatorregel ist nach Satz Satz 3.1.13 hinreichend dafür, daß x_* Minimum ist.
Sei umgekehrt x_* Minimum von f auf P. Da die Slater-Bedingung nach Bemerkung 3.1.10 und Satz 3.1.4 die Guignard-Bedingung impliziert, ist nach Satz 3.1.4 die Multiplikatorregel erfüllt. □

Lemma 3.1.15. Die Slater-Bedingung sei nicht erfüllt. Dann existieren $a \in \mathbf{R}^m_{\geqslant}$ und $b \in \mathbf{R}^p$ mit $\|a\| + \|b\| > 0$ und

$$a^T g(x) + b^T h(x) \geqslant 0 \text{ für alle } x \in D.$$

Beweis. Setze $\Lambda(x) := \left\{ \begin{pmatrix} y \\ z \end{pmatrix} \,\middle|\, g(x) < y,\ h(x) = z \right\}$ für $x \in D$ und $\Lambda := \bigcup_{x \in D} \Lambda(x)$.
Behauptung: Λ ist konvex. Es sei

$$\begin{pmatrix} y \\ z \end{pmatrix} \in \Lambda(x) \text{ und } \begin{pmatrix} y' \\ z' \end{pmatrix} \in \Lambda(x') \text{ und } \lambda \in [0,1]$$

mit $x, x' \in D$. Weil D konvex ist, ist auch $\bar{x} = \lambda x + (1-\lambda)x' \in D$. Nun ist

$$g(\bar{x}) \leqslant \lambda g(x) + (1-\lambda) g(x') < \lambda y + (1-\lambda) y'$$

und analog

$$h(\bar{x}) = \lambda h(x) + (1-\lambda)h(x') = \lambda z + (1-\lambda)z',$$

also ist $\lambda\begin{pmatrix} y \\ z \end{pmatrix} + (1-\lambda)\begin{pmatrix} y' \\ z' \end{pmatrix} \in \Lambda(\bar{x}) \subseteq \Lambda$. Also ist Λ konvex.

Da $0 \notin \Lambda$ (sonst wäre die Slater-Bedingung erfüllt), kann man aus dem Trennungssatz die Existenz von $\begin{pmatrix} a \\ b \end{pmatrix} \neq 0$ mit $a^T y + b^T z \geqslant 0$ für alle $\begin{pmatrix} y \\ z \end{pmatrix} \in \Lambda$ für $x \in D$ folgern. Weil y beliebig groß gewählt werden kann, muß hierbei $a \geqslant 0$ gelten.

Sei nun $\varepsilon > 0$ beliebig und $x \in D$ beliebig. Setze $y := g(x) + \varepsilon\mathbf{1}$ und $z := h(x)$. Dann ist $\begin{pmatrix} y \\ z \end{pmatrix} \in \Lambda(x)$. Folglich gilt $a^T\Big(g(x) + \varepsilon\mathbf{1}\Big) + b^T h(x) \geqslant 0$, bzw.

$$a^T g(x) + b^T h(x) \geqslant -\mathbf{1}^T a\varepsilon \text{ für alle } x \in D,\ \varepsilon > 0.$$

Also gilt $a^T g(x) + b^T h(x) \geqslant 0$ für alle $x \in D$. □

Bemerkung 3.1.11. (Zusätze zu Lemma 3.1.15)

1. Ist $D = \mathbf{R}^n$ und $h(x) = 0$ lösbar, so gibt es Paare (a, b) mit $a \neq 0$. Hat weiterhin ∇h maximalen Rang, so ist stets $a \neq 0$.

   Beweis. (indirekt) Annahme: $a = 0$. Also gilt $b \neq 0$. Folglich ist

   $$0 \leqslant b^T h(x) = b^T Hx - b^T h_0 \text{ für alle } x \in \mathbf{R}^n.$$

   Speziell für $x = 0$ folgt $b^T h_0 \leqslant 0$.

   Ist $b^T H \neq 0$, so folgt der Widerspruch, weil x beliebig unbeschränkt wählbar ist. Also gilt $b^T H = 0$. Ist H von maximalem Rang, so folgt damit $b = 0$, ein Widerspruch. Also ist $a = 0$ in diesem Fall nicht möglich.

   Ist H nicht von maximalem Rang, so gibt es $b \neq 0$ mit $b^T H = 0$. Es sei o. B. d. A.

   $$H = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix} \text{ mit } \operatorname{rang} H_1 = \operatorname{rang} H \text{ und } h_0 = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}.$$

   Dann gilt $h(x) = 0$ nur dann, wenn auch $H_1 x - h_1 = 0$. Anwendung von Lemma 3.1.15 auf $g(x)$ und $\bar{h}(x) := H_1 x - h_1$ liefert $\bar{a}$ und $\bar{b}_1$ mit $\bar{a} \neq 0$ und

   $$\bar{a}^T g(x) + \bar{b}_1^T \bar{h}(x) \geqslant 0 \text{ für alle } x \in \mathbf{R}^n.$$

   Mit $\bar{b} := \begin{pmatrix} \bar{b}_1 \\ 0 \end{pmatrix}$ folgt die Behauptung. □

2. Gilt die Slater-Bedingung in x_0, so gibt es für alle $a \in \mathbf{R}^m_{\geqslant}$, $b \in \mathbf{R}^p$ mit $a \neq 0$ ein $x \in D$ mit $a^T g(x) + b^T h(x) < 0$.

   Beweis. trivial (Man setze $x = x_0$.) □

Bemerkung 3.1.12. Lemma 3.1.15 ergibt zusammen mit Bemerkung 3.1.11 den folgenden Alternativsatz:

Es sei $D = \mathbf{R}^n$ und H von maximalem Rang. Dann ist genau eine der beiden folgenden Aussagen richtig:

1. Es existieren $a \in \mathbf{R}^m_{\geqslant} \setminus \{0\}$ und $b \in \mathbf{R}^p$, so daß für alle $x \in D$
$$a^T g(x) + b^T h(x) \geqslant 0$$
gilt.

2. Es existiert ein x_0 in D, so daß für alle $a \in \mathbf{R}^m_{\geqslant} \setminus \{0\}$ und $b \in \mathbf{R}^p$
$$a^T g(x_0) + b^T h(x_0) < 0$$
gilt.

Satz 3.1.16. (Sattelpunktbedingung für konvexe Probleme)
Es sei $D = \mathbf{R}^n$. Die Slater-Bedingung sei erfüllt. Dann ist x_* globales Minimum von f auf P genau dann, wenn $\lambda_* \in \mathbf{R}^m_{\geqslant}$ und $\mu_* \in \mathbf{R}^p$ existieren, so daß (x_*, λ_*, μ_*) Sattelpunkt der Lagrangefunktion ist.

Beweis. Es sei (x_*, λ_*, μ_*) Sattelpunkt der Lagrangefunktion. Dann ist x_* nach Satz 3.1.10 die globale Optimallösung.
Sei umgekehrt x_* globales Minimum von f auf P. Es sei $g_0(x) := f(x) - f(x_*) \geqslant 0$. Also gilt
$$\left\{ x \;\middle|\; g_0(x) < 0,\ g(x) < 0,\ h(x) = 0 \right\} = \emptyset.$$
Also gibt es $\begin{pmatrix} a_0 \\ a \end{pmatrix} \geqslant 0$, $\begin{pmatrix} a_0 \\ a \end{pmatrix} \neq 0$ und $b \in \mathbf{R}^p$ derart, daß
$$a_0\Big(f(x) - f(x_*)\Big) + a^T g(x) + b^T h(x) \geqslant 0 \text{ für alle } x \in \mathbf{R}^n$$
gilt.
Falls $a_0 = 0$, dann folgt $a \neq 0$, was aber wegen der Slater-Bedingung auf einen Widerspruch führt. Also ist a_0 echt positiv. Division durch a_0 ergibt mit $\lambda_* := \frac{1}{a_0} a$ und $\mu_* := \frac{1}{a_0} b$:
$$f(x_*) \leqslant \underbrace{f(x) + \lambda_*^T g(x) + \mu_*^T h(x)}_{=\ell(x,\lambda_*,\mu_*)} \text{ für alle } x \in \mathbf{R}^n.$$
Da $g(x_*) \leqslant 0$, $h(x_*) = 0$ folgt
$$\ell(x_*, \lambda, \mu) = f(x_*) + \underbrace{\lambda^T g(x_*)}_{\leqslant 0} + \mu^T \underbrace{h(x_*)}_{=0} \leqslant f(x_*) \text{ für alle } \lambda \in \mathbf{R}^m_{\geqslant},\ \mu \in \mathbf{R}^p.$$
Wir haben damit
$$\ell(x_*, \lambda, \mu) \leqslant f(x_*) \leqslant \ell(x, \lambda_*, \mu_*) \text{ für alle } x \in \mathbf{R}^n,\ \lambda \in \mathbf{R}^m_{\geqslant},\ \mu \in \mathbf{R}^p.$$
Setzen wir speziell $\lambda = \lambda_*$, $\mu = \mu_*$ und $x = x_*$, so gilt also Gleichheit, d. h. $f(x_*) = f(x_*) + \lambda_*^T g(x_*) + \mu_*^T h(x_*)$ und (x_*, λ_*, μ_*) ist Sattelpunkt von ℓ. ☐

Der Spezialfall der definiten quadratischen Optimierungsaufgaben

Es sei das folgende Problem gegeben:

$$\min\left\{\frac{1}{2}x^TQx - q^Tx \;\middle|\; Ax \leqslant b,\ Hx = h_0\right\},$$

wobei $Q \in \mathbf{R}^{n\times n}$ positiv definit. Es sei $A \in \mathbf{R}^{m\times n}$ mit $m < n$ und $H \in \mathbf{R}^{p\times n}$ mit $p < n$. Die Slater-Bedingung sei erfüllt. P ist abgeschlossen. Das Optimum existiert, da f nach unten beschränkt ist. Die Lagrangefunktion lautet dann

$$\ell(x,\lambda,\mu) := \frac{1}{2}x^TQx - q^Tx + \lambda^T(Ax-b) + \mu^T(Hx-h_0)$$

und ist konvex in x. Die Aufgabe

$$\Psi(\lambda,\mu) := \inf_{x\in\mathbf{R}^n} \ell(x,\lambda,\mu)$$

ist eine konvexe Aufgabe ohne Restriktionen. Die Optimallösung werde bei $x_*(\lambda,\mu)$ angenommen. Dann gilt

$$0 = \nabla_x\ell(x_*,\lambda,\mu) = (Qx_* - q)^T + \lambda^TA + \mu^TH,$$

was wie folgt nach x_* aufgelöst werden kann:

$$x_*^T(\lambda,\mu) = (q^T - \lambda^TA - \mu^TH)Q^{-1}.$$

Damit folgt durch Einsetzen

$$\begin{aligned}
\Psi(\lambda,\mu) =& \frac{1}{2}x_*^TQx_* - q^Tx_* + \lambda^T(Ax_*-b) + \mu^T(Hx_*-h_0)\\
=& -\frac{1}{2}x_*^TQx_* + \underbrace{(x_*^TQ - q^T + \lambda^TA + \mu^TH)}_{=0}x_* - \lambda^Tb - \mu^Th_0\\
=& -\frac{1}{2}(q^T - \lambda^TA - \mu^TH)Q^{-1}(q - A^T\lambda - H^T\mu) - \lambda^Tb - \mu^Th_0\\
=& -\frac{1}{2}(\lambda^TA + \mu^TH)Q^{-1}(H^T\mu + A^T\lambda) + (q^TQ^{-1}A^T - b^T)\lambda\\
&+ (q^TQ^{-1}H^T - h_0^T)\mu - \frac{1}{2}q^TQ^{-1}q\\
=& -\frac{1}{2}\begin{pmatrix}\lambda\\ \mu\end{pmatrix}^T\tilde{Q}\begin{pmatrix}\lambda\\ \mu\end{pmatrix} - \tilde{q}^T\begin{pmatrix}\lambda\\ \mu\end{pmatrix} - \tilde{q}_0
\end{aligned}$$

mit $\tilde{Q} = \begin{pmatrix}A\\ H\end{pmatrix}Q^{-1}\begin{pmatrix}A\\ H\end{pmatrix}^T$, $\tilde{q} = \begin{pmatrix}b^T - q^TQ^{-1}A^T\\ h_0^T - q^TQ^{-1}H^T\end{pmatrix}$ und $\tilde{q}_0 = \frac{1}{2}q^TQ^{-1}q$. Die Maximierung von Ψ ist äquivalent zur Minimierung von $-\Psi$, die Konstante ist unerheblich. Also lautet die zu lösende Aufgabe

$$\min_{\lambda\in\mathbf{R}^m_{\geqslant},\,\mu\in\mathbf{R}^p} \frac{1}{2}\begin{pmatrix}\lambda\\ \mu\end{pmatrix}^T\tilde{Q}\begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \tilde{q}^T\begin{pmatrix}\lambda\\ \mu\end{pmatrix}. \tag{3.9}$$

Falls die Zeilen der Restriktionen linear unabhängig sind, dann ist $\tilde{Q}$ positiv definit. Die duale Aufgabe (3.9) ist dann einfacher zu lösen, weil die Restriktionen $\lambda \geqslant 0$ i. allg. leichter zu behandeln sind. Falls (λ_*,μ_*) optimal für (3.9) ist, dann ist $x_*(\lambda_*,\mu_*)$ optimal für die Ausgangsaufgabe.

3.2 Quadratische Optimierungsaufgaben

Es sei die quadratische Optimierungsaufgabe

$$\min\left\{f(x)\;\middle|\;g(x)\leqslant 0,\ h(x)=0\right\} \tag{3.10}$$

mit $f(x)=\frac{1}{2}x^TAx-b^Tx+\gamma$, $g(x)=Gx-g_0$ und $h(x)=Hx-h_0$ gegeben. A sei positiv definit und es gelte die Slater-Bedingung. Dann gilt der folgende Satz:

Satz 3.2.1. Es sei rang$H$ maximal und $x_0\in\mathbf{R}^n$. Dann ist die Lösung x_* der konvexen Optimierungsaufgabe

$$f(x_*):=\min\left\{f(x)\;\middle|\;h(x)=0\right\}$$

mit zugehörigen Lagrange-Parametern μ_* durch

$$\begin{bmatrix} A & H^T \\ H & 0 \end{bmatrix}\begin{pmatrix} x_*-x_0 \\ \mu_* \end{pmatrix}=-\begin{pmatrix} \nabla f(x_0)^T \\ h(x_0) \end{pmatrix} \tag{3.11}$$

gegeben.

Beweis. Aus (3.11) folgt $H(x_*-x_0)=-h(x_0)$, also $Hx_*=h_0$, also ist x_* zulässig. Für konvexe Aufgaben ist die Gültigkeit der Multiplikatorregel in x_* notwendig und hinreichend für die Optimalität von x_*. Wegen

$$\nabla f(x)=x^TA-b^T \text{ und } \nabla h(x)=H$$

folgt die Multiplikatorregel aus

$$\begin{aligned}
\nabla f(x_*)+\mu_*^T\nabla h(x_*) &= x_*^TA-b^T+\mu_*^TH \\
&= (x_*^T-x_0^T)A+\underbrace{x_0^TA-b^T}_{=\nabla f(x_0)}+\mu_*^TH \\
&= \underbrace{\left(A(x_*-x_0)+H^T\mu_*\right)^T}_{=-\nabla f(x_0)\text{ wegen (3.11)}}+\nabla f(x_0) \\
&= 0.
\end{aligned}$$

□

Das System (3.11) kann mit Hilfe der QR-Zerlegung numerisch stabil gelöst werden. Dazu berechnet man zunächst mittels Householder-Transformationen die Darstellung

$$QH^T=\begin{pmatrix} R \\ 0 \end{pmatrix},$$

wobei R rechte obere Dreieckmatrix ist, und dann

$$c:=Q\nabla f(x_0)^T=\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \text{ mit } c_1\in\mathbf{R}^p \text{ und } c_2\in\mathbf{R}^{n-p}$$

und

$$C := QAQ^T = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \text{ mit } C_{11} \in \mathbf{R}^{p \times p} \text{ und } C_{22} \in \mathbf{R}^{n-p \times n-p}.$$

Bestimme nun eine Lösung s_1 des Dreiecksystems $R^T s_1 = -h(x_0)$ und eine Lösung s_2 des Systems $C_{22}s_2 = -c_2 - C_{21}s_1$ mittels Cholesky-Zerlegung von C_{22} (C_{22} ist positiv definit). Bestimme nun eine Lösung μ_* des Dreiecksystems $R\mu_* := -c_1 - C_{11}s_1 - C_{12}s_2$ und bestimme abschließend $x_* := x_0 - Q^T\begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$ durch Einsetzen.

3.2.1 Das PQP-Verfahren

Das PQP-Verfahren bestimmt eine optimale Lösung der quadratischen Optimierungsaufgabe (3.10), wobei der zuvor beschriebene Algorithmus zur Lösung der Optimallösung des Problems mit Gleichheitsrestriktionen als Unterprogramm verwendet wird, wobei die zur Zeit aktiven Ungleichheitsrestriktionen wie Gleichheitsrestriktionen behandelt werden und die inaktiven Restriktionen fortgelassen werden.
Das PQP-Verfahren lautet wie folgt:

Algorithmus 3.2.1. (Das PQP-Verfahren)

1. Wähle ein $x_0 \in P$ mittels Phase 1 des Simplexverfahrens.
   Setze $k := 0$.

2. Setze $\mathcal{A}_k := \mathcal{A}(x_k)$.

3. Löse

   $$x'_{k+1} := \operatorname{argmin}\left\{ f(x) \;\middle|\; x \in \mathbf{R}^n,\ g_{\mathcal{A}_k}(x) = 0,\ h(x) = 0 \right\}$$

   mittels QR-Zerlegung wie oben angegeben. μ sei der Multiplikator zu h, λ zu $g_{\mathcal{A}_k}$.

4. Führe die folgende Fallunterscheidung durch:

   (a) Falls $x'_{k+1} = x_k$ und $\lambda \geqslant 0$,
   STOP.

   (b) Falls $x'_{k+1} \neq x_k$,
   setze $\sigma_k := \max\left\{ \sigma \;\middle|\; 0 \leqslant \sigma \leqslant 1,\ \sigma x'_{k+1} + (1-\sigma)x_k \in P \right\}$,
   setze $x_{k+1} := x_k + \sigma_k(x'_{k+1} - x_k)$,
   setze $k := k + 1$,
   gehe zu 2.

   (c) Anderenfalls ($x'_{k+1} = x_k$, $\lambda_i := \min \lambda_j < 0$)
   setze $\mathcal{A}_{k+1} := \mathcal{A}_k \setminus \{i\}$,
   setze $x_{k+1} := x_k$,
   setze $k := k + 1$,
   gehe zu 3.

Lemma 3.2.2. Die aktiven Restriktionen in allen $x \in P$ seien linear unabhängig. Dann gilt in Schritt 4 von Algorithmus 3.2.1:

1. Im Fall (b) ist $x'_{k+1} - x_k$ eine Abstiegsrichtung.

2. Im Fall (c) findet tritt in der nächsten Iteration Fall (b) ein und es gilt $\sigma_{k+1} > 0$.

3. Im Fall (a) ist x_k optimal.

Beweis. Die Fälle werden getrennt bewiesen.

1. Wegen der Gültigkeit der Multiplikatorregel in x'_{k+1} ist mit $d_k := x'_{k+1} - x_k$

$$\nabla f(x_k) = x_k^T A - b^T = -d_k^T A + \underbrace{{x'_{k+1}}^T A - b^T}_{=\nabla f(x'_{k+1})} = -d_k^T A - \lambda^T G_{\mathcal{A}_k} - \mu^T H,$$

also folgt

$$\nabla f(x_k) d_k = -d_k^T A d_k - \lambda^T \underbrace{G_{\mathcal{A}_k} d_k}_{=0} - \mu^T \underbrace{H d_k}_{=0} < 0,$$

weil A positiv definit und $d_k \neq 0$ ist.

2. Die Multiplikatoren in der $(k+1)$-ten Iteration seien mit $\bar{\mu}$ und $\bar{\lambda}$ bezeichnet. O. B. d. A. sei i der letzte Index in $\mathcal{A}_k$. Falls $d_{k+1} = x'_{k+2} - x_{k+1} = 0$, so folgt analog zum 1. Fall

$$-\nabla f(x_{k+1}) = 0^T A + \begin{pmatrix} \bar{\lambda} \\ 0 \end{pmatrix}^T G_{\mathcal{A}_k} + \bar{\mu}^T H.$$

Dies ist ein Widerspruch zu $\lambda_i < 0$, weil die Multiplikatoren aufgrund der linearen Unabhängigkeit der aktiven Gradienten eindeutig bestimmt sind.

Wir müssen nun noch zeigen, daß $\sigma_{k+1} > 0$ gilt. Nach Konstruktion ist $H d_{k+1} = 0$ und $G_{\mathcal{A}_{k+1}} d_{k+1} = 0$. Zu zeigen ist also $G_i d_{k+1} \leqslant 0$. Dies folgt aber wegen $x_{k+1} = x_k$ und $\lambda_i < 0$ sofort aus

$$0 > \nabla f(x_{k+1}) d_{k+1} = \nabla f(x_k) d_{k+1} = \Big(-\underbrace{d_k^T A}_{=0} - \lambda^T G_{\mathcal{A}_k} - \mu^T H \Big) d_{k+1} = -\lambda_i G_i d_{k+1}.$$

(Wie man leicht sieht, gilt im Fall (b) $\sigma_k > 0$ auch dann, wenn nicht Fall (c) vorausging.)

3. Weil im vorangegangen Schritt nicht Fall (c) eingetreten ist, gilt $\mathcal{A}_k = \mathcal{A}(x_k)$. Ferner erfüllt x_k die Kuhn-Tucker-Bedingungen (3.2) und (3.3), wenn die fehlenden Komponenten λ_j auf null gesetzt werden, und es gilt $\lambda \geqslant 0$, woraus Optimalität folgt.

$\square$

Satz 3.2.3. Der Algorithmus PQP bricht nach endlich vielen Schritten mit der eindeutig bestimmten Minimalstelle x_* von f auf P ab, falls die aktiven Restriktionen für alle $x \in P$ linear unabhängig sind.

Beweis. Es ist nur zu zeigen, daß Fall (b) nur endlich oft auftreten kann. Sei also $x'_{k+1} \neq x_k$. Jedesmal, wenn Fall (b) eintritt, gilt $f(x_{k+1}) < f(x_k)$, da $\sigma_k > 0$ und da in Richtung $d_k := x'_{k+1} - x_k$ die quadratische Funktion streng monoton fällt. Im Fall (c) kann $f(x_k)$ nicht wieder ansteigen.
Im Fall (b) gibt es zwei Möglichkeiten:

1. Es sei $\sigma_k = 1$. Dann wird in x_{k+1} das Minimum von f auf der Menge
$$\left\{x \;\middle|\; h(x) = 0,\; g_{\mathcal{A}_k}(x) = 0\right\}$$
angenommen. Also kann dieser Fall für jede mögliche Teilmenge $\mathcal{A}_k$ von $\{1, \dots, m\}$ nur einmal auftreten.

2. Es sei $\sigma_k < 1$. Dann gilt $|\mathcal{A}_{k+1}| > |\mathcal{A}_k|$, weil mindestens eine neue Restriktion aktiv wird. Dieser Fall kann in nicht mehr als $\min\{m, n\}$ aufeinanderfolgenden Schritten auftauchen.

Damit ist die Gesamtzahl aller Iterationen endlich. □

3.2.2 Lösung von quadratischen Problemen mit Innere-Punkte-Methoden

Ein quadratisches Optimierungsproblem liegt in **Standardform** vor, wenn es die folgende Gestalt hat:
$$\min\left\{\frac{1}{2}x^T Qx + c^T x \;\middle|\; \underbrace{Ax = b,\; x \geqslant 0}_{=P}\right\}, \tag{3.12}$$
wobei wir in den folgenden Abschnitten für $Q \in \mathbf{R}^{n\times n}$ nur positive Semidefinitheit fordern. Ferner sei rang$A$ maximal und die Slater-Bedingung sei erfüllt, d. h. es existiere ein $x_0 > 0$ mit $Ax_0 = b$. Die Kuhn-Tucker-Bedingungen (3.2) und (3.3) lauten für die Multiplikatoren $z \in \mathbf{R}^m_{\geqslant}$ und $y \in \mathbf{R}^m$
$$Qx + c - z - A^T y = 0 \tag{3.13}$$
und
$$z^T x = 0. \tag{3.14}$$
Weil Q nur semidefinit und damit nicht unbedingt invertierbar ist, können wir die zu (3.12) duale Aufgabe nicht explizit als Minimierung über z und y wie in (3.9) lösen, sondern erhalten hier
$$\begin{aligned}\Psi(z, y) =& \frac{1}{2}x_*^T Qx_* + c^T x_* - z^T x_* - y^T(Ax_* - b)\\ =& -\frac{1}{2}x_*^T Qx_* + y^T b + \Big(\underbrace{x^T Q + c^T - z^T - y^T A}_{=0 \text{ wegen } (3.13)}\Big)x_*,\end{aligned}$$
weshalb die zu (3.12) duale Aufgabe
$$\max\left\{-\frac{1}{2}x^T Qx + y^T b \;\middle|\; \underbrace{Qx - A^T y - z = -c,\; x \in \mathbf{R}^n,\; y \in \mathbf{R}^m,\; z \in \mathbf{R}^n_{\geqslant}}_{=D}\right\} \tag{3.15}$$
lautet. Wir fordern, daß die Slater-Bedingung für die duale Aufgabe ebenfalls erfüllt ist. Beide Aufgaben sind dann beschränkt, es gibt also einen Sattelpunkt (x_*, y_*, z_*), wobei x_* Optimallösung zu (3.12) ist.

3.2.3 Primale Innere-Punkte-Verfahren

Definition 3.2.1. Ein Verfahren zur Lösung der Aufgabe (3.12) heißt **primale Innere-Punkte-Methode**, wenn es eine Folge von primal zulässigen Vektoren $(x_k) \subseteq \mathbf{R}^n_>$ erzeugt.

Ein Beispiel für ein primales Innere-Punkte-Verfahren, welches zur Klasse der Affine-Scaling-Methoden gehört, ist der folgende Algorithmus (Dowling, 1992):

Algorithmus 3.2.2. (Primaler Affine-Scaling-Ansatz für quadratische Probleme)
Es sei $x_0 \in P$ mit $x_0 > 0$ und $\alpha \in [0, 1)$.
Für $k = 0, 1, \ldots$
  setze $X := \operatorname{diag}((x_k)_1, \ldots, (x_k)_n)$,
  setze $P_k := I - XA^T(AX^2A^T)^{-1}AX$,
  setze $d_k := -XP_kX\nabla f(x_k)^T$,
  falls $d_k = 0$
    STOP,
  setze $\alpha_k := \alpha \cdot \|X^{-1}d_k\|_2^{-1}$,
  setze $\lambda_k := \operatorname{argmin}\Big\{ f(x_k + td_k) \;\Big|\; t \in [0, \alpha_k] \Big\}$,
  setze $x_{k+1} := x_k + \lambda_k d_k$.

Bemerkung 3.2.1. Da $\operatorname{rang} A$ maximal ist, gilt für alle $b \in \mathbf{R}^m \setminus \{0\}$ die Aussage $A^T b \neq 0$ und somit ist $b^T AX^2A^T b > 0$, mit anderen Worten, die Matrix AX^2A^T ist positiv definit und somit auch invertierbar, denn wegen $x_k > 0$ ist X^2 positiv definit.

Wir können die folgenden Aussagen über den Algorithmus 3.2.2 machen:

Lemma 3.2.4. Es sei $x_k \in P$, $x_k > 0$, d_k wie in Algorithmus 3.2.2 beschrieben und $d_k \neq 0$. Dann gilt:

1. $x_{k+1} > 0$ und $x_{k+1} \in P$,
2. P_k ist symmetrisch und idempotent, d. h. $P_k^2 = P_k$,
3. d_k ist eine Abstiegsrichtung.

Beweis. Wir beweisen die Aussagen separat.

1. Zur Beweis der Aussage $x_{k+1} > 0$ müssen wir lediglich Komponenten i mit $(d_k)_i < 0$ betrachten. Nach Konstruktion von α_k gilt

$$\lambda_k \leqslant \alpha_k = \alpha \cdot \sqrt{\sum_{i=1}^{n} \left(\frac{(d_k)_i}{(x_k)_i} \right)^2}^{\,-1} \leqslant \alpha \frac{(x_k)_i}{(d_k)_i},$$

woraus wegen $\alpha < 1$ sofort $(x_{k+1})_i > 0$ folgt.

Nun ist noch zu zeigen, daß $Ad_k = 0$ gilt. Dies folgt aus

$$\begin{aligned} Ad_k &= -AX\Big(I - XA^T(AX^2A^T)^{-1}AX\Big)X\nabla f(x_k)^T \\ &= -\underbrace{\Big(AX - \underbrace{AX^2A^T(AX^2A^T)^{-1}}_{=I} AX \Big)}_{=0} X\nabla f(x_k)^T. \end{aligned}$$

2. Die Symmetrie von P_k ist klar.
   Es sei $\hat{P}_k := XA^T(AX^2A^T)^{-1}AX$. Damit gilt $P_k = I - \hat{P}_k$ und
   $$\hat{P}_k^2 = XA^T \underbrace{(AX^2A^T)^{-1}AX\ XA^T}_{=I}(AX^2A^T)^{-1}AX = XA^T(AX^2A^T)^{-1}AX = \hat{P}_k.$$
   Durch Ausmultiplizieren ergibt sich damit
   $$P_k^2 = (I - \hat{P}_k)(I - \hat{P}_k) = I - 2\hat{P}_k + \hat{P}_k^2 = I - \hat{P}_k = P_k.$$

3. Wegen $P_k = P_k^2 = P_k X X^{-2} X P_k$ gilt
   $$\begin{aligned}\nabla f(x_k)d_k &= -\nabla f(x_k)XP_kX\nabla f(x_k)^T \\ &= -\underbrace{\nabla f(x_k)XP_kX}_{=-d_k^T}X^{-2}\underbrace{XP_kX\nabla f(x_k)^T}_{=-d_k} \\ &= -d_k^TX^{-2}d_k < 0,\end{aligned}$$
   denn X^{-2} ist positiv definit und $d_k \neq 0$.

□

Lemma 3.2.5. Es sei $x_k \in P$, $x_k > 0$ und $d_k = 0$. Dann ist x_k globales Minimum von f auf P.

Beweis. Setze $\mu := -(AX^2A^T)^{-1}AX^2\nabla f(x_k)^T$. Dann gilt
$$\begin{aligned}X^2\nabla f(x_k)^T + X^2A^T\mu &= X^2\nabla f(x_k)^T - X^2A^T(AX^2A^T)^{-1}AX^2\nabla f(x_k)^T \\ &= X\Big(\underbrace{I - XA^T(AX^2A^T)^{-1}AX}_{=P_k}\Big)X\nabla f(x_k)^T = -d_k = 0,\end{aligned}$$
woraus sich nach Multiplikation von X^{-2} von links die Gültigkeit der Multiplikatorregel (3.2) in transponierter Form ergibt. Da es sich um ein konvexes Optimierungsproblem handelt, folgt hieraus nach Satz 3.1.13 die Optimalität von x_k. □

Aufgrund der vorigen Lemmata erzeugt Algorithmus 3.2.2 also eine Folge $(x_k) \subset P$ von inneren Punkten mit streng monoton abnehmenden Funktionswerten $f(x_k)$.

Definition 3.2.2. Das Problempaar (3.12) und (3.15) heißt **stark dual nichtentartet**, wenn es keinen Vektor $(x^T, y^T, z^T)^T \in D$ gibt, so daß z mehr als m Nichtnullkomponenten hat.

Es gilt der folgende Satz:

Satz 3.2.6. Es sei $x_0 \in P$, $x_0 > 0$ gegeben. Unter den zusätzlichen Voraussetzungen, daß

1. P beschränkt ist und
2. entweder Q positiv definit ist oder das Problem stark dual nichtentartet ist,

konvergiert die von Algorithmus 3.2.2 erzeugte Folge (x_k) gegen ein globales Minimum von f auf P.

Beweis. (Siehe [2].) □

Sind die Eingabedaten gut konditioniert, so kann die Berechnung der Abstiegsrichtung d_k wie folgt in $O(m^2 n)$ Schritten vorgenommen werden:
Es seien x_k, $\nabla f(x_k)$ und A gegeben, es sei $X := \operatorname{diag}((x_k)_1, \dots, (x_k)_n)$.

1. Berechne $H_1 := AX$. (Spalte i von A mit $(x_k)_i$ multiplizieren.) $O(nm)$
2. Berechne $H_2 := H_1 H_1^T$. (H_2 ist positiv definit, weil H_1 vollen Rang hat.) $O(m^2 n)$
3. Bestimme die wurzelfreie Cholesky-Zerlegung von H_2: $H_2 = LDL^T$. $O(m^3)$
4. Setze $a_1 := X \nabla f(x_k)^T$ und $a_2 := H_1 a_1 \in \mathbf{R}^m$. $O(nm)$
5. Löse $LDL^T a_3 = a_2$. (2 Dreiecksysteme) $O(m^2)$
6. Setze $a_4 := H_1^T a_3 - a_1$ und berechne $d_k := X a_4$. $O(nm)$

3.2.4 Primal-duale Innere-Punkte-Verfahren

Es sei

$$W = \left\{ w = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \,\middle|\, x \in P,\ w \in D,\ x > 0,\ z > 0 \right\} \neq \emptyset.$$

Weil w sowohl einen Vektor des primalen wie auch des dualen Zulässigkeitsbereiches definiert, kann w eine Dualitätslücke $g(w)$ zugeordnet werden, deren Wert

$$\begin{aligned} g(w) &:= \frac{1}{2} x^T Q x + c^T x - \left(-\frac{1}{2} x^T Q x + y^T b \right) \\ &= x^T Q x + c^T x - y^T b \\ &= \left(z^T + y^T A \right) x - y^T b \\ &= z^T x + y^T \Big(\underbrace{Ax - b}_{=0} \Big) > 0 \end{aligned}$$

ist, d. h. $g(w)$ ist gerade der Fehler in der Komplementaritätsbedingung (3.14). **Primal-duale Innere-Punkte-Verfahren** benutzen nur Vektoren aus W als Iterationspunkte und verkleinern iterativ die Dualitätslücke. Bei Konvergenz gegen ein $w_* \in \partial W$ mit $z_*^T x_* = 0$ ist die Optimallösung von (3.12) und (3.15) durch w_* gegeben, denn W_* ist die Lösung des Systems

$$H(w) := \begin{pmatrix} Ax - b \\ -Qx + A^T y + z - c \\ ZX\mathbf{1} \end{pmatrix} = 0,\ x \geqslant 0,\ z \geqslant 0, \tag{3.16}$$

wobei $X = \operatorname{diag}(x_1, \dots, x_n)$ und $Z = \operatorname{diag}(z_1, \dots, z_n)$.

Definition 3.2.3. Für $\tau > 0$ bezeichne $w_\tau = \begin{pmatrix} x_\tau \\ y_\tau \\ z_\tau \end{pmatrix} \in W$ einen Vektor mit der Eigenschaft $(x_\tau)_i (z_\tau)_i = \tau$ für $1 \leqslant i \leqslant n$. Dann heißt die Menge

$$\mathcal{C} := \left\{ w_\tau \in W \,\middle|\, \tau > 0 \right\}$$

zentraler Pfad.

Bemerkung 3.2.2. w_τ ist unter den gegebenen Voraussetzungen (d. h. wenn die Slater-Bedingung sowohl für (3.12) als auch für (3.15) erfüllt ist) durch $\tau > 0$ eindeutig bestimmt und Lösung von

$$H(w) = \begin{pmatrix} 0 \\ 0 \\ \tau\mathbf{1} \end{pmatrix}.$$

Der zentrale Pfad spielt eine wichtige Rolle bei der Berechnung einer Lösung von (3.16). Der Ansatz, (3.16) mit Hilfe des Newton-Verfahrens direkt zu lösen, ist zwar auch hier möglich, jedoch muß durch Wahl einer geeigneten Schrittweite gewährleistet sein, daß die Restriktionen $x \geqslant 0$ und $z \geqslant 0$ gültig bleiben.

Die sogenannten „Path-Following"-Methoden drehen die Suchrichtung in Richtung des zentralen Pfades, um die Positivität zu gewährleisten. Dies läßt sich etwa erreichen, indem die Zielfunktion derart verändert wird, daß sie für negative x und z sehr schlecht wird, etwa durch Addition eines sogenannten Barriereterms der Form

$$-\tau \sum_{i=1}^{n} \log x_i.$$

Wir erhalten also eine von x und τ abhängige modifizierte Zielfunktion g mit

$$g_\tau(x) = \frac{1}{2} x^T Q x + c^T x - \tau \sum_{i=1}^{n} \log x_i$$

und den Ableitungen

$$\nabla g_\tau(x) = x^T Q + c^T - \tau \mathbf{1}^T X^{-1} \text{ sowie } G_\tau(x) = Q + \tau X^{-2}.$$

Die zweite Ableitung ist offensichtlich für alle $\tau > 0$ positiv definit, d. h. g_τ ist streng konvex. Wir zeigen nun, daß die Einführung des Barriereterms in der Tat die Positivität von x gewährleistet.

Lemma 3.2.7. Es sei $\tau > 0$. Dann wird das eindeutig bestimmte globale Minimum $x_*(\tau)$ von g_τ auf P in einem $x_* > 0$ angenommen.

Ist $x \in P \cap \mathbf{R}^n_>$ und d eine Abstiegsrichtung mit $Ad = 0$, so gibt es ein $\lambda > 0$ mit $x + \lambda d > 0$ und $\nabla g_\tau(x + \lambda d)d = 0$, d. h. $x + \lambda d$ ist globales Minimum in Richtung d.

Beweis. Es sei $x_*(\tau)$ das Minimum von g_τ auf P und $(x_k) \subset P \cap \mathbf{R}^n_>$ eine gegen x_* konvergente Folge mit $g_\tau(x_{k+1}) < g_\tau(x_k)$. Falls $(x_*)_i = 0$, so würde der Barriereterm für $k \to \infty$ unbeschränkt wachsen. Dies ist ein Widerspruch, also gilt $x_*(\tau) > 0$.

Sei nun $x \in P \cap \mathbf{R}^n_>$ und d eine Abstiegsrichtung mit $Ad = 0$. Dann gilt

$$\begin{aligned} \nabla g_\tau(x + \lambda d)d &= (x + \lambda d)^T Q d + c^T d - \tau \left(\frac{1}{x_1 + \lambda d_1}, \ldots, \frac{1}{x_n + \lambda d_n} \right) d \\ &= x^T Q d + c^T d + \lambda \underbrace{d^T Q d}_{\geqslant 0} - \tau \sum_{i=1}^{n} \frac{d_i}{x_i + \lambda d_i}. \end{aligned}$$

Für alle i mit $d_i < 0$ geht der obige Ausdruck für $\lambda \to \frac{x_i}{-d_i}$ gegen positiv unendlich, für $\lambda = 0$ ist er hingegen negativ, weil d Abstiegsrichtung ist. Laut Zwischenwertsatz gibt es also ein $\bar\lambda$, so daß $\nabla g_\tau(x + \bar\lambda d)d = 0$. Nach Konstruktion gilt ferner $x + \bar\lambda d > 0$.

Sei nun $d \geqslant 0$. Es gibt Dualvariablen $y \in \mathbf{R}^m$ und $z \in \mathbf{R}^n_>$ mit $0 < z^T = x^T Q + c^T - y^T A$, also $(x^T, y^T, z^T)^T \in W$. Multiplikation mit d ergibt nun

$$0 < z^T d = x^T Q d + c^T d - y^T \underbrace{Ad}_{=0}.$$

Also gilt

$$\nabla g_\tau(x + \lambda d)d = \underbrace{x^T Q d + c^T d}_{>0} + \underbrace{\lambda d^T Q d}_{\geqslant 0} - \underbrace{\tau \sum_{i=1}^{n} \frac{d_i}{x_i + \lambda d_i}}_{>0},$$

wobei der letzte Terme mit wachsendem λ gegen null strebt. Deshalb wird auch in diesem Fall für ein endliches $\bar\lambda > 0$ der Wert 0 angenommen, also ist $x + \bar\lambda d$ lokales Minimum in Richtung d. Weil g_τ konvex ist, ist $x + \bar\lambda d$ sogar globales Minimum. □

Im allgemeinen kann auch eine Kombination aus Schrittweitensteuerung und Addition einer Barrierefunktion benutzt werden. Die allgemeine Form der primal-dualen Innere-Punkte-Methoden zur Lösung von (3.12) ist:

Algorithmus 3.2.3. (Allgemeine Form der primal-dualen Innere-Punkte-Methoden)
Es seien $\sigma_k \in [0, 1]$ für $k \in \mathbf{N}_0$ vorgegebene Verfahrensparameter.
Es sei $w_0 \in W$ gegeben.
Für $k = 0, 1, \ldots$

wähle $\tau_k > 0$, z. B. $\tau_k := \frac{1}{n} z_k^T x_k$,
berechne die Newton-Richtung $d_k = (\Delta x^T, \Delta y^T, \Delta z^T)^T$ zur Lösung des Systems

$$H(w) = \begin{pmatrix} 0 \\ 0 \\ \sigma_k \tau_k \end{pmatrix}$$

durch Lösung von

$$\begin{pmatrix} A & 0 & 0 \\ -Q & A^T & I \\ Z_k & 0 & X_k \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -Z_k X_k \mathbf{1} + \sigma_k \tau_k \mathbf{1} \end{pmatrix},$$

berechne eine Schrittweite λ_k derart, daß

$$w_{k+1} := w_k + \lambda_k d_k$$

positiv in den x- und z-Komponenten ist.

Die σ_k sind die Zentrierparameter des Verfahrens. Für $\sigma_k \equiv 0$ erhalten wir die affine Variante des Verfahrens, je größer σ_k wird, umso mehr werden die Suchrichtungen in das Innere von W gedrückt.

Lemma 3.2.8. Es sei $w_k \in W$. Dann ist bei Wahl von d_k nach Algorithmus 3.2.3 auch $w_{k+1} \in W$.

Beweis. (übungsaufgabe) □

Die folgende Variante des Verfahrens orientiert sich stark am zentralen Pfad ($\sigma_k \equiv 1$) und kommt ohne Schrittweitensteuerung aus (Monteiro, Adler, 1989):

Algorithmus 3.2.4. (Path-Following-Variante der primal-dualen Innere-Punkte-Methode)
Gegeben seien Verfahrensparameter $\vartheta := 0.1$, $\delta := 0.1$ und $\varepsilon > 0$.
Es sei $\tau_0 > 0$ und $w_0 \in W$ mit $\|Z_0 X_0 \mathbf{1} - \tau_0 \mathbf{1}\| \leqslant \vartheta \tau_0$.
Setze $k := 0$.
Solange $z_k^T x_k > \varepsilon$
  setze $\tau_{k+1} := \tau_k \left(1 - \frac{\delta}{\sqrt{n}}\right)$,
  berechne d_k durch Lösung von

$$\nabla H(w_k) d_k = \begin{pmatrix} 0 \\ 0 \\ -Z_k X_k \mathbf{1} + \tau_{k+1} \mathbf{1} \end{pmatrix},$$

  setze $w_{k+1} := w_k + d_k$,
  setze $k := k + 1$.

Dann gilt der folgende Satz:

Satz 3.2.9. Für alle $k \in \mathbf{N}$ gilt

1. $w_k \in W$,
2. $$\begin{aligned} \|X_k Z_k \mathbf{1} - \tau_k \mathbf{1}\| &\leqslant \vartheta \tau_k, \\ \|X_k^{-1}(x_{k+1} - x_k)\| &\leqslant 0.28, \\ \|Z_k^{-1}(z_{k+1} - z_k)\| &\leqslant 0.28, \end{aligned}$$
3. $g(w_k) = x_k^T z_k \leqslant 1.1 n \tau_k$.

Beweis. (Siehe [6].) □

Bemerkung 3.2.3. Die maximale Schrittzahl, die Algorithmus 3.2.4 durchführt, ist nach Satz 3.2.9 $\log\left(\frac{1.1 n \tau_0}{\varepsilon}\right) \frac{\sqrt{n}}{\delta} + 1$. Bei ganzzahligen Eingabedaten der Gesamtgröße L zeigen Monteiro und Adler, daß eine Optimallösung in $O(\sqrt{n} L)$ Schritten gefunden werden kann. Der Gesamtaufwand bei Newtonschritten ist damit $O(n^{3.5} L)$; mit einer modifizierten Methode kann dies auf $O(n^3 L)$ heruntergedrückt werden.

3.3 Exakte nichtdifferenzierbare Penalty-Funktionen

Die Idee der Optimierung durch Penalty-Funktionen ist es, durch Änderung der Zielfunktion eine Optimierungsaufgabe mit Restriktionen in eine Optimierungsaufgabe ohne Restriktionen zu verwandeln.

Definition 3.3.1. Es sei $\alpha \in \mathbf{R}$. Dann heißt $\alpha^+ := \max\{0, \alpha\}$ **positiver Teil** und $\alpha^- := \min\{0, \alpha\}$ **negativer Teil** von α.

Es sei eine Optimierungsaufgabe (3.1) gegeben. Für $b \in \mathbf{R}^m_\geqslant$ und $c \in \mathbf{R}^p_\geqslant$ sei

$$\Phi(x; b, c) := f(x) + b^T g^+(x) + c^T |h(x)|. \tag{3.17}$$

Bei welcher festen Wahl von b und c wird das Minimum von (3.1) an der gleichen Stelle x_* wie das Minimum von $\Phi(x; b, c)$ angenommen?

Beispiel 3.3.1. Es seien $\alpha, \beta \in \mathbf{R}$ mit $\alpha < \beta$ und $g(x) := (x-\alpha)(x-\beta)$. Dann ist $g(x) \leqslant 0$ genau dann, wenn $x \in [\alpha, \beta]$. Für $\mu \in \mathbf{R}_\geqslant$ setze

$$\Phi(x; \mu) := \mu f(x) + g^+(x) = \mu f(x) + \max\Big\{0,\ (x-\alpha)(x-\beta)\Big\}.$$

Damit ist $\Phi|_{[\alpha,\beta]} = \mu f|_{[\alpha,\beta]}$, also ist das lokale Minimum von f auf (α, β) auch das Minimum von Φ.
Sei $x_* \in \{\alpha, \beta\}$ lokales Minimum von f auf $[\alpha, \beta]$, aber nicht auf $\mathbf{R}$, z. B. $x_* = \alpha$ und $f'(x_*) > 0$. Dann folgt durch Taylorentwicklung von f um α für $x \in [x_* - \varepsilon, x_*]$

$$\begin{aligned}
\Phi(x;\mu) &= \mu f(x) + (x-\alpha)(x-\beta) \\
&= \mu f(\alpha) + \mu f'(\alpha)(x-\alpha) + \frac{\mu}{2} f''(\alpha + \vartheta(x-\alpha))(x-\alpha)^2 + (x-\alpha)(x-\beta) \\
&= \mu f(\alpha) + \underbrace{(\alpha - x)}_{\geqslant 0}\Big((\beta - x) - \mu f'(\alpha) + \frac{\mu}{2} f''(\alpha + \vartheta(x-\alpha))(\alpha - x)\Big) \\
&\geqslant \Phi(\alpha;\mu) + (\alpha - x)\Big((\beta - x) - \mu\Big(f'(\alpha) + \frac{M}{2}\varepsilon\Big)\Big),
\end{aligned}$$

wobei $|f''(x)| \leqslant M$, d. h. falls $0 < \mu \leqslant \frac{\beta - x}{f'(\alpha) + \frac{M}{2}\varepsilon}$, so ist x_* lokales Minimum von $\Phi(\cdot;\mu)$.

Die Gewichtung mit Vektoren b und c skaliert die Komponenten von g und h im Ansatz (3.17). Der Ansatz kann noch allgemeiner formuliert werden: Gegeben sei eine Norm $\|\cdot\|_v$ und eine differenzierbare Funktion $Q_v \colon \mathbf{R}_\geqslant \to \mathbf{R}_\geqslant$ mit

$$Q_v(0) = 0,\ Q_v'(0) > 0,\ Q_v(\tau) > 0 \text{ für alle } \tau > 0$$

und

$$\Phi(x;\mu) := f(x) + \mu Q_v(\|g^+(x),\ h(x)\|_v).$$

Ansatz (3.17) ergibt sich bei Verwendung der skalierten Betragssummennorm:

$$\|x\|_v := \|\mathrm{diag}(b^T,\ c^T)x\|_1.$$

Die Maximumnorm $\|\cdot\|_v := \|\cdot\|_\infty$ liefert hingegen

$$\Phi(x;\alpha) = f(x) + \alpha \max\left\{g_1^+(x), \ldots, g_m^+(x), |h_1(x)|, \ldots, |h_p(x)|\right\}.$$

Wegen der Äquivalenz der Normen handelt es sich um theoretisch gleichwertige Ansätze, was auch durch das folgende Lemma quantitativ ausgedrückt wird.

Lemma 3.3.1. Es seien durch $\|\cdot\|_v$, Q_v, $\|\cdot\|_w$, Q_w zwei Penalty-Funktionen des allgemeinen Typs definiert, etwa Φ_v und Φ_w. Es sei $\bar{x} \in P$. Ist $\bar{x}$ lokales Minimum von $\Phi_v(x;\alpha)$ für alle $\alpha \geqslant \alpha_v$, so ist $\bar{x}$ lokales Minimum von $\Phi_w(x;\alpha)$ für alle $\alpha \geqslant \alpha_w$, wobei

$$\alpha_w := \frac{(1+\varepsilon)Q_v'(0)}{(1-\varepsilon)Q_w'(0)} \alpha_v \gamma_{wv}^{-1}$$

für beliebiges $\varepsilon \in (0,1)$ und Normäquivalenzkonstante

$$\gamma_{wv} := \min\left\{ \frac{\|x\|_w}{\|x\|_v} \;\middle|\; x \in \mathbf{R}^{m+p} \setminus \{0\} \right\} > 0$$

ist.

Beweis. Ist $\bar{x}$ lokales Minimum von Φ_v für alle $\alpha \geqslant \alpha_v$, dann gilt

$$\Phi_v(\bar{x};\alpha) \leqslant \Phi_v(x;\alpha) \text{ für alle } x \in U_\delta(\bar{x}) \text{ und alle } \alpha \geqslant \alpha_v.$$

Sei $\varepsilon \in (0,1)$. Da $Q_v'(0) > 0$ und $Q_w'(0) > 0$, gilt

$$Q_v(t) \leqslant (1+\varepsilon)Q_v'(0)t \text{ für } t > 0 \text{ hinreichend klein}$$

und

$$Q_w(t) \geqslant (1-\varepsilon)Q_w'(0)t \text{ für } t > 0 \text{ hinreichend klein,}$$

etwa für $t \in [0,\hat{t}]$. Da $\bar{x}$ zulässig ist, gibt es ein $\delta_1 > 0$ mit $\delta_1 \leqslant \delta$ und

$$\|g^+(x), h(x)\|_v \leqslant \hat{t} \text{ und } \|(g^+(x), h(x))\|_w \leqslant \hat{t} \text{ für alle } x \in U_{\delta_1}(\bar{x}).$$

Sei $\alpha \geqslant \alpha_w$ und $x \in U_{\delta_1}(\bar{x})$. Dann gilt

$$\begin{aligned}
\Phi_w(x;\alpha) &= f(x) + \alpha Q_w(\|g^+(x), h(x)\|_w) \\
&\geqslant f(x) + \alpha(1-\varepsilon)Q_w'(0)\|g^+(x), h(x)\|_w \\
&\geqslant f(x) + \alpha_w(1-\varepsilon)Q_w'(0)\gamma_{wv}\|g^+(x), h(x)\|_v \\
&\geqslant f(x) + \alpha_v(1+\varepsilon)Q_v'(0)\|g^+(x), h(x)\|_v \\
&\geqslant f(x) + \alpha_v Q_v(\|g^+(x), h(x)\|_v) \\
&\geqslant \Phi_v(x;\alpha_v) \geqslant \Phi_v(\bar{x};\alpha_v) \\
&= f(\bar{x}) + \alpha_v \underbrace{Q_v(\|g^+(\bar{x}), h(\bar{x})\|)}_{=0,\ \text{da } \bar{x}\in P} \\
&= f(\bar{x}) = \Phi_w(\bar{x};\alpha),
\end{aligned}$$

was zu zeigen war. $\square$

Lemma 3.3.2. Es sei $\bar{x} \in P$ und $\alpha \geqslant 0$. Es sei $\bar{x}$ lokales Minimum von Φ_v, d. h. für ein $\delta > 0$ sei

$$\Phi_v(x;\alpha) \geqslant \Phi_v(\bar{x};\alpha) \text{ für alle } x \in U_\delta(\bar{x}).$$

Dann ist $\bar{x}$ lokales Minimum von f auf P, d. h. es gilt $f(x) \geqslant f(\bar{x})$ für alle $x \in U_\delta(\bar{x}) \cap P$.

Beweis. Falls $x \in P$, dann gilt $\|(g^+(x), h(x))\| = 0$, also folgt

$$f(x) = \Phi_v(x;\alpha) \geqslant \Phi_v(\bar{x};\alpha) = f(\bar{x}).$$

□

3.3.1 Die erweiterte Mangasarian-Fromowitz-Bedingung

Definition 3.3.2. Für $\tau \geqslant 0$ sei

$$P(\tau) := \left\{ x \,\middle|\, g(x) \leqslant \tau \mathbf{1},\ |h(x)| \leqslant \tau \mathbf{1} \right\} \text{ und } \mathcal{V}(x) := \left\{ i \,\middle|\, g_i(x) > 0 \right\}.$$

Dann erfüllt $x \in P(\tau)$ die **erweiterte Mangasarian-Fromowitz-Bedingung**, wenn es ein $z \in \mathbf{R}^n$ gibt mit

1. $\nabla g_i(x) z < 0$ für alle $i \in \mathcal{A}(x) \cup \mathcal{V}(x)$,
2. $\nabla h(x) z = 0$,
3. $\nabla h(x)$ hat maximalen Rang.

Für ein gegebenes Problem ist die erweiterte Mangasarian-Fromowitz-Bedingung erfüllt, wenn es ein $\tau_0 > 0$ gibt, so daß die erweiterte Mangasarian-Fromowitz-Bedingung für alle $x \in P(\tau_0)$ erfüllt ist.

Bemerkung 3.3.1. Für $x \in P$ ist wegen $\mathcal{V}(x) = \emptyset$ die erweiterte Mangasarian-Fromowitz-Bedingung genau dann erfüllt wenn die herkömmliche Mangasarian-Fromowitz-Bedingung erfüllt ist.

Lemma 3.3.3. Es sei $\tau_0 > 0$. Dann sind die folgenden Aussagen äquivalent:

1. Für alle $\bar{x} \in P(\tau_0)$ ist die erweiterte Mangasarian-Fromowitz-Bedingung erfüllt.
2. Für alle $\bar{x} \in P(\tau_0)$ gilt: Ist $b\colon U_\varepsilon(\bar{x}) \to \mathbf{R}^p$ beschränkt, so existiert ein $\varepsilon' \in (0, \varepsilon]$ und $d\colon U_{\varepsilon'}(\bar{x}) \to \mathbf{R}^n$ mit

$$\nabla g_{\mathcal{A}(\bar{x}) \cup \mathcal{V}(\bar{x})}(x) d(x) \leqslant -\mathbf{1} \text{ und } \nabla h(x) d(x) = b(x) \text{ für alle } x \in U_{\varepsilon'}(\bar{x}).$$

Beweis. Die erweiterte Mangasarian-Fromowitz-Bedingung sei für $\tau_0 > 0$ und alle $\bar{x} \in P(\tau_0)$ erfüllt. Aus der Rangbedingung für ∇h folgt die Existenz von stetigen $Q, R\colon P(\tau_0) \to \mathbf{R}^{p \times p}$ mit

$$Q(x) \nabla h(x)^T = \begin{pmatrix} R(x) \\ 0 \end{pmatrix} \text{ für alle } x \in P(\tau_0).$$

Definiere die stetigen Funktionen $G_1\colon \mathbf{R}^n \to \mathbf{R}^p$ und $G_2\colon \mathbf{R}^n \to \mathbf{R}^{n-p}$ durch

$$\begin{pmatrix} G_1(x) \\ G_2(x) \end{pmatrix} := Q(x)\nabla g_{\mathcal{A}(\bar{x})\cup\mathcal{V}(\bar{x})}(x)^T.$$

Es sei $\bar{x} \in P(\tau_0)$ und $\bar{z} \in \mathbf{R}^n$ wie in der erweiterten Mangasarian-Fromowitz-Bedingung gefordert. Setze

$$\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} := Q(\bar{x})\bar{z} \text{ mit } s_1 \in \mathbf{R}^p \text{ und } s_2 \in \mathbf{R}^{n-p}.$$

Wegen $0 = \nabla h(\bar{x})\bar{z} = R(\bar{x})^T s_1$ muß $s_1 = 0$ aufgrund der Regularität von $R(\bar{x})$ gelten. Nun folgt wegen $\nabla g_{\mathcal{A}(\bar{x})\cup\mathcal{V}(\bar{x})}(x) = \Big(G_1(x)^T,\, G_2(x)^T\Big)Q(x)$

$$0 > \nabla g_{\mathcal{A}(\bar{x})\cup\mathcal{V}(\bar{x})}(\bar{x})\bar{z} = \Big(G_1(\bar{x})^T,\, G_2(\bar{x})^T\Big)\underbrace{Q(\bar{x})\,Q(\bar{x})^T}_{=I}\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = G_2(\bar{x})^T s_2.$$

Da alle vorkommenden Funktionen stetig sind, gibt es ein $\varepsilon'' > 0$ und ein $\beta > 0$ mit

$$G_2(x)^T s_2 \leqslant -\beta\mathbf{1} \text{ für alle } x \in U_{\varepsilon''}(\bar{x}).$$

Sei nun $b\colon U_\varepsilon(\bar{x}) \to \mathbf{R}^p$ eine beschränkte Funktion. Dann gilt mit $\varepsilon' := \min\{\varepsilon, \varepsilon''\}$

$$|G_1(x)^T R(x)^{-T} b(x)| \leqslant \gamma\mathbf{1} \text{ für alle } x \in U_{\varepsilon'}(\bar{x}) \text{ und ein festes } \gamma > 0.$$

Setze nun

$$d(x) := Q(x)^T\begin{pmatrix} R(x)^{-T} b(x) \\ \frac{\gamma+1}{\beta} s_2 \end{pmatrix}.$$

Die so konstruierte Funktion d ist offensichtlich beschränkt auf $U_{\varepsilon'}(\bar{x})$. Sei $x \in U_{\varepsilon'}(\bar{x})$. Dann ist

$$\nabla h(x)d(x) = \Big(R(x)^T,\, 0^T\Big)Q(x)d(x) = b(x)$$

und

$$\begin{aligned}\nabla g_{\mathcal{A}(\bar{x})\cup\mathcal{V}(\bar{x})}(x)d(x) &= \Big(G_1(x)^T,\, G_2(x)^T\Big)Q(x)d(x) \\ &= G_1(x)^T R(x)^{-T} b(x) + G_2(x)^T s_2\frac{\gamma+1}{\beta} \\ &\leqslant \gamma\mathbf{1} - \beta\mathbf{1}\frac{\gamma+1}{\beta} = -\mathbf{1}.\end{aligned}$$

Beweis der Umkehrung: Die Lösbarkeit von $\nabla h(x)d(x) = b(x)$ für alle b liefert die Rangbedingung. Es sei d_0 die Funktion zur Funktion $b_0 = 0$. Die Wahl von $\bar{z} := d_0(\bar{x})$ erfüllt nun alle Forderungen der erweiterten Mangasarian-Fromowitz-Bedingung. □

Satz 3.3.4. Es sei $\tau_0 > 0$ und $P(\tau_0)$ kompakt. Die erweiterte Mangasarian-Fromowitz-Bedingung gelte für alle $x \in P(\tau_0)$. Es sei $Q_v \colon \mathbf{R}_\geqslant \to \mathbf{R}_\geqslant$ stetig differenzierbar mit

$$0 < \mu_1 := \inf_{\zeta \in \mathbf{R}_\geqslant} Q_v'(\zeta) < \sup_{\zeta \in \mathbf{R}_\geqslant} Q_v'(\zeta) =: \mu_2.$$

Ferner gebe es eine Konstante $k_v > 0$, so daß

$$\|z\|_v \leqslant \|y\|_v - k_v \|y - z\|_v \text{ für alle } 0 \leqslant z \leqslant y \text{ mit } z_i < y_i \text{ falls } y_i \neq 0$$

gilt. Dann gibt es ein $\alpha_v > 0$, so daß $\Phi_v(x;\alpha)$ für alle $\alpha \geqslant \alpha_v$ kein lokales Minimum auf der Menge $P(\tau_0) \setminus P$ besitzt.

Beweis. (Siehe [7], Satz 3.6.4; Seite 466–468.) $\square$

Bemerkung 3.3.2. Die in Satz 3.3.4 benötigte Konstante k_v kann man z. B. für die Betragssummennorm bestimmen ($k_1 = 1$ ist möglich), aber für die Maximumnorm existiert keine, denn die Wahl von

$$z_\varepsilon := \begin{pmatrix} 0 \\ 1 - \varepsilon \end{pmatrix} \leqslant \begin{pmatrix} 1 \\ 1 \end{pmatrix} =: y$$

führt auf die Bedingung

$$1 - \varepsilon \leqslant 1 - k_\infty \text{ für alle } \varepsilon > 0,$$

welche nur für $k_\infty \leqslant 0$ erfüllbar ist.

Satz 3.3.5. Es sei die erweiterte Mangasarian-Fromowitz-Bedingung für $P(\tau_0)$ erfüllt. Für $x_* \in P$ seien die hinreichenden Bedingungen 2. Ordnung mit Multiplikatoren $\lambda_* \in \mathbf{R}^m_\geqslant$ und $\mu_* \in \mathbf{R}^p$ erfüllt (siehe Satz 3.1.8).
Dann gilt: Falls $b > \lambda_*$ und $c > |\mu_*|$, so ist x_* strenges lokales Minimum von $\Phi(\,\cdot\,; b, c)$.

Beweis. Es sei $b > \lambda_*$ und $c > |\mu_*|$. Nach Satz 3.1.8 ist x_* strenges lokales Minimum von f auf P; also gilt für ein $\delta > 0$

$$\Phi(x; b, c) = f(x) > f(x_*) = \Phi(x_*; b, c) \text{ für alle } x \in P \cap U_\delta(x_*).$$

Der Beweis wird indirekt geführt. Annahme: Es existiert eine gegen x_* konvergente Folge $(x_k) \subseteq U_\delta(x_*) \setminus P$ mit

$$\Phi(x_k; b, c) \leqslant \Phi(x_*; b, c), \text{ wobei } \mathcal{V}(x_k) \neq \emptyset \text{ oder } h(x_k) \neq 0 \text{ für alle } k \in \mathbf{N}.$$

Wir wählen o. B. d. A. eine Teilfolge aus, für die alle $\mathcal{V}(x_k)$ identisch sind, etwa $\mathcal{V}$. (Dies ist möglich, weil es nur endlich viele (nämlich 2^m) Möglichkeiten für $\mathcal{V}(x_k)$ gibt.)
Weiterhin sei durch Auswahl von Teilfolgen o. B. d. A. $\left(\frac{x_k - x_*}{\|x_k - x_*\|}\right) \to s$ und $\operatorname{sgn}(h(x_k)) = \sigma$. Nun gilt

$$\Phi(x_k; b, c) = f(x_k) + \sum_{i \in \mathcal{V}} b_i g_i(x_k) + \sum_{j=1}^{p} c_j \sigma_j h_j(x_k).$$

Mittels Taylorentwicklung folgt aus der Annahme

$$
\begin{aligned}
0 &\geqslant \Phi(x_k;b,c) - \Phi(x_*;b,c) \\
&= \nabla f(x_*)(x_k - x_*) + \sum_{i\in\mathcal{V}} b_i \nabla g_i(x_*)(x_k - x_*) + \sum_{j=1}^{p} c_j \sigma_j \nabla h_j(x_*)(x_k - x_*) + O(\|x_k - x_*\|^2)
\end{aligned}
$$

Division durch $\|x_k - x_*\|$ und nachfolgender Grenzübergang $k \to \infty$ liefert

$$
0 \geqslant \nabla f(x_*)s + \sum_{i\in\mathcal{V}} b_i \nabla g_i(x_*)s + \sum_{j=1}^{p} c_j \sigma_j \nabla h_j(x_*)s.
$$

Es ist $\mathcal{V} \subseteq \mathcal{A}(x_*)$. Ferner gilt $\nabla g_{\mathcal{A}(x_*)\setminus\mathcal{V}}(x_*)s \leqslant 0$. Nach Multiplikatorregel (3.2) gilt:

$$
\begin{aligned}
0 &= \nabla f(x_*)s + \lambda_*^T \nabla g(x_*)s + \mu_*^T \nabla h(x_*)s \\
&= \nabla f(x_*)s + \sum_{i\in\mathcal{A}(x_*)} (\lambda_*)_i \nabla g_i(x_*)s + \sum_{j=1}^{p} (\mu_*)_j \nabla h_j(x_*)s.
\end{aligned}
$$

Subtraktion liefert nun mit $\tilde{\sigma} := \operatorname{sgn}(\nabla h(x_*)s)$

$$
0 \geqslant \sum_{i\in\mathcal{A}(x_*)\setminus\mathcal{V}} \underbrace{-(\lambda_*)_i \nabla g_i(x_*)s}_{\geqslant 0} + \sum_{i\in\mathcal{V}} \underbrace{(b_i - (\lambda_*)_i)}_{>0} \underbrace{\nabla g_i(x_*)s}_{\geqslant 0} + \sum_{j=1}^{p} (c_j\sigma_j\tilde{\sigma}_j - (\mu_*)_j\tilde{\sigma}_j) \underbrace{|\nabla h_j(x_*)s|}_{\geqslant 0}.
$$

Nach Konstruktion von σ gilt $\sigma_j h_j(x_k) \geqslant 0$, also folgt aus $\sigma_j h_j(x_*) = 0$ sofort $\sigma_j \nabla h_j(x_*)s \geqslant 0$, also $\sigma_j\tilde{\sigma}_j \in \{0,1\}$. Weil bei $\sigma_j\tilde{\sigma}_j = 0$ sofort $\tilde{\sigma}_j = 0$ folgt, ist damit ist die Nichtnegativität der letzten Koeffizienten gezeigt, folglich muß in obiger Ungleichung Gleichheit gelten.
Als Folgerung ergibt sich

$$
\nabla g_{\mathcal{V}}(x_*)s = 0 \text{ und } \nabla h(x_*)s = 0, \text{ sowie } \nabla g_i(x_*)s \text{ für } i \in \left\{ j \in \mathcal{A}(x_*) \;\middle|\; (\lambda_*)_j > 0 \right\}.
$$

Dies besagt genau, daß $s \in \hat{L}_0(x_*)$.
Die Taylorentwicklung von $\ell(x,\lambda_*,\mu_*)$ um x_* ergibt nun wegen $\nabla_x \ell(x_*,\lambda_*,\mu_*)s = 0$ für x_k nahe genug an x_*

$$
\begin{aligned}
\ell(x_k,\lambda_*,\mu_*) &= \ell(x_*,\lambda_*,\mu_*) + \frac{1}{2}\|x_* - x_k\|^2 \underbrace{s^T \mathcal{L}(x_*)s}_{>0} + O(\|x_k - x_*\|^3) \\
&> \ell(x_*,\lambda_*,\mu_*) = f(x_*) = \Phi(x_*;b,c).
\end{aligned}
$$

Aufgrund der Wahl von b und c gilt andererseits

$$
\begin{aligned}
\Phi(x_k;b,c) &= f(x_k) + b^T g^+(x_k) + c^T |h(x_k)| \\
&\geqslant f(x_k) + \lambda_*^T g(x_k) + \mu_*^T h(x_k) \\
&= \ell(x_k,\lambda_*,\mu_*), \\
&> \Phi(x_*;b,c)
\end{aligned}
$$

was einen Widerspruch zur Annahme über die Folge (x_k) darstellt. □

Wir werden nun eine Verallgemeinerung des letzten Satzes beweisen. Zunächst benötigen wir allerdings noch den Begriff der dualen Normen.

Definition 3.3.3. Es sei $\|\cdot\|$ eine Vektornorm auf dem $\mathbf{R}^n$. Dann ist die durch

$$\|x\|_D := \sup \left\{ y^T x \;\middle|\; \|y\| = 1 \right\}$$

definierte Norm die zu $\|\cdot\|$ **duale Norm**.
Erfüllt eine Norm $\|\cdot\|$ die Gleichung

$$\|\,|x|\,\| = \|x\| \text{ für alle } x \in \mathbf{R}^n,$$

so heißt sie **absolute Norm**.

Bemerkung 3.3.3. (Aussagen über duale Normen)

1. Die euklidische Vektornorm ist selbstdual.
2. Die Maximumnorm ist dual zur Betragssummennorm.
3. Ist A positiv definit, und ist $\|\cdot\|_A$ definiert durch

$$\|x\|_A := \sqrt{x^T A x},$$

   so ist die zu $\|\cdot\|_A$ duale Norm $\|\cdot\|_{A^{-1}}$.
4. Sind $\|\cdot\|_v$ und $\|\cdot\|_w$ duale Normen, dann gilt die verallgemeinerte Cauchy-Schwarz-Ungleichung

$$|x^T y| \leqslant \|x\|_v \cdot \|y\|_w \text{ für alle } x, y \in \mathbf{R}^n.$$

5. Ist $\|\cdot\|_v$ eine absolute Norm, dann ist auch die zu $\|\cdot\|_v$ duale Norm absolut.

Satz 3.3.6. Es gelte die erweiterte Mangasarian-Fromowitz-Bedingung für $P(\tau_0)$. Für $x_* \in P$ seien die hinreichenden Bedingungen 2. Ordnung mit Multiplikatoren $\lambda_* \in \mathbf{R}^m_{\geqslant}$ und $\mu_* \in \mathbf{R}^p$ erfüllt. Q sei konvex und differenzierbar auf $\mathbf{R}_{\geqslant}$ mit $Q(0) = 0$ und $Q'(0) > 0$. $\|\cdot\|_v$ sei eine absolute Vektornorm und $\Phi_v(x;\alpha) := f(x) + \alpha Q(\|g^+(x), h(x)\|_v)$. Es sei $\|\cdot\|_w$ die zu $\|\cdot\|_v$ duale Norm und $\bar{\alpha} := \frac{1}{Q'(0)}\|\lambda_*, \mu_*\|_w$.
Dann ist x_* strenges lokales Minimum von $\Phi_v(\,\cdot\,;\alpha)$, falls $\alpha > \bar{\alpha}$.

Beweis. Die Norm $\|\cdot\|_w$ ist absolut. Es gilt

$$\alpha Q'(0) > \bar{\alpha} Q'(0) = \|\lambda_*, \mu_*\|_w.$$

Wähle $b > \lambda_*$ und $c > |\mu_*|$ mit

$$\alpha Q'(0) > \|b, c\|_w > \|\lambda_*, \mu_*\|_w$$

(etwa $b := \sqrt{\frac{\alpha}{\bar{\alpha}}}\lambda_*$ und $c := \sqrt{\frac{\alpha}{\bar{\alpha}}}|\mu_*|$). Nach Satz 3.3.5 ist x_* strenges lokales Minimum von $\Phi(x; b, c)$.

Für $x \notin P$ ist $\|g^+(x), h(x)\|_v > 0$. Also folgt mit Zwischenwertsatz und Konvexität von Q für ein $\vartheta > 0$

$$Q(\|g^+(x), h(x)\|_v) = Q'(\vartheta)\|g^+(x), h(x)\|_v \geqslant Q'(0)\|g^+(x), h(x)\|_v.$$

Nun folgt für $x \in U_\delta(x_*) \setminus P$ unter Verwendung der verallgemeinerten Cauchy-Schwarz-Ungleichung

$$\begin{aligned}
\Phi_v(x;\alpha) &= f(x) + \alpha Q(\|g^+(x), h(x)\|_v) \\
&\geqslant f(x) + \alpha Q'(0)\|g^+(x), h(x)\|_v \\
&> f(x) + \|b, c\|_w \, \|g^+(x), h(x)\|_v \\
&= f(x) + \|b, c\|_w \, \|g^+(x), |h(x)|\|_v \\
&\geqslant f(x) + b^T g^+(x) + c^T |h(x)| \\
&= \Phi(x; b, c) > \Phi(x_*; b, c) = f(x_*) = \Phi_v(x_*; \alpha)
\end{aligned}$$

Für $x \in U_\delta(x_*) \cap P \setminus \{x_*\}$ gilt

$$\Phi_v(x;\alpha) = f(x) > f(x_*) = \Phi_v(x_*;\alpha),$$

weil x_* strenges lokales Minimum von f auf P ist.
Insgesamt ist damit gezeigt, daß x_* strenges lokales Minimum von $\Phi_v(\,\cdot\,;\alpha)$ ist. □

Bemerkung 3.3.4. Neben der Äquivalenz der Probleme

$$\min\left\{ f(x) \;\middle|\; x \in P \right\} \text{ und } \min\left\{ \Phi(x;\alpha) \;\middle|\; x \in \mathbf{R}^n \right\}$$

haben wir *berechenbare* Schranken für die Penalty-Parameter abgeleitet, die sich an den Lagrange-Multiplikatoren orientieren. Dies werden wir im folgenden zur Konstruktion eines Algorithmus ausnutzen.

3.3.2 Die Methode der sequentiellen quadratischen Optimierung

Der Methode der sequentiellen quadratischen Optimierung liegt die folgende Idee zugrunde: Konstruiere eine Folge $(x_k) \subseteq P(\tau_0)$ mit

$$\Phi(x_{k+1}; b, c) - \Phi(x_k; b, c) \leqslant -\Psi(\|g^+(x_k), h(x_k)\|) - \|\nabla \ell(x_k, \lambda_k, \mu_k)\| \leqslant 0, \tag{3.18}$$

wobei λ_k und μ_k Schätzungen der Multiplikatoren sind, die die Komplementaritätseigenschaft $\lambda_k^T g(x_k) = 0$ besitzen. Für die Funktion $\Psi\colon \mathbf{R}_\geqslant \to \mathbf{R}_\geqslant$ gelte, daß aus $(\Psi(t_k)) \to 0$ stets $(t_k) \to 0$ folgt.
Dann gilt

- Jeder Häufungspunkt von (x_k) ist zulässig, weil $P(\tau_0)$ kompakt ist, und in Häufungspunkten gilt die Multiplikatorregel.
- Die Penalty-Parameter b und c werden an λ und $|\mu|$ gekoppelt, z. B. sei für festes $\varepsilon > 0$

$$(b_{k+1})_i := \begin{cases} (\lambda_k)_i + 2\varepsilon & \text{falls } (\lambda_k)_i + \varepsilon \geqslant (b_k)_i, \\ (b_k)_i & \text{anderenfalls} \end{cases} \quad \text{für } i = 1, \ldots, m$$

$$(c_{k+1})_j := \begin{cases} |(\mu_k)_j| + 2\varepsilon & \text{falls } |(\mu_k)_j| + \varepsilon \geqslant (c_k)_j, \\ (c_k)_j & \text{anderenfalls} \end{cases} \quad \text{für } j = 1, \dots, p$$

Als Startwerte sind z. B. $b = \varepsilon\mathbf{1}$, $c = \varepsilon\mathbf{1}$ möglich.

Dieser Ansatz führt auf das *SQP*-Verfahren.

Quadratische Approximation $QP(\bar{x}, A)$ (Han)

Im Verfahren von Han wird die Zielfunktion des Optimierungsproblems lokal quadratisch approximiert und die Restriktionen linear. Es sei

$$\begin{aligned} f^k(x) &:= f(x_k) + \nabla f(x_k)(x - x_k) + \tfrac{1}{2}(x - x_k)^T A_k (x - x_k), \\ g^k(x) &:= g(x_k) + \nabla g(x_k)(x - x_k) \leqslant 0 \text{ und} \\ h^k(x) &:= h(x_k) + \nabla h(x_k)(x - x_k) = 0, \end{aligned}$$

wobei A_k positiv definit. Ist x_* eine Optimallösung zu $QP(x_k, A)$, so wähle die Abstiegsrichtung $d_k := x_* - x_k$ und bestimme eine geeignete Schrittweite.
Der folgende Algorithmus ist bereits eine Weiterentwicklung des Algorithmus von Han:

Algorithmus 3.3.1. (Grundversion des SQP-Algorithmus)
Eingabeparameter: $\varepsilon > 0$, $\delta, \alpha \in (0,1)$, $\rho_2 \gg 1 \gg \rho_1 > 0$, $\tau_0 > 0$.
Es sei $x_0 \in P(\tau_0)$.
Setze $b_0 := \varepsilon\mathbf{1}$ und $c_0 := \varepsilon\mathbf{1}$.
Für $k = 1, 2, \dots$

1. Wähle A_k symmetrisch mit $\rho_1 y^T y \leqslant y^T A_k y \leqslant \rho_2 y^T y$ für alle $y \in \mathbf{R}^n$.

2. Bestimme die Optimallösung $\hat{x}_{k+1}$ und Lagrangemultiplikatoren $(\hat{x}_{k+1}, \lambda_k, \mu_k)$ zum Problem QP(x_k, A_k). Setze $d_k := \hat{x}_{k+1} - x_k$.

3. Berechne das Update der Penalty-Parameter

$$(b_{k+1})_i := \begin{cases} (\lambda_k)_i + 2\varepsilon & \text{falls } (\lambda_k)_i + \varepsilon \geqslant (b_k)_i, \\ (b_k)_i & \text{anderenfalls} \end{cases} \quad \text{für } i = 1, \dots, m$$

$$(c_{k+1})_j := \begin{cases} |(\mu_k)_j| + 2\varepsilon & \text{falls } |(\mu_k)_j| + \varepsilon \geqslant (c_k)_j, \\ (c_k)_j & \text{anderenfalls} \end{cases} \quad \text{für } j = 1, \dots, p.$$

4. Bestimme $l_k \in \mathbf{N}_0$ minimal, so daß

$$\Phi(x_k + \alpha^{l_k} d_k; b_{k+1}, c_{k+1}) - \Phi(x_k; b_{k+1}, c_{k+1}) \leqslant -\delta\alpha^{l_k}\left[d_k^T A_k d_k + \varepsilon\|h(x_k)\|_1 + \varepsilon\|g^+(x_k)\|_1\right]$$

und im Falle $x_k \in P(\tau_0) \setminus P(\frac{1}{2}\tau_0)$ gleichzeitig

$$\|h(x_k + \alpha^{l_k} d_k)\|_1 + \|g^+(x_k + \alpha^{l_k} d_k)\|_1 - \|h(x_k)\|_1 - \|g^+(x_k)\|$$

$$\leqslant -\delta\alpha^{l_k}\left[\|h(x_k)\|_1 + \|g^+(x_k)\|_1\right].$$

5. Setze $x_{k+1} := x_k + \alpha^{l_k} d_k$.

Das Problem im Algorithmus 3.3.1 ist, daß die quadratische Approximation $QP(x_k, A)$ keine Lösung besitzen muß, wie das folgende Beispiel zeigt.

Beispiel 3.3.2. Es sei $x_0 = (2,2)^T$ und g gegeben durch

$$\begin{aligned}
g_1(x) &= x_1 + x_2 - 1 && \leqslant 0,\\
g_2(x) &= (x_1 - 2) + (x_2 - 2) - \tfrac{3}{2}(x_1 - 2)^2 + 5 && \leqslant 0,\\
g_3(x) &= -x_1 && \leqslant 0,\\
g_4(x) &= -x_2 && \leqslant 0.
\end{aligned}$$

Damit gilt

$$\nabla g(x) = \begin{pmatrix} 1 & 1 \\ 7 - 3x_1 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Die lineare Approximation von g in x_0 ist also

$$g^0(d) = g(x_0) + \nabla g(x_0) d = \begin{pmatrix} 3 + d_1 + d_2 \\ 5 + d_1 + d_2 \\ -2 - d_1 \\ -2 - d_2 \end{pmatrix},$$

hier erfüllt kein $d \in \mathbf{R}^2$ die Vorschrift $g^0(d) \leqslant 0$, wie man durch Addition der Ungleichungen sofort sieht. Das Ausgangsproblem hat aber den zulässigen Punkt $\bar{x} = (\frac{1}{2}, \frac{1}{2})^T$.

Ist jedoch $QP(x_k, A_k)$ lösbar, dann ist die in Algorithmus 3.3.1 bestimmte Suchrichtung d_k tatsächlich eine Abstiegsrichtung für die Penalty-Funktion, wie wir im folgenden zeigen werden. Hierzu benötigen wir den Begriff der einseitigen Richtungsableitung.

Definition 3.3.4. Es sei $f\colon \mathbf{R}^n \to \mathbf{R}$ und $d \in \mathbf{R}^n$, $d \neq 0$. Existiert der Grenzwert

$$\lim_{t \to 0, t > 0} \frac{f(x + td) - f(x)}{t},$$

dann ist f bei x **einseitig in Richtung** d **differenzierbar** und der Grenzwert heißt **einseitige Richtungsableitung von** f **bei** x **in Richtung** d und wird mit $f'(x; d)$ bezeichnet.

Lemma 3.3.7. Für $i \in \{1, \dots, k\}$ sei $\varphi_i \in C_1(\mathbf{R}^n, \mathbf{R})$. Dann ist

$$\Psi(x) := \max \left\{ \varphi_i(x) \;\middle|\; i = 1, \dots, k \right\}$$

bei allen $x \in \mathbf{R}^n$ für alle $d \in \mathbf{R}^n$ einseitig richtungsdifferenzierbar und

$$\Psi'(x; d) = \max \left\{ \nabla \varphi_i(x) d \;\middle|\; \Psi(x) = \varphi_i(x),\ i = 1, \dots, k \right\}.$$

Beweis. Setze

$$\mathcal{A} := \left\{ i \;\middle|\; \varphi_i(x) = \Psi(x) \right\} \text{ und } \gamma(x) := \max \left\{ \varphi_i(x) \;\middle|\; i \notin \mathcal{A} \right\} < \Psi(x).$$

Weil die φ_i stetig differenzierbar sind, gilt

$$\varphi_i(x+td) = \varphi_i(x) + t\nabla\varphi_i(x)d + t\varepsilon_i(t) \text{ mit } \varepsilon_i(t) = o(t) \text{ für } 1 \leqslant i \leqslant k.$$

Wegen der Stetigkeit der φ_i und Ψ gibt es ein $t_0 > 0$ mit

$$\varphi_i(x+td) < \frac{1}{2}\Big(\Psi(x+td) + \gamma(x)\Big) \text{ für alle } t \in [-t_0, t_0] \text{ und } i \notin \mathcal{A}.$$

Es folgt

$$\Psi(x+td) = \max\Big\{ \underbrace{\varphi_i(x)}_{=\Psi(x)} + t\nabla\varphi_i(x)d + t\varepsilon_i(t) \;\Big|\; i \in \mathcal{A}\Big\} \text{ für } t \in [-t_0, t_0]$$

und damit ist

$$\frac{\Psi(x+td) - \Psi(x)}{t} = \max\Big\{\nabla\varphi_i(x)d + \varepsilon_i(t) \;\Big|\; i \in \mathcal{A}\Big\} \text{ für } t \in (0, t_0].$$

Jetzt kann der Grenzübergang $t \to 0$ durchgeführt werden, wonach die Behauptung folgt. □

Bemerkung 3.3.5. Mit Hilfe von Lemma 3.3.7 können wir die einseitigen Richtungsableitungen der nichtdifferenzierbaren Penalty-Funktionen bestimmen.

1. Die einseitige Ableitung einer Betragsfunktion $\Psi(x) = |h_j(x)| = \max\{h_j(x), -h_j(x)\}$ ist also

$$\Psi'(x,d) = \begin{cases} \nabla h_j(x)d & \text{falls } h_j(x) > 0, \\ |\nabla h_j(x)d| & \text{falls } h_j(x) = 0, \\ -\nabla h_j(x)d & \text{falls } h_j(x) < 0. \end{cases}$$

2. Die einseitige Richtungsableitung des positiven Teils einer Funktion $\Psi(x) = g_i^+(x) = \max\{g_i(x), 0\}$ ist

$$\Psi'(x,d) = \begin{cases} \nabla g_i(x)d & \text{falls } g_i(x) > 0, \\ (\nabla g_i(x)d)^+ & \text{falls } g_i(x) = 0, \\ 0 & \text{falls } g_i(x) < 0. \end{cases}$$

Satz 3.3.8. Es sei $\bar{x} \in \mathbf{R}^n$ und A positiv definit. Für ein $x_0 \in \mathbf{R}^n$ gelte

$$g(\bar{x}) + \nabla g(\bar{x})(x_0 - \bar{x}) < 0$$

und

$$h(\bar{x}) + \nabla h(\bar{x})(x_0 - \bar{x}) = 0.$$

Es sei x_* die Optimallösung von $QP(\bar{x}, A)$

$$\begin{array}{ll} \min & \bar{f}(x) = f(\bar{x}) + \nabla f(\bar{x})(x-\bar{x}) + \frac{1}{2}(x-\bar{x})^T A(x-\bar{x}) \\ \text{unter} & \qquad \bar{g}(x) = g(\bar{x}) + \nabla g(\bar{x})(x-\bar{x}) \;\leqslant\; 0, \\ & \qquad \bar{h}(x) = h(\bar{x}) + \nabla h(\bar{x})(x-\bar{x}) \;=\; 0 \end{array} \tag{3.19}$$

und $d := x_* - \bar{x} \neq 0$.

Dann gibt es positive Vektoren $b \in \mathbf{R}^m_>$, $c \in \mathbf{R}^p_>$ und ein $t_0 > 0$, so daß $\Phi'(\bar{x}; d; b, c) < 0$, d. h.

$$\Phi(\bar{x} + td; b, c) < \Phi(\bar{x}; b, c) \text{ für alle } t \in (0, t_0).$$

Beweis. Es sei A positiv definit. Damit ist $\bar{f}$ strikt konvex. Insbesondere ist (3.19) ein konvexes Optimierungsproblem, für das die Slater-Bedingung erfüllt ist (wähle $x_0 = \bar{x}$). (3.19) hat also eine eindeutig bestimmte Optimallösung x_*, die durch die Multiplikatorregeln

$$\underbrace{\nabla f(\bar{x}) + (x_* - \bar{x})^T A}_{=\nabla \bar{f}(x_*)} + \lambda_*^T \nabla g(\bar{x}) + \mu_*^T \nabla h(\bar{x}) = 0$$

$$\lambda_*^T \Big(\underbrace{g(\bar{x}) + \nabla g(\bar{x})(x_* - \bar{x})}_{=\bar{g}(x_*)} \Big) = 0$$

für $\lambda_* \in \mathbf{R}^m_{\geqslant}$, $\mu_* \in \mathbf{R}^p$ charakterisiert ist.
Setze

$$B_g = \Big\{ i \;\Big|\; g_i(\bar{x}) < 0 \Big\}, \quad B_h = \Big\{ i \;\Big|\; h_i(\bar{x}) < 0 \Big\},$$

$$A_g = \Big\{ i \;\Big|\; g_i(\bar{x}) = 0 \Big\}, \quad A_h = \Big\{ i \;\Big|\; h_i(\bar{x}) = 0 \Big\},$$

$$V_g = \Big\{ i \;\Big|\; g_i(\bar{x}) > 0 \Big\}, \quad V_h = \Big\{ i \;\Big|\; h_i(\bar{x}) > 0 \Big\}.$$

Dann gilt nach Bemerkung 3.3.5

$$\begin{aligned} \Phi'(\bar{x}; d; b, c) &= \nabla f(\bar{x}) d + \sum_{i \in V_g} b_i \nabla g_i(\bar{x}) d + \sum_{i \in A_g} b_i (\nabla g_i(\bar{x}) d)^+ \\ &\quad + \sum_{j \in B_h} c_j \nabla h_j(\bar{x}) d + \sum_{j \in A_h} c_j |\nabla h_j(\bar{x}) d| + \sum_{j \in V_h} c_j (-\nabla h_j(\bar{x}) d) \\ &= \nabla f(\bar{x}) d + \sum_{i \in V_g} b_i \nabla g_i(\bar{x}) d + \sum_{i \in A_g} b_i (\nabla g_i(\bar{x}) d)^+ - \sum_{j=1}^{p} c_j |h_j(\bar{x})|, \end{aligned}$$

weil x_* zulässig für $QP(\bar{x}, A)$ ist, also $0 = \bar{h}_j(x_*) = -\nabla h_j(\bar{x}) d$ gilt.
Ebenfalls aus der Zulässigkeit von x_* für $QP(\bar{x}, A)$ folgt $\nabla g(\bar{x}) d \leqslant -g(\bar{x})$, was für $c \geqslant |\mu_*|$ und $b \geqslant \lambda_*$ die Abschätzung

$$\begin{aligned} &\Phi'(\bar{x}; d; b, c) \\ &\leqslant \underbrace{-d^T A d - \lambda_*^T \nabla g(\bar{x}) d - \mu_*^T \nabla h(\bar{x}) d}_{=\nabla f(\bar{x}) d} + \sum_{i \in V_g} b_i \nabla g_i(\bar{x}) d + \sum_{i \in A_g} b_i \underbrace{g_i(\bar{x})^+}_{=0} - \sum_{j=1}^{p} c_j |h_j(\bar{x})| \\ &= -d^T A d - \sum_{j=1}^{p} \Big(\underbrace{c_j |h_j(\bar{x})| - (\mu_*)_j h_j(\bar{x})}_{\geqslant 0 \text{ falls } c \geqslant |\mu_*|} \Big) + \sum_{i \in V_g} \Big(b_i - (\lambda_*)_i \Big) \nabla g_i(\bar{x}) d - \sum_{i \in A_g \cup B_g} (\lambda_*)_i \nabla g_i(\bar{x}) d \\ &\leqslant -\underbrace{d^T A d}_{>0} - \sum_{i \in V_g} \Big(\underbrace{b_i - (\lambda_*)_i}_{\geqslant 0} \Big) \underbrace{g_i(\bar{x})}_{>0} + \sum_{i \in B_g} (\lambda_*)_i \underbrace{g_i(\bar{x})}_{<0} \end{aligned}$$

ermöglicht. Der obige Ausdruck ist offensichtlich echt negativ, weil $d^T A d > 0$ gilt. □

Bemerkung 3.3.6. (Schrittweitensteuerung für Algorithmus 3.3.1)
Das Verfahren von Han benutzt die Penalty-Funktion Φ zur Bestimmung einer Schrittweite. Von Han wurde ursprünglich gefordert, daß Φ in Richtung d nahezu exakt minimiert wird, genauer:

$$\Phi(x_{k+1}; b, c) \leqslant \min \left\{ \Phi(x_k + \sigma d; b, c) + \varepsilon_k \;\middle|\; 0 \leqslant \sigma \leqslant \delta \right\}$$

für ein festes $\delta > 0$ und $\sum_{k=1}^{\infty} \varepsilon_k < \infty$. Dies erfordert jedoch einen zu großen Aufwand pro Iteration. Man kann zeigen, daß die in Algorithmus 3.3.1 angegebene Methode, die der Armijo-Goldstein-Suche entspricht, ausreichend ist. Die Schrittweitenbestimmung erfordert also zu $\alpha \in (0,1)$ die Wahl eines minimalen $j \in \mathbf{N}_0$ mit

$$\Phi(x_k + \alpha^j d_k; b_{k+1}, c_{k+1}) - \Phi(x_k; b_{k+1}, c_{k+1}) \leqslant -\delta\alpha^j \left[d_k^T A_k d_k + \varepsilon \|h(x_k)\|_1 + \varepsilon \|g^+(x_k)\|_1 \right].$$

Die Folge der j ist beschränkt, wenn die folgenden Voraussetzungen erfüllt sind:

- (x_k) ist beschränkt,
- alle $QP(x_k, A_k)$ besitzen zulässige Lösungen,
- (A_k) ist eine Folge gleichmäßig beschränkter Matrizen,
- (d_k) ist beschränkt,
- die Penalty-Parameter b_k und c_k sind stets um ein festes $\varepsilon > 0$ größer als λ_k und $|\mu_k|$.

Unter diesen Voraussetzungen folgt weiterhin, daß $(x_k) \subseteq P(\tau_0)$. Da $P(\tau_0)$ kompakt ist, gilt dann auch $(d_k) \to 0$, $(h(x_k)) \to 0$ und $(g^+(x_k)) \to 0$. Insbesondere existiert eine gegen ein $x_* \in P$ konvergente Teilfolge $(x_{k'})$ mit $(A_{k'}) \to A_*$.
Falls die Mangasarian-Fromowitz-Bedingung in P gilt, so ist x_* das Minimum von $QP(x_*, A_*)$, also erfüllt x_* die Multiplikatorregel.

Lösung konvexer Aufgaben mit dem SQP-Verfahren

Wir betrachten wieder die konvexe Optimierungsaufgabe (3.8), also die Aufgabe

$$\min \left\{ f(x) \;\middle|\; g(x) \leqslant 0,\; h(x) = 0 \right\}$$

für konvexes offenes $D \subseteq \mathbf{R}^n$ und auf D konvexe und stetig differenzierbare Funktionen f und g, sowie affin lineares h von der Form $h(x) = Hx - h_0$ mit rang$H$ maximal.
Im folgenden sei die Slater-Bedingung erfüllt. Ferner gebe es ein $\tau_0 > 0$ mit $P(\tau_0) \subseteq D$ und kompakt.
Unter diesen Voraussetzungen gilt

Lemma 3.3.9. Es sei $(x_k) \subseteq P(\tau_0)$ und (A_k) eine beschränkte, gleichmäßig positiv definite Matrixfolge. Dann besitzt $QP(x_k, A_k)$ eine eindeutige Optimallösung $\hat{x}_{k+1}$, und die Folgen der zugehörigen Lagrange-Parameter sowie der Suchrichtungen $d_k := \hat{x}_{k+1} - x_k$ sind beschränkt.

Beweis. Weil Slater-Bedingung für P gilt, gibt es ein $\hat{x} \in P$ mit

$$h(\hat{x}) = 0 \text{ und } g(\hat{x}) < 0.$$

Damit folgt

$$h(x_k) + \nabla h(x_k)(\hat{x} - x_k) = Hx_k - h_0 + H(\hat{x} - x_k) = h(\hat{x}) = 0$$

und weil g konvex ist

$$g(x_k) + \nabla g(x_k)(\hat{x} - x_k) \leqslant g(\hat{x}) < 0.$$

Also ist $\hat{x}$ zulässig für $QP(x_k, A_k)$ und die Slaterbedingung für $QP(x_k, A_k)$ ist erfüllt. Das Problem $QP(x_k, A_k)$ besitzt also stets eine eindeutige Optimallösung $\hat{x}_{k+1}$, die die Multiplikatorregel erfüllt, d. h. mit geeigneten Lagrange-Parametern $\lambda_k \in \mathbf{R}^m_{\geqslant}$ und $\mu_k \in \mathbf{R}^p$ gilt

$$\begin{aligned} \nabla f(x_k) + d_k^T A_k + \lambda_k^T \nabla g(x_k) + \mu_k^T \nabla h(x_k) &= 0 \\ \lambda_k^T (g(x_k) + \nabla g(x_k) d_k) &= 0 \\ g(x_k) + \nabla g(x_k) d_k &\leqslant 0 \\ h(x_k) + \nabla h(x_k) d_k &= 0 \end{aligned}$$

Wir betrachten nun den Wert von $d := x - x_k$ anstelle von x als Variable der quadratischen Approximation. In d ausgedrückt lautet die Lagrangefunktion von $QP(x_k, A_k)$

$$\ell(d, \lambda, \mu) = f(x_k) + \nabla f(x_k) d + \frac{1}{2} d^T A_k d + \lambda^T \Big(g(x_k) + \nabla g(x_k) d \Big) + \mu^T \Big(h(x_k) + \nabla h(x_k) d \Big)$$

und damit lautet das zu $QP(x_k, A_k)$ gehörende Sattelpunktproblem nach Satz 3.1.16

$$\ell(d_k, \lambda, \mu) \leqslant \ell(d_k, \lambda_k, \mu_k) \leqslant \ell(d, \lambda_k, \mu_k) \text{ für alle } d \in \mathbf{R}^n,\ \lambda \in \mathbf{R}^m_{\geqslant} \text{ und } \mu \in \mathbf{R}^p. \tag{3.20}$$

Sei $\hat{d}_k := \hat{x} - x_k$. Damit ist

$$\begin{aligned} 0 &\geqslant \lambda_k^T g(\hat{x}) \geqslant \lambda_k^T \Big(g(x_k) + \nabla g(x_k) \hat{d}_k \Big) \\ &= \underbrace{f(x_k) + \nabla f(x_k) \hat{d}_k + \frac{1}{2} \hat{d}_k^T A_k \hat{d}_k + \lambda_k^T \Big(g(x_k) + \nabla g(x_k) \hat{d}_k \Big)}_{=\ell(\hat{d}_k, \lambda_k, \mu_k), \text{ weil } h(x_k) + \nabla h(x_k) \hat{d}_k = 0} - f(x_k) - \nabla f(x_k) \hat{d}_k - \frac{1}{2} \hat{d}_k^T A_k \hat{d}_k \\ &\geqslant \ell(d_k, \lambda_k, \mu_k) - f(x_k) - \nabla f(x_k) \hat{d}_k - \frac{1}{2} \hat{d}_k^T A_k \hat{d}_k \\ &= \nabla f(x_k)(d_k - \hat{d}_k) + \frac{1}{2} d_k^T A_k d_k - \frac{1}{2} \hat{d}_k^T A_k \hat{d}_k. \end{aligned}$$

Nun ist $\gamma := \max \Big\{ g_i(\hat{x}) \;\Big|\; i = 1, \ldots, m \Big\} < 0$, weil $g(\hat{x}) < 0$. Damit folgt aus der vorigen Ungleichungskette die Abschätzung

$$\|\lambda_k\|_1 \leqslant \frac{1}{|\gamma|} \left[\nabla f(x_k)(\hat{d}_k - d_k) - \frac{1}{2} d_k^T A_k d_k + \frac{1}{2} \hat{d}_k^T A_k \hat{d}_k \right].$$

Das unrestringierte Minimum der Zielfunktion von $QP(x_k, A_k)$ wird bei $-A_k^{-1}\nabla f(x_k)^T$ angenommen, also gilt

$$f(x_k) - \frac{1}{2}\nabla f(x_k)A_k^{-1}\nabla f(x_k)^T \leqslant f(x_k) + \nabla f(x_k)d_k + \frac{1}{2}d_k^T A_k d_k.$$

Diese Ungleichung können wir in der Abschätzung von $\|\lambda_k\|_1$ verwenden und erhalten

$$\|\lambda_k\|_1 \leqslant \frac{1}{|\gamma|}\left[\nabla f(x_k)\hat{d}_k + \frac{1}{2}\hat{d}_k^T A_k \hat{d}_k + \frac{1}{2}\nabla f(x_k)A_k^{-1}\nabla f(x_k)^T\right].$$

Weil $P(\tau_0)$ kompakt ist, sind alle $\hat{d}_k$ beschränkt. Nach Voraussetzung sind auch A_k und A_k^{-1} gleichmäßig beschränkt, d. h. es gilt $\|\lambda_k\|_1 \leqslant c_1$ für ein genügend großes $c_1 > 0$.
Wir führen nun eine QR-Zerlegung von H^T durch:

$$QH^T = \begin{pmatrix} R \\ 0 \end{pmatrix},$$

wobei R rechte obere $(p\times p)$-Dreieckmatrix ist. Sei $r_k^i := \operatorname{sgn}((\mu_k)_i)e_i$ und $R^T s_k^i := -h(x_k)+r_k^i$ für $i = 1, \ldots, p$. Die s_k^i sind gleichmäßig beschränkt, also auch $d_k^i := Q^T\begin{pmatrix} s_k^i \\ 0 \end{pmatrix}$. Damit gilt

$$h(x_k) + \nabla h(x_k)d_k^i = r_k^i,$$

wobei $\|d_k\| \leqslant c_2$ für $c_2 > 0$ hinreichend groß.
Nun können wir mit Hilfe der Sattelpunktbedingung (3.20) mit $d = d_k^i$ wie folgt abschätzen:

$$\begin{aligned} |(\mu_k)_i| &= \mu_k^T r_k^i \\ &= \mu_k^T\Big(h(x_k) + \nabla h(x_k)d_k^i\Big) \\ &\leqslant \nabla f(x_k)d_k^i + \frac{1}{2}(d_k^i)^T A_k d_k^i - \lambda^T\Big(g(x_k) + \nabla g(x_k)d_k^i\Big) + \frac{1}{2}\nabla f(x_k)A_k^{-1}\nabla f(x_k)^T \end{aligned}$$

Wegen der (gleichmäßigen) Beschränktheit aller auftretenden Faktoren gibt es ein hinreichend großes $c_3 > 0$ mit $\|\mu_k\| \leqslant c_3$. Weil $P(\tau_0)$ kompakt ist, ist auch d_k beschränkt. □

Damit ist gezeigt, daß das SQP-Verfahren für konvexe Aufgaben, die die Slater-Bedingung erfüllen, ein durchführbares Verfahren ist, weil die quadratischen Approximationsprobleme stets lösbar sind.
Dabei ist noch offen, welche Wahl der Matrizen A_k sinnvoll ist.

Wahl der Matrizen A_k im SQP-Algorithmus

Wilson zeigte 1963, daß die Wahl von $A_{k+1} = \mathcal{L}(x_{k+1})$ bei Schrittweite konstant 1 in der Nähe des Optimums ein quadratische konvergentes Verfahren ergibt, wenn f, g und h dreifach stetig differenzierbar sind.
Ist $\mathcal{L}(x_*, \lambda_*, \mu_*)$ positiv definit, und wird für genügend große k die Schrittweite identisch 1 gesetzt, dann ist die Konvergenz nach Han superlinear, wenn das BFGS-Update

$$A_{k+1} := \Psi_{\mathrm{BFGS}}\Big(A_k, x_{k+1} - x_k, \nabla_x\ell(x_{k+1}, \lambda_k, \mu_k) - \nabla_x\ell(x_k, \lambda_k, \mu_k)\Big)$$

benutzt wird, wobei

$$\Psi_{\mathrm{BFGS}}(A, s, y) = A - \frac{Ass^T A}{s^T As} + \frac{yy^T}{y^T s}.$$

Update der Penaltyparameter im SQP-Algorithmus

Bemerkung 3.3.7. In der vorgestellten Version des SQP-Algorithmus werden die Penaltyparameter b und c nie verkleinert und sind deshalb in der Regel zu groß.
In einer Modifikation von Powell wählt man

$$(b_{k+1})_i := \max\left\{(\lambda_k)_i, \frac{1}{2}\Big[(b_k)_i + (\lambda_k)_i\Big]\right\} \text{ für } 1 \leqslant i \leqslant m$$

und

$$(c_{k+1})_j := \max\left\{|\mu_k|_j, \frac{1}{2}\Big[(c_k)_j + |\mu_k|_j\Big]\right\} \text{ für } 1 \leqslant i \leqslant m.$$

Das resultierende Verfahren hat sich bewährt, weil es im allgemeinen weniger Schritte zur Bestimmung einer optimalen Lösung benötigt als die Grundversion, es konvergiert jedoch nicht immer.

Literaturverzeichnis

[1] John E. Dennis, Jr. und Robert B. Schnabel: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall Inc., 1983, Prentice-Hall Series in Computational Mathematics, ISBN 0-13-627216-9.

[2] Michael L. Dowling: *An affine scaling algorithm for quadratic programming.* Forschungsbericht, Abteilung für Mathematische Optimierung, Institut für Angewandte Mathematik, Technische Universität Carolo-Wilhelmina zu Braunschweig, 1992.

[3] Roger Fletcher: *Practical Methods of Optimization.* 2. Auflage, John Wiley & Sons, Chichester, 1987, ISBN 0-471-91547-5.

[4] Gene H. Golub und Charles F. van Loan: *Matrix Computations.* 2. Auflage, The John Hopkins University Press, Baltimore, 1989, ISBN 0-8018-3739-1.

[5] Roger A. Horn und Charles A. Johnson: *Topics in Matrix Analysis.* Cambridge University press, Cambridge, 1991, ISBN 0-521-30587-X.

[6] Renato. D. C. Monteiro und Ilan Adler: *Interior path-following primal-dual algorithms — Part II. Convex quadratic programming.* Math. Programming **44** (1989), 43–66.

[7] Peter Spellucci: *Numerische Verfahren der nichtlinearen Optimierung.* Birkhäuser Verlag, Berlin, 1993, Internationale Schriftenreihe zur numerischen Mathematik, ISBN 3-7643-2854-1.

[8] Josef Stoer: *Einführung in die Numerische Mathematik I.* 3. Auflage, Heidelberger Taschenbücher, Vol. 15, Springer-Verlag, Berlin, 1979, ISBN 3-540-09346-X.

[9] Josef Stoer und Roland Bulirsch: *Einführung in die Numerische Mathematik II.* Heidelberger Taschenbücher, Vol. 15, Springer-Verlag, Berlin, 1978, ISBN 3-540-08840-7.

[10] Uwe Zimmermann: *Einführung in die Mathematische Optimierung.* Vorlesungsskriptum, Abteilung für Mathematische Optimierung, Institut für Angewandte Mathematik, Technische Universität Carolo-Wilhelmina zu Braunschweig, 1993.

Index