Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Nichtlineare Optimierung

Vorlesungsskriptum im SoSe2023

7. März 2023

Winnifried Wollner

winnifried.wollner@uni-hamburg.de
Universität Hamburg,
MIN Fakultät,
Fachbereich Mathematik

Inhaltsverzeichnis

1 Einleitung

1.1 Einleitung

Vorwort Dieses Skript ist eine Weiterentwicklung des Skripts zur Vorlesung "Optimierung" an der Universität Hamburg im Sommer 2014, 2015, 2022 sowie der Vorlesung "Nichtlineare Optimierung" an der Technischen Universität Darmstadt im Winter 2016/2017, 2018/2019 sowie 2021/2022. Die Vorlesung ist in weiten Teilen an das Buch Ulbrich and Ulbrich [2012] angelehnt. Dieses Skriptum dient daher vor allem zur Planung der Vorlesung und ist nicht zur Weitergabe bestimmt.

Einleitung & Grundlagen Die Vorlesung gibt einen Einstieg in die nichtlineare Optimierung. Sie basiert auf dem Buch Ulbrich and Ulbrich [2012]. Ergänzend zur Vorlesung kann es hilfreich sein sich mit anderen Standardwerken zum Thema auseinander zu setzen. Als Beispiele seien hier Bertsekas [1999], Geiger and Kanzow [1999, 2002], Nocedal and Wright [1999] erwähnt.

Wir werden uns im folgenden mit der Lösung von Problemen der Form

$$\min f(x) \text{ unter der Nebenbedingung (u.d.N.) } x \in X \tag{1.1}$$

befassen. Dabei nehmen wir an, die *Zielfunktion* $f : X \to \mathbb{R}$ wäre stetig auf dem nichtleeren *zulässigen Bereich (Menge)* $X \subset \mathbb{R}^n$. Durch Wechsel des Vorzeichen enthält diese offenkundig auch Maximierungsprobleme.

Wir können nun einige grundlegende Sprachkonventionen treffen:

Definition 1.1.1. Ein Punkt (Vektor) $x \in X$ heißt *zulässig* für (1.1), andernfalls ($x \notin X$) *unzulässig*.

Ein Punkt $\overline{x} \in \mathbb{R}^n$ ist ein

1. *lokales Minimum* von (1.1) falls $\overline{x}$ zulässig ist, und für ein $\varepsilon > 0$ gilt

$$f(x) \geq f(\overline{x}) \qquad \forall x \in X \cap B_\varepsilon(\overline{x}).$$

2. *striktes lokales Minimum* (auch *strenges*) von (1.1) falls $\overline{x}$ zulässig ist, und für ein

$\varepsilon > 0$ gilt

$$f(x) > f(\overline{x}) \qquad \forall x \in (X \cap B_\varepsilon(\overline{x})) \setminus \{\overline{x}\}.$$

3. *isoliertes lokales Minimum* von (1.1) falls $\overline{x}$ für ein $\varepsilon > 0$ das einzige lokale Minimum auf $X \cap B_\varepsilon(\overline{x})$ ist.
4. *globales Minimum* von (1.1) falls in 1. jedes beliebige $\varepsilon > 0$ gewählt werden kann.
5. *striktes globales Minimum* von (1.1) falls 2. für jedes beliebige $\varepsilon > 0$ gilt.

Die Existenz von mindestens einer Lösung des Problems (1.1) folgt aus dem folgenden Theorem:

Theorem 1.1.2 (Existenz von Minimierern)**.** *Sei $f : X \to \mathbb{R}$ stetig und sei $x^0 \in X$, so dass die Niveaumenge*

$$N_f(x^0) := \{x \in X \,|\, f(x) \leq f(x^0)\}$$

kompakt ist. Dann besitzt (1.1) *mindestens ein globales Minimum.*

Beweis. Es genügt offenbar das äquivalente Problem

$$\min f(x) \quad \text{u.d.N. } x \in X \cap N_f(x^0)$$

zu betrachten. Da der zulässige Bereich kompakt ist folgt die Existenz eines globalen Minimums aus dem Satz vom Extremum (Satz von Weierstraß). □

Bemerkung 1.1.3. Im Gegensatz zur Existenz ist die Suche nach globalen Extrema i.A. sehr aufwändig, da das Auffinden aller lokalen Minima und der Vergleich der entsprechenden Zielfunktionswerte notwendig ist. Wir beschränken uns daher für die folgende Vorlesung auf die Suche nach lokalen Minima (bzw. stationären Punkten).

Bemerkung 1.1.4. Neben der hier behandelten *kontinuierlichen Optimierung* (d.h. $\text{int}(X) \neq \emptyset$ bzw. das relative innere ist nicht leer) gibt es weitere Fälle die wir hier nicht behandeln werden. Dies sind *diskrete Optimierung* $X \subset \mathbb{Z}^n$ sowie die *unendlich dimensionale Optimierung* falls $X \subset V$ in einem unendlich dimensionalen Raum V.

Für die folgende Veranstaltung werden wir zunächst in Teil 2 *unrestringierte Optimierungsprobleme* ($X = \mathbb{R}^n$) betrachten. Anschließend werden wir in Teil 3 den Fall *restringierter Optimierungsprobleme* betrachten. Hierfür ist es zweckmäßig den zulässigen Bereich in der Form

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

mit stetigen Funktionen $h\colon \mathbb{R}^n \to \mathbb{R}^p$ und $g\colon \mathbb{R}^n \to \mathbb{R}^m$ anzugeben. Dabei verwenden wir für Vektoren im $\mathbb{R}^m$ die Konvention, dass die Ungleichungszeichen komponentenweise zu verstehen sind. Wir erhalten also das *nichtlineare Optimierungsproblem (Nonlinear Program, NLP)*

$$\min f(x) \quad \text{u.d.N} \begin{cases} h(x) = 0, \\ g(x) \leq 0. \end{cases} \tag{1.2}$$

In der Klasse der Probleme (1.2) gibt es viele Spezialfälle die sich in der Literatur finden. dies sind

1. *Gleichungsrestringierte Probleme* falls $m = 0$.
2. *Lineare Optimierungsprobleme* falls f, g, h (affin) lineare Abbildungen sind.
3. *Quadratische Optimierungsprobleme* in denen f quadratisch, sowie g, h (affin) linear sind.
4. *Konvexe Optimierungsprobleme* in denen f, g konvexe Funktionen sind und h affin-linear ist.
5. *Minmax-Probleme* in denen f von der speziellen Form $f = \max_i f_i$ mit glatten Funktionen f_i ist. Da f dann i. A. nicht differenzierbar ist, verzichten wir im folgenden auf diesen Fall.

Es gibt nun im Wesentlichen zwei Quellen für Optimierungsprobleme. Zum Einen glauben wir, dass viele physikalische Systeme einen energieminimalen Zustand annehmen. Ein prototypisches Beispiel hierfür ist das folgende:

Beispiel 1.1.5. Die Auslenkung u einer eingespannten Membran über der Fläche Ω unter gegebener Kraft f (welche nur in normalen Richtung auf die Membran wirkt) ist beispielsweise gegeben durch die Lösung des Problems

$$\min_{v \in H_0^1(\Omega)} f(v) := \frac{1}{2} \int_\Omega |\nabla v|^2 \, \mathrm{d}x - \int_\Omega f v \, \mathrm{d}x.$$

Dieses Problem ist zunächst unendlich dimensional, kann jedoch durch Wahl einer approximierenden Folge von endlich dimensionalen Räumen $V_h \subset H_0^1(\Omega)$ in eine Folge von endlich dimensionalen Problemen der Form

$$\min_{v \in V_h} f(v)$$

überführt werden. Eine typische Wahl für V_h ist z.B. gegeben durch eine Zerlegung $\mathcal{T}_h$

von Ω in Dreiecke T und Setzung

$$V_h = \{v \in C(\overline{\Omega}) \,|\, v\big|_T \in P_1(T),\ v\big|_{\partial\Omega} = 0\}.$$

Dies ist eine typische Quelle von hochdimensionalen Optimierungsproblemen.

Natürlich können in dieser Form auch Schranken an die Lösung eingebaut werden. Darf die Membran beispielsweise nicht unter ein Hindernis ψ fallen, so kann das Problem zu

$$\min_{v \in V_h} f(v) \quad \text{u.d.N. } v \geq \psi.$$

abgeändert werden.

Eine weitere Quelle für Optimierungsprobleme entspringt dem Wunsch des Menschen Prozess zu optimieren, wie z.B.

Beispiel 1.1.6. Zur optimalen Bestimmung des Ortes x für die Platzierung eines Objektes (Produktionsanlage, Mikrochip, usw.) ist der Abstand zu bereits vorhandenen Komponenten an den Orten y_i zu Minimieren. Dies führt auf das Problem

$$\min \frac{1}{2} \sum_i w_i \|x - y_i\|^2 \quad \text{u.d.N. } \|x - y_i\| \geq r_i$$

mit Gewichten $w_i > 0$, die die relative Bedeutung der Entfernung beschreiben, sowie Radien $r_i \geq 0$, die ggf. Schranken an den Abstand (Bauteilgröße) enthalten.

1.2 Notation

Wir bezeichnen mit $x \in \mathbb{R}^n$ stets Spaltenvektoren; der zugehörige Zeilenvektor ist dann x^T.

Auch wenn im Allgemeinen andere Normen betrachtet werden können, werden wir uns für die Zwecke der Vorlesung auf die euklidische Norm $\|x\| = \langle x, x\rangle^{1/2}$ mit zugehörigem Skalarprodukt $\langle x, y\rangle = x^T y$ beziehen. (Dies ist keine Einschränkung der späteren Konvergenzaussagen, da im $\mathbb{R}^n$ alle Normen äquivalent sind!) Entsprechend definieren wir die offene Kugel

$$B_\varepsilon(x) = \{y \in \mathbb{R}^n \,|\, \|x - u\| < \varepsilon\}.$$

Für Matrizen $M \in \mathbb{R}^{m \times n}$ verwenden wir i.A. die induzierte Operatornorm

$$\|M\| = \sup_{\|x\|=1} \|Mx\|.$$

Für eine stetig differenzierbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ (d.h. $f \in C^1(\mathbb{R}^n)$) ist die Ableitung $f'(x) \in \mathscr{L}(\mathbb{R}^n; \mathbb{R}) = (\mathbb{R}^n)^*$ im Punkt x ein Element des Dualraumes (Zeilenvektor). Bei

der gegebenen Wahl des Skalarproduktes weisen wir diesem in kanonischer Weise einen Spaltenvektor, den *Gradienten*

$$\nabla f(x) = \begin{pmatrix} \partial_1 f(x) \\ \vdots \\ \partial_n f(x) \end{pmatrix}$$

von f im Punkt x zu. Dieser ist durch die Gleichung $\langle \nabla f(x), y \rangle = f'(x)y$ definiert. Für zweimal stetig differenzierbares f nennen wir

$$\nabla^2 f(x) = \big(\partial_i \partial_j f(x)\big)_{i,j} = \begin{pmatrix} \partial_1\partial_1 f(x) & \dots & \partial_1\partial_n f(x) \\ \vdots & & \vdots \\ \partial_n\partial_1 f(x) & \dots & \partial_n\partial_n f(x) \end{pmatrix} = (\nabla\partial_1 f(x) \ \dots \ \nabla\partial_n f(x)) \in \mathbb{R}^{n\times n}$$

die *Hesse-Matrix* von f im Punkt x. Da $f \in C^2$ ist $\nabla^2 f(x)$ symmetrisch. Für eine stetig differenzierbare vektorwertige Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$ bezeichnen wir die Ableitung $F'(x) \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ als *Jacobi-Matrix*, d.h.,

$$F'(x) = \begin{pmatrix} \partial_1 F_1(x) & \dots & \partial_n F_1(x) \\ \vdots & & \vdots \\ \partial_1 F_m(x) & \dots & \partial_n F_m(x) \end{pmatrix} = \begin{pmatrix} F_1'(x) \\ \vdots \\ F_m'(x) \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

Kompatibel zur Notation im skalaren Fall setzen wir

$$\nabla F(x) = F'(x)^T.$$

Mit dieser Notation erinnern wir nun an den aus der Analysis bekannten

Theorem 1.2.1 (Taylor)**.** *Sei $\mathcal{O} \subset \mathbb{R}^n$ offen. Dann gilt für $x \in \mathcal{O}$ und hinreichend kleines $d \in \mathbb{R}^n$ sowie*

1. *$F \in C^1(\mathcal{O}; \mathbb{R}^m)$ $(m \geq 1)$:*

$$F(x+d) = F(x) + F'(x)d + o(\|d\|).$$

2. *$f \in C^2(\mathcal{O}; \mathbb{R})$:*

$$f(x+d) = f(x) + f'(x)d + \frac{1}{2}\langle d, \nabla^2 f(x)d\rangle + o(\|d\|^2).$$

Dabei verwenden wir die übliche Landau-Notation

$$\phi(d) = o(\|d\|^k) \text{ genau dann, wenn } \lim_{d\to 0} \frac{\phi(d)}{\|d\|^k} = 0.$$

1.3 Konvexe Optimierung

Bevor wir mit der Betrachtung der unrestringierten und restringierten Optimierung beginnen, wollen wir uns zunächst einen besonderen Spezialfall ansehen. Bei dieser Klasse von Funktionen werden wir sehen, dass lokale Minima auch globale sind. Dies ist für andere Problemklassen natürlich falsch!

Definition 1.3.1. Eine Menge $X \subset \mathbb{R}^n$ heißt *konvex*, falls für alle $x, y \in X$ und $\lambda \in [0,1]$ gilt

$$\lambda x + (1-\lambda)y \in X.$$

Bemerkung 1.3.2. Zu jeder Menge $X \subset \mathbb{R}^n$ gibt es eine (eindeutig bestimmte) kleinste konvexe Menge $\mathrm{co}(X)$, mit $X \subset \mathrm{co}(X) \subset \mathbb{R}^n$. Diese heißt *konvexe Hülle* von X, und besitzt die folgende Darstellung

$$\begin{aligned}\mathrm{co}(X) &= \bigcap_{X \subset K \text{ konvex}} K \\ &= \Big\{ \sum_{i=1}^{m} \lambda_i x_i \,\Big|\, x_i \in X, \lambda_i \geq 0, \sum_{i=1}^{m} \lambda_i = 1 \Big\}.\end{aligned}$$

In der Tat ist $(m = n+1)$ wegen des Satzes von Carathéodory.

Definition 1.3.3. Sei $X \subset \mathbb{R}^n$ konvex und $f : X \to \mathbb{R}$ gegeben. Wir sagen f ist

1. *konvex*, falls für alle $x, y \in X$ und $\lambda \in [0,1]$ gilt
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$
2. *strikt konvex*, falls für alle $x, y \in X$ und $\lambda \in (0,1)$ gilt
$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y).$$
3. *gleichmäßig konvex*, falls es ein $\mu > 0$ gibt, so dass für alle $x, y \in X$ und $\lambda \in [0,1]$ gilt
$$f(\lambda x + (1-\lambda)y) + \mu\lambda(1-\lambda)\|x-y\|^2 \leq \lambda f(x) + (1-\lambda) f(y).$$

Besitzt eine konvexe Funktion zusätzlich Differenzierbarkeit, so können wir obige Begriffe auch anders charakterisieren.

Theorem 1.3.4 (Charakterisierung konvexer C^1-Funktionen)**.** *Sei $X \subset \mathbb{R}^n$ konvex, und $f : X \to \mathbb{R}$ stetig differenzierbar auf einer Umgebung von X. Dann gilt:*

1. *f ist genau dann konvex, wenn für alle $x, y \in X$ gilt*
$$\langle \nabla f(x), y - x \rangle \leq f(y) - f(x).$$
2. *f ist genau dann strikt konvex, wenn für alle $x, y \in X$ mit $x \neq y$ gilt*
$$\langle \nabla f(x), y - x \rangle < f(y) - f(x).$$
3. *f ist genau dann gleichmäßig konvex, wenn es ein $\mu > 0$ gibt, so dass für alle $x, y \in X$ gilt*
$$\langle \nabla f(x), y - x \rangle + \mu \|x - y\|^2 \leq f(y) - f(x).$$

Beweis. **1.** $\Rightarrow$: Sei zunächst f konvex. Dann folgt für $x, y \in X$ und $0 < \lambda \leq 1$

$$\begin{aligned} f(y) - f(x) &= \frac{(1-\lambda)f(x) + \lambda f(y) - f(x)}{\lambda} \\ &\geq \frac{f((1-\lambda)x + \lambda y) - f(x)}{\lambda} \\ &\to \langle \nabla f(x), y - x \rangle \quad (\lambda \downarrow 0). \end{aligned}$$

$\Leftarrow$: Umgekehrt, seien $x, y \in X$ und $\lambda \in [0,1]$ beliebig. Setze $x_\lambda = (1-\lambda)x + \lambda y$. Dann gilt nach Voraussetzung

$$\begin{aligned} (1-\lambda)f(x) + \lambda f(y) - f(x_\lambda) &= (1-\lambda)(f(x) - f(x_\lambda)) + \lambda(f(y) - f(x_\lambda)) \\ &\geq (1-\lambda)\langle \nabla f(x_\lambda), x - x_\lambda \rangle + \lambda \langle \nabla f(x_\lambda), y - x_\lambda \rangle \\ &= \langle \nabla f(x_\lambda), (1-\lambda)x + \lambda y - x_\lambda \rangle \\ &= 0. \end{aligned}$$

Dies zeigt 1.
2. $\Rightarrow$: Sei zunächst f strikt konvex. Seien $x, y \in X$ mit $x \neq y$ gegeben und sei $z = \frac{1}{2}(x + y)$. Dann ist

$$f(z) - f(x) < \frac{1}{2}(f(x) + f(y)) - f(x) = \frac{1}{2}(f(y) - f(x)).$$

Unter Verwendung von 1. erhalten wir somit

$$\begin{aligned} \langle \nabla f(x), y - x \rangle &= 2\langle \nabla f(x), z - x \rangle \\ &\leq 2(f(z) - f(x)) \\ &< f(y) - f(x). \end{aligned}$$

⇐: Die umgekehrte Richtung folgt wie in Fall 1. durch Verwendung der strikten Ungleichung.
3. ⇒: Sei zunächst f gleichmäßig konvex. Es folgt dann durch Differenzenquotienten analog zu 1. mit $x_\lambda = (1-\lambda)x + \lambda y$

$$\begin{aligned}\langle \nabla f(x), y-x\rangle &= \lim_{\lambda \downarrow 0} \frac{f(x_\lambda)-f(x)}{\lambda} \\ &\leq \lim_{\lambda \downarrow 0} \frac{(1-\lambda)f(x)+\lambda f(y)-\mu\lambda(1-\lambda)\|x-y\|^2-f(x)}{\lambda} \\ &= f(y)-f(x)-\mu\|x-y\|^2.\end{aligned}$$

⇐: Umgekehrt erhalten wir analog zu 1.

$$\begin{aligned}(1-\lambda)f(x)+\lambda f(y)-f(x_\lambda) &= (1-\lambda)(f(x)-f(x_\lambda))+\lambda(f(y)-f(x_\lambda)) \\ &\geq (1-\lambda)\Big(\langle \nabla f(x_\lambda), x-x_\lambda\rangle + \mu\|x-x_\lambda\|^2\Big) \\ &\quad + \lambda\Big(\langle \nabla f(x_\lambda), y-x_\lambda\rangle + \mu\|y-x_\lambda\|^2\Big) \\ &= \mu\Big((1-\lambda)\|x-x_\lambda\|^2 + \lambda\|y-x_\lambda\|^2\Big) \\ &= \mu\Big((1-\lambda)\lambda^2 + \lambda(1-\lambda)^2\Big)\|x-y\|^2 \\ &= \mu\lambda(1-\lambda)\|x-y\|^2.\end{aligned}$$

□

Theorem 1.3.5 (Charakterisierung konvexer C^2-Funktionen)**.** *Sei $X \subset \mathbb{R}^n$ offen und konvex sowie $f : X \to \mathbb{R}$ zweimal stetig differenzierbar. Dann gilt:*

1. *Die Funktion f ist genau dann konvex, wenn $\nabla^2 f(x)$ für alle $x \in X$ positiv semidefinit ist, d.h. für alle $x \in X$ und $d \in \mathbb{R}^n$ gilt*
$$\langle d, \nabla^2 f(x) d\rangle \geq 0.$$
2. *Die Funktion f ist strikt konvex, wenn $\nabla^2 f(x)$ für alle $x \in X$ positiv definit ist, d.h. für alle $x \in X$ und $d \in \mathbb{R}^n \setminus \{0\}$ gilt*
$$\langle d, \nabla^2 f(x) d\rangle > 0.$$
3. *Die Funktion f ist genau dann gleichmäßig konvex, wenn $\nabla^2 f(x)$ für alle $x \in X$ gleichmäßig positiv definit ist, d.h. genau dann, wenn es ein $\mu > 0$ gibt, so dass für alle $x \in X$ und $d \in \mathbb{R}^n$ gilt*
$$\langle d, \nabla^2 f(x) d\rangle \geq \mu\|d\|^2.$$

Beweis. **1.** $\Rightarrow$: Sei zunächst f konvex, $x \in X$ und $d \in \mathbb{R}^n$ beliebig. Da X offen ist, ist $x + td \in X$ für alle hinreichend kleinen $t > 0$. Das Theorem von Taylor 1.2.1 zusammen mit Theorem 1.3.4.1 liefert

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

Der Grenzübergang $t \to 0$ liefert die gewünschte Aussage.
$\Leftarrow$: Umgekehrt liefert uns Taylorentwicklung mit Darstellung des Restgliedes für ein $\theta \in [0,1]$

$$\begin{aligned} f(y) - f(x) &= \langle \nabla f(x), y-x\rangle + \frac{1}{2}\langle y-x, \nabla^2 f(x+\theta(y-x))(y-x)\rangle \\ &\geq \langle \nabla f(x), y-x\rangle \end{aligned}$$

und somit folgt Konvexität nach Theorem 1.3.4.1.
2. Analog zu obiger Formel erhalten wir die strikte Ungleichung falls $x \neq y$.
3. $\Rightarrow$: Analog zu 1. folgt aus der gleichmäßigen Konvexität

$$0 \leq f(x+td) - f(x) - t\langle \nabla f(x), d\rangle - \mu\|td\|^2 = \frac{t^2}{2}\langle d, \nabla^2 f(x) d\rangle - \mu\|td\|^2 + o(t^2)$$

und somit nach Grenzübergang $t \downarrow 0$

$$2\mu\|d\|^2 \leq \langle d, \nabla^2 f(x) d\rangle$$

$\Leftarrow$: Umgekehrt folgt analog zu 1. für ein $\theta \in (0,1)$

$$\begin{aligned} f(y) - f(x) &= \langle \nabla f(x), y-x\rangle + \frac{1}{2}\langle (y-x), \nabla^2 f(x+\theta(y-x))(y-x)\rangle \\ &\geq \langle \nabla f(x), y-x\rangle + \frac{\mu}{2}\|(y-x)\|^2 \end{aligned}$$

und somit folgt gleichmäßige Konvexität nach Theorem 1.3.4.3. □

Bemerkung 1.3.6. Man beachte, dass die Umkehrung der Aussage 2. in Theorem 1.3.5 nicht richtig ist, wie die strikt konvexe Funktion $f(x) = x^4$ zeigt.

Bemerkung 1.3.7. Die Bedeutung der gleichmäßigen Konvexität liegt in der hierdurch erzwungenen „Mindestkrümmung" des Funktionsgraphen. Es kann daher für gleichmäßig konvexe Funktionen keine „flachen" Minima geben.

Theorem 1.3.8 (Minima konvexer Funktionen)**.** *Sei $X \subset \mathbb{R}^n$ konvex und $f : X \to \mathbb{R}$ konvex. Dann gilt:*

1. *Jedes lokale Minimum $\overline{x}$ von f auf X ist auch globales Minimum.*
2. *Ist f strikt konvex, so besitzt f höchstens ein lokales Minimum auf X. Im Falle der Existenz ist dieses dann auch das strikte globale Minimum von f.*
3. *Ist X offen, $f \in C^1(X, \mathbb{R})$ und $\overline{x}$ ein Punkt mit $\nabla f(\overline{x}) = 0$ (ein sog.* stationärer Punkt *von f), so ist $\overline{x}$ ein globales Minimum von f auf X.*

Beweis. **1.** Angenommen es gäbe ein $x \in X$ mit $f(x) < f(\overline{x})$ dann wäre für alle $\lambda \in (0, 1]$

$$f(\overline{x} + \lambda(x - \overline{x})) \leq (1 - \lambda) f(\overline{x}) + \lambda f(x) < f(\overline{x})$$

im Widerspruch zur lokalen Optimalität von $\overline{x}$.
2. Seien $\overline{x}, \overline{y} \in X$ zwei Minima, dann liefert

$$f\Big(\frac{\overline{x} + \overline{y}}{2}\Big) < \frac{f(\overline{x}) + f(\overline{y})}{2} \leq f\Big(\frac{\overline{x} + \overline{y}}{2}\Big)$$

einen Widerspruch.
3. Nach Theorem 1.3.4.1 folgt

$$f(x) - f(\overline{x}) \geq \langle \nabla f(\overline{x}), x - \overline{x} \rangle = 0.$$

Folglich ist $\overline{x}$ ein globales Minimum. □

Bemerkung 1.3.9. Man mache sich anhand der Funktion $f : \{x > 0\} \to \mathbb{R}$ mit $x \mapsto x^{-1}$ klar, dass konvexe Funktionen nicht notwendig ein Minimum besitzen.

2 Unrestringierte Optimierung

Wir werden uns nun zunächst mit dem Fall der unrestringierten Optimierung, d.h. Problemen der Form

$$\min_{x\in\mathbb{R}^n} f(x) \tag{$\mathscr{U}$}$$

befassen.

2.1 Optimalitätsbedingungen

Theorem 2.1.1 (Notwendige Bedingung erster Ordnung)**.** *Sei $f : O \subset \mathbb{R}^n \to \mathbb{R}$ differenzierbar auf der offenen Menge O. Ist dann $\overline{x} \in O$ ein lokales Minimum von f, so ist notwendig $\nabla f(\overline{x}) = 0$.*

Beweis. Offenkundig ist für jedes $d \in \mathbb{R}^n$ und hinreichend kleines $t > 0$

$$f(\overline{x} + td) \geq f(\overline{x}).$$

Damit folgt durch Betrachten der Differenzenquotienten

$$\langle \nabla f(\overline{x}), d\rangle = \lim_{t\downarrow 0} \frac{f(\overline{x}+td) - f(x)}{t} \geq 0 \quad \forall\, d \in \mathbb{R}^n.$$

Durch Wahl von $\pm d$ folgt

$$\langle \nabla f(\overline{x}), d\rangle = 0 \quad \forall\, d \in \mathbb{R}^n$$

und somit die Behauptung. □

Definition 2.1.2. Sei $O \subset \mathbb{R}^n$ offen und $f \in C^1(O;\mathbb{R})$. Dann nennen wir einen Punkt $\overline{x} \in O$ mit $\nabla f(\overline{x}) = 0$ *stationären Punkt* von f.

Bemerkung 2.1.3. Wie die Funktionen $x^2, -x^2, x^3$ zeigen ist das Vorliegen stationärer Punkte, im Gegensatz zu konvexen Funktionen, im Allgemeinen nicht hinreichend für die Existenz eines Minimums.

Definition 2.1.4. Ein stationärer Punkt $\overline{x}$ von f, der weder ein lokales Minimum noch ein lokales Maximum ist nennen wir *Sattelpunkt*.

Bemerkung 2.1.5. Achtung! Der Begriff Sattelpunkt wird in der Literatur nicht einheitlich Definiert. Für unsere Vorlesung ist es aber zweckmäßig diese als „Restmenge" der stationären Punkte zu behandeln.

Um zwischen den verschiedenen Typen von stationären Punkten unterscheiden zu können sind Informationen der ersten Ableitung in nur einem Punkt nicht ausreichend. Wir zeigen daher zunächst folgendes

Theorem 2.1.6 (Notwendige Bedingung zweiter Ordnung)**.** *Sei $f : O \subset \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar auf der offenen Menge O. Ist dann $\overline{x} \in O$ ein lokales Minimum von f, so ist notwendig*

1. *$\overline{x}$ ist ein stationärer Punkt von f, d.h. $\nabla f(\overline{x}) = 0$,*
2. *die Hesse-Matrix positiv semidefinit, d.h.*

$$\langle d, \nabla^2 f(\overline{x}) d \rangle \geq 0 \quad \forall\, d \in \mathbb{R}^n.$$

Beweis. Es ist nur noch 2. zu zeigen, da 1. bereits in Theorem 2.1.1 gezeigt wurde. Zu gegebenem $d \in \mathbb{R}^n$ sei wieder $t > 0$ hinreichend klein. Dann erhalten wir durch Taylorentwicklung 1.2.1, da $\overline{x}$ ein lokales Minimum ist,

$$0 \leq f(\overline{x} + td) - f(\overline{x}) = t\langle \nabla f(\overline{x}), d \rangle + \frac{t^2}{2}\langle d, \nabla^2 f(\overline{x}) d \rangle + o(t^2).$$

Da $\overline{x}$ ein stationärer Punkt ist liefert der Grenzübergang $t \to 0$ für

$$\frac{-2o(t^2)}{t^2} \leq \langle d, \nabla^2 f(\overline{x}) d \rangle$$

das Gewünschte. □

Bemerkung 2.1.7. Wie das Beispiel der Funktionen $x^3, -x^4$ zeigt, sind auch diese Bedingungen noch nicht hinreichend für die Existenz eines Minimums, da wir noch immer nicht zuverlässig zwischen den verschiedenen Typen von stationären Punkten

unterscheiden können.

Das in der Tat nicht mehr viel fehlt zeigen die folgenden hinreichenden Bedingungen die lediglich eine leichte Verschärfung der notwendigen Bedingungen zweiter Ordnung in Theorem 2.1.6 sind.

Theorem 2.1.8 (Hinreichende Bedingungen zweiter Ordnung). *Sei $f : O \subset \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar auf der offenen Menge O. Sei dann $\overline{x} \in O$ ein Punkt in dem gilt*

1. *$\overline{x}$ ist ein stationärer Punkt von f, d.h. $\nabla f(\overline{x}) = 0$,*
2. *die Hesse-Matrix positiv definit, d.h.*

$$\langle d, \nabla^2 f(\overline{x}) d \rangle > 0 \quad \forall\, d \in \mathbb{R}^n \setminus \{0\}.$$

Dann ist $\overline{x}$ ein striktes lokales Minimum von f.

Beweis. Seien die Bedingungen des Theorems erfüllt. Dann gilt für den kleinsten Eigenwert μ von $\nabla^2 f(\overline{x})$

$$\mu \|d\|^2 \leq \langle d, \nabla^2 f(\overline{x}) d \rangle.$$

Sei nun $d \neq 0$. Durch Taylorentwicklung 1.2.1 und Stationarität erhalten wir

$$f(\overline{x} + d) - f(\overline{x}) = \frac{1}{2} \langle d, \nabla^2 f(\overline{x}) d \rangle + o(\|d\|^2).$$

Nun gibt es ein $\varepsilon > 0$, so dass für alle $\|d\| < \varepsilon$ gilt

$$|o(\|d\|^2)| \leq \frac{\mu}{4} \|d\|^2$$

und somit ist

$$f(\overline{x} + d) - f(\overline{x}) \geq \frac{\mu}{2} \|d\|^2 - \frac{\mu}{4} \|d\|^2 = \frac{\mu}{4} \|d\|^2 > 0.$$

□

Wir werden später in Lemma 2.5.4 sehen, dass dieses Minimum auch isoliert ist.

Bemerkung 2.1.9. Es verbleiben lokale Minima für die die hinreichenden Bedingungen zweiter Ordnung aus Theorem 2.1.8 nicht anwendbar sind. Das dies der Fall ist zeigt die Funktion x^4.

Bemerkung 2.1.10. Wie die Betrachtung der Funktionen $\pm\exp(-x^{-2})$ zeigt, kann die Existenz einer Lücke zwischen hinreichenden und notwendigen Bedingungen auch durch die Betrachtung höherer Ableitungen nicht geschlossen werden. Da selbst die Betrachtung sämtlicher Ableitungen in einem Punkt, und damit die ganze Taylorreihe, nicht für alle Funktionen ausreichend viele Informationen bereitstellt.

Bemerkung 2.1.11. Um die Differenz zwischen notwendigen und hinreichenden Bedingungen noch etwas zu erhellen lässt sich folgendes beobachten. Falls $\overline{x}$ die notwendigen Bedingungen zweiter Ordnung für $f : \mathbb{R}^n \to \mathbb{R}$ erfüllt, so ist $\overline{x}$ für beliebiges $\epsilon > 0$ ein lokales Minimum der gestörten Funktion $f_\epsilon(x) := f(x) + \epsilon\|x - \overline{x}\|^2$ da die hinreichenden Bedingungen 2. Ordnung für f_ϵ in $\overline{x}$ erfüllt sind.

2.2 Abstiegsverfahren

Dies legt den folgenden Algorithmus nahe

Algorithmus 2.2.1 (Allgemeines Abstiegsverfahren)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$.

for k = 0, 1, ... **do**

- Prüfe auf Abbruch.
  (Für die Beweise $\nabla f(x^k) = 0$, in der Praxis vgl. Abschnitt 2.2.4)
- Berechne *Abstiegsrichtung* $d^k \in \mathbb{R}^n$ von f in x^k.
- Bestimme *Schrittweite* $t_k > 0$, so dass $f(x^k + t_k d^k) < f(x^k)$.
- Setze $x^{k+1} = x^k + t_k d^k$.

end for

Definition 2.2.2. Ein Vektor $d \in \mathbb{R}^n$ heißt *Abstiegsrichtung* der stetig differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$ im Punkt x, falls

$$\langle \nabla f(x), d \rangle < 0.$$

Solche Abstiegsrichtungen sind hilfreich, da uns der Satz von Taylor 1.2.1 wegen

$$\lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{\|td\|} = \frac{\langle \nabla f(x), d \rangle}{\|d\|} < 0$$

einen Abstieg im Funktionswert in der Richtung d garantiert.

Bemerkung 2.2.3. Achtung! Die Bezeichnung Abstiegsrichtung wird in der Literatur nicht einheitlich verwendet!

Bemerkung 2.2.4. Man beachte, dass eine Richtung in der sich der Funktionswert verringert nicht notwendig eine Abstiegsrichtung ist. Man betrachte hierzu die Funktion x^3 im Punkt $x = 0$. Dies zeigt, dass i.A. nicht mehr als Stationarität von Häufungspunkten der Iterierten x^k von Algorithmus 2.2.1 erwartet werden kann.

In den einzelnen Punkten des Algorithmus kann es zu offenkundigen Problemen kommen:

1. Die Abstiegsrichtungen können so gewählt sein, dass $\langle \nabla f(x^k), d^k \rangle \to 0$ ohne dass $\nabla f(x^k) \to 0$. Dann ist entweder $d^k \to 0$ oder $d^k \to d$ mit einem d das orthogonal zum Gradienten ist. Beide Varianten können die Konvergenz des Verfahrens gegen einen stationären Punkt verhindern. (Ungeschickte Wahl von d^k)
2. Die Schrittweiten $t_k \to 0$, so schnell, dass x^{k+1} gegen einen nicht stationären Punkt konvergiert. (Ungeschickte Wahl von t_k)
3. Die Folge x^k konvergiert nicht; und hat auch keinen Häufungspunkt. (Problem von f; es gibt z.B. kein Minimum)

Wir werden uns daher im Folgenden damit befassen wie diese Fälle ausgeschlossen werden können, um schließlich die Konvergenzaussage

$$\overline{x} \text{ ist Häufungspunkt der von Algorithmus 2.2.1 erzeugten Folge } x^k \Rightarrow \nabla f(\overline{x}) = 0$$

zu zeigen.

Zur Lösung des unrestringierten Problems ($\mathscr{U}$) definieren wir zunächst

2.2.1 Zulässige Richtungen

Um oben genannte Probleme mit der Wahl der Schrittweiten zu umgehen definieren wir

Definition 2.2.5 (Zulässige Richtungen)**.** Eine Teilfolge $K \subset \mathbb{N}$ der in Algorithmus 2.2.1 verwendeten Richtungen $(d^k)_K$ heißt *zulässig*, falls:

$$\langle \nabla f(x^k), d^k \rangle < 0 \quad \forall\, k \geq 0, \tag{2.1}$$

$$\left(\frac{\langle \nabla f(x^k), d^k \rangle}{\|d^k\|} \right)_K \to 0 \quad \Rightarrow \quad (\nabla f(x^k))_K \to 0. \tag{2.2}$$

Die obige Bedingung wird durch die *Winkelbedingung*

$$\cos \angle(-\nabla f(x^k), d^k) := \frac{-\langle \nabla f(x^k), d^k \rangle}{\|\nabla f(x^k)\| \|d^k\|} \geq c \quad \forall\, k \in K \tag{2.3}$$

für ein $c \in (0,1)$ impliziert, denn für eine Abstiegsrichtung d^k ist dann

$$c\|\nabla f(x^k)\| \leq \cos \angle(-\nabla f(x^k), d^k)\| \|\nabla f(x^k)\| = \frac{|\langle \nabla f(x^k), d^k \rangle|}{\|d^k\|}$$

und somit folgt unter der Winkelbedingung aus der Konvergenz des Quotienten $\frac{|\langle \nabla f(x^k), d^k \rangle|}{\|d^k\|}$ auch die Konvergenz von $\nabla f(x^k)$ gegen Null.

Dies lässt sich verallgemeinern zur *verallgemeinerten Winkelbedingung*. Diese ist erfüllt, falls es eine in Null stetige Funktion $\phi : [0, \infty) \to [0, \infty)$ mit $\phi(0) = 0$ gibt für die

$$\|\nabla f(x^k)\| \leq \phi\left(\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|}\right) \tag{2.4}$$

gilt.

Es ist also möglich zulässige Richtungen zu konstruieren solange man nur „nahe" am negativen Gradienten bleibt.

Es gilt in der Tat

Theorem 2.2.6. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und $(d^k)_{k \in K}$ eine Teilfolge der von Algorithmus 2.2.1 erzeugte Folge d^k. Dann gilt*

$$\begin{aligned} & d^k \text{ erfüllt die Winkelbedingung (2.3)} \\ \Rightarrow\ & d^k \text{ erfüllt die verallgemeinerte Winkelbedingung (2.4)} \\ \Rightarrow\ & d^k \text{ erfüllt die Bedingung (2.2)} \end{aligned}$$

Beweis. Die erste Implikation folgt offenbar durch die Wahl $\phi(t) = t/c$. Die zweite Implikation folgt sofort aus der Stetigkeit von ϕ in 0, da damit

$$\|\nabla f(x^k)\| \leq \phi\left(\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|}\right) \to 0 \quad \text{falls} \quad \left(\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|} \to 0\right)$$

□

2.2.2 Zulässige Schrittweiten

Nun befassen wir uns mit den Bedingungen an die Schrittweite.

Definition 2.2.7 (Zulässig Schrittweite)**.** Wir nennen eine von Algorithmus 2.2.1 erzeugte Teilfolge $(t_k)_K$ der Schrittweiten t_k *zulässig*, falls:

$$f(x^k + t_k d^k) \leq f(x^k) \quad \forall\, k \geq 0, \tag{2.5}$$

$$f(x^k + t_k d^k) - f(x^k) \to 0 \quad \Rightarrow \quad \left(\frac{\langle \nabla f(x^k), d^k \rangle}{\|d^k\|} \right)_K \to 0. \tag{2.6}$$

Im Gegensatz zur Wahl der Richtungen ist es hier nicht offenkundig, dass es möglich ist zulässige Schrittweiten zu konstruieren. Das dies in der Tat geht zeigt uns der nachfolgende Begriff, der einen gewissen Mindestabstieg erzwingt.

Definition 2.2.8 (Effiziente Schrittweite)**.** Sei $(d^k)_K$ eine Teilfolge von Abstiegsrichtungen von f in den Punkten x^k. Die Schrittweitenfolge $t_k > 0$ heißt *effizient*, falls

$$f(x^k + t_k d^k) \leq f(x^k) - \theta \left(\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|} \right)^2 \quad \forall\, k \in K$$

mit einem $\theta > 0$ gilt.

Theorem 2.2.9. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Sind x^k, d^k und t_k die durch Algorithmus 2.2.1 erzeugten Folgen. Ist dann $(t_k)_K$ eine effiziente Teilfolge (im Sinn von Definition 2.2.8), dann ist $(t_k)_K$ zulässig (im Sinn von Definition 2.2.7).*

Beweis. Nach Definition des Algorithmus 2.2.1 ist stets $f(x^k + t_k d^k) \leq f(x^k)$, d.h. (2.5) ist erfüllt. Es ist also nur noch zu zeigen, dass (2.6) erfüllt ist.

Sei also $f(x^k + t_k d^k) - f(x^k) \to 0$ für $K \ni k \to \infty$ angenommen. Wegen der angenommenen Effizienz der Schrittweiten ist dann

$$\left(\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|} \right)^2 \leq \frac{1}{\theta} (f(x^k) - f(x^k + t_k d^k)) \to 0$$

was zu zeigen war. □

Wir vertagen die Auswahl einer effizienten Schrittweitenfolge auf später und widmen uns zunächst der Konvergenz des Verfahrens.

2.2.3 Globale Konvergenz

Wir werden für das Folgenden annehmen, dass wir das Verfahren aus Algorithmus 2.2.1 solange durchführen, wie $\nabla f(x^k) \neq 0$.

Theorem 2.2.10 (Konvergenz des allgemeinen Abstiegsverfahrens). *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Angenommen Algorithmus 2.2.1 terminiere nicht nach endlich vielen Schritten und generiere hierbei die Folgen x^k, d^k, t_k. Ist dann $\overline{x}$ ein Häufungspunkt von x^k und $(x^k)_K$ eine gegen diesen konvergente Teilfolge mit zulässigen Richtungen $(d^k)_K$ und Schrittweiten t_k. Dann ist $\overline{x}$ ein stationärer Punkt von f.*

Beweis. Nach Konstruktion ist die Folge $f(x_k)$ monoton fallend. Somit gilt

$$\lim_{k\to\infty} f(x^k) = \lim_{K\ni k\to\infty} f(x^k) = f(\overline{x}).$$

Durch Betrachten von Teleskopsummen erhalten wir

$$-\infty < f(\overline{x}) - f(x^0) = \lim_{k\to\infty} f(x^k) - f(x^0) = \sum_{k=0}^{\infty} (f(x^{k+1}) - f(x^k)) \leq 0.$$

und somit notwendig $f(x^k + t_k d^k) - f(x^k) \to 0$. Die Zulässigkeit der Schrittfolge garantiert somit wegen (2.6)

$$\left(\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|} \right)_K \to 0.$$

Die Zulässigkeit der Richtungen garantiert nun wegen (2.2) sowie der Stetigkeit von ∇f

$$0 = \lim_{K\ni k\to\infty} \nabla f(x^k) = \nabla f(\overline{x}).$$

□

Bemerkung 2.2.11. Man beachte, dass Theorem 2.2.10 nicht die Existenz eines Häufungspunktes impliziert. In der Tat können die folgenden Fälle eintreten:

- Abbruch nach endliche vielen Iterationen. (Dann ist $\nabla f(x^k) = 0$.)
- $f(x^k) \to -\infty$.
- $\nabla f(x^k) \to 0$ (Dies impliziert nicht die Existenz von Häufungspunkten!).

2.2.4 Praktische Wahl der Abbruchkriterien

In der Praxis ist es i.A. nicht möglich den Algorithmus 2.2.1 bis zur Erfüllung der Bedingung $\nabla f(x) = 0$ durchzuführen. Wir geben hier nur kurz die gebräuchlichsten Kriterien wieder, man vergleiche hierzu [Gill et al., 1981, Sektion 8.2.3.2].

Aufgrund inexakter Arithmetik ist die erreichbare Genauigkeit in Funktionswerten und Iterierten beschränkt; und die jeweiligen Genauigkeiten sind miteinander Verknüpft, vgl. hierzu auch [Gill et al., 1981, Sektion 8.2.2.1].

Wir beginnen mit einer allgemeinen Beobachtung. Angenommen $\overline{x}$ löst $\nabla f(\overline{x}) = 0$ dann wird die bestmögliche Approximation $\tilde{x}$ mit numerisch ausgewertetem Gradienten $\widetilde{\nabla f}(\tilde{x}) \approx \nabla f(\tilde{x})$ im Allgemeinen $\widetilde{\nabla f}(\tilde{x}) \neq 0$ erfüllen. Hierbei beachte man, dass Nähe von $\tilde{x}$ zu $\overline{x}$ nicht zwingend mit einem kleinen $\widetilde{\nabla f}(\tilde{x})$ einhergeht. Deshalb wird man in praktischen Rechnungen nach Punkten $\tilde{x}$ suchen, die „Vernünftig" sind in dem Sinne, dass

$$\|\nabla f(\tilde{x})\| \leq \epsilon_g$$

für ein „angemessen kleines" $\epsilon_g > 0$ gilt. Dabei ist „angemessen klein" problemspezifisch und kann für schlecht Konditionierte Probleme recht groß sein. In diesem Sinne betrachten wir alle Punkte $\tilde{x}$ die diese Bedingung erfüllen als „ununterscheidbar" von einander da Sie die Stationaritätsbedingungen mit „angemessener" Genauigkeit erfüllen.

Analog kann man den optimalen Funktionswert nicht exakt erhalten sondern sollte für ein kleines ϵ_f Punkte akzeptieren die der Bedingung

$$|f(\tilde{x}) - f(\overline{x})| \leq \epsilon_f$$

genügen. Ist dann $f \in C^2(\mathbb{R}^n)$, und setzen wir $t = \|\tilde{x} - \overline{x}\|$ und $d = \frac{\tilde{x}-\overline{x}}{t}$ erhalten wir aus einer Taylorentwicklung

$$\begin{aligned} f(\tilde{x}) - f(\overline{x}) &= t\langle \nabla f(\overline{x}), d\rangle + \frac{t^2}{2}\langle d, \nabla^2 f(\overline{x}) d\rangle + o(t^2) \\ &\approx \frac{t^2}{2}\langle d, \nabla^2 f(\overline{x}) d\rangle \end{aligned}$$

für kleine t. Durch Auflösen nach t erhält man eine Relation der erreichbaren Genauigkeiten in den Argumenten:

$$\|\tilde{x} - \overline{x}\| = t \approx \sqrt{\frac{2\epsilon_f}{\langle d, \nabla^2 f(\overline{x}) d\rangle}}$$

sofern $\langle d, \nabla^2 f(\overline{x}) d\rangle \neq 0$ ist.

Dies motiviert z.B. als Abbruchkriterium die simulaten Erfüllung der Bedingungen

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

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

zu fordern. Hierbei ist TOL > 0 eine benutzerdefinierte relative Toleranz. Die Motivation ist hierbei folgendes: Die Iteration kann/sollte angehalten werden sobald die Änderung der Funktionswerte und Iterierten unterhalb der „angemessenen" Genauigkeit liegt und die Stationarität hinreichend gut erfüllt ist. Tasächlich würde eine Taylor entwicklung auch $\sqrt{\text{TOL}}$ in der dritten Bedingung nahelegen, aber [Gill et al., 1981, Sektion 8.2.3.2] Argumentieren dies wäre für praktische Anwendungen zu restriktiv.

Ergänzend sollten die Abbruchbedingungen

- $\|\nabla f(x^k)\| \leq \varepsilon_{\text{Maschine}}$
- $k \geq k_{\max}$

berücksichtigt werden um sicherzustellen, dass die Iteration endet, wenn die Stationarität in Maschienengenauigkeit nicht von Null zu unterscheiden ist, oder falls eine maximale Iterationszahl erreicht wurde. Letzteres ist eine Sicherung für Situationen in denen kein „nahezu" Stationärerpunkt gefunden werden kann.

2.3 Gradientenverfahren

Wir werden nun unseren Ersten praktikablen Algorithmus zur Lösung des Problems ($\mathscr{U}$) vorstellen. Zu diesem Zweck müssen wir die Wahl von Richtung und Schrittweite in Algorithmus 2.2.1 spezifizieren.

2.3.1 Richtung des steilsten Abstiegs

Die naheliegendste Wahl für die Abstiegsrichtung ist zunächst die Richtung des steilsten Abstiegs.

Definition 2.3.1. Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und $x \in \mathbb{R}^n$ beliebig mit $\nabla f(x) \neq 0$. Sei $d \in \mathbb{R}^n$ eine Lösung des Problems

$$\min_{\|d\|_X = 1} \langle \nabla f(x), d \rangle = \min_{\|d\|_X = 1} f'(x)d$$

für eine gegebene Norm $\|\cdot\|_X$ auf dem $\mathbb{R}^n$. Dann bezeichnen wir alle Vektoren der Form λd für $\lambda > 0$ als *Richtung des steilsten Abstiegs* von f in x bezüglich $\|\cdot\|_X$. Ist $\|\cdot\|_X = \|\cdot\|$, die euklidische Norm, so verzichten wir auf die Nennung der Norm.

Theorem 2.3.2. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und $x \in \mathbb{R}^n$ beliebig mit $\nabla f(x) \neq 0$. Dann sind alle Richtungen des steilsten Abstiegs (in der euklidischen Norm) von der Form*

$$-\lambda \nabla f(x), \quad \lambda > 0.$$

Beweis. Diese Aussage ist bereits aus der Analysis bekannt. Sie folgt sofort aus der Ungleichung

$$\langle \nabla f(x), d \rangle \geq -\|\nabla f(x)\|\|d\| = -\|\nabla f(x)\| \quad \forall \|d\| = 1.$$

Zusammen mit der Tatsache, dass Gleichheit nur für linear abhängige Vektoren gilt. □

Bemerkung 2.3.3. Ist $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, und sei

$$\|x\|_A^2 = \langle x, Ax \rangle$$

die durch A induzierte Norm. Dann sind die Richtungen des steilsten Abstiegs bezüglich $\|\cdot\|_A$ gegeben durch

$$-\lambda A^{-1} \nabla f(x), \quad \lambda > 0.$$

Theorem 2.3.4. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und $x \in \mathbb{R}^n$ beliebig mit $\nabla f(x) \neq 0$. Dann ist die Richtungen $d = -\nabla f(x)$ des steilsten Abstiegs zulässig im Sinne von Definition 2.2.5*

Beweis. Die Bedingung (2.1) folgt sofort aus

$$\langle \nabla f(x), d \rangle = -\|\nabla f(x)\|^2 < 0.$$

Die Bedingung (2.2) ist ebenfalls offenkundig, da

$$-\|\nabla f(x)\| = \frac{\langle \nabla f(x), d \rangle}{\|d\|} \to 0 \quad \Leftrightarrow \|\nabla f(x)\| \to 0.$$

□

2.3.2 Armijo-Schrittweitenregel

Wir müssen nun eine zulässige Schrittweitenregel für die obige Wahl der Richtungen konstruieren. Hierzu betrachten wir die folgende Definition

Definition 2.3.5 (Armijo-Schrittweitenbedingung)**.** Sei $\gamma \in (0,1)$ (z.B. $\gamma = 10^{-2}$) gewählt. Wir sagen dann, eine Schrittweite t_k genügt der Armijo-Bedingung falls

$$f(x^k + t_k d^k) - f(x^k) \leq t_k \gamma \langle \nabla f(x^k), d^k \rangle \qquad (\mathscr{A})$$

gilt.

Die Armijo-Bedingung sichert, im Vergleich mit dem durch den Gradienten vorhergesagten Abstieg, eine gewisse Abnahme im Funktionalwert. Hierbei ist zu beachten, dass $\gamma \approx 1$ im allgemeinen sehr kleine Schrittweiten impliziert, da nur dann die lineare Approximation durch den Gradienten gut genug ist um den Abstieg zu erhalten. Um nun eine solche Richtung zu finden verwenden wir den folgenden

Algorithmus ALS [2.3.6] (Liniensuche (Backtracking Linesearch))**.**

- Wähle $\beta \in (0,1)$ (z.B. $\beta = 0.5$) und $\gamma \in (0,1)$.

for $l = 0, 1, \ldots$ **do**
Falls $t_k = \beta^l$ die Armijo-Bedingung ($\mathscr{A}$) erfüllt → Ende!
end for

Zunächst stellen wir fest, dass der Liniensuchalgorithmus ALS [2.3.6] durchführbar ist.

Theorem 2.3.7. *Sei $O \subset \mathbb{R}^n$ offen, $f : O \to \mathbb{R}$ stetig differenzierbar und $\gamma \in (0,1)$ gegeben. Ist dann $x \in O$ und $d \in \mathbb{R}^n$ eine Abstiegsrichtung von f in x, so gibt es $\bar{t}$, so dass*

$$f(x + td) - f(x) \leq t\gamma \langle \nabla f(x), d \rangle \quad \forall t \in [0, \bar{t}].$$

Insbesondere terminiert Algorithmus ALS [2.3.6] nach endlich vielen Schritten mit einer Schrittweite $t_k > 0$, die der Armijo-Bedingung ($\mathscr{A}$) genügt.

Beweis. Für $t = 0$ ist nichts zu zeigen. Sei also $t > 0$, klein genug damit $x + td \in O$. Taylorentwicklung 1.2.1 impliziert für $t \to 0$

$$\frac{f(x+td) - f(x)}{t} - \gamma \langle \nabla f(x), d \rangle \to \langle \nabla f(x), d \rangle - \gamma \langle \nabla f(x), d \rangle = (1-\gamma)\langle \nabla f(x), d \rangle < 0.$$

□

Nun können wir für die spezielle Wahl der Richtung $d^k = -\nabla f(x^k)$ die Zulässigkeit der durch Algorithmus ALS [2.3.6] erzeugten Schrittweiten zeigen.

Theorem 2.3.8. *$f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und $\beta, \gamma \in (0,1)$ gegeben. Sei dann $d^k = -\nabla f(x^k)$ eine Richtung des steilsten Abstiegs und die Schrittweitenfolge t_k durch Algorithmus ALS [2.3.6] erzeugt. Ist dann $(x^k)_K$ eine konvergente Teilfolge, so ist $(t_k)_K$ zulässig im Sinne von Definition 2.2.7.*

Beweis. Offenkundig ist (2.5) erfüllt. Wir müssen also nur noch (2.6) zeigen. Nach Annahme ist $(x^k)_K \to \overline{x}$. Wegen der Monotonie der Funktionswerte folgt $f(x^k + t_k d^k) - f(x^k) \to 0$. Im Falle $(\nabla f(x^k))_K \to \nabla f(\overline{x}) = 0$ ist nichts zu zeigen. Wir nehmen also an

$$\|\nabla f(\overline{x})\| > 0.$$

Aus der Armijo-Bedingung erhalten wir

$$t_k \gamma \|\nabla f(x^k)\|^2 = -t_k \gamma \langle \nabla f(x^k), d^k \rangle \leq f(x^k) - f(x^{k+1}) \to 0 \quad (k \to \infty).$$

In Kombination erhalten wir $(t_k)_K \to 0$. O.B.d.A. können wir also $t_k < 1$ annehmen. Entsprechend der Vorschrift von Algorithmus ALS [2.3.6] ist dann $\beta^{-1} t_k \leq 1$ und die Schrittweite $\beta^{-1} t_k$ erfüllt die Armijo-Bedingung ($\mathscr{A}$) nicht, d.h.

$$f(x^k + \beta^{-1} t_k d^k) - f(x^k) > \gamma \beta^{-1} t_k \langle \nabla f(x^k), d^k \rangle.$$

Offenkundig ist mit t_k auch $\beta^{-1} t_k$ eine Nullfolge, und wir erhalten nach dem Mittelwertsatz für $\theta_k \in (0,1)$

$$\begin{aligned}
-\|\nabla f(\overline{x})\|^2 &= \lim_{K \ni k \to \infty} -\|\nabla f(x^k)\|^2 \\
&< -\gamma \lim_{K \ni k \to \infty} \|\nabla f(x^k)\|^2 \\
&= \gamma \lim_{K \ni k \to \infty} \langle \nabla f(x^k), d^k \rangle \\
&\leq \lim_{K \ni k \to \infty} \frac{f(x^k + \beta^{-1} t_k d^k) - f(x^k)}{\beta^{-1} t_k} \\
&= \lim_{K \ni k \to \infty} \frac{\beta^{-1} t_k \langle \nabla f(x^k + \theta_k \beta^{-1} t_k d^k), d^k \rangle}{\beta^{-1} t_k} \\
&= \lim_{K \ni k \to \infty} \Big(\langle \nabla f(x^k), d^k \rangle + \langle \nabla f(x^k + \theta_k \beta^{-1} t_k d^k) - \nabla f(x^k), d^k \rangle \Big) \\
&= - \lim_{K \ni k \to \infty} \|\nabla f(x^k)\|^2 \\
&= -\|\nabla f(\overline{x})\|^2 < 0.
\end{aligned}$$

Hierbei haben wir ausgenutzt, dass $d^k = -\nabla f(x^k) \to \nabla f(\overline{x})$ gilt und daher $\theta_k \beta^{-1} t_k d^k \to 0$ folgt.

Das strikte Ungleichheitszeichen liefert einen Widerspruch. Die Annahme $\nabla f(\overline{x}) \neq 0$ kann also nicht zutreffen. □

Bemerkung 2.3.9. Man beachte, das wir in obigem Beweis mehrfach die Eigenschaft $d = -\nabla f(x)$ ausgenutzt haben. Für allgemeine Richtungen werden wir also noch einmal arbeiten müssen.

2.3.3 Globale Konvergenz

Wir können nun das Gradientenverfahren als Spezialfall des allgemeinen Abstiegsverfahrens 2.2.1 formulieren.

Algorithmus GV [2.3.10] (Gradientenverfahren)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$.

for k = 0, 1, ... **do**

- Prüfe auf Abbruch ($\nabla f(x^k) = 0$).
- Setze $d^k = -\nabla f(x^k)$.
- Bestimme $t_k > 0$ durch die Armijo-Liniensuche (Algorithmus ALS [2.3.6]).
- Setze $x^{k+1} = x^k + t_k d^k$.

end for

Für diesen haben wir folgendes Konvergenzresultat

Korollar 2.3.11. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Dann terminiert Algorithmus GV [2.3.10] entweder nach endlich vielen Schritten mit einem stationären Punkt x^k, oder er erzeugt eine unendliche Folge (x^k) mit den Eigenschaften*

1. *$f(x^{k+1}) < f(x^k)$ für alle k.*
2. *Jeder Häufungspunkt von (x^k) ist ein stationärer Punkt.*

Beweis. Im Falle des Abbruchs ist nichts zu zeigen. Ansonsten ist die Eigenschaft 1. eine unmittelbare Konsequenz aus der Armijo-Bedingung ($\mathscr{A}$). Die Eigenschaft 2. folgt dann aus der Zulässigkeit der Richtungen d^k nach Theorem 2.3.4 sowie der Zulässigkeit der Schrittweiten t_k für die konvergente Teilfolge nach Theorem 2.3.8 durch Anwendung des Konvergenzsatzes für das Abstiegsverfahren aus Theorem 2.2.10. □

2.3.4 Konvergenzgeschwindigkeit

Ein wesentlicher Punkt für die Betrachtung allgemeiner Abstiegsverfahren in Abschnitt 2.2 ist die Tatsache, das das Gradientenverfahren i.A. nur sehr langsam konvergiert. Um dies auch

formal einzusehen und die Geschwindigkeit zu charakterisieren werden wir uns für diesen Abschnitt auf eine streng konvexe quadratische Zielfunktion, d.h.

$$f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle \tag{2.7}$$

mit $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, beschränken. In diesem einfachen Fall können wir auf das Backtracking verzichten und die Schrittweite exakt über den Ansatz

$$t_k = \arg\min_{t>0} f(x^k + td^k)$$

bestimmen. Hieraus erhalten wir über die notwendigen und hinreichenden Bedingungen für Minima konvexer Funktionen aus Theorem 1.3.8 unter Verwendung von $\nabla f(x^k) = Ax^k + b = -d^k$

$$\begin{aligned} 0 &= \langle Ax^k + b, d^k\rangle + t_k\langle d^k, Ad^k\rangle \\ &= -\|d^k\|^2 + t_k\langle d^k, Ad^k\rangle \end{aligned}$$

bzw.

$$t_k = \frac{\|d^k\|^2}{\langle d^k, Ad^k\rangle}. \tag{2.8}$$

Für den folgenden Konvergenzsatz benötigen wir den folgenden Hilfssatz

Lemma 2.3.12 (Kantorowitsch-Ungleichung). *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit mit Eigenwerten*

$$0 < \lambda = \lambda_1 \le \ldots \le \lambda_n = \Lambda.$$

Dann gilt für alle $x \in \mathbb{R}^n \setminus \{0\}$ die Kantorowitsch-Ungleichung

$$\frac{\|x\|^4}{\langle x, Ax\rangle\langle x, A^{-1}x\rangle} \ge \frac{4\lambda\Lambda}{(\lambda+\Lambda)^2}.$$

Beweis. Sei $v_i \in \mathbb{R}^n$ eine Orthonormalbasis aus Eigenvektoren von A bzw. A^{-1} zu den Eigenwerten λ_i bzw. λ_i^{-1}. Dann folgt mit der Basisdarstellung $x = \sum_{i=1}^n x_i v_i$, $x_i \in \mathbb{R}$:

$$\begin{aligned} 4\lambda\Lambda\langle x, Ax\rangle\langle x, A^{-1}x\rangle &= 4\lambda\Lambda \sum_{i=1}^n \lambda_i x_i^2 \sum_{i=1}^n \lambda_i^{-1} x_i^2 \\ &= 4\lambda\Lambda\|x\|^4 \sum_{i=1}^n \lambda_i \frac{x_i^2}{\|x\|^2} \sum_{i=1}^n \lambda_i^{-1} \frac{x_i^2}{\|x\|^2}. \end{aligned} \tag{2.9}$$

Nun erfüllen die Werte $0 \le \frac{x_i^2}{\|x\|^2} \le 1$, $\sum_{i=1}^n \frac{x_i^2}{\|x\|^2} = 1$ und die Abbildung $\lambda \mapsto \lambda^{-1}$ ist konvex auf $\mathbb{R}_{>0} \to \mathbb{R}_{>0}$. Wir erhalten somit

$$\left(\sum_{i=1}^n \lambda_i \frac{x_i^2}{\|x\|^2}\right)^{-1} \le \sum_{i=1}^n \lambda_i^{-1} \frac{x_i^2}{\|x\|^2}.$$

Sei nun $\mu = \sum_{i=1}^n \lambda_i \frac{x_i^2}{\|x\|^2} \in [\lambda, \Lambda]$. Dann ist $\left(\sum_{i=1}^n \lambda_i \frac{x_i^2}{\|x\|^2}\right)^{-1} = \mu^{-1}$ weiter ist $\sum_{i=1}^n \lambda_i^{-1} \frac{x_i^2}{\|x\|^2}$ eine Konvexkombination der inversen Eigenwerte. Somit folgt wegen $\lambda_i^{-1} \in [\Lambda^{-1}, \lambda^{-1}]$

$$\mu^{-1} \leq \sum_{i=1}^n \lambda_i^{-1} \frac{x_i^2}{\|x\|^2} \leq \frac{(\lambda + \Lambda - \mu)}{\lambda\Lambda} =: g(\mu)$$

dabei gilt die Ungleichung, da die rechten Seite gerade die lineare Interpolation zwischen den Werten Λ^{-1} und λ^{-1} ist. Somit ergibt sich in obiger Gleichung (2.9)

$$\begin{aligned}
4\lambda\Lambda\langle x, Ax\rangle\langle x, A^{-1}x\rangle &= 4\lambda\Lambda\|x\|^4 \sum_{i=1}^n \lambda_i \frac{x_i^2}{\|x\|^2} \sum_{i=1}^n \lambda_i^{-1} \frac{x_i^2}{\|x\|^2} \\
&= 4\lambda\Lambda\|x\|^4 \frac{\sum_{i=1}^n \lambda_i^{-1} \frac{x_i^2}{\|x\|^2}}{\left(\sum_{i=1}^n \lambda_i \frac{x_i^2}{\|x\|^2}\right)^{-1}} \\
&\leq 4\lambda\Lambda\|x\|^4 \max_{\mu \in [\lambda, \Lambda]} \frac{(\lambda + \Lambda - \mu)}{\lambda\Lambda\mu^{-1}} \\
&= 4\|x\|^4 \max_{\mu \in [\lambda, \Lambda]} \left((\lambda + \Lambda)\mu - \mu^2\right).
\end{aligned}$$

Das Maximum wird in $(\lambda + \Lambda)/2$ angenommen, und es ist somit

$$\begin{aligned}
4\lambda\Lambda\langle x, Ax\rangle\langle x, A^{-1}x\rangle &\leq 4\|x\|^4 \left(\frac{(\lambda + \Lambda)^2}{2} - \frac{(\lambda + \Lambda)^2}{4}\right) \\
&= 4\|x\|^4 \frac{(\lambda + \Lambda)^2}{4}
\end{aligned}$$

dies zeigt die Behauptung. □

Theorem 2.3.13 (Konvergenzgeschwindigkeit des Gradientenverfahrens). *Sei $f : \mathbb{R}^n \to \mathbb{R}$ streng konvex und quadratisch, also von der Form (2.7). Seien die Folgen x^k, d^k und t_k durch das Gradientenverfahren mit exakter Liniensuche erzeugt. Dann gilt*

$$f(x^{k+1}) - f(\overline{x}) \leq \left(\frac{\Lambda - \lambda}{\Lambda + \lambda}\right)^2 (f(x^k) - f(\overline{x}))$$

$$\|x^k - \overline{x}\| \leq \sqrt{\frac{\Lambda}{\lambda}} \left(\frac{\Lambda - \lambda}{\Lambda + \lambda}\right)^k \|x^0 - \overline{x}\|.$$

Hierbei ist, wie immer, $\overline{x} = -A^{-1}b$ das globale Minimum von f und λ, Λ bezeichnen den

kleinsten und größten Eigenwert von A.

Beweis. Für die Funktionswerte liefert uns Taylorentwicklung 1.2.1

$$f(x^{k+1}) = f(x^k) + t_k\langle \nabla f(x^k), d^k\rangle + \frac{t_k^2}{2}\langle d^k, Ad^k\rangle.$$

Man beachte hierbei, dass der Restgliedterm für eine quadratische Funktion verschwindet. Verwendung der Schrittweitenformel (2.8) liefert damit

$$\begin{aligned} f(x^{k+1}) - f(\overline{x}) &= f(x^k) - f(\overline{x}) + t_k\langle \nabla f(x^k), d^k\rangle + \frac{t_k^2}{2}\langle d^k, Ad^k\rangle \\ &= f(x^k) - f(\overline{x}) - t_k\|d^k\|^2 + \frac{t_k^2}{2}\langle d^k, Ad^k\rangle \\ &= f(x^k) - f(\overline{x}) - \frac{\|d^k\|^4}{\langle d^k, Ad^k\rangle} + \frac{1}{2}\frac{\|d^k\|^4}{\langle d^k, Ad^k\rangle} \\ &= f(x^k) - f(\overline{x}) - \frac{1}{2}\frac{\|d^k\|^4}{\langle d^k, Ad^k\rangle}. \end{aligned}$$

Erneute Taylorentwicklung, und $\nabla f(\overline{x}) = 0$ sowie $d^k = -\nabla f(x^k) = -(Ax^k + b) = -A(x^k - \overline{x})$ liefert

$$\begin{aligned} f(x^k) - f(\overline{x}) &= \frac{1}{2}\langle x^k - \overline{x}, A(x^k - \overline{x})\rangle \\ &= \frac{1}{2}\langle A(x^k - \overline{x}), A^{-1}A(x^k - \overline{x})\rangle \\ &= \frac{1}{2}\langle d^k, A^{-1}d^k\rangle. \end{aligned}$$

Kombination dieser beiden Rechnungen liefert mit der Kantorowitsch-Ungleichung 2.3.12

$$\begin{aligned} f(x^{k+1}) - f(\overline{x}) &= f(x^k) - f(\overline{x}) - \frac{1}{2}\frac{\|d^k\|^4}{\langle d^k, Ad^k\rangle} \\ &= \left(1 - \frac{\|d^k\|^4}{\langle d^k, Ad^k\rangle\langle d^k, A^{-1}d^k\rangle}\right)(f(x^k) - f(\overline{x})) \\ &\leq \left(1 - \frac{4\lambda\Lambda}{(\lambda+\Lambda)^2}\right)(f(x^k) - f(\overline{x})) \\ &= \frac{(\lambda-\Lambda)^2}{(\lambda+\Lambda)^2}(f(x^k) - f(\overline{x})). \end{aligned}$$

Für die zweite Ungleichung iteriert man die Erste bis zu $f(x^0) - f(\overline{x})$ und nutzt

$$\frac{\lambda}{2}\|x^k - \overline{x}\|^2 \leq \frac{1}{2}\langle x^k - \overline{x}, A(x^k - \overline{x})\rangle = f(x^k) - f(\overline{x}) \leq \frac{\Lambda}{2}\|x^k - \overline{x}\|^2.$$

□

Bemerkung 2.3.14. Für Schrittweiten t_k, die durch die Armijo-Liniensuche ALS [2.3.6] bestimmt werden folgt, durch Vergleich der Funktionswerte mit exakter Schrittweite t_{opt}

$$f(x + t_{\text{opt}} d) \leq f(x + t_k d) \leq f(x)$$

und somit für ein $\kappa \geq \frac{(\lambda-\Lambda)^2}{(\lambda+\Lambda)^2}$

$$f(x + t_{\text{opt}} d) - f(\overline{x}) \leq \kappa(f(x) - f(\overline{x})).$$

Für allgemeine Zielfunktionen f mit $\nabla^2 f(\overline{x})$ positiv definit folgt zumindest in einer Umgebung von $\overline{x}$ eine Konvergenzrate die durch die Eigenwerte von $\nabla^2 f(\overline{x})$ dominiert wird.

Beispiel 2.3.15. Wir wollen nun einmal den Einfluss der Wahl der Norm auf das Gradientenverfahren betrachten. Hierzu betrachten wir das folgende Problem der Suche nach Anfangswerten (q_u, q_v) für die gedämpfte Wellengleichung

$$\begin{aligned} \partial_t^2 u - \mu\Delta u - \lambda\Delta\partial_t u &= f && \text{in } \Omega \times (0,T), \\ u(0) &= q_u && \text{in } \Omega, \\ \partial_t u(0) &= q_v && \text{in } \Omega, \\ u &= 0 && \text{auf } \partial\Omega \times (0,T), \end{aligned} \tag{2.10}$$

so dass $u(T) = u(0)$ und $\partial_t u(T) = \partial_t u(0)$ auf einem Gebiet Ω mit Parametern $\mu, \lambda > 0$ gilt.

Dies führt auf das Optimierungsproblem

$$\begin{aligned} &\min \frac{1}{2}\|q_u - u(T)\|^2_{L^2(\Omega)} + \frac{1}{2}\|q_v - \partial_t u(T)\|^2_{L^2(\Omega)} \\ &\quad \text{u.d.N (2.10) ist erfüllt.} \end{aligned}$$

Die Gleichung (2.10) lässt sich als System 1.-Ordnung in den Variablen u und $v = \partial_t u$ schreiben und anschließend, z.B. mittels Finiter Elemente, diskretisieren. Hiernach sucht man Vektoren $\vec{q}_u, \vec{q}_v \in \mathbb{R}^N$ und $\vec{u}, \vec{v} \in (\mathbb{R}^N)^M$ (mit beliebig großen Freiheitsgraden im Ort N und in der Zeit M), so dass diese das Problem

$$\min \frac{1}{2}(\vec{q}_u - \vec{u}_M)^T M(\vec{q}_u - \vec{u}_M) + \frac{1}{2}(\vec{q}_v - \vec{v}_M)^T M(\vec{q}_v - \vec{v}_M)$$
$$\text{u.d.N } A\begin{pmatrix}\vec{u}\\ \vec{v}\end{pmatrix} = \begin{pmatrix}\vec{q}_u\\ \vec{q}_v\\ \vec{f}\end{pmatrix}$$

lösen. Hierbei ist $A \in \mathbb{R}^{2MN \times 2MN}$ invertierbar, $M \in \mathbb{R}^{N \times N}$ die sog. Massematrix, und $\vec{f} \in \mathbb{R}^{2(M-1)N}$ eine Diskretisierung der rechten Seite f. Ersetzt man nun

$$\begin{pmatrix}\vec{u}\\ \vec{v}\end{pmatrix} = A^{-1}\begin{pmatrix}\vec{q}_u\\ \vec{q}_v\\ \vec{f}\end{pmatrix}$$

im Funktional, so erhält man ein unbeschränktes quadratisches Minimierungsproblem in der unbekannten $(\vec{q}_u, \vec{q}_v)$.

Wir schauen uns nun das Verhalten des Verfahrens des steilsten Abstiegs für dieses Problem ansehen. Hierzu wählen wir $N = 81$, $M = 26$ und erhalten die folgenden Werte für das Kostenfunktional (welches im Optimum den Wert 0 hat). In der folgenden Tabelle sind die Funktionalwerte für die Richtung des steilsten Abstiegs bezüglich der euklidischen Norm im $\mathbb{R}^{2N}$ sowie der durch die Matrix M induzierten Norm. $\mathbb{R}^{2N}$

Iteration	euklidische Norm	M-Norm
0	0.0025	0.0025
1	$2.52 \cdot 10^{-3}$	$1.28 \cdot 10^{-3}$
2	$2.51 \cdot 10^{-3}$	$7.26 \cdot 10^{-5}$
3	$2.49 \cdot 10^{-3}$	$7.01 \cdot 10^{-6}$
4	$2.47 \cdot 10^{-3}$	$1.35 \cdot 10^{-6}$
5	$2.46 \cdot 10^{-3}$	$6.86 \cdot 10^{-7}$
6	$2.44 \cdot 10^{-3}$	$8.11 \cdot 10^{-8}$
7	$2.42 \cdot 10^{-3}$	$2.46 \cdot 10^{-8}$
8	$2.41 \cdot 10^{-3}$	$1.07 \cdot 10^{-8}$
9	$2.39 \cdot 10^{-3}$	$3.27 \cdot 10^{-9}$
10	$2.37 \cdot 10^{-3}$	$1.60 \cdot 10^{-9}$
11	$2.36 \cdot 10^{-3}$	$5.12 \cdot 10^{-10}$
12	$2.34 \cdot 10^{-3}$	$2.45 \cdot 10^{-10}$
13	$2.33 \cdot 10^{-3}$	$8.16 \cdot 10^{-11}$
14	$2.31 \cdot 10^{-3}$	$3.77 \cdot 10^{-11}$
15	$2.30 \cdot 10^{-3}$	$1.31 \cdot 10^{-11}$
16	$2.28 \cdot 10^{-3}$	$5.81 \cdot 10^{-12}$
17	$2.27 \cdot 10^{-3}$	$2.09 \cdot 10^{-12}$
18	$2.25 \cdot 10^{-3}$	$8.98 \cdot 10^{-13}$
19	$2.24 \cdot 10^{-3}$	$3.35 \cdot 10^{-13}$

Dies zeigt eindrücklich die Sensitivität des Algorithmus GV [2.3.10] auf die Wahl der Richtung des steilsten Abstiegs. In der Tat braucht das Verfahren mit der Richtung des steilsten Abstiegs aus der euklidischen Norm etwa 3000 Iterationen, um die selbe Genauigkeit wie das Verfahren bezüglich der M-Norm zu erreichen.

2.4 Schrittweitenregeln

Wie wir im vorherigen Abschnitt gesehen haben ist das Konvergenzverhalten des Gradientenverfahrens i.A. unbefriedigend. Wir möchten uns daher mit Alternativen zur Wahl des steilsten Abstiegs befassen.

Bei der speziellen Wahl $d^k = -\nabla f(x^k)$ im Gradientenverfahren konnten wir zeigen, dass die durch die Armijo-Regel gefundenen Schrittweiten zulässig sind, vgl. Theorem 2.3.8. Dies ist bei allgemeinen Suchrichtungen nicht mehr richtig

Beispiel 2.4.1. Man betrachte die Zielfunktion $f(x) = \frac{x^2}{8}$, $f : \mathbb{R} \to \mathbb{R}$. Für Startpunkte $x^0 > 0$ und Suchrichtungen $d^k = -2^{-k}\nabla f(x^k)$ führe man das allgemeine Abstiegsverfahren mit Armijo-Liniensuche ALS [2.3.6] durch. Die so erzeugte Folge x^k

ist monoton fallend gegen einen Limes $\overline{x} \geq x^0/2$. Wegen der Stetigkeit von f folgt $f(x^k + t_k d^k) - f(x^k) \to 0$, es ist aber

$$\frac{\langle f(x^k), d^k \rangle}{\|d^k\|} = \|\nabla f(x^k)\| \to \|\nabla f(\overline{x})\| > 0.$$

Die Schrittweitenfolge ist also nicht zulässig.

Die Problematik des obigen Beispiels lässt sich schnell erahnen, da $\|d^k\| \to 0$, obwohl wir nicht in der Nähe eines stationären Punktes sind. Damit müssten wir Schrittweiten $t_k > 1$ zulassen, was unsere Armijo-Regel nicht zulässt. Das dies in der Tat den Kern der Sache trifft zeigt der folgendes Theorem:

Theorem 2.4.2. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und die Teilfolge $(x^k)_K$ sei beschränkt. Angenommen es gibt eine streng monoton wachsende Funktion $\phi : [0, \infty) \to [0, \infty)$, so dass die Abstiegsrichtungen d^k der Bedingung*

$$\|d^k\| \geq \phi\left(\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|}\right)$$

genügen, dann ist die durch die Armijo-Regel ALS [2.3.6] erzeugte Schrittweitenfolge t_k zulässig im Sinne von Definition 2.2.7.

Beweis. Wir verallgemeinern nun den Beweis von Theorem 2.3.8 auf den Fall allgemeiner Suchrichtungen.

Im Falle

$$\left(\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|}\right)_K \to 0$$

ist nichts zu zeigen. Wir nehmen daher an

$$\left(\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|}\right)_K \not\to 0.$$

Dann gibt es notwendig ein $\varepsilon > 0$, so dass

$$\frac{-\langle \nabla f(x^k), d^k \rangle}{\|d^k\|} \geq \varepsilon \quad \forall\, k \in K$$

gilt, wobei wir ggf. zu einer weiteren Teilfolge übergehen. Aufgrund unserer Bedingung an die Suchrichtung d^k ist damit

$$\|d^k\| \geq \phi(\varepsilon) > 0.$$

Da die Armijo-Bedingung erfüllt ist gilt

$$t_k\gamma\varepsilon\phi(\varepsilon) \le t_k\gamma\varepsilon\|d^k\| \le -t_k\gamma\langle\nabla f(x^k), d^k\rangle \le f(x^k) - f(x^{k+1}) \to 0 \quad (K \ni k \to 0)$$

und somit $(t_k d^k)_K \to 0$ wie auch $(t_k)_K \to 0$. O.B.d.A. können wir also $t_k < 1$ annehmen. Entsprechend der Vorschrift von Algorithmus ALS [2.3.6] ist dann $\beta^{-1}t_k \le 1$ und die Schrittweite $\beta^{-1}t_k$ erfüllt die Armijo-Bedingung ($\mathscr{A}$) nicht, d.h.

$$f(x^k + \beta^{-1}t_k d^k) - f(x^k) > \gamma\beta^{-1}t_k\langle\nabla f(x^k), d^k\rangle.$$

Offenkundig ist mit t_k auch $\beta^{-1}t_k$ eine Nullfolge, und wir erhalten durch Taylorentwicklung mit $\theta_k \in (0,1)$

$$\begin{aligned}\gamma\langle\nabla f(x^k), d^k\rangle &\le \frac{f(x^k + \beta^{-1}t_k d^k) - f(x^k)}{\beta^{-1}t_k} \\ &= \frac{\beta^{-1}t_k\langle\nabla f(x^k + \theta_k\beta^{-1}t_k d^k), d^k\rangle}{\beta^{-1}t_k} \\ &= \langle\nabla f(x^k), d^k\rangle + \langle\nabla f(x^k + \theta_k\beta^{-1}t_k d^k) - \nabla f(x^k), d^k\rangle\end{aligned}$$

Durch Umstellen erhalten wir

$$\begin{aligned}0 &\le (1-\gamma)\frac{\langle\nabla f(x^k), d^k\rangle}{\|d^k\|} + \frac{\langle\nabla f(x^k + \theta_k\beta^{-1}t_k d^k) - \nabla f(x^k), d^k\rangle}{\|d^k\|} \\ &\le -(1-\gamma)\varepsilon + \frac{\langle\nabla f(x^k + \theta_k\beta^{-1}t_k d^k) - \nabla f(x^k), d^k\rangle}{\|d^k\|} \\ &\to -(1-\gamma)\varepsilon < 0\end{aligned}$$

wobei wir im letzten Schritt die gleichmäßige Stetigkeit von ∇f auf der kompakten Menge $\operatorname{clos}\bigcup B_1(x^k)$ ausgenutzt haben. Da die Ungleichungskette einen offenkundigen Widerspruch erzeugt, kann die Annahme

$$\left(\frac{-\langle\nabla f(x^k), d^k\rangle}{\|d^k\|}\right)_K \not\to 0$$

daher nicht zutreffen. □

Bemerkung 2.4.3. Da die Zulässigkeit der Suchrichtungen d^k, Definition 2.2.5 invariant unter Skalierungen mit $\lambda > 0$ ist, kann man die Bedingung aus Theorem 2.4.2 durch geeignetes Skalieren von d^k stets erreichen. Wir müssten hierzu jedoch eine Funktion ϕ geeignet wählen. Allerdings ist dieses Vorgehen nicht so schön, da das reskalieren von d^k ein Problem der Schrittweitenwahl auf die Suchrichtung verschiebt. Zudem ist nicht so klar, was ein „geeignetes" ϕ ist.

2.4.1 Powell-Wolfe-Schrittweitenregel

Anstatt, wie in der obigen Bemerkung ausgeführt, die Suchrichtung umzuskalieren können wir alternativ auch Schrittweiten $t_k > 1$ zulassen. Offenkundig brauchen wir dann neben der Armijo-Regel die eine obere Schranke an t_k definiert auch eine Regel für eine untere Schranke an t_k.

Wir betrachten hierzu die folgende Schrittweitenregel nach Powell und Wolfe

Definition 2.4.4 (Powell-Wolfe-Schrittweitenbedingung)**.** Wir sagen eine Schrittweite t_k erfüllt die Powell-Wolfe-Schrittweitenbedingung, falls Sie zu gegebenen Parametern $\gamma \in (0, 1/2)$ und $\eta \in (\gamma, 1)$ sowohl die Armijo-Bedingung ($\mathscr{A}$)

$$f(x^k + t_k d^k) - f(x^k) \leq t_k \gamma \langle \nabla f(x^k), d^k \rangle$$

wie auch

$$\langle \nabla f(x^k + t_k d^k), d^k \rangle \geq \eta \langle \nabla f(x^k), d_k \rangle \qquad (\mathscr{P}W)$$

erfüllt.

Bemerkung 2.4.5. Die zweite Ungleichung in der Powell-Wolfe Bedingung besagt gerade, dass die aktuelle Suchrichtung in dem neuen Punkt keine besonders gute Abstiegsrichtung mehr ist. Dies hat hat zwei wichtige Funktionen: Zum Einen garantiert dies, dass keine zu kleinen Schritte gemacht werden können, da ansonsten $\langle \nabla f(x^k + t_k d^k), d^k \rangle \approx \langle \nabla f(x^k), d_k \rangle$ gilt.

Zum Anderen ermöglicht es diese Regel das exakte Minimum von $t \mapsto f(x^k + t d^k)$ zu finden, zumindest solange dies nicht durch die Armijo-Regel ausgeschlossen wird. Dies ist ein Vorteil gegenüber alternativen Regeln, wie z.B. der Goldstein-Regel, in der neben der Armijo-Regel die Schrittlänge durch

$$f(x^k + t_k d^k) \geq f(x^k) + \eta t_k \langle \nabla f(x^k), d^k \rangle$$

nach unten beschränkt wird.

In der Tat, ist die obige Powell-Wolfe-Bedingung stets erfüllbar

Lemma 2.4.6. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, $x \in \mathbb{R}^n$ ein gegebener Punkt und $d \in \mathbb{R}^n$ eine Abstiegsrichtung von f in x mit*

$$\inf_{t \geq 0} f(x + td) > -\infty.$$

Zu gegebenen $\gamma \in (0,1/2)$ und $\eta \in (\gamma,1)$ gibt es dann ein $t>0$ so dass die Powell-Wolfe-Bedingung 2.4.4 erfüllt ist.

Beweis. Die Abbildung

$$\phi(t) = f(x+td) - f(x) - t\gamma\langle \nabla f(x), d\rangle$$

erfüllt $\phi'(0) = (1-\gamma)\langle \nabla f(x), d\rangle < 0$. Somit ist die stetig differenzierbare Funktion ϕ in einer Umgebung von 0 streng monoton fallend und daher ist $\phi < 0$ auf $(0,\varepsilon)$ für hinreichend kleines ε und damit ist die Menge

$$M = \{t>0 \mid \phi(t)=0\} = \{t \geq \varepsilon \mid \phi(t)=0\}$$

abgeschlossen. Da $f(x+td)$ nach unten beschränkt ist, und $-t\gamma\langle \nabla f(x), d\rangle \to \infty$ für $t\to\infty$ gibt es folglich ein kleinstes Element $t^* \in M$. Damit ist $\phi(t) \leq 0$ auf $(0,t^*]$ und somit ist die Armijo-Bedingung ($\mathscr{A}$) für alle $t \in (0,t^*]$ erfüllt. Weiterhin ist

$$\langle \nabla f(x+t^*d), d\rangle - \gamma\langle \nabla f(x), d\rangle = \phi'(t^*) = \lim_{t\downarrow 0} \frac{\phi(t^*-t)-\phi(t^*)}{-t} \geq 0$$

und somit

$$\langle \nabla f(x+t^*d), d\rangle \geq \gamma\langle \nabla f(x), d\rangle \geq \eta\langle \nabla f(x), d\rangle.$$

Man beachte hierbei $\langle \nabla f(x), d\rangle < 0$. □

Bemerkung 2.4.7. Im Gegensatz zur Armijo-Regel ($\mathscr{A}$) wählen wir nun $\gamma \in (0,1/2)$. Dies hat den Zweck, dass zumindest für eine quadratische Funktion das exakte Minimum entlang der Linie auch tatsächlich die Powell-Wolfe Bedingungen erfüllt.

Bemerkung 2.4.8. Die Konstruktion des folgenden Algorithmus ist angelehnt an ein Intervallschachtelungsverfahren zur Konstruktion des Punktes t^* aus dem Beweis von Lemma 2.4.6. Man beachte, das in der Konstruktion die untere Schranke t_- stets die Armijo-Bedingung ($\mathscr{A}$) erfüllt, und ggf. vergrößert werden muss, solange die Schrittweite zu klein ist um ($\mathscr{PW}$) zu erfüllen. Die Schranke t_+ ist jeweils so gewählt, dass hier die Armijo-Bedingung ($\mathscr{A}$) verletzt ist. Diese Schranke muss ggf. nach unten korrigiert werden. Es ist dann nach Konstruktion

$$t_- \nearrow t^* \swarrow t_+.$$

Algorithmus 2.4.9 (Powell-Wolfe Schrittweitenbestimmung)**.**

- Wähle $\gamma \in (0,1/2)$ und $\eta \in (\gamma,1)$.

1. Falls $t = 1$ der Armijo-Bedingung ($\mathscr{A}$) genügt gehe zu 3.
2. Bestimme die größte Zahl $t_- \in \{2^{-1}, 2^{-2}, \ldots\}$, so dass $t = t_-$ die Armijo-Bedingung ($\mathscr{A}$) erfüllt. Setze $t_+ = 2t_-$ und gehe zu 5.
3. Falls $t = 1$ die Bedingung ($\mathscr{PW}$) erfüllt: Stopp mit $t = 1$.
4. Bestimme die kleinste Zahl $t_+ \in \{2, 2^2, 2^3, \ldots\}$, so dass die Armijo-Bedingung ($\mathscr{A}$) für $t = t_+$ verletzt ist. Setze $t_- = t_+/2$.
5. Solange $t = t_-$ die Bedingung ($\mathscr{PW}$) verletzt:
  a) Berechne $t = \frac{t_- + t_+}{2}$.
  b) Falls t der Bedingung ($\mathscr{A}$) genügt setze $t_- = t$ ansonsten setze $t_+ = t$.
6. Stopp mit $t = t_-$.

Folglich ist der überraschende Punkt, das der obige Algorithmus nicht nur gegen einen Punkt der der Powell-Wolfe-Bedingung genügt konvergiert, sondern in der Tat bereits nach endlich vielen Schritten abbricht.

Theorem 2.4.10. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, $x \in \mathbb{R}^n$ ein gegebener Punkt und $d \in \mathbb{R}^n$ eine Abstiegsrichtung von f in x mit*

$$\inf_{t \geq 0} f(x + td) > -\infty.$$

Zu gegebenen $\gamma \in (0, 1/2)$ und $\eta \in (\gamma, 1)$ terminiert der Algorithmus 2.4.9 nach endlich vielen Schritten mit einer Schrittweite $t > 0$ die den Powell-Wolfe-Bedingungen 2.4.4 genügt.

Beweis. Die Zeile 2. entspricht der bereits bekannten Armijo-Liniensuche ALS [2.3.6]. Nach Theorem 2.3.7 endet die Suche nach endlich vielen Schritten. Da nach Voraussetzung $\inf_{t \geq 0} f(x + td) > -\infty$ und $\langle \nabla f(x), d \rangle < 0$ folgt

$$\phi(t) = f(x + td) - f(x) - t\gamma \langle \nabla f(x), d \rangle \to \infty$$

somit bricht die Suche in Zeile 4. nach endlich vielen Schritten ab.

Zu Beginn der Iteration in Schritt 5 sind nun $t_- < t_+$ gegeben, so dass $t = t_-$ die Armijo-Bedingung ($\mathscr{A}$) erfüllt, aber $t = t_+$ dies nicht tut. Dies heißt $\phi(t_-) \leq 0 < \phi(t_+)$. Angenommen die Iteration in Schritt 5 bricht nicht ab, dann konstruieren wir Folgen $t_-^k < t_+^k$ (und $\phi(t_-^k) \leq 0 < \phi(t_+^k)$) mit $|t_+^k - t_-^k| \leq 2^{-k} |t_+^0 - t_-^0|$. Dabei ist t_-^k monoton nicht fallend und t_+^k monoton nicht wachsend. Es gibt daher einen Grenzwert

$$t^* = \lim_{k \to \infty} t_-^k = \lim_{k \to \infty} t_+^k.$$

Es gilt nun $\phi(t_+^k) > 0$ und daher $0 \leq \phi'(t^*) = \langle \nabla f(x + t^* d), d \rangle - \gamma \langle \nabla f(x), d \rangle$. Es folgt also wie im Beweis von Lemma 2.4.6

$$\langle \nabla f(x + t^* d), d \rangle \geq \gamma \langle \nabla f(x), d \rangle > \eta \langle \nabla f(x), d \rangle.$$

Da $t \mapsto \nabla f(x + td)$ stetig ist dann auch

$$\langle \nabla f(x + t_-^k d), d\rangle \geq \eta \langle \nabla f(x), d\rangle$$

für k hinreichend groß im Widerspruch zur Annahme. Zeile 5 des Algorithmus bricht also nach endlich vielen Schritten ab. □

Wir werden nun sehen, dass die Powell-Wolfe-Bedingung in der Tat zulässige Schrittweiten erzeugt.

Theorem 2.4.11. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und nach unten beschränkt, d.h. $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$. Dann ist das Abstiegsverfahren 2.2.1 mit Powell-Wolfe Schrittweitenregel 2.4.9 durchführbar. Ist dann $(x^k)_K$ eine konvergente Teilfolge mit Limes $\overline{x}$, so ist die zugehörige Schrittweitenteilfolge t_k zulässig im Sinn von Definition 2.2.7.*

Beweis. Da f nach unten beschränkt ist, ist Theorem 2.4.10 für jeden Punkt x^k und jede Richtung d^k anwendbar. Das Abstiegsverfahren ist also durchführbar.

Offenkundig ist die erzeugte Folge $f(x^k)$ monoton fallend und wir erhalten (2.5). Um die Zulässigkeit der Schrittweiten zu sehen nehmen wir nun an

$$\left(\frac{\langle \nabla f(x^k), d^k\rangle}{\|d^k\|}\right)_K \not\to 0.$$

Es gibt dann ein $\varepsilon > 0$, so dass (ggf. mit einer weiteren Teilfolge)

$$\frac{-\langle \nabla f(x^k), d^k\rangle}{\|d^k\|} \geq \varepsilon \quad \forall k \in K.$$

Aus der Bedingung ($\mathscr{P}W$) ergibt sich dann

$$\begin{aligned}\|\nabla f(x^k + t_k d^k) - \nabla f(x^k)\|\|d^k\| &\geq \langle \nabla f(x^k + t_k d^k) - \nabla f(x^k), d^k\rangle \\ &\geq (\eta - 1)\langle \nabla f(x^k), d^k\rangle \\ &= -(1-\eta)\langle \nabla f(x^k), d^k\rangle \\ &\geq (1-\eta)\varepsilon\|d^k\|\end{aligned}$$

für alle $k \in K$. Wegen der Stetigkeit von ∇f gibt es ein $\delta > 0$, so dass

$$\|\nabla f(x) - \nabla f(\overline{x})\| \leq (1-\eta)\varepsilon/3$$

für alle $x \in B_\delta(\overline{x})$. Da für fast alle $k \in K$ gilt $x^k \in B_{\delta/2}(\overline{x})$ folgt somit $x^k + t_k d^k \notin B_\delta(\overline{x})$, ansonsten wäre $\|\nabla f(x^k + t_k d^k) - \nabla f(x^k)\| < (1-\eta)\varepsilon$. Damit ist dann notwendig $t_k\|d^k\| \geq$

$\delta/2$ für fast alle k und es folgt aus der Armijo-Regel ($\mathscr{A}$)

$$\begin{aligned} f(x^k) - f(x^k + t_k d^k) &\geq -t_k \gamma \langle \nabla f(x_k), d^k \rangle \\ &= -t_k \|d^k\| \gamma \frac{\langle \nabla f(x_k), d^k \rangle}{\|d^k\|} \\ &\geq \delta/2 \gamma \varepsilon. \end{aligned}$$

Somit folgt $f(x^k) - f(x^k + t_k d^k) \not\to 0$ und damit wegen (2.6) die Zulässigkeit der Schrittweiten. □

2.5 Das Newton-Verfahren

Wir wenden uns nun einer alternativen Möglichkeit zur Bestimmung einer passenden Suchrichtung zu. Dies ist das sogenannte Newton-Verfahren für das wir zunächst einmal zwei Herleitungen betrachten wollen.

Sicht 1: Zunächst einmal überlegen wir uns, dass wir zur Bestimmung einer geeigneten Suchrichtung versuchen können ein geeignetes Modell von f zu Minimieren. Im einfachsten Fall nehmen wir hierzu eine lineare Approximation, d.h.

$$\tilde{f}(z) = f(x) + \langle \nabla f(x), z - x \rangle.$$

Um dieses zu Minimieren benötigt man zunächst eine geeignete Schranke an z, da das lineare Modell auf $\mathbb{R}^n$ nach unten unbeschränkt ist. Betrachtet man z.B. zur Bestimmung von $d = z - x$ das Problem

$$\min \tilde{f}(z) \quad \text{u.d.N.} \quad \|z - x\|_2 \leq 1$$

so erhält man die bekannte Richtung des steilsten Abstiegs,

$$d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

Wir müssen also zu einem besseren Modell übergehen. Wählt man eine quadratische Approximation an f so ergibt sich das folgende Problem

$$\min_{z \in \mathbb{R}^n} \tilde{f}(z) = f(x) + \langle \nabla f(x), z - x \rangle + \frac{1}{2} \langle z - x, \nabla^2 f(x)(z - x) \rangle$$

zur Bestimmung von $d = z - x$. Ist $\nabla^2 f(x)$ positiv definit, so besitzt dieses Problem eine Lösung, und diese erfüllt notwendig $\nabla \tilde{f}(z) = 0$ bzw.

$$\nabla^2 f(x) d = -\nabla f(x).$$

Sicht 2: Alternativ, kann man sich das Newton-Verfahren auch, wie dies oft in der Numerik geschieht, direkt für das Lösen der Gleichung

$$0 = \nabla f(x) =: F(x)$$

ansehen. Taylorentwicklung 1.2.1 ergibt dann

$$F(x+d) = F(x) + F'(x)d + o(\|d\|).$$

Dies legt die folgende Iterationsvorschrift

$$0 = F(x) + F'(x)d$$

zur Bestimmung von d nahe. Wegen $F(x) = \nabla f(x)$ und $F'(x) = \nabla^2 f(x)$ ist dies identisch zu unserer ersten Herleitung. Die beiden Ansätze liefern zwei verschiedene Sichtweisen auf das Problem der Richtungsbestimmung die beide gewisse Vorteile haben. Wir werden uns zunächst auf die zweite Sichtweise einlassen.

Bemerkung 2.5.1. Bevor wir dies tun betrachten wir kurz die Frage, ob die oben bestimmte Richtung d überhaupt eine vernünftige Wahl, also eine Abstiegsrichtung ist. Durch Testen der obigen Vorschrift mit d erhalten wir

$$\langle \nabla f(x), d\rangle = -\langle d, \nabla^2 f(x)d\rangle.$$

Wir sehen also, dass d eine Abstiegsrichtung ist, falls $\nabla^2 f(x)$ positiv definit ist. Ist dies nicht der Fall, so kann es passieren, dass die Gleichung $\nabla^2 f(x)d = -\nabla f(x)$ nicht, oder nicht eindeutig, lösbar ist ($\nabla^2 f(x)$ ist singulär), oder sogar, dass d eine Aufstiegsrichtung ist ($\nabla^2 f(x)$ hat negative Eigenwerte). Es ist also im Allgemeinen weder die Bestimmbarkeit von d noch ein Abstieg entlang von d garantiert. Wir benötigen also hierfür zusätzliche Bedingungen.

2.5.1 Das lokale Newton-Verfahren für Gleichungssysteme

Für den folgenden Abschnitt stellen wir uns auf den Standpunkt ein nichtlineares Gleichungssystem

$$F(x) = 0$$

für eine Funktion $F \in C^1(\mathbb{R}^n;\mathbb{R}^n)$ lösen zu wollen. Wir erhalten dann aus den vorherigen Überlegungen:

Algorithmus NVG [2.5.2] (Lokales Newton-Verfahren für Gleichungssysteme)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$.

for k = 0, 1, ... **do**

- Prüfe auf Abbruch: Stopp falls $F(x^k) = 0$
- Berechne den Newtonschritt $d^k \in \mathbb{R}^n$ durch Lösen der Gleichung

$$F'(x^k)d^k = -F(x^k).$$

- Setze $x^{k+1} = x^k + d^k$.

end for

Wir erinnern an folgende Aussage

Lemma 2.5.3 (Menge der invertierbaren Matrizen)**.** *Die Menge $\mathscr{M} \subset \mathbb{R}^{n\times n}$ der invertierbaren Matrizen ist offen und die Abbildung $\mathscr{M} \ni M \mapsto M^{-1}$ ist stetig. Genauer ist für $A \in \mathscr{M}$ und beliebiges $B \in \mathbb{R}^{n\times n}$ mit $\|A^{-1}B\| < 1$ die Matrix $A+B$ invertierbar und es ist*

$$\|(A+B)^{-1}\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}B\|},$$
$$\|(A+B)^{-1} - A^{-1}\| \leq \frac{\|A^{-1}\|\|A^{-1}B\|}{1-\|A^{-1}B\|}.$$

Beweis. Die Neumannreihe $S = \sum_{k=0}^{\infty}(-A^{-1}B)^k$ ist nach Voraussetzung normal konvergent, da

$$\|S\| \leq \sum_{k=0}^{\infty}\|A^{-1}B\|^k = \frac{1}{1-\|A^{-1}B\|} < \infty.$$

Es folgt

$$I = (I + A^{-1}B)\sum_{k=0}^{\infty}(-A^{-1}B)^k$$

folglich ist $I + A^{-1}B \in \mathscr{M}$ mit $(I + A^{-1}B)^{-1} = S$ und wir erhalten

$$\begin{aligned}
\|(A+B)^{-1}\| &= \|(A+AA^{-1}B)^{-1}\| \\
&= \|(A(I+A^{-1}B))^{-1}\| \\
&= \|(I+A^{-1}B)^{-1}A^{-1}\| \\
&\leq \|A^{-1}\|\|S\| \\
&\leq \frac{\|A^{-1}\|}{1-\|A^{-1}B\|}, \\
\|(A+B)^{-1} - A^{-1}\| &= \|SA^{-1} - A^{-1}\| \\
&\leq \|A^{-1}\|\sum_{k=1}^{\infty}\|A^{-1}B\|^k \\
&= \|A^{-1}\|\|A^{-1}B\|\sum_{k=0}^{\infty}\|A^{-1}B\|^k \\
&= \frac{\|A^{-1}\|\|A^{-1}B\|}{1-\|A^{-1}B\|}.
\end{aligned}$$

□

Lemma 2.5.4 (Nullstellen differenzierbarer Funktionen). *Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar. Sei $F(\overline{x}) = 0$ und $F'(\overline{x}) \in \mathscr{M}$ (also invertierbar). Dann gibt es $\varepsilon > 0$ und $\gamma > 0$ mit*

$$\|F(x)\| \geq \gamma \|x - \overline{x}\| \quad \forall\, x \in B_\varepsilon(\overline{x}).$$

Insbesondere ist $\overline{x}$ eine isolierte Nullstelle von F.

Beweis. Es ist offenkundig

$$\|x - \overline{x}\| \leq \|F'(\overline{x})^{-1}\| \|F'(\overline{x})(x - \overline{x})\|.$$

Setzen wir nun

$$\gamma = \frac{1}{2\|F'(\overline{x})^{-1}\|}$$

so folgt

$$2\gamma \|x - \overline{x}\| \leq \|F'(\overline{x})(x - \overline{x})\|.$$

Wegen der Taylorentwicklung 1.2.1 gibt es ein $\varepsilon > 0$, so dass

$$\|F(x) - F(\overline{x}) - F'(\overline{x})(x - \overline{x})\| \leq \gamma \|x - \overline{x}\| \quad \forall\, x \in B_\varepsilon(\overline{x}).$$

Da $F(\overline{x}) = 0$ folgt

$$\begin{aligned} 2\gamma \|x - \overline{x}\| &\leq \|F'(\overline{x})(x - \overline{x})\| \\ &= \|F(x) - [F(x) - F(\overline{x}) - F'(\overline{x})(x - \overline{x})]\| \\ &\leq \|F(x)\| + \|F(x) - F(\overline{x}) - F'(\overline{x})(x - \overline{x})\| \\ &\leq \|F(x)\| + \gamma \|x - \overline{x}\| \end{aligned}$$

und somit die Behauptung. □

Bevor wir uns an die Analyse von Algorithmus NVG [2.5.2] machen benötigen wir die folgenden Begriffe

Definition 2.5.5. Eine Folge $x^k \in \mathbb{R}^n$ konvergiert

1. *q-linear* (Quotienten-linear) mit Rate $0 < \gamma < 1$ gegen $\overline{x} \in \mathbb{R}^n$, falls es ein $l \geq 0$ gibt, so dass
$$\|x^{k+1} - \overline{x}\| \leq \gamma \|x^k - \overline{x}\| \quad \forall\, k \geq l.$$

2. *q-superlinear* gegen $\overline{x} \in \mathbb{R}^n$, falls es eine Folge $\mathbb{R} \ni c_k \to 0$ gibt, so dass
$$\|x^{k+1} - \overline{x}\| \leq c_k \|x^k - \overline{x}\|.$$

3. *q-quadratisch* gegen $\overline{x} \in \mathbb{R}^n$, falls es ein $C > 0$ gibt, so dass

$$\|x^{k+1} - \overline{x}\| \leq C\|x^k - \overline{x}\|^2.$$

Wir können nun die lokale Durchführbarkeit und Konvergenz des Newtonverfahrens NVG [2.5.2] zeigen.

Theorem 2.5.6 (Lokale Konvergenz des Newton-Verfahrens). *Sei $F\colon \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\overline{x}$ ein Punkt mit $F(\overline{x}) = 0$, in dem $F'(\overline{x}) \in \mathscr{M}$ ist. Dann gibt es $\delta, C > 0$, so dass gilt:*

1. *$\overline{x}$ ist die einzige Nullstelle von F auf $B_\delta(\overline{x})$.*
2. *$F'(x) \in \mathscr{M}$ für alle $x \in B_\delta(\overline{x})$ und*

$$\sup_{\|x-\overline{x}\|<\delta} \|F'(x)^{-1}\| \leq C.$$

   Insbesondere ist ein Schritt des Newtonverfahrens NVG [2.5.2] durchführbar falls $x^k \in B_\delta(\overline{x})$.
3. *Für alle $x^0 \in B_\delta(\overline{x})$ erfüllen die Iterierten des Newtonverfahrens NVG [2.5.2] $x^k \in B_\delta(\overline{x})$. Insbesondere ist das Newtonverfahren durchführbar.*
4. *Das Newtonverfahren bricht dann entweder nach endlich vielen Iterationen mit $x^k = \overline{x}$ ab, oder es erzeugt eine Folge $x^k \subset B_\delta(\overline{x})$, die q-superlinear gegen $\overline{x}$ konvergiert.*
5. *Ist F' sogar Lipschitz stetig auf $B_\delta(\overline{x})$, d.h. es gibt ein $L > 0$, so dass*

$$\|F'(x) - F'(y)\| \leq L\|x - y\| \quad \forall\, x, y \in B_\delta(\overline{x}).$$

   Dann ist die Konvergenz q-quadratisch mit

$$\|x^{k+1} - \overline{x}\| \leq \frac{CL}{2}\|x^k - \overline{x}\|^2 \quad \forall k \geq 0.$$

Beweis. **1.** Aufgrund von Lemma 2.5.4 ist die Existenz einer isolierten Nullstelle auf $B_\delta(\overline{x})$ gesichert.
2. Nach Annahme ist $F'(\overline{x}) \in \mathscr{M}$. Da $\mathscr{M}$ nach Lemma 2.5.3 offen ist können wir obiges $\delta > 0$ auch so klein wählen, dass $F'(x) \in \mathscr{M}$ für alle $x \in B_\delta(\overline{x})$ gilt, und wegen der ersten Ungleichung in Lemma 2.5.3 auch

$$\sup_{x \in B_\delta} \|F'(x)^{-1}\| \leq C$$

für alle $\delta \leq 1$. Damit ist die Newtoniteration für beliebiges $x \in B_\delta(\overline{x})$ ausführbar. Da $F \in C^1$ folgt zudem

$$\sup_{x,y\in B_\delta(\overline{x})} C\|F'(y)-F'(x)\| := c_\delta \to 0 \quad (\delta \to 0)$$

3. Wegen $F(\overline{x}) = 0$ erhalten wir für gegebenes $x^k \in B_\delta(\overline{x})$ für die nächste Newtoniterierte $x^{k+1} = x^k + d^k$ die folgende Abschätzung

$$\begin{aligned}\|x^{k+1}-\overline{x}\| &= \|F'(x^k)^{-1}F'(x^k)(x^{k+1}-\overline{x})\|\\ &\le C\|F'(x^k)(x^{k+1}-\overline{x})\|\\ &= C\|F'(x^k)(x^k+d^k-\overline{x})\|\\ &= C\|F'(x^k)(x^k-\overline{x})-F(x^k)\|\\ &= C\|F'(x^k)(x^k-\overline{x})+F(\overline{x})-F(x^k)\|.\end{aligned}$$

Nun setzen wir $G(t) := F(x^k + t(\overline{x}-x^k))$ und damit $G'(t) = F'(x^k + (t(\overline{x}-x^k)))(\overline{x}-x^k)$. Damit ist durch den Fundamentalsatz der Differential- und Integralrechnung

$$F(\overline{x})-F(x^k) = G(1)-G(0) = \int_0^1 G'(s)\,\mathrm{d}s$$

und wir erhalten

$$\begin{aligned}\|x^{k+1}-\overline{x}\| &\le C\left\|\left(\int_0^1 F'(x^k+s(\overline{x}-x^k))\,\mathrm{d}s - F'(x^k)\right)(\overline{x}-x^k)\right\| \qquad (2.11)\\ &\le \sup_{y\in B_\delta(\overline{x})} C\|F'(y)-F'(x^k)\|\|\overline{x}-x^k\|\\ &\le c_\delta\|\overline{x}-x^k\|.\end{aligned}$$

In dem wir nun $\delta > 0$ ggf. verkleinern können wir sicherstellen, dass

$$c_\delta \le \frac{1}{2}.$$

Dann folgt aus obiger Rechnung wegen $x^k \in B_\delta(\overline{x})$

$$\|x^{k+1}-\overline{x}\| \le \frac{1}{2}\|x^k-\overline{x}\| \le \frac{\delta}{2}$$

und somit per Induktion über k die Behauptung von 3.
4. Detailliertere Betrachtung der Argumentation in 3. zeigt in der Tat

$$\|x^k-\overline{x}\| \le \frac{1}{2}\|x^{k-1}-\overline{x}\| \le \left(\frac{1}{2}\right)^k\|x^0-\overline{x}\| \le \left(\frac{1}{2}\right)^k\delta =: \delta_k \to 0 \quad (k\to 0)$$

und damit weiter wegen (2.11)

$$\frac{\|x^{k+1}-\overline{x}\|}{\|x^k-\overline{x}\|} \le c_{\delta_k} \to 0 \quad (k\to\infty).$$

5. Ist nun F' L-stetig mit Konstante $L > 0$, so folgt aus der obigen Ausführung, insbesondere (2.11),

$$\begin{aligned}
\|x^{k+1} - \overline{x}\| &\leq C \left\| \left(\int_0^1 F'(x^k + s(\overline{x} - x^k))\, \mathrm{d}s - F'(x^k) \right)(\overline{x} - x^k) \right\| \\
&\leq C \int_0^1 \left\| F'(x^k + s(\overline{x} - x^k)) - F'(x^k) \right\| \mathrm{d}s \, \|\overline{x} - x^k\| \\
&\leq CL \int_0^1 \|x^k + s(\overline{x} - x^k) - x^k\| \, \mathrm{d}s \, \|\overline{x} - x^k\| \\
&\leq CL \int_0^1 s \, \mathrm{d}s \|\overline{x} - x^k\|^2 \\
&= \frac{CL}{2} \|\overline{x} - x^k\|^2
\end{aligned}$$

und damit die q-quadratische Konvergenz. □

2.5.2 Das lokale Newton-Verfahren für Optimierungsprobleme

Wir wollen nun die Anwendung des Newton-Verfahrens für Optimierungsprobleme diskutieren. Hierzu wenden wir das soeben besprochene Verfahren auf die notwendige Bedingung

$$F(x) := \nabla f(x) = 0$$

an. Ist $f \in C^2(\mathbb{R}^n; \mathbb{R})$, so ist $\nabla f \in C^1(\mathbb{R}^n; \mathbb{R}^n)$. Wir erhalten dadurch aus Algorithmus NVG [2.5.2] das Folgende.

Algorithmus NVO [2.5.7] (Lokales Newton-Verfahren für unrestringierte Optimierung)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$.

for k = 0, 1, ... **do**

- Prüfe auf Abbruch: Stopp falls $\nabla f(x^k) = 0$
- Berechne den Newtonschritt $d^k \in \mathbb{R}^n$ durch Lösen der Gleichung

$$\nabla^2 f(x^k) d^k = -\nabla f(x^k).$$

- Setze $x^{k+1} = x^k + d^k$.

end for

Bemerkung 2.5.8. Man beachte, das der Beweis der lokalen Konvergenz des Newtonverfahrens für Gleichungen 2.5.6 keinerlei Informationen über die Vorzeichen der Eigenwerte von $F'(\overline{x})$ verwendet. Das Verfahren konvergiert also lokal ebenso q-superlinear gegen

Maxima mit negativ definiter Hesse-Matrix, oder auch Sattelpunkte, solange hier alle Eigenwerte von $\nabla^2 f(\overline{x})$ von Null verschieden sind.

Wie wir bereits in der Einleitung zum Newton-Verfahren gesehen haben liefert uns das obige Verfahren nur dann Abstiegsrichtungen, wenn $\nabla^2 f(x^k)$ positiv definit ist, bzw. der kleinste Eigenwert $\lambda_{\min}(\nabla^2 f(x^k)) > 0$. Das diese Eigenschaft stetig bezüglich der Koeffizienten ist zeigt folgendes Lemma.

Lemma 2.5.9. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit. Dann gilt für alle $\mu \in (0, \lambda_{\min}(A))$ und alle symmetrischen $B \in \mathbb{R}^{n\times n}$ mit $\|B\| \leq \lambda_{\min}(A) - \mu$*

$$\mu \leq \lambda_{\min}(A+B) \leq 2\lambda_{\min}(A) - \mu$$

Beweis. Es gilt

$$\begin{aligned}
\lambda_{\min}(A+B) &= \min_{\|d\|=1} \langle d, (A+B)d\rangle \\
&\geq \min_{\|d\|=1} \langle d, Ad\rangle + \min_{\|d\|=1} \langle d, Bd\rangle \\
&\geq \lambda_{\min}(A) - \max_{\|d\|=1} \langle d, Bd\rangle \\
&\geq \lambda_{\min}(A) - \max_{\|d\|=1} \|d\|\|Bd\| \\
&= \lambda_{\min}(A) - \|B\| \geq \mu.
\end{aligned}$$

Für die obere Schranke rechnen wir

$$\begin{aligned}
\lambda_{\min}(A+B) &= \max_{\|d\|=1} -\langle d, (A+B)d\rangle \\
&\leq \max_{\|d\|=1} -\langle d, Ad\rangle + \max_{\|d\|=1} -\langle d, Bd\rangle \\
&\leq \lambda_{\min}(A) + \max_{\|d\|=1} -\langle d, Bd\rangle \\
&\leq \lambda_{\min}(A) + \|-B\| \leq 2\lambda_{\min}(A) - \mu.
\end{aligned}$$

□

Insbesondere ist für $\|B\| \to 0$, bzw. $\mu \to \lambda_{\min}(A)$

$$\lambda_{\min}(A+B) \to \lambda_{\min}(A) \quad (\|B\| \to 0).$$

Wir können nun folgendes Theorem zeigen

Theorem 2.5.10 (Lokale Konvergenz des Newton-Verfahrens für Optimierungsprobleme)**.**

Sei $f \in C^2(\mathbb{R}^n;\mathbb{R})$ und $\overline{x} \in \mathbb{R}^n$ ein lokales Minimum von f in dem die hinreichenden Bedingungen zweiter Ordnung 2.1.8 gelten. Dann gibt es $\delta > 0$ und $\mu > 0$, so dass

1. *$\overline{x}$ ist der einzige stationäre Punkt auf $B_\delta(\overline{x})$.*
2. *$\lambda_{\min}(\nabla^2 f(x)) \geq \mu$ für alle $x \in B_\delta(\overline{x})$.*
3. *Für alle Startwerte $x^0 \in B_\delta(\overline{x})$ terminiert Algorithmus NVO [2.5.7] entweder mit $x^k = \overline{x}$, oder er erzeugt eine Folge $x^k \in B_\delta(\overline{x})$, die q-superlinear gegen $\overline{x}$ konvergiert.*
4. *Ist $\nabla^2 f$ L-stetig auf $B_\delta(\overline{x})$ mit Konstante L, d.h.*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \quad \forall x, y \in B_\delta(\overline{x}),$$

so ist die Konvergenzrate sogar q-quadratisch mit

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

Beweis. **1.** Wegen der angenommenen positiv Definitheit von $\nabla^2 f(\overline{x})$ ist $\nabla^2 f(\overline{x}) \in \mathscr{M}$ und die Behauptung folgt aus Lemma 2.5.4 für die Abbildung $F(x) = \nabla f(x)$.
2. Wegen Lemma 2.5.9 und der angenommenen Stetigkeit von $\nabla^2 f$ ist $\lambda_{\min}(\nabla^2 f(x)) \geq \mu$ auf $B_\delta(\overline{x})$, wobei wir ggf. δ verkleinern müssen.
3. Wegen 2. ist nun

$$\|\nabla^2 f(x)^{-1}\| \leq \frac{1}{\lambda_{\min}(\nabla^2 f(x))} \leq \frac{1}{\mu} \quad \forall x \in B_\delta(\overline{x}).$$

Die Behauptung folgt damit aus Theorem 2.5.6 Teil 3.-5. für die Funktion $F(x) = \nabla f(x)$ mit $C = \frac{1}{\mu}$. □

Wir haben bereits in der Herleitung des Newtonverfahrens gesehen, dass es i.A. keine Lösung d der Newton-Gleichung $\nabla^2 f(x)d = -\nabla f(x)$ geben muss. Selbst wenn es eine Lösung gibt, so ist diese nicht notwendig eine Abstiegsrichtung, es sei denn $\nabla^2 f(x)$ ist positiv definit. Wir können daher für beliebige Startwerte weder die Durchführbarkeit noch die Konvergenz des obigen lokalen Newtonverfahrens garantieren.

Wie das folgende Beispiel zeigt kann auch für positiv definites $\nabla^2 f(x)$ lediglich lokale Konvergenz erwartet werden.

Beispiel 2.5.11. Sie $f : \mathbb{R} \to \mathbb{R}$ durch $f(x) = \sqrt{x^2+1}$ gegeben. Dann ist

$$\nabla f(x) = \frac{x}{\sqrt{x^2+1}}, \quad \nabla^2 f(x) = \frac{1}{\sqrt{x^2+1}} - \frac{x^2}{(x^2+1)^{3/2}} = \frac{1}{(x^2+1)^{3/2}} > 0.$$

Die Hesse-Matrix ist also stets positiv definit und die Newton-Iteration

$$\frac{1}{((x^k)^2+1)^{3/2}} d^k = -\frac{x^k}{\sqrt{(x^k)^2+1}}, \quad x^{k+1} = x^k + d^k$$

durchführbar und es ist $d^k = -x^k((x^k)^2+1)$ und damit $x^{k+1} = x^k + d^k = -(x^k)^3$. Es ist somit

1. $x^k \to 0$ für $|x^0| < 1$, die Konvergenz ist dann sogar q-kubisch.
2. $|x^k| \to \infty$ für $|x^0| > 1$ (Divergenz).
3. $x^k = (-1)^k x^0$ für $|x^0| = 1$ (Oszillation).

Um dieses, und auch die anderen Probleme der Durchführbarkeit des Newtonverfahrens zu lösen benötigen wir sowohl eine geeignete Schrittweite, und ggf. auch eine alternative Suchrichtung. Wir werden dies nun im folgenden Abschnitt erreichen.

2.5.3 Globalisiertes Newtonverfahren

Algorithmus NV [2.5.12] (Globales Newton-Verfahren)**.**

• Wähle Startpunkt $x^0 \in \mathbb{R}^n$, $\beta \in (0,1)$, $\gamma \in (0,1)$, $\alpha_1, \alpha_2 > 0$ und $p > 0$.
(Es ist zweckmäßig für die schnelle Konvergenz $\gamma \in (0, 1/2)$ zu wählen)
for k = 0, 1, … **do**
  • Prüfe auf Abbruch: Stopp falls $\nabla f(x^k) = 0$
  • Berechne den Newtonschritt $\tilde{d}^k \in \mathbb{R}^n$ durch Lösen der Gleichung

$$\nabla^2 f(x^k)\tilde{d}^k = -\nabla f(x^k).$$

  Falls dies eindeutig möglich ist, und $\tilde{d}^k$ die Bedingung

$$-\langle \nabla f(x^k), \tilde{d}^k \rangle \geq \min(\alpha_1, \alpha_2 \|\tilde{d}^k\|^p) \|\tilde{d}^k\|^2$$

  erfüllt, so setzte $d^k = \tilde{d}^k$. Andernfalls setze $d^k = -\nabla f(x^k)$.
  • Bestimme eine Schrittweite $t_k > 0$ mithilfe der Armijo-Liniensuche ALS [2.3.6] mit Parametern β und γ.
  • Setze $x^{k+1} = x^k + t_k d^k$.
end for

Bemerkung 2.5.13. Für die Bedingung

$$-\langle \nabla f(x^k), d\rangle \geq \min(\alpha_1, \alpha_2 \|d\|^p)\|d\|^2$$

gibt es eine Vielzahl von Varianten, die alle implizieren, dass die akzeptierten Suchrichtungen auch tatsächlich Abstiegsrichtungen sind. Die nicht ganz so offenkundige zweite Eigenschaft der Bedingungen ist die, dass Sie es uns ermöglichen unter gewissen Bedingungen die Newtonrichtung zu akzeptieren – Die Bedingung sollte also erfüllbar sein, wenn wir nahe an einem Minimum mit positiv definiter Hesse-Matrix sind. Wir werden dies später noch im Detail untersuchen.

Wir werden nun sehen, dass das obige globalisierte Newtonverfahren global konvergent ist. Es gilt nämlich

Theorem 2.5.14. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Dann terminiert Algorithmus NV [2.5.12] entweder mit $\nabla f(x^k) = 0$, oder er erzeugt eine unendliche Folge x^k, deren Häufungspunkte stationäre Punkte von f sind.*

Beweis. 1. Zunächst müssen wir sehen, dass die Richtungen d^k auch tatsächlich Abstiegsrichtungen sind. Im Falle $d^k = -\nabla f(x^k)$ haben wir dies bereits gesehen. Im anderen Fall erfüllt d^k die Gleichung $\nabla^d f(x^k)d^k = -\nabla f(x^k)$. Da diese nach Annahme eindeutig lösbar ist, folgt $d^k \neq 0$ und aus der zweiten Bedingung erhalten wir

$$\frac{-\langle \nabla f(x^k), d^k\rangle}{\|d^k\|} \geq \min(\alpha_1, \alpha_2\|d^k\|^p)\|d^k\| > 0.$$

Die Richtung d^k ist also eine Abstiegsrichtung, folglich ist Algorithmus NV [2.5.12] durchführbar. Wir definieren nun

$$K_G = \{k \mid d^k = -\nabla f(x^k)\},$$
$$K_N = \mathbb{N}_0 \setminus K_G.$$

2. Wir zeigen nun die Konvergenz des Verfahrens. Im Falle $\nabla f(x^k) = 0$ für ein k ist nichts zu zeigen. Ansonsten sei $(x^k)_{k\in K}$ eine konvergente Teilfolge mit Limes $\overline{x}$.

3. Wir zeigen nun, dass die Richtungen d^k zulässig sind. Da $x^k \to \overline{x}$ für $K \ni k \to \infty$ folgt

$$\|\nabla^2 f(x^k)\| \leq C \quad k \in K$$

und damit für $k \in K_N \cap K$ notwendig

$$\|\nabla f(x^k)\| = \|\nabla^2 f(x^k)d^k\| \leq C\|d^k\|.$$

Sei nun zum Nachweis der Zulässigkeit angenommen, dass

$$\left(\frac{\langle \nabla f(x^k), d^k\rangle}{\|d^k\|}\right)_K \to 0.$$

Dann folgt – falls jeweils $K_G \cap K$ bzw. $K_N \cap K$ nicht endlich sind – aus der Auswahlbedingung für die Suchrichtung und obiger Rechnung

$$\begin{aligned}
0 \leftarrow -\frac{\langle \nabla f(x^k), d^k\rangle}{\|d^k\|} &= \|\nabla f(x^k)\| \quad (k \in K_G \cap K),\\
0 \leftarrow \frac{-\langle \nabla f(x^k), d^k\rangle}{\|d^k\|} &\quad (k \in K_N \cap K)\\
&\geq \min(\alpha_1, \alpha_2 \|d^k\|^p)\|d^k\|\\
&\geq \frac{\min\left(\alpha_1, \frac{\alpha_2}{C^p}\|\nabla f(x^k)\|^p\right)}{C}\|\nabla f(x^k)\|.
\end{aligned}$$

und somit die Zulässigkeit der Suchrichtungen d^k.

4. Für die Schrittweiten t_k ist zunächst

$$\begin{aligned}
\|d^k\| &= -\frac{\langle \nabla f(x^k), d^k\rangle}{\|d^k\|} \quad (k \in K_G \cap K),\\
\|d^k\| &\geq \frac{1}{C}\|\nabla f(x^k)\| \geq \frac{-1}{C}\frac{\langle \nabla f(x^k), d^k\rangle}{\|d^k\|} \quad (k \in K_N \cap K).
\end{aligned}$$

Damit ist also

$$\|d^k\| \geq \phi\left(-\frac{\langle \nabla f(x^k), d^k\rangle}{\|d^k\|}\right)$$

für die streng monoton wachsende Funktion $\phi(t) = \min(t, t/C)$. Nach Theorem 2.4.2 liefert die Armijo-Liniensuche ALS [2.3.6] zulässige Schrittweiten t_k.

5. Die Konvergenz folgt damit aus dem globalen Konvergenzresultat Theorem 2.2.10 für das allgemeine Abstiegsverfahren.

□

2.5.4 Übergang zu schneller lokaler Konvergenz

Es bleibt die Frage, ob das gedämpfte Newton-Verfahren auch tatsächlich irgendwann die Newton-Richtung verwendet und für diese die Schrittweite $t_k = 1$ bestimmt. Um dies zu untersuchen, benötigen wir einige Hilfsaussagen.

Lemma 2.5.15. *Sei $\overline{x} \in \mathbb{R}^n$ ein isolierter Häufungspunkt der Folge $x^k \in \mathbb{R}^n$. Für jede gegen $\overline{x}$ konvergente Teilfolge $(x^k)_K$ gelte $(x^{k+1} - x^k)_K \to 0$. Dann konvergiert die gesamte Folge gegen $\overline{x}$.*

Beweis. Wie zeigen die Aussage durch ein Widerspruchsargument. Angenommen x^k konvergiert nicht gegen $\overline{x}$. Dann gäbe es $\varepsilon > 0$, so dass $\overline{x}$ der einzige Häufungspunkt in $\overline{B_\varepsilon(\overline{x})}$ ist, und gleichzeitig unendlich viele Folgeglieder außerhalb von $B_\varepsilon(\overline{x})$ liegen. Folglich gäbe es eine Teilfolge $(x^k)_K$ mit der Eigenschaft

$$x^k \in \overline{B_\varepsilon(\overline{x})}, \quad x^{k+1} \notin B_\varepsilon(\overline{x}) \quad \forall k \in K.$$

Da $\overline{B_\varepsilon(\overline{x})}$ kompakt ist besitzt die Teilfolge $(x^k)_K$ einen Häufungspunkt in $\overline{B_\varepsilon(\overline{x})}$. Da $\overline{x}$ der einzige Kandidat hierfür ist folgt

$$x^k \to \overline{x} \quad (K \ni k \to \infty).$$

Aus der Annahme folgt nun

$$x^{k+1} - \overline{x} = x^{k+1} - x^k + x^k - \overline{x} \to 0 \quad (K \ni k \to \infty)$$

im Widerspruch zu $x^{k+1} \notin B_\varepsilon(\overline{x})$. □

Lemma 2.5.16. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Algorithmus NV [2.5.12] erzeuge eine unendliche Folge x^k und $\overline{x} \in \mathbb{R}^n$ sei ein Häufungspunkt von x^k, in dem die Hesse-Matrix positiv definit ist. Dann ist $\overline{x}$ ein isoliertes lokales Minimum von f und die gesamte Folge x^k konvergiert gegen $\overline{x}$.*

Beweis. Nach Voraussetzung ist Theorem 2.5.14 anwendbar, folglich ist $\overline{x}$ ein stationärer Punkt. Da dann nach Voraussetzung in $\overline{x}$ die hinreichenden Bedingungen zweiter Ordnung (Theorem 2.1.8) gelten ist also $\overline{x}$ ein isoliertes lokales Minimum.

Aufgrund des Theorems 2.5.10 zur lokalen Konvergenz des Newton-Verfahrens wissen wir, dass es $\delta > 0$ und $\mu > 0$ gibt, so dass $\overline{x}$ der einzige stationäre Punkt in $B_\delta(\overline{x})$ ist und

$$\lambda_{\min}(\nabla^2 f(x)) \geq \mu \quad \forall\, x \in B_\delta(\overline{x}).$$

Es bleibt also nur noch zu zeigen, dass die ganze Folge gegen $\overline{x}$ konvergiert. O.B.d.A. können wir annehmen, dass $x^k \in B_\delta(\overline{x})$ für alle $k \in K$ eine konvergente Teilfolge mit Limes $\overline{x}$ ist.

Wir können dann für beliebiges $k \in K$ folgendes sehen:

1. Falls $d^k = -\nabla f(x^k)$, so ist

$$\|x^{k+1} - x^k\| = t_k \|d_k\| \leq \|d_k\| = \|\nabla f(x^k)\|.$$

2. Falls $\nabla^2 f(x^k)d^k = -\nabla f(x^k)$, so ist

$$\|x^{k+1} - x^k\| \leq \|d_k\| \leq \frac{1}{\mu}\|\nabla f(x^k)\|.$$

Da $\nabla f(x^k) \to 0$ für $K \ni k \to \infty$ folgt somit $\|x^{k+1} - x^k\| \to 0$ für $K \ni k \to \infty$. Aufgrund des vorherigen Lemmas 2.5.15 folgt die Konvergenz der gesamten Folge $x^k \to \overline{x}$. □

Lemma 2.5.17. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $\overline{x} \in \mathbb{R}^n$ ein lokales Minimum von f, in dem die hinreichenden Bedingungen 2. Ordnung erfüllt sind und $\gamma \in (0, 1/2)$ gegeben. Dann gibt es ein $\epsilon > 0$, so dass für alle $x \in B_\epsilon(\overline{x}) \setminus \{\overline{x}\}$ gilt:*

1. *Die Lösung d von $\nabla^2 f(x)d = -\nabla f(x)$ existiert und ist eine Abstiegsrichtung von f in x.*
2. *Die Armijo-Bedingung ist für alle Schrittweiten $t \in (0, 1]$ erfüllt.*

Beweis. Aus dem vorherigen Lemma 2.5.16 erhalten wir zusammen mit Theorem 2.5.10, dass für hinreichend kleines $\epsilon > 0$ das Minimum $\overline{x}$ die einzige Nullstelle von ∇f auf $B_\epsilon(\overline{x})$ ist und

$$\lambda_{\min}(\nabla^2 f(x)) \geq \mu > 0 \quad \forall\, x \in B_\epsilon(\overline{x}).$$

Indem wir ggf. $\epsilon > 0$ weiter verkleinern finden wir $\delta \geq \epsilon > 0$, so dass simultan

$$\begin{aligned} \frac{1}{2}\|\nabla^2 f(y) - \nabla^2 f(x)\| &\leq \Big(\frac{1}{2} - \gamma\Big)\mu && \forall\, x, y \in B_{2\delta}(\overline{x}), \\ \|\nabla f(x)\| &\leq \mu\delta && \forall\, x \in B_\epsilon(\overline{x}). \end{aligned}$$

1. Entsprechend ist die Newton-Gleichung $\nabla^2 f(x)d = -\nabla f(x)$ auf $B_\epsilon(\overline{x})$ eindeutig lösbar. Für beliebiges $x \in B_\epsilon(\overline{x}) \setminus \{\overline{x}\}$ ist dann $d \neq 0$, da $\nabla f(x) \neq 0$ und d ist eine Abstiegsrichtung von f in x, da

$$\langle \nabla f(x), d\rangle = -\langle d, \nabla^2 f(x)d\rangle \leq -\mu\|d\|^2 < 0.$$

2. Es ist zu zeigen, dass $\frac{f(x+td)-f(x)}{t} - \gamma\langle \nabla f(x), d\rangle \leq 0$ gilt.

Taylorentwicklung liefert für beliebiges $t \in (0, 1]$ ein $\theta = \theta_t \in (0, t)$, so dass

$$\begin{aligned} \frac{f(x+td)-f(x)}{t} - \gamma\langle \nabla f(x), d\rangle &= (1-\gamma)\langle \nabla f(x), d\rangle + \frac{t}{2}\langle d, \nabla^2 f(x+\theta d)d\rangle \\ &= -(1-\gamma)\langle d, \nabla^2 f(x)d\rangle + \frac{t}{2}\langle d, \nabla^2 f(x+\theta d)d\rangle \\ &\leq -\Big(1-\gamma-\frac{t}{2}\Big)\langle d, \nabla^2 f(x)d\rangle + \frac{t}{2}\|\nabla^2 f(x+\theta d) - \nabla^2 f(x)\|\|d\|^2 \\ &\leq -\Big(\frac{1}{2}-\gamma\Big)\mu\|d\|^2 + \frac{1}{2}\|\nabla^2 f(x+\theta d) - \nabla^2 f(x)\|\|d\|^2. \end{aligned}$$

Nach Konstruktion ist

$$\|d\| = \|\nabla^2 f(x)^{-1}\nabla f(x)\| \le \frac{1}{\mu}\|\nabla f(x)\| \le \delta.$$

Folglich ist $x + \theta d \in B_{2\delta}(\overline{x})$ und wir erhalten aus der Wahl von ϵ und δ

$$\begin{aligned}\frac{f(x+td)-f(x)}{t} - \gamma\langle\nabla f(x), d\rangle &\le -\Big(\frac{1}{2}-\gamma\Big)\mu\|d\|^2 + \Big(\frac{1}{2}-\gamma\Big)\mu\|d\|^2 \\ &\le 0.\end{aligned}$$

Die Armijo-Bedingung ist also erfüllt. □

Wir haben nun alles beisammen, um den Übergang zu schneller Konvergenz für das globalisierte Newton-Verfahren NV [2.5.12] zu zeigen.

Theorem 2.5.18. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Die vom globalisierten Newton-Verfahren NV [2.5.12] mit $\gamma \in (0, 1/2)$ erzeugte Folge x^k habe einen Häufungspunkt $\overline{x}$, in dem die Hesse-Matrix positiv definit ist. Dann gilt:*

1. *$\overline{x}$ ist ein isoliertes Minimum.*
2. *Die gesamte Folge x^k konvergiert gegen $\overline{x}$.*
3. *Es gibt ein $l \ge 0$, so dass das Verfahren für $k \ge l$ in das Newton-Verfahren mit Schrittweite $t_k = 1$ (Lokales Newton-Verfahren NVO [2.5.7]) übergeht. Insbesondere ist Algorithmus NV [2.5.12] q-superlinear konvergent. Die Rate ist q-quadratisch, falls $\nabla^2 f$ in einer Umgebung von $\overline{x}$ Lipschitz-stetig ist.*

Beweis. **1. & 2.** Sind bereits in Lemma 2.5.16 gezeigt.

3. In Hinblick auf Lemma 2.5.17 betrachten wir $x^k \in B_\epsilon(\overline{x})$ für $\epsilon > 0$ wie in Lemma 2.5.17 gegeben. Dann ist die Newton-Gleichung

$$\nabla^2 f(x^k)\tilde{d}^k = -\nabla f(x^k)$$

lösbar, und es gilt mit den Bezeichnungen aus dem Beweis von Lemma 2.5.17

$$\|\tilde{d}_k\| \le \frac{1}{\mu}\|\nabla f(x^k)\| \to 0.$$

Folglich gibt es ein $l \ge 0$ mit den Eigenschaften

$$x^k \in B_\epsilon(\overline{x}), \quad \|\tilde{d}^k\| \le \left(\frac{\mu}{\alpha_2}\right)^{1/p} \qquad \forall\, k \ge l.$$

Damit gilt nun,

$$\begin{aligned} -\langle \nabla f(x^k), \tilde{d}^k \rangle &= \langle \tilde{d}^k, \nabla^2 f(x^k)\tilde{d}^k \rangle \\ &\geq \mu \|\tilde{d}^k\|^2 \\ &\geq \alpha_2 \|\tilde{d}^k\|^p \|\tilde{d}^k\|^2 \\ &\geq \min(\alpha_1, \alpha_2 \|\tilde{d}^k\|^p) \|\tilde{d}^k\|^2. \end{aligned}$$

Folglich akzeptiert das globalisierte Newton-Verfahren für $k \geq l$ die Suchrichtung $d^k = \tilde{d}^k$. Nach Lemma 2.5.17 wird dann die Schrittweite $t_k = 1$ gewählt.

Die Konvergenzrate des Verfahrens folgt nun aus der Aussage für das lokale Newtonverfahren 2.5.10. □

Bemerkung 2.5.19. Wir haben nun gesehen, dass das gedämpfte Newton-Verfahren global konvergiert. Allerdings kann die häufige Wahl der Gradientenrichtung i.A. nicht verhindert werden, da wir bereits gesehen haben, dass die Newton-Richtung i.A. keine Abstiegsrichtung ist. Um diesem Problem entgegen zu wirken gibt es u.A. den Ansatz von *Levenberg und Marquardt*:

Ist $\nabla^2 f(x)$ nicht positiv definit, so bestimme man ein (kleines) $\lambda > 0$, so dass die geshiftete Matrix

$$\nabla^2 f(x) + \lambda \mathrm{Id}$$

positiv definit ist. Man bestimmt dann eine Abstiegsrichtung! d als Lösung der Gleichung

$$(\nabla^2 f(x) + \lambda \mathrm{Id}) d = -\nabla f(x)$$

Im Falle $\lambda \downarrow 0$ konvergiert d gegen die Newton-Richtung. Im Fall $\lambda \to \infty$ konvergiert d gegen die Richtung des steilsten Abstiegs. Das Levenberg-Marquardt Verfahren erlaubt also die Wahl von „Zwischenrichtungen". Zur Bestimmung eines geeigneten Wertes λ und zur Entscheidung ob $\nabla^2 f(x)$ positiv definit ist bietet sich die Bestimmung der Cholesky-Zerlegung

$$\nabla^2 f(x) = LDL^T$$

mit einer unteren Dreiecksmatrix L (mit 1en auf der diagonalen) und einer diagonal Matrix D an. Es ist $\nabla^2 f(x)$ positiv definit, falls die Zerlegung existiert, und D positiv definit ist.

Dieses Verfahren ist insbesondere für *Least-Squares* Probleme

$$\min_{x \in \mathbb{R}^n} f(x) := \|F(x)\|^2$$

mit einem $F : \mathbb{R}^n \to \mathbb{R}^m$ populär. Hierbei ist, dann

$$\nabla f(x) = F'(x)^T F(x)$$
$$\nabla^2 f(x) = F'(x)^T F'(x) + \sum_{i=1}^{m} F_i(x)\nabla^2 F_i(x).$$

Da für Least-Squares Probleme die Annahme $F(\overline{x}) \approx 0$ vernünftig ist bietet sich hier die *Gauß-Newton* Approximation

$$\nabla^2 f(x) \approx F'(x)^T F'(x)$$

an.

Bemerkung 2.5.20. Des weiteren kann es auch hier wieder passieren, dass wir in einem stationären Punkt enden, welcher kein Minimum ist. Um dieses Problem zu umgehen, kann man im Fall $\nabla f(x^k) = 0$ mit nicht positiv semi-definiter Hesse-Matrix $\nabla^2 f(x^k)$ eine sog. *Richtung negativer Krümmung* d als neue Suchrichtung wählen. Dies ist eine Richtung mit $\langle d, \nabla^2 f(x)d\rangle < 0$. Für eine solche ist dann nach Taylorentwicklung mit $\nabla f(x) = 0$

$$f(x + td) = f(x) + \frac{t^2}{2}\langle d, \nabla^2 f(x)d\rangle + o(t^2)$$

und somit für $t > 0$ hinreichend klein

$$f(x + td) < f(x).$$

Auch diese Richtung d kann wieder aus einer Cholesky-Zerlegung von $\nabla^2 f(x)$ gewonnen werden.

2.6 Newton-artige Verfahren

Im Allgemeinen ist die Berechnung (und Invertierung) der Hesse-Matrix im Newton-Verfahren zu aufwändig – insbesondere wenn im globalisierten Verfahren nach der Bestimmung der Newtonrichtung dann doch der negative Gradient verwendet wird. Um dieses Problem zu umgehen verwenden wir anstelle der Newton-Gleichung

$$F'(x)d = -F(x)$$

lediglich eine Approximation M_k an die Jacobi-Matrix. Die Schritte werden dann durch Lösen der Gleichung

$$M_k d^k = -F(x^k)$$

bestimmt. Wir erhalten damit

Algorithmus NA [2.6.1] (Lokales Newton-artiges Verfahren).

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$.

for k = 0, 1, ... **do**

- Prüfe auf Abbruch: Stopp falls $F(x^k) = 0$

Wähle eine invertierbare Matrix $M_k \in \mathbb{R}^{n\times n}$.

- Berechne den Schritt $d^k \in \mathbb{R}^n$ durch Lösen der Gleichung

$$M_k d^k = -F(x^k).$$

- Setze $x^{k+1} = x^k + d^k$.

end for

Wir werden nun untersuchen, unter welchen Bedingungen an M_k wir auch von diesem Verfahren wieder lokal q-superlineare Konvergenz erwarten können. Zuvor erinnern wir uns an folgende Aussage aus der Analysis.

Lemma 2.6.2. *Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar. $X \subset \mathbb{R}^n$ sei kompakt und konvex. Dann ist F auf X Lipschitz-stetig mit der Konstante $L = \max_{x\in X} \|F'(x)\|$.*

Beweis. Da F' stetig ist, nimmt $\|F'\|$ auf X ein Maximum $L = \max_{x\in X} \|F'(x)\|$ an. Da X konvex ist, liegt für alle $x, y \in X$ auch die Verbindungsstrecke $x + t(y-x)$ für $t \in (0,1)$ in X und wir erhalten aus dem Fundamentalsatz der Differential- und Integralrechnung

$$\begin{aligned}\|F(y)-F(x)\| &= \left\| \int_0^1 F'(x+t(y-x))(y-x)\,\mathrm{d}t \right\| \\ &\le \int_0^1 \|F'(x+t(y-x))\|\,\mathrm{d}t \|y-x\| \\ &\le L\|y-x\|.\end{aligned}$$

□

Wir erhalten damit nun die folgende Aussage

Theorem 2.6.3 (Charakterisierung von q-superlinearer Konvergenz). *Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\overline{x} \in \mathbb{R}^n$ ein Punkt, so dass $F'(\overline{x})$ invertierbar ist. Sei $x^k \to \overline{x}$ mit $x^k \neq \overline{x}$ für alle k gegeben. Dann sind die folgenden Aussagen äquivalent:*

a) x^k konvergiert q-superlinear gegen $\overline{x}$ und es ist $F(\overline{x}) = 0$.

b) Für $k \to \infty$ gilt

$$\frac{\|F(x^k) + F'(\overline{x})(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} \to 0.$$

c) Für $k \to \infty$ gilt

$$\frac{\|F(x^k) + F'(x^k)(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} \to 0.$$

Beweis. Analog zum obigen Lemma erhalten wir für beliebiges $y \in \mathbb{R}^n$

$$\begin{aligned} F(x^{k+1}) &= F(x^k) + \int_0^1 F'(x^k + t(x^{k+1} - x^k))(x^{k+1} - x^k)\,\mathrm{d}t \\ &= \int_0^1 (F'(x^k + t(x^{k+1} - x^k)) - F'(y))(x^{k+1} - x^k)\,\mathrm{d}t \\ &\quad + F(x^k) + F'(y)(x^{k+1} - x^k). \end{aligned} \tag{2.12}$$

Wegen der Konvergenz der Folge x^k gibt es eine kompakte (konvexe) Kugel K in der alle Folgeglieder enthalten sind. Nach dem vorherigen Lemma 2.6.2 ist F also Lipschitz-stetig mit Konstante $L = \max_{y \in K} \|F'(y)\|$.

a) $\Rightarrow$ b) Nach a) ist für alle hinreichend großen k die Bedingung

$$\|x^{k+1} - \overline{x}\| \leq \frac{1}{2}\|x^k - \overline{x}\|$$

erfüllt. Damit folgt

$$\|x^k - \overline{x}\| \leq \|x^{k+1} - x^k\| + \|x^{k+1} - \overline{x}\| \leq \|x^{k+1} - x^k\| + \frac{1}{2}\|x^k - \overline{x}\|$$

und somit

$$\|x^k - \overline{x}\| \leq 2\|x^{k+1} - x^k\|.$$

Wir nutzen nun (2.12) mit $y = \overline{x}$ und erhalten

$$\begin{aligned}
&\frac{\|F(x^k) + F'(\overline{x})(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} \\
&\leq \frac{\|F(x^{k+1})\|}{\|x^{k+1} - x^k\|} + \int_0^1 \|F'(x^k + t(x^{k+1} - x^k)) - F'(\overline{x})\| \, \mathrm{d}t \\
&\leq \frac{\|F(x^{k+1}) - F(\overline{x})\|}{\|x^{k+1} - x^k\|} + \int_0^1 \|F'(x^k + t(x^{k+1} - x^k)) - F'(\overline{x})\| \, \mathrm{d}t \\
&\leq \frac{L\|x^{k+1} - \overline{x}\|}{\|x^{k+1} - x^k\|} + \int_0^1 \|F'(x^k + t(x^{k+1} - x^k)) - F'(\overline{x})\| \, \mathrm{d}t \\
&\leq \frac{L\|x^k - \overline{x}\|}{\|x^{k+1} - x^k\|} \frac{\|x^{k+1} - \overline{x}\|}{\|x^k - \overline{x}\|} + \int_0^1 \|F'(x^k + t(x^{k+1} - x^k)) - F'(\overline{x})\| \, \mathrm{d}t \\
&\leq 2L \frac{\|x^{k+1} - \overline{x}\|}{\|x^k - \overline{x}\|} + \int_0^1 \|F'(x^k + t(x^{k+1} - x^k)) - F'(\overline{x})\| \, \mathrm{d}t \\
&\to 0 \qquad (k \to \infty)
\end{aligned}$$

wegen der angenommenen q-superlinearen Konvergenz von x^k und der gleichmäßigen Stetigkeit von F' auf der kompakten Menge K.

b) ⇒ c) Die Implikation folgt sofort aus der Konvergenz $x^k \to \overline{x}$ und der Stetigkeit von F'. Denn es ist

$$\begin{aligned}
\frac{\|F(x^k) + F'(x^k)(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} &\leq \frac{\|F(x^k) + F'(\overline{x})(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} + \|F'(\overline{x}) - F'(x^k)\| \\
&\to 0 \qquad (k \to \infty).
\end{aligned}$$

c) ⇒ a) Wir verwenden (2.12) mit $y = x^k$ sowie die Gültigkeit von c) und erhalten

$$\begin{aligned}
\frac{\|F(x^{k+1})\|}{\|x^{k+1} - x^k\|} &\leq \int_0^1 \|F'(x^k + t(x^{k+1} - x^k)) - F'(x^k)\| \, \mathrm{d}t \\
&\quad + \frac{\|F(x^k) + F'(x^k)(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} \\
&\to 0 \qquad (k \to 0)
\end{aligned}$$

wegen der gleichmäßigen Stetigkeit von F' auf der kompakten Menge K.

Es gibt somit eine Folge $\epsilon_k > 0$ mit $\epsilon_k \to 0$, so dass

$$\|F(x^{k+1})\| \leq \epsilon_k \|x^{k+1} - x^k\|.$$

Insbesondere ist wegen der Stetigkeit von F notwendig $F(\overline{x}) = 0$. Nach Annahme ist $F'(\overline{x})$ invertierbar und nach Lemma 2.5.4 gibt es folglich ein $\gamma > 0$, so dass

$$\|F(x^{k+1})\| \geq \gamma \|x^{k+1} - \overline{x}\|$$

für alle hinreichend großen k. Damit folgt

$$\begin{aligned}\|x^{k+1}-\overline{x}\| &\leq \frac{1}{\gamma}\|F(x^{k+1})\| \\ &\leq \frac{\epsilon_k}{\gamma}\|x^{k+1}-x^k\| \\ &\leq \frac{\epsilon_k}{\gamma}\|x^{k+1}-\overline{x}\| + \frac{\epsilon_k}{\gamma}\|x^k-\overline{x}\|.\end{aligned}$$

Sei nun k hinreichend groß, damit $\frac{\epsilon_k}{\gamma} \leq \frac{1}{2}$ ist, so folgt

$$\|x^{k+1}-\overline{x}\| \leq \frac{2\epsilon_k}{\gamma}\|x^k-\overline{x}\|$$

und somit die q-superlineare Konvergenz von x^k gegen $\overline{x}$. □

Hieraus ergibt sich nun ein Kriterium wie die Wahl von M_k in Algorithmus NA [2.6.1] geschehen sollte

Korollar 2.6.4 (Dennis-Moré-Bedingung)**.** *Die unendliche Folge x^k sei durch Algorithmus NA [2.6.1] erzeugt und konvergiere gegen einen Punkt $\overline{x}$, in dem $F'(\overline{x})$ invertierbar ist. Dann sind die folgenden Aussagen äquivalent:*

a) x^k konvergiert q-superlinear gegen $\overline{x}$ und es ist $F(\overline{x}) = 0$.

b) Für $k \to \infty$ gilt

$$\frac{\|(M_k - F'(\overline{x}))(x^{k+1}-x^k)\|}{\|x^{k+1}-x^k\|} \to 0.$$

c) Für $k \to \infty$ gilt

$$\frac{\|(M_k - F'(x^k))(x^{k+1}-x^k)\|}{\|x^{k+1}-x^k\|} \to 0.$$

Beweis. Wegen

$$-F(x^k) = M_k d^k = M_k(x^{k+1}-x^k)$$

sind die Bedingungen b) und c) identisch zu denen in Theorem 2.6.3. Es ist also nichts zu zeigen. □

Die wesentliche Bedeutung dieser Beobachtung ist, dass wir zur Sicherung der q-superlinearen Konvergenz nicht F' im ganzen, sondern nur die Wirkung auf die Richtung d gut approximieren müssen.

Bemerkung 2.6.5. Erzeugt der Algorithmus NA [2.6.1] Folgen $x^k \to \overline{x}$ und $M_k \to F'(\overline{x})$, so folgt direkt

$$\frac{\|(M_k - F'(\overline{x}))(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} \leq \|M_k - F'(\overline{x})\| \to 0.$$

Ist dann $F'(\overline{x})$ invertierbar, so sind die Dennis-Moré-Bedingungen erfüllt, und die q-superlineare Konvergenz folgt.

In den folgenden beiden Abschnitten werden wir uns mit konkreten Verfahren zur Wahl von M_k befassen. Dies sind einerseits *inexakte Newton-Verfahren* und andererseits *Quasi-Newton-Verfahren*.

Natürlich bedürfen auch Newton-artige Verfahren einer Globalisierung, um die Konvergenz für beliebige Startwerte zu sichern. Wir verfahren hier analog zum globalisierten Newton-Verfahren NV [2.5.12].

Algorithmus 2.6.6 (Globales Newton-artiges Verfahren)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$, $\beta \in (0,1)$, $\gamma \in (0,1)$, $\alpha_1, \alpha_2 > 0$ und $p > 0$.

for k $= 0, 1, \ldots$ **do**

- Prüfe auf Abbruch: Stopp falls $\nabla f(x^k) = 0$

Wähle eine invertierbare symmetrische Matrix $M_k \in \mathbb{R}^{n \times n}$.

- Berechne den Schritt $d \in \mathbb{R}^n$ durch Lösen der Gleichung

$$M_k d = -\nabla f(x^k).$$

Falls d die Bedingung

$$-\langle \nabla f(x^k), d \rangle \geq \min(\alpha_1, \alpha_2 \|d\|^p) \|d\|^2$$

erfüllt, so setzte $d^k = d$. Andernfalls setze $d^k = -\nabla f(x^k)$.

- Bestimme eine Schrittweite $t_k > 0$ mithilfe der Armijo-Liniensuche ALS [2.3.6] mit Parametern β und γ.
- Setze $x^{k+1} = x^k + t_k d^k$.

end for

Die Wahl einer symmetrischen Matrix M_k wird hier getätigt, da wir uns direkt auf das Lösen eines Optimierungsproblems beziehen, und dann $M_k \approx \nabla^2 f(x^k)$ eine symmetrische Wahl nahelegt.

Wir sehen (Übung), das sich das globale Konvergenzresultat 2.5.14 für das globalisierte Newton-Verfahren NV [2.5.12] auch auf Newton-artige Verfahren 2.6.6 übertragen werden kann sofern $\|M_k\|$ gleichmäßig beschränkt ist.

2.7 Inexakte Newton-Verfahren

Als erstes Newton-artiges Verfahren werden wir uns mit sog. inexakten Newton-Verfahren, in denen die Newton-Gleichung lediglich Approximativ gelöst werden kann, befassen. Ein solches Vorgehen ist insbesondere bei sehr großen Systemen $n > 10\,000$, bei denen bereits das Speichern der Einträge von $\nabla^2 f$ Probleme bereitet, angebracht.

Solche inexakten Löser erhält man aus diversen iterativen Lösungsverfahren

- Jacobi-/Gauß-Seidel-Verfahren
- Gradientenverfahren (zur Lösung von $\nabla^2 f(x)d = -\nabla f(x)$)
- CG-Verfahren (Konjugierte Gradienten)
- ...

Wobei man das Verfahren abbricht, sobald das Residuum $\|-\nabla f(x) - \nabla^2 f(x)d\|$ hinreichend klein geworden ist. Wir erhalten damit

Algorithmus IN [2.7.1] (Lokales inexaktes Newton-Verfahren für Gleichungssysteme)**.**

• Wähle Startpunkt $x^0 \in \mathbb{R}^n$.
for k = 0, 1, ... **do**
• Prüfe auf Abbruch: Stopp falls $F(x^k) = 0$
• Wähle $\eta_k > 0$
• Berechne den Newtonschritt $d^k \in \mathbb{R}^n$ durch approximatives Lösen der Gleichung

$$F'(x^k)d^k = -F(x^k).$$

Hierbei sei durch das Residuum

$$\|F(x^k) + F'(x^k)d^k\| \leq \eta_k \|F(x^k)\| \tag{IN}$$

erfüllt.
• Setze $x^{k+1} = x^k + d^k$.
end for

Wir erhalten den folgenden Konvergenzsatz

Theorem 2.7.2. *Sei $F \in C^1(\mathbb{R}^n; \mathbb{R}^n)$ und $\overline{x}$ eine Nullstelle von F, in der F' invertierbar ist. Dann gibt es $\delta > 0$ und $\eta > 0$, so dass gilt:*

1. *Ist $x^0 \in B_\delta(\overline{x})$ und erfüllen die Schritte d^k in Algorithmus IN [2.7.1] die Bedingung (IN) mit $\eta_k \leq \eta$, so terminiert Algorithmus IN [2.7.1] entweder mit $x^k = \overline{x}$,*

oder er erzeugt eine Folge x^k, die q-linear gegen $\overline{x}$ konvergiert.

2. *Falls zusätzlich $\eta_k \to 0$, so ist die Konvergenz q-superlinear.*
3. *Ist zusätzlich $\eta_k = O(\|F(x^k)\|)$ und ist F' Lipschitz-stetig auf $B_\delta(\overline{x})$, so ist die Konvergenzrate q-quadratisch.*

Beweis. 1.) Analog zum Konvergenzbeweis für das lokale Newton-Verfahren (Theorem 2.5.6) können wir $\delta > 0$ so wählen, dass $\overline{x}$ die einzige Nullstelle in $B_\delta(\overline{x})$ ist und F' auf $B_\delta(\overline{x})$ invertierbar. Weiter sei δ so gewählt, dass

$$\sup_{x,y\in B_\delta(\overline{x})} \|F'(x)^{-1}\|\|(F'(y)-F'(x))\| \le \frac{1}{3}$$

Nun setzen wir

$$C = \sup_{x\in B_\delta(\overline{x})} \|F'(x)^{-1}\|, \qquad L = \sup_{x\in B_\delta(\overline{x})} \|F'(x)\|.$$

Dann folgt die Behauptung mit diesen Werten und $\eta = \frac{1}{CL3}$

Mit diesen Voraussetzungen erhalten wir analog zu Theorem 2.5.6 (lokale Konvergenz des Newton-Verfahrens)

$$\begin{aligned}
\|x^{k+1}-\overline{x}\| &= \|F'(x^k)^{-1}F'(x^k)(x^{k+1}-\overline{x})\| \\
&\le \|F'(x^k)^{-1}\|\|F'(x^k)(x^k-\overline{x})+F'(x^k)d^k+F(x^k)-F(x^k)+F(\overline{x})\| \\
&= \|F'(x^k)^{-1}\|\|F'(x^k)(x^k-\overline{x})-F(x^k)+F(\overline{x})\| + C\|F'(x^k)d^k+F(x^k)\| \\
&\le \|F'(x^k)^{-1}\|\left\|\left(\int_0^1 F'(x^k+s(\overline{x}-x^k))\,\mathrm{d}s - F'(x^k)\right)(\overline{x}-x^k)\right\| \\
&\quad + \eta_k C\|F(x^k)\| \\
&\le \sup_{y\in B_\delta(\overline{x})} \|F'(x^k)^{-1}\|\|F'(y)-F'(x^k)\|\|\overline{x}-x^k\| + \eta_k C\|F(x^k)-F(\overline{x})\| \\
&\le \frac{1}{3}\|\overline{x}-x^k\| + \eta_k CL\|\overline{x}-x^k\| \\
&\le \frac{2}{3}\|\overline{x}-x^k\|
\end{aligned}$$

und somit die q-lineare Konvergenz.

2.) Nach Teil 1 gibt es, da $\eta_k \le \eta$ ist, ein $\gamma \in (0,1)$, so dass

$$\|\overline{x}-x^k\| \le \|x^{k+1}-x^k\| + \|x^{k+1}-\overline{x}\| \le \|x^{k+1}-x^k\| + \gamma\|x^k-\overline{x}\|$$

und folglich

$$\|\overline{x}-x^k\| \le \frac{1}{1-\gamma}\|x^{k+1}-x^k\|.$$

Für die q-superlineare Konvergenz beobachten wir, dass nach Voraussetzung

$$\|F(x^k)+F'(x^k)(x^{x+1}-x^k)\| \le \eta_k\|F(x_k)\| \le \eta_k L\|\overline{x}-x^k\|.$$

Dies zeigt

$$\frac{\|F(x^k)+F'(x^k)(x^{x+1}-x^k)\|}{\|x^{k+1}-x^k\|} \leq \frac{\eta_k L}{1-\gamma} \to 0.$$

Aufgrund der Charakterisierung der q-superlinearen Konvergenz (Theorem 2.6.3) folgt die Behauptung.

3.) Unter den gegebenen Voraussetzungen sei $\tilde{L}$ die L-Konstante für F' und für k hinreichend groß

$$\eta_k \leq c\|F(x^k)\|.$$

Aus der Rechnung in 1. folgt

$$\begin{aligned}
\|x^{k+1}-\overline{x}\| &\leq C\left\|\left(\int_0^1 F'(x^k+s(\overline{x}-x^k))\,\mathrm{d}s - F'(x^k)\right)(\overline{x}-x^k)\right\| \\
&\quad + \eta_k C\|F(x^k)\| \\
&\leq \frac{C\tilde{L}}{2}\|x^k-\overline{x}\|^2 + cC\|F(x^k)\|^2 \\
&\leq \frac{C\tilde{L}}{2}\|x^k-\overline{x}\|^2 + cC\|F(x^k)-F(\overline{x})\|^2 \\
&\leq \left(\frac{C\tilde{L}}{2} + cCL^2\right)\|x^k-\overline{x}\|^2
\end{aligned}$$

und somit die Behauptung. □

Bemerkung 2.7.3. Wir stellen fest, dass das inexakte Newton-Verfahren in der Tat ein Newton-artiges Verfahren ist. Sei dazu

$$r^k = -F'(x^k)d^k - F(x^k).$$

Dann ist d^k die exakte Lösung des Newton-artigen Verfahrens mit

$$M_k = F'(x^k) + \frac{r^k(d^k)^T}{\|d^k\|^2}.$$

Zur Globalisierung des Verfahrens geht man genauso wie für das Newton-Verfahren vor.

2.8 Quasi-Newton-Verfahren

Nachdem wir uns nun mit den inexakten Verfahren eine erste Möglichkeit angesehen haben wie ein Newton-artiges Verfahren zur Reduktion des Lösungsaufwandes durchgeführt werden kann werden wir nun eine zweite Variante kennenlernen, bei der die Bestimmung der Hesse-Matrix $\nabla^2 f(x^k)$ entfällt, und stattdessen lediglich eine Approximation $H_k \in \mathbb{R}^{n\times n}$ für diese

verwendet wird. Damit wir nicht trotzdem die zweiten Ableitungen bestimmen müssen, werden wir H_k durch Update-Formeln aus H_{k-1}, die ohne zweite Ableitungen auskommen, bestimmen. Durch Taylorentwicklung des Gradienten erhalten wir folgendes

$$\nabla f(x^k) = \nabla f(x^{k+1}) + \nabla^2 f(x^{k+1})(x^k - x^{k+1}) + o(\|x^k - x^{k+1}\|).$$

Dies legt die folgende *Quasi-Newton-Gleichung* nahe

$$H_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k). \tag{QN}$$

Bemerkung 2.8.1. Man beachte, dass die Hesse-Matrix $\nabla^2 f(x^{k+1})$ i.A. die Quasi-Newton-Gleichung *nicht* erfüllt.

Damit sind wir nun wieder in der Lage unseren lokalen Newton-artigen Algorithmus NA [2.6.1] für diese Wahl aufzustellen.

Algorithmus 2.8.2 (Lokales Quasi-Newton Verfahren)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$ und eine symmetrische, invertierbare Matrix $H_0 \in \mathbb{R}^{n\times n}$.

for k = 0, 1, ... **do**

- Prüfe auf Abbruch: Stopp falls $\nabla f(x^k) = 0$
- Berechne den Quasi-Newton-Schritt $d^k \in \mathbb{R}^n$ durch Lösen der Gleichung

$$H_k d^k = -\nabla f(x^k).$$

- Setze $x^{k+1} = x^k + d^k$.
- Berechne mit einer Aufdatierungsformel eine symmetrische, invertierbare Matrix $H_{k+1} = H(H_k, x^{k+1} - x^k, \nabla f(x^{k+1}) - \nabla f(x^k))$, welche der Quasi-Newton-Gleichung (QN) genügt.

end for

In der Tat folgt die lokale q-superlineare Konvergenz des Verfahrens direkt aus unseren Aussagen für Newton-artige Verfahren. Es gilt nämlich folgendes:

Lemma 2.8.3. *$\overline{x} \in \mathbb{R}^n$ erfülle die hinreichenden Bedingungen zweiter Ordnung. Sei x^k eine von Algorithmus 2.8.2 erzeugte Folge mit $x^k \to \overline{x}$. Gilt dann für die erzeugten Matrizen*

$$\lim_{k\to\infty} \|H_{k+1} - H_k\| = 0,$$

dann erfüllen die Matrizen H_k die Dennis-Moré-Bedingung 2.6.4. Folglich konvergiert x^k q-superlinear gegen $\overline{x}$.

Beweis. Aus Taylorentwicklung ergibt sich wegen der Quasi-Newton-Bedingung und $d^k := x^{k+1} - x^k$

$$\begin{aligned}\frac{\|(H_k - \nabla^2 f(x^k))d^k\|}{\|d^k\|} &\leq \frac{\|(H_k - H_{k+1})d^k\| + \|(H_{k+1} - \nabla^2 f(x^k))d^k\|}{\|d^k\|} \\ &\leq \|H_k - H_{k+1}\| + \frac{\|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)d^k\|}{\|d^k\|} \\ &\to 0 \qquad (k \to \infty).\end{aligned}$$

□

2.8.1 Quasi-Newton-Aufdatierungsformeln

Wir haben also die Aufgabe Aufdatierungsformeln zu finden, für die H_{k+1} möglichst nahe an H_k liegt und welche die Quasi-Newton-Gleichung (QN) erfüllt.

2.8.1.1 Broyden-Approximation

Das obige Lemma legt es nahe, die Matrix H_{k+1} so zu wählen, dass $H_{k+1}v = H_k v$ für alle $v \perp d^k = x^{k+1} - x^k$. Entsprechend ist $H_{k+1} - H_k$ eine Matrix vom Rang Eins. Das Update hat also notwendig die Form

$$H_{k+1} = H_k + \frac{u(d^k)^T}{\|d^k\|^2}$$

mit einem noch zu bestimmenden Vektor $u \in \mathbb{R}^n$. Aus der Quasi-Newton-Gleichung (QN) ergibt sich

$$H_{k+1}d^k = H_k d^k + \frac{u(d^k)^T}{\|d^k\|^2}d^k = H_k d^k + u = \nabla f(x^{k+1}) - \nabla f(x^k) = y^k$$

und somit

$$u = \nabla f(x^{k+1}) - \nabla f(x^k) - H_k d^k = y^k - H_k d^k.$$

In der Tat ist diese Wahl optimal in Bezug auf die Größe

$$\|H_{k+1} - H_k\|,$$

es gilt nämlich

Theorem 2.8.4. *Die Broyden-Approximation*

$$H_k + \frac{(y^k - H_k d^k)(d^k)^T}{\|d^k\|^2}$$

ist eine Lösung des Problems

$$\min_{H \in \mathbb{R}^{n\times n}} \|H - H_k\|$$

u.d.N. H genügt der Quasi-Newton-Bedingung (QN).

Das wesentlich Problem dieser Wahl ist, dass die Broyden-Approximation weder Symmetrie noch Definitheits erhaltend ist. Für die Anwendung auf Optimierungsprobleme ist diese also nicht geeignet.

2.8.1.2 Symmetrie erhaltende Update-Formeln

Rang-1 Updates Wir wollen nun Symmetrie erhaltende Update-Formeln suchen. Der einfachste Fall ist ein *symmetrische Rang-1-Update* **SR1**.

Wir definieren dazu die folgenden Abkürzungen:

$$y^k = \nabla f(x^{k+1}) - \nabla f(x^k), \qquad d^k = x^{k+1} - x^k.$$

Bemerkung 2.8.5. Man beachte, dass in unserem obigen Algorithmus $t_k = 1$ ist und daher ist in der Tat $d^k = x^{k+1} - x^k$. Im Falle globalisierter Verfahren ist die nachfolgende Argumentation jedoch für den Schritt $x^{k+1} - x^k$ zu betrachten.

Um die Symmetrie von H_k zu erhalten muss unser Rang-1 Update die Form

$$H_{k+1} = H_k + \gamma_k u^k (u^k)^T$$

mit geeigneten $\gamma_k \in \mathbb{R}$ und $u^k \in \mathbb{R}^n$ mit $\|u^k\| = 1$ haben. Analog zur Broyden-Approximation gibt uns die Quasi-Newton-Bedingung (QN)

$$H_{k+1} d^k = H_k d^k + \gamma_k \langle u^k, d^k \rangle u^k = y^k.$$

Wäre nun $y^k - H_k d^k = 0$, so gilt (mit Schrittweite $t_k = 1$)

$$\nabla f(x^{k+1}) = \nabla f(x^k) + y^k = \nabla f(x^k) + H_k d^k = 0$$

und entsprechend ist nichts mehr aufzudatieren. Ist jedoch $t_k \neq 1$, so kann es passieren, dass $y^k - H_k d^k = 0$, ohne dass das Verfahren beendet ist.

Wir nehmen nun an $y^k - H_k d^k \neq 0$. So erhalten wir bei willkürlicher Festlegung des Vorzeichens

$$u^k = \frac{y^k - H_k d^k}{\|y^k - H_k d^k\|}.$$

Nimmt man nun weiter an, es wäre $\langle y^k - H_k d^k, d^k\rangle \neq 0$, so folgt aus der Quasi-Newton Gleichung und obigem u^k

$$\gamma_k \langle u^k, d^k\rangle = \|y^k - H_k d^k\|.$$

Dies impliziert

$$\gamma_k = \frac{\|y^k - H_k d^k\|^2}{\langle y^k - H_k d^k, d^k\rangle}$$

und wir erhalten somit die Aufdatierungsformel

$$H_{k+1} = H_k + \frac{(y^k - H_k d^k)(y^k - H_k d^k)^T}{\langle y^k - H_k d^k, d^k\rangle}.$$

Bemerkung 2.8.6. Wir wollen die in der Herleitung festgestellten Probleme noch einmal zusammenfassen. Zunächst einmal ist nicht gesichert, dass wir das Update überhaupt ausführen können, da es passieren kann, dass $\langle y^k - H_k d^k, d^k\rangle = 0$. Weiterhin ist H_{k+1} im Falle der Existenz zwar symmetrisch. Doch falls $\langle y^k - H_k d^k, d^k\rangle < 0$ ist, ist selbst für positiv definites H_k, die aufdatierte Matrix H_{k+1} nicht notwendig positiv definit (und ggf. nicht einmal invertierbar.)

Rang-2 Updates Da unsere Herleitung gezeigt hat, dass jedes Rang-1 Update welches die Quasi-Newton-Bedingung (QN) erfüllt Probleme mit sich bringt, müssen wir etwas mehr Aufwand in unser Aufdatierungsformel stecken. Dies führt auf die Betrachtung der (erfolgreichen) Rang-2-Updates

$$H_{k+1} = H_k + \gamma_{k1} u_1^k (u_1^k)^T + \gamma_{k2} u_2^k (u_2^k)^T$$

mit $\gamma_{k1}, \gamma_{k2} \in \mathbb{R}$ und $u_1^k, u_2^k \in \mathbb{R}^n$.

Wiedereinmal können wir die Quasi-Newton-Bedingung verwenden, und erhalten

$$H_k d^k + \gamma_{k1} \langle u_1^k, d^k\rangle u_1^k + \gamma_{k2} \langle u_2^k, d^k\rangle u_2^k = y^k.$$

Wir sehen damit, dass $y^k - H_k d^k$ eine Linearkombination von u_1^k und u_2^k ist. Indem wir die Wahl

$$u_1^k = y^k, \qquad u_2^k = H_k d^k$$

treffen erhalten wir (sofern die Terme wohldefiniert sind)

$$\gamma_{k1} = \frac{1}{\langle u_1^k, d^k\rangle} = \frac{1}{\langle y^k, d^k\rangle}, \qquad \gamma_{k2} = \frac{-1}{\langle u_2^k, d^k\rangle} = \frac{-1}{\langle H_k d^k, d^k\rangle}.$$

Dieses ist die sogenannte **BFGS**-Update-Formel (*Broyden, Fletcher, Goldfarb und Shanno*). Eine weitere wichtige Rang-2 Update-Formel erhalten wir durch analoge Rechnung für ein Rang zwei Update der Inversen $B_k = H_k^{-1}$. Dies ist die sog. **DFP**-Update-Formel (*Davidon, Fletcher, Powell*). Mit diesen ergeben sich die folgenden Klassen von Aufdatierungsformeln

- Die BFGS-Formel
$$H_{k+1}^{BFGS} = H_k + \frac{y^k (y^k)^T}{\langle y^k, d^k \rangle} - \frac{H_k d^k (H_k d^k)^T}{\langle d^k, H_k d^k \rangle}.$$

- Die DFP-Formel
$$H_{k+1}^{DFP} = H_k + \frac{(y^k - H_k d^k)(y^k)^T + y^k (y^k - H_k d^k)^T}{\langle y^k, d^k \rangle} - \frac{\langle y^k - H_k d^k, d^k \rangle}{\langle y^k, d^k \rangle^2} y^k (y^k)^T.$$

- Die Broyden-Klasse
$$H_{k+1}^{\lambda} = (1-\lambda) H_{k+1}^{BFGS} + \lambda H_{k+1}^{DFP}$$
hierbei erhalten wir $H_{k+1}^0 = H_{k+1}^{BFGS}$ und $H_{k+1}^1 = H_{k+1}^{DFP}$. Für
$$\lambda = \frac{\langle y^k, d^k \rangle}{\langle y^k, d^k \rangle - \langle d^k, H_k d^k \rangle}$$
ergibt sich das SR1-Update.

- Die konvexe Broyden-Klasse: H_{k+1}^{λ} mit $\lambda \in [0,1]$.

Wir werden nun sehen, dass die Broyden-Klasse unter gewissen Bedingungen in der Tat symmetrische und positiv definite Matrizen erzeugt.

Theorem 2.8.7 (Quasi-Newton-Matrizen)**.**

1. *Ist $\langle y^k, d^k \rangle \neq 0$ und $\langle d^k, H_k d^k \rangle \neq 0$, so sind die Matrizen H_{k+1}^{λ} für $\lambda \in \mathbb{R}$ wohldefiniert, symmetrisch und erfüllen die Quasi-Newton-Gleichung* (QN).
2. *Ist H_k positiv definit, und gilt $\langle y^k, d^k \rangle > 0$, so sind auch die Matrizen H_{k+1}^{λ} für $\lambda \geq 0$ positiv definit.*

Beweis. 1. Nach Definition von H_{k+1}^{BFGS} und H_{k+1}^{DFP} sind Wohldefiniertheit und Symmetrie unter den gegebenen Bedingungen klar.

Aus der Herleitung wissen wir bereits, dass die BFGS-Update-Formel die Quasi-Newton-Gleichung erfüllt, d.h.
$$H_{k+1}^{BFGS} d^k = y^k.$$

Für die Broyden-Klasse beobachten wir nun, dass für einen beliebigen Vektor $v \in \mathbb{R}^n$ gilt

$$\begin{aligned}
H_{k+1}^{\lambda} v &= (1-\lambda) H_{k+1}^{BFGS} v + \lambda H_{k+1}^{DFP} v \\
&= (1-\lambda)\Big(H_k v + \frac{\langle y^k, v\rangle}{\langle y^k, d^k\rangle} y^k - \frac{\langle H_k d^k, v\rangle}{\langle d^k, H_k d^k\rangle} H_k d^k\Big) \\
&\quad + \lambda\Bigg(H_k v + \Big(\frac{\langle y^k, v\rangle}{\langle y^k, d^k\rangle}(y^k - H_k d^k) + \frac{\langle y^k - H_k d^k, v\rangle}{\langle y^k, d^k\rangle} y^k\Big) \\
&\qquad - \Big(\frac{\langle y^k - H_k d^k, d^k\rangle \langle y^k, v\rangle}{\langle y^k, d^k\rangle^2} y^k\Big)\Bigg) \\
&= H_k v + \frac{\langle y^k, v\rangle}{\langle y^k, d^k\rangle} y^k - \frac{\langle H_k d^k, v\rangle}{\langle d^k, H_k d^k\rangle} H_k d^k \\
&\quad + \lambda\Bigg[\frac{\langle H_k d^k, v\rangle}{\langle d^k, H_k d^k\rangle} H_k d^k - \frac{\langle y^k, v\rangle}{\langle y^k, d^k\rangle} H_k d^k + \frac{\langle y^k - H_k d^k, v\rangle}{\langle y^k, d^k\rangle} y^k \\
&\qquad - \frac{\langle y^k - H_k d^k, d^k\rangle \langle y^k, v\rangle}{\langle y^k, d^k\rangle^2} y^k\Bigg] \\
&= H_{k+1}^{BFGS} v + \lambda\Bigg[\frac{\langle H_k d^k, v\rangle}{\langle d^k, H_k d^k\rangle} H_k d^k - \frac{\langle y^k, v\rangle}{\langle y^k, d^k\rangle} H_k d^k \\
&\qquad\qquad - \frac{\langle H_k d^k, v\rangle}{\langle y^k, d^k\rangle} y^k + \frac{\langle H_k d^k, d^k\rangle \langle y^k, v\rangle}{\langle y^k, d^k\rangle^2} y^k\Bigg] \\
&=: H_{k+1}^{BFGS} v + \lambda R v
\end{aligned}$$

Im Falle $v = d^k$ verschwindet die letzte Klammer hinter dem Faktor λ und somit folgt die Behauptung.

2. Sei nun $v \in \mathbb{R}^n \setminus \{0\}$ beliebig gegeben. Wir möchten nun $\langle v, H_{k+1}^{\lambda} v\rangle$ nach unten abschätzen. Hierzu verwenden wir die in 1. getätigte Rechnung und betrachten zunächst die letzte Klammer wenn wir diese mit v skalar multiplizieren. Wir erhalten

$$\begin{aligned}
\langle Rv, v\rangle &= \frac{\langle H_k d^k, v\rangle^2}{\langle d^k, H_k d^k\rangle} - \frac{\langle y^k, v\rangle \langle H_k d^k, v\rangle}{\langle y^k, d^k\rangle} - \frac{\langle H_k d^k, v\rangle \langle y^k, v\rangle}{\langle y^k, d^k\rangle} + \frac{\langle H_k d^k, d^k\rangle \langle y^k, v\rangle^2}{\langle y^k, d^k\rangle^2} \\
&= \frac{\langle H_k d^k, v\rangle^2}{\langle d^k, H_k d^k\rangle} - 2\frac{\langle y^k, v\rangle \langle H_k d^k, v\rangle}{\langle y^k, d^k\rangle} + \frac{\langle H_k d^k, d^k\rangle \langle y^k, v\rangle^2}{\langle y^k, d^k\rangle^2} \\
&= \langle H_k d^k, d^k\rangle \left(\frac{\langle H_k d^k, v\rangle}{\langle d^k, H_k d^k\rangle} - \frac{\langle y^k, v\rangle}{\langle y^k, d^k\rangle}\right)^2 \\
&\geq 0.
\end{aligned}$$

Wegen $\lambda \geq 0$ ist also

$$\langle v, H_{k+1}^{\lambda} v\rangle \geq \langle v, H_{k+1}^{BFGS} v\rangle.$$

Um dies abzuschätzen, berechnen wir mit der Cholesky-Zerlegung $H_k = LL^T$

$$\begin{aligned}
\langle v, H_{k+1}^{BFGS} v\rangle &= \langle v, H_k v\rangle + \frac{\langle y^k, v\rangle^2}{\langle y^k, d^k\rangle} - \frac{\langle H_k d^k, v\rangle^2}{\langle d^k, H_k d^k\rangle} \\
&= \|L^T v\|^2 + \frac{\langle y^k, v\rangle^2}{\langle y^k, d^k\rangle} - \frac{\langle L^T d^k, L^T v\rangle^2}{\|L^T d^k\|^2} \\
&\geq \|L^T v\|^2 + \frac{\langle y^k, v\rangle^2}{\langle y^k, d^k\rangle} - \frac{\|L^T d^k\|^2 \|L^T v\|^2}{\|L^T d^k\|^2} \\
&= \|L^T v\|^2 + \frac{\langle y^k, v\rangle^2}{\langle y^k, d^k\rangle} - \|L^T v\|^2 \\
&= \frac{\langle y^k, v\rangle^2}{\langle y^k, d^k\rangle}.
\end{aligned}$$

Man beachte nun, dass wir in dieser Rechnung eine strikte Ungleichung haben, falls $L^T d^k$ und $L^T v$ linear unabhängig sind, d.h. in diesem Fall ist

$$\langle v, H_{k+1}^{BFGS} v\rangle > 0.$$

Ansonsten gibt es ein $t \in \mathbb{R} \setminus \{0\}$ (da $v \neq 0$), so dass

$$L^T v = t L^T d^k$$

und somit

$$\frac{\langle y^k, v\rangle^2}{\langle y^k, d^k\rangle} = \frac{\langle y^k, L^{-T} L^T v\rangle^2}{\langle y^k, d^k\rangle} = t^2 \frac{\langle y^k, L^{-T} L^T d^k\rangle^2}{\langle y^k, d^k\rangle} = t^2 \langle y^k, d^k\rangle > 0.$$

Somit ist H_{k+1}^{λ} positiv definit.

□

2.8.1.3 Effizienz der Update-Formeln

Nachdem wir nun gesehen haben, dass wir in der Tat eine Update-Formel für die Hesse-Matrix-Approximationen H_k finden können, die lediglich die Evaluation von H_k, y^k und d^k benötigt stellt sich die Frage, wie wir die so erzeugten Matrizen invertieren können. Das dies in der Tat einfach geht wird durch folgendes Lemma garantiert, welches sichert, dass sich die Inverse eines niedrig-Rang Updates auch durch ein niedrig-Rang Update der Inversen darstellen lässt (sofern die Matrizen invertierbar sind):

Lemma 2.8.8 (Sherman-Morrison-Woodbury Formel)**.** *Sei $A \in \mathbb{R}^{n\times n}$ invertierbar und $u, v \in \mathbb{R}^n$. Dann ist $A + uv^T$ genau dann invertierbar, wenn $1 + v^T A^{-1} u \neq 0$. In diesem Fall*

ist dann

$$(A+uv^T)^{-1} = \left(I - \frac{A^{-1}uv^T}{1+v^TA^{-1}u}\right)A^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}.$$

Beweis. Offenbar ist

$$\begin{aligned}(A+uv^T)\Big[(1+v^TA^{-1}u)A^{-1} - A^{-1}uv^TA^{-1}\Big] &= (1+v^TA^{-1}u)I + (1+v^TA^{-1}u)uv^TA^{-1} \\ &\quad - uv^TA^{-1} - uv^TA^{-1}uv^TA^{-1} \\ &= (1+v^TA^{-1}u)I \\ &\quad + (v^TA^{-1}u)uv^TA^{-1} - u(v^TA^{-1}u)v^TA^{-1} \\ &= (1+v^TA^{-1}u)I.\end{aligned}$$

Im Falle $(1+v^TA^{-1}u) \neq 0$ ergibt sich also gerade die angegebene Formel für die Inverse.

Ist $(1+v^TA^{-1}u) = 0$, erhalten wir

$$\begin{aligned}(A+uv^T)A^{-1}u &= (I+uv^TA^{-1})u \\ &= u + uv^TA^{-1}u \\ &= (1+v^TA^{-1}u)u \\ &= 0.\end{aligned}$$

Die Matrix $(A+uv^T)A^{-1}$ besitzt also einen mindestens eindimensionalen Kern, da $u \neq 0$ (sonst wäre $(1+v^TA^{-1}u) \neq 0$). Da A^{-1} invertierbar ist, folgt die nicht Invertierbarkeit von $(A+uv^T)$. □

Per Induktion kann dies nun auf beliebige Updates vom Rang-n erweitert werden. Für unsere BFGS und DFP Matrizen erhalten wir folgende Darstellung der Inversen

Theorem 2.8.9. *Sei H_k positiv definit und $B_k = H_k^{-1}$. Ist dann $\langle y^k, d^k\rangle > 0$, so gelten für $B_{k+1}^{BFGS} = (H_{k+1}^{BFGS})^{-1}$ und $B_{k+1}^{DFP} = (H_{k+1}^{DFP})^{-1}$ die Aufdatierungsformeln*

$$B_{k+1}^{BFGS} = B_k + \frac{(d^k - B_ky^k)(d^k)^T + d^k(d^k - B_ky^k)^T}{\langle d^k, y^k\rangle} - \frac{\langle d^k - B_ky^k, y^k\rangle}{\langle d^k, y^k\rangle^2} d^k(d^k)^T,$$

$$B_{k+1}^{DFP} = B_k + \frac{d^k(d^k)^T}{\langle d^k, y^k\rangle} - \frac{B_ky^k(B_ky^k)^T}{\langle y^k, B_ky^k\rangle}.$$

Beweis. Es ist leicht einzusehen, dass die angegebenen Formeln in der Tat die Inversen sind. □

Bemerkung 2.8.10. Entsprechend unserer Bemerkung zur DFP-Update-Formel sehen wir hier nun die behauptete Entsprechung der inversen und direkten Update-Formeln. Es ergibt sich hierbei die inverse DFP-Formel aus der direkten BFGS-Formel und umgekehrt die inverse BFGS-Formel aus der direkten DFP-Formel durch Ersetzen von (H_k, d^k, y^k) durch (B_k, y^k, d^k). In der Tat ist die inverse Quasi-Newton Gleichung

$$d^k = By^k \qquad \text{anstatt} \qquad Hd^k = y^k$$

In der Tat sind die BFGS und DFP Update-Formeln geeignet im Sinne der Approximation zwischen H_{k+1} und H_k. Es gilt das folgende Theorem

Theorem 2.8.11. *Sei H_k symmetrisch und positiv definit und gelte $\langle y^k, d^k\rangle > 0$. Dann gibt es mindestens eine positiv definite Matrix W mit $W^2 d^k = y^k$. Für jede solche Matrix gilt dann*

- *H_{k+1}^{DFP} löst das Problem*

$$\min_{H\in\mathbb{R}^{n\times n}} \|W^{-1}(H-H_k)W^{-1}\|_F$$
$$\text{u.d.N. } H = H^T, \quad Hd^k = y^k.$$

- *H_{k+1}^{BFGS} löst das Problem*

$$\min_{H\in\mathbb{R}^{n\times n}} \|W(H^{-1}-H_k^{-1})W\|_F$$
$$\text{u.d.N. } H = H^T, \quad Hd^k = y^k.$$

Dabei ist $\|\cdot\|_F$ die Frobenius-Norm, d.h. $\|H\|_F^2 = \sum_{i,j=1}^n H_{ij}^2$.

Insbesondere sehen wir, dass die DFP und BFGS Update-Formel invariant unter affinen Transformationen sind, diese Eigenschaft des Newton-Verfahrens überträgt sich also auch auf die Broyden-Klasse.

Um die Invarianz zu sehen sei $f \in C^1$, $\tilde{x} = Ax + b$ bzw. $x = A^{-1}(\tilde{x} - b)$. Ist dann

$$g(\tilde{x}) := f(A^{-1}(\tilde{x} - b))$$

und somit

$$\nabla g(\tilde{x}) = A^{-T}\nabla f(x).$$

Wir setzen nun

$$H^{-1} = A^{-1}\widetilde{H}^{-1}A^{-T} \quad \text{bzw.} \quad \widetilde{H} = A^{-T}HA^{-1}.$$

Dann ist

$$\begin{aligned} d &= -H^{-1}\nabla f(x), \\ \tilde{d} &= -\widetilde{H}^{-1}\nabla g(\tilde{x}) \\ &= -\widetilde{H}^{-1}A^{-T}\nabla f(x) \\ &= -AH^{-1}\nabla f(x) = Ad \end{aligned}$$

und es folgt die Invarianz der Iterierten x^+ bzw. $\tilde{x}^+$ aus

$$\begin{aligned} A^{-1}(\tilde{x}^+ - b) &= A^{-1}(\tilde{x} - b + \tilde{d}) \\ &= x + d \\ \tilde{y} &= \nabla g(\tilde{x}^+) - \nabla g(\tilde{x}) \\ &= A^{-T}\Big(\nabla f(x^+) - \nabla f(x)\Big) \\ &= A^{-T}y. \end{aligned}$$

Dies impliziert für die Quasi-Newton Gleichungen

$$\widetilde{H}\tilde{d} = \tilde{y} \quad \Leftrightarrow \quad A^{-T}HA^{-1}Ad = A^{-T}y \quad \Leftrightarrow \quad Hd = y.$$

In der Tat gilt somit für das BFGS-Update

$$H^+ = H + \frac{yy^T}{\langle y, d\rangle} - \frac{Hd(Hd)^T}{\langle Hd, d\rangle}$$

die Beziehung

$$\widetilde{H}^+ = \widetilde{H} + \frac{\tilde{y}\tilde{y}^T}{\langle \tilde{y}, \tilde{d}\rangle} - \frac{\widetilde{H}\tilde{d}(\widetilde{H}\tilde{d})^T}{\langle \widetilde{H}\tilde{d}, \tilde{d}\rangle} = A^{-T}H^+A^{-1}.$$

Dies zeigt die Invarianz des BFGS-Updates unter affinen Transformationen.

2.8.2 Ein lokales BFGS-Verfahren

Wir werden nun ein lokales BFGS-Verfahren angeben, bei dem wir die inversen Update-Formeln verwenden.

Algorithmus 2.8.12 (Lokales inverses BFGS-Verfahren)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$, sowie eine symmetrisch positiv definite Matrix $B_0 \in \mathbb{R}^{n\times n}$ (Möglichst $B_0 \approx \nabla^2 f(\overline{x})^{-1}$).
- Stopp, falls $\nabla f(x^0) = 0$.

for k = 0, 1, ... **do**

- Bestimme $d^k = -B_k\nabla f(x^k)$.
- Setze $x^{k+1} = x^k + d^k$.
- Prüfe auf Abbruch: Stopp, falls $\nabla f(x^{k+1}) = 0$
- Setze $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$.
- Stopp, falls $\langle y^k, d^k\rangle \leq 0$ (Fehler!!!)

- Bestimme

$$B_{k+1} = B_{k+1}^{BFGS} = B_k + \frac{(d^k - B_k y^k)(d^k)^T + d^k(d^k - B_k y^k)^T}{\langle d^k, y^k \rangle} - \frac{\langle d^k - B_k y^k, y^k \rangle}{\langle d^k, y^k \rangle^2} d^k (d^k)^T$$

end for

Wir haben nun das folgende repräsentative Konvergenzresultat.

Theorem 2.8.13. *Sei $f \in C^2(\mathbb{R}^n;\mathbb{R})$ mit lokal L-stetiger Hesse-Matrix $\nabla^2 f$. Weiter seien in $\overline{x}$ die hinreichenden Bedingungen zweiter Ordnung erfüllt. Dann gibt es $\delta > 0$, $\epsilon > 0$, so dass Algorithmus 2.8.12 für jeden Startwert $x^0 \in B_\delta(\overline{x})$ und jede symmetrisch positiv definite Startmatrix $B_0 \in \mathbb{R}^{n\times n}$ mit $\|B_0 - \nabla^2 f(\overline{x})^{-1}\| \leq \epsilon$ entweder mit $x^k = \overline{x}$ abbricht, oder eine unendliche Folge $x^k \in B_\delta(\overline{x})$ erzeugt, die q-superlinear gegen $\overline{x}$ konvergiert.*

Beweis. Wir verzichten im Rahmen der Vorlesung auf den umfangreichen Beweis. Die Teilnehmer sind im Selbststudium angehalten sich diesen anzusehen. Einen Beweis findet man u.A. in [Geiger and Kanzow, 1999, Satz 11.33]. □

2.8.3 Globalisierte Quasi-Newton Verfahren

Wie wir im lokalen inversen-BFGS-Verfahren gesehen haben, kann es passieren, dass die Update-Formel nicht durchführbar ist wenn $\langle y^k, d^k \rangle \leq 0$. Wir müssen dies in einer globalisierten Variante des Verfahrens ausschließen, damit wir nicht immer wieder eine neue Startmatrix B_0 wählen müssen.

Algorithmus 2.8.14 (Globalisiertes inverses BFGS-Verfahren)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$, sowie eine symmetrisch positiv definite Matrix $B_0 \in \mathbb{R}^{n\times n}$, sowie Parameter $\gamma \in (0, 1/2)$ und $\eta \in (\gamma, 1)$.
- Stopp, falls $\nabla f(x^0) = 0$.

for k = 0, 1, ... **do**

- Bestimme $d^k = -B_k \nabla f(x^k)$.
- Bestimme $t_k > 0$ durch die Powell-Wolfe-Regel 2.4.9
- Setze $x^{k+1} = x^k + t_k d^k$.
- Prüfe auf Abbruch: Stopp, falls $\nabla f(x^{k+1}) = 0$
- Setze $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ und $\tilde{d}^k = t_k d^k$.
- Bestimme

$$B_{k+1} = B_{k+1}^{BFGS} = B_k + \frac{(\tilde{d}^k - B_k y^k)(\tilde{d}^k)^T + \tilde{d}^k(\tilde{d}^k - B_k y^k)^T}{\langle \tilde{d}^k, y^k \rangle} - \frac{\langle \tilde{d}^k - B_k y^k, y^k \rangle}{\langle \tilde{d}^k, y^k \rangle^2} \tilde{d}^k (\tilde{d}^k)^T$$

end for

Zunächst sehen wir, dass die Powell-Wolfe-Schrittweite in der Tat die Bedingung

$$\langle y^k, \tilde{d}^k \rangle \geq 0$$

sicherstellt, daher ist das inverse BFGS-Update in jedem Schritt durchführbar.

Lemma 2.8.15. *Sei $f \in C^1(\mathbb{R}^n;\mathbb{R})$. Ist B_k positiv definit, und ist die Bestimmung einer Schrittweite t_k in der Powell-Wolfe-Liniensuche im globalisierten inversen BFGS-Verfahren 2.8.14 erfolgreich, so ist*

$$\langle y^k, \tilde{d}^k \rangle > 0$$

und B_{k+1} ist positiv definit.

Beweis. Nach Annahme ist B_k und damit auch $H_k = B_k^{-1}$ positiv definit. Ist dann $\langle y^k, \tilde{d}^k \rangle > 0$, so folgt nach dem Theorem zu den Quasi-Newton-Matrizen 2.8.7 die positiv Definitheit von H_{k+1} und damit auch $B_{k+1} = H_{k+1}^{-1}$. Es bleibt also nur noch das Vorzeichen von $\langle y^k, \tilde{d}^k \rangle$ zu zeigen.

Aufgrund der Powell-Wolfe-Bedingung ($\mathscr{PW}$) ist

$$\begin{aligned}
\langle y^k, \tilde{d}^k \rangle &= t_k \langle \nabla f(x^{k+1}) - \nabla f(x^k), d^k \rangle \\
&\geq t_k \eta \langle \nabla f(x^k), d^k \rangle - t_k \langle \nabla f(x^k), d^k \rangle \\
&= -t_k (1-\eta) \langle \nabla f(x^k), d^k \rangle \\
&= t_k (1-\eta) \langle \nabla f(x^k), B_k \nabla f(x^k) \rangle \\
&> 0
\end{aligned}$$

da B_k positiv definit ist. □

Wir haben nun den folgenden globalen Konvergenzsatz

Theorem 2.8.16. *Sei $f \in C^1(\mathbb{R}^n;\mathbb{R})$ und $x^0 \in \mathbb{R}^n$ so, dass die Niveaumenge $N_f(x^0)$ kompakt ist. Dann ist das globalisierte inverse BFGS-Verfahren 2.8.14 durchführbar. Ist zudem die Kondition $\mathrm{cond}(B_k) = \frac{\lambda_{\max}(B_k)}{\lambda_{\min}(B_k)} \leq C < \infty$, so ist jeder Häufungspunkt von x^k ein stationärer Punkt von f.*

Beweis. Ist $\nabla f(x^k) \neq 0$ und B_k positiv definit, so ist

$$\langle \nabla f(x^k), d^k \rangle = -\langle \nabla f(x^k), B_k \nabla f(x^k) \rangle < 0$$

und nach Theorem 2.4.10 zur Powell-Wolfe Schrittweitenbestimmung erhalten wir in der Liniensuche eine Schrittweite $t_k > 0$, welche der Powell-Wolfe-Bedingung 2.4.4 genügt. Aufgrund des vorherigen Lemmas 2.8.15 ist damit auch B_{k+1} positiv definit. Die Durchführbarkeit des Verfahrens folgt somit per Induktion.

Ist nun zudem $\operatorname{cond}(B_k) \leq C < \infty$, so erhalten wir

$$\begin{aligned}
-\langle \nabla f(x^k), d^k \rangle &= \langle d^k, B_k^{-1} d^k \rangle \\
&\geq \frac{1}{\lambda_{\max}(B_k)} \|d^k\|^2 \\
&= \frac{1}{\lambda_{\max}(B_k)} \|B_k \nabla f(x^k)\| \|d^k\| \\
&\geq \frac{\lambda_{\min}(B_k)}{\lambda_{\max}(B_k)} \|\nabla f(x^k)\| \|d^k\| \\
&= \operatorname{cond}(B_k)^{-1} \|\nabla f(x^k)\| \|d^k\| \\
&\geq C^{-1} \|\nabla f(x^k)\| \|d^k\|.
\end{aligned}$$

Dies ist gerade die Winkelbedingung

$$\cos \angle(-\nabla f(x^k), d^k) := \frac{-\langle \nabla f(x^k), d^k \rangle}{\|\nabla f(x^k)\| \|d^k\|} \geq C^{-1}$$

somit ist die Richtung d^k zulässig (vgl. hierzu auch den Kommentar nach Definition 2.2.5). □

Bemerkung 2.8.17.

1. Im Allgemeinen ist es nicht möglich zu zeigen, dass die Kondition von B_k beschränkt bleibt. In praktischen Implementierungen ist es daher notwendig die Kondition von B_k zu überprüfen (oder einen Winkeltest durchzuführen) und ggf. die Matrix B_k zu verwerfen und durch eine geeignete Matrix (z.B. B_0) neu zu initialisieren und anschließend die Suchrichtung d^k erneut zu bestimmen.
2. In der Praxis ist es nicht selten der Fall, dass die Folge der erzeugten Approximationen $H_k = B_k^{-1} \to \nabla^2 f(\overline{x})$ für eine Lösung $\overline{x}$ des Optimierungsproblems. In diesem Fall folgt die q-superlineare Konvergenz im Falle $t_k = 1$ aus den Dennis-Moré-Bedingungen 2.6.4. Ist dann $\nabla^2 f(\overline{x})$ positiv definit, so lässt sich zeigen, dass für x^k in einer Umgebung von $\overline{x}$ und B_k in einer Umgebung von $\nabla^2 f(\overline{x})^{-1}$ auch tatsächlich $t_k = 1$ als Schrittweite akzeptiert wird.
3. Ist $f \in C^2(\mathbb{R}^n; \mathbb{R})$ gleichmäßig konvex auf $N_f(x_0)$, so gilt sogar die stärkere Konvergenzaussage: Die vom globalisierten inversen BFGS-Verfahren erzeugt Folge x^k konvergiert für beliebige symmetrische und positiv definite B_0 gegen das (eindeutig bestimmte) Minimum $\overline{x}$ von f.
4. Um das Abspeichern der vollen Inversen zu vermeiden ist es üblich lediglich B_0 sowie die zur Anwendung von B_k notwendigen Vektoren $(\tilde{d}^k, \tilde{d}^k - B_k y^k)$ zu

speichern. Für $k \geq n/2$ ist der Speicheraufwand dann jedoch bereits größer als der von B_k. Um diesen Aufwand zu reduzieren ist es möglich den verwendeten Speicher zu limitieren, sog. *Limited Memory Quasi-Newton-Verfahren*, und lediglich die letzten $m \ll n$ Iterationen zum Update von B_0 zu verwenden. Die Konvergenzaussagen lassen sich entsprechend auch auf diesen Fall übertragen. Man beachte hierbei, dass jede Suchrichtung $d^k = -B_k \nabla f(x^k)$ mit positiv definitem B_k eine Abstiegsrichtung von f ist.

5. Für Verfahren die mit H_k anstelle von B_k arbeiten ist es notwendig direkt ein Update der Cholesky-Zerlegung von H_k durchzuführen und nicht eine neue Zerlegung von H_{k+1} zu bestimmen, um ein effizientes Verfahren zu erhalten.

2.9 Trust-Region-Verfahren

Wir werden nun die bisher betrachtete Klasse von Abstiegsverfahren verlassen und jetzt Suchrichtung und Schrittweite simultan bestimmen. Wir wollen also nun das Update

$$x^{k+1} = x^k + d^k$$

durchführen, ohne eine zusätzliche Schrittweite zu bestimmen. Hierzu beobachten wir, dass sich $f \in C^2(\mathbb{R}^n; \mathbb{R})$ lokal in x^k durch ein *quadratisches Modell*

$$m_k(d) = f_k + \langle g^k, d\rangle + \frac{1}{2}\langle d, H_k d\rangle$$

mit $f_k = f(x^k)$, $g^k = \nabla f(x^k)$ und $H_k = \nabla^2 f(x^k)$ approximieren lässt. Nach Taylor-Entwicklung wissen wir, dass

$$f(x^k + d) = m_k(d) + o(\|d\|^2),$$

wir sollten dem Modell also nur für hinreichend kleine Argumente d vertrauen. Wir definieren daher den *Vertrauensbereich* (oder *Trust-Region* (TR))

$$\{d \mid \|d\| \leq \Delta_k\}$$

auf dem wir den Vorhersagen von m_k vertrauen wollen. Dabei ist $\Delta_k > 0$ der *Trust-Region-Radius* ein noch zu bestimmender Parameter. Wir bestimmen dann den Schritt d durch Lösen des **Trust-Region-Problems**

$$\min_{d \in \mathbb{R}^n} m_k(d), \qquad \text{u.d.N. } \|d\| \leq \Delta_k. \tag{TR}$$

Anschließend sollten wir prüfen, ob unser Parameter Δ_k gut genug gewählt wurde. Hierzu bestimmt man die Modellvorhersage (predicted reduction)

$$\text{pred}_k(d^k) := m_k(0) - m_k(d^k) = f_k - m_k(d^k) = -\langle g^k, d^k\rangle - \frac{1}{2}\langle d^k, H_k d^k\rangle$$

mit der tatsächlichen Abnahme des Funktionswertes (actual reduction)

$$\text{ared}_k(d^k) := f_k - f(x^k + d^k).$$

Wir bestimmen dann

$$\rho_k(d^k) := \frac{\text{ared}_k(d^k)}{\text{pred}_k(d^k)}. \tag{2.13}$$

Für einen Parameter $\eta_1 \in (0,1)$ (z.B. $\eta_1 = 0.1$) prüfen wir dann, ob uns diese Verhältnis noch akzeptabel erscheint.

- Ist
$$\rho_k(d^k) \leq \eta_1$$
so erscheint uns die Vorhersage zu ungenau. Wir verwerfen dann d^k und setzen $x^{k+1} = x^k$. Für den nächsten Versuch wählen wir $\Delta_{k+1} < \Delta_k$.
- Im anderen Fall
$$\rho_k(d^k) > \eta_1$$
akzeptieren wir den Schritt und setzen $x^{k+1} = x^k + d^k$. Der nächste Trust-Region-Radius Δ_k wird dann abhängig von $\rho_k(d^k)$ mit $0 \leq \Delta_{\min} \leq \Delta_{k+1}$ angepasst.

Bemerkung 2.9.1.
1. Offenbar können wir auch hier wieder alle in den vorherigen Kapiteln gewonnenen Modifikationen vornehmen. Es kann z.B. H_k anstelle der Hesse-Matrix auch durch eine Quasi-Newton-Approximation bestimmt werden.
2. Das exakte Lösen des Trust-Region-Problems (TR) ist i.A. zu aufwändig. Es genügt für den Nachweis der globalen Konvergenz, wenn die Schritte d^k der Cauchy-Abstiegs-Bedingung genügen.

Definition 2.9.2 (Cauchy-Abstiegsbedingung (Fraction of Cauchy-Decrease))**.** Für gegebene Parameter $\alpha \in (0,1)$ und $\beta \geq 1$ genügt d^k der *Cauchy-Abstiegsbedingung*, falls

$$\|d^k\| \leq \beta \Delta_k, \qquad \text{pred}_k(d^k) \geq \alpha \, \text{pred}_k(d_c^k), \tag{CA}$$

wobei der *Cauchy-Schritt* d_c^k die eindeutige Lösung des (eindimensionalen) Problems

$$\min_{d \in \mathbb{R}^n} m_k(d), \qquad \text{u.d.N. } d = -t g^k, \quad t \geq 0, \quad \|d\| \leq \Delta_k \tag{2.14}$$

ist.

Bemerkung 2.9.3. Offenbar genügt der Cauchy-Schritt auch der Cauchy-Abstiegsbedingung. Da es sich hierbei jedoch wieder nur um eine Richtung des steilsten Abstiegs handelt sollte man diese i.A. nicht verwenden.

Wir fassen unsere Beobachtungen zusammen.

Algorithmus TRU [2.9.4] (Update des Trust-Region-Radius)**.**

Zu gegebenen Parameter $0 < \eta_1 < \eta_2 < 1$, $0 < \gamma_0 < \gamma_1 < 1 < \gamma_2$ und $\Delta_{\min} \geq 0$ (z.B. aus Algorithmus TR [2.9.5]). Bestimme Δ_{k+1} wie folgt

$$\Delta_{k+1} \in \begin{cases} [\gamma_0 \Delta_k, \gamma_1 \Delta_k] & \rho_k(d^k) \leq \eta_1, \\ [\max(\Delta_{\min}, \gamma_1 \Delta_k), \max(\Delta_{\min}, \Delta_k)] & \eta_1 < \rho_k(d^k) \leq \eta_2, \\ [\max(\Delta_{\min}, \Delta_k), \max(\Delta_{\min}, \gamma_2 \Delta_k)] & \eta_2 < \rho_k(d^k), \end{cases}$$

Algorithmus TR [2.9.5] (Trust-Region-Verfahren)**.**

- Wähle Parameter $\alpha \in (0,1]$, $\beta \geq 1$, $0 < \eta_1 < \eta_2 < 1$, $0 < \gamma_0 < \gamma_1 < 1 < \gamma_2$ und $\Delta_{\min} \geq 0$.
- Wähle Startpunkt $x^0 \in \mathbb{R}^n$ und Trust-Region-Radius $\Delta_0 > 0$ und $\Delta_0 \geq \Delta_{\min}$.

for k = 0, 1, ... **do**

- Falls $g^k = 0$, Stopp mit Lösung x^k.
- Wähle eine symmetrische Matrix $H_k \in \mathbb{R}^{n \times n}$.
- Berechne einen Schritt d^k welcher der Cauchy-Abstiegsbedingung (CA) genügt.
- Bestimme $\rho_k(d^k)$ aus (2.13).
- Berechne x^{k+1} durch

$$x^{k+1} = x^k + \begin{cases} d^k & \rho_k(d^k) > \eta_1, \\ 0 & \rho_k(d^k) \leq \eta_1. \end{cases}$$

- Bestimme Δ_{k+1} mittels Algorithmus TRU [2.9.4].

end for

Definition 2.9.6. Wir nennen einen Schritt d^k *erfolgreich*, falls $\rho_k(d^k) > \eta_1$ gilt und damit $x^{k+1} = x^k + d^k$ gesetzt wird. Mit $\mathscr{S} \subset \mathbb{N}_0$ bezeichnen wir die Menge aller Indizes erfolgreicher Schritte.

2.9.1 Globale Konvergenz

Wir wollen nun die globale Konvergenz des Trust-Region-Verfahrens zeigen. Hierzu benötigen wir zunächst Kontrolle über die Modellvorhersage im Vergleich zur Größe des Gradienten.

Lemma 2.9.7 (Modell-Abnahme). *Sei $f \in C^1(\mathbb{R}^n;\mathbb{R})$ und $\|H_k\| \le C_H$ für alle $k \in \mathbb{N}_0$. Ist dann $g^k \neq 0$ und genügt d^k der Cauchy-Abstiegsbedingung (CA), so gilt*

$$\operatorname{pred}_k(d^k) \ge \frac{\alpha}{2}\|g^k\| \min(\Delta_k, \|g^k\|/C_H).$$

Beweis. Es ist nach Annahme

$$\begin{aligned}
\operatorname{pred}_k(d^k) &\ge \alpha \operatorname{pred}_k(d_c^k) \\
&= \alpha\Big(f_k - \min_{t\in[0,\Delta_k/\|g^k\|]} m_k(-t g^k)\Big) \\
&= \alpha\Big(-\min_{t\in[0,\Delta_k/\|g^k\|]}\Big[-t\|g^k\|^2 + \frac{t^2}{2}\langle g^k, H_k g^k\rangle\Big]\Big) \\
&= \alpha \max_{t\in[0,\Delta_k/\|g^k\|]}\Big(t\|g^k\|^2 - \frac{t^2}{2}\langle g^k, H_k g^k\rangle\Big) \\
&=: \alpha \max_{t\in[0,\Delta_k/\|g^k\|]} \Phi(t).
\end{aligned}$$

Sei nun $t^* \in [0, \Delta_k/\|g^k\|] =: I$ der Maximierer von Φ.

- Ist dann $\langle g^k, H_k g^k\rangle \le 0$, so ist $\Phi\colon I \to \mathbb{R}$ monoton wachsend, und somit ist $t^* = \Delta_k/\|g^k\|$ und es folgt
$$\Phi(t^*) \ge t^*\|g^k\|^2 = \Delta_k\|g^k\|.$$

- Ist hingegen $\langle g^k, H_k g^k\rangle > 0$, so liegt das globale Maximum von $\Phi(t)$ im Punkt t_+ mit $\Phi'(t_+) = 0$, also
$$t_+ = \frac{\|g^k\|^2}{\langle g^k, H_k g^k\rangle},$$
und somit $t^* = \min\left(\frac{\Delta_k}{\|g^k\|}, \frac{\|g^k\|^2}{\langle g^k, H_k g^k\rangle}\right)$.
  - Ist der erste Term kleiner, so erhalten wir, wegen
$$t^* = \frac{\Delta_k}{\|g^k\|} \le \frac{\|g^k\|^2}{\langle g^k, H_k g^k\rangle}$$

die folgende Abschätzung

$$\Phi(t^*) = t^*\|g^k\|^2 - \frac{t^*}{2}\, t^*\langle g^k, H_k g^k\rangle \geq \frac{t^*}{2}\|g^k\|^2 = \frac{1}{2}\Delta_k\|g^k\|.$$

– Ansonsten ist

$$t^* = t_+ = \frac{\|g^k\|^2}{\langle g^k, H_k g^k\rangle} \leq \frac{\Delta_k}{\|g^k\|}$$

und somit

$$\begin{aligned}\Phi(t^*) &= \frac{\|g^k\|^2}{\langle g^k, H_k g^k\rangle}\|g^k\|^2 - \frac{\|g^k\|^4}{2\langle g^k, H_k g^k\rangle^2}\langle g^k, H_k g^k\rangle \\ &= \frac{\|g^k\|^4}{2\langle g^k, H_k g^k\rangle} \\ &\geq \frac{1}{2}\frac{\|g^k\|^2}{\|H_k\|} \\ &\geq \frac{1}{2}\frac{\|g^k\|^2}{C_H}.\end{aligned}$$

Dies zeigt die behauptete Abschätzung. □

Nun wollen wir sehen, dass für $g^k \neq 0$ auch irgendwann ein Schritt akzeptiert wird.

Lemma 2.9.8 (Wohldefiniertheit der Trust-Region-Iteration). *Sei $f \in C^1(\mathbb{R}^n;\mathbb{R})$ und $x \in \mathbb{R}^n$ mit $\nabla f(x) \neq 0$. Dann gibt es zu jedem $\eta \in (0,1)$ Konstanten $\delta = \delta(x,\eta) > 0$ und $\Delta = \Delta(x,\eta) > 0$, so dass Folgendes gilt.*

Ist $x^k \in \mathbb{R}^n$ mit $\|x^k - x\| \leq \delta$ und genügt $d^k \in \mathbb{R}^n$ der Cauchy-Abstiegsbedingung (CA) für das Trust-Region-Problem (TR) in x^k mit $\Delta_k \in (0,\Delta]$ und symmetrischen Matrizen H_k mit $\|H_k\| \leq C_H$. So gilt die Ungleichung

$$\rho_k(d^k) > \eta.$$

Beweis. Es ist

$$\rho^k(d^k) = 1 - \frac{\mathrm{pred}_k(d^k) - \mathrm{ared}_k(d^k)}{\mathrm{pred}_k(d^k)}.$$

Für eine untere Schranke an $\rho^k(d^k)$ benötigen wir somit eine obere Schranke an den Zähler des Bruchs sowie eine untere Schranke an den Nenner.

Obere Schranke: Nach dem Mittelwertsatz ist für ein $\theta \in (0,1)$

$$\begin{aligned}\mathrm{pred}_k(d^k) - \mathrm{ared}_k(d^k) &= f(x^k + d^k) - m_k(d^k) \\ &= \langle \nabla f(x^k + \theta d^k), d^k \rangle - \langle g^k, d^k \rangle - \frac{1}{2}\langle d^k, H_k d^k \rangle \\ &\leq \beta \Delta_k \|\nabla f(x^k + \theta d^k) - g^k\| + \frac{\beta^2 C_H}{2}\Delta_k^2.\end{aligned}$$

Wir beobachten weiter, dass

$$\|x^k - x\| \leq \delta, \quad \|x^k + \theta d^k - x\| \leq \|x^k - x\| + \theta\|d^k\| \leq \delta + \beta\Delta_k \leq \delta + \beta\Delta.$$

Somit folgt für $\delta + \Delta \to 0$ auch $g^k \to \nabla f(x)$ und $\nabla f(x^k + \theta d^k) \to \nabla f(x)$. Wir können also für gegebenes $\epsilon := \frac{\|\nabla f(x)\|}{2} > 0$ die Parameter δ und Δ so klein wählen, dass

$$\beta\Delta_k\|\nabla f(x^k + \theta d^k) - g^k\| + \frac{\beta^2 C_H}{2}\Delta_k^2 < (1-\eta)\frac{\alpha\epsilon}{2}\Delta_k.$$

Untere Schranke: Wir können nun $\delta > 0$ so klein wählen, dass

$$\|g^k\| = \|\nabla f(x^k)\| \geq \epsilon = \frac{\|\nabla f(x)\|}{2} \qquad \forall\, \|x^k - x\| \leq \delta.$$

Dann ist für $0 < \Delta \leq \frac{\epsilon}{C_H}$ und alle $\Delta_k \leq \Delta$ aufgrund des vorhergehenden Lemmas 2.9.7

$$\begin{aligned}\mathrm{pred}_k(d^k) &\geq \frac{\alpha}{2}\|g^k\| \min(\Delta_k, \|g^k\|/C_H) \\ &= \frac{\alpha}{2}\|g^k\|\Delta_k \\ &\geq \frac{\alpha\epsilon}{2}\Delta_k.\end{aligned}$$

Es folgt damit

$$\begin{aligned}\rho_k(d^k) &= 1 - \frac{\mathrm{pred}_k(d^k) - \mathrm{ared}_k(d^k)}{\mathrm{pred}_k(d^k)} \\ &> 1 - \frac{(1-\eta)\frac{1}{2}\alpha\epsilon\Delta_k}{\frac{1}{2}\alpha\epsilon\Delta_k} \\ &= \eta.\end{aligned}$$

□

Es ergibt sich hieraus unmittelbar das folgende

Korollar 2.9.9. *Angenommen das Trust-Region-Verfahren TR [2.9.5] terminiert nicht endlich. Sind dann $f \in C^1(\mathbb{R}^n;\mathbb{R})$ und $\|H_k\| \le C_H$ für alle k, dann erzeugt das Trust-Region-Verfahren unendlich viele erfolgreiche Schritte.*

Beweis. Angenommen, die Aussage wäre falsch. Dann gäbe es $l \ge 0$ mit $x^k = x^l$ und $\rho_k(d^k) \le \eta_1$ für alle $k \ge l$. Weiter ist

$$\Delta_k \le \gamma_1 \Delta_{k-1} \le \ldots \le \gamma_1^{k-l}\Delta_l \to 0 \qquad (k \to \infty).$$

Da $g^k \neq 0$ nach Voraussetzung gibt es nach dem vorhergehenden Lemma 2.9.8 mit $x = x^l$ und $\eta = \eta_1$ ein $\Delta > 0$, so dass $\rho(d^k) > \eta_1$ für alle $k \ge l$ mit $\Delta_k \le \Delta$. Dies ist ein Widerspruch, da ein solcher Schritt erfolgreich wäre! □

Lemma 2.9.10. *Sei die Folge x^k durch das Trust-Region-Verfahren TR [2.9.5] mit $f \in C^1(\mathbb{R}^n;\mathbb{R})$ nach unten beschränkt und $\|H_k\| \le C_H$ erzeugt. Ist dann $\mathcal{K} \subset \mathcal{S}$ eine unendliche Menge mit $\|g^k\| \ge \epsilon > 0$ für alle $k \in \mathcal{K}$, so ist*

$$\sum_{k\in\mathcal{K}} \Delta_k < \infty.$$

Beweis. Für $k \in \mathcal{K} \subset \mathcal{S}$ ist der Schritt d^k erfolgreich und folglich ist nach dem Lemma zur Modell-Abnahme 2.9.7

$$\begin{aligned} f(x^k) - f(x^{k+1}) &= \operatorname{ared}_k(d^k) \\ &> \eta_1 \operatorname{pred}_k(d^k) \\ &\ge \eta_1 \frac{\alpha}{2} \|g^k\| \min(\Delta_k, \|g^k\|/C_H) \\ &\ge \eta_1 \frac{\alpha}{2} \epsilon \min(\Delta_k, \epsilon/C_H). \end{aligned}$$

Da $f(x^k) \ge f(x^{k+1})$ für alle $k \in \mathbb{N}_0$ ist

$$\begin{aligned} \infty > f(x^0) - \inf_{x\in\mathbb{R}^n} f(x) &\ge f(x^0) - f(x^k) \\ &= \sum_{l\in\mathcal{S},l<k} \Big(f(x^l) - f(x^{l+1})\Big) \\ &\ge \sum_{l\in\mathcal{K},l<k} \Big(f(x^l) - f(x^{l+1})\Big) \\ &\ge \eta_1 \frac{\alpha}{2} \epsilon \sum_{l\in\mathcal{K},l<k} \min(\Delta_l, \epsilon/C_H). \end{aligned}$$

Es folgt

$$\sum_{l \in \mathcal{K}} \min(\Delta_l, \epsilon / C_H) < \infty$$

somit ist insbesondere für fast alle $k \in \mathcal{K}$

$$\Delta_k = \min(\Delta_k, \epsilon / C_H)$$

und somit folgt die Behauptung. □

Als Konsequenz erhalten wir das folgende Konvergenzresultat

Theorem 2.9.11 (Globale Konvergenz des TR-Verfahrens I). *Sei $f \in C^1(\mathbb{R}^n; \mathbb{R})$ nach unten beschränkt und $\|H_k\| \leq C_H$ für alle $k \in \mathbb{N}_0$. Dann terminiert das Trust-Region-Verfahren entweder in einem stationären Punkt x^k, oder es erzeugt eine unendliche Folge x^k mit*

$$\liminf_{k \to \infty} \|g^k\| = 0.$$

Ist ∇f gleichmäßig stetig auf einer Menge $\Omega \subset \mathbb{R}^n$ mit $x^k \in \Omega$, so ist sogar

$$\lim_{k \to \infty} \|g^k\| = 0.$$

Beweis. Im Falle des Abbruchs nach endlich vielen Iterationen ist nichts zu zeigen. Ansonsten wissen wir bereits (Korollar 2.9.9), dass $|\mathcal{S}| = \infty$.

1.) Zum Nachweis der Aussage $\liminf \|g^k\| = 0$ nehmen wir zum Widerspruch an, dies wäre falsch. Dann gibt es ein $\epsilon > 0$, so dass $\|g^k\| \geq \epsilon$ für alle $k \geq 0$. Aus dem vorhergehenden Lemma 2.9.10 wissen wir das folglich

$$\sum_{k \in \mathcal{S}} \Delta_k < \infty.$$

Damit ist insbesondere $\Delta_k \to 0$ für $\mathcal{S} \ni k \to \infty$. Ist dann $k > l$, so folgt

$$\|x^k - x^l\| \leq \sum_{i \in \mathcal{S}, l \leq i < k} \|d^i\| \leq \beta \sum_{i \in \mathcal{S}, l \leq i < k} \Delta_i \leq \beta \sum_{i \in \mathcal{S}, l \leq i} \Delta_i \to 0 \quad (l \to \infty).$$

Die Folge x^k ist also eine Cauchy-Folge. Wegen der Vollständigkeit des $\mathbb{R}^n$ gibt es somit einen Limes $\overline{x}$ und es ist

$$\|\nabla f(\overline{x})\| \geq \epsilon$$

da ∇f stetig ist.

Wir können nun das Lemma zur Wohldefiniertheit des TR-Verfahrens 2.9.8 mit $x = \overline{x}$ und $\eta = \eta_2$ anwenden. Dies liefert uns $\delta, \Delta > 0$, und wegen der Konvergenz $x^k \to \overline{x}$ ist $\|x^k - \overline{x}\| \leq \delta$ für alle $k \geq k_0$. Folglich ist dann nach Lemma 2.9.8 für $k \geq k_0$

$$\rho_k(d^k) > \eta_2 \qquad \text{falls } \Delta_k \leq \Delta.$$

Dies liefert uns nun eine untere Schranke an Δ_k. Denn es ist

$$\Delta_k \geq \min(\Delta_{k_0}, \gamma_0 \Delta) \quad \forall\, k \geq k_0.$$

Für $k = k_0$ ist dies offenkundig. Per Induktion sehen wir den Rest. Sei dazu $\Delta_k \geq \min(\Delta_{k_0}, \gamma_0 \Delta)$ für ein $k \geq k_0$.
Ist nun $\Delta_k \geq \Delta$, so erhalten wir

$$\Delta_{k+1} \geq \gamma_0 \Delta_k \geq \gamma_0 \Delta.$$

Andernfalls ist $\Delta_k < \Delta$ und folglich ist $\rho_k(d^k) \geq \eta_2$ und nach Konstruktion des TR-Radius folgt

$$\Delta_{k+1} \geq \Delta_k \geq \min(\Delta_{k_0}, \gamma_0 \Delta).$$

Dies steht im Widerspruch zu $\lim_{\mathscr{S} \ni k \to \infty} \Delta_k = 0$, und folglich ist $\liminf_{k \to \infty} \|g^k\| = 0$.

2.) Sei nun ∇f gleichmäßig stetig auf Ω. Aus Teil 1 wissen wir bereits $\liminf_{k \to \infty} \|g^k\| = 0$. Sei nun zum Widerspruch angenommen $\lim_{k \to \infty} \|g^k\|$ existiere nicht. Dann gibt es ein $\epsilon > 0$, so dass $\|g^k\| \geq 2\epsilon$ für unendlich viele $k \in \mathscr{S}$. Wegen $\liminf \|g^k\| = 0$ gibt es ebenfalls unendlich viele $k \in \mathscr{S}$ mit $\|g^k\| \leq \epsilon$. Wir können somit aufsteigende Folgen k_i, $l_i \in \mathscr{S}$ finden, so dass

$$\begin{aligned} k_1 < l_1 < k_2 < l_2 < \ldots & \\ \|g^{k_i}\| \geq 2\epsilon, & \\ \|g^k\| \geq \epsilon \quad \forall k \in \mathscr{K}_i := \{k_i, \ldots, l_i - 1\} \cap \mathscr{S}, & \\ \|g^{l_i}\| < \epsilon. & \end{aligned}$$

Die Menge $\mathscr{K} = \bigcup_{i=1}^{\infty} \mathscr{K}_i$ enthält unendlich viele Indizes, und es ist $\|g^k\| \geq \epsilon$ für $k \in \mathscr{K}$. Nach dem vorhergehenden Lemma 2.9.10 ist also

$$\sum_{k \in \mathscr{K}} \Delta_k < \infty.$$

Da die Mengen $\mathscr{K}_i$ disjunkt sind folgt

$$\sum_{k \in \mathscr{K}_i} \Delta_k \to 0 \qquad (i \to \infty).$$

Wir erhalten damit analog zu Teil 1.

$$\|x^{l_i} - x^{k_i}\| \leq \sum_{k \in \mathscr{K}_i} \|d^k\| \leq \beta \sum_{k \in \mathscr{K}_i} \Delta_k \to 0 \qquad (i \to \infty).$$

Andererseits ist

$$\|g^{l_i} - g^{k_i}\| \geq \left| \|g^{l_i}\| - \|g^{k_i}\| \right| > \epsilon$$

für alle $i \in \mathbb{N}_0$ im Widerspruch zur gleichmäßigen Stetigkeit von ∇f. Folglich ist $\lim_{k\to\infty} \|g^k\| = 0$.

□

In der Tat lässt sich unter der Annahme $\Delta_{\min} > 0$ noch folgende stärkere Aussage zeigen. Man beachte hierbei, das wir die gleichmäßige Stetigkeit von ∇f nicht voraussetzen müssen!

Theorem 2.9.12 (Globale Konvergenz des TR-Verfahrens II). *Es sei $f \in C^1(\mathbb{R}^n;\mathbb{R})$ nach unten beschränkt und $\|H_k\| \leq C_H$ für alle $k \in \mathbb{N}_0$. Ist dann $\Delta_{\min} > 0$, so terminiert der Trust-Region Algorithmus TR [2.9.5] entweder mit einem stationären Punkt, oder er erzeugt eine unendliche Folge x^k deren Häufungspunkte stationäre Punkte sind.*

Beweis. Im Fall des endlichen Abbruchs ist wie immer nichts zu zeigen. Ist dann $\overline{x}$ ein Häufungspunkt und $\mathcal{K} \subset \mathcal{S}$ so, dass $x^k \to \overline{x}$ für $\mathcal{K} \ni k \to \infty$. Angenommen es wäre $\nabla f(\overline{x}) \neq 0$. Dann gibt es $\epsilon > 0$, so dass $\|g^k\| \geq \epsilon$ für alle $k \in \mathcal{K}$. Nach dem Lemma 2.9.10 ist dann

$$\sum_{k\in\mathcal{K}} \Delta_k < \infty$$

und nach dem Lemma zur Wohldefiniertheit der TR-Iteration 2.9.8 gibt es $\delta, \Delta > 0$, so dass $l \in \mathcal{S}$ falls $x^l \in B_\delta(\overline{x})$ und $\Delta_l \leq \Delta$.
Ist dann $k \in \mathcal{K}$ hinreichend groß, so ist $x^k \in B_\delta(\overline{x})$. Ist dann $k-1 \in \mathcal{S}$, so ist $\Delta_k > \Delta_{\min} > 0$. Ansonsten ist $k-1 \notin \mathcal{S}$, dann ist jedoch $x^{k-1} = x^k$ und daher $\Delta_{k-1} > \Delta$, sonst wäre der Schritt $k-1$ erfolgreich gewesen. Damit folgt nun

$$\Delta_k \geq \gamma_0 \Delta_{k-1} > \gamma_0 \Delta.$$

Für fast alle $k \in \mathcal{K}$ ist damit $\Delta_k \geq \min(\gamma_0\Delta, \Delta_{\min})$ im Widerspruch zu $\Delta_k \to 0$ für $\mathcal{K} \ni k \to \infty$. □

2.9.2 Schnelle Lokale Konvergenz

Wir wollen uns nun überlegen, dass wir auch für das TR-Verfahren schnelle lokale Konvergenz erhalten können. Zu diesem Zweck modifizieren wir unseren TR-Algorithmus, in dem wir bevorzugt die Newton-Richtung auswählen. Wir erhalten also die folgende Variante unseres TR-Algorithmus TR [2.9.5]:

Algorithmus TRN [2.9.13] (Trust-Region-Newton-Verfahren).

- Wähle Parameter $\alpha \in (0,1)$, $\beta \geq 1$, $0 < \eta_1 < \eta_2 < 1$, $0 < \gamma_0 < \gamma_1 < 1 < \gamma_2$ und $\Delta_{\min} \geq 0$.
- Wähle Startpunkt $x^0 \in \mathbb{R}^n$ und Trust-Region-Radius $\Delta_0 > 0$ und $\Delta_0 \geq \Delta_{\min}$.

for $k = 0, 1, \ldots$ **do**

- Falls $g^k = 0$, Stopp mit Lösung x^k.
- **Wähle $\mathbf{H_k = \nabla^2 f(x^k)}$.**
- **Falls der Newton-Schritt $d_n^k = -H_k^{-1} g^k$ existiert und (CA) erfüllt, so wähle $\mathbf{d^k = d_n^k}$.** Ansonsten bestimme ein d^k, so dass (CA) erfüllt ist.
- Bestimme $\rho_k(d^k)$ aus (2.13).
- Berechne x^{k+1} durch

$$x^{k+1} = x^k + \begin{cases} d^k & \rho_k(d^k) > \eta_1, \\ 0 & \rho_k(d^k) \leq \eta_1. \end{cases}$$

- Bestimme Δ_{k+1} mittels Algorithmus TRU [2.9.4].

end for

Bemerkung 2.9.14. Anstelle des Newton-Schrittes könnten wir natürlich auch irgendeine andere Variante mit schneller lokaler Konvergenz wählen.

Theorem 2.9.15. *Sei $f \in C^2(\mathbb{R}^n;\mathbb{R})$ und die Niveaumenge $N_f(x^0)$ kompakt. Erzeugt das Trust-Region-Newton Verfahren TRN [2.9.13] eine Folge x^k mit einem Häufungspunkt $\overline{x}$, in dem $\nabla^2 f(\overline{x})$ positiv definit ist, so gilt*

1. *$x^k \to \overline{x}$, $g^k \to 0$ und damit auch $\nabla f(\overline{x}) = 0$.*
2. *Es gibt ein $l \geq 0$, so dass für alle $k \geq l$ jeder Schritt d^k erfolgreich ist und $\Delta_k \geq \Delta_l > 0$ gilt.*
3. *Es gibt ein $l' \geq l$ mit $d^k = d_n^k$ für alle $k \geq l'$. Das Verfahren geht also für $k \geq l'$ in das lokale Newton-Verfahren NVO [2.5.7] über und die Folge x^k konvergiert q-superlinear gegen $\overline{x}$ (sogar q-quadratisch falls $\nabla^2 f$ L-stetig in einer Umgebung von $\overline{x}$).*

Beweis. 1.) Nach Konstruktion des Verfahrens liegen alle Iterierten $x^k \in N_f(x^0)$. Da $N_f(x^0)$ kompakt ist, ist die stetige Funktion f auf $N_f(x^0)$ nach unten beschränkt, ∇f gleichmäßig stetig und es gilt $\|H_k\| = \|\nabla^2 f(x^k)\| \leq C_H$.

Aus dem globalen Konvergenzsatz 2.9.11 folgt somit

$$\lim_{k\to\infty} g^k = 0.$$

Wegen der Stetigkeit von ∇f ist dann $\nabla f(\overline{x}) = 0$. Damit sind in $\overline{x}$ die hinreichenden Bedingungen zweiter Ordnung erfüllt, und nach Lemma 2.5.9 gibt es dann $\mu, \epsilon > 0$, so dass

$$\langle d, \nabla^2 f(x) d\rangle \geq \mu \|d\|^2 \quad \forall\, x \in B_\epsilon(\overline{x}), \forall\, d \in \mathbb{R}^n.$$

Auf Grund des Lemmas über Nullstellen differenzierbarer Funktionen 2.5.4 ist somit $\overline{x}$ die einzige Nullstelle von ∇f auf $B_\epsilon(\overline{x})$ wobei wir ggf. ϵ verkleinern. Da $g^k \to 0$ ist damit $\overline{x}$ auch der einzige Häufungspunkt der Folge x^k auf $B_\epsilon(\overline{x})$.

Da $d^k \neq 0$ ist

$$0 > -\mathrm{pred}_k(d^k) = \langle g^k, d^k\rangle + \frac{1}{2}\langle d^k, \nabla^2 f(x^k) d^k\rangle \geq \Big(-\|g^k\| + \frac{\mu}{2}\|d^k\|\Big)\|d^k\|$$

und somit

$$0 < \frac{\mu}{2}\|d^k\| < \|g^k\|.$$

Ist nun $x^k \to \overline{x}$ für $\mathscr{K} \ni k \to \infty$ eine gegen $\overline{x}$ konvergente Teilfolge, so ist deshalb offenbar

$$0 < \|d^k\| < \frac{2}{\mu}\|g^k\| \to 0 \qquad (\mathscr{K} \ni k \to \infty). \tag{2.15}$$

Da $x^{k+1} - x^k \in \{0, d^k\}$ ist damit

$$x^{k+1} - x^k \to 0 \qquad (\mathscr{K} \ni k \to \infty).$$

Da $\overline{x}$ ein isolierter Häufungspunkt ist konvergiert damit wegen Lemma 2.5.15 bereits die ganze Folge x^k gegen $\overline{x}$.

2.) Um zu sehen, dass für hinreichend großes k nur noch erfolgreiche Schritte vorkommen zeigen wir nun $\rho_k(d^k) \to 1$. Sei dazu zunächst l so gewählt, dass $x^k \in B_\epsilon(\overline{x})$ für alle $k \geq l$, mit der in Teil 1 definierten Umgebung $B_\epsilon(\overline{x})$. Wir setzten nun mit den Bezeichnungen aus Teil 1 und dem Algorithmus

$$c = \frac{\alpha\mu}{4}\min(1/\beta, \mu/(2C_H)).$$

Damit folgt aus der Cauchy-Abstiegsbedingung (CA), dem Lemma zur Modell-Abnahme 2.9.7 sowie (2.15)

$$\begin{aligned}
\mathrm{pred}_k(d^k) &\geq \frac{\alpha}{2}\|g^k\|\min(\Delta_k, \|g_k\|/C_H) \\
&\geq \frac{\alpha\mu}{4}\|d^k\|\min\Big(\frac{\|d^k\|}{\beta}, \frac{\mu}{2C_H}\|d^k\|\Big) \\
&= c\|d^k\|^2.
\end{aligned}$$

Nach Taylor-Entwicklung gibt es $\theta_k \in (0,1)$, so dass

$$\begin{aligned}
|1-\rho_k(d^k)| &= \frac{|\operatorname{pred}_k(d^k)-\operatorname{ared}_k(d^k)|}{\operatorname{pred}_k(d^k)} \\
&= \frac{|(f(x^k+d^k)-f(x^k)-\langle g^k,d^k\rangle)-1/2\langle d^k,H_k,d^k\rangle|}{\operatorname{pred}_k(d^k)} \\
&\leq \frac{|\langle d^k,(\nabla^2 f(x^k+\theta_k d^k)-\nabla^2 f(x^k))d^k\rangle|}{2c\|d^k\|^2} \\
&\leq \frac{1}{2c}\|\nabla^2 f(x^k+\theta_k d^k)-\nabla^2 f(x^k)\| \\
&\to 0
\end{aligned}$$

da $d^k \to 0$ und $\nabla^2 f$ auf der kompakten Menge $N_f(x^0)$ gleichmäßig stetig ist. In dem wir l ggf. vergrößern können wir erreichen, dass $\rho_k(d^k) \geq \eta_2$ für alle $k \geq l$ gilt. Dann ist auch d^k erfolgreich und entsprechend der Update-Regeln für den Trust-Region-Radius in Algorithmus TRU [2.9.4] ist auch $\Delta_k \geq \Delta_l > 0$.

3.) Für $k \geq l$ ist $x^k \in B_\epsilon(\overline{x})$ und somit ist H_k invertierbar mit $\|H_k^{-1}\| \leq \frac{1}{\mu}$. Die Newton-Schritte existieren also und es ist

$$\|d_n^k\| \leq \|H_k^{-1}\|\|g^k\| \leq \frac{1}{\mu}\|g^k\| \to 0 \qquad (k \to \infty).$$

Wir können daher ein $l' \geq l$ wählen, so dass

$$\|d_n^k\| \leq \Delta_l \leq \Delta_k$$

für alle $k \geq l'$ gilt. Da der Newton-Schritt d_n^k das globale Minimum von m_k auf dem $\mathbb{R}^n$ ist (man beachte hierzu, dass für $k \geq l'$ notwendig H_k positiv-definit ist) ist d_n^k auch die Lösung des Trust-Region-Problems (TR) und erfüllt somit die Cauchy-Abstiegsbedingung (CA). Für $k \geq l'$ führen wir also das lokale Newton-Verfahren NVO [2.5.7] durch und die Konvergenzgeschwindigkeit folgt aus dem Konvergenzsatz 2.5.10 für dieses.

□

2.9.3 Lösung des Trust-Region-Problems

Wir haben bereits im Lemma zur Abnahme im Modell 2.9.7 gesehen, dass wir den Cauchy-Punkt durch

$$d_c^k = -t_* \Delta_K \frac{g^k}{\|g^k\|} \tag{2.16}$$

mit

$$t^* = \begin{cases} 1 & \langle g^k, H_k g^k\rangle \leq 0, \\ \min\left(1, \frac{\|g^k\|^3}{\Delta_k \langle g^k, H_k g^k\rangle}\right) & \text{sonst} \end{cases}$$

darstellen können. Offenkundig ist dies nicht die beste Wahl, da wir ansonsten stets ein Gradientenverfahren mit spezieller Schrittweite durchführen und dieses bereits bei optimaler Schrittweitenwahl nicht besonders schnell konvergiert.

2.9.3.1 Charakterisierung von Lösungen des Trust-Region Problems

In der Tat ist das Trust-Region Problem sogar (nahezu) exakt Lösbar. Hierzu betrachten wir zunächst die notwendigen, und erstaunlicher Weise auch hinreichenden, Bedingungen für eine Lösung des Trust-Region Problems (TR).

Theorem 2.9.16. *Es gelten folgende Aussagen:*

1. *Das TR-Problem (TR) besitzt mindestens eine (globale) Lösung.*
2. *Ein Vektor $d^k \in \mathbb{R}^n$ ist genau dann eine globale Lösung des TR-Problems (TR), wenn es ein $\lambda \in \mathbb{R}$ gibt, so dass*

$$\begin{aligned} \|d^k\| &\leq \Delta_k, \\ \lambda &\geq 0, \\ \lambda(\|d^k\| - \Delta_k) &= 0 \\ (H_k + \lambda I)d^k &= -g^k \\ H_k + \lambda I &\text{ ist positiv semidefinit.} \end{aligned} \tag{2.17}$$

3. *Gilt (2.17) und ist $H_k + \lambda I$ positiv definit, so ist die Lösung d^k des TR-Problems (TR) sogar eindeutig.*

Beweis. 1. Da $\{d \mid \|d\| \leq \Delta_k\}$ kompakt und m_k stetig ist, gibt es mindestens eine Lösung des Trust-Region Problems (TR).

2. „⇒" Sei also zunächst d^k als globale Lösung des TR-Problems angenommen. Die Ungleichung $\|d^k\| \leq \Delta_k$ ist somit erfüllt.

   Wir unterscheiden nun zwei Fälle.

   a) Ist $\|d^k\| < \Delta_k$, so ist d^k ein lokales Minimum von m_k bzgl. $\mathbb{R}^n$ und mit $\lambda = 0$ folgen die restlichen Bedingungen in (2.17) aus den notwendigen Bedingungen zweiter Ordnung (Theorem 2.1.6).

   b) Im anderen Fall ist $\|d^k\| = \Delta_k$. Angenommen es gäbe kein λ, so dass die Bedingungen 2-4 von (2.17) gelten. Wir setzen $y^k = \nabla m_k(d^k) = H^k d^k + g^k \neq 0$ (sonst erfüllt $\lambda = 0$ die Bedingung 2-4). Ferner gibt es kein $t \geq 0$, mit $y^k = -td^k$ (sonst wäre ja $(H_k + tI)d^k = -g^k$) und somit folgt (strikte Ungleichung!)

$$\langle y^k, d^k \rangle > -\|y^k\|\|d^k\|.$$

Somit ist also

$$\cos\angle(y^k, d^k) = \frac{\langle y^k, d^k\rangle}{\|y^k\|\|d^k\|} > -1$$

bzw. $\angle(y^k, d^k) \neq \pi$. Es folgt für die Winkelhalbierende

$$v^k = -\frac{y^k}{\|y^k\|} - \frac{d^k}{\|d^k\|}$$

(zwischen $-y^k$ und $-d^k$) die Ungleichung

$$\langle y^k, v^k\rangle = -\|y^k\| - \frac{\langle y^k, d^k\rangle}{\|d^k\|} = -\|y^k\|(1+\cos\angle(y^k, d^k)) < 0.$$

Somit ist v^k eine Abstiegsrichtung von m_k in d^k. Ferner ist

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|d^k + tv^k\|^2\Big|_{t=0} &= \langle v^k, d^k\rangle \\ &= -\left(\frac{\langle y^k, d^k\rangle}{\|y^k\|} + \|d^k\|\right) \\ &= -\|d^k\|(1+\cos\angle(y^k, d^k)) < 0.\end{aligned}$$

Für hinreichend kleines $t > 0$ ist somit

$$\|d^k + tv^k\| < \|d^k\| = \Delta_k$$

im Widerspruch zur Optimalität von d^k. Es gibt also ein λ welches den Bedingungen 2-4 von (2.17) genügt. Es bleibt die positiv Semidefinitheit von $H_k + \lambda I$ zu zeigen. Sei dazu zunächst $s \in \mathbb{R}^n$ beliebig mit $\langle s, d^k\rangle < 0$ gewählt. Wir setzen

$$t = -2\frac{\langle s, d^k\rangle}{\|s\|^2} > 0.$$

Dann ist

$$\|d^k + ts\|^2 = \|d^k\|^2 + 2t\langle s, d^k\rangle + t^2\|s\|^2 = \|d^k\|^2 \leq \Delta_k^2$$

und es folgt unter Nutzung von (2.17)-4. sowie der Definition von t

$$\begin{aligned}0 &\leq m_k(d^k + ts) - m_k(d^k) \\ &= t\langle y^k, s\rangle + \frac{t^2}{2}\langle s, H_k s\rangle \\ &= -\lambda t\langle d^k, s\rangle + \frac{t^2}{2}\langle s, H_k s\rangle \\ &= \lambda\frac{t^2}{2}\|s\|^2 + \frac{t^2}{2}\langle s, H_k s\rangle \\ &= \frac{t^2}{2}\langle s, (H_k + \lambda I)s\rangle.\end{aligned}$$

Aufgrund der Stetigkeit ist damit auch für $\langle s, d^k \rangle \leq 0$

$$\langle s, (H_k + \lambda I)s \rangle \geq 0.$$

Da das Vorzeichen von s in der Ungleichung keine Rolle spielt folgt damit die positiv Semidefinitheit von $H_k + \lambda I$.

„⇐" Für die umgekehrte Richtung nehmen wir an, es seien für $d^k \in \mathbb{R}^n$ und λ die Bedingungen (2.17) erfüllt. Für beliebiges $h \in \mathbb{R}^n$ mit $\|h\| \leq \Delta_k$ setzen wir $s = h - d^k$ und $y^k = \nabla m_k(d^k) = H_k d^k + g^k$ und es folgt mit (2.17)-4. und der Semidefinitheit von $H + \lambda I$

$$\begin{aligned} m_k(h) - m_k(d^k) &= m_k(d^k + s) - m_k(d^k) \\ &= \langle y^k, s \rangle + \frac{1}{2}\langle s, H_k s \rangle \\ &= -\lambda \langle d^k, s \rangle + \frac{1}{2}\langle s, H_k s \rangle \\ &\geq -\lambda \langle d^k, s \rangle - \frac{1}{2}\lambda \|s\|^2 \\ &= -\frac{\lambda}{2}\Big(2\langle d^k, s \rangle + \|s\|^2\Big) \\ &= -\frac{\lambda}{2}\Big(\|d^k + s\|^2 - \|d^k\|^2\Big) \\ &= -\frac{\lambda}{2}\Big(\|h\|^2 - \|d^k\|^2\Big). \end{aligned} \tag{2.18}$$

Ist nun $\lambda = 0$, so folgt aus (2.18) sofort, dass $m_k(h) - m_k(d^k) \geq 0$ ist. Ansonsten ist $\lambda > 0$ und dann notwendig wegen der 3. Bedingung in (2.17)

$$\|d^k\| = \Delta_k \geq \|h\|$$

und damit wieder $m_k(h) - m_k(d^k) \geq 0$. Somit ist d^k ein globales Minimum von m_k auf der Trust-Region.

3. Ist $H_k + \lambda I$ sogar positiv definit, so folgt in (2.18) eine strikte Ungleichung und damit die Behauptung.

□

Bemerkung 2.9.17. An dieser Stelle ist Bemerkenswert, dass die Bedingungen in Theorem 2.9.16 nicht nur notwendig sondern sogar hinreichend sind, und das obwohl m_k nicht notwendig konvex ist!

Bemerkung 2.9.18. In der Tat erfüllt der TR-Schritt d^k gerade die Gleichung

$$(H_k + \lambda I)d^k = -g^k$$

des Levenberg-Marquardt Schrittes mit der speziellen Skalierung λ.

2.9.3.2 Dogleg-Verfahren

Eine relativ einfache Variante, um eine approximative Lösung des Trust-Region-Problems zu finden welche besser als der Cauchy-Punkt ist, ist das sog. *Dogleg-Verfahren*. Wir beachten dazu, dass wir bereits im Satz zur schnellen lokalen Konvergenz des Verfahrens gesehen haben, dass in der Umgebung eines Punktes in dem die hinreichenden Bedingungen zweiter Ordnung erfüllte sind die Wahl der Newton-Richtung geeignet ist. Da diese jedoch i.A außerhalb der Trust-Region liegt müssen wir ggf. zwischen dem Cauchy-Punkt und diesem interpolieren. Hierzu definieren wir für positiv definites H_k den Dogleg-Pfad

$$\tilde{d}^k : [0,2] \to \mathbb{R}^n, \qquad \tilde{d}^k(t) = \begin{cases} t d_g^k & 0 \le t \le 1, \\ d_g^k + (t-1)(d_n^k - d_g^k) & 1 \le t \le 2. \end{cases}$$

Hierbei ist $d_g^k = \operatorname{argmin} m^k(d)$ u.d.N. $d \in \{-\lambda g^k \mid \lambda \ge 0\}$ der (unbeschränkte) Minimierer in Gradientenrichtung. und $d_n^k = -H_k^{-1} g^k$ der volle Newton-Schritt in x^k.

Das dies in der Tat eine vernünftige Wahl ist sagt uns das folgende Lemma.

Lemma 2.9.19 (Dogleg-Pfad)**.** *Ist H_k positiv definit und $g^k \neq 0$, so gibt es genau ein Minimum von m_k auf $\tilde{d}^k([0,2])$.*

Beweis. Übung! □

2.9.3.3 Steihaug-CG

Der Dogleg-Pfad kann in der Tat modifizierte werden, für den Fall, dass H_k nicht positiv definit ist. Da dann d_n^k jedoch nicht das globale Minimum von m_k ist, ist die Berechnung von d_n^k jedoch nicht sinnvoll.

Alternativ, kann man das CG-Verfahren zur Lösung der Newton-Gleichung $H_k d^k = -g^k$ modifizieren, damit es auch für nicht positiv definite Matrizen H_k zur Bestimmung einer Richtung d^k in unserem Trust-Region Algorithmus brauchbar ist. Die Idee ist hierbei einen Polygonzug zur Approximation des exakten Minimums aus dem CG-Verfahren zu konstruieren.

Algorithmus 2.9.20 (Steihaug CG)**.**

- Wähle TOL ≥ 0.
- Wähle Startpunkt $d_0^k = 0 \in \mathbb{R}^n$.
- Setze $r_0 = g^k, s_0 = -r_0 \in \mathbb{R}^n$.
- Stopp, falls $\|r_0\| <$ TOL.

for j $= 0, 1, \ldots$ **do**
  if $\langle s_j, H_k s_j \rangle \leq 0$ **then**
    - Finde $t_* > 0$, so dass $d = d_j^k + t_* s_j$, mit $\|d\| = \Delta_k$ das Problem

$$\min_{t>0} m_k(d_j^k + t s_j), \quad \|d_j^k + t s_j\| \leq \Delta_k$$

    löst.
  return $d^k = d$.
  end if
  - Setze

$$\alpha_j = \frac{\|r_j\|^2}{\langle s_j, H_k s_j \rangle}$$

  - Setze

$$d_{j+1}^k = d_j^k + \alpha_j s_j$$

  if $\|d_{j+1}^k\| \geq \Delta_k$ **then**
    - Finde $t \geq 0$, so dass $d = d_j^k + t s_j$ die Bedingung $\|d\| = \Delta_k$ erfüllt.
  return $d^k = d$.
  end if
  - Setze

$$r_{j+1} = r_j + \alpha_j H_k s_j$$

  if $\|r_{j+1}\| <$ TOL$\|r_0\|$
then return $d^k = d_{j+1}^k$
  end if
  - Setze

$$\beta_{j+1} = \frac{\|r_{j+1}\|^2}{\|r_j\|^2}$$

  - Setze

$$s_{j+1} = r_{j+1} + \beta_{j+1} s_j$$

end for

Hierbei ist wichtig zu sehen, dass wegen $d_0^k = 0$ nach einer Iteration der Cauchy-Punkt

$$d_1^k = \alpha_0 s_0 = -\frac{\|g^k\|^2}{\langle g^k, H_k g^k \rangle} g^k$$

ggf. mit Skalierung, um $\|d_1^k\| \leq \Delta_k$ zu erreichen gewählt wird. Da das CG Verfahren in jedem Schritt das Modell m_k entlang der Richtung s_j minimiert erfüllt der Rückgabewert d^k stets die Cauchy-Abstiegsbedingung (CA).

Das der Abbruch des Verfahrens sobald $\|d_j^k\| \geq \Delta_k$ in der Tat vernünftig ist zeigt das folgende Lemma

Lemma 2.9.21. *Die Iterierten des Algorithmus 2.9.20 erfüllen*

$$0 = \|d_0^k\| < \ldots < \|d_j^k\| \leq \|d^k\| \leq \Delta.$$

Beweis. Übung, bzw. Literatur, z.B. [Nocedal and Wright, 1999, Theorem 4.2]. □

3 Restringierte Optimierung

Wie in der Einleitung angekündigt betrachten wir nun das *nichtlineare Optimierungsproblem (Nonlinear Program, NLP)*

$$\min f(x) \quad \begin{cases} g(x) \leq 0, \\ h(x) = 0 \end{cases} \tag{NLP}$$

mit stetig differenzierbaren Funktionen $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ und $h : \mathbb{R}^n \to \mathbb{R}^p$.

Definition 3.0.1. Wir nennen die Menge

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

den *zulässigen Bereich* von (NLP). Ein Punkt $x \in \mathbb{R}^n$ heißt *zulässig* (*unzulässig*), falls $x \in X$ ($x \notin X$) ist.
Wir definieren die Abkürzungen

$$\mathcal{U} = \{1, \ldots, m\}, \qquad \mathcal{G} = \{1, \ldots, p\}.$$

Für einen zulässigen Punkt $x \in X$ definieren wir die *Indexmenge der aktiven Ungleichungsnebenbedingungen* $\mathcal{A}(x)$ sowie die *Indexmenge der inaktiven Ungleichungsnebenbedingungen* $\mathcal{I}(x)$ durch

$$\mathcal{A}(x) = \{i \in \mathcal{U} \mid g_i(x) = 0\},$$
$$\mathcal{I}(x) = \mathcal{U} \setminus \mathcal{A}(x) = \{i \in \mathcal{U} \mid g_i(x) < 0\}.$$

Da wir im folgenden öfters einmal gewisse Komponenten eines Vektors auswählen müssen ist folgende Notation hilfreich

Definition 3.0.2. Für einen Vektor $v \in \mathbb{R}^n$ und einer Indexmenge $\mathcal{J} \subset \{1, \ldots, n\}$ definieren wir mit $v_{\mathcal{J}} \in \mathbb{R}^{|\mathcal{J}|}$ den aus den Komponenten v_j, $j \in \mathcal{J}$, bestehenden Vektor.

Bemerkung 3.0.3. Man beachte, dass beschränkte Optimierungsprobleme oft viele

(äquivalente) Reformulierungen haben. Die Eigenschaften dieser Reformulierungen können jedoch erheblich von einander abweichen. Um dies einzusehen überlegt man sich sofort, dass z.B. im $\mathbb{R}^2$ die Beschränkung

$$|x|_1 \leq 1 \quad \text{d.h.} \quad (|x_1| + |x_2| \leq 1)$$

mit der nicht differenzierbaren Funktion $g : \mathbb{R}^2 \to \mathbb{R}; g(x) = |x|_1 - 1$ äquivalent durch die Funktion $\tilde{g} : \mathbb{R}^2 \to \mathbb{R}^4$ mit

$$\tilde{g}(x) = \begin{pmatrix} x_1 + x_2 - 1 \\ x_1 - x_2 - 1 \\ -x_1 + x_2 - 1 \\ -x_1 - x_2 - 1 \end{pmatrix} \leq 0$$

dargestellt werden kann.

Dies gilt natürlich auch für unbeschränkte Probleme, bei denen nicht differenzierbare Zielfunktionen f durch Einführung geeigneter Hilfsvariablen in differenzierbare Probleme überführt werden können. So sind z.B. die beiden Probleme

$$\min_{x \in \mathbb{R}^2} \|x\|_\infty = \min_{x \in \mathbb{R}^2} \max(|x_1|, |x_2|)$$

und

$$\min_{(x_1, x_2, s) \in \mathbb{R}^3} s$$
$$\text{u.d.N.} \begin{pmatrix} x_1 - s \\ -x_1 - s \\ x_2 - s \\ -x_2 - s \end{pmatrix} \leq 0$$

äquivalent. Das Erste besitzt jedoch eine nicht differenzierbare Zielfunktion, während das Zweiten nur differenzierbare (sogar lineare) Funktionen benötigt.

Bemerkung 3.0.4. Der Rand der zulässigen Menge $\{x \in \mathbb{R}^n \mid g(x) \leq 0\}$ besteht aus dem Schnitt der Mengen

$$M_i = \{x \in \mathbb{R}^n \mid g_i(x) = 0\}.$$

Gilt dann in einem Punkt $\overline{x}$ das $\nabla g_i(\overline{x}) \neq 0$, so ist M_i nahe $\overline{x}$ eine $(n-1)$-dimensionale C^1-Untermannigfaltigkeit, also der Graph einer C^1 Funktion $\mathbb{R}^n \to \mathbb{R}$. Dann steht $\nabla g_i(\overline{x})$ senkrecht auf dem Tangentialraum an M_i; es ist sogar $\nabla g_i(\overline{x})$ eine äußere Normale an die zulässige Menge.

Bsp. $g_i(x) = x_1^2 + x_2^2 - 1$ für $n = 2$. Dann ist in $\overline{x} = (1,0)^T$ der Gradient $\nabla g_i(x) =$

$(1,0)^T \neq 0$ und in der Tat ist M_i lokal der Graph der Funktion $x_1 = \sqrt{1-x_2^2}$.

3.1 Optimalitätsbedingungen

3.1.1 Notwendige Optimalitätsbedingungen erster Ordnung

Wir wollen nun notwendige Optimalitätsbedingungen für lokale Lösungen $\overline{x}$ von (NLP) herleiten. Hierzu benötigen wir zunächst einige Definitionen:

Definition 3.1.1. Eine Menge $K \subset \mathbb{R}^n$ heißt *Kegel*, falls für beliebiges $x \in K$ auch

$$\lambda x \in K, \qquad \forall \lambda > 0.$$

Definition 3.1.2. Zu einer nichtleeren Menge $M \subset \mathbb{R}^n$ definieren wir den *Tangentialkegel* an M im Punkt $x \in M$ durch

$$T(M,x) = \left\{ d \in \mathbb{R}^n \mid \exists \mathbb{R} \ni \eta_k > 0, x^k \in M : \lim_{k\to\infty} x^k = x, \lim_{k\to\infty} \eta_k (x^k - x) = d \right\}.$$

Hiermit erhalten wir eine erste Fassung der notwendigen Optimalitätsbedingungen:

Theorem 3.1.3. *[Notwendige Bedingungen erster Ordnung] Sei $\overline{x} \in \mathbb{R}^n$ eine lokale Lösung von (NLP). Dann gilt*

1. $\overline{x} \in X$,
2. $\langle \nabla f(\overline{x}), d \rangle \geq 0$ *für alle* $d \in T(X, \overline{x})$.

Beweis. 1.) gilt per Definition einer lokalen Lösung. Für 2.) nehmen wir $d \in T(X,\overline{x})$. Es gibt also Folgen $x^k \in X$ und $\eta_k \in \mathbb{R}_> 0$ mit

$$x^k \to \overline{x}, \quad d^k := \eta_k (x^k - x) \to d.$$

Nach Definition eines lokalen Minimums ist für k hinreichend groß $f(x^k) \geq f(\overline{x})$ und somit folgt

$$\begin{aligned} 0 &\leq \eta_k (f(x^k) - f(\overline{x})) \\ &= \eta_k \langle \nabla f(\overline{x}), x^k - \overline{x} \rangle + \eta_k o(\|x^k - \overline{x}\|) \\ &= \langle \nabla f(\overline{x}), d^k \rangle + \|d^k\| \frac{o(\|x^k - \overline{x}\|)}{\|x^k - \overline{x}\|} \\ &\to \langle \nabla f(\overline{x}), d \rangle \end{aligned}$$

und somit die Behauptung. □

Problematisch ist an dieser Darstellung, dass der Tangentialkegel i.A. nur schwer zu bestimmen ist, und auch kein 'schönes' Objekt ist. Insbesondere ist $T(X,x)$ i.A. kein Polyeder, d.h. er lässt sich i.A. nicht durch endlich viele lineare Ungleichungen beschreiben. Um dieses zu umgehen betrachten wir den linearisierten Tangentialkegel

Definition 3.1.4. Wir definieren mit

$$T_l(g,h,x) = \left\{ d \in \mathbb{R}^n \,\middle|\, \langle \nabla g_i(x), d\rangle \leq 0, i \in \mathscr{A}(x);\ \langle \nabla h_i(x), d\rangle = 0, i \in \mathscr{G} \right\}$$

den *linearisierten Tangentialkegel* in $x \in X$ zur Darstellung (NLP) von X.

Bemerkung 3.1.5. Man beachte, dass der Tangentialkegel selbst lediglich von der Gestalt des zulässigen Bereichs abhängt. Der linearisierte Kegel ist jedoch von der gewählten Darstellung des zulässigen Bereichs abhängig.

Bemerkung 3.1.6. Der linearisierte Tangentialkegel in $\overline{x}$ kann auch als Tangentialkegel der im Punkt $\overline{x}$ linearisieren Nebenbedingungen erhalten werden. Hierzu definiert man die Menge

$$X_l(\overline{x}) = \left\{ x \in \mathbb{R}^n \,\middle|\, g(\overline{x}) + \left(\langle \nabla g_i(\overline{x}), x - \overline{x}\rangle\right)_{i=1}^m \leq 0,\ h(\overline{x}) + \left(\langle \nabla h_i(\overline{x}), x - \overline{x}\rangle\right)_{i=1}^p = 0 \right\}.$$

Dann ist

$$T_l(g,h,\overline{x}) = T(X_l(\overline{x}), \overline{x}).$$

Da der linearisiert Tangentialkegel von der Wahl der Darstellung der Menge X abhängt stellt sich natürlich die Frage, ob dieser überhaupt eine vernünftige Wahl sein kann.

Bemerkung 3.1.7. Wie das Beispiel der Menge $X = (-\infty, 0]$ mit dem Kegel

$$T(X,0) = \{d \in \mathbb{R} \mid d \leq 0\}$$

und den Darstellungen $X = \{x \in \mathbb{R} \mid x =: g_1(x) \leq 0\} = \{x \in \mathbb{R} \mid x^3 =: g_2(x) \leq 0\}$ und den zugehörigen linearisierten Kegeln

$$T_l(g_1,0) = \{d \in \mathbb{R} \mid g_1'(0)d = d \leq 0\} = T(X,0)$$
$$T_l(g_2,0) = \{d \in \mathbb{R} \mid g_2'(0)d = 0d \leq 0\} = \mathbb{R} \supsetneq T(X,0)$$

zeigt, ist die Darstellung in der Tat relevant für die Form des linearisierten Kegels.

Das obige Beispiel ist generisch, in dem Sinne, das der Kegel $T_l(g,h,x)$ zumindest nicht beliebig klein werden kann. Dies zeigt uns

Lemma 3.1.8. *Es gilt für alle* $x \in X$

$$T(X,x) \subset T_l(g,h,x).$$

Beweis. Sei $d \in T(X,\overline{x})$. Es gibt also Folgen $x^k \in X$ und $\eta_k \in \mathbb{R}_>0$ mit

$$x^k \to \overline{x}, \quad d^k := \eta_k(x^k - x) \to d.$$

Dann folgt

$$\begin{aligned}
0 &\geq \eta_k g_i(x^k) \\
&= \eta_k(g_i(x^k) - g_i(x)) \\
&= \langle \nabla g_i(x), d^k \rangle + \eta_k o(\|x^k - x\|) \qquad \forall i \in \mathscr{A}(x), \\
0 &= \eta_k(h_i(x^k) - h_i(x)) \\
&= \langle \nabla h_i(x), d^k \rangle + \eta_k o(\|x^k - x\|) \qquad \forall i \in \mathscr{G}.
\end{aligned}$$

Durch Grenzübergang $k \to \infty$ folgt somit

$$\begin{aligned}
0 &\geq \langle \nabla g_i(x), d \rangle \qquad \forall i \in \mathscr{A}(x), \\
0 &= \langle \nabla h_i(x), d \rangle \qquad \forall i \in \mathscr{G}.
\end{aligned}$$

Folglich ist $d \in T_l(g,h,x)$. □

Man beachte, dass i.A. keine Gleichheit zwischen den beiden Kegeln gilt, und damit eine analoge Ungleichung wie in Theorem 3.1.3 nicht notwendig für Richtungen aus der größeren Menge T_l gilt.

Wir benötigen also Bedingungen, welche uns garantieren, dass wir immer noch eine zu Theorem 3.1.3 analoge Aussage für den linearisierten Tangentialkegel erhalten. Solche Bedingungen nennen wir *Regularitätsbedingung* oder auch *Constraint Qualification*. Eine offenkundige Bedingung dieser Art ist die Folgende

Definition 3.1.9. In einem Punkt $x \in X$ gilt die *Abadie Constraint Qualification* (ACQ), falls

$$T_l(g,h,x) = T(X,x).$$

Hiermit ist offenkundig

Theorem 3.1.10 (Notwendige Bedingungen mit (ACQ)). *Sei $\overline{x}$ eine lokale Lösung von* (NLP) *in der die (ACQ) gilt. Dann gilt*

1. $\overline{x} \in X$,
2. $\langle \nabla f(\overline{x}), d \rangle \geq 0$ *für alle* $d \in T_l(g, h, \overline{x})$.

Beweis. Die Aussage folgt sofort aus der notwendigen Bedingung in Theorem 3.1.3 zusammen mit der (ACQ). □

In der Tat können wir diese Bedingung noch etwas abschwächen. Hierzu benötigen wir eine weitere Definition

Definition 3.1.11. Sei $K \subset \mathbb{R}^n$ ein nicht leerer Kegel. Dann ist

$$K^\circ = \{v \in \mathbb{R}^n \mid \langle v, d \rangle \leq 0 \quad \forall d \in K\}$$

der *Polarkegel* von K.

Wir können nun die folgende Bedingung formulieren

Definition 3.1.12. In einem Punkt $x \in X$ gilt die *Guignard Constraint Qualification* (GCQ), falls

$$T_l(g, h, x)^\circ = T(X, x)^\circ.$$

Lemma 3.1.13. *Die Bedingung (ACQ) impliziert (GCQ).*

Beweis. Offensichtlich. □

Die Bedingung (GCQ) ist aber wie wir im Beispiel gesehen haben echt schwächer als (ACQ) (d.h. GCQ impliziert nicht ACQ). Trotzdem gilt noch

Theorem 3.1.14. *[Notwendige Bedingungen mit (GCQ)] Sei $\overline{x}$ eine lokale Lösung von* (NLP) *in der die (GCQ) gilt. Dann gilt*

1. $\overline{x} \in X$,
2. $\langle \nabla f(\overline{x}), d \rangle \geq 0$ *für alle* $d \in T_l(g, h, \overline{x})$.

Beweis. Wegen der notwendigen Bedingung in Theorem 3.1.3 ist in $\overline{x}$ notwendig

$$-\nabla f(\overline{x}) \in T(X,\overline{x})^\circ = T_l(g,h,\overline{x})^\circ.$$

Dies impliziert die Behauptung. □

Beispiel 3.1.15. In der Tat ist diese fast trivial anmutende Erweiterung von (ACQ) auf (GCQ) von Interesse bei der Behandlung sogenannter Komplementaritätsbedingungen, wenn von zwei Funktionen mindestens eine Null sein soll.

Um dies zu verstehen betrachten wir die Bedingung

$$h(x) = (x_1 - x_2)(x_1 + x_2) = 0$$

Es ist leicht einzusehen, dass gilt $X = \{x \in \mathbb{R}^2 \,|\, h(x) = 0\} = T(X,0)$. Es ist aber auch

$$\nabla h(x) = \begin{pmatrix} (x_1 - x_2) + (x_1 + x_2) \\ (x_1 - x_2) + (x_1 + x_2) \end{pmatrix}, \quad \text{und somit} \quad \nabla h(0) = 0$$

und somit $T_l(h,0) = \mathbb{R} \supsetneq T(X,0)$ und folglich gilt die (ACQ) nicht.

Andererseits ist

$$T(X,0)^\circ = \{0\} = T_l(h,0)^\circ$$

und damit gilt die (GCQ).

Wir werden uns später noch mit weiteren Constraint Qualifications befassen, die (GCQ) implizieren jedoch leichter zu überprüfen sind.

Wir möchten die notwendigen Bedingungen nun noch in eine Standardform überführen. Hierzu benötigen wir jedoch noch etwas Vorbereitung, damit wir die definierenden Bedingungen in $T_l(g,h,x)$ besser verstehen.

3.1.1.1 Das Lemma von Farkas

Theorem 3.1.16 (Trennungssatz von Hahn-Banach)**.** *Sei $M \subset \mathbb{R}^n$ nicht-leer, abgeschlossen und konvex sowie $c \in M^c = \mathbb{R}^n \setminus M$. Dann gibt es $v \in \mathbb{R}^n$ und $\alpha \in \mathbb{R}$ mit*

$$\begin{aligned} \langle v, c\rangle &> \alpha, \\ \langle v, x\rangle &\leq \alpha \quad \forall x \in M. \end{aligned}$$

Zusatz: *Ist M ein Kegel, so kann $\alpha = 0$ gewählt werden.*

Beweis. Wegen $M \neq \emptyset$ gibt es ein $x^0 \in M$ und

$$\widehat{M} = \{x \in M \mid \|x - c\| \leq \|x^0 - c\|\}$$

ist kompakt und $x^0 \in \widehat{M} \neq \emptyset$. Da $\|\cdot\|$ stetig ist gibt es daher ein $\overline{x} \in \widehat{M} \subset M$ mit

$$\|\overline{x} - c\| = \inf_{x \in \widehat{M}} \|x - c\| = \inf_{x \in M} \|x - c\|.$$

Wir setzen nun $v = c - \overline{x}$ und $\alpha = \langle v, \overline{x} \rangle$. Da $c \notin M$ ist gilt

$$0 < \|v\|^2 = \langle v, c - \overline{x} \rangle = \langle v, c \rangle - \alpha.$$

Dies zeigt die erste Ungleichung. Sei nun $x \in M$ beliebig. Da M konvex ist, ist für jedes $\lambda \in [0,1]$ auch $x_\lambda = \lambda x + (1-\lambda)\overline{x} \in M$. Wir definieren nun die stetig differenzierbare Funktion $\phi : [0,1] \to \mathbb{R}$ durch

$$\phi(\lambda) = \frac{1}{2}\|c - x_\lambda\|^2.$$

Es ist offenbar

$$\phi(0) = \frac{1}{2}\|c - \overline{x}\|^2 = \min_{0 \leq \lambda \leq 1} \phi(\lambda)$$

und somit

$$0 \leq \phi'(0) = \langle c - \overline{x}, \overline{x} - x \rangle = \langle v, \overline{x} - x \rangle = \alpha - \langle v, x \rangle.$$

Dies zeigt die zweite Ungleichung.

Für den Zusatz sei nun M ein Kegel. Wegen der Abgeschlossenheit ist dann $0 \in M$. Somit ist nach dem bereits gezeigten

$$0 = \langle v, 0 \rangle \leq \alpha.$$

Für beliebiges $x \in M$ ist nun für $\lambda > 0$ auch $\lambda x \in M$ und daher

$$\lambda \langle v, x \rangle \leq \alpha.$$

Durch den Grenzübergang $\lambda \to \infty$ folgt

$$\langle v, x \rangle \leq 0.$$

Wir können daher $\alpha = 0$ wählen ohne die Aussage des Theorems zu stören. □

Lemma 3.1.17. *Seien $A \in \mathbb{R}^{n \times m}$ und $B \in \mathbb{R}^{n \times p}$. Dann ist*

$$K = \left\{ x \in \mathbb{R}^n \,\middle|\, x = Au + Bv, u \in \mathbb{R}^m, u \geq 0, v \in \mathbb{R}^p \right\}$$

ein abgeschlossener konvexer Kegel.

Beweis. Wir können K äquivalent als

$$K = \left\{ x \in \mathbb{R}^n \,\middle|\, x = \begin{pmatrix} A & B & (-B) \end{pmatrix} u, u \in \mathbb{R}^{m+2p}, u \geq 0 \right\}$$

darstellen. Es genügt daher die Aussage für die Menge

$$K(A) = \left\{ x \in \mathbb{R}^n \,\middle|\, x = Au, u \in \mathbb{R}^m, u \geq 0 \right\}$$

zu zeigen. Es ist offenkundig K ein konvexer Kegel.

Die Abgeschlossenheit zeigen wir per Induktion über m. Ist $m = 0$ so ist nichts zu zeigen. Ist $m = 1$, so ist $A = a_1 \in \mathbb{R}^n$ und $K(A) = \{\lambda a_1 \mid \lambda \geq 0\}$ ist offenbar abgeschlossen.

Für den Induktionsschritt sei nun angenommen, dass alle Kegel $K(\widetilde{A})$ der obigen Form für $\widetilde{A} \in \mathbb{R}^{n \times p}$ mit $p < m$ abgeschlossen seien. Sei nun $A \in \mathbb{R}^{n \times m}$ beliebig und $x^k \in K(A)$ mit $x^k \to x$. Per Definition gibt es also $u^k \in [0, \infty)^m$ mit

$$x^k = Au^k.$$

Wir machen nun eine Fallunterscheidung:

a) Sind die Spalten von A linear unabhängig, so ist $A^T A$ regulär und

$$u^k = (A^T A)^{-1} A^T x^k.$$

Folglich gibt es

$$u = \lim_{k \to \infty} (A^T A)^{-1} A^T x^k = \lim_{k \to \infty} u^k$$

und es folgt $u \geq 0$. Nach Definition ist dann auch

$$x = \lim_{k \to \infty} x^k = \lim_{k \to \infty} Au^k = Au \in K(A).$$

b) Andernfalls sind die Spalten von A linear abhängig, es gibt also $w \in \mathbb{R}^m \setminus \{0\}$, so dass

$$Aw = 0.$$

Zu jedem k gibt es dann eine Zahl α_k mit minimalem Betrag, so dass

$$\tilde{u}^k = u^k + \alpha_k w$$

mindestens eine Komponente i_k mit $\tilde{u}^k_{i_k} = 0$ besitzt, es ist dann auch $\tilde{u}^k \geq 0$ (Sonst gäbe es ein α_k mit kleinerem Betrag). Ferner gibt es einen Wert $i \in \{1, \ldots, m\}$, der unendlich oft durch i_k angenommen wird. Sei nun $\mathcal{K} \subset \mathbb{N}$ diejenige (unendliche) Menge mit $i_k = i$ für alle $k \in \mathcal{K}$. Per Definition ist auch $x^k \to x$ für $\mathcal{K} \ni k \to \infty$ und es ist $\tilde{u}^k_{i_k} = \tilde{u}^k_i = 0$ für alle $k \in \mathcal{K}$. Wir können daher die i-te Spalte aus A entfernen, und erhalten so

$$\begin{aligned} \overline{A} &= (a_1; \ \ldots \ a_{i-1} \ a_{i+1} \ \ldots \ a_m) \in \mathbb{R}^{n \times m-1}, \\ \bar{u}^k &= \tilde{u}^k_{\{1,\ldots,m\} \setminus i} = (\tilde{u}^k_1 \ \ldots \ \tilde{u}^k_{i-1} \ \tilde{u}^k_{i+1} \ \ldots \ \tilde{u}^k_m)^T \geq 0. \end{aligned}$$

Damit ist für alle $k \in \mathscr{K}$

$$x^k \in K(\bar{A}) \subset K(A),$$

denn

$$x^k = Au^k = Au^k + \alpha_k Aw = A\tilde{u}^k = \bar{A}\bar{u}^k.$$

Nach Induktionsannahme ist $K(\bar{A})$ abgeschlossen, es ist also

$$x = \lim_{\mathscr{K} \ni k \to \infty} x^k \in K(\bar{A}) \subset K$$

und folglich ist K abgeschlossen.

□

Lemma 3.1.18 (Farkas). *Seien $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{n\times p}$ und $c \in \mathbb{R}^n$ gegeben. Dann ist äquivalent*

1. *Für alle $d \in \mathbb{R}^n$ mit $A^T d \leq 0$ und $B^T d = 0$ gilt $\langle c, d\rangle \leq 0$.*
2. *Es gibt ein $u \in \mathbb{R}^m$, $u \geq 0$ und $v \in \mathbb{R}^p$ mit $c = Au + Bv$.*

Die gilt ebenfalls sinngemäß, wenn $m = 0$ oder $p = 0$ wobei dann $A = 0$ oder $B = 0$.

Beweis. „2 ⇒ 1": Sei also $c = Au + Bv$ mit $u \geq 0$. Dann ist für beliebiges $d \in \mathbb{R}^n$ mit $A^T d \leq 0$ und $B^T d = 0$ auch

$$\langle c, d\rangle = \langle Au, d\rangle + \langle Bv, d\rangle = \langle u, A^T d\rangle + \langle v, B^T d\rangle = \langle u, A^T d\rangle \leq 0.$$

„1 ⇒ 2" Aufgrund des vorherigen Lemmas 3.1.17 ist der Kegel

$$K = \{x \in \mathbb{R}^n \mid x = Au + Bv, u \geq 0\}$$

abgeschlossen und konvex. Angenommen 2. gelte nicht, dann ist $c \notin K$ und nach dem Trennungssatz von Hahn-Banach 3.1.16 gibt es ein $w \in \mathbb{R}^n$ mit den Eigenschaften

$$\begin{aligned} \langle w, c\rangle &> 0, \\ \langle w, x\rangle &\leq 0 \qquad \forall x \in K. \end{aligned}$$

Weiter ist offenbar $a_i \in K$ für jede Spalte a_i von A und $\pm b_i \in K$ für jede Spalte b_i von B und wir erhalten

$$A^T w \leq 0, \quad B^T w = 0.$$

da aber $\langle c, w\rangle > 0$ ist 1. nicht erfüllt. □

3.1.1.2 Karush-Kuhn-Tucker-Bedingungen

Wir haben nun unsere Vorbereitungen abgeschlossen und können die Standardform der notwendigen Optimalitätsbedingungen angeben.

Theorem 3.1.19. *[Notwendige Bedingungen erster Ordnung, Karush-Kuhn-Tucker (KKT) Bedingungen] Sei $\overline{x} \in \mathbb{R}^n$ eine lokale Lösung von (NLP), in der die (GCQ) gilt. Dann gelten die folgenden* KKT-Bedingungen*:*

Es gibt Lagrange-Multiplikatoren $\overline{\lambda} \in \mathbb{R}^m$ und $\overline{\mu} \in \mathbb{R}^p$, so dass

1. *(Multiplikatorregel/Stationarität der Lagrangefunktion)*

$$\nabla f(\overline{x}) + \nabla g(\overline{x})\overline{\lambda} + \nabla h(\overline{x})\overline{\mu} = 0,$$

2. $h(\overline{x}) = 0$,
3. *(Komplementaritätsbedingung)*

$$\overline{\lambda} \geq 0, \qquad g(\overline{x}) \leq 0, \qquad \langle \overline{\lambda}, g(\overline{x}) \rangle = 0.$$

Beweis. Sei nun $\overline{x} \in \mathbb{R}^n$ eine lokale Lösung in der (GCQ) erfüllt ist. Dann ist nach den notwendigen Bedingungen mit (GCQ) 3.1.14 notwendig $\overline{x} \in X$ und damit $h(\overline{x}) = 0$ und $g(\overline{x}) \leq 0$ und es ist

$$-\langle \nabla f(\overline{x}), d \rangle \leq 0$$

für alle

$$d \in T_l(g,h,\overline{x}) = \{d \in \mathbb{R}^n \mid \langle \nabla g_i(\overline{x}), d \rangle \leq 0, \quad i \in \mathscr{A}(\overline{x}) \quad \text{und} \quad \nabla h(\overline{x})^T d = 0\}.$$

Nach dem Lemma von Farkas, mit $c = -\nabla f(\overline{x})$, $A = \nabla g_{\mathscr{A}(\overline{x})}(\overline{x})$ und $B = \nabla h(\overline{x})$, gibt es $u \geq 0$ und $v \in \mathbb{R}^p$ mit

$$-\nabla f(\overline{x}) = c = Au + Bv.$$

Wir setzen nun $\overline{\lambda} \in \mathbb{R}^m$ und $\overline{\mu} \in \mathbb{R}^p$ als

$$\overline{\lambda}_{\mathscr{A}(\overline{x})} = u, \quad \overline{\lambda}_{\mathscr{I}(\overline{x})} = 0, \quad \overline{\mu} = v.$$

Damit ergibt sich die Multiplikatorregel sowie die Komplementaritätsbedingung. □

Definition 3.1.20. Erfüllt ein Tripel $(\overline{x}, \overline{\lambda}, \overline{\mu}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ die KKT-Bedingungen, so nennen wir $\overline{x}$ einen *KKT-Punkt* von (NLP) und $(\overline{x}, \overline{\lambda}, \overline{\mu})$ ein *KKT-Tripel* von (NLP).

Bemerkung 3.1.21.

1. Die Komplementaritätsbedingung besagt gerade, dass für jede Komponente $i \in \mathcal{U}$ gilt $\overline{\lambda}_i = 0$ oder $g_i(\overline{x}) = 0$. Insbesondere ist $\overline{\lambda}_i = 0$ für alle $i \in \mathcal{I}(\overline{x})$.
2. Ist stets einer der beiden Werte $\overline{\lambda}_i$ oder $g_i(\overline{x})$ von Null verschieden, so spricht man von *strikter Komplementarität*. Im Falle strikter Komplementarität ist also $\overline{\lambda}_i > 0$ für alle $i \in \mathcal{A}(\overline{x})$.
3. Die erste der KKT-Bedingungen kann mithilfe der *Lagrange-Funktion* $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$ gegeben durch
$$\mathcal{L}(x, \lambda, \mu) = f(x) + \langle \lambda, g(x) \rangle + \langle \mu, h(x) \rangle$$
kompakter in der Form
$$\nabla_x \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) = 0$$
geschrieben werden.
4. Die KKT-Bedingungen besagen gerade, dass der negative Gradient von f in einem KKT-Punkt im Kegel der Gradienten aktiver Nebenbedingungen liegen muss.

3.1.1.3 Constraint Qualifications

Wir wollen uns nun Bedingungen zuwenden, welche die (GCQ) (Definition 3.1.12) implizieren, und damit natürlich auch die Existenz von Lagrange-Multiplikatoren zu den notwendigen Optimalitätsbedingungen.

Der einfachste Fall ist eine spezielle Struktur der Nebenbedingungen:

Theorem 3.1.22. *Ist $x \in X$ und sind*
$$g_i \text{ konkav}, \ i \in \mathcal{A}(x) \quad \text{sowie} \quad h \text{ affin-linear},$$
so gilt die (ACQ) in x. Hierbei nennen wir g_i konkav, falls $-g_i$ konvex ist.

Beweis. Sei also $x \in X$, so dass die Voraussetzungen des Theorems erfüllt sind. Ist dann $d \in T_l(g, h, x)$ beliebig, so definieren wir
$$t_k = 1/k.$$
Ist dann $k \geq k_0$ hinreichend groß, so ist
$$g_i(x + t_k d) \leq 0, \quad \forall i \in \mathcal{I}(x).$$

Da ferner für $i \in \mathscr{A}(x)$ die Funktion g_i konkav ist, folgt (vgl. Theorem 1.3.4 zur Charakterisierung konvexer Funktionen)

$$\begin{aligned} g(x_i + t_k d) &\leq g_i(x) + t_k \langle \nabla g_i(x), d \rangle \\ &= t_k \langle \nabla g_i(x), d \rangle \leq 0 \end{aligned}$$

nach Definition von $T_l(g,h,x)$. Ebenso ergibt die Linearität von h

$$h(x + t_k d) = h(x) + t_k \langle \nabla h(x), d \rangle = t_k \langle \nabla h(x), d \rangle = 0.$$

Somit ist für $k \geq k_0$ der Punkt $x^k = x + t_k d \in X$ und durch die Setzung $\eta_k = \frac{1}{t_k}$ folgt

$$d = \lim_{k\to\infty} \eta_k (x^k - x) = \lim_{k\to\infty} d,$$

und somit $d \in T(X,x)$, bzw. $T_l(g,h,x) \subset T(X,x)$. Wegen Lemma 3.1.8 folgt somit $T(X,x) = T_l(g,h,x)$ und damit die (ACQ) in x. □

Bemerkung 3.1.23. Diese Bedingung ist natürlich insbesondere für affin-lineares g erfüllt, weshalb man für *lineare Programme* keine Constraint Qualifications voraussetzen muss. Was sich natürlich auch aus der Tatsache ergibt, dass dann $T_l(g,h,x) = T(X,x)$ natürlicher Weise gilt.

Wir wollen uns nun wieder allgemeineren Constraint Qualifications zuwenden. Eine der häufigsten ist die folgende

Definition 3.1.24. Ein Punkt $x \in X$ genügt der *Mangasarian-Fromovitz Constraint Qualification* (MFCQ), falls

1. $\nabla h(x)$ hat vollen Spaltenrang (insbesondere ist $p \leq n$),
2. es gibt ein $d \in \mathbb{R}^n$ mit

$$\langle \nabla g_i(x), d \rangle < 0, \quad \forall i \in \mathscr{A}(x), \qquad \nabla h(x)^T d = 0.$$

Im Falle $m = 0$ oder $\mathscr{A}(x) = \emptyset$ entfällt 2.), im Fall $p = 0$ entfallen 1.) und die Bedingung $\nabla h(x)^T d = 0$ in 2.).

Bemerkung 3.1.25. Die (MFCQ) kann auch leicht abgewandelt formuliert werden. Im Falle einer affin-linearen Funktion h kann auf die Bedingung 1.) in der (MFCQ) verzichtet werden.

Lemma 3.1.26. *Sei $x \in X$ gegeben. Hat $\nabla h(x)$ vollen Spaltenrang, oder ist h affin-linear, so gibt es ein $\epsilon > 0$ und eine stetig differenzierbare Funktion $\phi : \mathbb{R}^n \supset B_\epsilon(0) \to \mathbb{R}^n$ mit folgenden Eigenschaften*

$$\begin{aligned} h(x + w + \phi(w)) &= 0 && \forall\, w \in B_\epsilon(0) \cap \mathrm{Kern}(\nabla \mathrm{h}(\mathrm{x})^{\mathrm{T}}), \\ \phi(0) &= 0, \\ \nabla\phi(0)^T z &= 0 && \forall\, z \in \mathrm{Kern}(\nabla h(x)^T). \end{aligned}$$

Theorem 3.1.27. *Für ein $x \in X$ impliziert die (MFCQ) auch (ACQ). Das Selbe gilt auch für die modifizierte Bedingung aus Bemerkung 3.1.25.*

Bemerkung 3.1.28. Der Folgende technische Beweis hat folgende Struktur. Für gegebenes $s \in T_l(g,h,x)$ ist i.A. $x + \epsilon s \notin X$ für beliebiges $\epsilon > 0$. Daher müssen wir eine kleine Korrektur td vornehmen, mit der wir sicherstellen können, dass $x + \epsilon(s + td) = x + w$ zumindest die Ungleichungsbedingungen sicherstellt. Nun benötigen wir eine weitere Korrektur $\phi(w)$, damit $x + w + \phi(w) \in X$ liegt.

Beweis. Aufgrund des vorhergegangenen Lemmas 3.1.26 existiert zu $x \in X$ eine Funktion $\phi : B_\epsilon(0) \to \mathbb{R}^n$. mit den dort gegebenen Eigenschaften.

Ist dann $w \in B_\epsilon(0) \cap \mathrm{Kern}(\nabla \mathrm{h}(\mathrm{x})^{\mathrm{T}})$, so folgt durch Taylorentwicklung

$$\phi(w) = \phi(0) + \nabla\phi(0)^T w + o(\|w\|) = o(\|w\|). \tag{3.1}$$

Wir können daher o.B.d.A. $0 < \epsilon \leq \frac{2}{3}$ so klein wählen, dass

$$\|\phi(w)\| \leq \frac{\|w\|}{2} \leq \frac{1}{3} \quad \forall w \in B_\epsilon(0) \cap \mathrm{Kern}(\nabla \mathrm{h}(\mathrm{x})^{\mathrm{T}}).$$

Sei nun $s \in T_l(g,h,x)$ beliebig. Wir wollen zeigen, dass $s \in T(X,x)$ liegt. Da $0 \in T(X,x)$ ($x^k = x$ in der Definition) können wir $s \neq 0$ annehmen. Da weiter $T(X,x)$ ein Kegel ist, können wir zudem voraussetzen, dass $\|s\| \leq \frac{\epsilon}{2}$. Für die Richtung d aus der (MFCQ) können wir ebenfalls annehmen, dass $\|d\| \leq \frac{\epsilon}{2}$.

Nach Definition der Differenzierbarkeit und obiger Beobachtung (3.1) ist

$$\begin{aligned} |g_i(x+v) - g_i(x) - \langle \nabla g_i(x), v\rangle| &= o(\|v\|) \\ \|\phi(w)\| &= o(\|w\|) \qquad \forall w \in B_\epsilon(0) \cap \mathrm{Kern}(\nabla \mathrm{h}(\mathrm{x})^{\mathrm{T}}). \end{aligned}$$

Es gibt also eine Nullfolge $t_k > 0$ mit

$$\frac{|g_i(x+v) - g_i(x) - \langle \nabla g_i(x), v\rangle|}{\|v\|} := r_i(v) \leq t_k^2 \quad \forall\, \|v\| \leq \frac{1}{k},\ \forall i \in \mathscr{A}(x).$$
$$\frac{\|\phi(w)\|}{\|w\|} \leq t_k^2 \qquad \forall\, w \in B_{\min(1/k,\epsilon)}(0) \cap \mathrm{Kern}(\nabla \mathrm{h(x)}^{\mathrm{T}}).$$

O.B.d.A. können wir $t_k \in (0,1)$ annehmen, indem wir ggf. k hinreichend groß wählen. Wir setzen nun $w^k = (s + t_k d)/k$ und erhalten

$$\|w^k\| \leq \frac{\|s\|}{k} + \frac{t_k}{k}\|d\| \leq \frac{\epsilon}{2k} + \frac{\epsilon}{2k} = \frac{\epsilon}{k} \leq \frac{1}{k}.$$

Da $s, d \in \mathrm{Kern}(\nabla h(x)^T)$ ist auch $w^k \in \mathrm{Kern}(\nabla h(x)^T)$ und damit können wir $s^k := w^k + \phi(w^k)$ definieren, und erhalten

$$\|s^k\| \leq \|w^k\| + \|\phi(w^k)\| \leq \frac{3}{2}\|w^k\| \leq \frac{3\epsilon}{2k} \leq \frac{1}{k}.$$

Damit ergibt sich für $v = s^k$ und $w = w^k$

$$t_k^2 \geq r_i(s^k) \quad \forall i \in \mathscr{A}(x),$$
$$t_k^2 \geq \frac{\|\phi(w^k)\|}{\|w^k\|}.$$

Damit ist für beliebiges $i \in \mathscr{A}(x)$ (und k hinreichend groß)

$$\begin{aligned} g_i(x+s^k) &\leq g_i(x) + \langle \nabla g_i(x), s^k\rangle + r_i(s^k)\|s^k\| \\ &= \underbrace{\frac{\langle \nabla g_i(x), s\rangle}{k}}_{\leq 0} + \frac{t_k\langle \nabla g_i(x), d\rangle}{k} + \langle \nabla g_i(x), \phi(w^k)\rangle + r_i(s^k)\|s^k\| \\ &\leq \frac{t_k}{k}\langle \nabla g_i(x), d\rangle + t_k^2\|\nabla g_i(x)\|\|w^k\| + t_k^2\|s^k\| \\ &\leq \frac{t_k}{k}\langle \nabla g_i(x), d\rangle + \frac{t_k^2}{k}(\|\nabla g_i(x)\| + 1). \end{aligned}$$

Indem wir k ggf. weiter vergrößern ist o.B.d.A für alle k die Bedingung $t_k(1 + \|\nabla g_i(x)\|) \leq -\langle \nabla g_i(x), d\rangle$ erfüllt. Somit ist dann

$$g(x + s^k) \leq 0.$$

Für die inaktiven Ungleichungsbedingungen kann Zulässigkeit, durch weitere Vergrößerung von k, ebenfalls erreicht werden. Nach Konstruktion von s^k ist außerdem

$$h(x + s^k) = h(x + w^k + \phi(w^k)) = 0.$$

Damit ist $x^k := x + s^k \in X$ für hinreichend großes k.

Nun ist $x^k \to x$ für $k \to \infty$, da $s^k \to 0$. Mit der Wahl $\eta_k = k$ folgt nun

$$\|\eta_k(x^k - x) - s\| = \|t_k d + k\phi(w^k)\| \leq \frac{t_k}{2} + ko(\|w^k\|) = \frac{t_k}{2} + k\|w^k\| \frac{o(\|w^k\|)}{\|w^k\|} \to 0.$$

Somit ist $s \in T(X,x)$ und folglich $T_l(g,h,x) \subset T(X,x)$. Mit Lemma 3.1.8 folgt $T_l(g,h,x) = T(X,x)$ also die (ACQ) in x. □

Damit kommen wir nun zum noch fehlenden Detail, dem Beweis von Lemma 3.1.26

von Lemma 3.1.26. Ist h affin-linear, so erfüllt $\phi \equiv 0$ offenbar die gewünschten Eigenschaften.

Ansonsten hat die Matrix $\nabla h(x)^T \in \mathbb{R}^{p \times n}$ vollen Zeilenrang p. Sei nun $A \in \mathbb{R}^{n \times (n-p)}$ so gegeben, dass Ihre Spalten eine Basis von $\text{Kern}(\nabla h(x)^T)$ bilden. Die Abbildung

$$\Psi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n; \quad \Psi(w,y) = \begin{pmatrix} h(x+w+y) \\ A^T y \end{pmatrix}$$

ist stetig differenzierbar (in einer Umgebung von 0). Es ist $\Psi(0,0) = 0$ und

$$\nabla_y \Psi(0,0)^T = \begin{pmatrix} \nabla h(x)^T \\ A^T \end{pmatrix} \in \mathbb{R}^{n \times n}$$

ist invertierbar (man beachte, das die Zeilen in $\nabla h(x)^T$ und A^T nach Konstruktion linear unabhängig sind).

Nach dem Satz über implizite Funktionen gibt es also ein $\epsilon > 0$ und ein $\phi : \mathbb{R}^n \supset B_\epsilon(0) \to \mathbb{R}^n$ mit $\phi(0) = 0$ und

$$\Psi(w, \phi(w)) = 0$$

für alle $w \in B_\epsilon(0)$. Ferner können wir für $z \in \text{Kern}(\nabla h(x)^T)$ nachrechnen

$$\begin{aligned} \nabla\phi(0)^T z &= -(\nabla_y \Psi(0,0))^{-T} \nabla_w \Psi(0,0)^T z \\ &= -(\nabla_y \Psi(0,0))^{-T} \begin{pmatrix} \nabla h(x)^T \\ 0 \end{pmatrix} z \\ &= 0. \end{aligned}$$

Dies zeigt die Behauptung. □

Neben der (MFCQ) gibt es natürlich weitere Constraint Qualifications. Zur Vorbereitung geben wir hier ein kurzes Lemma wieder

Lemma 3.1.29 (Alternativlemma). *Sind $A \in \mathbb{R}^{n \times m}$ und $B \in \mathbb{R}^{n \times p}$. Dann ist äquivalent:*

1. *Es gibt $d \in \mathbb{R}^n$ mit $A^T d < 0$ und $B^T d = 0$*

2. *Es gibt keine Vektoren* $(u, v) \in \mathbb{R}^m \times \mathbb{R}^p$ *mit* $0 \neq u \geq 0$ *und*

$$Au + Bv = 0.$$

Beweis. Übung! □

Damit können wir eine zur (MFCQ) äquivalente Bedingung angeben

Definition 3.1.30. Ein Punkt $x \in X$ genügt der *Positive Linear Independence Constraint Qualification* (PLICQ), falls

1. $\nabla h(x)$ hat vollen Spaltenrang (insbesondere ist $p \leq n$),
2. es gibt keine Vektoren $u \in \mathbb{R}^m$, $v \in \mathbb{R}^p$ mit

$$\nabla g(x)u + \nabla h(x)v = 0, \quad 0 \neq u_{\mathcal{A}(x)} \geq 0, u_{\mathcal{I}(x)} = 0.$$

Im Falle $m = 0$ oder $\mathcal{A}(x) = \emptyset$ entfällt 2.), im Fall $p = 0$ entfallen 1.) und der Summand $\nabla h(x)v$ in 2.).

In der Tat ist diese äquivalent zur (MFCQ)

Theorem 3.1.31. *Die Bedingungen (PLICQ) und (MFCQ) sind äquivalent. Dies gilt auch für die verallgemeinerten Fassungen in denen jeweils h affin-linear ist.*

Beweis. Die Bedingung 1. in (PLICQ) und (MFCQ) sind identisch. Die Bedingung 2. in (PLICQ) und (MFCQ) entsprechen sich wegen des Alternativlemmas 3.1.29. □

Definition 3.1.32. Ein Punkt $x \in X$ genügt der *Linear Independence Constraint Qualification* (LICQ), falls die Spalten der Matrix

$$(\nabla g_{\mathcal{A}(x)}(x) \quad \nabla h(x))$$

linear unabhängig sind. In diesem Fall nennt man den Punkt x *regulär*.

Lemma 3.1.33. *(LICQ) impliziert (PLICQ).*

Beweis. Offensichtlich wegen der Definition linearer Unabhängigkeit. □

Wir haben damit die folgende Kette an Constraint Qualifications gezeigt.

$$\text{LICQ} \Longrightarrow \text{PLICQ} \Longleftrightarrow \text{MFCQ} \Longrightarrow \text{ACQ} \Longrightarrow \text{GCQ}.$$

3.1.1.4 KKT-Bedingungen für konvexe Probleme

Im Falle eines konvexen NLPs, d.h. f sowie alle g_i sind konvex und h ist affin-linear, sind die KKT-Bedingungen nicht nur notwendig, sondern auch hinreichend.

Theorem 3.1.34. *Sei (NLP) konvex. Dann ist jede lokale Lösung $\overline{x} \in X$ von (NLP) auch eine globale Lösung. Gilt in $\overline{x} \in X$ die (GCQ), so sind die KKT-Bedingungen erfüllt.*

Ist umgekehrt $\overline{x}$ ein KKT-Punkt, so ist $\overline{x}$ eine globale Lösung von (NLP).

Beweis. Wir wissen bereits aus dem Theorem zu Minima konvexer Funktionen 1.3.8, dass jedes lokale Minimum auch global ist. Gilt dann die (GCQ), so sind in einem Minimum auch die KKT-Bedingungen erfüllt, vgl. Theorem 3.1.19.

Umgekehrt sei nun $\overline{x}$ ein KKT-Punkt, $x \in X$ beliebig und $d = x - \overline{x}$. Für beliebiges $i \in \mathcal{A}(\overline{x})$ folgt wegen der Komplementaritätsbedingung sowie der Charakterisierung konvexer Funktionen (Theorem 1.3.4)

$$\overline{\lambda}_i \langle \nabla g_i(\overline{x}), d \rangle \leq \overline{\lambda}_i (g_i(x) - g_i(\overline{x})) = \overline{\lambda}_i g_i(x) \leq 0.$$

Ebenso ist $\nabla h(\overline{x})^T d = h(x) - h(\overline{x}) = 0$. Aus der Konvexität von f (Theorem 1.3.4) und den KKT-Bedingungen folgt

$$\begin{aligned} f(x) - f(\overline{x}) &\geq \langle \nabla f(\overline{x}), d \rangle \\ &= -\langle \nabla g(\overline{x})\overline{\lambda}, d \rangle - \langle \nabla h(\overline{x})\overline{\mu}, d \rangle \\ &= -\sum_{i \in \mathcal{U}} \overline{\lambda}_i \langle \nabla g_i(\overline{x}), d \rangle - \langle \overline{\mu}, \nabla h(\overline{x})^T d \rangle \\ &= -\sum_{i \in \mathcal{A}(\overline{x})} \overline{\lambda}_i \langle \nabla g_i(\overline{x}), d \rangle \\ &\geq 0. \end{aligned}$$

□

3.1.2 Optimalitätsbedingungen zweiter Ordnung

3.1.2.1 Hinreichende Bedingungen zweiter Ordnung

Um nun hinreichende Bedingungen angeben zu können möchten wir, wie im Falle der unbeschränkten Optimierung auf Informationen der zweiten Ableitungen zurückgreifen. Es

ist hierbei natürlich möglich die positiv Definitheit der Hesse-Matrix in $\overline{x}$ zu fordern. Dies wird i.A. allerdings nicht erfüllt sein! Wir möchten diese Bedingung daher abschwächen, und positiv Definitheit nur auf einer möglichst kleinen Menge voraussetzen. Hierzu ist es Hilfreich den Kegel der kritischen Richtungen zu betrachten:

Definition 3.1.35. Zu gegebenen Punkten $x \in X$ und $\lambda \in \mathbb{R}^m_{\geq 0}$ definieren wir den Kegel

$$T_+(g,h,x,\lambda) = \left\{ d \in \mathbb{R}^n \,\middle|\, \nabla h(x)^T d = 0, \quad \langle \nabla g_i(x), d\rangle \begin{cases} = 0 & \text{falls } i \in \mathcal{A}(x), \lambda_i > 0, \\ \leq 0 & \text{falls } i \in \mathcal{A}(x), \lambda_i = 0. \end{cases} \right\}$$

Bemerkung 3.1.36. Der Kegel $T_+(g,h,x,\lambda)$ liegt zwischen dem linearisierten Tangentialkegel und dem Tangentialraum der aktiven Nebenbedingungen

$$T_a(g,h,x) = \left\{ d \in \mathbb{R}^n \,\middle|\, \nabla h(x)^T d = 0, \quad \langle \nabla g_i(x), d\rangle = 0, i \in \mathcal{A}(x) \right\},$$

d.h.

$$T_a(g,h,x) \subset T_+(g,h,x,\lambda) \subset T_l(g,h,x).$$

Im Falle strikter Komplementarität gilt

$$T_a(g,h,x) = T_+(g,h,x,\lambda).$$

Wir erhalten damit das folgende Resultat.

Theorem 3.1.37 (Hinreichende Bedingungen 2. Ordnung)**.** *Seien $f \in C^2(\mathbb{R}^n;\mathbb{R})$, $g \in C^2(\mathbb{R}^n;\mathbb{R}^m)$, $h \in C^2(\mathbb{R}^n;\mathbb{R}^p)$ und $(\overline{x},\overline{\lambda},\overline{\mu}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ sei ein KKT-Tripel des* (NLP)*, d.h. Sie erfüllen die KKT-Bedingung 3.1.19. Gilt zudem*

$$\langle d, \nabla^2_{xx} \mathcal{L}(\overline{x},\overline{\lambda},\overline{\mu}) d\rangle > 0 \quad \forall\, d \in T_+(g,h,\overline{x},\overline{\lambda}) \setminus \{0\},$$

so ist $\overline{x}$ eine isolierte lokale Lösung von (NLP).

Bemerkung 3.1.38. Man beachte hierbei, dass wir die positiv Definitheit für die zweite Ableitung der Lagrangefunktion

$$\mathcal{L}(x,\lambda,\mu) = f(x) + \langle \lambda, g(x)\rangle + \langle \mu, h(x)\rangle$$

und nicht etwa für die Hesse-Matrix von f fordern!. Dies ist vernünftig, da die zweite Ableitung von f in einem Minimum von (NLP) nicht einmal positiv semi-definit sein

muss. vgl. Übung!

Beweis. Angenommen das KKT-Tripel $(\overline{x}, \overline{\lambda}, \overline{\mu})$ ist keine isolierte lokale Lösung von (NLP). Dann gäbe es eine Folge $x^k \in X$ mit den Eigenschaften

$$x^k \neq \overline{x}, \qquad x^k \to \overline{x}, \qquad f(x^k) \leq f(\overline{x}).$$

Wir können nun $d^k = x^k - \overline{x}$ setzen. O.B.d.A ist dann (ggf. durch Wahl einer Teilfolge)

$$y^k := \frac{d^k}{\|d^k\|} \to y.$$

Damit folgt

$$\begin{aligned}
0 &\geq \frac{f(x^k) - f(\overline{x})}{\|d^k\|} = \langle \nabla f(\overline{x}), y^k \rangle + \frac{o(\|d^k\|)}{\|d^k\|} \to \langle \nabla f(\overline{x}), y \rangle, \\
0 &\geq \frac{g_i(x^k) - g_i(\overline{x})}{\|d^k\|} \to \langle \nabla g_i(\overline{x}), y \rangle, \quad \forall i \in \mathcal{A}(\overline{x}), \\
0 &= \frac{h_i(x^k) - h_i(\overline{x})}{\|d^k\|} \to \langle \nabla h_i(\overline{x}), y \rangle, \quad \forall i \in \mathcal{G}.
\end{aligned}$$

Damit ist $y \in T_l(g, h, \overline{x})$ und die KKT-Bedingungen liefern uns

$$0 = \langle \nabla_x \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\mu}), y \rangle = \underbrace{\langle \nabla f(x), y \rangle}_{\leq 0} + \sum_{i \in \mathcal{A}(\overline{x})} \underbrace{\overline{\lambda}_i \langle \nabla g_i(\overline{x}), y \rangle}_{\leq 0} + \sum_{i \in \mathcal{G}} \underbrace{\overline{\mu}_i \langle \nabla h_i(\overline{x}), y \rangle}_{=0}.$$

Damit ist für $\overline{\lambda}_i > 0$ notwendig $\langle \nabla g_i(\overline{x}), y \rangle = 0$ sonst wäre die rechte Seite negativ. Wir erhalten also $y \in T_+(g, h, \overline{x}, \overline{\lambda})$.

Es folgt damit aus den KKT-Bedingungen ($h_i(\overline{x}) = 0$, $\lambda^T g(\overline{x}) = 0$) und der Zulässigkeit von x^k

$$\mathcal{L}(x^k, \overline{\lambda}, \overline{\mu}) = f(x^k) + \sum_{i \in \mathcal{A}(\overline{x})} \overline{\lambda}_i g_i(x^k) \leq f(x^k) \leq f(\overline{x}) = \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\mu}).$$

Da $\nabla_x \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) = 0$ folgt durch Taylorentwicklung

$$\begin{aligned}
0 &\geq \frac{\mathcal{L}(x^k, \overline{\lambda}, \overline{\mu}) - \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\mu})}{\|d^k\|^2} \\
&= \frac{1}{2} \langle y^k, \nabla^2_{xx} \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) y^k \rangle + \frac{o(\|d^k\|^2)}{\|d^k\|^2} \\
&\to \frac{1}{2} \langle y, \nabla^2_{xx} \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) y \rangle.
\end{aligned}$$

Dies steht im Widerspruch zur Annahme

$$\frac{1}{2} \langle y, \nabla^2_{xx} \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) y \rangle > 0.$$

□

Bemerkung 3.1.39. Man beachte, dass die Verwendung des Tangentialraumes $T_a(g,h,x)$ im Theorem 3.1.37, im Falle nicht strikter Komplementarität, i.A. nicht ausreichend ist, um die Existenz einer lokalen Lösung von (NLP) zu garantieren, da das im Beweis konstruierte y i.A. nicht in $T_a(g,h,\bar{x})$ liegt, da wir für Indizes i mit $\overline{\lambda}_i = 0$ nicht die Identität $\langle \nabla g_i(\overline{x}), y\rangle = 0$ zeigen können.

In der Tat ist die positiv Definitheit von $\nabla^2_{xx}\mathscr{L}(\overline{x},\overline{\lambda},\overline{\mu})$ auf $T_a(g,h,\overline{x})$ auch nicht hinreichend wie uns das Beispiel

$$\begin{aligned}\min_{x\in\mathbb{R}^2} f(x) &= x_1^2 - x_2^2\\ \text{u.d.N } g(x) &= x_1^2 - x_2 \le 0\end{aligned}$$

zeigt. Im zulässigen Punkt $\overline{x} = 0$ ist

$$\nabla f(0) = 0, \qquad \nabla^2 f(0) = \begin{pmatrix} 2 & 0\\ 0 & -2\end{pmatrix}, \qquad \nabla g(0) = \begin{pmatrix} 0\\ -1\end{pmatrix}$$

und somit ist $(\overline{x},\overline{\lambda}) = (0,0)$ ein KKT-Punkt und

$$T_a(g,0) = \{d\in\mathbb{R}^2 \mid \langle d, \nabla g(0)\rangle = 0\} = \mathbb{R}\times\{0\}.$$

Damit ist jedes $d\in T_a(g,0)$ von der Form $d = \begin{pmatrix} d_1\\ 0\end{pmatrix}$ und es folgt

$$\begin{aligned}\langle d, \nabla^2_{xx}\mathscr{L}(\overline{x},\overline{\lambda})d\rangle &= \langle d, \nabla^2 f(0)d\rangle + \overline{\lambda}\langle d, \nabla^2 g(0)d\rangle\\ &= 2d_1^2 - 2d_2^2\\ &= 2d_1^2 > 0\end{aligned}$$

für alle $d\in T_a(g,0)\setminus\{0\}$. Allerdings ist $\overline{x} = 0$ kein isoliertes lokales Minimum des Problems, denn die Punkte $x^\tau = \begin{pmatrix} 0\\ \tau\end{pmatrix}$ mit $\tau > 0$ erfüllen

$$g(x^\tau) = -\tau < 0, \quad f(x^\tau) = -\tau^2 < 0 = f(\overline{x}), \quad x^\tau \to \overline{x} \text{ für } \tau\to 0.$$

3.1.2.2 Notwendige Bedingungen zweiter Ordnung

Für die Betrachtung notwendiger Bedingungen müssen wir analog zu den Bedingungen erster Ordnung sicherstellen, dass wir für eine gegebene Richtung d eine, nun zweimal differenzierbare, Kurve γ mit $\gamma'(0) = d$ finden können welche in der zulässigen Menge X bleibt. Die aus dem konkaven Fall bekannte Idee durch $x+td$ einen neuen zulässigen Punkt zu erhalten kann i.A. nicht funktionieren. Daher versuchen wir dies durch eine (kleine) Korrektur zu retten, dies gelingt durch den Ansatz einer (differenzierbaren) Kurve $\gamma(t) = x + td + R(t)$

mit einer noch zu definierenden Abbildung $R\colon \mathbb{R} \to \mathbb{R}^n$. Dies geht nicht immer, daher benötigen wir eine weitere Annahme:

Definition 3.1.40. In $(x,\lambda,\mu) \in X \times \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$ ist die *Constraint Qualification 2. Ordnung* (CQ2) erfüllt, falls es zu jedem $d \in T_+(g,h,x,\lambda)$ ein $\epsilon > 0$ und eine zweimal stetig differenzierbare Kurve $\gamma\colon (-\epsilon,\epsilon) \to \mathbb{R}^n$ mit den Eigenschaften

$$\begin{aligned} \gamma(0) &= x, \\ \gamma'(0) &= d \\ g_i(\gamma(t)) &= 0, \quad \forall t \in (-\epsilon,\epsilon), i \in \{j \in \mathscr{A}(x) \mid \langle \nabla g_j(x), d\rangle = 0\}, \\ h(\gamma(t)) &= 0, \quad \forall t \in (-\epsilon,\epsilon) \end{aligned}$$

gibt.

Lemma 3.1.41. *Seien $g \in C^2(\mathbb{R}^n;\mathbb{R}^m)$, $h \in C^2(\mathbb{R}^n;\mathbb{R}^p)$ und $(x,\lambda,\mu) \in X \times \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$. Gilt dann die (LICQ) in x (d.h. x ist regulär), so gilt (CQ2).*

Beweis. Sei $d \in T_+(g,h,x,\lambda)$ beliebig. Wir setzen $\mathscr{A}_0 = \{i \in \mathscr{A}(x) \mid \langle \nabla g_i(x), d\rangle = 0\}$ und $l = |A_0| + p$ und definieren damit $G\colon \mathbb{R}^n \to \mathbb{R}^l$ durch

$$G(y) = \begin{pmatrix} g_{\mathscr{A}_0}(y) \\ h(y) \end{pmatrix}.$$

Damit machen wir den Ansatz $\gamma(t) = x + td + \nabla G(x)z(t)$ mit einer noch zu findenden Funktion $z\colon \mathbb{R} \to \mathbb{R}^l$. Es ist damit aus der (CQ2) ersichtlich, dass $G(\gamma(t)) = 0$ gelten muss, diese implizite Funktion müssen wir nun bestimmen.

Hierzu setzen wir $\psi\colon \mathbb{R} \times \mathbb{R}^l \to \mathbb{R}^l$ durch

$$\psi(t,z) = G(x + td + \nabla G(x)z).$$

Damit ist $\psi \in C^2(\mathbb{R} \times \mathbb{R}^l;\mathbb{R}^l)$, und es ist

$$\psi(0,0) = 0, \qquad (\nabla_z \psi(0,0))^T = \nabla G(x)^T \nabla G(x).$$

Da die (LICQ) erfüllt ist können wir $(\nabla_z \psi(0,0))^T$ invertieren, und nach dem Satz über implizite Funktionen gibt es $\epsilon > 0$ und eine C^2-Funktion $z\colon (-\epsilon,\epsilon) \to \mathbb{R}^l$ mit den Eigenschaften

$$z(0) = 0, \quad \psi(t,z(t)) = 0 \ \forall\, t \in (-\epsilon,\epsilon), \quad \nabla z(0)^T = -(\nabla_z \psi(0,z(0)))^{-T} \nabla_t \psi(0,z(0))^T.$$

Da $d \in T_+(g,h,x,\lambda)$ liegt, und somit $\nabla G(x)^T d = 0$, folgt daher

$$\nabla z(0)^T = -(\nabla G(x)^T \nabla G(x))^{-1} \nabla G(x)^T d = 0.$$

Damit hat die C^2-Kurve $\gamma(t) = x + td + \nabla G(x)z(t)$ die gewünschten Eigenschaften, denn

$$\gamma(0) = x, \quad G(\gamma(t)) = \psi(t, z(t)) = 0, \ \forall\, t \in B_\epsilon(0) \quad \gamma'(0) = d + \nabla G(x)\nabla z(0)^T = d.$$

□

Wir können nun die notwendigen Bedingungen 2. Ordnung angeben

Theorem 3.1.42 (Notwendige Optimalitätsbedingungen 2. Ordnung)**.** *Seien f, g und h zweimal stetig differenzierbar. Sei $\overline{x}$ eine lokale Lösung von (NLP) in der die (GCQ) erfüllt ist. Dann gibt es Lagrange-Multiplikatoren $\overline{\lambda} \in \mathbb{R}^m$ und $\overline{\mu} \in \mathbb{R}^p$, so dass $(\overline{x}, \overline{\lambda}, \overline{\mu})$ ein KKT-Tripel ist.*
Ist zudem in $(\overline{x}, \overline{\lambda}, \overline{\mu})$ die (CQ2) erfüllt, so gilt

$$\langle d, \nabla^2_{xx} \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) d\rangle \geq 0 \quad \forall\, d \in T_+(g, h, \overline{x}, \overline{\lambda}).$$

Beweis. Wegen der KKT-Bedingungen in Theorem 3.1.19 ist nur noch der Zusatz für die (CQ2) zu zeigen.

Sei daher $d \in T_+(g, h, \overline{x}, \overline{\lambda})$ gegeben und $\gamma\colon \mathbb{R} \supset B_\epsilon(0) \to \mathbb{R}^n$ eine durch (CQ2) gegebene Kurve mit

$$\mathcal{A}_0 = \{i \in \mathcal{A}(\overline{x}) \mid \langle \nabla g_i(\overline{x}), d\rangle = 0\}.$$

Wir können die Ableitungen berechnen, und erhalten

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t)) &= \langle \nabla f(\gamma(t)), \gamma'(t)\rangle, \\
\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\gamma(t)) &= \langle \gamma'(t), \nabla^2 f(\gamma(t))\gamma'(t)\rangle + \langle \nabla f(\gamma(t)), \gamma''(t)\rangle.
\end{aligned}$$

Analoge Formeln gelten natürlich auch für die Komponenten von g und h.

Wir können nun $\epsilon > 0$ ggf. verkleinern, so dass für alle $t \in [0, \epsilon]$ gilt

$$\begin{aligned}
g_{\mathcal{I}(\overline{x})}(\gamma(t)) &< 0, \\
g_{\mathcal{A}_0}(\gamma(t)) &= 0, \\
h(\gamma(t)) &= 0, \\
g_i(\gamma(t)) &= g_i(\gamma(0)) + t\langle \nabla g_i(\gamma(0)), \gamma'(0)\rangle + o(t) \\
&= g_i(\overline{x}) + t\langle \nabla g_i(\overline{x}), d\rangle + o(t) \\
&= t \underbrace{\langle \nabla g_i(\overline{x}), d\rangle}_{<0} + o(t) \qquad\qquad \forall i \in \mathcal{A}(\overline{x}) \setminus \mathcal{A}_0 \\
&\leq 0.
\end{aligned}$$

Folglich ist $\gamma([0, \epsilon]) \subset X$.

Für $i \notin \mathscr{A}_0$ ist entweder $i \in \mathscr{I}(\overline{x})$, und damit $\overline{\lambda}_i = 0$, oder $i \in \mathscr{A}(\overline{x})$ mit $\langle \nabla g_i(\overline{x}), d\rangle < 0$, und es ist ebenfalls $\overline{\lambda}_i = 0$ nach Definition von $T_+(g,h,\overline{x},\overline{\lambda})$. Es ist also

$$\overline{\lambda}_i = 0 \qquad \forall i \notin \mathscr{A}_0.$$

Nach Definition von $\mathscr{A}_0$ und der Multiplikatorregel der KKT-Bedingungen folgt

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\bigg|_{t=0} &= \langle \nabla f(\overline{x}), d\rangle \\ &= \langle \nabla f(\overline{x}), d\rangle + \langle \overline{\lambda}_{\mathscr{A}_0}, \nabla g_{\mathscr{A}_0}(\overline{x})^T d\rangle + \langle \overline{\mu}, \nabla h(\overline{x})^T d\rangle \\ &= \langle \nabla_x \mathscr{L}(\overline{x},\overline{\lambda},\overline{\mu}), d\rangle \\ &= 0.\end{aligned}$$

Da $\overline{x}$ ein lokales Minimum von f ist, ist $t = 0$ ein lokales und, nach evtl. Verkleinern von ϵ, auch globales Minimum von

$$[0,\epsilon] \ni t \mapsto f(\gamma(t)).$$

Da $\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\Big|_{t=0} = 0$ folgt

$$\begin{aligned}0 &\le 2\frac{f(\gamma(t)) - f(\gamma(0))}{t^2} \\ &= 0 + \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\gamma(t))\bigg|_{t=0} + \frac{o(t^2)}{t^2} \\ &\to \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\gamma(t))\bigg|_{t=0} \\ &= \langle d, \nabla f(\overline{x}) d\rangle + \langle \nabla f(\overline{x}), v\rangle\end{aligned}$$

mit $d = \gamma'(0)$ und $v := \gamma''(0)$. Analog folgt aus $g_i(\gamma(t)) = h_j(\gamma(t)) = \langle \nabla g_i(\overline{x}), d\rangle = \langle \nabla h_j(\overline{x}), d\rangle = 0$ für $t \in [0,\epsilon]$ und $i \in \mathscr{A}_0$, $j \in \mathscr{G}$ dass

$$\begin{aligned}\langle d, \nabla^2 g_i(\overline{x}) d\rangle + \langle \nabla g_i(\overline{x}), v\rangle &= 0 \qquad && \forall\, i \in \mathscr{A}_0, \\ \langle d, \nabla^2 h_i(\overline{x}) d\rangle + \langle \nabla h_i(\overline{x}), v\rangle &= 0 && \forall\, i \in \mathscr{G}.\end{aligned}$$

Damit erhalten wir in Kombination unter Verwendung der KKT-Bedingungen und $\overline{\lambda}_i = 0$ für $i \notin \mathscr{A}_0$

$$\begin{aligned}\langle d, \nabla^2_{xx} \mathscr{L}(\overline{x},\overline{\lambda},\overline{\mu}) d\rangle &= \langle d, \nabla^2_{xx} \mathscr{L}(\overline{x},\overline{\lambda},\overline{\mu}) d\rangle + \langle \nabla_x \mathscr{L}(\overline{x},\overline{\lambda},\overline{\mu}), v\rangle \\ &= \langle d, \nabla^2 f(\overline{x}) d\rangle + \langle \nabla f(\overline{x}), v\rangle \\ &\quad + \sum_{i \in .\mathscr{A}_0} \overline{\lambda}_i \Big(\langle d, \nabla^2 g_i(\overline{x}) d\rangle + \langle \nabla g_i(\overline{x}), v\rangle \Big) \\ &\quad + \sum_{i \in \mathscr{G}} \overline{\mu}_i \Big(\langle d, \nabla^2 h_i(\overline{x}) d\rangle + \langle \nabla h_i(\overline{x}), v\rangle \Big) \\ &= \langle d, \nabla^2 f(\overline{x}) d\rangle + \langle \nabla f(\overline{x}), v\rangle \ge 0.\end{aligned}$$

□

3.2 Lagrange-Dualität

Wir erinnern uns daran, dass wir zum (NLP) eine Lagrange-Funktion

$$\mathscr{L}(x,\lambda,\mu) = f(x) + \langle \lambda, g(x)\rangle + \langle \mu, h(x)\rangle$$

definiert haben. Neben Ihrem Nutzen für die Bestimmung notwendiger und hinreichender Optimalitätsbedingungen sehen wir einen weiteren Nutzen dieser Funktion aus der folgenden Beobachtung

$$p(x) := \sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} \mathscr{L}(x,\lambda,\mu) = \begin{cases} f(x) & x \in X, \\ +\infty & \text{sonst.} \end{cases}$$

Damit ist unser restringiertes Optimierungsproblem (NLP) offenbar äquivalent zur Lösung des unrestringierten Problem

$$\min_{x\in\mathbb{R}^n} p(x) = \inf_{x\in\mathbb{R}^n} \sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} \mathscr{L}(x,\lambda,\mu). \tag{$\mathbb{P}$}$$

Damit definieren wir

Definition 3.2.1. Das Problem

$$\sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} \inf_{x\in\mathbb{R}^n} \mathscr{L}(x,\lambda,\mu) \tag{$\mathbb{D}$}$$

ist das *(Lagrange) duale Problem* zum *primalen Problem* ($\mathbb{P}$). Dabei nennen wir

$$p(x) = \sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} \mathscr{L}(x,\lambda,\mu)$$

die *primale Zielfunktion* und

$$d(\lambda,\mu) = \inf_{x\in\mathbb{R}^n} \mathscr{L}(x,\lambda,\mu)$$

die *duale Zielfunktion*.

Bemerkung 3.2.2.

1. Man beachte, das hierbei $d(\lambda,\mu)$ auch den Wert $-\infty$ annehmen kann.
2. Für gegebenes x ist $(\lambda,\mu) \mapsto \mathscr{L}(x,\lambda,\mu)$ affin-linear, und damit ist $d(\lambda,\mu)$ als Infimum von linearen Funktionen konkav. Das Maximierungsproblem

$$\sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} d(\lambda,\mu)$$

ist damit zu einem konvexen Problem

$$\inf_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} -d(\lambda,\mu)$$

äquivalent.

Der wesentliche Nutzen des dualen Problems liegt darin, dass es uns stets eine untere Schranke an den gesuchten Minimalwert von f gibt.

Theorem 3.2.3 (Schwacher Dualitätssatz)**.** *Ist $(\tilde{\lambda},\tilde{\mu})$ zulässig für das duale Problem ($\mathbb{D}$), d.h. $\tilde{\lambda}\geq 0$, so gilt*

$$p(\tilde{x})\geq d(\tilde{\lambda},\tilde{\mu}).$$

Ist zudem $\tilde{x}$ zulässig für das primale Problem ($\mathbb{P}$), d.h. $\tilde{x}\in X$, so ist

$$p(\tilde{x}) = f(\tilde{x})$$

Beweis. Nach Definition ist

$$d(\tilde{\lambda},\tilde{\mu}) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x,\tilde{\lambda},\tilde{\mu}) \leq \mathcal{L}(\tilde{x},\tilde{\lambda},\tilde{\mu}) \leq \sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} \mathcal{L}(\tilde{x},\lambda,\mu) = p(\tilde{x}).$$

Den Zusatz für $\tilde{x}\in X$ haben wir bereits gesehen. □

In der Tat folgt aus der Rechnung sofort,

$$\sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} \inf_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda,\mu) = \sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} d(\lambda,\mu) \leq \sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} \mathcal{L}(\tilde{x},\lambda,\mu)$$

auch falls $\tilde{x}\notin X$. Damit ist $\tilde{x}$ beliebig, und wir erhalten als sofortige Konsequenz

$$\sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} \inf_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda,\mu) \leq \inf_{x\in\mathbb{R}^n} \sup_{(\lambda,\mu)\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p} \mathcal{L}(x,\lambda,\mu). \tag{3.2}$$

In der Tat ist in vielen Fällen Gleichheit der beiden Werte zu sehen, und zwar genau dann wenn $(\overline{x},\overline{\lambda},\overline{\mu})$ ein Sattelpunkt der Lagrange-Funktion ist.

Definition 3.2.4. Ein Punkt $(\overline{x},\overline{\lambda},\overline{\mu})\in\mathbb{R}^n\times\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p$ heißt *Sattelpunkt der Lagrange-Funktion*, falls

$$\mathcal{L}(\overline{x},\lambda,\mu)\leq\mathcal{L}(\overline{x},\overline{\lambda},\overline{\mu})\leq\mathcal{L}(x,\overline{\lambda},\overline{\mu}) \qquad \forall(x,\lambda,\mu)\in\mathbb{R}^n\times\mathbb{R}^m_{\geq 0}\times\mathbb{R}^p.$$

Theorem 3.2.5. *Es ist äquivalent:*

1. $(\overline{x}, \overline{\lambda}, \overline{\mu})$ *ist ein Sattelpunkt der Lagrange-Funktion.*
2. $\overline{x}$ *ist ein globales Minimum von* (ℙ), $(\overline{\lambda}, \overline{\mu})$ *ist ein globales Maximum von* (𝔻) *und*

$$f(\overline{x}) = d(\overline{\lambda}, \overline{\mu}).$$

Beweis. „1 ⇒ 2" Aus (3.2) folgt

$$\begin{aligned}
\mathscr{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) &= \inf_{x \in \mathbb{R}^n} \mathscr{L}(x, \overline{\lambda}, \overline{\mu}) \\
&\leq \sup_{(\lambda,\mu) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p} \inf_{x \in \mathbb{R}^n} \mathscr{L}(x, \lambda, \mu) \\
&\leq \inf_{x \in \mathbb{R}^n} \sup_{(\lambda,\mu) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p} \mathscr{L}(x, \lambda, \mu) \\
&\leq \sup_{(\lambda,\mu) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p} \mathscr{L}(\overline{x}, \lambda, \mu) \\
&\leq \mathscr{L}(\overline{x}, \overline{\lambda}, \overline{\mu}).
\end{aligned}$$

Dies zeigt

$$\infty > \mathscr{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) = \inf_{x \in \mathbb{R}^n} \mathscr{L}(x, \overline{\lambda}, \overline{\mu}) = d(\overline{\lambda}, \overline{\mu}) = \sup_{(\lambda,\mu) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p} \mathscr{L}(\overline{x}, \lambda, \mu) = p(\overline{x}).$$

Also ist $\overline{x}$ zulässig, $p(\overline{x}) = f(\overline{x})$ und wegen $d(\overline{\lambda}, \overline{\mu}) = p(\overline{x})$ folgt die globale Optimalität von $(\overline{\lambda}, \overline{\mu})$ und $\overline{x}$ aus dem schwachen Dualitätssatz 3.2.3.

„2 ⇒ 1" Umgekehrt ist

$$\mathscr{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) \leq \sup_{(\lambda,\mu) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p} \mathscr{L}(\overline{x}, \lambda, \mu) = p(\overline{x}) = f(\overline{x}) = d(\overline{\lambda}, \overline{\mu}) = \inf_{x \in \mathbb{R}^n} \mathscr{L}(x, \overline{\lambda}, \overline{\mu}) \leq \mathscr{L}(\overline{x}, \overline{\lambda}, \overline{\mu})$$

und somit ist $(\overline{x}, \overline{\lambda}, \overline{\mu})$ ein Sattelpunkt von $\mathscr{L}$. □

Bemerkung 3.2.6. Sind f, g_i konvex und differenzierbar und h_i affin-linear, so ist

$$x \mapsto \mathscr{L}(x, \lambda, \mu)$$

konvex für beliebige $\lambda \in \mathbb{R}^m_{\geq 0}$ und $\mu \in \mathbb{R}^p$. Wird dann $\min_{x \in \mathbb{R}^n} \mathscr{L}(x, \lambda, \mu)$ für alle $(\lambda, \mu) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$ angenommen, dann sind die Minima gerade diejenigen $x \in \mathbb{R}^n$ für die

gilt $\nabla_x \mathscr{L}(x,\lambda,\mu) = 0$. Damit ist dann (𝔻) äquivalent zu

$$\sup_{(\lambda,\mu)\in\mathbb{R}^m\times\mathbb{R}^p} \quad \mathscr{L}(x,\lambda,\mu) \tag{3.3}$$

$$\text{u.d.N} \begin{cases} \lambda \geq 0 \\ \nabla_x \mathscr{L}(x,\lambda,\mu) = 0 \end{cases}. \tag{3.4}$$

Dies ist das sog. *Wolfe-Dualproblem*.

Bemerkung 3.2.7. Im Allgemeinen ist $\sup d(\lambda,\mu) < p(x)$, d.h. die Ungleichung in Theorem 3.2.3 ist i.A. strikt. Für das Vorliegen eines Sattelpunktes sind daher weitere Bedingungen erforderlich! Dies gilt auch wenn beide Werte endliche sind!

Um sich dies klar zumachen betrachten wir das konvexe! Problem

$$\min_{(x,y)\in\mathbb{R}^2} e^{-x}$$
$$\text{u.d.N. } \frac{x^2}{e^y} \leq 0.$$

Die zulässige Menge ist hier gerade $X = 0 \times \mathbb{R}$ und somit sehen wir sofort, dass

$$\inf_{(x,y)\in\mathbb{R}^2} p(x,y) = e^{-0} = 1$$

für jeden zulässigen Punkt angenommen wird. Für das duale Problem (𝔻) ist die Lagrange-Funktion gerade $\mathscr{L}(x,y,\lambda) = e^{-x} + \lambda\frac{x^2}{e^y}$ und es ist

$$d(\lambda) = \inf_{(x,y)\in\mathbb{R}^2} \mathscr{L}(x,y,\lambda) = \begin{cases} 0 & \lambda \geq 0 \\ -\infty & \lambda < 0 \end{cases}.$$

Damit ist offenkundig

$$\sup_{\lambda\in\mathbb{R}} d(\lambda) = 0 < 1 = \inf_{(x,y)\in\mathbb{R}^2} p(x,y).$$

3.3 Strafterm-Verfahren

Um nun erste Verfahren für restringierte Optimierungsprobleme zu erhalten, ist es hilfreich sich zu überlegen, dass das (NLP) äquivalent durch das (formal unrestringierte Problem)

$$\min_{x\in\mathbb{R}^n} f(x) + I_X(x)$$

gegeben ist, dabei ist

$$I_X(x) = \begin{cases} 0 & x \in X, \\ \infty & \text{sonst} \end{cases}$$

die *Indikatorfunktion* von X (nicht zu verwechseln mit der charakteristischen Funktion χ_X). Um den unangenehmen, da nicht reellwertigen, Term I_X im Kostenfunktional zu beseitigen, ersetzt man diesen durch einen Strafterm (Penalty-Term). D.h. für eine Straffunktion $\pi : \mathbb{R}^n \to \mathbb{R}$ mit

$$\pi(x) \begin{cases} = 0 & x \in X, \\ > 0 & \text{sonst} \end{cases}$$

und einem Penalty-Parameter $\alpha > 0$ betrachtet man das Teilproblem

$$\min_{x \in \mathbb{R}^n} f(x) + \alpha \pi(x).$$

Da diese i.A. für große α nur schlecht lösbar sind, wählt man nun eine (monoton wachsende) Folge α_k und findet sukzessive Lösungen x^k der jeweiligen Probleme, wobei man mit der jeweils vorherigen Lösung x^{k-1} einen Startwert hat.

3.3.1 Quadratische Strafterm-Verfahren

Die gebräuchlichste Variante mit differenzierbarem π ist das *quadratische Penalty-Verfahren* mit der Zielfunktion

$$\begin{aligned} P_\alpha(x) &= f(x) + \frac{\alpha}{2} \sum_{i \in \mathcal{U}} \max(0, g_i(x))^2 + \frac{\alpha}{2} \sum_{i \in \mathcal{G}} h_i(x)^2 \\ &= f(x) + \frac{\alpha}{2} \|g(x)^+\|^2 + \frac{\alpha}{2} \|h(x)\|^2 \end{aligned}$$

wobei $v^+ = (\max(0, v_i))_i$ ist. Für den Gradienten ist

$$\nabla P_\alpha(x) = \nabla f(x) + \alpha \sum_{i \in \mathcal{U}} g_i(x)^+ \nabla g_i(x) + \alpha \sum_{i \in \mathcal{G}} h_i(x) \nabla h_i(x).$$

Damit ist $P_\alpha \in C^1(\mathbb{R}^n; \mathbb{R})$ und

$$P_\alpha(x) = f(x), \qquad \nabla P_\alpha(x) = \nabla f(x) \qquad \forall\, x \in X.$$

Insbesondere können lokale Minimierer $\overline{x}$ von P_α nur dann in X liegen, wenn diese bereits stationäre Punkte von f auf $\mathbb{R}^n$ sind.

Wir können hiermit nun ein Lösungsverfahren angeben

Algorithmus 3.3.1 (Idealisiertes Penalty-Verfahren)**.**

- Wähle $\alpha_0 \geq 0$.

for k = 0, 1, …, **do**

- Berechne eine globale Lösung x^k des Penalty-Problems $\min_{x\in\mathbb{R}^n} P_{\alpha_k}(x)$. Hierbei verwendet man im Fall $k > 0$ meist x^{k-1} als Startwert.
- Stopp, falls $x^k \in X$ (oder zumindest nahe genug).
- Wähle $\alpha_{k+1} > \alpha_k$.

end for

Bemerkung 3.3.2. Man beachte, dass die Berechnung von x^k kein endlicher Schritt ist, da alle unsere Algorithmen bisher

- nur stationäre Punkte suchen,
- i.A. nicht nach endlich vielen Iterationen abbrechen.

In der Realität wird man daher diesen Schritt nur inexakt ausführen! Trotzdem ist es hilfreich sich zuerst diese idealisierte Situation anzusehen.

Bemerkung 3.3.3. Achtung, es ist i.A. nicht klar, das P_α ein globales Minimum besitzt. Dies ist auch dann der Fall, wenn das eigentliche Problem (NLP) ein eindeutiges globales Minimum besitzt. Hierfür betrachte man das Beispiel:

$$\min_{x\le 0} -e^x$$

mit dem Minimum $x = 0$. Für das zugehörige Penalty-Problem ist aber

$$P_\alpha(x) = -e^x + \alpha(x^+)^2 \to -\infty \qquad (x \to \infty).$$

Theorem 3.3.4 (Globale Konvergenz des Penalty-Verfahrens)**.** *Seien f, g, und h stetig und der zulässige Bereich X sei nicht leer. Die Folge $\alpha_k \in (0,\infty)$ sei streng monoton wachsend mit $\alpha_k \to \infty$. Angenommen das Penalty-Verfahren 3.3.1 erzeugt eine unendliche Folge (die Lösbarkeit der Teilprobleme wird also angenommen!), so gilt:*

1. *$P_{\alpha_k}(x^k)$ ist monoton wachsend.*
2. *$\|g(x^k)^+\|^2 + \|h(x^k)\|^2$ ist monoton fallend.*
3. *$f(x^k)$ ist monoton wachsend.*
4. *$\lim_{k\to\infty} g(x^k)^+ = 0$ und $\lim_{k\to\infty} h(x^k) = 0$.*
5. *Jeder Häufungspunkt der Folge x^k ist eine globale Lösung des (NLP).*

Beweis. Wir setzen $\pi(x) = \frac{1}{2}\Big(\|g(x^k)^+\|^2 + \|h(x^k)\|^2\Big)$.

1. Aus der globalen Optimalität von x^k und $\alpha_k < \alpha_{k+1}$ folgt

$$\begin{aligned} P_{\alpha_k}(x^k) &\leq P_{\alpha_k}(x^{k+1}) \\ &= f(x^{k+1}) + \alpha_k \pi(x^{k+1}) \\ &\leq f(x^{k+1}) + \alpha_{k+1} \pi(x^{k+1}) \\ &= P_{\alpha_{k+1}}(x^{k+1}). \end{aligned}$$

2. Durch Addition der Ungleichungen $P_{\alpha_k}(x^k) \leq P_{\alpha_k}(x^{k+1})$ und $P_{\alpha_{k+1}}(x^{k+1}) \leq P_{\alpha_{k+1}}(x^k)$ erhalten wir

$$\alpha_k \pi(x^k) + \alpha_{k+1} \pi(x^{k+1}) \leq \alpha_k \pi(x^{k+1}) + \alpha_{k+1} \pi(x^k).$$

   Nun ist $\alpha_k < \alpha_{k+1}$ und damit folgern wir aus obiger Ungleichung

$$\pi(x^{k+1}) \leq \pi(x^k).$$

3. Aus 2. und der Definition von x^k folgt sofort

$$0 \leq P_{\alpha_k}(x^{k+1}) - P_{\alpha_k}(x^k) = f(x^{k+1}) - f(x^k) + \alpha_k(\pi(x^{k+1}) - \pi(x^k)) \leq f(x^{k+1}) - f(x^k).$$

4. Wir zeigen $\pi(x^k) \to 0$. Da $X \neq \emptyset$ gibt es $\hat{x} \in X$ und damit ist für alle $k \in \mathbb{N}_0$

$$P_{\alpha_k}(x^k) \leq P_{\alpha_k}(\hat{x}) = f(\hat{x}).$$

   Nach 3. ist f monoton wachsend und damit

$$f(\hat{x}) \geq P_{\alpha_k}(x^k) = f(x^k) + \alpha_k \pi(x^k) \geq f(x^0) + \alpha_k \pi(x^k).$$

   Da $\alpha_k \to \infty$ ist damit notwendig $\pi(x^k) \to 0$.

5. Ist $\overline{x}$ ein Häufungspunkt von x^k, so ist $\overline{x} \in X$ wegen 4. und der Stetigkeit von π. Ist nun $x^k \to \overline{x}$ für $k \in \mathcal{K}$ eine konvergente Teilfolge, so folgt für $k \in \mathcal{K}$ und beliebiges $x \in X$

$$f(\overline{x}) \leftarrow f(x^k) \leq P_{\alpha_k}(x^k) \leq P_{\alpha_k}(x) = f(x)$$

   und folglich ist $\overline{x}$ eine globale Lösung von (NLP).

□

Im Falle des Abbruchs des Verfahrens in $x^k \in X$ sind wir wie schon beobachtet fertig, da wir ein globales Minimum gefunden haben. Andernfalls ist jedes x^k ein stationärer Punkt von P_{α_k}. Wir können dies nutzen um Approximationen an die Lagrange-Multiplikatoren $\overline{\lambda}$ und $\overline{\mu}$ zu gewinnen, da

$$\begin{aligned} 0 &= \nabla P_{\alpha_k}(x^k) \\ &= \nabla f(x^k) + \alpha_k \sum_{i \in \mathcal{U}} g_i(x^k)^+ \nabla g_i(x^k) + \alpha_k \sum_{i \in \mathcal{G}} h_i(x^k) \nabla h_i(x^k) \\ &= \nabla f(x^k) + \nabla g(x^k) \lambda^k + \nabla h(x^k) \mu^k \end{aligned}$$

mit der Setzung

$$\lambda_i^k = \alpha_k \max(0, g_i(x^k)), \qquad \mu_i^k = \alpha_k h_i(x^k). \tag{3.5}$$

Damit ist dann

$$\nabla_x \mathscr{L}(x^k, \lambda^k, \mu^k) = 0$$

und wir können in dieser Bedingung zum Grenzwert übergehen.

Theorem 3.3.5. *[Globale Konvergenz des Penalty-Verfahrens] Seien f, g und h stetig differenzierbar, und $X \neq \emptyset$. Sei $(0, \infty) \ni \alpha_k \to \infty$ streng monoton wachsend. Angenommen das Penalty-Verfahren 3.3.1 erzeugt eine unendliche Folge x^k und (λ^k, μ^k) sind durch (3.5) gegeben. Dann gilt:*

1. *Ist $(x^k, \lambda^k, \mu^k)_{\mathscr{K}}$ eine konvergente Teilfolge mit Limes $(\overline{x}, \overline{\lambda}, \overline{\mu})$, so ist $\overline{x}$ eine globale Lösung des (NLP) und $(\overline{x}, \overline{\lambda}, \overline{\mu})$ ist ein KKT-Tripel von (NLP).*
2. *Ist $\overline{x}$ ein Häufungspunkt von x^k und $(x^k)_{\mathscr{K}}$ eine gegen $\overline{x}$ konvergente Teilfolge. Ist dann in $\overline{x}$ die (LICQ) erfüllt, so konvergiert auch $(x^k, \lambda^k, \mu^k)_{\mathscr{K}}$ gegen ein KKT-Tripel $(\overline{x}, \overline{\lambda}, \overline{\mu})$ des (NLP), und $\overline{x}$ ist eine globale Lösung von (NLP).*

Beweis. 1. Durch Grenzübergang folgt

$$\nabla_x \mathscr{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) = \lim_{\mathscr{K} \ni k \to \infty} \nabla_x \mathscr{L}(x^k, \lambda^k, \mu^k) = 0.$$

Aus dem vorhergehenden Theorem 3.3.4 wissen wir bereits $\overline{x} \in X$, es ist also nur noch die Komplementarität nachzuweisen. Es ist

$$\overline{\lambda} = \lim_{\mathscr{K} \ni k \to \infty} \lambda^k \geq 0$$

und es bleibt nur noch die Komplementarität zu zeigen. Ist also $i \in \mathscr{I}(\overline{x})$, so ist für k hinreichend groß auch $i \in \mathscr{I}(x^k)$ und damit

$$\overline{\lambda}_i = \lim_{\mathscr{K} \ni k \to \infty} \lambda_i^k = \lim_{\mathscr{K} \ni k \to \infty} \alpha_k \max(0, g_i(x^k)) = 0.$$

Damit ist $(\overline{x}, \overline{\lambda}, \overline{\mu})$ ein KKT-Tripel von (NLP).

2. Wir wissen bereits, dass $\overline{x}$ eine globale Lösung ist. Es genügt daher nach 1. die Konvergenz von $(x^k, \lambda^k, \mu^k)_{\mathscr{K}}$ zu zeigen.

   Ist $i \in \mathscr{I}(\overline{x})$, so ist wie in 1. $\lambda_i^k = 0$ für alle hinreichend großen k. Dies zeigt

   $$(\lambda^k_{\mathscr{I}(\overline{x})})_{\mathscr{K}} \to \overline{\lambda}_{\mathscr{I}(\overline{x})} = 0.$$

   Nach (LICQ) hat die Matrix

   $$A = (\nabla g_{\mathscr{A}(\overline{x})}(\overline{x}) \quad \nabla h(\overline{x})) \in \mathbb{R}^{n \times (|\mathscr{A}(\overline{x})| + p)}$$

vollen Spaltenrang und somit ist A^TA invertierbar. Damit ist für $k \in \mathcal{K}$ hinreichend groß und

$$A_k = (\nabla g_{\mathcal{A}(\overline{x})}(x^k) \quad \nabla h(x^k))$$

wegen des Lemmas zu invertierbaren Matrizen 2.5.3 auch $A_k^TA_k$ invertierbar sowie $(A_k^TA_k)^{-1} \to (A^TA)^{-1}$. Wir erhalten damit für $k \in \mathcal{K}$ hinreichend groß

$$0 = A_k^T \nabla_x \mathcal{L}(x^k, \lambda^k, \mu^k) = A_k^T \nabla f(x^k) + A_k^T A_k \begin{pmatrix} \lambda^k_{\mathcal{A}(\overline{x})} \\ \mu^k \end{pmatrix}$$

und somit

$$\begin{pmatrix} \lambda^k_{\mathcal{A}(\overline{x})} \\ \mu^k \end{pmatrix} = -(A_k^TA_k)^{-1}A_k^T \nabla f(x^k) \to -(A^TA)^{-1}A^T \nabla f(\overline{x}) \quad (\mathcal{K} \ni k \to \infty).$$

□

Wir wollen nun kurz einsehen wie sich aus den obigen Diskussionen ein tatsächlich realisierbares Verfahren entwickeln lässt. Hierfür ersetzen wir alle unmöglichen Schritte durch realisierbare Approximationen.

Algorithmus 3.3.6 (Inexaktes Penalty-Verfahren)**.**

- Wähle $\alpha_0 \geq 0$, $\mathrm{TOL}_0 > 0$.

for k = 0, 1, …, **do**

- Bestimme einen approximativ stationären Punkt x^k des Penalty-Problems, so dass $\|\nabla P_{\alpha_k}(x^k)\| \leq \mathrm{TOL}_k$ gilt.
- Stopp, falls x^k gut genug.
- Wähle $\alpha_{k+1} > \alpha_k$.
- Wähle $0 < \mathrm{TOL}_{k+1} \leq \mathrm{TOL}_k$.

end for

Für diesen lässt sich nun folgendes zeigen

Theorem 3.3.7 (Globale Konvergenz des Inexakten Penalty-Verfahrens)**.** *Seien f, g, und h stetig differenzierbar. Gilt $\mathrm{TOL}_k \to 0$ und $\alpha_k \to \infty$, $X \neq \emptyset$ und terminiert Algorithmus 3.3.6 nicht endlich, so gilt für jeden Häufungspunkt $\overline{x}$ der Folge x^k in dem (LICQ) erfüllt ist (hierbei fordern wir formal auch die Unabhängigkeit der Gradienten der Nebenbedingungen $g_i(\overline{x}) > 0$), dass folgende: Ist $(x^k)_K \to \overline{x}$ eine konvergente Teilfolge, so gilt*

$$(x^k, \lambda^k, \mu^k) \to (\overline{x}, \overline{\lambda}, \overline{\mu})$$

und $(\overline{x}, \overline{\lambda}, \overline{\mu})$ ist ein KKT-Punkt von (NLP).

Bemerkung 3.3.8. Die Aussage des Theorems ist bis auf den hier fehlenden Zusatz $\overline{x}$ ist globale Lösung von (NLP) identisch zu der von Theorem 3.3.5-2..

Beweis. Der Beweis ist identisch zu dem von Theorem 3.3.5 mit der Ausnahme, das wir zeigen müssen $\overline{x} \in X$. Hierzu beobachten wir, dass

$$\mathrm{TOL}_k \geq \|\nabla P_{\alpha_k}(x^k)\| = \|\nabla f(x^k) + \alpha_k \nabla \pi(x^k)\|$$

gilt und somit

$$\|\nabla \pi(x^k)\| \leq \frac{1}{\alpha_k}\Big(\mathrm{TOL}_k + \|\nabla f(x^k)\|\Big)$$

erfüllt ist. Da $x^k \to \overline{x}$ folgt $\|\nabla f(x^k)\| \leq c$ und wegen $\alpha_k \to \infty$ folgt

$$\|\nabla \pi(x^k)\| \to 0.$$

Nun gilt aber auch

$$\begin{aligned}\nabla \pi(x^k) &= \sum_{i \in \mathcal{U}} g_i(x^k)^+ \nabla g_i(x^k) + \sum_{i \in \mathcal{G}} h_i(x^k) \nabla h_i(x^k) \\ &\to \sum_{i \in \mathcal{U}} g_i(\overline{x})^+ \nabla g_i(\overline{x}) + \sum_{i \in \mathcal{G}} h_i(\overline{x}) \nabla h_i(\overline{x}) = 0.\end{aligned}$$

Damit haben wir eine Linearkombination der Null aus den Gradienten $\nabla g_i(\overline{x})$ (für $g_i(\overline{x}) \geq 0$) und $\nabla h_i(\overline{x})$. Wegen der angenommenen (LICQ) sind die Koeffizienten $g_i(\overline{x})^+ = h_i(\overline{x}) = 0$ und somit ist $\overline{x} \in X$.

Der Rest des Beweises überträgt sich mutatis mutandis von Theorem 3.3.4. □

Bemerkung 3.3.9. Wir wollen nun noch einsehen, dass in der Tat für wachsendes α_k die Kondition der zu den penalisierten Problemen gehörigen Hesse-Matrizen wie α_k wächst.

Der Einfachheit halber seien nur Gleichungsrestriktionen gegeben sowie f zweimal stetig differenzierbar und h affin-linear, d.h. $h(x) = A^T x + b$ für ein $A \in \mathbb{R}^{n \times p}$ und $b \in \mathbb{R}^p$. Es ist dann, wie bereits gesehen,

$$\begin{aligned}\nabla^2 P_\alpha(x) &= \nabla^2 f(x) + \alpha \nabla h(x) \nabla h(x)^T + \alpha \sum_{i \in \mathcal{G}} h_i(x) \nabla^2 h_i(x) \\ &= \nabla^2 f(x) + \alpha A A^T\end{aligned}$$

Ist dann $v \in R^n$ beliebig mit $A^T v \neq 0$, so ist

$$\langle v, \nabla^2 P_\alpha(x) v \rangle = \langle v, \nabla^2 f(x) v \rangle + \alpha \|A^T v\|^2 = O(\alpha) \qquad (\alpha \to \infty).$$

Ist hingegen $w \in \mathbb{R}^n \setminus \{0\}$ beliebig mit, mit $A^T w = 0$, so ist

$$\langle w, \nabla^2 P_\alpha(x) w \rangle = \langle w, \nabla^2 f(x) w \rangle = O(1) \qquad (\alpha \to \infty).$$

Man beachte, dass es i.A. sowohl ein $v \neq 0$, wie auch $w \neq 0$ mit den obigen Eigenschaften gibt. Denn gibt es kein solches v, so ist $h(x) = b$ für alle x und dann $X = \mathbb{R}^n$ oder $X = \emptyset$. Gibt es kein solches w, so ist $X = \{x_0\}$ oder $X = \emptyset$ wobei x_0 die eindeutige Lösung des Problems

$$A^T x = -b$$

ist, sofern diese existiert. Damit ist

$$\operatorname{cond}(\nabla^2 P_\alpha(x)) = O(\alpha) \quad (\alpha \to \infty).$$

Dies beeinflusst Gradienten-basierte Verfahren i.A. sehr negativ und führt bei Newton-artigen Verfahren zu einer zunehmenden Reduktion des Bereichs lokal schneller Konvergenz.

3.3.2 Exakte Penalty-Verfahren

Aufgrund obiger Ausführungen ist es erstrebenswert Penalty-Verfahren zu verwenden, bei denen man den Penalty-Parameter α nicht gegen unendlich laufen lassen muss.

Definition 3.3.10. Sei $\overline{x} \in \mathbb{R}^n$ eine lokale Lösung des (NLP). Die Penalty-Funktion $P : \mathbb{R}^n \to \mathbb{R}$ ist *exakt* in $\overline{x}$, falls $\overline{x}$ ein lokales Minimum von P ist.

Wir setzen nun die l_1-*Penalty-Funktion*

$$P_\alpha^1(x) = f(x) + \alpha \sum_{i \in \mathscr{U}} g_i(x)^+ + \alpha \sum_{i \in \mathscr{G}} |h_i(x)| = f(x) + \alpha\big(\|g(x)^+\|_1 + \|h(x)\|_1\big).$$

Ist dann das (NLP) konvex, so ist P_α^1 eine exakte Penalty-Funktion, falls α hinreichend groß ist.

Theorem 3.3.11. *Sei $(\overline{x}, \overline{\lambda}, \overline{\mu})$ ein KKT-Tripel des Optimierungsproblems (NLP) mit konvexen Funktionen $f \in C^1(\mathbb{R}^n; \mathbb{R})$ und $g \in C^1(\mathbb{R}^n; \mathbb{R}^m)$ sowie einem affin-linearen $h : \mathbb{R}^n \to \mathbb{R}^p$. Dann ist $\overline{x}$ eine globale Lösung von (NLP) und zudem ist $\overline{x}$ für jedes*

$$\alpha \geq \max\{\overline{\lambda}_1, \ldots, \overline{\lambda}_m, |\overline{\mu}_1|, \ldots, |\overline{\mu}_p|\} = \max\Big(\|\overline{\lambda}\|_\infty, \|\overline{\mu}\|_\infty\Big)$$

auch ein globales Minimum von P_α^1 auf $\mathbb{R}^n$.

Beweis. Die globale Optimalität von $\overline{x}$ folgt aus Theorem 3.1.34 zu KKT-Bedingungen für konvexe Probleme.

Für den Rest erkennen wir, dass $\overline{\lambda} \geq 0$ und somit $\mathscr{L}(\cdot, \overline{\lambda}, \overline{\mu})$ konvex ist. Da $\nabla_x \mathscr{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) = 0$ ist nach Theorem 1.3.8 der Punkt $\overline{x}$ ein globales Minimum von $x \mapsto \mathscr{L}(x, \overline{\lambda}, \overline{\mu})$ auf $\mathbb{R}^n$.

Sei nun α wie in den Voraussetzungen des Theorems gewählt. So folgt, da $\overline{x} \in X$ zulässig ist, für beliebiges $x \in \mathbb{R}^n$

$$\begin{aligned}
P_\alpha^1(\overline{x}) &= f(\overline{x}) + \alpha \|g(\overline{x})^+\|_1 + \alpha \|h(\overline{x})\|_1 \\
&= f(\overline{x}) \\
&= f(\overline{x}) + \langle \overline{\lambda}, g(\overline{x}) \rangle + \langle \overline{\mu}, h(\overline{x}) \rangle \\
&= \mathscr{L}(\overline{x}, \overline{\lambda}, \overline{\mu}) \\
&\leq \mathscr{L}(x, \overline{\lambda}, \overline{\mu}) \\
&= f(x) + \langle \overline{\lambda}, g(x) \rangle + \langle \overline{\mu}, h(x) \rangle \\
&\leq f(x) + \langle \overline{\lambda}, g(x)^+ \rangle + \langle |\overline{\mu}|, |h(x)| \rangle \\
&\leq f(x) + \max_{i \in \mathscr{U}} \overline{\lambda}_i \|g(x)^+\|_1 + \max_{i \in \mathscr{G}} |\overline{\mu}_i| \|h(x)\|_1 \\
&\leq P_\alpha^1(x).
\end{aligned}$$

□

Bemerkung 3.3.12.

1. In der Tat lässt sich auch zeigen, dass für isolierte lokale Minima $\overline{x}$ des allgemeinen (NLP) welche die (MFCQ) erfüllen P_α^1 eine exakte Penalty-Funktion ist, sofern α hinreichend groß gewählt wurde.
2. Die Exaktheit der Penalty-Funktion erscheint zunächst wünschenswert, liefert jedoch das Problem, dass P_α^1 nicht differenzierbar ist.
3. In der Tat lässt sich sogar stärker zeigen: Ist

$$P_\alpha(x) = f(x) + \alpha \pi_{\mathscr{U}}(x) + \alpha \pi_{\mathscr{G}}(x)$$

mit $\pi_{\mathscr{U}}, \pi_{\mathscr{G}} \geq 0$ und $\pi_{\mathscr{U}}(x) = 0$ genau dann, wenn $g(x) \leq 0$ und $\pi_{\mathscr{G}}(x) = 0$ genau dann, wenn $h(x) = 0$.
Ist dann $P_\alpha(x)$ in einem lokalen Minimum $\overline{x}$ mit $\nabla f(\overline{x}) \neq 0$ exakt, so ist $P_\alpha(x)$ nicht differenzierbar.

3.4 Sequential Quadratic Programming (SQP)

Wir werden uns nun zunächst auf den Fall eines gleichungsrestringierten Problems

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{u.d.N.} \quad h(x) = 0 \tag{NLP$_h$}$$

beschränken. Die KKT-Bedingungen dieses Problems legen es nahe das zugehörige Gleichungssystem für (x, μ) zu lösen.

3.4.1 Lagrange-Newton-Verfahren für Gleichungsrestriktionen

Ist $\overline{x}$ eine Lösung von (NLP_h), in der die (GCQ) gilt, so gibt es ein $\overline{\mu}$ welches den KKT-Bedingungen

$$\begin{aligned} \nabla_x \mathscr{L}(\overline{x}, \overline{\mu}) &= 0, \\ h(\overline{x}) &= 0 \end{aligned}$$

genügt. Dies legt es Nahe für $(\overline{x}, \overline{\mu})$ das Gleichungssystem für $\mathscr{F}\colon \mathbb{R}^{n+p} \to \mathbb{R}^{n+p}$ von der Form

$$\mathscr{F}(x, \mu) := \begin{pmatrix} \nabla_x \mathscr{L}(x, \mu) \\ h(x) \end{pmatrix} = 0$$

mit einem Newton-Verfahren zu lösen. Sind f und h jeweils zweimal stetig differenzierbar, so ist $\mathscr{F}$ einmal stetig differenzierbar, und es ist

$$\mathscr{F}'(x, \mu) = \begin{pmatrix} \nabla^2_{xx} \mathscr{L}(x, \mu) & \nabla^2_{x\mu} \mathscr{L}(x, \mu) \\ \nabla h(x)^T & 0 \end{pmatrix} = \begin{pmatrix} \nabla^2_{xx} \mathscr{L}(x, \mu) & \nabla h(x) \\ \nabla h(x)^T & 0 \end{pmatrix}.$$

In einem gegebenen Punkt (x^k, μ^k) ist damit der Newton-Schritt d^k gegeben durch

$$\mathscr{F}'(x^k, \mu^k) d^k = -\mathscr{F}(x^k, \mu^k)$$

bzw. für $d^k = \begin{pmatrix} d^k_x \\ d^k_\mu \end{pmatrix}$ in mehr Detail

$$\begin{pmatrix} \nabla^2_{xx} \mathscr{L}(x^k, \mu^k) & \nabla h(x^k) \\ \nabla h(x^k)^T & 0 \end{pmatrix} \begin{pmatrix} d^k_x \\ d^k_\mu \end{pmatrix} = \begin{pmatrix} -\nabla_x \mathscr{L}(x^k, \mu^k) \\ -h(x^k) \end{pmatrix}. \tag{LN}$$

Zusammenfassend erhalten wir das folgende lokale Lagrange-Newton-Verfahren

Algorithmus 3.4.1 (Lokales Lagrange-Newton-Verfahren)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$, $\mu^0 \in \mathbb{R}^p$.

for k = 0, 1, ..., **do**

- Prüfe auf Abbruch: Stopp falls $\mathscr{F}(x^k, \mu^k) = 0$ (d.h. x^k, μ^k sind KKT-Paar)
- Berechne $d^k \in \mathbb{R}^n$ durch Lösen der Gleichung (LN)
- Setze $x^{k+1} = x^k + d^k_x$ und $\mu^{k+1} = \mu^k + d^k_\mu$.

end for

Natürlich müssen wir i.A. eine Globalisierungsstrategie verwenden, um mehr als nur lokale Eigenschaften des Verfahrens zu erhalten. Zunächst aber wollen wir einsehen, dass wir zumindest in der Nähe eines *regulären KKT-Paares* die Lagrange-Newton-Gleichung (LN) lösen können.

Lemma 3.4.2. *Seien f und h zweimal stetig differenzierbar, $x \in \mathbb{R}^n$, $\mu \in \mathbb{R}^p$ gegeben. Ist dann*

$$\begin{aligned} &\operatorname{Rang} \nabla h(x) = p, \\ &\langle d, \nabla^2_{xx} \mathscr{L}(x,\mu) d \rangle > 0 \qquad \forall d \in \{d \in \mathbb{R}^n \setminus \{0\} \mid \nabla h(x)^T d = 0\}, \end{aligned} \tag{3.6}$$

so ist die Matrix

$$\mathscr{F}'(x,\mu) = \begin{pmatrix} \nabla^2_{xx} \mathscr{L}(x,\mu) & \nabla^2_{x\mu} \mathscr{L}(x,\mu) \\ \nabla h(x)^T & 0 \end{pmatrix} = \begin{pmatrix} \nabla^2_{xx} \mathscr{L}(x,\mu) & \nabla h(x) \\ \nabla h(x)^T & 0 \end{pmatrix}$$

invertierbar.

Bemerkung 3.4.3. Die erste Bedingung in (3.6) ist gerade die (LICQ) bzw. (MFCQ). Die zweite Bedingung ist gerade die zweite Ordnungsbedingung aus den hinreichenden Bedingungen zweiter Ordnung.

Beweis. Es genügt die Injektivität von $\mathscr{F}'(x,\mu)$ zu zeigen. Sei also $\begin{pmatrix} v \\ w \end{pmatrix} \in \operatorname{Kern} \mathscr{F}'(x,\mu)$. Die Multiplikation in der zweiten Blockzeile liefert

$$\nabla h(x)^T v = 0.$$

Die erste Blockzeile, skalar mit v multipliziert, liefert

$$0 = \langle v, \nabla^2_{xx} \mathscr{L}(x,\mu) v \rangle + \langle v, \nabla h(x) w \rangle = \langle v, \nabla^2_{xx} \mathscr{L}(x,\mu) v \rangle.$$

Aufgrund der zweiten Bedingung in (3.6) ist somit $v = 0$ (sonst wäre die rechte Seite positiv). Somit folgt aus der ersten Blockzeile

$$0 = \nabla^2_{xx} \mathscr{L}(x,\mu) v + \nabla h(x) w = \nabla h(x) w.$$

Da die Spalten von $\nabla h(x)$ linear unabhängig sind ist damit $w = 0$. □

Wir können somit einen Schritt des Verfahrens in der Nähe eines regulären KKT-Punktes durchführen und erhalten darüber hinaus

Theorem 3.4.4 (Lokale Konvergenz des Lagrange-Newton-Verfahrens)**.** *Seien $f \in C^2(\mathbb{R}^n;\mathbb{R})$ und $h \in C^2(\mathbb{R}^n;\mathbb{R}^p)$ und $(\overline{x},\overline{\mu})$ ein KKT-Paar mit*

$$\begin{aligned} &\operatorname{Rang} \nabla h(\overline{x}) = p, \\ &\langle d, \nabla^2_{xx} \mathscr{L}(\overline{x},\overline{\mu}) d \rangle > 0 \qquad \forall d \in \{d \in \mathbb{R}^n \setminus \{0\} | \nabla h(\overline{x})^T d = 0\}. \end{aligned}$$

Dann gibt es $\delta > 0$, so dass das lokale Lagrange-Newton-Verfahren 3.4.1 für alle $(x^0,\mu^0) \in$

$B_\delta(\overline{x},\overline{\mu})$ entweder mit $(x^k,\mu^k)=(\overline{x},\overline{\mu})$ terminiert, oder eine Folge (x^k,μ^k) erzeugt, die q-superlinear gegen $(\overline{x},\overline{\mu})$ konvergiert, d.h.

$$\|(x^{k+1}-\overline{x},\mu^{k+1}-\overline{\mu})\| = o(\|(x^k-\overline{x},\mu^k-\overline{\mu})\| \quad (k\to\infty).$$

Sind darüber hinaus $\nabla^2 f$ und $\nabla^2 h_i$ $(i\in\mathcal{G})$ Lipschitz-stetig auf $B_\delta(\overline{x},\overline{\mu})$, so ist die Konvergenz q-quadratisch.

Beweis. Aufgrund des vorhergehenden Lemmas 3.4.2 können wir Theorem 2.5.6 über das lokale Newton-Verfahren anwenden und erhalten die Aussagen des Theorems. □

Bemerkung 3.4.5. Analoge Vorgehensweisen können auch für Ungleichungsrestriktionen verwendet werden. Hierzu verwendet man sogenannte *Komplementaritätsfunktionen* $\Psi:\mathbb{R}^m\times\mathbb{R}^m\to\mathbb{R}^m$ mit

$$\Psi(x,y)=0 \quad\Leftrightarrow\quad x\le 0, y\le 0, xy=0.$$

Damit sind die allgemeinen KKT-Bedingungen äquivalent zu

$$\mathcal{F}(x,\lambda,\mu)=\begin{pmatrix}\nabla_x\mathcal{L}(x,\lambda,\mu)\\ h(x)\\ \Psi(g(x),-\lambda)\end{pmatrix}=0.$$

Ein typisches Beispiel einer solchen Komplementaritätsfunktion ist für beliebiges $c>0$ die Funktion

$$\Psi:\mathbb{R}\times\mathbb{R}\to\mathbb{R},\quad \Psi(x,y)=x+\max(0,cy-x).$$

Da solche Funktionen i.A. nicht differenzierbar gewählt werden (sonst kann es passieren, dass die Ableitung $\mathcal{F}'(x,\lambda,\mu)$, im Falle nicht strikter Komplementarität, auch in der Nähe eines KKT-Punktes singulär ist) können wir unsere bisherigen Resultate hierauf nicht ohne weiteres anwenden. Sie können jedoch, wie das obige Beispiel zeigt, schiefdifferenzierbar (Newton-Differenzierbar) sein, so dass ein lokales Newton-Verfahren, unter geeigneten Bedingungen, nach wie vor q-superlineare Konvergenz zeigt.

In der Tat ist für differenzierbares Ψ wegen

$$\Psi(x,0)=0 \qquad \forall x\le 0 \qquad \text{und} \qquad \Psi(0,y)=0 \qquad \forall y\le 0$$

notwendig $\nabla\Psi(0,0)=0$ und damit kann $\mathcal{F}'$ nicht invertierbar sein falls $g(x)=\lambda=0$ gilt.

3.4.2 Schief-Differenzierbarkeit

Als Exkurs zum Stoff der Vorlesung wollen wir noch etwas hinsichtlich des obigen Kommentars ausholen und sehen, das Funktionen vom Typ max durchaus in einem gewissen Sinne differenzierbar sind.

Definition 3.4.6. Eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$ ist *schief-differenzierbar* in einem Punkt $x \in \mathbb{R}^n$, falls es eine offene Umgebung $U(x)$ und eine Abbildung $G : U(x) \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m) \simeq \mathbb{R}^{m\times n}$ gibt, so dass

$$\lim_{\|h\|\to 0} \frac{\|F(x+h) - F(x) - G(x+h)h\|}{\|h\|} = 0$$

gilt.

Bemerkung 3.4.7. Im Vergleich zur normalen Differenzierbarkeit ist hierbei das Argument der „Ableitung" G nicht x sondern $x + h$.

In der Tat ist dies bereits ausreichend, um lokal schnelle Konvergenz des Newton-Verfahrens zu sehen. Deshalb nennt man eine solche Abbildung manchmal auch *Newton-differenzierbar*

Theorem 3.4.8. *Sei $\overline{x} \in \mathbb{R}^n$ eine Lösung der Gleichung $F(x) = 0$ für eine gegebene Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$. Sei F schief-differenzierbar in $\overline{x}$ mit zugehöriger Umgebung $U(\overline{x})$ und $G : U(\overline{x}) \to \mathbb{R}^{n\times m}$. Zudem sei $G(x) \in \mathscr{M}$ (invertierbar) für alle $x \in U(\overline{x})$ und $\sup_{x\in\mathscr{U}(\overline{x})} \|G^{-1}(x)\| \le C < \infty$. Dann ist die Newton-Iteration*

$$G(x^k)(x^{k+1} - x^k) = -F(x^k)$$

für $\|x^0 - \overline{x}\|$ hinreichend klein wohldefiniert und konvergiert q-superlinear gegen $\overline{x}$.

Beweis. O.B.d.A sei $x^k \in B_r(\overline{x}) \subset U(\overline{x})$ gegeben. Dann folgt (wie schon beim Newton-Verfahren)

$$x^{k+1} - \overline{x} = -G^{-1}(x^k)\Big(F(x^k) - F(\overline{x}) - G(x^k)(x^k - \overline{x})\Big).$$

Aus $x^k = \overline{x} + (x^k - \overline{x})$ und der Definition der Schief-Differenzierbarkeit erhalten wir somit

$$\begin{aligned}\|x^{k+1} - \overline{x}\| &\le \|G^{-1}(x^k)\|\|F(x^k) - F(\overline{x}) - G(x^k)(x^k - \overline{x})\| \\ &\le C\|F(\overline{x} + (x^k - \overline{x})) - F(\overline{x}) - G(\overline{x} + (x^k - \overline{x}))(x^k - \overline{x})\| \\ &\le Cc_r\|x^k - \overline{x}\|\end{aligned}$$

mit einer Konstante $c_r \to 0$ für $r \to 0$. Für $r > 0$ hinreichend klein ist damit

$$\|x^{k+1} - \overline{x}\| < \frac{1}{2}\|x^k - \overline{x}\| < \frac{r}{2}$$

und somit ist die Newton-Iteration wohldefiniert. Die q-superlineare Konvergenz folgt exakt wie im Beweis zur Konvergenz des lokalen Newtonverfahrens in Theorem 2.5.6. □

Offenkundig ist jede stetig-differenzierbare Funktion auch schief-differenzierbar. Aber auch einige nicht differenzierbare Funktionen sind schief-differenzierbar. Für die obige Komplementaritätsfunktion ist besonders wichtig, das dies für die Abbildung $x \mapsto \max(0, x)$ gilt. Diese ist in der Tat schief-differenzierbar in 0 mit Ableitung

$$G(x) := \begin{cases} 1 & x > 0 \\ 0 & x < 0 \\ c & x = 0 \end{cases}$$

für ein beliebiges $c \in \mathbb{R}$.

Wie bereits dieses Beispiel zeigt bringt die Schief-Differenzierbarkeit einige Probleme mit sich, so ist zum einen $G(x)$ nicht eindeutig definiert (oben haben wir einen freien Parameter), zudem hängt $G(x)$ auf $U(\overline{x})$ i.A. unstetig vom Punkt $\overline{x}$ ab, so dass wir auf eine vertiefte Analyse globalisierter Verfahren verzichten wollen.

3.4.3 Das lokale SQP-Verfahren

Die Lagrange-Newton-Gleichung (LN) ist nichts anderes als das KKT-System des quadratischen Problems

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \ & \frac{1}{2}\langle d, H_k d\rangle + \langle \nabla f(x^k), d\rangle \\ \text{u.d.N. } & h(x^k) + \nabla h(x^k)^T d = 0 \end{aligned} \tag{SQP$_h$}$$

wobei nun $H_k = \nabla^2_{xx}\mathscr{L}(x^k, \mu^k)$ ist! Dies ist also gerade eine quadratische Zielfunktion

$$m_k(d) = \frac{1}{2}\langle d, H_k d\rangle + \langle \nabla f(x^k), d\rangle.$$

Da die Nebenbedingung affin-linear ist, ist die (MFCQ) erfüllt, und folglich erfüllt jede lokale Lösung von (SQP$_h$) die KKT-Bedingungen. Somit gibt es zu einer Lösung d^k_{qp} (so es diese denn gibt) von (SQP$_h$) ein μ^k_{qp}, so dass

$$\begin{aligned} \nabla f(x^k) + H_k d^k_{qp} + \nabla h(x^k)\mu^k_{qp} &= 0, \\ h(x^k) + \nabla h(x^k)^T d^k_{qp} &= 0. \end{aligned}$$

Setzt man dann $d_x^k = d_{qp}^k$ und $d_\mu^k = \mu_{qp}^k - \mu^k$, so folgt

$$\begin{aligned} H_k d_x^k + \nabla h(x^k) d_\mu^k &= -\nabla f(x^k) - \nabla h(x^k)\mu^k, \\ \nabla h(x^k)^T d_x^k &= -h(x^k). \end{aligned}$$

Dies ist gerade, wie behauptet das System (LN) mit $d^k = \begin{pmatrix} d_x^k \\ d_\mu^k \end{pmatrix}$. Umgekehrt ist offenbar für jede Lösung d^k von (LN) durch $d_{qp}^k = d_x^k$ und $\mu_{qp}^k = \mu^k + d_\mu^k$ ein KKT-Paar von (SQP_h) gegeben. Zusammenfassend erhalten wir

Lemma 3.4.9. *Sind $f \in C^2(\mathbb{R}^n;\mathbb{R})$ und $h \in C^2(\mathbb{R}^n;\mathbb{R}^p)$ sowie $x^k \in \mathbb{R}^n$ und $\mu^k \in \mathbb{R}^p$ gegeben. Dann ist $d^k = \begin{pmatrix} d_x^k \\ d_\mu^k \end{pmatrix}$ genau dann eine Lösung von (LN), wenn $(d_{qp}^k, \mu_{qp}^k) = (d_x^k, \mu^k + d_\mu^k)$ ein KKT-Paar von (SQP_h) ist.*

Anhand der Lösung des SQP-Teilproblems (SQP_h) und dem zugehörigen Multiplikator können wir entscheiden, ob x^k eine lokale Lösung ist, denn es gilt:

Theorem 3.4.10. *Sind f und h zweimal stetig differenzierbar und $x^k \in \mathbb{R}^n$ und $\mu^k \in \mathbb{R}^p$ beliebig. Dann ist äquivalent*

1. *(x^k, μ^k) ist ein KKT-Paar von (NLP_h), in dem die hinreichenden Bedingungen 2. Ordnung 3.1.37 gelten.*
2. *$d_{qp}^k = 0$ ist eine isolierte lokale Lösung von (SQP_h) und $\mu_{qp}^k = \mu^k$ ist ein zugehöriger Lagrange-Multiplikator.*

Beweis. „1. ⇒ 2.": Es sei also 1. gegeben. Die Lagrange-Funktion für das QP-Problem (SQP_h) ist gerade

$$\mathcal{L}_k^{qp}(d_{qp}, \mu_{qp}) = m_k(d_{qp}) + \langle \mu_{qp}, h(x^k) + \nabla h(x^k)^T d_{qp} \rangle.$$

Daher ist

$$\begin{aligned} \nabla_d \mathcal{L}_k^{qp}(d_{qp}, \mu_{qp}) &= \nabla f(x^k) + H_k d_{qp} + \nabla h(x^k)\mu_{qp} \\ &= \nabla f(x^k) + \nabla h(x^k)\mu_{qp} + H_k d_{qp} \\ &= \nabla_x \mathcal{L}(x^k, \mu_{qp}) + H_k d_{qp}, \\ \nabla_{dd}^2 \mathcal{L}_k^{qp}(d_{qp}, \mu_{qp}) &= H_k \\ &= \nabla_{xx}^2 \mathcal{L}(x^k, \mu^k). \end{aligned}$$

(Man beachte, dass ja nun $H_k = \nabla^2_{xx}\mathcal{L}(x^k,\mu^k)$). Damit ist für $(d^k_{qp},\mu^k_{qp}) = (0,\mu^k)$ wegen 1.

$$\nabla_d \mathcal{L}^{qp}_k(0,\mu^k) = \nabla_x \mathcal{L}(x^k,\mu^k) = 0$$
$$\langle d, \nabla^2_{dd}\mathcal{L}^{qp}_k(d^k_{qp},\mu^k_{qp})d\rangle = \langle d, H_k d\rangle > 0 \quad \forall d \in \{d \in \mathbb{R}^n \setminus \{0\} \mid \nabla h(x^k)^T d = 0\}.$$

Dies sind gerade die hinreichenden Bedingungen 2. Ordnung für (SQP_h) und damit ist 2. erfüllt.

„2. ⇒ 1.": Nach Voraussetzung ist $d^k_{qp} = 0$ zulässig für (SQP_h) und somit ist

$$0 = h(x^k) + \nabla h(x^k)^T d^k_{qp} = h(x^k).$$

Da $\mu^k_{qp} = \mu^k$ ein zugehöriger Lagrange-Multiplikator ist folgt mit der im ersten Teil berechneten Ableitung wegen $d^k_{qp} = 0$

$$0 = \nabla_d \mathcal{L}^{qp}_k(0,\mu^k) = \nabla_x \mathcal{L}(x^k,\mu^k).$$

Damit ist (x^k,μ^k) ein KKT-Paar von (NLP_h).

Ist nun $d \in \text{Kern}(\nabla h(x^k)^T) \setminus \{0\}$ beliebig, so ist td für jedes $t \in \mathbb{R}$ zulässig für (SQP_h). Da $d^l_{qp} = 0$ ein isoliertes Minimum von m_k ist, ist $t = 0$ ein isoliertes lokales Minimum der quadratischen Funktion $\phi(t) = m_k(td)$. Somit ist $\phi''(0) > 0$ und somit

$$0 < \phi''(0) = \langle d, \nabla^2 m_k(0)d\rangle = \langle d, H_k d\rangle = \langle d, \nabla^2_{xx}\mathcal{L}(x^k,\mu^k)d\rangle.$$

Somit sind in (x^k,μ^k) die hinreichenden Bedingungen 2. Ordnung erfüllt. Dies zeigt 1. □

Damit können wir das lokale Lagrange-Newton-Verfahren umschreiben zu

Algorithmus 3.4.11 (Lokales SQP-Verfahren für Gleichungsrestriktionen)**.**

- Wähle Startpunkt $x^0 \in \mathbb{R}^n$, $\mu^0 \in \mathbb{R}^p$.

for k = 0, 1, . . . , **do**

- Prüfe auf Abbruch: Stopp falls (x^k,μ^k) ein KKT-Paar von (NLP_h) sind.
- Berechne $d^k_{qp} \in \mathbb{R}^n$ als eine Lösung von (SQP_h) und wähle μ^k_{qp} als zugehörigen Multiplikator.
- Setze $x^{k+1} = x^k + d^k_{qp}$ und $\mu^{k+1} = \mu^k_{qp}$.

end for

Analog lassen sich natürlich Varianten mit geeigneten Approximationen H_k der Hesse-Matrix $\nabla^2_{xx}\mathcal{L}(x^k,\mu^k)$ definieren.

3.4.4 SQP-Verfahren für Gleichungs- und Ungleichungsrestriktionen

Wir wenden uns nun wieder dem Problem (NLP) zu. Analog zu (SQP$_h$) können wir nun ein quadratisches Teilproblem definieren wobei wir jetzt natürlich $H_k := \nabla^2_{xx}\mathcal{L}(x^k, \lambda^k, \mu^k)$ wählen

$$\begin{aligned} \min_{d\in\mathbb{R}^n} \ & \frac{1}{2}\langle d, H_k d\rangle + \langle \nabla f(x^k), d\rangle \\ \text{u.d.N. } & g(x^k) + \nabla g(x^k)^T d \leq 0, \\ & h(x^k) + \nabla h(x^k)^T d = 0 \end{aligned} \tag{SQP}$$

Damit erhalten wir nun das

Algorithmus 3.4.12 (Lokales SQP-Verfahren)**.**

- Wähle Startpunkte $x^0 \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$ und $\mu^0 \in \mathbb{R}^p$.

for k = 0, 1, ..., **do**

- Prüfe auf Abbruch: Stopp falls (x^k, λ^k, μ^k) ein KKT-Tripel von (NLP) sind.
- Berechne $d^k_{qp} \in \mathbb{R}^n$ als eine Lösung (KKT-Tripel) von (SQP) und wähle λ^k_{qp} und μ^k_{qp} als zugehörigen Multiplikatoren.
- Setze $x^{k+1} = x^k + d^k_{qp}$, $\lambda^{k+1} = \lambda^k_{qp}$und $\mu^{k+1} = \mu^k_{qp}$.

end for

Auch für dieses Verfahren lässt sich die schnelle lokale Konvergenz zeigen

Theorem 3.4.13 (Konvergenz des lokalen SQP-Verfahrens)**.** *Seien $f \in C^2(\mathbb{R}^n;\mathbb{R})$, $g \in C^2(\mathbb{R}^n;\mathbb{R}^m)$ und $h \in C^2(\mathbb{R}^n;\mathbb{R}^p)$ gegeben.*

Im lokalen SQP-Verfahren 3.4.12 sei $H_k = \nabla^2_{xx}\mathcal{L}(x^k, \lambda^k, \mu^k)$ und im Falle einer mehrdeutigen Lösung des Teilproblems (SQP) wähle man unter allen möglichen KKT-Tripeln dasjenige $(d^k_{qp}, \lambda^k_{qp}, \mu^k_{qp})$ aus, welches den Abstand

$$\|(x^k + d^k_{qp}, \lambda^k_{qp}, \mu^k_{qp}) - (x^k, \lambda^k, \mu^k)\|$$

minimiert.

Ist dann $(\overline{x}, \overline{\lambda}, \overline{\mu})$ ein KKT-Tripel von (NLP) in dem (LICQ), strikte Komplementarität ($g_i(\overline{x}) + \overline{\lambda}_i \neq 0$) sowie die hinreichenden Bedingungen 2. Ordnung 3.1.37 gelten.

Dann gibt es ein $\delta > 0$, so dass das lokale SQP-Verfahren 3.4.12 für alle $(x^0, \lambda^0, \mu^0) \in B_\delta(\overline{x}, \overline{\lambda}, \overline{\mu})$ entweder mit $(x^k, \lambda^k, \mu^k) = (\overline{x}, \overline{\lambda}, \overline{\mu})$ terminiert, oder eine Folge (x^k, λ^k, μ^k) erzeugt, welche q-superlinear gegen $(\overline{x}, \overline{\lambda}, \overline{\mu})$ konvergiert. Sind darüber hinaus $\nabla^2 f$, $\nabla^2 g$ und $\nabla^2 h$ Lipschitz-stetig auf $B_\delta(\overline{x})$, so ist die Konvergenz q-quadratisch.

Beweis. Der Beweis erfolgt durch Anwenden des lokalen Newton-Verfahrens auf das Problem

$$\mathscr{F}(x,\lambda,\mu) := \begin{pmatrix} \nabla_x L(x,\lambda,\mu) \\ h(x) \\ \min(-g(x),\lambda) \end{pmatrix} = 0.$$

Hierbei beachte man, dass die dritte Blockzeile wegen der strikten Komplementarität in der Nähe von $(\overline{x},\overline{\lambda})$ in der Tat differenzierbar ist. Anschließend zeigt man, dass die Iterierten mit denen des SQP-Verfahrens übereinstimmen. Für Details siehe [Geiger and Kanzow, 2002, Satz 5.31]. □

3.4.5 Globalisiertes SQP-Verfahren

Für die Globalisierung des SQP-Verfahrens werden wir wieder eine Liniensuche verwenden. Um dies durchzuführen können wir jetzt allerdings nicht mehr den Funktionalwert von f heranziehen, da ein Schritt den Wert von f reduzieren kann, hierbei jedoch die Zulässigkeit verletzt. Wir müssen deshalb Abstieg bezüglich einer geeigneten Funktion (*Engl.: merit function*) messen. Es ist dabei hilfreich, wenn die hierzu verwendete Funktion die selben Minimierer wie f besitzt. Dies legt die Verwendung der exakten l_1-Penalty-Funktion

$$P_\alpha^1(x) = f(x) + \alpha\big(\|g(x)^+\|_1 + \|h(x)\|_1\big)$$

mit hinreichend großem α nahe. Diese ist natürlich nicht differenzierbar, allerdings besitzt P_α^1 noch Richtungsableitungen, was für unsere Armijo-Liniensuche ausreichend ist.

Definition 3.4.14 (Richtungsableitung)**.** Eine stetige Funktion $\phi : \mathbb{R}^n \to \mathbb{R}$ heißt *richtungsdifferenzierbar* in einem Punkt $x \in \mathbb{R}^n$, falls für alle $d \in \mathbb{R}^n$ die *Richtungsableitung*

$$\phi'(x;d) = \lim_{t \downarrow 0} \frac{\phi(x+td)-\phi(x)}{t} \in \mathbb{R}$$

existiert.

Bemerkung 3.4.15. Man beachte, dass $\phi'(x;\cdot)$ nicht linear in d sein muss. Die Richtungsableitung ist lediglich positiv homogen.

Ein Beispiel ist die Funktion $\phi : \mathbb{R} \to \mathbb{R}$ mit $\phi(t) = |t|$ und $\phi'(0,1) = \phi'(0,-1) = 1$.

Ist $\phi \in C^1(\mathbb{R}^n;\mathbb{R})$, so ist die Richtungsableitung durch die Ableitung der Funktion gegeben, und es gilt

$$\phi'(x)d = \phi'(x;d) \quad \forall d \in \mathbb{R}^n.$$

Lemma 3.4.16 (Richtungsdifferenzierbarkeit von P_α^1)**.** *Sind $f \in C^1(\mathbb{R}^n;\mathbb{R})$, $g \in C^1(\mathbb{R}^n;\mathbb{R}^m)$ und $h \in C^1(\mathbb{R}^n;\mathbb{R}^p)$ und $\alpha > 0$. Dann ist P_α^1 in jedem Punkt $x \in \mathbb{R}^n$ richtungsdifferenzierbar mit*

$$\begin{aligned}(P_\alpha^1)'(x;d) = \langle \nabla f(x), d\rangle &+ \alpha \sum_{\{i\in\mathscr{U}\,|\,g_i(x)>0\}} \langle \nabla g_i(x), d\rangle + \alpha \sum_{\{i\in\mathscr{U}\,|\,g_i(x)=0\}} \langle \nabla g_i(x), d\rangle^+ \\ &+ \alpha \sum_{\{i\in\mathscr{G}\,|\,h_i(x)>0\}} \langle \nabla h_i(x), d\rangle - \alpha \sum_{\{i\in\mathscr{G}\,|\,h_i(x)<0\}} \langle \nabla h_i(x), d\rangle \\ &+ \alpha \sum_{\{i\in\mathscr{G}\,|\,h_i(x)=0\}} |\langle \nabla h_i(x), d\rangle|\end{aligned}$$

Beweis. Aufgrund der Definition der Richtungsableitung gilt die Summenregel sofern alle Summanden richtungsdifferenzierbar sind, d.h. es ist

$$(P_\alpha^1)'(x;d) = f'(x;d) + \alpha \sum_{i\in\mathscr{U}} \max(0, g_i(\cdot))'(x;d) + \alpha \sum_{i\in\mathscr{G}} |h_i(\cdot)|'(x;d).$$

Nun ist $f \in C^1$ und damit

$$f'(x;d) = \langle \nabla f(x), d\rangle.$$

Ist $g_i(x) < 0$ oder $g_i(x) > 0$, so ist $\max(0,\cdot)$ in $g_i(x)$ stetig differenzierbar und somit

$$\begin{aligned}\max(0, g_i(\cdot))'(x;d) &= 0'(x;d) = 0 && g_i(x) < 0,\\ \max(0, g_i(\cdot))'(x;d) &= g_i'(x;d) = \langle \nabla g_i(x), d\rangle && g_i(x) > 0.\end{aligned}$$

Analog folgt für $h_i(x) \neq 0$

$$\begin{aligned}|h_i(\cdot)|'(x;d) &= -h_i'(x;d) = -\langle \nabla h(x), d\rangle && h_i(x) < 0,\\ |h_i(\cdot)|'(x;d) &= h_i'(x;d) = \langle \nabla h_i(x), d\rangle && h_i(x) > 0.\end{aligned}$$

Es bleiben also gerade die Fälle in denen $g_i(x) = 0$ oder $h_i(x) = 0$. Es folgt dann durch

Taylorentwicklung für g_i bzw. h_i

$$\begin{aligned}
\max(0, g_i(\cdot))'(x;d) &= \lim_{t\downarrow 0} \frac{g_i(x+td)^+ - g_i(x)^+}{t} \\
&= \lim_{t\downarrow 0} \frac{g_i(x+td)^+}{t} \\
&= \lim_{t\downarrow 0} \Big(\frac{g_i(x+td) + g_i(x)}{t} \Big)^+ \\
&= \lim_{t\downarrow 0} \Big(\langle \nabla g_i(x), d\rangle + \frac{o(t)}{t} \Big)^+ \\
&= \langle \nabla g_i(x), d\rangle^+ \\
|h_i(\cdot)|'(x;d) &= \lim_{t\downarrow 0} \frac{|h_i(x+td)| - |h_i(x)|}{t} \\
&= \lim_{t\downarrow 0} \frac{|h_i(x+td)|}{t} \\
&= \lim_{t\downarrow 0} \Big| \langle \nabla h_i(x), d\rangle + \frac{o(t)}{t} \Big| \\
&= |\langle \nabla h_i(x), d\rangle|.
\end{aligned}$$

Dies zeigt die Behauptung. □

In der Tat liefert eine Lösung von (SQP) eine Abstiegsrichtung für P^1_α, sofern α hinreichend groß ist und H_k positiv definit ist, denn es gilt:

Theorem 3.4.17. *Seien $f \in C^1(\mathbb{R}^n;\mathbb{R})$, $g \in C^1(\mathbb{R}^n;\mathbb{R}^m)$ und $h \in C^1(\mathbb{R}^n;\mathbb{R}^p)$ sowie $(d^k_{qp}, \lambda^k_{qp}, \mu^k_{qp})$ ein KKT-Tripel von (SQP). Ist dann*

$$\alpha \geq \max\Big(\|\lambda^k_{qp}\|_\infty, \|\mu^k_{qp}\|_\infty \Big),$$

so ist

$$(P^1_\alpha)'(x^k; d^k_{qp}) \leq -\langle d^k_{qp}, H_k d^k_{qp}\rangle.$$

Insbesondere ist d^k_{qp} eine Abstiegsrichtung für P^1_α, wenn H_k positiv definit ist.

Beweis. Aufgrund der Komplementarität

$$0 \leq \lambda^k_{qp} \quad \perp \quad g(x^k) + \nabla g(x^k)^T d^k_{qp} \leq 0$$

ist

$$\begin{aligned}\langle \lambda_{qp}^k, \nabla g(x^k)^T d_{qp}^k \rangle &= \sum_{\{i \in \mathscr{U} \mid g_i(x^k) > 0\}} (\lambda_{qp}^k)_i \langle \nabla g_i(x^k), d_{qp}^k \rangle \\ &\quad - \sum_{\{i \in \mathscr{U} \mid g_i(x^k) \leq 0\}} (\lambda_{qp}^k)_i g_i(x^k) \\ &\geq \sum_{\{i \in \mathscr{U} \mid g_i(x^k) > 0\}} (\lambda_{qp}^k)_i \underbrace{\langle \nabla g_i(x^k), d_{qp}^k \rangle}_{<0} \\ &\geq \alpha \sum_{\{i \in \mathscr{U} \mid g_i(x^k) > 0\}} \langle \nabla g_i(x^k), d_{qp}^k \rangle.\end{aligned}$$

Analog folgt aus $h(x^k) + \nabla h(x^k)^T d_{qp}^k = 0$

$$\begin{aligned}\langle \mu_{qp}^k, \nabla h(x^k)^T d_{qp}^k \rangle &= \sum_{\{i \in \mathscr{G} \mid h_i(x^k) > 0\}} (\mu_{qp}^k)_i \underbrace{\langle \nabla h_i(x^k), d_{qp}^k \rangle}_{<0} \\ &\quad + \sum_{\{i \in \mathscr{G} \mid h_i(x^k) < 0\}} (\mu_{qp}^k)_i \underbrace{\langle \nabla h_i(x^k), d_{qp}^k \rangle}_{>0} \\ &\geq \alpha \sum_{\{i \in \mathscr{G} \mid h_i(x^k) > 0\}} \langle \nabla h_i(x^k), d_{qp}^k \rangle \\ &\quad - \alpha \sum_{\{i \in \mathscr{G} \mid h_i(x^k) < 0\}} \langle \nabla h_i(x^k), d_{qp}^k \rangle.\end{aligned}$$

Aus der Stationarität der Lagrange-Funktion

$$0 = \nabla_d \mathscr{L}_k^{qp}(d_{qp}^k, \lambda_{qp}^k, \mu_{qp}^k) = \nabla f(x^k) + H_k d_{qp}^k + \nabla g(x^k) \lambda_{qp}^k + \nabla h(x^k) \mu_{qp}^k$$

ergibt sich mit den vorhergehenden Rechnungen

$$\begin{aligned}\langle \nabla f(x^k), d_{qp}^k \rangle &= -\langle d_{qp}^k, H_k d_{qp}^k \rangle - \langle \nabla g(x^k) \lambda_{qp}^k, d_{qp}^k \rangle - \langle \nabla h(x^k) \mu_{qp}^k, d_{qp}^k \rangle \\ &\leq -\langle d_{qp}^k, H_k d_{qp}^k \rangle - \alpha \sum_{\{i \in \mathscr{U} \mid g_i(x^k) > 0\}} \langle \nabla g_i(x^k), d_{qp}^k \rangle \\ &\quad - \alpha \sum_{\{i \in \mathscr{G} \mid h_i(x^k) > 0\}} \langle \nabla h_i(x^k), d_{qp}^k \rangle + \alpha \sum_{\{i \in \mathscr{G} \mid h_i(x^k) < 0\}} \langle \nabla h_i(x^k), d_{qp}^k \rangle\end{aligned}$$

Durch Einsetzen dieser Ungleichung in die Darstellung der Richtungsableitung in Lem-

ma 3.4.16 ergibt sich

$$\begin{aligned}(P_\alpha^1)'(x^k;d_{qp}^k) &= \langle \nabla f(x^k), d_{qp}^k\rangle + \alpha \sum_{\{i\in\mathscr{U}\,|\,g_i(x^k)>0\}} \langle \nabla g_i(x^k), d_{qp}^k\rangle \\ &\quad + \alpha \sum_{\{i\in\mathscr{U}\,|\,g_i(x^k)=0\}} \langle \nabla g_i(x^k), d_{qp}^k\rangle^+ + \alpha \sum_{\{i\in\mathscr{G}\,|\,h_i(x^k)>0\}} \langle \nabla h_i(x^k), d_{qp}^k\rangle \\ &\quad - \alpha \sum_{\{i\in\mathscr{G}\,|\,h_i(x^k)<0\}} \langle \nabla h_i(x^k), d_{qp}^k\rangle + \alpha \sum_{\{i\in\mathscr{G}\,|\,h_i(x^k)=0\}} |\langle \nabla h_i(x^k), d_{qp}^k\rangle| \\ &\le -\langle d_{qp}^k, H_k d_{qp}^k\rangle \\ &\quad + \alpha \sum_{\{i\in\mathscr{U}\,|\,g_i(x^k)=0\}} \langle \nabla g_i(x^k), d_{qp}^k\rangle^+ + \alpha \sum_{\{i\in\mathscr{G}\,|\,h_i(x^k)=0\}} |\langle \nabla h_i(x^k), d_{qp}^k\rangle|.\end{aligned}$$

Nun ist

$$\begin{aligned}\langle \nabla g_i(x^k), d_{qp}^k\rangle &\le -g_i(x^k),\\ \langle \nabla h_i(x^k), d_{qp}^k\rangle &= -h_i(x^k).\end{aligned}$$

Entsprechend sind die beiden Summen auf der rechten Seite Null, und die Behauptung gezeigt. □

Mit diesen Überlegungen scheint sich zunächst das folgende globalisierte SQP-Verfahren anzubieten

Algorithmus 3.4.18 (Entwurf eines globalisiertes SQP-Verfahren)**.**

- Wähle Startpunkte $x^0 \in \mathbb{R}^n$, $\lambda^0 \in \mathbb{R}^m$ und $\mu^0 \in \mathbb{R}^p$.
- Wähle $H_k \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit, $\alpha > 0$ hinreichend groß und $0 < \gamma < 1/2$.

for k = 0, 1, . . . , **do**

- Prüfe auf Abbruch: Stopp falls (x^k, λ^k, μ^k) ein KKT-Tripel von (NLP) ist.
- Berechne $d_{qp}^k \in \mathbb{R}^n$ als eine Lösung von (SQP) und wähle λ_{qp}^k und μ_{qp}^k als zugehörigen Multiplikatoren.

  Bestimme die größte Zahl $t_k \in \{1, 2^{-1}, 2^{-2}, \ldots\}$, so dass

$$P_\alpha^1(x^k + t_k d_{qp}^k) - P_\alpha^1(x^k) \le \gamma t_k (P_\alpha^1)'(x^k; d_{qp}^k)$$

- Setze $x^{k+1} = x^k + t_k d_{qp}^k$, und berechne neue Multiplikatoren, z.B., $\lambda^{k+1} = \lambda_{qp}^k$ und $\mu^{k+1} = \mu_{qp}^k$.
- Wähle $H_{k+1} \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit.

end for

Bemerkung 3.4.19.

1. In der Tat kann gezeigt werden das die Armijo-artige Liniensuche nach endlich vielen Schritten terminiert. (Es ist $P^1_\alpha(x+td) = P^1_\alpha(x) + t(P^1_\alpha)'(x;d) + o(t)$)
2. Die Wahl von α hinreichend groß wird in der Praxis durch geeignete Aufdatierung, z.B.
$$\alpha_{k+1} = \max\{\alpha_k, \max(\|\lambda^{k+1}_{qp}\|_\infty, \|\mu^{k+1}_{qp}\|_\infty) + \delta\}$$
mit einem gegebenen $\delta > 0$ ersetzt werden.
3. Um eine positiv definite Matrix $H_k \approx \nabla^2_{xx}\mathcal{L}(x^k, \lambda^k, \mu^k)$ zu wählen können wir natürlich wieder BFGS (oder andere) Aufdatierungsformeln verwenden. Da wir hierfür jedoch nicht mit einer Armijo-Liniensuche auskommen müssen wir hierfür noch weitere Arbeit investieren.
4. Es ist i.A. nicht klar, dass (SQP) eine Lösung besitzt!
5. Der Übergang zu schneller lokaler Konvergenz ist i.A. nicht gegeben.

3.4.6 Probleme und Modifikationen des SQP-Verfahrens

Wie wir in der vorhergehenden Bemerkung gesehen haben gibt es noch einige Probleme in unserem Entwurf für ein SQP-Verfahren. Wir wollen diese im folgenden Diskutieren.

3.4.6.1 Unzulässige Teilprobleme

Man kann leicht einsehen, dass das SQP-Teilproblem (SQP) i.A. keine zulässigen Punkte hat (*wir sagen das Problem ist unzulässig*), und somit das globalisierte SQP-Verfahren 3.4.18 nicht durchführbar ist.

Beispiel 3.4.20. Man betrachte für beliebiges f das Problem
$$\min_{x\in\mathbb{R}} f(x) \quad \text{u.d.N.} \quad g(x) := 1 - x^2 \le 0.$$
Dann ist im Punkt $x^k = 0$ der Gradient $\nabla g(x^k) = -2x^k = 0$ und damit erhalten wir die unerfüllbare Nebenbedingung
$$g(x^k) + \nabla g(x^k) d^k_{qp} = 1 \le 0$$
für das Problem (SQP).

Um dieses Problem zu umgehen können wir das Problem (SQP) relaxieren in dem wir die Verletzung der Zulässigkeit bestrafen. Wir führen hierzu *Schlupfvariablen* v, w^p, w^m ein und betrachten für einen Penalty-Parameter $\alpha > 0$ das Problem

$$\begin{aligned} \min_{d\in\mathbb{R}^n, v\in\mathbb{R}^m, w^p, w^m\in\mathbb{R}^p} \ & \frac{1}{2}\langle d, H_k d\rangle + \langle \nabla f(x^k), d\rangle + \alpha \sum_{i\in\mathscr{U}} v_i + \alpha \sum_{i\in\mathscr{G}} (w_i^p + w_i^m) \\ \text{u.d.N. } & g(x^k) + \nabla g(x^k)^T d - v \le 0, \\ & h(x^k) + \nabla h(x^k)^T d - w^p + w^m = 0, \\ & v \ge 0, \quad w^p \ge 0, \quad w^m \ge 0. \end{aligned} \tag{SQP+}$$

Durch die Relaxierung besitzt (SQP+) immer zulässige Punkt.

Man beachte, das der Penalty-Term wegen der Nichtnegativität von v, w^p, w^m in der Tat mit der l_1-Penalty Funktion für die Nebenbedingung $v = 0$, $w^p = w^m = 0$ übereinstimmt. Es ist daher leicht das folgende Einzusehen

Lemma 3.4.21.

1. *Ist* $(d_{qp}^k, \lambda_{qp}^k, \mu_{qp}^k)$ *ein KKT-Tripel von* (SQP)*, so ist für*

   $$\alpha \ge \max\left(\|\lambda_{qp}^k\|_\infty, \|\mu_{qp}^k\|_\infty\right)$$

   der Vektor $(d_{qp}^k, v, w^p, w^m, \lambda_{qp}^k, \mu_{qp}^k, \xi^v, \xi^p, \xi^m)$ *mit*

   $$v = 0, \quad w^p = w^m = 0, \quad \xi^v = \alpha e - \lambda_{qp}^k, \quad \xi^p = \alpha e - \mu_{qp}^k, \quad \xi^m = \alpha e + \mu_{qp}^k$$

   ein KKT-Tupel von (SQP+)*. Hierbei ist* $e = (1, \dots, 1)^T$*,* ξ^v *der Multiplikator zu* $-v \le 0$*, sowie* ξ^p *und* ξ^m *die Multiplikatoren zu* $-w^p, -w^m \le 0$*.*
2. *Ist* $(d_{qp}^k, 0, 0, 0, \lambda_{qp}^k, \mu_{qp}^k, \xi^v, \xi^p, \xi^m)$ *ein KKT-Tupel von* (SQP+)*, so ist* $(d_{qp}^k, \lambda_{qp}^k, \mu_{qp}^k)$ *ein KKT-Tripel von* (SQP)*.*

Somit stimmen die Lösungen von (SQP) (so sie denn existieren) mit Lösungen von (SQP+) für hinreichend große α überein.

Bemerkung 3.4.22. Für ein globalisiertes SQP-Verfahren 3.4.18 mit (SQP+) anstelle von (SQP) kann man globale Konvergenz, unter geeigneten Bedingungen an H_k, zeigen, vgl. [Geiger and Kanzow, 2002, Abschnitt 5.5.8]. Genauer, gilt

$$c\|d\|^2 \le \langle d, H_k d\rangle \le C\|d\|^2 \quad \forall d \in \mathbb{R}^n, \forall k \in \mathbb{N}$$

so ist jeder Häufungspunkt der mit dem globalisierte SQP-Verfahren und (SQP+) als

Teilproblem erzeugten Folge ein stationärer Punkt von (NLP).

Bemerkung 3.4.23. Da in der Zielfunktion von (SQP+) die Variablen v, w^p, w^m nur linear auftreten ist (SQP+) nicht strikt konvex. Da dies bei manchen Lösungsverfahren zu Problemen führen kann wird manchmal ein zusätzlicher quadratischer Term

$$\tilde{\alpha}\Big(\|v\|^2 + \|w^p\|^2 + \|w^m\|^2\Big)$$

zur Zielfunktion addiert.

3.4.6.2 Der Maratos-Effekt

Zunächst einmal erinnern wir uns das wir von den globalisierten Lösungsverfahren stets zeigen konnten, dass diese hinreichend Nahe an eine geeigneten Lösung wieder in die zugrunde liegende lokale Variante übergehen, da z.B. die Schrittweite $t_k = 1$ akzeptiert wird. Hierzu ist es notwendig, dass irgendwann einmal

$$P^1_\alpha(x^k + d^k_{qp}) < P^1_\alpha(x^k)$$

ist. Das dies i.A. nicht erfüllt wird ist der *Maratos-Effekt*. Der Einfachheit halber betrachten wir den Fall reiner Gleichungsrestriktionen:

Beispiel 3.4.24 (Maratos-Effekt)**.** Wir betrachten das Problem

$$\begin{aligned} \min_{x\in\mathbb{R}^2} f(x) &:= 2(x_1^2 + x_2^2 - 1) - x_1 \\ \text{u.d.N. } h(x) &:= x_1^2 + x_2^2 - 1 = 0. \end{aligned}$$

Damit ergibt sich

$$\begin{aligned} \nabla f(x) &= 4x - \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & \nabla^2 f(x) &= 4I, \\ \nabla h(x) &= 2x, & \nabla^2 h(x) &= 2I, \\ \nabla^2_{xx}\mathcal{L}(x,\mu) &= (4+2\mu)I. \end{aligned}$$

Für $x \in X$ ist dann $\|x\| = 1$ und somit

$$f(x) = 2h(x) - x_1 = -x_1 \begin{cases} > -1 & x \neq (1,0)^T \\ = -1 & x = (1,0)^T \end{cases}.$$

Das globale Minimum ist daher $\overline{x} = (1,0)^T$ mit $f(\overline{x}) = -1$ und aus der Multiplikatorregel

folgt

$$0 = \nabla f(\overline{x}) + \nabla h(\overline{x})\overline{\mu} = \begin{pmatrix} 3 \\ 0 \end{pmatrix} + \overline{\mu}\begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

bzw. $\overline{\mu} = -\frac{3}{2}$.

Sei nun $x^k \in X$ aber $x^k \neq \pm\overline{x}$ und $-2 < \mu^k < -1$. Dann ist das SQP-Teilproblem (SQP_h) lösbar! Sei also d_{qp}^k die Lösung von (SQP_h) und μ_{qp}^k ein zugehöriger Multiplikator, es gilt also

$$\begin{aligned} \nabla f(x^k) + \nabla_{xx}^2 \mathscr{L}(x^k, \mu^k) d_{qp}^k + \nabla h(x^k)\mu_{qp}^k &= 0, \\ h(x^k) + \nabla h(x^k)^T d_{qp}^k &= 0. \end{aligned}$$

Wir zeigen zunächst, dass $d_{qp}^k \neq 0$, denn ansonsten wäre

$$\begin{aligned} 0 &= \nabla f(x^k) + \nabla_{xx}^2 \mathscr{L}(x^k, \mu^k) d_{qp}^k + \nabla h(x^k)\mu_{qp}^k \\ &= (4 + 2\mu_{qp}^k)x^k - \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \end{aligned}$$

Die Vektoren x^k und $\overline{x} = (1,0)^T$ wären also linear abhängig. Wegen unserer Wahl $x^k \in X \setminus \{\pm\overline{x}\}$ ist dies nicht möglich, folglich ist $d_{qp}^k \neq 0$.

Aus der Definition von f ersehen wir zusammen mit den KKT-Bedingungen für d_{qp}^k und μ_{qp}^k

$$\begin{aligned} f(x^k + d_{qp}^k) - f(x^k) &= \langle \nabla f(x^k), d_{qp}^k \rangle + \frac{1}{2}\langle d_{qp}^k, \nabla^2 f(x^k) d_{qp}^k \rangle \\ &= \langle \nabla f(x^k), d_{qp}^k \rangle + 2\|d_{qp}^k\|^2 \\ &= -\langle \mu_{qp}^k, \nabla h(x^k)^T d_{qp}^k \rangle - \langle d_{qp}^k, \nabla_{xx}^2 \mathscr{L}(x^k, \mu^k) d_{qp}^k \rangle + 2\|d_{qp}^k\|^2 \\ &= -\langle \mu_{qp}^k, \nabla h(x^k)^T d_{qp}^k \rangle - (4 + 2\mu^k)\|d_{qp}^k\|^2 + 2\|d_{qp}^k\|^2 \\ &= \mu_{qp}^k h(x^k) - (2 + 2\mu^k)\|d_{qp}^k\|^2. \end{aligned}$$

Aufgrund unserer Wahl ist $h(x^k) = 0$, $\mu^k < -1$ und $d_{qp}^k \neq 0$ und somit erhalten wir aus obiger Rechnung

$$f(x^k + d_{qp}^k) - f(x^k) > 0.$$

Taylorentwicklung von h liefert wegen $h(x^k) = 0$ sowie der Zulässigkeit von d_{qp}^k für (SQP_h)

$$\begin{aligned} |h(x^k + d_{qp}^k)| - |h(x^k)| &= |h(x^k + d_{qp}^k)| \\ &\geq 0. \end{aligned}$$

Für unsere Merit-Funktion folgt somit für jedes $\alpha \geq 0$

$$P_\alpha^1(x^k + d_{qp}^k) - P_\alpha^1(x^k) = f(x^k + d_{qp}^k) - f(x^k) + \alpha\big(|h(x^k + d_{qp}^k)| - |h(x^k)|\big) > 0.$$

Ursächlich liegt der Maratos-Effekt darin, dass sich der Schritt $x^k + d_{qp}^k$ zu sehr von der gekrümmten Mannigfaltigkeit der zulässigen Punkte entfernt. Um dieses Problem zu umgehen bieten sich zwei Auswege

Second-Order Correction Zu dem Schritt d_{qp}^k kann eine Korrektur zweiter Ordnung

$$d_{\text{soc}}^k = -(\nabla h(x^k)\nabla h(x^k)^T)^{-1}\nabla h(x^k)h(x^k + d_{qp}^k)$$

bestimmt werden (sofern die Inverse existiert). Es ist dann per Definition der von d_{qp}^k erfüllten Nebenbedingung sowie der Tatsache dass h in unserem Beispiel quadratisch und skalar ist

$$\begin{aligned}
\|d_{\text{soc}}^k\| &\leq C\|h(x^k + d_{qp}^k)\| \\
&= C\|h(x^k) + \nabla h(x^k)^T d_{qp}^k + 1/2\langle d_{qp}^k, \nabla^2 h(x^k) d_{qp}^k\rangle\| \\
&= C\|\langle d_{qp}^k, \nabla^2 h(x^k) d_{qp}^k\rangle\| \\
&\leq C\|d_{qp}^k\|^2.
\end{aligned}$$

Der Schritt d_{soc}^k erfüllt per Definition die linearisierte Bedingung

$$h(x^k + d_{qp}^k) + \nabla h(x^k)^T d_{\text{soc}}^k = 0$$

und verbessert hierdurch die Zulässigkeit. Es ist nämlich

$$\begin{aligned}
h(x^k + d_{qp}^k + d_{\text{soc}}^k) &= h(x^k) + \nabla h(x^k)^T(d_{qp}^k + d_{\text{soc}}^k) + 1/2\langle d_{qp}^k + d_{\text{soc}}^k, \nabla^2 h(x^k)(d_{qp}^k + d_{\text{soc}}^k)\rangle \\
&= -h(x^k + d_{qp}^k) + 1/2\langle d_{qp}^k + d_{\text{soc}}^k, \nabla^2 h(x^k)(d_{qp}^k + d_{\text{soc}}^k)\rangle \\
&= -1/2\langle d_{qp}^k, \nabla^2 h(x^k) d_{qp}^k\rangle + 1/2\langle d_{qp}^k + d_{\text{soc}}^k, \nabla^2 h(x^k)(d_{qp}^k + d_{\text{soc}}^k)\rangle \\
&\leq \langle d_{qp}^k, \nabla^2 h(x^k) d_{\text{soc}}^k\rangle + 1/2\langle d_{\text{soc}}^k, \nabla^2 h(x^k) d_{\text{soc}}^k\rangle \\
&\leq C\|d_{qp}^k\|^3.
\end{aligned}$$

Damit kann man nun zusätzlich zum Schritt $x^k + d_{qp}^k$ auch den Schritt $x^k + d_{qp}^k + d_{\text{soc}}^k$ auf Abstieg prüfen bevor man mit der Liniensuche beginnt.

Für diese modifizierte Strategie lässt sich der Übergang zu lokal schneller Konvergenz unter geeigneten Bedingungen zeigen. Man beachte hierzu, das $\|d_{\text{soc}}^k\| = O(\|d_{qp}^k\|^2)$ ist, dieser zusätzliche Schritt die Konvergenzgeschwindigkeit also nicht stört, aber bei der Akzeptanz der Schrittweite $t_k = 1$ hilfreich ist, vgl. Fukushima [1986].

Nicht-monotone Liniensuche Alternativ kann man nicht monotone Liniensuchstrategien verwenden, bei der ein Schritt d^k_{qp} in x^k testweise auch dann akzeptiert wird, wenn die Merit-Funktion keinen Abstieg zeigt. Man beobachtet dann die nächsten l-Iterierten.

Zeigt sich auch nach l-Schritten noch kein Abstieg, d.h.

$$P^1_\alpha(x^k) \leq P^1_\alpha(x^{k+i}) \quad \forall i \in \{1,\dots,l\},$$

so verwirft man die letzten l-Schritte und führt die übliche Liniensuche in x^k durch.

3.4.6.3 BFGS-Updates

Im Rahmen des BFGS-Updates benötigten wir die Bedingung

$$\langle y^k, d^k\rangle > 0$$

wobei jetzt $d^k = x^{k+1} - x^k$, $y^k = \nabla_x \mathscr{L}(x^{k+1},\lambda^k,\mu^k) - \nabla_x \mathscr{L}(x^k,\lambda^k,\mu^k)$, um sicherzustellen, dass H^{BFGS}_{k+1} positiv definit bleibt. Wir hatten die Powell-Wolfe-Regel verwendet, um diese Bedingung zu sichern. Da jedoch $(P^1_\alpha)'(x^k;d^k)$ nicht linear in d^k zu sein braucht, und i.A. $P'_\alpha \neq \nabla_x \mathscr{L}$ können wir die Existenz ein Powell-Wolfe Schrittweite nicht wie in Lemma 2.4.6 erschließen.

Stattdessen wird oft ein „gedämpftes BFGS-Update"

$$H_{k+1} = H^{BFGS}(H_k, d^k, y^k_{\text{mod}})$$

mit

$$y^k_{\text{mod}} = \theta_k y^k + (1-\theta_k) H_k d^k$$

für

$$\theta_k = \begin{cases} 1 & \langle y^k, d^k\rangle \geq 0.2\langle d^k, H_k d^k\rangle, \\ 0.8\frac{\langle d^k, H_k d^k\rangle}{\langle d^k, H_k d^k\rangle - \langle y^k, d^k\rangle} & \text{sonst} \end{cases}$$

verwendet. Es ist dann in der Tat $\langle y^k_{\text{mod}}, d^k\rangle > 0$ und damit überträgt sich, nach Theorem 2.8.15, positiv Definitheit von H_k auf H_{k+1}.

In der Tat ist $\theta_k = 1$, so gilt $y^k = y^k_{\text{mod}}$, $\langle y^k, d^k\rangle \geq 0.2\langle d^k, H_h d^k\rangle > 0$ ansonsten gilt

$$\begin{aligned}
\langle y^k_{\text{mod}}, d^k\rangle &= \underbrace{0.8\frac{\langle d^k, H_h d^k\rangle}{\langle d^k, H_h d^k\rangle - \langle y^k, d^k\rangle}}_{=\theta_k}\langle y^k, d^k\rangle \\
&\quad + \underbrace{\frac{\langle d^k, H_h d^k\rangle - \langle y^k, d^k\rangle - 0.8\langle d^k, H_h d^k\rangle}{\langle d^k, H_h d^k\rangle - \langle y^k, d^k\rangle}}_{=1-\theta_k}\langle d^k, H_k d^k\rangle \\
&= \langle d^k, H_k d^k\rangle \frac{0.8\langle y^k, d^k\rangle + 0.2\langle d^k, H_h d^k\rangle - \langle y^k, d^k\rangle}{\langle d^k, H_h d^k\rangle - \langle y^k, d^k\rangle} \\
&= 0.2\langle d^k, H_k d^k\rangle \frac{\langle d^k, H_h d^k\rangle - \langle y^k, d^k\rangle}{\langle d^k, H_h d^k\rangle - \langle y^k, d^k\rangle} \\
&= 0.2\langle d^k, H_k d^k\rangle \\
&> 0.
\end{aligned}$$

Wegen $\langle y^k, d^k\rangle = \langle H_{k+1} d^k, d^k\rangle$ und $H_{k+1} \approx H_k$ besteht die Hoffnung, dass nahe eines lokalen Minimums $\theta_k = 1$ gewählt wird und somit die schnelle lokale Konvergenz des Verfahrens erhalten bleibt.

3.4.6.4 Trust-Region Varianten

Das Teilproblem (SQP) bzw. (SQP+) lassen sich natürlich auch mit Trust-Region Verfahren koppeln. Bei diesen wird zusätzlich eine Schranke $\|d^k_{qp}\| \leq \Delta_k$ eingeführt.

Diese haben den Vorteil, dass Sie auf die Definitheit von H_k verzichten können und somit in der Tat $\nabla^2_{xx}\mathscr{L}$ im Algorithmus verwendet werden kann.

Man beachte, dass hierbei die Schranke $\|d^k_{qp}\| \leq \Delta_k$ wieder zur Unzulässigkeit von (SQP) nicht jedoch von (SQP+) führen kann. Die Schranke an den Radius aber in jedem Fall nicht aufgeweicht werden darf (sonst ist der TR-Gedanke hinfällig).

3.5 Quadratische Optimierungsproblem

Wir betrachten nun ein quadratischen Optimierungsproblem, welchen z.B. durch das SQP-Teilproblem (SQP) gegeben sein könnte. Hierzu sei $c \in \mathbb{R}^n$, $H = H^T \in \mathbb{R}^{n\times n}$ positiv definit, $A \in \mathbb{R}^{n\times m}$, $\alpha \in \mathbb{R}^m$, $B \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$. Dann betrachten wir

$$\begin{aligned}
&\min_{x\in\mathbb{R}^n} q(x) := \frac{1}{2}\langle x, Hx\rangle + \langle c, x\rangle \\
&\text{u.d.N. } \alpha + A^T x \leq 0, \\
&\qquad \beta + B^T x = 0.
\end{aligned} \tag{QP}$$

Bemerkung 3.5.1. Wir beschränken uns hier auf den Fall eines konvexen QP, da in diesem Fall notwendige und hinreichende Bedingungen übereinstimmen. Ferner kann es im nicht konvexen Fall multiple (lokale) Minima geben, wie das Problem

$$\min_{x\in\mathbb{R}^n} -\frac{1}{6}\sum_{l=1}^{n} 2^l((x_l-1)^2-1) \quad \text{u.d.N.} \quad 0\le x_i\le 3 \quad i=1,\dots,n$$

zeigt. Es besitzt in jeder der 2^n Ecken des Würfels $[0,3]^n$ ein isoliertes lokales Minimum mit jeweils verschiedenen Funktionswerten $0,-1,-2,\dots,-2^n+1$.

Da wir H als positiv definit angenommen haben ist das (QP) strikt konvex, und es existiert eine Lösung solange das Problem einen zulässigen Punkt besitzt x_0, denn $q(x)\to\infty$ für $\|x\|\to\infty$ und somit ist die Niveaumenge $N_q(x_0)$ kompakt.

Da das Problem konvex ist können wir uns auf das Lösen der notwendigen Bedingungen erster Ordnung konzentrieren. Hierzu beobachten wir, dass die KKT-Bedingungen für den gleichungsrestringierten Fall gerade

$$\begin{pmatrix} H & B \\ B^T & 0 \end{pmatrix}\begin{pmatrix} \overline{x} \\ \overline{\mu} \end{pmatrix} = \begin{pmatrix} -c \\ -\beta \end{pmatrix}$$

ist. Ist $X\neq\emptyset$ so gibt es eine Lösung dieses Problems, ist darüber hinaus $\operatorname{Rang}(B)=p$, so ist diese eindeutig. Wir können ein gleichungsrestringiertes QP also durch das Lösen eines linearen Gleichungssystems behandeln. Wir werden nun die Lösung von (QP) auf eine Folge von gleichungsrestringierten Problemen zurückführen. Hierzu wählt man im Punkt x^k eine geeignete innere Approximation $\mathscr{A}_k\subset\mathscr{A}(x^k)$, z.B. damit $\operatorname{Rang}(A_{\mathscr{A}_k}\ B)=|\mathscr{A}_k|+p$ gilt, und löst das Problem

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} q(x) := \frac{1}{2}\langle x,Hx\rangle + \langle c,x\rangle \\ &\text{u.d.N. } \alpha_{\mathscr{A}_k} + A_{\mathscr{A}_k}^T x = 0, \\ &\qquad \beta + B^T x = 0 \end{aligned} \tag{QP$_k$}$$

wobei $A_{\mathscr{A}_k}$ die aus den Spalten a_i der Matrix A mit $i\in\mathscr{A}_k$ gebildet wird.

Algorithmus 3.5.2 ((Primale) Strategie der aktiven Mengen)**.**

- Wähle Startpunkt $x^0\in\mathbb{R}^n$ welcher zulässig für (QP) ist. Setzt $\mathscr{A}_0=\mathscr{A}(x^0)$.

for k = 0, 1, ..., **do**

- Setze $\mathscr{I}_k=\mathscr{U}\setminus\mathscr{A}_k$, $\lambda_{\mathscr{I}_k}^{k+1}=0$ und berechne ein KKT-Tripel $(\hat{x}^{k+1},\lambda_{\mathscr{A}_k}^{k+1},\mu^{k+1})$ von (QP$_k$).
- Setze $d^k=\hat{x}^{k+1}-x^k$.
- Ist $d^k=0$ und $\lambda^{k+1}\ge 0$, so setze $x^{k+1}=x^k$. STOPP da $(x^{k+1},\lambda^{k+1},\mu^{k+1})$ ein KKT-Tripel von (QP) ist.

- Ist $d^k = 0$ und gibt es ein $j \in \mathscr{A}_k$ mit $\lambda_j^{k+1} = \min_{i \in \mathscr{A}_k} \lambda_i^{k+1} < 0$,
  so setze $x^{k+1} = x^k$, $\mathscr{A}_{k+1} = \mathscr{A}_k \setminus \{j\}$ und gehe in die nächste Iteration.
- Ist $d^k \neq 0$ und ist $\hat{x}^{k+1}$ zulässig für (QP),
  so setzte $x^{k+1} = \hat{x}^{k+1}$, $\mathscr{A}_{k+1} = \mathscr{A}_k$ und gehe in die nächste Iteration.
- Ist $d^k \neq 0$ und ist $\hat{x}^{k+1}$ unzulässig für (QP), so bestimme

$$t_k = \max\left\{t \geq 0 \,|\, x^k + t d^k \quad \text{zulässig für (QP)}\right\}$$

  sowie einen Index $j \in \mathscr{I}_k$ mit $a_j^T(x^k + t_k d^k) + \alpha_j = 0$.
  Setze $x^{k+1} = x^k + t_k d^k$ und $\mathscr{A}_{k+1} = \mathscr{A}_k \cup \{j\}$.

end for

Theorem 3.5.3. *Für die aktive Mengen Strategie 3.5.2 gilt:*

1. *x^k ist zulässig für (QP_k) und (QP).*
2. *Ist $d^k \neq 0$ und ist $\hat{x}^{k+1}$ nicht zulässig für (QP), so gibt es eine Schrittweite $t_k \in [0,1)$ und einen zugehörigen Index j, so dass der letzte Fall des Algorithmus durchführbar ist. Es ist dann*
$$t_k = \min\left\{-\frac{a_i^T x^k + \alpha_i}{a_i^T d^k} \,\middle|\, i \in \mathscr{I}_k, a_i^T d^k > 0.\right\}.$$
3. *Gilt $d^k = 0$ und ist $\lambda^{k+1} \geq 0$, so ist in der Tat $(\hat{x}^{k+1}, \lambda^{k+1}, \mu^{k+1})$ ein KKT-Tripel von (QP).*
4. *Ist $d^k \neq 0$, so ist $\langle \nabla q(x^k), d^k \rangle < 0$, d.h. d^k ist eine Abstiegsrichtung für q im Punkt x^k. Ferner ist $q(x^{k+1}) < q(x^k)$, falls $x^{k+1} \neq x^k$.*
5. *Zu jedem vom Algorithmus erzeugten x^k gibt es ein $l \geq k$, so dass x^l die eindeutige globale Lösung von (QP_l) ist.*
6. *Terminiert der Algorithmus nicht endlich, dann gibt es $l \geq 0$, mit $x^k = x^l$ für alle $k \geq l$.*
7. *Sind die Spalten der Matrix $(A_{\mathscr{A}_k} B)$ linear unabhängig, so sind die Spalten von $(A_{\mathscr{A}_{k+1}} B)$ ebenfalls linear unabhängig.*

Beweis. 1. Die Zulässigkeit für (QP) folgt sofort aus den Updatevorschriften. Die Zulässigkeit für (QP_k) folgt ebenfalls sofort, da x^{k+1} stets zwischen den für (QP_k) zulässigen Punkten x^k und $\hat{x}^{k+1}$ liegt, und für (QP_{k+1}) entweder $\mathscr{A}_{k+1} \subset \mathscr{A}_k$ oder $\mathscr{A}_{k+1} = A_k \cup \{j\} \subset \mathscr{A}(\hat{x}^{k+1})$ gilt.

2. Da $\hat{x}^{k+1}$ nicht zulässig für (QP) aber zulässig für (QP_k) ist, gibt es (mindestens) ein $i \in \mathscr{I}_k$ mit
$$a_i^T(x^k + d^k) + \alpha_i > 0.$$

Nach 1. ist x^k zulässig für (QP), d.h.

$$a_i^T x^k + \alpha_i \leq 0$$

und somit ist $a_i^T d^k > 0$. Für diejenigen Nebenbedingungen i mit $a_i^T d^k \leq 0$ ist $x^k + td^k$ für alle $t \in [0,1]$ zulässig. Wir können daher t_k als die größte Zahl mit

$$a_i^T (x^k + t_k d^k) + \alpha_i \leq 0 \quad \forall\, i \in \mathscr{I}_k, a_i^T d^k > 0$$

setzen. Damit folgt die behauptete Formel und die Existenz des (nicht notwendig eindeutigen) Index j

3. Nach Definition ist nun $d^k = 0$, $x^{k+1} = x^k$ und $\lambda_{\mathscr{I}_k}^{k+1} = 0$. Nach den KKT-Bedingungen von (QP_k) folgt

$$\nabla q(x^{k+1}) + A\lambda^{k+1} + B\mu^{k+1} = \nabla q(x^{k+1}) + A_{\mathscr{A}_k}\lambda_{\mathscr{A}_k}^{k+1} + B\mu^{k+1} = 0.$$

Ferner ist nach 1. x^{k+1} zulässig für (QP), $\lambda^{k+1} \geq 0$ und

$$\langle \lambda^{k+1}, g(x^{k+1})\rangle = \langle \lambda_{\mathscr{A}_k}^{k+1}, g_{\mathscr{A}_k}(x^{k+1})\rangle = 0.$$

Damit ist wie behauptet $(\hat{x}^{k+1}, \lambda^{k+1}, \mu^{k+1})$ ein KKT-Tripel von (QP).

4. Ist $d^k \neq 0$, so ist $x^k \neq \hat{x}^{k+1}$. Da $\hat{x}^{k+1}$ die eindeutige globale Lösung von (QP_k) ist, folgt

$$q(\hat{x}^{k+1}) < q(x^k).$$

Die Konvexität von q liefert damit

$$\langle \nabla q(x^k), d^k\rangle \leq q(\hat{x}^{k+1}) - q(x^k) < 0.$$

Ist $x^{k+1} \neq x^k$, so ist $x^{k+1} = x^k + t_k d^k$ mit $t_k \in (0,1]$. Damit folgt

$$q(x^{k+1}) - q(x^k) \leq (1-t_k)q(x^k) + t_k q(\hat{x}^{k+1}) - q(x^k) = t_k(q(\hat{x}^{k+1}) - q(x^k)) < 0.$$

5. Für jede Iterierte x^k gilt einer der folgenden Fälle:

   a) $d^k = 0$, so ist $x^k = \hat{x}^{k+1}$ die eindeutige Lösung von (QP_k).

   b) $d^k \neq 0$, $x^{k+1} = \hat{x}^{k+1}$ und $\mathscr{A}_{k+1} = \mathscr{A}_k$. So ist x^{k+1} die eindeutige Lösung von (QP_{k+1}).

   c) $d^k \neq 0$, $x^{k+1} = x^k + t_k d^k$ und $\mathscr{A}_{k+1} = \mathscr{A}_k \cup \{j\}$. Da $\mathscr{U}$ endlich ist, kann dies nur endlich oft hintereinander auftreten.

6. Terminiert der Algorithmus nicht endlich, so gibt es nach 5. eine unendliche Folge l_i, so dass x^{l_i} die eindeutige Lösung von (QP_{l_i}) ist. Da es nur endlich viele Möglichkeiten gibt $\mathscr{A}_{l_i}$ zu wählen können wir o.B.d.A. annehmen $\mathscr{A}_{l_i} = \mathscr{A}_{l_0}$ und damit notwendig $x^{l_i} = x^{l_0}$. Angenommen es gäbe ein $k \geq l_0$, mit $x^k \neq x^{k+1}$, so ergibt sich aus 4. der Widerspruch

$$q(x^{l_0}) \leq q(x^{k+1}) < q(x^k) \leq q(x^{l_0}).$$

Damit ist also $x^k = x^{l_0}$ für alle $k \geq l_0$.

7. Interessant ist nur der Fall $\mathscr{A}_{k+1} \not\subset \mathscr{A}_k$. Dann ist nach Definition des Algorithmus und 2. $d^k \neq 0$ und $\mathscr{A}_{k+1} = \mathscr{A}_k \cup \{j\}$ mit $a_j^T d^k > 0$. Angenommen es wäre $a_j = A_{\mathscr{A}_k} v + Bw$ mit geeigneten v und w. Dann folgt wegen $A_{\mathscr{A}_k}^T d^k = 0$ und $B^T d^k = 0$ der Widerspruch

$$0 < a_j^T d^k = v^T A_{\mathscr{A}_k}^T d^k + w^T B^T d^k = 0.$$

□

Bemerkung 3.5.4.

1. Im Falle eines strikt konvexen QPs lässt sich – unter der Annahme, dass das (QP) nicht degeneriert ist, d.h. für $d^k \neq 0$ stets $t_k \neq 0$ ist – die Existenz von sog. Zyklen ausschließen – da der Wert von q monoton und gelegentlich sogar strikt monoton sein muss ($x^{k+1} \neq x^k$ mindestens alle m-Iterationen). Die aktive Mengenstrategie endet also nach endlich vielen Schritten. Im Allgemeinen kann Degeneriertheit nicht ausgeschlossen werden. In diesem Fall kann durch geeignete Auswahlregeln das Auftreten von Zyklen in $\mathscr{A}_k$ vermieden werden.
2. Da in jedem Schritt die Menge $\mathscr{A}_k$ um höchstens ein Element modifiziert wird, besitzt die aktive Mengen Strategie eine Worst-Case Komplexität von mindestens m Iterationen! Bei der Verwendung von SQP-Verfahren kann dies durch einen sog. *Warmstart* – bekannte Lösung aus dem vorherigen SQP-Teilproblem liefert eine gute Schätzung für $\mathscr{A}$ – in Teilen umgangen werden. Sind jedoch viele Ungleichungen involviert, so sind alternative Verfahren oft hilfreicher, z.B. Penalty-Verfahren um einen guten Startwert zu „raten" oder primal-duale Aktive Mengenstrategien und Projektionsverfahren, in denen mehrerer Indizes simultan aktiviert und inaktiviert werden können.
3. Im Fall nicht konvexer Probleme kann i.A. nicht mit der Modifikation nur eines Index j für $\mathscr{A}$ gearbeitet werden, da hierdurch nicht immer ein Abstieg garantiert werden kann, vgl. [Nocedal and Wright, 1999, Seite 468ff].

3.5.1 Primal-Duale Aktive Mengenstrategie

Ein wesentliches Problem der bisher besprochenen primalen aktiven Mengen Strategie ist, das in jeder Iteration genau ein Index zwischen aktiver und inaktiver Menge wechseln kann. Hierdurch kommt es gerade bei sehr großen Problemen (m groß) zu sehr langen Laufzeiten. Aus diesem Grund gibt es Varianten, bei denen versucht wird die Anzahl der wechselnden Indizes zu vergrößern (bzw. die Menge der Ungleichungsbedingungen zu verkleinern). Hierfür betrachten wir das Quadratische Problem

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \frac{1}{2} \langle x, Hx \rangle + \langle c, x \rangle \\ &\text{u.d.N.} \, B^T x + \beta = 0, \\ &\qquad x \leq 0. \end{aligned} \tag{3.7}$$

Wir zerlegen nun die Menge $\mathcal{U} = \{1, \dots, n\}$ in die (disjunkten) Teilmengen $\mathcal{A}_k \subset A(x^k)$ sowie $\mathcal{F}_k = \mathcal{I}_k \cup \mathcal{U}_k$. Hiermit ergibt sich das Teilproblem

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \frac{1}{2} \langle x, Hx \rangle + \langle c, x \rangle \\ &\text{u.d.N.} \, B^T x + \beta = 0, \\ &\qquad x_{\mathcal{A}_k} = 0, \\ &\qquad x_{\mathcal{U}_k} \leq 0. \end{aligned} \tag{PDA^k}$$

Eine Lösung dieses Problems ergibt ein KKT-Tripel $(x^{k+1}, \lambda^{k+1}_{\mathcal{U}_k}, \mu^{k+1})$ aus dem wie bisher

$$\lambda^{k+1}_{\mathcal{A}_k} = -\left[Hx^{k+1} + c + B\mu^{k+1} \right]_{\mathcal{A}_k}, \qquad \lambda^{k+1}_{\mathcal{I}_k} = 0$$

gesetzt werden kann. Nun kann das Update der aktiven Menge anders als beim primalen Verfahren durch

$$\begin{aligned} \mathcal{A}_{k+1} &= \{ i \in \mathcal{U} \mid x^{k+1}_i > 0, i \in \mathcal{I}_k, \text{ oder } \lambda^{k+1}_i > 0, i \in \mathcal{A}_k \} \\ \mathcal{F}_{k+1} &= \mathcal{U} \setminus \mathcal{A}_{k+1} \end{aligned}$$

bestimmt werden.

Bemerkung 3.5.5. Hierzu einige Bemerkungen Curtis et al. [2015]:

1. Die naheliegende Zerlegung $\mathcal{F}_k = \mathcal{I}_k$, und somit $\mathcal{U}_k = \emptyset$ kann, auch bei positiv definitem H, zu Zyklen ('Kreiseln') des Algorithmus führen. Lediglich für den Fall das H nahezu ein M-Matrix ist, d.h. $H_{i,j} \leq 0$ für $i \neq j$ und $H^{-1}_{i,j} \geq 0$ für alle i, j, lässt sich Konvergenz dieses Verfahrens mit $\mathcal{U}_k = \emptyset$ zeigen. Dies ist insbesondere für die Lösung mancher diskretisierter unendlich dimensionaler Optimierungsprobleme von großem Interesse (→ Vorlesungen 'Optimierung im Funktionenraum' oder 'Optimierung mit partiellen Differentialgleichungen').

2. Für allgemeine Probleme kann das Kreisel z.B. durch ein Monitoring der Indexwechsel verhindert werden. Hierbei definiert man

   $$q^{k+1}_i = q^k_i + 1$$

   falls $i \in \mathcal{A}_{k+1}$ und $i \notin \mathcal{A}_k$ oder $i \in \mathcal{I}_{k+1}$ und $i \notin \mathcal{I}_k$, sowie $q^{k+1}_i = q^k_i$ in allen anderen Fällen. Ist dann $q^k_i \geq q_{\max} \in \mathbb{N}$ für ein i, so setzt man $U_{k+1} = U_k \cup \{i\}$, und $\mathcal{U}_0 = \emptyset$.

   Das hierbei entstehende Problem (PDA^k) ist dann ein quadratisches Problem welches z.B. mit der primalen aktiven Mengen Strategie 3.5.2 gelöst werden kann, wobei hoffentlich $|\mathcal{U}_k| \ll |\mathcal{U}|$ gilt.

3.6 Projektionsverfahren

Ist $X \neq \emptyset$ konvex und abgeschlossen und $f \in C^1$, dann ist die notwendige Optimalitätsbedingung aus Theorem 3.1.3 gerade

$$\nabla f(\overline{x})^T(x-\overline{x}) \geq 0 \quad \forall x \in X, \tag{3.8}$$

wobei hierfür keine CQ erforderlich ist! Nun ist (3.8) für beliebiges $t > 0$ äquivalent zur Bedingung

$$\overline{x} = \mathrm{Proj}_X(\overline{x} - t\nabla f(\overline{x})) \tag{3.9}$$

wobei $\mathrm{Proj}_X : \mathbb{R}^n \to X$ die Bestapproximation auf die Menge X darstellt, d.h.,

$$\|\mathrm{Proj}_X(x) - x\| = \min_{y \in X} \|y - x\|.$$

Dies legt eine Fixpunktiteration (Verallgemeinerung des Gradientenverfahrens) der Form

$$x^{k+1} = \mathrm{Proj}_X(x^k - t_k \nabla f(x^k))$$

nahe. Dabei kann t_k z.B. durch die Armijo-artige Liniensuche $t_k = \max\{\beta^l \mid l = 0, 1, \ldots\}$ mit

$$f(\mathrm{Proj}_X(x^k - t_k\nabla f(x^k))) \leq f(x^k) - \gamma \nabla f(x^k)^T\Big(x^k - \mathrm{Proj}_X(x^k - t_k \nabla f(x^k))\Big)$$

bestimmt werden.

3.7 Barriere-Verfahren

Eine Alternative zu den Strafterm-Verfahren bieten sogenannte Barriere-Verfahren. Bei diesen wird eine innere Approximation der Nebenbedingungen sichergestellt. Man wählt hierzu eine Funktion mit der Eigenschaft:

$$b(x) = \infty \quad x \notin X, \qquad b(x) \to \infty \quad (X \ni x \to \partial X).$$

Die am häufigsten verwendete Barriere Funktion ist

$$b : (-\infty, 0) \to \mathbb{R}, \quad b(t) = -\ln(-t)$$

für die Ungleichung $t \leq 0$. Da Barriere-Methoden für Gleichungsnebenbedingungen nicht verwendet werden können betrachte wir nun lediglich

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{u.d.N.} \quad g(x) \leq 0. \tag{NLP$_g$}$$

Bemerkung 3.7.1. Gleichungsnebenbedingungen können natürlich auch mit einer der bereits besprochenen Varianten behandelt werden (Penalty, oder die Gleichungsnebenbedingungen beibehalten).

Damit wir unsere obige Barriere verwenden können benötigen wir die zusätzliche Annahme, dass es einen strikt zulässigen Punkt (insbesondere inneren) Punkt gibt, d.h.

$$X^\circ := \{x \in \mathbb{R}^n \mid g(x) < 0\} \neq \emptyset.$$

Bemerkung 3.7.2. Die Menge X° ist das *strikte Innere* von X. Es kann sich vom Inneren $\operatorname{int}(X)$ unterscheiden. Beispielsweise sind für $g(x) = \max(0,x)^2$ die Mengen $X = (-\infty, 0]$, $\operatorname{int}(X) = (-\infty, 0)$ aber $X^\circ = \emptyset$.

Wir können nun die *logarithmische Barriere-Funktion* $B_\alpha \colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ durch

$$\begin{aligned} B_\alpha(x) &= f(x) - \alpha \sum_{i \in \mathscr{U}} \ln(-g_i(x)) =: f(x) + \alpha b(x) && x \in X^\circ, \\ B_\alpha(x) &= \infty && x \notin X^\circ \end{aligned}$$

definieren. Ist dann $\alpha_k \to 0$ monoton fallend, so können wir analog zum Penalty-Verfahren 3.3.1 die Barriere-Probleme

$$\min_{x \in \mathbb{R}^n} B_\alpha(x) \tag{B_α}$$

lösen und erhalten damit eine Folge $x^k \subset X^\circ$. Wir bekommen so das folgende Verfahren:

Algorithmus 3.7.3 (Barriere-Verfahren)**.**

- Wähle Startpunkt $x^0 \in X^\circ$ und $\alpha_0 > 0$.

for k = 0, 1, ..., **do**

- Berechne eine globale Lösung x^k des Barriere-Problems (B_α).
  Hierbei verwendet man für $k > 0$ meist x^{k-1} als Startpunkt (nicht optimal!).
  Da B_{α_k} nur auf X° definiert ist benötigt die Lösung von (B_α)
  besondere Vorsicht (insbesondere bei der Schrittweitenwahl).
- Wähle $0 < \alpha_{k+1} < \alpha_k$.

end for

Theorem 3.7.4. *[Globale Konvergenz des Barriere-Verfahrens] Seien f und g stetig und das strikte Innere $X^\circ \neq \emptyset$. Erzeugt das Barriere-Verfahren 3.7.3 eine unendliche Folge, so gilt:*

1. *Die Folge $b(x^k) = -\sum_{i\in\mathcal{U}} \ln(-g_i(x^k))$ ist monoton wachsend.*
2. *Die Folge $f(x^k)$ ist monoton fallend.*
3. *Ist $X = \operatorname{clos} X^\circ$, so ist jeder Häufungspunkt $\overline{x}$ der Folge x^k eine globale Lösung von (NLP_g) und es gilt*

$$f(x^k) \to f(\overline{x}), \quad B_{\alpha_k}(x^k) \to f(\overline{x}).$$

Beweis. 1. Wir addieren die Ungleichungen $B_{\alpha_k}(x^k) \leq B_{\alpha_k}(x^{k+1})$ und $B_{\alpha_{k+1}}(x^{k+1}) \leq B_{\alpha_{k+1}}(x^k)$ und erhalten

$$\alpha_k b(x^k) + \alpha_{k+1} b(x^{k+1}) \leq \alpha_k b(x^{k+1}) + \alpha_{k+1} b(x^k).$$

Da $\alpha_k > \alpha_{k+1}$ folgt $b(x^k) \leq b(x^{k+1})$.

2. Wegen 1. erhalten wir

$$\begin{aligned} 0 &\leq B_{\alpha_{k+1}}(x^k) - B_{\alpha_{k+1}}(x^{k+1}) \\ &= f(x^k) - f(x^{k+1}) + \alpha_{k+1}(b(x^k) - b(x^{k+1})) \\ &\leq f(x^k) - f(x^{k+1}). \end{aligned}$$

3. O.B.d.A können wir die Konvergenz der gesamten Folge annehmen. Angenommen $\overline{x}$ ist keine globale Lösung von (NLP_g). Dann gibt es ein $\hat{x} \in X$ mit $f(\hat{x}) < f(\overline{x})$. Wegen $X = \operatorname{clos} X^\circ$ gibt es eine Folge $X^\circ \ni y^k \to \hat{x}$ und somit wegen der Stetigkeit von f auch ein $y \in X^\circ$ mit $f(y) < f(\overline{x})$. Damit folgt wegen 1.

$$\begin{aligned} f(x^k) + \alpha_k b(x^0) &\leq f(x^k) + \alpha_k b(x^k) \\ &= B_\alpha(x^k) \\ &\leq B_\alpha(y). \end{aligned}$$

Aus 2. erhalten wir damit für beliebiges k

$$\begin{aligned} f(\overline{x}) &= \lim_{l\to\infty} f(x^l) \\ &\leq f(x^k) \\ &\leq f(y) + \alpha_k(b(y) - b(x^0)) \\ &\to f(y) \qquad (k \to \infty) \\ &< f(\overline{x}). \end{aligned}$$

Dies ist ein Widerspruch und somit ist $\overline{x}$ eine globale Lösung.

Es beleibt nur noch $B_{\alpha_k}(x^k) \to f(\overline{x})$ zu zeigen. Sei dazu $\epsilon > 0$ beliebig. Dann gibt es $y_\epsilon \in X^\circ$ mit $f(y_\epsilon) \leq f(\overline{x}) + \epsilon$. Damit folgt aus 1.

$$\begin{aligned} f(\overline{x}) &\overset{k\to\infty}{\longleftarrow} f(x^k) + \alpha_k b(x^0) \\ &\leq f(x^k) + \alpha_k b(x^k) \\ &= B_{\alpha_k}(x^k) \\ &\leq f(y_\epsilon) + \alpha_k b(y_\epsilon) \\ &\overset{k\to\infty}{\longrightarrow} f(y_\epsilon) \\ &\leq f(\overline{x}) + \epsilon. \end{aligned}$$

Da $\epsilon > 0$ beliebig war folgt die Behauptung.

□

Sind die Funktionen f und g differenzierbar, so können wir auch hier analog zu den Penalty-Verfahren Lagrange-Multiplikatoren definieren. Dies geschieht durch die Beobachtung

$$0 = \nabla B_{\alpha_k}(x^k) = \nabla f(x) + \alpha_k \sum_{i \in \mathscr{U}} \frac{-1}{g_i(x^k)} \nabla g_i(x^k)$$

und die hieraus resultierende Setzung

$$\lambda_i^k = \frac{-\alpha_k}{g_i(x^k)}.$$

Analoge Aussagen zum Theorem 3.3.5 zur Konvergenz der Lagrange-Multiplikatoren für das Penalty-Verfahren lassen sich auch für die Barriere-Verfahren zeigen.

3.7.1 Primal-Duale-Innere-Punkte-Verfahren (PDIP)

Sei nun zur Vereinfachung angenommen (NLP_g) wäre konvex und f und g wären differenzierbar. Ist die Lösung von (B_α) eindeutig, so definieren die bisherigen Überlegungen eine Abbildung

$$(0,1) \ni \alpha \mapsto (x(\alpha), \lambda(\alpha)) \in \mathbb{R}^n \times \mathbb{R}^m$$

den *primal-dualen zentralen Pfad*. Hierbei ist $x(\alpha)$ die globale Lösung von (B_α) zum Parameter α, und

$$\lambda(\alpha)$$

die zugehörige Lagrange-Multiplikator Approximation. Das Paar erfüllt dann die gestörten KKT-Bedingungen

$$\begin{aligned} \nabla_x \mathscr{L}(x,\lambda) &= 0, \\ -\lambda_i g_i(x) &= \alpha, \qquad (i \in \mathscr{U}). \end{aligned}$$

Für die Lösung des Problems bietet sich ein globalisiertes Newtonverfahren an, bei dem nach jedem Schritt der Parameter α verkleinert wird. Bei der Globalisierung wie auch der Anpassung von α ist darauf zu achten, dass $(x(\alpha), \lambda(\alpha))$ in einer geeigneten Umgebung des zentralen Pfades bleibt. Hierbei ist insbesondere zu beachten, dass $-\lambda_i g_i(x) \gg 0$ bleibt, so dass typische Umgebungen von der Form

$$N_{-\infty}(\gamma) = \left\{ (x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^m \,|\, x \in X^0, \lambda > 0, -g_i(x)\lambda_i \geq -\gamma \frac{\langle g(x), \lambda \rangle}{m} \right\}$$

für ein $\gamma \in (0,1)$ sind. Hierbei füllt $N_{-\infty}(\gamma)$ für $\gamma \to 0$ den ganzen primal-dual zulässigen Bereich aus.

Für konvexes f, g_i sind X und X° konvex. Dies ist für X unmittelbar klar. Für X° seien $x_1, x_2 \in X^\circ$, $t \in [0,1]$ so gilt

$$g_i((1-t)x_1 + tx_2) \leq (1-t)g_i(x_1) + tg(x_2) < 0, \quad i = 1, \ldots, m.$$

Ferner ist $-\ln(-t)\colon (-\infty, 0) \to \mathbb{R}$ streng monoton wachsend und streng konvex, denn

$$\Big(-\ln(-t)\Big)' = -\frac{1}{t} > 0, \quad \text{und} \quad \Big(-\ln(-t)\Big)'' = \frac{1}{t^2} > 0.$$

Damit ist auch

$$X^\circ \ni x \mapsto -\ln(-g_i(x)))$$

konvex und somit auch $B_\alpha(x)$ auf X° konvex. Ist nun $X^\circ \neq \emptyset$, so folgt $X = \overline{X^\circ}$, denn ist $x_0 \in X^\circ$ und $x \in X$ beliebig, so ist

$$x_t = x + t(x_0 - x) \in X \quad \forall\, t \in (0,1]$$

und

$$g_i(x_t) \leq (1-t)g_i(x) + t\,g_i(x_0) \leq t\,g_i(x_0) < 0 \quad \forall\, t \in (0,1].$$

Also gilt $x_t \in X^\circ$ und $x_t \to x$ für $t \to 0$ nach Definition.

Lemma 3.7.5. *Sei der zulässige Bereich X von (NLP_g) so beschrieben, dass $X^\circ \neq \emptyset$ ist. Ferner sei die Lösungsmenge X_{opt} von (NLP_g) nicht-leer und beschränkt sowie f konvex. Dann ist für alle $\nu \in \mathbb{R}$ die Niveaumenge*

$$X(\nu) = \{x \in X \,|\, f(x) \leq \nu\}$$

beschränkt.

Beweis. Angenommen es gäbe ein $\nu \in \mathbb{R}$ für das $X(\nu)$ unbeschränkt wäre. Sei dann $\overline{x} \in X_{\text{opt}}$ beliebig. Da X_{opt} kompakt ist gibt es ein $r > 0$, so dass

$$\underbrace{X \cap \delta B_r(\overline{x})}_{=:M_r} \cap X_{\text{opt}} = \emptyset.$$

Sei nun $z_j \in X(\nu)$ mit $r < \|z_j - \overline{x}\| \to \infty$ und setze

$$t_j = \frac{r}{\|z_j - \overline{x}\|} \in (0,1), \quad \text{und} \quad \overline{z}_j := \overline{x} + t_j(z_j - \overline{x}).$$

Nach Definition gilt $t_j \to 0$ und $\overline{z}_j \in M_r$, also ist $M_r \neq \emptyset$ und f nimmt auf der kompakten Menge M_r ein Minimum in $x^* \in M_r$ an. Da $x^* \in X \setminus X_{\text{opt}}$ liegt folgt $f(x^*) > f(\overline{x})$. Damit erhalten wir aus der Konvexität von f die Abschätzung

$$\begin{aligned} f(x^*) &\leq f(\overline{z}_j) \\ &= f((1-t_j)\overline{x} + t_j z_j) \\ &\leq (1-t_j) f(\overline{x}) + t_j f(z_j) \\ &\leq (1-t_j) f(\overline{x}) + t_j \nu. \end{aligned}$$

Dies ergibt

$$\nu - f(\overline{x}) \geq \frac{f(x^*) - f(\overline{x})}{t_j} \to \infty$$

und somit einen Widerspruch. □

Theorem 3.7.6. *Seien $f, g_i \in C^0(\mathbb{R}^n;\mathbb{R})$ konvex, $X^\circ \neq \emptyset$ und α_k streng monoton fallende Nullfolge von Barriere-Parametern. Ist die Lösungsmenge X_{opt} von (NLP$_g$) nicht-leer und beschränkt, so gilt*

1. *Für alle $\alpha > 0$ ist B_α konvex auf X° und für jedes $x^0 \in X^\circ$ ist die Niveaumenge*

$$N(\alpha, x^0) = \{x \in \mathbb{R}^n \,|\, B_\alpha(x) \leq B_\alpha(x^0)\}$$

kompakt und konvex.

2. *Für jedes $\alpha > 0$ hat das Barriere Problem (B$_\alpha$) eine nicht-leere, kompakte und konvexe Lösungsmenge $\Omega(\alpha)$.*

3. *Es gibt $\Omega \subset \mathbb{R}^n$ kompakt mit $\Omega(\alpha_k) \subset \Omega$ für alle $k \in \mathbb{N}$. Insbesondere besitzt die von Algorithmus 3.7.3 erzeugte Folge x^k mindestens einen Häufungspunkt $\overline{x}$. Jeder Häufungspunkt $\overline{x}$ der Folge ist Lösung von (NLP$_g$) und es gilt*

$$\lim_{k\to\infty} f(x^k) = f(\overline{x}), \qquad \lim_{k\to\infty} B_{\alpha_k}(\overline{x}) = f(\overline{x})$$

mit monoton fallenden Werten $f(x^k)$.

Beweis. 1. Wir haben schon gesehen, dass X° und B_α konvex sind. Damit ist auch für jedes $x^0 \in X^\circ$ die Niveaumenge $N(\alpha, x^0)$ konvex.

Für die Kompaktheit beachten wir, dass B_α auf X° stetig ist und somit die Niveaumenge $N(\alpha, x^0) \subset X^\circ$ abgeschlossen.

Es bleibt noch die Beschränktheit zu sehen. Sei hierzu zum Widerspruch angenommen es gäbe eine Folge $y^j \in N(\alpha, x^0)$ mit $\|y^j\| \to \infty$. O.B.d.A (Teilfolge) können wir annehmen, dass

$$d = \lim_{j\to\infty} \frac{y^j - x^0}{\|y^j - x_0\|}$$

existiert. Dann ist $\|d\| = 1$. Ferner ist für $0 \le t < \|y^j - x^0\|$

$$x^0 + \frac{t}{\|y^j - x^0\|} \in N(\alpha, x^0)$$

da $x^0, y^j \in N(\alpha, x^0)$ und $N(\alpha, x^0)$ konvex. Wegen der Abgeschlossenheit von $N(\alpha, x^0)$ ist somit auch

$$x^0 + td = \lim_{j\to\infty} x^0 + t\frac{y^j - x^0}{\|y^j - x_0\|} \in N(\alpha, x^0) \tag{3.10}$$

für jedes $t \ge 0$. Nun ist aber nach Lemma 3.7.5 die Niveaumenge $X(f(x^0)+1)$ beschränkt. Damit gibt es ein $\bar{t} > 0$, mit $x^0 + td \notin X(f(x^0)+1)$ für alle $t \ge \bar{t}$, aber $x^0 \in X(f(x^0)+1)$. Damit erhalten wir für die konvexe Funktion

$$\psi : \mathbb{R} \to \mathbb{R}; \quad \psi(t) = f(x^0 + td)$$

die Ungleichung $\psi(\bar{t}) \ge f(x^0) + 1 > f(x^0) = \psi(0)$ und somit

$$\gamma := \frac{\psi(\bar{t}) - \psi(0)}{\bar{t}} > 0.$$

Für jedes $t \ge \bar{t}$ ist jetzt $\bar{t} = (1-\sigma)0 + \sigma t$ mit $\sigma = \frac{\bar{t}}{t} \in (0, 1]$ und somit folgt

$$\psi(\bar{t}) = \psi((1-\sigma)0 + \sigma t) \le (1-\sigma)\psi(0) + \sigma\psi(t).$$

Wir erhalten hieraus eine Wachstumsabschätzung für f, und zwar

$$\begin{aligned} f(x^0 + td) = \psi(t) & \\ &\ge \frac{\psi(\bar{t})}{\sigma} - \psi\frac{1-\sigma}{\sigma} \\ &= \psi(\bar{t}) + \big(\psi(\bar{t}) - \psi(0)\big)\frac{1-\sigma}{\sigma} \\ &= \psi(\bar{t}) + \big(\psi(\bar{t}) - \psi(0)\big)\frac{t-\bar{t}}{\bar{t}} \\ &= \psi(\bar{t}) + \gamma(t - \bar{t}). \end{aligned}$$

Damit wächst $f(x^0 + td)$ mindestens linear für $t \ge \bar{t}$. Analog folgt aus der Konvexität von $\psi_i(t) := g_i(x^0 + td)$ für $t \ge \bar{t}$ die Abschätzung

$$\begin{aligned} \psi_i(t) &\ge \psi_i(\bar{t}) + \frac{\psi_i(\bar{t}) - \psi_i(0)}{\bar{t}}(t - \bar{t}) \\ &=: \xi_i + \gamma_i(t - \bar{t}). \end{aligned}$$

Da $x^0 + td \in N(\alpha, x^0) \subset X^\circ$ liegt ist $\psi_i(t) < 0$ für alle $t \geq 0$ und es folgt $\xi_i < 0$ sowie $\gamma_i \leq 0$. Die Abbildung $s \mapsto -\ln(-s)$ ist monoton wachsend, und daher folgt für $t \geq \overline{t}$

$$\begin{aligned} B_\alpha(x^0 + td) &= \psi(t) - \alpha \sum_{i \in \mathscr{U}} \ln(-\psi_i(t)) \\ &\geq \psi(\overline{t}) + \gamma(t - \overline{t}) - \alpha \sum_{i \in \mathscr{U}} \ln(-\xi_i - \gamma_i(t - \overline{t})) \\ &=: I \end{aligned}$$

Nun ist $\gamma_i \leq 0$, $\xi_i < 0$ und es folgt

$$\lim_{t \to \infty} \frac{\ln(-\xi_i - \gamma_i(t - \overline{t}))}{t - \overline{t}} = 0.$$

Wegen $\gamma > 0$ folgt damit $I \to \infty$ und damit

$$B_\alpha(x^0 + td) \to \infty$$

im Widerspruch zu (3.10); somit ist $N(\alpha, x^0)$ auch beschränkt.

2. Ist $x^0 \in X^\circ$ beliebig, so folgt $x^0 \in N(\alpha, x^0)$. Damit ist, nach 1.) die Menge $N(\alpha, x^0)$ nicht-leer und kompakt. Daher hat die stetige Funktion B_α auf $N(\alpha, x^0)$ ein Minimum x^*, und daher ist $\Omega(\alpha)$ nicht-leer, und es gilt $\Omega(\alpha) = N(\alpha, x^*)$ und somit ist $\Omega(\alpha)$ nach 1. konvex und kompakt.

3. Sei nun $x^0 \in \Omega(\alpha_0)$. Dann ist nach Theorem 3.7.4-2. für jedes $x^k \in \Omega(\alpha_k)$, $k \geq 1$, der Funktionswert $f(x^k) \leq f(x^0)$ und damit $\Omega(\alpha_k) \subset X(x^0)$ für alle $k \geq 1$. Nach Lemma 3.7.5 ist $X(x^0)$ kompakt, und damit hat die Folge x^k mindestens einen Häufungspunkt $\overline{x}$.

   Alle weiteren Aussagen sind bereits in Theorem 3.7.4 bewiesen.

□

Literaturverzeichnis

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

F. E. Curtis, Z. Han, and D. P. Robinson. A globally convergent primal-dual active-set framework for large-scale convex quadratic optimization. *Comput. Optim. Appl.*, 60(2):311–341, 2015.

M. Fukushima. A successive quadratic programming algorithm with global and superlinear convergence properties. *Math. Program.*, 35(3):253–264, 1986.

C. Geiger and C. Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer Berlin Heidelberg, 1999.

C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer Berlin Heidelberg, 2002.

P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, 1981.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999.

M. Ulbrich and S. Ulbrich. *Nichtlineare Optimierung*. Mathematik Kompakt. Birkhäuser Basel, 2012.