Nichtlineare Optimierung

Vorlesungsskriptum

Stefan Ulbrich

Fachbereich Mathematik, AG 10
Technische Universität Darmstadt

Wintersemester 2007/2008

Das Skript ist in Zusammenarbeit mit Prof. Michael Ulbrich (TU München) entstanden und basiert auf Vorlesungen von M. Ulbrich an der Universität Hamburg und TU München sowie von S. Ulbrich an der TU München und TU Darmstadt.

Inhaltsverzeichnis

1 Einführung **1**

1.1 Problemstellung 1

1.2 Beispiele 3

1.3 Notationen 6

2 Optimierung ohne Nebenbedingungen **8**

2.1 Optimalitätsbedingungen 8

2.1.1 Notwendige Optimalitätsbedingungen 9

2.1.2 Hinreichende Optimalitätsbedingung 11

2.2 Konvexität 12

2.3 Grundkonzept von Abstiegsverfahren 15

2.4 Das Gradientenverfahren 17

2.4.1 Richtung des steilsten Abstiegs 17

2.4.2 Die Schrittweitenregel von Armijo 18

2.4.3 Globale Konvergenz des Gradientenverfahrens 20

2.4.4 Konvergenzgeschwindigkeit für quadratische Zielfunktion 21

2.5 Konvergenztheorie allgemeiner Abstiegsverfahren 25

2.5.1 Zulässige Suchrichtungen 26

2.5.2 Zulässige Schrittweiten 28

2.5.3 Ein globaler Konvergenzsatz 30

2.6 Schrittweitenregeln 31

2.6.1 Die Armijo-Regel 31

2.6.2 Die Powell-Wolfe-Regel 33
2.7 Das Newton-Verfahren 36
2.7.1 Das Newton-Verfahren für Gleichungssysteme 36
2.7.2 Superlineare und quadratische lokale Konvergenz des Newton-Verfahrens 37
2.7.3 Das Newton-Verfahren für Minimierungsprobleme 41
2.7.4 Globalisierung des Newton-Verfahrens 43
2.7.5 Übergang zu schneller lokaler Konvergenz 46
2.8 Newton-artige Verfahren 48
2.8.1 Superlineare Konvergenz und Dennis-Moré-Bedingung 49
2.8.2 Globalisierung Newton-artiger Verfahren 51
2.8.3 Dennis-Moré-artige Bedingung für quadratische Konvergenz 54
2.9 Inexakte Newton-Verfahren 55
2.9.1 Ein lokaler Konvergenzsatz 57
2.9.2 Zusammenhang von Newton-artigen Verfahren und inexakten Newton-Verfahren 59
2.10 Quasi-Newton-Verfahren 60
2.10.1 Updates minimaler Änderung als Grundprinzip bei der Konstruktion von Quasi-Newton-Updates 63
2.10.2 Wichtige Quasi-Newton-Updates 64
2.10.3 Eigenschaften von DFP-, BFGS- und PSB-Update 67
2.10.4 Globale Konvergenz des BFGS-Verfahrens 69
2.10.5 Lokale Konvergenzaussagen 73
2.11 Richtungen negativer Krümmung 75
2.12 Trust-Region-Verfahren 76
2.12.1 Motivation von Trust-Region-Verfahren 76
2.12.2 Globale Konvergenz 79
2.12.3 Übergang zu schneller lokaler Konvergenz 83
2.12.4 Charakterisierung der Lösung des Trust-Region-Problems 85
2.12.5 Näherungsweise Lösung des Trust-Region-Problems 86

3 Optimierung mit Nebenbedingungen **89**

3.1 Einführung 89

3.2 Notwendige Optimalitätsbedingungen 90

3.2.1 Die Karush-Kuhn-Tucker-Bedingungen 90

3.2.2 Constraint Qualifications 94

3.2.3 Konvexe Probleme und die KKT-Bedingungen 98

3.3 Hinreichende Optimalitätsbedingungen 98

3.3.1 Beweis des Lemmas von Farkas 100

3.4 Penalty-Verfahren 102

3.4.1 Das quadratische Penalty-Verfahren 102

3.4.2 Exakte Penalty-Verfahren 106

3.5 Innere-Punkte-Verfahren 107

3.5.1 Konvergenzeigenschaften des Innere-Punkte-Verfahrens 110

3.5.2 Anwendung des Newton-Verfahrens auf das Barriereproblem 113

3.6 Sequential Quadratic Programming Verfahren 114

3.6.1 Lagrange-Newton- und lokales SQP-Verfahren 114

3.6.2 SQP-Verfahren für Probleme mit Ungleichungsrestriktionen 117

3.6.3 Globalisiertes SQP-Verfahren 119

3.6.4 Probleme beim SQP-Verfahren und mögliche Lösungen 122

3.6.5 BFGS-Updates für SQP-Verfahren 126

3.7 Lösung quadratischer Optimierungsprobleme 126

3.8 Dualität 129

3.8.1 Das duale Problem 129

3.9 Augmented-Lagrange-Verfahren (Ergänzung) 132

3.9.1 Motivation des Augmented-Lagrange-Verfahrens 132

3.9.2 Globale Konvergenz 135

3.9.3 Lokale Konvergenz 136

Kapitel 1

Einführung

1.1 Problemstellung

Die Vorlesung behandelt die Theorie und iterative numerische Lösung von endlichdimensionalen nichtlinearen stetigen Optimierungsproblemen der Form

$$(\mathrm{P}_Z) \qquad \min f(x) \quad \text{u.d. Nebenbedingung} \quad x \in Z.$$

Hierbei ist $Z \subset \mathbb{R}^n$ der *zulässige Bereich*, $x = (x_1, \ldots, x_n)^T \in Z$ ein Vektor von zu optimierenden Parametern und $f : Z \to \mathbb{R}$ eine zumindest stetige *Zielfunktion*. Die Bedingung $x \in Z$ heißt *Nebenbedingung* des Optimierungsproblems. Natürlich können Maximierungsprobleme durch Verwendung der Zielfunktion $-f$ statt f als Minimierungsproblem geschrieben werden.

In der Regel wird der zulässige Bereich durch Gleichungen und Ungleichungen beschrieben, also

$$(1.1) \qquad Z = \{x \in \mathbb{R}^n \ : \ h(x) = 0, \ \ c(x) \leq 0\}$$

mit stetigen Funktionen $h = (h_1, \ldots, h_p)^T : \mathbb{R}^n \to \mathbb{R}^p$ und $c = (c_1, \ldots, c_m)^T : \mathbb{R}^n \to \mathbb{R}^m$. Hierbei ist die Ungleichung $c(x) \leq 0$ komponentenweise zu verstehen, also

$$c(x) \leq 0 :\Longleftrightarrow c_i(x) \leq 0, \quad i = 1, \ldots, m.$$

Wir unterscheiden die folgenden Problemklassen:

Ist $Z = \mathbb{R}^n$ in (P_Z), dann liegen keine Nebenbedingungen vor und wir erhalten ein

Unrestringiertes Optimierungsproblem:

$$(\mathrm{P}) \qquad \min_{x \in \mathbb{R}^n} \ f(x).$$

Ist Z von der Form (1.1), dann ergibt sich ein

Nichtlineares Optimierungsproblem (NLP):

(NLP) $\quad \min_{x\in\mathbb{R}^n} f(x) \quad$ u.d. Nebenbedingung $\quad h(x) = 0, \quad c(x) \le 0.$

Bemerkung: Offensichtlich ist wegen

$$c(x) \ge 0 \Longleftrightarrow -c(x) \le 0$$

die Verwendung von Ungleichungen der Form $c(x) \le 0$ keine Einschränkung. $\square$

Wir können (NLP) weiter klassifizieren:

- Treten keine Ungleichungsnebenbedingungen auf, ist also $m = 0$, dann liegt ein *gleichungsrestringiertes Optimierungsproblem* vor.
- Sind alle Funktionen (affin) linear, also

  $$f(x) = g^T x, \quad c(x) = Ax - b, \quad h(x) = Bx - d$$

  mit $g \in \mathbb{R}^n$, $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$, $B \in \mathbb{R}^{p,n}$, $d \in \mathbb{R}^p$, dann ist (NLP) ein *lineares Optimierungsproblem* (LP).
- Ist die Zielfunktion quadratisch, also

  $$f(x) = g^T x + \frac{1}{2} x^T H x$$

  mit $g \in \mathbb{R}^n$, $H \in \mathbb{R}^{n,n}$ symmetrisch, und sind die Nebenbedingungen linear, also $c(x) = Ax - b$, $h(x) = Bx - d$, dann ist (NLP) ein *quadratisches Optimierungsproblem* (QP).
- Sind f und c_i, $i = 1, \ldots, m$ konvex und ist h linear, dann ist (NLP) ein *konvexes Optimierungsproblem.*

Wir führen zunächst einige Grundbegriffe ein.

Definition 1.1.1 *a) Ein Punkt $x \in \mathbb{R}^n$ heißt* zulässig *für das Problem (P_Z), falls $x \in Z$ gilt.*

b) Ein Punkt $\bar{x} \in \mathbb{R}^n$ heißt lokales Minimum *oder* lokale Lösung *von (P_Z), falls $\bar{x} \in Z$ gilt und es $\varepsilon > 0$ gibt mit*

$$f(x) \ge f(\bar{x}) \quad \forall\, x \in Z \cap B_\varepsilon(\bar{x}).$$

Hierbei bezeichne

$$B_\varepsilon(\bar{x}) := \{x \in \mathbb{R}^n \;:\; \|x - \bar{x}\| < \varepsilon\}$$

die ε-Kugel um $\bar{x}$ mit der euklidischen Norm $\|x\| := \sqrt{x^T x}$.

$\bar{x} \in \mathbb{R}^n$ heißt isoliertes lokales Minimum *von (P_Z), falls $\bar{x} \in Z$ gilt und es $\varepsilon > 0$ gibt mit*

$$f(x) > f(\bar{x}) \quad \forall\, x \in (Z \cap B_\varepsilon(\bar{x})) \setminus \{\bar{x}\}.$$

c) *Ein Punkt $\bar{x} \in \mathbb{R}^n$ heißt* globales Minimum *von (P_Z), falls $\bar{x} \in Z$ gilt und*

$$f(x) \geq f(\bar{x}) \quad \forall\, x \in Z.$$

$\bar{x} \in \mathbb{R}^n$ heißt isoliertes globales Minimum *von (P_Z), falls $\bar{x}$ ein globales und zudem ein isoliertes lokales Minimum ist.*

d) *Zu $x_0 \in Z$ heißt*

$$N_f(x_0) := \{x \in Z \;:\; f(x) \leq f(x_0)\}$$

die Niveaumenge von (P_Z) zu x_0.

Die Existenz eines globalen Minimums läßt sich unter recht allgemeinen Voraussetzungen sicherstellen.

Satz 1.1.2 *Es sei $Z \subset \mathbb{R}^n$ nicht leer und $f : Z \to \mathbb{R}$ stetig. Existiert ein $x_0 \in Z$, so dass die zugehörige Niveaumenge $N_f(x_0) := \{x \in Z \;:\; f(x) \leq f(x_0)\}$ kompakt ist, dann besitzt (P_Z) ein globales Minimum $\bar{x}$.*

Beweis: Natürlich kommen für das globale Minimum von (P_Z) nur Punkte $x \in N_f(x_0)$ in Betracht. Nach dem Satz von Weierstraß nimmt die stetige Funktion f auf dem Kompaktum $N_f(x_0)$ ihren Minimalwert in einem Punkt $\bar{x} \in N_f(x_0)$ an und $\bar{x}$ ist auch globales Minimum auf Z. □

Im nichtkonvexen Fall kann (P_Z) viele lokale Minima besitzen. Die Bestimmung des globalen Minimums (dies ist Gegenstand der Globalen Optimierung) kann beliebig aufwendig sein und ist nur für gewisse Problemklassen effizient durchführbar. Wir beschäftigen uns mit Algorithmen, die lokale Minima von (P_Z) bestimmen.

1.2 Beispiele

Die Optimierung von Vorgängen ist ein grundlegendes Anliegen der Menschheit, sei es zur Maximierung von Kapitalerträgen (z.B. Portfoliooptimierung), zur Optimierung von Produktionsprozessen, zur Verbesserung von medizinischen Therapien, ... Darüberhinaus führt auch die Simulation physikalischer, biologischer oder chemischer Vorgänge häufig auf Optimierungsprobleme, da sich als stabile Zustände in der Natur (lokale) Energieminima einstellen.

Wir geben im Folgenden ein paar Beispiele für Optimierungsprobleme an.

Beispiel 1.2.1 Portfoliooptimierung
Ein Investor möchte einen Betrag $B > 0$ so in ein Portfolio aus n Aktien investieren, dass die erwartete Rendite mindestens ρ und das Risiko minimal ist. Bezeichne r_i die Rendite der i-ten Aktie nach einem Jahr (dies ist eine Zufallsvariable) und $x \in \mathbb{R}^n$ mit

$$\sum_{i=1}^{n} x_i = 1, \quad x \geq 0,$$

die Anteile der Aktien am Portfolio (der Anleger investiert $x_i B$ in Aktie i), dann ist die Rendite des Portfolios

$$R(x) = r^T x, \quad r = (r_1, \ldots, r_n)^T.$$

Wir nehmen an, dass der Zufallsvektor $r = (r_1, \ldots, r_n)^T$ den Erwartungswert $\mu \in \mathbb{R}^n$ und die Kovarianzmatrix $\Sigma \in \mathbb{R}^{n,n}$ habe. Dann ist die erwartete Rendite des Portfolios

$$E(R(x)) = \mu^T x$$

und seine Varianz

$$V(R(x)) = x^T \Sigma x.$$

Suchen wir nun das Portfolio mit erwarteter Rendite $\geq \rho$, das minimale Varianz hat, so führt dies auf das Optimierungsproblem

$$\min \ x^T \Sigma x \quad \text{u. d. Nebenbedingung} \quad \sum_{i=1}^{n} x_i = 1, \quad x \geq 0, \quad \mu^T x \geq \rho.$$

Dies ist ein konvexes quadratisches Optimierungsproblem.

Suchen wir alternativ das Portfolio mit Varianz $\leq \nu$, das die maximale erwartete Rendite hat, so erhalten wir das Optimierungsproblem

$$\max \ \mu^T x \quad \text{u. d. Nebenbedingung} \quad \sum_{i=1}^{n} x_i = 1, \quad x \geq 0, \quad x^T \Sigma x \leq \nu.$$

Dies ist ein konvexes Optimierungsproblem mit linearen und quadratischen Nebenbedingungen.

Beispiel 1.2.2 Regression, Parameterschätzung, Data Assimilation
Ein (physikalischer, technischer, wirtschaftlicher, ...) Vorgang liefere zu Eingangsgrößen $u \in \mathbb{R}^r$ eine Systemantwort $y \in \mathbb{R}^s$. Das Systemverhalten soll durch einen parameterabhängigen Ansatz $u \mapsto g(u; x)$ mit Parametern $x \in Z \subset \mathbb{R}^n$ approximiert werden. Anhand von Messungen y_i zu Eingangsgrößen u_i, $i = 1, \ldots, N$, sollen hierzu die Parameter $x \in Z$ so gewählt werden, dass $g(u_i; x)$ möglichst gut mit den Messungen y_i übereinstimmen. Die Bedeutung von "möglichst gut" kann durch Verwendung einer geeigneten Norm festgelegt werden.

Bei der *Methode der kleinsten Quadrate* verwendet man die euklidische Norm und bestimmt x als Lösung des Minimierungsproblems

$$\min \quad \sum_{i=1}^{N} \|y_i - g(u_i; x)\|^2 \quad \text{u. d. Nebenbedingung} \quad x \in Z.$$

Im Fall $Z = \mathbb{R}^n$ ergibt sich das klassische Problem der *Nichtlinearen Regression*.

Anwendungsbeispiele: Computertomographie, Seismische Inversion zum Auffinden von Bodenschatzen, Kalibirierung von Wettermodellen zur Wettervorhersage anhand von meteorologischen Messdaten, Bestimmung der Volatilitt von Wertpapieren anhand von Marktdaten, . . .

Beispiel 1.2.3 Data Mining, Support Vector Machine
Die Klassifikation von Daten auf Basis von Lernmustern ist heutzutage von immenser Bedeutung und der Entwurf guter Klassifikationsmechanismen kann viel Geld einbringen. Typische Beispiele sind SPAM-Filter (SPAM-email oder nicht), Krebsprognose (Heilung aussichtsreich oder nicht), DNA-Sequencing (Unterscheidung von Promoter-Sequenzen, die den Beginn von Genen markieren, und Nonpromoter-Sequenzen).

In den letzten Jahren haben sich *Support Vector Machines* als sehr erfolgreich bei der Klassifizierung von Daten erwiesen. Gegeben seien m_+ Trainingsdaten $x_+^{(i)} \in \mathbb{R}^n$, $1 \le i \le m_+$, der Klasse K_+ (z.B. SPAM-email) und $x_-^{(i)} \in \mathbb{R}^n$, $1 \le i \le m_-$, der Klasse K_- (keine SPAM-email). Zu einem Datenpunkt $x \in \mathbb{R}^n$ soll nun entschieden werden, ob er zur Klasse K_+ oder K_- gehört. Support Vector Machines versuchen hierzu, die Mengen

$$M_+ = \left\{ x_+^{(i)} \;:\; 1 \le i \le m_+ \right\}, \quad M_- = \left\{ x_-^{(i)} \;:\; 1 \le i \le m_- \right\}$$

durch zwei Hyperbenen

$$H_+ : w^T x = \gamma + 1, \quad H_- : w^T x = \gamma - 1$$

mit möglichst großem Abstand – gegeben durch $\frac{2}{\|w\|}$, $\|w\| = \sqrt{w^T w}$ (warum?) – zu trennen, also

$$\begin{aligned} &w^T x \ge \gamma + 1 \quad \forall\, x \in M_+ \\ &w^T x \le \gamma - 1 \quad \forall\, x \in M_-, \\ &\|w\| \;\; \text{minimal.} \end{aligned}$$

Die Klassifikation eines neuen Datenpunktes erfolgt nun danach, auf welcher Seite der Hyperbene $H : w^T x = \gamma$ der Punkt liegt:

$$x \;\; \text{gehört zu Klasse} \;\; \begin{cases} K_+ & \text{falls } w^T x \ge \gamma \\ K_- & \text{falls } w^T x \le \gamma \end{cases}$$

Die Sicherheit der Klassifikation kann aus der Größe von $w^T x$ abgelesen werden.

Da eine vollständige Trennung von M_+ und M_- unmöglich sein kann, gestattet man in diesem Fall eine unvollständige Trennung durch einen sogenannten "soft margin". Dies führt auf folgende *standard support vector machine*, ein konvexes quadratisches Optimierungsproblem:

$$\begin{aligned}
&\min_{(w,\gamma,y)\in\mathbb{R}^{n+1+m_++m_-}} \nu \sum_{i=1}^{m_++m_-} y_i + \frac{1}{2} w^T w \\
&\text{u. d. Nebenbedingung} \quad w^T x_+^{(i)} + y_i \geq \gamma + 1, \quad 1 \leq i \leq m_+, \\
&\qquad\qquad\qquad\qquad\qquad w^T x_-^{(i)} - y_{i+m_+} \leq \gamma - 1, \quad 1 \leq i \leq m_-, \\
&\qquad\qquad\qquad\qquad\qquad y \geq 0.
\end{aligned}$$

Hierbei ist ν ein Strafparameter, der den Trennungsfehler y bestraft.

Oft werden die tatsächlichen Datenpunkte $z \in \mathbb{R}^N$ durch eine nichtlineare Abbildung $\Phi : z \mapsto x = \Phi(z) \in \mathbb{R}^n$ zunächst in den sogenannten *Feature space* transformiert und dann getrennt.

Beispiel 1.2.4 Optimale Plazierung von Komponenten
Unter anderem bei der Anordnung von Funktionsmodulen auf einem Mikroprozessorchip sollten Module, die durch Signalleitungen verbunden sind, möglichst nahe beeinander liegen, um die Signallaufzeiten minimal zu halten. Seien die Module der Einfachheit halber Kreise mit Mittelpunkt $(x_i, y_i) \in \mathbb{R}^2$ und Radius r_i, $1 \leq i \leq n$. Die Kantenmenge $E \subset \{1, \ldots, n\} \times \{1, \ldots, n\}$ gebe an, welche Module miteinander verbunden sind und zu jeder Kante $e = (i, j) \in E$ existiere ein Gewicht $w_{ij} \geq 0$, das die Wichtigkeit der Verbindung von Modul i mit Modul j angibt (z.B. Zahl der Verbindungen). Eine sinnvolle Plazierung ergibt sich durch Minimierung der gewichteten Abstände unter der Nebenbedingung, dass sich die Module nicht überlappen:

$$\begin{aligned}
&\min_{x,y\in\mathbb{R}^n} \sum_{(i,j)\in E} w_{ij} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \\
&\text{u. d. Nebenbedingung} \quad (x_i - x_j)^2 + (y_i - y_j)^2 \geq (r_i + r_j)^2, \quad 1 \leq i, j \leq n.
\end{aligned}$$

Verwendet man für die Module andere Geometrien (z.B. Rechtecke), dann ergeben sich etwas kompliziertere Nebenbedingungen.

1.3 Notationen

Wir fassen $x \in \mathbb{R}^n$ grundsätzlich als Spaltenvektor auf. Das euklidische Skalarprodukt zweier Vektoren $x, y \in \mathbb{R}^n$ ist dann gegeben durch $x^T y$. Wir bezeichnen die euklidische

Norm mit $\|\cdot\|$, also

$$\|x\| = \sqrt{x^T x} = \sqrt{\sum_{i=1}^n x_i^2}.$$

Für Matrizen $M \in \mathbb{R}^{m,n}$ verwenden wir die induzierte Norm

$$\|M\| := \max_{\|x\| \le 1} \|Mx\|.$$

Dann gilt $\|Mx\| \le \|M\|\|x\|$.

Zu $\varepsilon > 0$ bezeichnen wir die ε-Kugel um $\bar{x} \in \mathbb{R}^n$ mit $B_\varepsilon(\bar{x})$.

Ist $f : \mathbb{R}^n \to \mathbb{R}$ differenzierbar, dann bezeichnen wir mit

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \in \mathbb{R}^n$$

den Gradienten von f in x. Ist f zweimal differenzierbar, dann bezeichnen wir mit

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{pmatrix} \in \mathbb{R}^{n,n}$$

die Hessematrix von f in x. Ist f zweimal stetig differenzierbar, dann ist $\nabla^2 f(x)$ bekanntlich symmetrisch.

Ist $c : \mathbb{R}^n \to \mathbb{R}^m$ differenzierbar, dann bezeichnet

$$c'(x) := \begin{pmatrix} \nabla c_1(x)^T \\ \vdots \\ \nabla c_m(x)^T \end{pmatrix} \in \mathbb{R}^{m,n}$$

dir Jacobi-Matrix (oder Funktionalmatrix) von c. Weiter setzen wir

$$\nabla c(x) := c'(x)^T.$$

Wir verwenden häufig die Landauschen Symbole $O(h^k)$ und $o(h^k)$, $k \in \mathbb{N}$. Die Schreibweisen

$$g(s) = O(\|s\|^k) \quad \text{für } s \to 0 \quad \text{bzw.} \quad g(s) = o(\|s\|^k) \quad \text{für } s \to 0$$

mit einer Funktion $g : \mathbb{R}^n \to \mathbb{R}^m$ bedeuten

$$\limsup_{s \to 0} \frac{\|g(s)\|}{\|s\|^k} < \infty \quad \text{bzw.} \quad \lim_{s \to 0} \frac{\|g(s)\|}{\|s\|^k} = 0.$$

Kapitel 2

Optimierung ohne Nebenbedingungen

Wir betrachten zunächst das unrestringierte Optimierungsproblem

$$\text{(P)} \qquad \min_{x \in \mathbb{R}^n} \; f(x)$$

mit einer Zielfunktion $f : \mathbb{R}^n \to \mathbb{R}$.

2.1 Optimalitätsbedingungen

Ein wichtiges analytisches Werkzeug zur Analyse von lokalen Minima sowie zur Konvergenzanalyse von Optimierungsalgorithmen ist die Taylorformel. Wir wiederholen kurz die wesentlichen Aussagen.

Satz 2.1.1 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Dann gilt für beliebige $x, s \in \mathbb{R}^n$*

$$f(x+s) = f(x) + \nabla f(x+ts)^T s = f(x) + \nabla f(x)^T s + o(\|s\|)$$

mit einem $t \in (0,1)$.
Ist $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, dann gilt

$$f(x+s) = f(x) + \nabla f(x)^T s + \frac{1}{2} s^T \nabla^2 f(x+ts) s = f(x) + \nabla f(x)^T s + \frac{1}{2} s^T \nabla^2 f(x) s + o(\|s\|^2)$$

mit einem $t \in (0,1)$. Zudem gilt

$$\nabla f(x+s) = \nabla f(x) + \int_0^1 \nabla^2 f(x+ts) s \, dt,$$

wobei das Integral komponentenweise zu berechnen ist.

Beweis: Wir stellen fest, dass die Funktion $\phi : t \mapsto f(x+ts)$ ein bzw. zweimal stetig differenzierbar ist mit Ableitungen

$$\phi'(t) = \nabla f(x+ts)^T s, \quad \phi''(t) = s^T \nabla^2 f(x+ts)s.$$

Taylorentwicklung in $t = 0$ liefert nun z.B.

$$\phi(1) = \phi(0) + \phi'(0) + \frac{1}{2}\phi''(t)$$

mit einem $t \in (0,1)$. Einsetzen ergibt die zweite Taylorformel.

Die letzte Formel ergibt sich aus der Beobachtung, dass gilt

$$\frac{d}{dt}\nabla f(x+ts) = \nabla^2 f(x+ts)s$$

und somit

$$\nabla f(x+s) - \nabla f(x) = \int_0^1 \frac{d}{dt}\nabla f(x+ts)\,dt = \int_0^1 \nabla^2 f(x+ts)s\,dt.$$

□

2.1.1 Notwendige Optimalitätsbedingungen

Der folgende Satz gibt eine wohlbekannte notwendige Bedingung für ein lokales Minimum an.

Satz 2.1.2 (Notwendige Bedingung erster Ordnung)
Es sei $\bar{x} \in \mathbb{R}^n$ lokales Minimum von (P) mit Zielfunktion $f : \mathbb{R}^n \to \mathbb{R}$ und es sei f differenzierbar in $\bar{x}$. Dann gilt

$$\nabla f(\bar{x}) = 0. \tag{2.1}$$

Beweis: Für $\varepsilon > 0$ klein genug gilt $f(x) \geq f(\bar{x})$ für alle $x \in B_\varepsilon(\bar{x})$. Sei nun $s \in \mathbb{R}^n$ beliebig. Dann gilt (mit der Konvention $\varepsilon/0 = \infty$)

$$f(\bar{x}+ts) - f(\bar{x}) \geq 0 \quad \forall\, t \in [0, \varepsilon/\|s\|)$$

und mit der Definition der Ableitung

$$0 \leq \lim_{t \searrow 0} \frac{f(\bar{x}+ts) - f(\bar{x})}{t} = \nabla f(\bar{x})^T s.$$

Die Wahl $s = -\nabla f(\bar{x})$ liefert

$$0 \leq -\nabla f(\bar{x})^T \nabla f(\bar{x}) = -\|\nabla f(\bar{x})\|^2 \leq 0,$$

also $\nabla f(\bar{x}) = 0$. □

Dies motiviert folgende

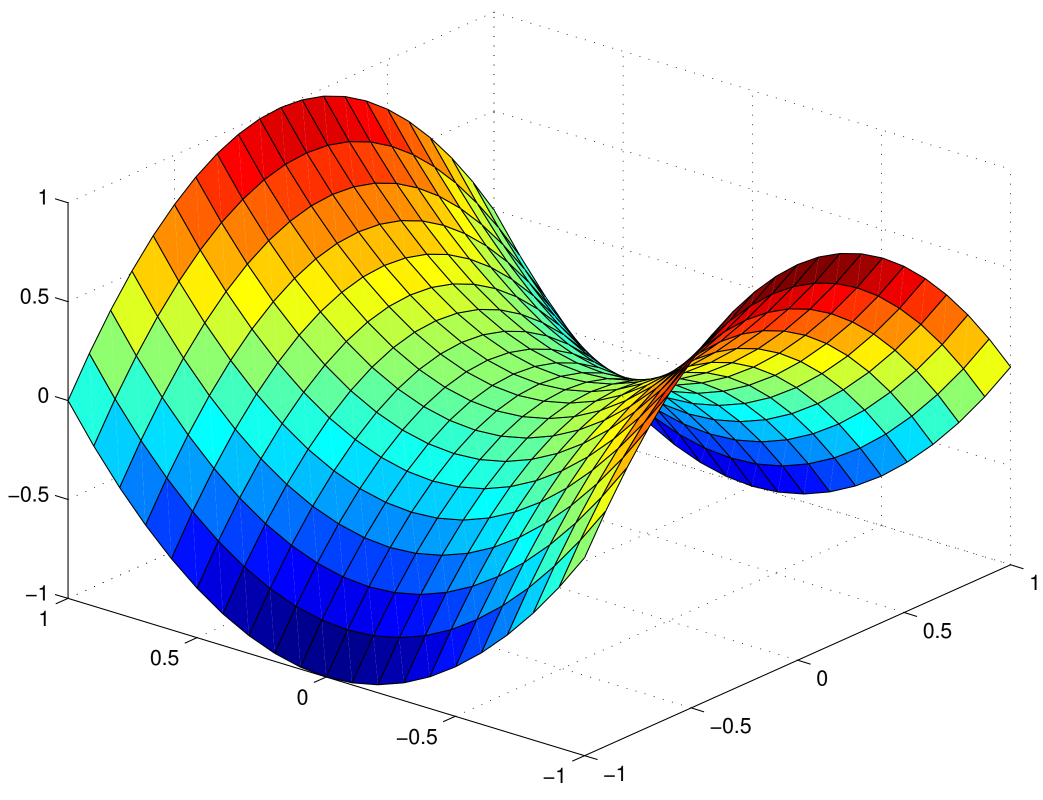
Definition 2.1.3 *Sei $\bar{x} \in \mathbb{R}^n$ und sei $f : U \to \mathbb{R}$ stetig differenzierbar in einer offenen Umgebung U von $\bar{x}$. Der Punkt $\bar{x}$ heißt* stationärer Punkt *von f, wenn $\nabla f(\bar{x}) = 0$ gilt.*

Natürlich muss ein stationärer Punkt nicht notwendigerweise ein lokales Minimum von f sein, denn $\bar{x}$ kann auch ein lokales Maximum oder ein Sattelpunkt sein.

Beispiel 2.1.1 Betrachte die Funktion

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

Dann ist $\nabla f(x) = \binom{-2x_1}{2x_2}$ und somit $\bar{x} = \binom{0}{0}$ der einzige stationäre Punkt. Aber die Funktion $x_1 \mapsto f(x_1, 0) = -x_1^2$ hat in $x_1 = 0$ ein Maximum, und $x_2 \mapsto f(0, x_2) = x_2^2$ hat in $x_2 = 0$ ein Minimum. $\bar{x} = \binom{0}{0}$ ist also weder ein Minimum noch ein Maximum von f, sondern ein sogenannter *Sattelpunkt*.



Graph der Funktion $f(x) = -x_1^2 + x_2^2$.

Bemerkung: Man nennt einen stationären Punkt, der weder lokales Minimum noch lokales Maximum ist, einen Sattelpunkt. □

Offensichtlich ist also neben der Stationarität das Krümmungsverhalten von Bedeutung.

Satz 2.1.4 (Notwendige Bedingung zweiter Ordnung)
Es sei $\bar{x} \in \mathbb{R}^n$ lokales Minimum von (P) und es sei $f : B_\varepsilon(\bar{x}) \to \mathbb{R}$ zweimal stetig differenzierbar für ein $\varepsilon > 0$. Dann gilt:

i) $\nabla f(\bar{x}) = 0$, *d.h.* $\bar{x}$ *ist stationär,*

ii) $\nabla^2 f(\bar{x})$ *is positiv semidefinit, d.h.*

$$s^T \nabla^2 f(\bar{x}) s \geq 0 \quad \forall\, s \in \mathbb{R}^n.$$

Beweis: Für $\varepsilon > 0$ klein genug ist $f : B_\varepsilon(\bar{x}) \to \mathbb{R}$ zweimal stetig differenzierbar und es gilt $f(x) \geq f(\bar{x})$ für alle $x \in B_\varepsilon(\bar{x})$. Sei nun $s \in \mathbb{R}^n$ beliebig. Dann liefert Taylorentwicklung

$$0 \leq f(\bar{x} + ts) - f(\bar{x}) = t \nabla f(\bar{x})^T s + \frac{t^2}{2} s^T \nabla^2 f(\bar{x}) s + o(t^2) \quad \forall\, t \in (-\varepsilon/\|s\|, \varepsilon/\|s\|).$$

Nun ist $\nabla f(\bar{x}) = 0$ nach Satz 2.1.2 und daher liefert Division durch $t^2/2$

$$0 \leq s^T \nabla^2 f(\bar{x}) s + \frac{o(t^2)}{t^2} \stackrel{t \to 0}{\longrightarrow} s^T \nabla^2 f(\bar{x}) s.$$

$\square$

Auch die notwendige Bedingung zweiter Ordnung ist nicht hinreichend für ein lokales Minimum.

Beispiel 2.1.2 Die Funktion $f(x) = x^3$ hat bei $\bar{x} = 0$ einen stationären Punkt und die Hessematrix $\nabla^2 f(0) = 0$ ist positiv semidefinit. Dennoch ist $\bar{x} = 0$ kein lokales Minimum.

2.1.2 Hinreichende Optimalitätsbedingung

Verschärfen wir die positive Semidefinitheitsbedingung in Satz 2.1.4, so erhalten wir eine hinreichende Bedingung.

Satz 2.1.5 (Hinreichende Bedingung zweiter Ordnung)
Es sei $f : U \to \mathbb{R}$ zweimal stetig differenzierbar auf einer offenen Menge $U \subset \mathbb{R}^n$. Gilt in einem Punkt $\bar{x} \in U$ die sogenannte hinreichende Bedingung zweiter Ordnung

i) $\nabla f(\bar{x}) = 0$, *d.h.* $\bar{x}$ *ist stationär,*

ii) $\nabla^2 f(\bar{x})$ *is positiv definit, d.h.*

$$s^T \nabla^2 f(\bar{x}) s > 0 \quad \forall\, 0 \neq s \in \mathbb{R}^n,$$

dann ist $\bar{x}$ ein isoliertes lokales Minimum von (P)*. Genauer gibt es $\varepsilon > 0$ und $\mu > 0$ mit*

$$f(x) - f(\bar{x}) \geq \frac{\mu}{4}\|x - \bar{x}\|^2 \quad \forall\, x \in B_\varepsilon(\bar{x}).$$

Beweis: Wegen ii) finden wir ein $\mu > 0$ mit

$$s^T \nabla^2 f(\bar{x}) s \geq \mu \|s\|^2 \quad \forall\, s \in \mathbb{R}^n \tag{2.2}$$

Zur Erinnerung: mit einer orthogonalen Matrix $U \in \mathbb{R}^{n,n}$ ist $\nabla^2 f(\bar{x}) = U^T D U$, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ mit den Eigenwerten $0 < \lambda_1 \leq \cdots \leq \lambda_n$ von $\nabla^2 f(\bar{x})$ und daher gilt mit $v = Us$

$$s^T \nabla^2 f(\bar{x}) s = s^T U^T D U s = v^T D v \geq \lambda_1 v^T v = \lambda_1 s^T U^T U s = \lambda_1 s^T s.$$

Für $\varepsilon > 0$ klein genug ist $B_\varepsilon(\bar{x}) \subset U$. Taylorentwicklung liefert für beliebiges $s \in B_\varepsilon(0)$ wegen i) und (2.2)

$$f(\bar{x} + s) - f(\bar{x}) = \nabla f(\bar{x})^T s + \frac{1}{2} s^T \nabla^2 f(\bar{x}) s + o(\|s\|^2) \geq \frac{\mu}{2}\|s\|^2 + o(\|s\|^2).$$

Wir können nun $\varepsilon > 0$ so verkleinern, dass folgt

$$f(\bar{x} + s) - f(\bar{x}) \geq \frac{\mu}{4}\|s\|^2 \quad \forall\, s \in B_\varepsilon(0).$$

□

Die hinreichende Bedingung zweiter Ordnung ist keine notwendige Bedingung.

Beispiel 2.1.3 Die Funktion $f(x) = x^{2k}$ hat für $k \in \mathbb{N} \setminus \{1\}$ in $\bar{x} = 0$ ihr eindeutiges globales Minimum, aber $\nabla^2 f(\bar{x}) = 2k(2k-1)\bar{x}^{2k-2} = 0$ ist nicht positiv definit.

2.2 Konvexität

Konvexität spielt in der Optimierung eine wichtige Rolle. Konvexe Optimierungsprobleme, d.h. Probleme (P_Z) mit konvexer Zielfunktion und konvexem zulässigen Bereich Z, treten in vielen Anwendungen auf und haben die wichtige Eigenschaft, dass jedes lokale Minimum ein globales Minimum ist.

Definition 2.2.1 *Eine Menge $Z \subset \mathbb{R}^n$ heißt* konvex*, falls für alle $x, y \in Z$ gilt*

$$(1-t)x + ty \in Z \quad \forall\, t \in [0, 1].$$

Anders ausgedrückt: mit zwei Punkten $x, y \in Z$ liegt auch ihre Verbindungsstrecke $[x, y]$ in Z.

Definition 2.2.2 *Es sei $Z \subset \mathbb{R}^n$ konvex.*

a) Eine Funktion $f : Z \to \mathbb{R}$ heißt konvex, *falls für alle $x, y \in Z$ gilt*

$$f((1-t)x + ty) \le (1-t)f(x) + tf(y) \quad \forall\, t \in [0,1].$$

b) Eine Funktion $f : Z \to \mathbb{R}$ heißt streng konvex, *falls für alle $x, y \in Z$ mit $x \neq y$ gilt*

$$f((1-t)x + ty) < (1-t)f(x) + tf(y) \quad \forall\, t \in (0,1).$$

Wir geben einige nützliche Charakterisierungen von (streng) konvexen Funktionen an.

Satz 2.2.3 *Es sei $Z \subset \mathbb{R}^n$ eine offene konvexe Menge und $f : Z \to \mathbb{R}$ stetig differenzierbar. Dann gilt:*

i) f ist konvex genau dann, wenn gilt

(2.3) $$f(y) - f(x) \ge \nabla f(x)^T (y - x) \quad \forall\, x, y \in Z.$$

ii) f ist streng konvex genau dann, wenn gilt

(2.4) $$f(y) - f(x) > \nabla f(x)^T (y - x) \quad \forall\, x, y \in Z,\ x \neq y.$$

Beweis: zu i) und ii): "$\Longrightarrow$": Sei $f : Z \to \mathbb{R}$ konvex. Dann gilt für beliebige $x, y \in Z$ und alle $0 < t \le 1$

$$\frac{f(x + t(y-x)) - f(x)}{t} \le \frac{tf(y) + (1-t)f(x) - f(x)}{t} = f(y) - f(x).$$

Übergang zum Limes $t \searrow 0$ liefert nun die Behauptung in i), denn

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

Zum Nachweis dieser Richtung in ii) sei f streng konvex. Dann gilt für alle $x, y \in Z$, $x \neq y$, und $0 < t < 1$:

$$f((1-t)x + ty) - f(x) \ge t\nabla f(x)^T (y - x),$$

und wegen der strengen Konvexität von f

$$f((1-t)x + ty) - f(x) < (1-t)f(x) + tf(y) - f(x) = t(f(y) - f(x)).$$

'Hintereinanderschalten' beider Ungleichungen und Teilen durch $t > 0$ gibt

$$f(y) - f(x) > \nabla f(x)^T (y - x).$$

"$\Longleftarrow$": Es gelte (2.3). Für beliebige $x, y \in Z$ und $0 \le t \le 1$ setzen wir $z = (1-t)x + ty$. Für i) müssen wir zeigen, dass gilt

$$(1-t)f(x) + tf(y) - f(z) \ge 0.$$

Um dies nachzuweisen, berechnen wir unter Benutzung von (2.3)

$$\begin{aligned}
(1-t)f(x) + tf(y) - f(z) &= (1-t)(f(x) - f(z)) + t(f(y) - f(z)) \\
(*) \qquad &\ge (1-t)\nabla f(z)^T(x-z) + t\nabla f(z)^T(y-z) \\
&= \nabla f(z)^T((1-t)x + ty - z) = 0.
\end{aligned}$$

Der Nachweis dieser Richtung in ii) folgt direkt durch Verwenden der strikten Ungleichung (2.4) für $x \neq y$ und $0 < t < 1$ in $(*)$. $\square$

Schließlich untersuchen wir den Zusammenhang zwischen Krümmungsverhalten und Konvexität.

Satz 2.2.4 *Es sei $Z \subset \mathbb{R}^n$ eine offene konvexe Menge und $f : Z \to \mathbb{R}$ sei zweimal stetig differenzierbar. Dann gilt:*

i) $f : Z \to \mathbb{R}$ ist konvex genau dann, wenn die Hessematrix $\nabla^2 f$ auf Z positiv semidefinit ist.

ii) Ist $\nabla^2 f$ auf Z positiv definit, dann ist $f : Z \to \mathbb{R}$ streng konvex (die Umkehrung gilt im allgemeinen nicht!).

Beweis: zu i): "$\Longrightarrow$": Sei f konvex und $x \in Z$ beliebig. Zu jedem $s \in \mathbb{R}^n$ gilt für alle $t \ge 0$ klein genug $x + ts \in Z$, da Z offen ist, und Satz 2.2.3, i) ergibt mit Taylorentwicklung

$$t\nabla f(x)^T s \le f(x+ts) - f(x) = t\nabla f(x)^T s + \frac{t^2}{2} s^T \nabla^2 f(x + \tau s)s.$$

mit einem $\tau \in [0, t]$. Dies liefert

$$0 \le s^T \nabla^2 f(x + \tau s)s \quad \overset{t \searrow 0}{\longrightarrow} \quad s^T \nabla^2 f(x)s.$$

"$\Longleftarrow$": Sei $\nabla^2 f$ positiv semidefinit auf Z. Für alle $x, y \in Z$ ist dann $(1-t)x + ty \in Z$ für alle $t \in [0,1]$ und Taylorentwicklung ergibt mit $s = y - x$

$$f(y) - f(x) = \nabla f(x)^T s + \frac{1}{2} s^T \nabla^2 f(x+ts)s \ge \nabla f(x)^T s = \nabla f(x)^T (y-x)$$

mit einem $t \in [0,1]$. Nach Satz 2.2.3, i) folgt die Konvexität von f.

zu ii): Ist $\nabla^2 f$ positiv definit auf Z, dann ergibt sich für $x, y \in Z$, $x \neq y$ wie eben

$$f(y) - f(x) = \nabla f(x)^T s + \frac{1}{2} s^T \nabla^2 f(x+ts)s > \nabla f(x)^T (y-x)$$

mit einem $t \in [0,1]$. Nach Satz 2.2.3, ii) folgt die strenge Konvexität von f. $\square$

Aus ii) ergibt sich die folgende nützliche Verschärfung von strenger Konvexität.

Definition 2.2.5 *Es sei $f : Z \to \mathbb{R}$ zweimal stetig differenzierbar auf einer offenen konvexen Menge $Z \subset \mathbb{R}^n$. Dann heißt f* gleichmäßig konvex*, wenn $\nabla^2 f$ auf Z gleichmäßig positiv definit ist, es also $\mu > 0$ gibt mit*

$$s^T \nabla^2 f(x) s \geq \mu \|s\|^2 \quad \forall\, s \in \mathbb{R}^n, \ \ \forall\, x \in Z.$$

Bemerkung: Nach Satz 2.2.4 ist eine gleichmäßig konvexe Funktion streng konvex. Die Umkehrung gilt nicht, denn $f(x) = x^4$ ist streng konvex, aber nicht gleichmäßig konvex. □

Konvexe Funktionen spielen in der Optimierung eine wichtige Rolle, da jedes lokale Minimum zugleich globales Minimum ist.

Satz 2.2.6 *Es sei $Z \subset \mathbb{R}^n$ konvex und $f : Z \to \mathbb{R}$ eine konvexe Funktion. Dann gilt:*

i) Jedes lokale Minimum von f auf Z (also von (P_Z)) ist auch globales Minimum. Die Lösungsmenge des Problems (P_Z) ist konvex.

ii) Ist f streng konvex, dann hat f auf Z höchstens ein lokales Minimum und dieses ist dann zugleich das einzige globale Minimum.

Beweis: Siehe Übung. □

2.3 Grundkonzept von Abstiegsverfahren

Wir betrachten das

Unrestringierte Optimierungsproblem:

(P) $$\min_{x \in \mathbb{R}^n} \ f(x)$$

mit einer zumindest stetig differenzierbaren Zielfunktion $f : \mathbb{R}^n \to \mathbb{R}$. In den letzten 30 Jahren wurden sehr leistungsfähige Algorithmen zur Lösung von (P) entwickelt.

Optimierungsverfahren sind in der Regel *Abstiegsverfahren*: Ausgehend von einem Startpunkt $x_0 \in \mathbb{R}^n$ generieren sie Punkte $x_k \in \mathbb{R}^n$, $k \geq 0$, mit

$$f(x_{k+1}) < f(x_k)$$

und terminieren, wenn ein geeignetes Abbruchkriterium erfüllt ist (in der Regel, falls x_k stationär ist).

Bemerkung: Es gibt auch *nichtmonotone* Abstiegsverfahren, die zumindest $f(x_{k+m_k}) < f(x_k)$ sicherstellen mit gewissen $m_k > 1$. □

Es haben sich zwei wichtige Klassen von Abstiegsverfahren etabliert: *Linesearch-Verfahren* und *Trust-Region-Verfahren*. Wir gehen in den folgenden Abschnitten zunächst ausgiebig auf *Linesearch-Verfahren* ein. *Linesearch-Verfahren* haben die folgende Struktur:

Algorithmus 1 Modellalgorithmus für ein Linesearch-Abstiegsverfahren

Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$.

Für $k = 0, 1, \dots$:

1. *Falls* x_k *stationär, d.h.* $\nabla f(x_k) = 0$*: STOP mit Ergebnis* x_k.
2. *Berechne eine* Abstiegsrichtung $s_k \in \mathbb{R}^n$, *d.h. ein* $s_k \in \mathbb{R}^n$ *mit* $\nabla f(x_k)^T s_k < 0$.
3. *Bestimme eine Schrittweite* $\sigma_k > 0$, *so dass gilt*

$$f(x_k + \sigma_k s_k) < f(x_k)$$

und zudem die Abnahme $f(x_k) - f(x_k + \sigma_k s_k)$ *ausreichend groß ist.*

4. *Setze* $x_{k+1} = x_k + \sigma_k s_k$.

Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Wir wollen erreichen, dass Algorithmus 1 folgende globale Konvergenzeigenschaften hat: entweder Algorithmus 1 terminiert endlich oder er erzeugt eine Folge (x_k) mit

- $f(x_{k+1}) < f(x_k)$
- Jeder Häufungspunkt $\bar{x}$ von (x_k) ist stationär, also $\nabla f(\bar{x}) = 0$.

Dazu müssen wir zwei Dinge sicherstellen:

- Die Abstiegsrichtungen s_k müssen hinreichend gut sein.
- Die Schrittweite σ_k muss einen ausreichenden Anteil des entlang $x_k + \sigma s_k$, $\sigma > 0$, möglichen Abstiegs liefern.

Beispiele für Abstiegsrichtungen:

- Die vielleicht naheliegendste Wahl der Abstiegsrichtung ist der negative Gradient $s_k = -\nabla f(x_k)$. Dies führt auf das Gradientenverfahren. Obwohl sich das Gradientenverfahren als nicht sehr effizient erweisen wird, lohnt es sich, seine Eigenschaften zu studieren.
- Die Basis sehr leistungsfähiger Verfahren ist der *Newton-Schritt* s_k. Sei $f : U \to \mathbb{R}$ zweimal stetig differenzierbar mit positiv definiter Hesse-Matrix auf einer offenen Menge $U \subset \mathbb{R}^n$, und sei $x_k \in U$. Zur Definition des Newton-Schritts approximieren wir f durch das Taylor-Polynom zweiter Ordnung in x_k

$$f(x_k + s) \approx f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s =: q_k(s).$$

Da $\nabla^2 f(x_k)$ positiv definit ist, ist q_k streng konvex und hat ein eindeutiges globales Minimum s_k gegeben durch

$$\nabla q_k(s_k) = \nabla^2 f(x_k)s_k + \nabla f(x_k) = 0.$$

damit ergibt sich der Newton-Schritt als Lösung des Gleichungssystems

$$\nabla^2 f(x_k)s_k = -\nabla f(x_k).$$

Der Newton-Schritt ist im Fall $\nabla f(x_k) \neq 0$ und $\nabla^2 f(x_k)$ positiv definit eine Abstiegsrichtung, da

$$s_k^T \nabla f(x_k) = -s_k^T \nabla^2 f(x_k) s_k < 0.$$

Ist die Hessematrix $\nabla^2 f(x_k)$ nicht positiv definit, dann muss sie modifiziert werden.

- Allgemein berechnen fast alle verbreiteten Abstiegsverfahren Suchrichtungen nach der Vorschrift
$$B_k s_k = -\nabla f(x_k)$$
mit $B_k \in \mathbb{R}^{n,n}$ symmetrisch positiv definit.

2.4 Das Gradientenverfahren

Wir betrachten zunächst das Abstiegsverfahren aus Algorithmus 1 mit der Wahl $s_k = -\nabla f(x_k)$. Wir beginnen mit dem Nachweis, dass dann s_k in Richtung des steilsten Abstiegs zeigt.

2.4.1 Richtung des steilsten Abstiegs

Definition 2.4.1 *Sei $f : U \to \mathbb{R}$ differenzierbar auf einer offenen Menge $U \subset \mathbb{R}^n$ und sei $x \in U$ beliebig mit $\nabla f(x) \neq 0$. Dann heißt die eindeutige Lösung $d \in \mathbb{R}^n$ des Problems*

$$\min_{\|d\|=1} \nabla f(x)^T d. \tag{2.5}$$

normierte Richtung des steilsten Abstiegs von f in x *und jeder Vektor $s = \lambda d$ mit $\lambda > 0$ heißt* Richtung des steilsten Abstiegs von f in x.

Tatsächlich ist $-\frac{\nabla f(x)}{\|\nabla f(x)\|}$ die normierte Richtung des steilsten Abstiegs.

Satz 2.4.2 *Sei $f : U \to \mathbb{R}$ differenzierbar auf einer offenen Menge $U \subset \mathbb{R}^n$ und sei $x \in U$ beliebig mit $\nabla f(x) \neq 0$. Dann hat* (2.5) *die eindeutige Lösung*

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

Insbesondere gilt

s *ist Richtung des steilsten Abstiegs von* f *in* x $\iff$ $s = -\lambda \nabla f(x)$ *mit einem* $\lambda > 0$.

Beweis: Nach der Cauchy-Schwarzschen Ungleichung ist

$$|u^T v| \leq \|u\| \|v\| \quad \forall\, u, v \in \mathbb{R}^n$$

und Gleichheit tritt genau dann auf, wenn u, v linear abhängig sind. Dies ergibt

$$\nabla f(x)^T d \geq -\|\nabla f(x)\| \|d\| = -\|\nabla f(x)\| \quad \forall\, d \in \mathbb{R}^n,\ \|d\|_2 = 1$$

mit Gleichheit genau dann, wenn $d = -\nabla f(x)/\|\nabla f(x)\|$. Somit ist $d = -\nabla f(x)/\|\nabla f(x)\|$ die eindeutige Lösung von (2.5) und $s = -\lambda \nabla f(x)$, $\lambda > 0$, liefert genau alle Abstiegsrichtungen. □

Bemerkung: Verwendet man auf $\mathbb{R}^n$ anstelle der euklidischen Norm eine andere Norm $\|\cdot\|_*$, dann sind die Richtungen des steilsten Abstiegs in $(\mathbb{R}^n, \|\cdot\|_*)$ gegeben durch λd, mit $\lambda > 0$ und d aus der (nun nicht mehr notwendigerweise einpunktigen) Lösungsmenge von $\min_{\|d\|_*=1} \nabla f(x)^T d$.

Ist speziell $\|x\|_* = \|x\|_M := x^T M x$ mit $M \in \mathbb{R}^{n,n}$ symmetrisch und positiv definit, dann sind die Richtungen des steilsten Abstiegs in $(\mathbb{R}^n, \|\cdot\|_M)$ gegeben durch $s = -\lambda M^{-1} \nabla f(x)$, $\lambda > 0$ (einfache Übungsaufgabe!). □

2.4.2 Die Schrittweitenregel von Armijo

Es wäre naheliegend, die Schrittweite σ_k zu wählen durch

Exakte Schrittweitensuche:

$$\sigma_k = \text{argmin}_{\sigma \in [0, \sigma_{max}]}\ f(x_k + \sigma s_k). \tag{2.6}$$

mit einem $\sigma_{max} \in (0, \infty]$. Diese Schrittweitenbestimmung ist jedoch – auch in einer inexakten Variante – zu aufwendig.

Wir beschreiben nun eine einfach zu implementierende Schrittweitenregel, die sogenannte Armijo-Regel. Sie bildet die Basis der meisten heute verwendeten Schrittweitenregeln und kann für eine beliebig Abstiegsrichtung s_k verwendet werden.

Schrittweitenregel von Armijo:

Es seien $\beta \in (0, 1)$ (häufig $\beta = 1/2$) und $\gamma \in (0, 1)$ (oft $\gamma \in [10^{-3}, 10^{-2}]$) fest gewählte Konstanten.

Bestimme die größte Schrittweite $\sigma_k \in \{1, \beta, \beta^2, \ldots\}$ mit

$$f(x_k) - f(x_k + \sigma_k s_k) \geq -\gamma \sigma_k \nabla f(x_k)^T s_k. \tag{2.7}$$

Interpretation: (2.7) ist äquivalent zu

$$(2.7)' \qquad f(x_k + \sigma_k s_k) \le f(x_k) + \gamma\sigma_k \nabla f(x_k)^T s_k.$$

Setzen wir $\phi_k(\sigma) = f(x_k + \sigma s_k)$, dann gilt $\phi_k'(\sigma) = \nabla f(x_k + \sigma s_k)^T s_k$ und wir können (2.7)′ schreiben als

$$\phi_k(\sigma_k) \le \phi_k(0) + \gamma\sigma_k\phi_k'(0).$$

Die rechte Seite ensteht aus der Taylorapproximation erster Ordnung von ϕ_k in $\sigma = 0$, durch Reduktion der Steigung auf das γ-fache. Siehe Abb. 2.1. $\square$

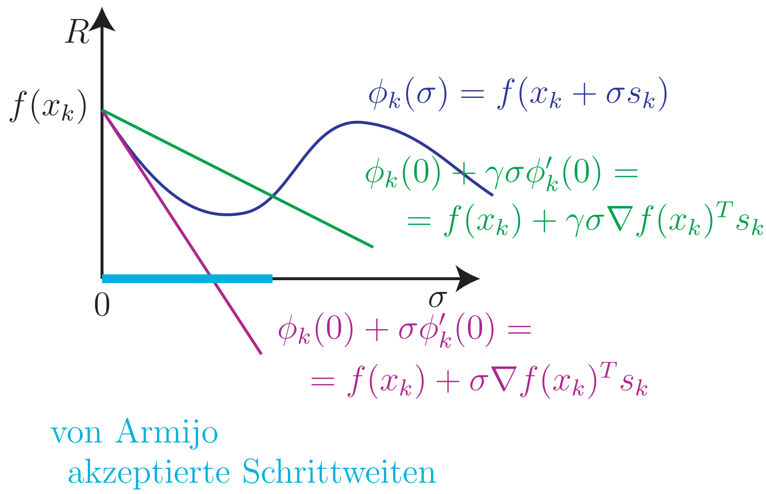


Abb. 2.1: Illustration der Armijo-Regel.

Bemerkung: Man kann auch bei einer Schrittweite $\sigma_{max} > 0$ starten, also das größte $\sigma_k \in \{\sigma_{max}, \sigma_{max}\beta, \sigma_{max}\beta^2, \dots\}$ suchen. $\square$

Wir zeigen zunächst, dass die Armijo-Regel wohldefiniert ist.

Lemma 2.4.3 *Es sei $f : U \to \mathbb{R}$ differenzierbar auf einer offenen Menge $U \subset \mathbb{R}^n$. Sei weiter $\gamma \in (0,1)$ beliebig fest. Dann gibt es zu jedem $x \in U$ mit $\nabla f(x) \neq 0$ und zu jeder Abstiegsrichtung s von f in x ein $\bar\sigma > 0$ mit*

$$f(x) - f(x + \sigma s) \ge -\gamma\sigma\nabla f(x)^T s \quad \forall \sigma \in [0, \bar\sigma].$$

Beweis: Es ist $-\nabla f(x)^T s > 0$. Für $\sigma > 0$ klein genug ist $x + \sigma s \in U$ und wir erhalten

$$\frac{f(x) - f(x+\sigma s)}{\sigma} \xrightarrow{\sigma \searrow 0} -\nabla f(x)^T s > -\gamma\nabla f(x)^T s.$$

Daher finden wir $\bar\sigma > 0$ mit

$$\frac{f(x) - f(x+\sigma s)}{\sigma} > -\gamma\nabla f(x)^T s \quad \forall \sigma \in [0, \bar\sigma].$$

$\square$

2.4.3 Globale Konvergenz des Gradientenverfahrens

Wir betrachten Algorithmus 1 mit $s_k = -\nabla f(x_k)$ und Armijo-Schrittweitenregel:

Algorithmus 2 Gradientenverfahren, Methode des steilsten Abstiegs

Wähle Parameter $\beta, \gamma \in (0, 1)$ *und einen Startpunkt* $x_0 \in \mathbb{R}^n$.

Für $k = 0, 1, \ldots$:

1. *Falls* x_k *stationär, d.h.* $\nabla f(x_k) = 0$*: STOP mit Ergebnis* x_k.
2. *Setze* $s_k = -\nabla f(x_k)$.
3. *Bestimme die Schrittweite* $\sigma_k > 0$ *nach der Armijo-Regel* (2.7).
4. *Setze* $x_{k+1} = x_k + \sigma_k s_k$.

Bemerkung: In der Praxis wird das Abbruchkriterium $\nabla f(x_k) = 0$ in Schritt 1 ersetzt durch $\|\nabla f(x_k)\| \leq \varepsilon$ mit einer vorab festgelegten Schranke $\varepsilon > 0$ (z.B. $\varepsilon = 10^{-8}$). $\square$

Es gilt folgender Konvergenzsatz.

Satz 2.4.4 *Es sei* $f : \mathbb{R}^n \to \mathbb{R}$ *stetig differenzierbar. Dann terminiert Algorithmus 2 entweder endlich oder er erzeugt eine Folge* (x_k) *mit*

i) $f(x_{k+1}) < f(x_k)$

ii) Jeder Häufungspunkt $\bar{x}$ *von* (x_k) *ist stationär, also* $\nabla f(\bar{x}) = 0$.

Beweis: Wir müssen nur den Fall diskutieren, dass der Algorithmus nicht endlich terminiert. Für alle k ist dann $\nabla f(x_k) \neq 0$ und somit ist $s_k = -\nabla f(x_k)$ eine Abstiegsrichtung wegen $\nabla f(x_k)^T s_k = -\|\nabla f(x_k)\|^2 < 0$.

zu i): Nach Lemma 2.4.3 liefert die Armijo-Regel daher Schrittweiten $\sigma_k > 0$ und es gilt

$$f(x_k) - f(x_{k+1}) \geq -\gamma\sigma_k \nabla f(x_k)^T s_k = \gamma\sigma_k \|\nabla f(x_k)\|^2 > 0.$$

Dies zeigt i).

zu ii): Sei $\bar{x}$ ein Häufungspunkt von (x_k) und $(x_k)_{k\in K}$ (kurz $(x_k)_K$) eine Teilfolge mit $(x_k)_K \to \bar{x}$. Da $f(x_k)$ monoton fällt, ist $\lim_{k\to\infty} f(x_k) = f^*$ mit $f^* \in \mathbb{R} \cup \{-\infty\}$. Aus Stetigkeitsgründen gilt aber auch

$$f^* = \lim_{k\in K\to\infty} f(x_k) = f(\bar{x}).$$

Es folgt $\lim_{k\to\infty} f(x_k) = f(\bar{x})$. Die Armijo-Bedingung (2.7) ergibt

$$0 \le \gamma\sigma_k \|\nabla f(x_k)\|^2 \le f(x_k) - f(x_{k+1}) \to 0,$$

also $\sigma_k \|\nabla f(x_k)\|^2 \to 0$ und insbesondere

$$\lim_{k\in K\to\infty} \sigma_k \|\nabla f(x_k)\|^2 = \|\nabla f(\bar{x})\|^2 \liminf_{k\in K\to\infty} \sigma_k = 0.$$

Annahme, es ist $\nabla f(\bar{x}) \neq 0$. Dann folgt

$$\lim_{k\in K\to\infty} \sigma_k = 0$$

und wir werden dies auf einen Widerspruch führen. Für alle $k \in K$ groß genug ist dann $\sigma_k \le \beta$ und daher die Armijo-Bedingung (2.7) mit $\tilde{\sigma}_k = \sigma_k/\beta$ anstelle σ_k noch nicht erfüllt, also

$$(2.8) \qquad f(x_k) - f(x_k + \sigma_k/\beta s_k) < \gamma\sigma_k/\beta \|\nabla f(x_k)\|^2 \quad \text{für alle } k \in K \text{ groß genug.}$$

Nach Division durch $\tilde{\sigma}_k$ ergibt sich nach dem Mittelwertsatz mit geeigneten $\tau_k \in [0, \tilde{\sigma}_k]$

$$\begin{aligned}\lim_{k\in K\to\infty} \frac{f(x_k) - f(x_k + \tilde{\sigma}_k s_k)}{\tilde{\sigma}_k} &= \lim_{k\in K\to\infty} -\nabla f(x_k + \tau_k s_k)^T s_k = \|\nabla f(\bar{x})\|^2 \\ &\overset{(2.8)}{\le} \lim_{k\in K\to\infty} \gamma \|\nabla f(x_k)\|^2 = \gamma \|\nabla f(\bar{x})\|^2.\end{aligned}$$

Aber $\|\nabla f(\bar{x})\|^2 \le \gamma \|\nabla f(\bar{x})\|^2$ ist wegen $\gamma \in (0,1)$ ein Widerspruch zu $\nabla f(\bar{x}) \neq 0$. □

Bemerkung: Dieselbe Konvergenzaussage gilt bei Verwendung der

Exakten lokalen Schrittweitensuche:

$$(2.9) \quad \sigma_k = \begin{cases} \text{kleinstes lokales Minimum von } 0 \le \sigma \mapsto f(x_k + \sigma s_k), \text{ falls es existiert,} \\ \sigma_{max} \quad \text{sonst} \end{cases}$$

mit festem $\sigma_{max} > 0$. Diese ist jedoch bei nichtlinearer Zielfunktion in der Praxis zu aufwendig. □

2.4.4 Konvergenzgeschwindigkeit für quadratische Zielfunktion

Praktische Tests mit dem Gradientenverfahren zeigen schnell, dass die Konvergenzgeschwindigkeit sehr langsam sein kann. Wir untersuchen im folgenden die Konvergenzgeschwindigkeit des Gradientenverfahrens für den Fall einer streng konvexen quadratischen Zielfunktion

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

$b \in \mathbb{R}^n$, $Q \in \mathbb{R}^{n,n}$ symmetrisch positiv definit.

Wir betrachten das Gradientenverfahren aus Algorithmus 2, allerdings der Einfachheit halber mit der exakten Schrittweitenregel (2.9) anstelle der Armijo-Regel.

Sei x_k nicht stationär. Dann ist die Suchrichtung gegeben durch

$$s_k = -\nabla f(x_k) = -(b + Qx_k).$$

Da f eine streng konvexe quadratische Funktion ist, können wir die exakte Schrittweite σ_k explizit berechnen: Die Funktion

$$\varphi(\sigma) := f(x_k + \sigma s_k)$$

ist eine Parabel mit

$$\varphi'(\sigma) = \nabla f(x_k + \sigma s_k)^T s_k, \quad \varphi''(\sigma) = s_k^T Q s_k > 0,$$

ist also insbesondere streng konvex. Die exakte Schrittweite σ_k nach (2.9) ergibt sich somit aus

$$0 = \varphi'(\sigma_k) = \nabla f(x_k + \sigma_k s_k)^T s_k = (b + Qx_k)^T s_k + \sigma_k s_k^T Q s_k, \tag{2.10}$$

also

$$\sigma_k = -\frac{(b + Qx_k)^T s_k}{s_k^T Q s_k} = -\frac{\nabla f(x_k)^T s_k}{s_k^T Q s_k} = \frac{\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_k)^T Q \nabla f(x_k)} > 0. \tag{2.11}$$

Da σ_k das einzige Minimum ist, liefert die exakte lokale Schrittweitensuche (2.9) dasselbe wie die exakte Schrittweitensuche (2.6) mit $\sigma_{max} = \infty$.

Bemerkung: (2.10) zeigt, dass bei exakter Schrittweite gilt

$$\nabla f(x_{k+1})^T s_k = \nabla f(x_k + \sigma s_k)^T s_k = 0.$$

Die neue Suchrichtung $s_{k+1} = -\nabla f(x_{k+1})$ steht also senkrecht auf der alten.

Im zweidimensionalen Fall $n = 2$ ergibt sich also ein rechtwinkliger Zickzack-Pfad und langsame Konvergenz ist offensichtlich, wenn $-\nabla f(x_0)$ nicht nahezu in Richtung der Lösung $\bar{x}$ zeigt. Siehe Abb. 2.2. $\square$

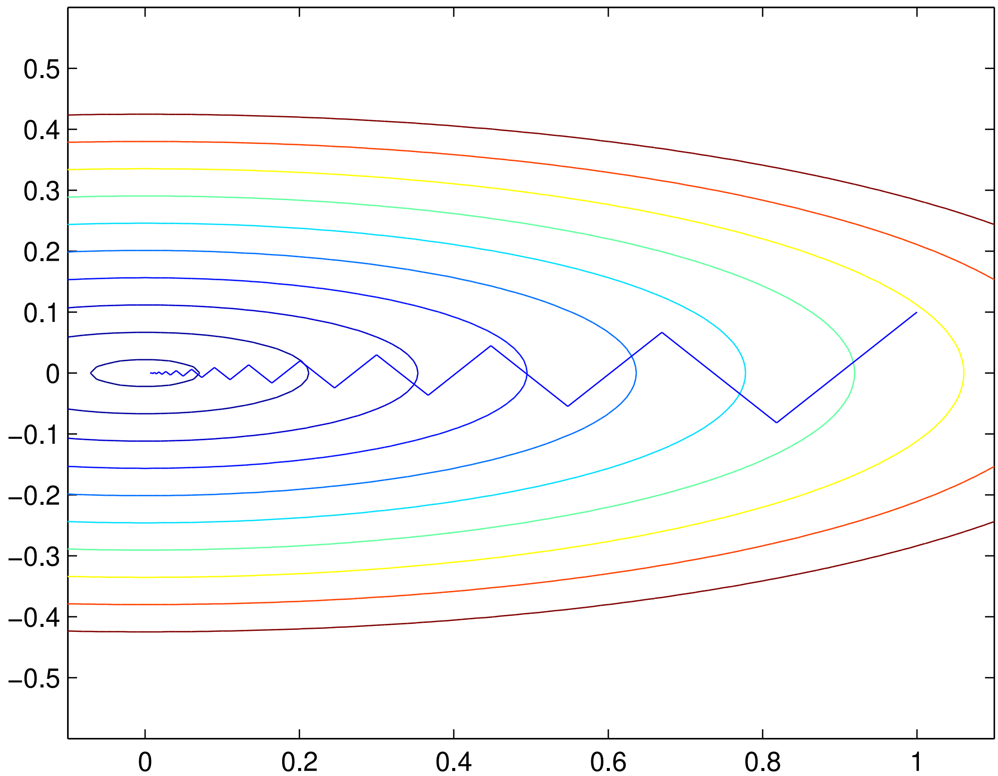


Abb. 2.2: Rechtwinkliger Zickzack-Pfad bei Anwendung des Gradientenverfahrens mit exakter Schrittweite für $f(\xi_1, \xi_2) = \frac{1}{2}(\xi_1^2 + 10\xi_2^2)$ mit Startpunkt $x_0 = \binom{1}{0.1}$.

Wir erhalten die folgende worst-case-Abschätzung der Konvergenzrate.

Satz 2.4.5 *Sei $f(x) = b^T x + \frac{1}{2} x^T Q x$ eine quadratische Funktion mit $Q \in \mathbb{R}^{n,n}$ symmetrisch positiv definit. Es bezeichne (x_k) und (σ_k) die vom Gradientenverfahren (Algorithmus refalg:grad) mit lokal exakter lokaler Schrittweitensuche* (2.9) *erzeugten Folgen (falls es nicht endlich konvergiert). Dann gilt mit dem eindeutigen globalen Minimum $\bar{x} = -Q^{-1} b$ von f*

$$f(x_{k+1}) - f(\bar{x}) \le \left(\frac{\lambda_{max}(Q) - \lambda_{min}(Q)}{\lambda_{max}(Q) + \lambda_{min}(Q)} \right)^2 (f(x_k) - f(\bar{x})) \tag{2.12}$$

$$\|x_k - \bar{x}\| \le \left(\frac{\lambda_{max}(Q)}{\lambda_{min}(Q)} \right)^{1/2} \left(\frac{\lambda_{max}(Q) - \lambda_{min}(Q)}{\lambda_{max}(Q) + \lambda_{min}(Q)} \right)^k \|x_0 - \bar{x}\|, \tag{2.13}$$

wobei $\lambda_{min}(Q)$ und $\lambda_{max}(Q)$ den kleinsten bzw. größten Eigenwert von Q bezeichnen.

Bemerkung: Der garantierte Konvergenzfaktor strebt für große Konditionszahl $\kappa(Q) = \frac{\lambda_{max}(Q)}{\lambda_{min}(Q)}$ gegen 1, die garantierte Konvergenzgeschwindigkeit kann also je nach Kondition beliebig langsam sein. □

Bemerkung: Man kann zeigen, dass die Abschätzung (2.12) für $f(\xi_1, \xi_2) = \frac{1}{2}(\xi_1^2 + \kappa\xi_2^2)$, $\kappa \ge 1$, und den Startpunkt $x_0 = \binom{1}{1/\kappa}$ scharf ist (Nachweis durch Nachrechnen). Siehe Abb. 2.2 für den Fall $\kappa = 10$. □

Beweis: Wir setzen $g_k := \nabla f(x_k) = b + Qx_k$. Terminiert der Algorithmus nicht endlich, dann ist $g_k \neq 0$ und die exakte Schrittweite nach (2.11) gegeben durch

$$\sigma_k = \frac{g_k^T g_k}{g_k^T Q g_k}. \tag{2.14}$$

Nun liefert Taylorentwicklung in $\bar{x}$

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

(das Restglied verschwindet, da $\nabla^2 f = Q$ konstant ist) und somit

$$\begin{aligned} f(x_{k+1}) - f(\bar{x}) &= f(x_{k+1}) - f(x_k) + f(x_k) - f(\bar{x}) \\ &= f(x_k) - f(\bar{x}) + \sigma_k g_k^T s_k + \frac{1}{2}\sigma_k^2 s_k^T Q s_k \\ &= f(x_k) - f(\bar{x}) - \sigma_k g_k^T g_k + \frac{1}{2}\sigma_k^2 g_k^T Q g_k. \end{aligned}$$

Einsetzen von (2.14) ergibt

$$f(x_{k+1}) - f(\bar{x}) = f(x_k) - f(\bar{x}) - \frac{1}{2}\frac{(g_k^T g_k)^2}{g_k^T Q g_k}. \tag{2.15}$$

Weiter ist wegen $g_k = Qx_k + b = Q(x_k - \bar{x})$

$$f(x_k) - f(\bar{x}) = \frac{1}{2}(x_k - \bar{x})^T Q (x_k - \bar{x}) = \frac{1}{2} g_k^T Q^{-1} g_k$$

und wir erhalten mit (2.15)

$$f(x_{k+1}) - f(\bar{x}) = \left(1 - \frac{(g_k^T g_k)^2}{(g_k^T Q g_k)(g_k^T Q^{-1} g_k)}\right)(f(x_k) - f(\bar{x})).$$

Die nachfolgende Kantorovich-Ungleichung ergibt nun (2.12).

Schließlich ist

$$f(x) - f(\bar{x}) = \frac{1}{2}(x - \bar{x})^T Q (x - \bar{x}) \begin{cases} \geq \frac{\lambda_{min}(Q)}{2}\|x - \bar{x}\|^2 \\ \leq \frac{\lambda_{max}(Q)}{2}\|x - \bar{x}\|^2 \end{cases}$$

Zur Erinnerung: mit einer orthogonalen Matrix $U \in \mathbb{R}^{n,n}$ ist $Q = U^T D U$, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ mit den Eigenwerten $0 < \lambda_1 \leq \cdots \leq \lambda_n$ von Q und daher gilt mit $v = Us$

$$s^T \nabla^2 f(\bar{x}) s = s^T U^T D U s = v^T D v \begin{cases} \geq \lambda_1 v^T v = \lambda_1 s^T U^T U s = \lambda_1 s^T s, \\ \leq \lambda_n v^T v = \lambda_n s^T s. \end{cases}$$

Daher folgt (2.13) unmittelbar aus (2.12). $\square$

Lemma 2.4.6 (Kontorovich-Ungleichung)
Sei $Q \in \mathbb{R}^{n,n}$ symmetrisch positiv definit. Dann gilt

$$\frac{(d^T d)^2}{(d^T Q d)(d^T Q^{-1} d)} \geq \frac{4\lambda_{max}(Q)\lambda_{min}(Q)}{(\lambda_{max}(Q) + \lambda_{min}(Q))^2} \quad \forall\, d \in \mathbb{R}^n \setminus \{0\}.$$

Beweis: Siehe zum Beispiel [Be99, Lem. 3.1] oder [GK99, S. 71]. □

Bemerkung: Ist f eine beliebige zweimal stetig differenzierbare Zielfunktion und erfüllt $\bar{x}$ die hinreichende Bedingung zweiter Ordnung, dann gilt im Falle $x_k \to \bar{x}$ die Abschätzung (2.12) zumindest asymptotisch mit $Q = \nabla^2 f(\bar{x})$, da $f(x) = f(\bar{x}) + \frac{1}{2}(x - \bar{x})^T Q(x - \bar{x}) + o(\|x - \bar{x}\|^2)$. □

2.5 Konvergenztheorie allgemeiner Abstiegsverfahren

Wegen der möglicherweise langsamen Konvergenz des Gradientenverfahrens liegt es nahe, im Abstiegsverfahren Algorithmus 1 Abstiegsrichtungen s_k zu verwenden, die schnellere Konvergenz liefern. Wir betrachten in diesem Abschnitt das allgemeine Abstiegsverfahren aus Algorithmus 1, das wir zur Erinnerung nochmals kompakt notieren:

Algorithmus 3 Allgemeines Abstiegsverfahren

Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$.

Für $k = 0, 1, \ldots$:

1. *Falls $\nabla f(x_k) = 0$: STOP mit Ergebnis x_k.*
2. *Berechne eine Abstiegsrichtung $s_k \in \mathbb{R}^n$, d.h. ein $s_k \in \mathbb{R}^n$ mit $\nabla f(x_k)^T s_k < 0$.*
3. *Bestimme eine Schrittweite $\sigma_k > 0$, so dass gilt $f(x_k + \sigma_k s_k) < f(x_k)$.*
4. *Setze $x_{k+1} = x_k + \sigma_k s_k$.*

Wir wollen zunächst Minimalanforderungen an die verwendeten Suchrichtungen s_k und Schrittweiten σ_k herleiten, welche die globale Konvergenz des allgemeinen Abstiegsverfahren aus Algorithmus 3 in folgendem Sinne sicherstellen: entweder Algorithmus 3 terminiert endlich oder er erzeugt eine Folge (x_k) mit

- $f(x_{k+1}) < f(x_k)$
- Jeder Häufungspunkt $\bar{x}$ von (x_k) ist stationär, also $\nabla f(\bar{x}) = 0$.

Bemerkung: Natürlich stellt diese Konvergenzaussage nicht notwendigerweise sicher, dass jeder Häufungspunkt $\bar{x}$ von (x_k) ein lokales Minimum ist. Wir werden später illustrieren, wie man zusätzlich erreichen kann, dass jeder Häufungspunkt $\bar{x}$ die notwendige Bedingung 2. Ordnung aus Satz 2.1.4 erfüllt, falls f zweimal stetig differenzierbar ist. $\Box$

Wir hatten bereits bemerkt, dass wir für die Konvergenz von Algorithmus 3 zwei Dinge garantieren müssen:

- die Abstiegsrichtungen s_k müssen hinreichend gut sein,
- die Schrittweiten σ_k müssen ausreichend Abstieg realisieren.

2.5.1 Zulässige Suchrichtungen

Um hinreichend gute Abstiegsrichtungen s_k zu garantieren, dürfen wir $\cos(\angle(-\nabla f(x_k), s_k)) \to 0$ nur im Fall $\nabla f(x_k) \to 0$ zulassen. Diese Anforderung ist tatsächlich sachgerecht und kann durch folgende Bedingung sichergestellt werden:

Definition 2.5.1 (Zulässige Suchrichtungen)
Die Folge der von Algorithmus 3 erzeugten Suchrichtungen (s_k) heißt zulässig*, falls gilt*

$$\nabla f(x_k)^T s_k < 0, \quad \text{d.h. die } s_k \text{ sind Abstiegsrichtungen} \tag{2.16}$$

$$\frac{\nabla f(x_k)^T s_k}{\|s_k\|} \overset{k\to\infty}{\longrightarrow} 0 \implies \nabla f(x_k) \overset{k\to\infty}{\longrightarrow} 0. \tag{2.17}$$

Die Bedingung (2.17) kann als *abstrakte Winkelbedingung* interpretiert werden: es gilt

$$\frac{-\nabla f(x_k)^T s_k}{\|s_k\|} = \frac{-\nabla f(x_k)^T s_k}{\|\nabla f(x_k)\|\,\|s_k\|}\|\nabla f(x_k)\| = \cos(\angle(-\nabla f(x_k), s_k))\|\nabla f(x_k)\|.$$

Offensichtlich ist (s_k) insbesondere zulässig, wenn mit einer Konstante $c_0 > 0$ die *Winkelbedingung* gilt

$$\begin{aligned}&\cos(\angle(-\nabla f(x_k), s_k)) \ge c_0 \quad \forall\, k \ge 0,\\ &\text{oder äquivalent}\\ &\frac{-\nabla f(x_k)^T s_k}{\|s_k\|} \ge c_0\|\nabla f(x_k)\| \quad \forall\, k \ge 0, \quad \text{(gleichmäßige Winkelbedingung)}\end{aligned} \tag{2.18}$$

Allgemeiner ist (s_k) zulässig, wenn die folgende *verallgemeinerte Winkelbedingung* gilt:

$$\begin{aligned}&\text{Mit einer stetigen, streng monoton wachsenden Funktion}\\ &\varphi : [0,\infty[\to [0,\infty[,\ \varphi(0) = 0 \text{ gilt}\\ &\frac{-\nabla f(x_k)^T s_k}{\|s_k\|} \ge \varphi(\|\nabla f(x_k)\|) \quad \forall\, k \ge 0.\end{aligned} \tag{2.19}$$

Genauer gilt der folgende

Satz 2.5.2 *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Bezeichne (s_k) die von Algorithmus 3 erzeugte Folge von Suchrichtungen. Dann gilt:*

$$(s_k) \text{ erfüllt } (2.18) \Longrightarrow (s_k) \text{ erfüllt } (2.19) \Longrightarrow (s_k) \text{ ist zulässig.}$$

Beweis: Natürlich folgt (2.19) aus (2.18) mit der Wahl $\varphi(t) = c_0 t$.

Es gelte nun (2.19). Wegen $\varphi(0) = 0$ folgt $\varphi(t) > 0$ für $t > 0$ und daher folgt aus (2.19)

$$-\nabla f(x_k)^T s_k \geq \|s_k\| \varphi(\|\nabla f(x_k)\|) > 0.$$

Somit gilt (2.16).

Zum Nachweis von (2.17) zeigen wir

$$\nabla f(x_k) \not\to 0 \implies \frac{\nabla f(x_k)^T s_k}{\|s_k\|} \not\to 0.$$

Im Fall $\nabla f(x_k) \not\to 0$ gibt es $\varepsilon > 0$ und eine Teilfolge $(x_k)_K$ mit $\|\nabla f(x_k)\| \geq \varepsilon$ für alle $k \in K$ und wir erhalten mit (2.19)

$$\begin{aligned} \|\nabla f(x_k)\| \geq \varepsilon \quad \forall\, k \in K \quad &\Longrightarrow \quad -\frac{\nabla f(x_k)^T s_k}{\|s_k\|} \geq \varphi(\|\nabla f(x_k)\|) \geq \varphi(\varepsilon) > 0 \quad \forall\, k \in K \\ &\Longrightarrow \quad \frac{\nabla f(x_k)^T s_k}{\|s_k\|} \not\to 0. \end{aligned}$$

□

Beispiele für zulässige Suchrichtungen:

1. Die Wahl $s_k = -\nabla f(x_k)$ (Gradientenverfahren) liefert zulässige Suchrichtungen. Tatsächlich gilt dann die gleichmäßige Winkelbedingung (2.18) mit $c_0 = 1$, da

   $$-\frac{\nabla f(x_k)^T s_k}{\|s_k\|} = \frac{\nabla f(x_k)^T \nabla f(x_k)}{\|\nabla f(x_k)\|} = \|\nabla f(x_k)\|.$$

2. Wie bereits erwähnt, wählt man bei Newton-artigen Verfahren s_k als Lösung von

   $$B_k s_k = -\nabla f(x_k),$$

   mit geeigneten symmetrischen, positiv definiten Matrizen $B_k \in \mathbb{R}^{n,n}$. Gilt nun mit Konstanten $0 < \mu \leq \eta$

   $$\lambda_{min}(B_k) \geq \mu, \quad \lambda_{max}(B_k) \leq \eta \quad \forall\, k \geq 0,$$

   dann ist (s_k) zulässig. Denn zunächst gilt

   $$\|\nabla f(x_k)\| = \|B_k s_k\| \leq \eta \|s_k\|$$

und somit

$$\frac{-\nabla f(x_k)^T s_k}{\|s_k\|} = \frac{s_k^T B_k s_k}{\|s_k\|} \geq \frac{\mu \|s_k\|^2}{\|s_k\|} = \mu \|s_k\| \geq \frac{\mu}{\eta} \|\nabla f(x_k)\|.$$

Also ist (2.18) mit $c_0 = \mu/\eta$ erfüllt.

Das folgende Beispiel zeigt, dass globale Konvergenz zerstört werden kann, falls $\cos(\angle(\nabla f(x_k), s_k))$ hinreichend schnell gegen 0 konvergiert.

Beispiel 2.5.1 *Siehe Übung.*

2.5.2 Zulässige Schrittweiten

Eine geeignete Schrittweitenwahl muss sicherstellen, dass $f(x_k) - f(x_k + \sigma_k s_k) \to 0$ nur auftreten kann, falls $\nabla f(x_k) \to 0$ (wie gewünscht) oder falls $\cos(\angle(-\nabla f(x_k), s_k)) \to 0$ (Abstiegsrichtungen werden unbrauchbar). Dies motiviert folgende Definition:

Definition 2.5.3 (Zulässige Schrittweiten)
Die Folge der von Algorithmus 3 erzeugten Schrittweiten (σ_k) heißt zulässig, *falls gilt*

$$f(x_k + \sigma_k s_k) < f(x_k) \quad \forall\, k \geq 0 \tag{2.20}$$

$$f(x_k) - f(x_k + \sigma_k s_k) \overset{k\to\infty}{\longrightarrow} 0 \quad \Longrightarrow \quad \frac{\nabla f(x_k)^T s_k}{\|s_k\|} \overset{k\to\infty}{\longrightarrow} 0. \tag{2.21}$$

Dieser Zulässigkeitsbegriff ist möglichst allgemein gehalten, damit die Bedingungen leicht überprüfbar sind.

Zulässige Schrittweiten werden insbesondere von sogenannten *effizienten Schrittweitenregeln* erzeugt. Obwohl wir ohne diesen Begriff auskommen werden, führen wir ihn ein, da er in der Literatur verbreitet ist.

Definition 2.5.4 (Effiziente Schrittweitenregel)
Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und sei $x_0 \in \mathbb{R}^n$.

a) Eine Schrittweitenregel *S ist ein Algorithmus, der zu jedem Punkt $x \in N_f(x_0) = \{x : f(x) \leq f(x_0)\}$ mit $\nabla f(x) \neq 0$ und jeder Abstiegsrichtung $s \in \mathbb{R}^n$ von f in x, also $\nabla f(x)^T s < 0$, eine Schrittweite $\sigma = S(x, s) > 0$ liefert.*

b) Eine Schrittweitenregel S heißt effizient, *wenn es eine Konstante $\theta > 0$ gibt, so dass zu jedem $x \in N_f(x_0)$ mit $\nabla f(x) \neq 0$ und jedem $s \in \mathbb{R}^n$ mit $\nabla f(x)^T s < 0$ für die geliefert Schrittweite $\sigma = S(x, s)$ gilt*

$$f(x) - f(x + \sigma s) \geq \theta \left(\frac{\nabla f(x)^T s}{\|s\|} \right)^2.$$

c) *Die von Algorithmus 3 erzeugte Schrittweitenfolge* (σ_k) *heißt* effizient, *wenn mit einer effizienten Schrittweitenregel* S *gilt* $\sigma_k = S(x_k, s_k)$.

Wir erhalten unmittelbar folgendes Resultat.

Lemma 2.5.5 *Sei* $f : \mathbb{R}^n \to \mathbb{R}$ *stetig differenzierbar. Verwendet Algorithmus 3 eine effiziente Schrittweitenregel* S, *gilt also* $\sigma_k = S(x_k, s_k)$, *und terminiert er nicht endlich, dann ist* (σ_k) *eine zulässige Folge von Schrittweiten.*

Beweis: Da die s_k Abstiegsrichtungen sind, ergibt sich

$$f(x_k) - f(x_{k+1}) \geq \theta \left(\frac{\nabla f(x_k)^T s_k}{\|s_k\|} \right)^2 > 0$$

und folglich $x_{k+1} \in N_f(x_0)$ zunächst für $k = 0$ und nun induktiv für alle $k \geq 0$. Dies ergibt $f(x_k) - f(x_k + \sigma_k s_k) > 0$, also (2.20). Zudem gilt auch (2.21) wegen

$$\begin{aligned} f(x_k) - f(x_k + \sigma_k s_k) \to 0 \quad &\Longrightarrow \quad 0 < \theta \left(\frac{\nabla f(x_k)^T s_k}{\|s_k\|} \right)^2 \leq f(x_k) - f(x_{k+1}) \to 0 \\ &\Longrightarrow \quad \frac{\nabla f(x_k)^T s_k}{\|s_k\|} \to 0. \end{aligned}$$

□

Wir ziehen es vor, lediglich die Zulässigkeit von Schrittweiten (σ_k) zu zeigen, da dies einfacher ist, als die Effizienz nachzuweisen.

Bemerkung: Die Armijo-Regel ist effizient, wenn mit einer Konstante $\nu > 0$ für die erzeugten Schrittweiten gilt

$$\sigma_k \geq \nu \frac{-\nabla f(x_k)^T s_k}{\|s_k\|^2}. \tag{2.22}$$

Tatsächlich liefert dann die Armijo-Bedingung

$$f(x_k) - f(x_k + \sigma_k s_k) \geq -\gamma \sigma_k \nabla f(x_k)^T s_k \geq \gamma \nu \left(\frac{\nabla f(x_k)^T s_k}{\|s_k\|} \right)^2.$$

Der Nachweis von (2.22) erfordert stärkere Voraussetzungen (∇f Lipschitz-stetig, $\|s_k\| \geq c \frac{-\nabla f(x_k)^T s_k}{\|s_k\|}$ mit einer Konstante $c > 0$) als der Nachweis, dass die Armijo-Regel zulässige Schrittweiten erzeugt (siehe Satz 2.6.1). □

2.5.3 Ein globaler Konvergenzsatz

Wir zeigen nun die globale Konvergenz von Algorithmus 3 für zulässige Suchrichtungen und Schrittweiten:

Satz 2.5.6 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Algorithmus 3 terminiere nicht endlich und erzeuge zulässige Suchrichtungen (s_k) und zulässige Schrittweiten (σ_k). Dann gilt:*

i) $f(x_{k+1}) < f(x_k) \quad \forall\, k \geq 0.$

ii) Jeder Häufungspunkt $\bar{x}$ von (x_k) ist stationär, also $\nabla f(\bar{x}) = 0$.

iii) Besitzt (x_k) einen Häufungspunkt $\bar{x}$, dann gilt $\lim\limits_{k\to\infty} \nabla f(x_k) = 0$.

iv) Ist $\bar{x}$ ein isolierter Häufungspunkt von (x_k) und gilt für jede Teilfolge $(x_k)_{k\in K}$ mit $(x_k)_{k\in K} \to \bar{x}$ zudem $(x_{k+1} - x_k)_{k\in K} \to 0$, dann konvergiert (x_k) gegen $\bar{x}$.

Beweis: zu i): Da die Folge $f(x_k)$ nach (2.20) streng monoton fällt, erhalten wir i).

zu ii): Es sei $\bar{x}$ ein Häufungspunkt von (x_k) und $(x_k)_K$ eine Teilfolge mit $(x_k)_K \to \bar{x}$. Da die Folge $f(x_k)$ nach (2.20) streng monoton fällt, gilt $\lim_{k\to\infty} f(x_k) = f^*$ mit $f^* \in \mathbb{R} \cup \{-\infty\}$. Aus Stetigkeitsgründen gilt aber auch

$$f^* = \lim_{k\to\infty} f(x_k) = \lim_{k\in K\to\infty} f(x_k) = f(\bar{x}).$$

Es folgt $\lim_{k\to\infty} f(x_k) = f(\bar{x})$ und insbesondere ($(f(x_k))$ ist eine Cauchy-Folge)

$$f(x_k) - f(x_k + \sigma_k s_k) = f(x_k) - f(x_{k+1}) \to 0$$

Dies ergibt

$$f(x_k) - f(x_k + \sigma_k s_k) \to 0 \overset{(\sigma_k)\ \text{zulässig}}{\Longrightarrow} \frac{-\nabla f(x_k)^T s_k}{\|s_k\|} \to 0$$
$$\overset{(s_k)\ \text{zulässig}}{\Longrightarrow} \nabla f(x_k) \to 0.$$

Wegen der Stetigkeit von ∇f folgt $\nabla f(\bar{x}) = \lim_{k\in K\to\infty} \nabla f(x_k) = 0$. Damit ist auch ii) gezeigt.

zu iii): Besitzt (x_k) einen Häufungspunkt $\bar{x}$, dann zeigt der Beweis zu ii), dass $\nabla f(x_k) \to 0$.

zu iv): Sei $\bar{x}$ ein isolierter Häufungspunkt. Angenommen, $x_k \not\to \bar{x}$. Da $\bar{x}$ ein isolierter Häufungspunkt ist, aber $x_k \not\to \bar{x}$ finden wir dann $\varepsilon > 0$, so dass $\bar{x}$ der einzige Häufungspunkt in der abgeschlossen Kugel $\overline{B_\varepsilon(\bar{x})}$ ist und $x_k \notin \overline{B_\varepsilon(\bar{x})}$ für unendlich viele k. Es existiert also eine Teilfolge $(x_k)_{k\in K} \subset \overline{B_\varepsilon(\bar{x})}$ mit $x_{k+1} \notin \overline{B_\varepsilon(\bar{x})}$ für alle $k \in K$. Nun hat $(x_k)_{k\in K}$ einen Häufungspunkt im Kompaktum $\overline{B_\varepsilon(\bar{x})}$. Dies kann nur $\bar{x}$ sein, da dies der einzige Häufungspunkt in $\overline{B_\varepsilon(\bar{x})}$ ist. Daher gilt $(x_k)_{k\in K} \to \bar{x}$ (sonst gäbe es einen weiteren Häufungspunkt im Kompaktum $\overline{B_\varepsilon(\bar{x})}$) und somit nach Voraussetzung $(x_k - x_{k+1})_{k\in K} \to 0$. Es folgt $(x_{k+1})_{k\in K} \to \bar{x}$ im Widerspruch zu $x_{k+1} \notin \overline{B_\varepsilon(\bar{x})}$ für alle $k \in K$. $\square$

Ist die Niveaumenge $N_f(x_0)$ für den verwendeten Startpunkt x_0 kompakt, dann können wir folgende Konvergenzeigenschaften ableiten.

Satz 2.5.7 *Ist in Satz 2.5.6 zudem die Niveaumenge $N_f(x_0)$ für den Startpunkt $x_0 \in \mathbb{R}^n$ kompakt, dann gelten i)–iv) und iii) kann verschärft werden zu*

iii)' $\lim\limits_{k\to\infty} \nabla f(x_k) = 0.$

Beweis: Einfache Übung.

$\square$

2.6 Schrittweitenregeln

2.6.1 Die Armijo-Regel

Wir haben die Armijo-Regel (2.7) bereits in Abschnitt 2.7 kennengelernt. Wir haben gesehen, dass sie wohldefiniert ist und die globale Konvergenz des Gradientenverfahrens sicherstellt.

Wir wollen nun die Armijo-Regel im allgemeinen Abstiegsverfahren aus Algorithmus 3 einsetzen. Wir werden zeigen, dass die Armijo-Regel zulässige Schrittweiten erzeugt, solange die verwendeten Abstiegsrichtungen s_k nicht "zu kurz" im Vergleich zu $\nabla f(x_k)$ werden. Anschließend werden wir die Powell-Wolfe-Regel behandeln, die mit beliebigen Abstiegsrichtungen auskommt und bei Variable-Metrik-Verfahren eine wichtige Rolle spielen wird.

Das folgende Resultat zeigt, dass die Armijo-Regel unter schwachen Voraussetzungen zulässige Schrittweiten erzeugt.

Satz 2.6.1 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Sei $x_0 \in \mathbb{R}^n$ ein beliebiger Startpunkt, so dass die Niveaumenge $N_f(x_0)$ kompakt ist. Verwendet man im Abstiegsverfahren Algorithmus 3 die Armijo-Regel, terminiert das Verfahren nicht endlich und gilt für die Suchrichtungen*

$$\|s_k\| \geq \varphi\left(\frac{-\nabla f(x_k)^T s_k}{\|s_k\|}\right) \quad \forall\, k \geq 0 \tag{2.23}$$

mit einer streng monoton wachsenden Funktion $\varphi : [0,\infty) \to [0,\infty)$, dann ist die von der Armijo-Regel erzeugte Schrittweitenfolge (σ_k) zulässig.

Beweis: Da der Algorithmus nicht endlich terminiert, gilt $\nabla f(x_k) \neq 0$ für alle $k \geq 0$. Nach Lemma 2.4.3 liefert die Armijo-Regel also Schrittweiten $\sigma_k > 0$, die (2.7) erfüllen, also

$$f(x_k) - f(x_k + \sigma_k s_k) \geq -\gamma\sigma_k \nabla f(x_k)^T s_k > 0.$$

Damit ist (2.20) bereits gezeigt. Zum Nachweis von (2.21) zeigen wir

$$\frac{\nabla f(x_k)^T s_k}{\|s_k\|} \not\to 0 \implies f(x_k) - f(x_k + \sigma_k s_k) \not\to 0.$$

Im Fall $\frac{\nabla f(x_k)^T s_k}{\|s_k\|} \not\to 0$ gibt es $\varepsilon > 0$ und eine Teilfolge $(x_k)_K$ mit $\frac{-\nabla f(x_k)^T s_k}{\|s_k\|} \geq \varepsilon$ für alle $k \in K$ (beachte, dass $-\nabla f(x_k)^T s_k > 0$). Dann gilt wegen (2.23)

$$\|s_k\| \geq \varphi\left(\frac{-\nabla f(x_k)^T s_k}{\|s_k\|}\right) \geq \varphi(\varepsilon) =: \delta > \varphi(0) \geq 0. \tag{2.24}$$

Wir zeigen zunächst, dass die Armijo-Bedingung für alle $k \in K$ erfüllt ist, sobald $\sigma_k\|s_k\| \leq \rho$ mit geeignetem $\rho > 0$ ist. Tatsächlich liefert Taylorentwicklung mit geeigneten $\tau_k \in [0, \sigma_k]$ für alle $k \in K$

$$\begin{aligned}
&f(x_k) - f(x_k + \sigma_k s_k) + \gamma\sigma_k \nabla f(x_k)^T s_k \\
&= -\nabla f(x_k + \tau_k s_k)^T(\sigma_k s_k) + \gamma\sigma_k \nabla f(x_k)^T s_k \\
&= (\nabla f(x_k) - \nabla f(x_k + \tau_k s_k))^T(\sigma_k s_k) + (1-\gamma)\sigma_k\|s_k\| \frac{-\nabla f(x_k)^T s_k}{\|s_k\|} \\
&\geq \sigma_k\|s_k\| \left(-\|\nabla f(x_k) - \nabla f(x_k + \tau_k s_k)\|\right) + (1-\gamma)\varepsilon).
\end{aligned} \tag{2.25}$$

Da $N_f(x_0)$ kompakt ist, finden wir ein konvexes Kompaktum $N \supset N_f(x_0)$ (z.B. $\overline{B_R(x_0)}$, R groß genug). Wegen der gleichmäßigen Stetigkeit von ∇f auf dem Kompaktum N finden wir $\rho > 0$ mit

$$\|\nabla f(x) - \nabla f(y)\| < (1-\gamma)\varepsilon \quad \forall\, x, y \in N,\ \|x - y\| \leq \rho.$$

Also ist die rechte Seite in (2.25) nichtnegativ, sobald gilt $\sigma_k\|s_k\| \leq \rho$ (beachte $\tau_k \in [0, \sigma_k]$). Wegen (2.24) erhalten wir also

$$\sigma_k\|s_k\| \geq \min\{\beta\rho, \delta\} =: \varepsilon' > 0$$

(entweder maximales $\beta^i\|s_k\| \leq \rho$ oder $\sigma_k = 1$). Nun erhalten wir mit der Armijo-Bedingung (2.7)

$$f(x_k) - f(x_k + \sigma_k s_k) \geq \gamma\sigma_k\|s_k\| \frac{-\nabla f(x_k)^T s_k}{\|s_k\|} \geq \gamma\varepsilon'\varepsilon \quad \forall\, k \in K. \tag{2.26}$$

Damit ist $f(x_k) - f(x_k + \sigma_k s_k) \not\to 0$ gezeigt. □

2.6.2 Die Powell-Wolfe-Regel

Um unabhängig von der Länge der Suchrichtungen zulässige Schrittweiten zu garantieren, fordert man bei der Powell-Wolfe-Regel neben der Armijo-Bedingung

$$f(x_k) - f(x_k + \sigma_k s_k) \geq -\gamma \sigma_k \nabla f(x_k)^T s_k \tag{2.27}$$

zudem die Bedingung

$$\nabla f(x_k + \sigma_k s_k)^T s_k \geq \theta \nabla f(x_k)^T s_k \tag{2.28}$$

mit Konstanten $0 < \gamma < \theta < 1$. Offensichtlich ist (2.28) wegen $\nabla f(x_k)^T s_k < 0$ nicht mehr erfüllt, wenn $\sigma_k \|s_k\|$ zu klein ist. Dies führt auf die

Schrittweitenregel von Powell-Wolfe:

Es seien $0 < \gamma < \theta < 1$ (oft $\gamma \in [10^{-3}, 10^{-2}]$, $\theta = 0.9$) fest gewählte Konstanten. Bestimme $\sigma_k > 0$, so dass (2.27) und (2.28) gilt.

Interpretation: Setzen wir $\phi_k(\sigma) = f(x_k + \sigma s_k)$, dann ist (2.28) äquivalent zu

$$\phi_k'(\sigma_k) \geq \theta \phi_k'(0).$$

Die Schrittweite muss also so groß sein, dass die Steigung von ϕ_k in σ_k mindestens das θ-fache der negativen Anfangs-Steigung $\phi_k'(0)$ ist. Dies legt mit der Armijo-Bedingung (2.27) den erlaubten Schrittweitenbereich fest. Siehe Abb. 2.3. □

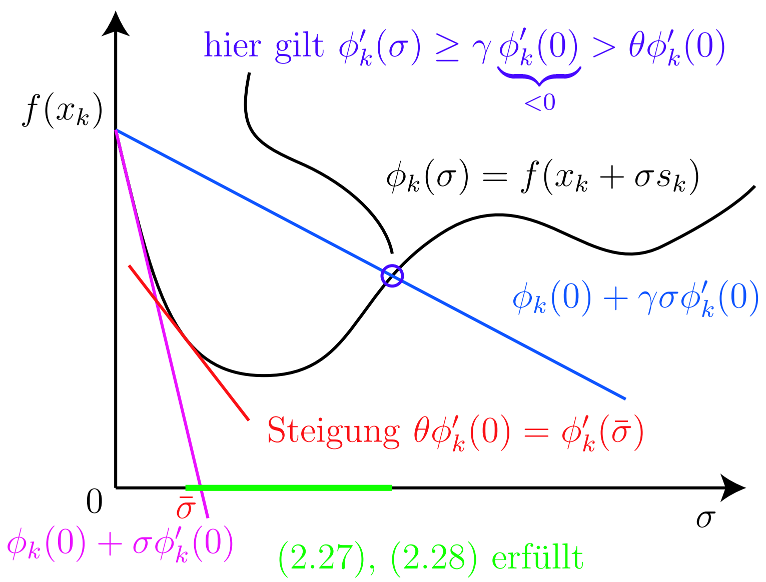


Abb. 2.3: Illustration der Powell-Wolfe-Regel.

Wir zeigen zunächst, dass die Powell-Wolfe-Regel wohldefiniert ist.

Lemma 2.6.2 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Seien weiter $0 < \gamma < \theta < 1$ beliebig fest. Dann gibt es zu jedem $x \in \mathbb{R}^n$ mit $\nabla f(x) \neq 0$ und zu jeder Abstiegsrichtung s von f in x, entlang der f nach unten beschränkt ist, also*

$$\inf_{\sigma>0} f(x+\sigma s) > -\infty$$

eine offene nichtleere Menge $\Sigma(x,s) \subset (0,\infty)$ mit

$$f(x) - f(x+\sigma s) \geq -\gamma\sigma\nabla f(x)^T s \tag{2.29}$$

$$\nabla f(x+\sigma s)^T s \geq \theta\nabla f(x)^T s \tag{2.30}$$

für alle $\sigma \in \Sigma(x,s)$.

Beweis: Betrachte die Funktion

$$\varphi(\sigma) = f(x+\sigma s) - f(x) - \gamma\sigma\nabla f(x)^T s.$$

Dann ist (2.29) äquivalent zu $\varphi(\sigma) \leq 0$. φ ist stetig differenzierbar mit

$$\varphi(0) = 0, \ \ \varphi'(0) = (1-\gamma)\nabla f(x)^T s < 0, \quad \lim_{\sigma\to\infty} \varphi(\sigma) = +\infty.$$

Also existieren $0 < \sigma_{min} < \sigma_{max}$ mit

$$\varphi(\sigma) < 0 \quad \forall\, \sigma \in (0, \sigma_{min}], \quad \varphi(\sigma) > 0 \quad \forall\, \sigma \geq \sigma_{max}.$$

Daher hat φ auf $[\sigma_{min}, \sigma_{max}]$ eine kleinste Nullstelle σ^*, denn die Menge der Nullstellen auf $[\sigma_{min}, \sigma_{max}]$ ist nichtleer, abgeschlossen und beschränkt. Für σ^* gilt nun

$$\varphi(\sigma^*) = 0, \quad \varphi(\sigma) < 0 \quad \forall\, \sigma \in (0, \sigma^*).$$

Also gilt (2.29) für alle $\sigma \in [0, \sigma^*]$ und zudem

$$\nabla f(x+\sigma^* s)^T s - \gamma\nabla f(x)^T s = \varphi'(\sigma^*) = \lim_{t\searrow 0} \frac{\varphi(\sigma^*) - \varphi(\sigma^* - t)}{t} \geq 0.$$

Dies ergibt

$$\nabla f(x+\sigma^* s)^T s \geq \gamma\nabla f(x)^T s > \theta\nabla f(x)^T s.$$

Somit gilt auch (2.30) für $\sigma = \sigma^*$ und aus Stetigkeitsgründen für alle $\sigma \in [\sigma^* - \varepsilon, \sigma^*]$ mit $\varepsilon > 0$ klein genug. □

Wir geben nun einen Algorithmus zur Bestimmung einer Schrittweite σ_k an, welche die Powell-Wolfe-Bedingungen (2.27), (2.28) erfüllt. Der Algorithmus beruht auf folgender Vorüberlegung:

Lemma 2.6.3 *Sind (σ_j^-) und (σ_j^+) Folgen mit $\sigma_j^- \nearrow \sigma^* > 0$, $\sigma_j^+ \searrow \sigma^*$, so dass die Armijo-Bedingung* (2.27) *für alle $\sigma_k = \sigma_j^-$ erfüllt und für alle $\sigma_k = \sigma_j^+$ verletzt ist, dann erfüllt $\sigma_k = \sigma_j^-$ die Powell-Wolfe-Bedingungen* (2.27), (2.28) *für alle j groß genug.*

Beweis: Setze $\varphi(\sigma) = f(x+\sigma s) - f(x) - \gamma\sigma\nabla f(x)^T s$. Wir haben $\varphi(\sigma_j^-) \le 0 < \varphi(\sigma_j^+)$, also mit dem Mittelwertsatz $\varphi'(\tilde{\sigma}_j) \ge 0$ für ein $\tilde{\sigma}_j \in [\sigma_j^-, \sigma_j^+]$. Grenzübergang liefert

$$0 \le \varphi'(\sigma^*) = \nabla f(x_k + \sigma^* s_k)^T s_k - \gamma g_k^T s_k < \nabla f(x_k + \sigma^* s_k)^T s_k - \theta g_k^T s_k.$$

Also gilt (2.28) für alle σ_k in einer offenen Umgebung von σ^*. □

Wir können Folgen (σ_j^-) und (σ_j^+) mit den Eigenschaften aus Lemma 2.6.3 ausgehend von einem $\sigma^- > 0$, das die Armijo-Bedingung erfüllt, und einem $\sigma^+ > \sigma^-$, das die Armijo-Bedingung verletzt, durch Bisektion gewinnen:

Algorithmus 4 Berechnung einer Powell-Wolfe-Schrittweite:

1. Wähle das maximale $\sigma^- \in \{1, 2^{-1}, \ldots\}$, so dass $\sigma_k = \sigma^-$ die Armijo-Bedingung (2.27) erfüllt. Setze $\sigma^+ := 2\sigma^-$.

   Falls $\sigma^- < 1$, gehe zu 3.

2. Wähle das minimale $\sigma^+ \in \{2^1, 2^2, \ldots\}$, so dass $\sigma_k = \sigma^+$ die Armijo-Bedingung (2.27) verletzt. Setze $\sigma^- := \sigma^+/2$.

3. Solange $\sigma_k = \sigma^-$ Bedingung (2.28) verletzt:

   Berechne $\sigma := (\sigma^+ + \sigma^-)/2$.
   Falls $\sigma_k = \sigma$ Bedingung (2.27) erfüllt, setze $\sigma^- := \sigma$, sonst setze $\sigma^+ := \sigma$.

4. STOP mit Ergebnis $\sigma_k := \sigma^-$.

Es ist nicht überraschend, dass der Algorithmus zum Ziel führt.

Satz 2.6.4 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Seien weiter $0 < \gamma < \theta < 1$ beliebig fest. Dann liefert Algorithmus 4 zu jedem $x_k \in \mathbb{R}^n$ mit $\nabla f(x_k) \neq 0$ und zu jeder Abstiegsrichtung s_k von f in x_k, entlang der f nach unten beschränkt ist, eine Schrittweite σ_k, die die Powell-Wolfe-Bedingungen* (2.27), (2.28) *erfüllt.*

Beweis: Übung. □

Wir zeigen abschließend, dass die Powell-Wolfe-Regel zulässige Schrittweiten erzeugt.

Satz 2.6.5 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und nach unten beschränkt. Sei $x_0 \in \mathbb{R}^n$ ein beliebiger Startpunkt, so dass ∇f gleichmäßig stetig auf der Niveaumenge $N_f(x_0)$ ist (z.B. erfüllt, falls $N_f(x_0)$ kompakt). Verwendet man im Abstiegsverfahren Algorithmus 3 die Powell-Wolfe-Regel und terminiert das Verfahren nicht endlich, dann ist die von der Powell-Wolfe-Regel erzeugte Schrittweitenfolge (σ_k) wohldefiniert und zulässig.*

Beweis: Übung. □

Bemerkung: Die zweite Powell-Wolfe-Bedingung (2.28) stellt sicher, dass gilt

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^T s_k > 0. \tag{2.31}$$

Denn (2.28) liefert

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^T s_k \geq (\theta - 1)\nabla f(x_k)^T s_k > 0.$$

Die Ungleichung (2.31) wird sich im Zusammenhang mit Quasi-Newton-Verfahren als wichtig erweisen. □

2.7 Das Newton-Verfahren

Das Newton-Verfahren ist eines der wichtigsten Verfahren der Numerik, da es die Basis von schnell lokal konvergenten Verfahren bildet. Das Newton-Verfahren kann sowohl zur Lösung linearer Gleichungssysteme als auch zur Minimierung nichtlinearer Funktionen verwendet werden.

Wir betrachten zunächst das Newton-Verfahren zur Lösung eines nichtlinearen Gleichungssystems

$$F(x) = 0 \tag{2.32}$$

mit $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar. Danach gehen wir auf das Newton-Verfahren zur Minimierung einer zweimal stetig differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ein. Es beruht auf den in §2.3 kurz angesprochenen Newton-Schritten, und erweist sich als dasselbe wie das Newton-Verfahren zur Lösung von

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

2.7.1 Das Newton-Verfahren für Gleichungssysteme

Zur Motivation des Newton-Verfahrens für (2.32) sei $x_k \in \mathbb{R}^n$ ein gegebener Punkt. Dann ist $\bar{x}$ eine Lösung von (2.32) genau dann, wenn $\bar{x} = x_k + s_k$ gilt mit einer Lösung $s = s_k$ von

$$F(x_k + s) = 0. \tag{2.33}$$

Die Idee des Newton-Verfahrens besteht darin, $F(x_k + s)$ durch die Taylorentwicklung erster Ordnung zu ersetzen: Es gilt

$$F(x_k + s) = F(x_k) + F'(x_k)s + o(\|s\|)$$

mit der Jacobi-Matrix $F'(x_k)$ von F in x_k und das Restglied wird für kurze s klein.

Bei der k-ten Iteration des Newton-Verfahrens ersetzt man daher (2.33) durch die linearisierte Gleichung

$$F(x_k) + F'(x_k)s = 0.$$

Dies ergibt

Algorithmus 5 Lokales Newton-Verfahren für Gleichungssysteme

Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$.

Für $k = 0, 1, \ldots$:

1. *Falls* $F(x_k) = 0$*: STOP mit Ergebnis* x_k.
2. *Berechne den Newton-Schritt* $s_k \in \mathbb{R}^n$ *durch Lösen der* Newton-Gleichung
$$F'(x_k)s_k = -F(x_k).$$
3. *Setze* $x_{k+1} = x_k + s_k$.

2.7.2 Superlineare und quadratische lokale Konvergenz des Newton-Verfahrens

Wir werden unter geeigneten Voraussetzungen die schnelle lokale Konvergenz des Newton-Verfahrens zeigen. Hierzu müssen wir zunächst einige Begriffe zur Charakterisierung von Konvergenzgeschwindigkeiten einführen:

Definition 2.7.1 (Konvergenzraten)

a) *Eine Folge* $(x_k) \subset \mathbb{R}^n$ *konvergiert* Q-linear mit Rate $\kappa \in (0,1)$ *gegen* $\bar{x}$*, wenn mit einem* $l \geq 0$ *gilt*
$$\|x_{k+1} - \bar{x}\| \leq \kappa \|x_k - \bar{x}\| \quad \forall\, k \geq l.$$

b) *Eine Folge* $(x_k) \subset \mathbb{R}^n$ *konvergiert* Q-superlinear *gegen* $\bar{x}$*, wenn* $x_k \to \bar{x}$ *gilt und zudem*
$$\|x_{k+1} - \bar{x}\| = o(\|x_k - \bar{x}\|) \quad \textit{für } k \to \infty.$$
Nach Definition bedeutet dies, dass gilt $\frac{\|x_{k+1}-\bar{x}\|}{\|x_k-\bar{x}\|} \to 0$.

c) *Eine Folge* $(x_k) \subset \mathbb{R}^n$ *konvergiert* Q-quadratisch *gegen* $\bar{x}$*, wenn* $x_k \to \bar{x}$ *gilt und zudem*
$$\|x_{k+1} - \bar{x}\| = O(\|x_k - \bar{x}\|^2) \quad \textit{für } k \to \infty.$$
Nach Definition bedeutet dies, dass es $C > 0$ *und* $l > 0$ *gibt mit*
$$\|x_{k+1} - \bar{x}\| \leq C\|x_k - \bar{x}\|^2 \quad \forall\, k \geq l.$$

Zur Analyse des Newton-Verfahrens benötigen wir das folgende Resultat:

Lemma 2.7.2 *Die Menge $\mathcal{I} \subset \mathbb{R}^{n,n}$ der invertierbaren Matrizen ist offen und die Abbildung $A \in \mathcal{I} \mapsto A^{-1}$ ist stetig, sogar unendlich oft stetig differenzierbar. Ist $A \in \mathcal{I}$, dann gilt auch $A + B \in \mathcal{I}$ für alle $B \in \mathbb{R}^{n,n}$ mit $\|A^{-1}B\| < 1$ (also insbesondere, falls $\|B\| < 1/\|A^{-1}\|$).*

Beweis: Bekanntlich gilt $A \in \mathcal{I}$ genau dann, wenn $\det(A) \neq 0$. Da $\det(A)$ ein Polynom in den Koeffizienten von A ist, ist die Abbildung $A \in \mathbb{R}^{n,n} \mapsto \det(A)$ unendlich oft stetig differenzierbar. Ist also $\det(A) \neq 0$, dann gibt es $\varepsilon > 0$ mit $\det(A + B) \neq 0$ für alle $B \in \mathbb{R}^{n,n}$ mit $\|B\| < \varepsilon$. Dies zeigt, dass $\mathcal{I}$ offen ist.

Aus der Cramerschen Regel folgt, dass zu $A = (a_{ij}) \in \mathcal{I}$ die Inverse A^{-1} gegeben ist durch

$$A^{-1} = \frac{1}{\det(A)}(a^*_{ij}) \quad \text{mit} \quad a^*_{ij} = (-1)^{i+j}\det(A_{ji}),$$

wobei A_{ij} die Matrix ist, die aus A durch Streichen der i-ten Zeile und j-ten Spalte entsteht. Die rechte Seite ist offensichtlich eine glatte Funktion auf $\mathcal{I}$.

Ist $A \in \mathcal{I}$ und $B \in \mathbb{R}^{n,n}$ beliebig mit $\|A^{-1}B\| < 1$, dann ist $A + B \in \mathcal{I}$, da gilt

$$\begin{aligned}\|A^{-1}(A+B)u\| = \|u + A^{-1}Bu\| &\geq \|u\| - \|A^{-1}Bu\| \\ &\geq (1 - \|A^{-1}B\|)\|u\| > 0 \quad \forall\, u \in \mathbb{R}^n \setminus \{0\}.\end{aligned}$$

□

Wir beweisen nun die superlineare lokale Konvergenz des Newton-Verfahrens.

Satz 2.7.3 (Schnelle lokale Konvergenz des Newton-Verfahrens)
Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und sei $\bar{x} \in \mathbb{R}^n$ ein Punkt mit $F(\bar{x}) = 0$ und $F'(\bar{x})$ nichtsingulär. Dann gibt es $\delta > 0$, so dass gilt:

i) *$F'(x)$ ist nichtsingulär mit $\|F'(x)^{-1}\| \leq 2\|F'(\bar{x})^{-1}\|$ für alle $x \in B_\delta(\bar{x})$.*

ii) *$\bar{x}$ ist die einzige Nullstelle von F auf $B_\delta(\bar{x})$ und es gilt*

$$\frac{1}{2\|F'(\bar{x})^{-1}\|}\|x - \bar{x}\| \leq \|F(x)\| \leq 2\|F'(\bar{x})\|\,\|x - \bar{x}\|$$

iii) *Für alle $x_0 \in B_\delta(\bar{x})$ terminiert Algorithmus 5 entweder mit $x_k = \bar{x}$ oder erzeugt eine Folge $(x_k) \subset B_\delta(\bar{x})$, die Q-superlinear gegen $\bar{x}$ konvergiert.*

iv) *Ist F' Lipschitz-stetig auf $B_\delta(\bar{x})$ mit Konstante L, dann konvergiert (x_k) Q-quadratisch gegen $\bar{x}$, wobei*

$$\|x_{k+1} - \bar{x}\| \leq \|F'(\bar{x})^{-1}\|L\|x_k - \bar{x}\|^2 \quad \forall\, k \geq 0.$$

Beweis: zu i): Da nach Lemma 2.7.2 die Menge $\mathcal{I}$ der invertierbaren Matrizen offen und die Abbildung $A \in \mathcal{I} \mapsto A^{-1}$ stetig ist, ist $x \mapsto F'(x)^{-1}$ wohldefiniert und stetig in einer Umgebung von $\bar{x}$. Daher finden wir aus Stetigkeitsgründen $\delta > 0$, so dass i) gilt.

zu ii): Wir können aus Stetigkeitsgründen $\delta > 0$ so verkleinern, dass neben i) gilt

$$\|F'(x)\| \leq 2\|F'(\bar{x})\| \quad \forall\, x \in B_\delta(\bar{x}).$$

Zunächst erinnern wir daran, dass mit $\phi(t) := F(\bar{x} + t(x - \bar{x}))$ für beliebiges $x \in B_\delta(\bar{x})$ gilt

$$F(x) - F(\bar{x}) = \phi(1) - \phi(0) = \int_0^1 \phi'(t)\, dt = \int_0^1 F'(\bar{x} + t(x - \bar{x}))(x - \bar{x})\, dt. \tag{2.34}$$

Nun gilt für jede stetige Abbildung $t \in [0,1] \mapsto v(t) \in \mathbb{R}^n$

$$\left\| \int_0^1 v(t)\, dt \right\| \leq \int_0^1 \|v(t)\|\, dt \tag{2.35}$$

(Beweis z.B. durch Approximation mit Riemann-Summen und Anwendung der Dreiecksungleichung). Dies liefert mit (2.34)

$$\|F(x) - F(\bar{x})\| \leq \int_0^1 \underbrace{\|F'(\bar{x} + t(x - \bar{x}))\|}_{\leq 2\|F'(\bar{x})\|} \|x - \bar{x}\|\, dt \leq 2\|F'(\bar{x})\|\,\|x - \bar{x}\| \quad \forall\, x \in B_\delta(\bar{x}).$$

Weiter können wir aus Stetigkeitsgründen $\delta > 0$ so klein wählen, dass zudem gilt

$$\|I - F'(x)^{-1}F'(y)\| \leq \frac{1}{2} \quad \forall\, x, y \in B_\delta(\bar{x}). \tag{2.36}$$

Insbesondere ist dann für alle $x, y \in B_\delta(\bar{x})$

$$\|s\|^2 - s^T F'(x)^{-1}F'(y)s \leq \|I - F'(x)^{-1}F'(y)\|\,\|s\|^2 \leq \frac{1}{2}\|s\|^2 \quad \forall\, s \in \mathbb{R}^n,$$

also $s^T F'(x)^{-1}F'(y)s \geq \frac{1}{2}\|s\|^2$. Skalarmultiplikation von (2.34) mit $F'(\bar{x})^{-T}(x - \bar{x})$ ergibt nun

$$(x-\bar{x})^T F'(\bar{x})^{-1}(F(x)-F(\bar{x})) = \int_0^1 (x-\bar{x})^T F'(\bar{x})^{-1}F'(\bar{x}+t(x-\bar{x}))(x-\bar{x})\, dt \geq \frac{1}{2}\|x - \bar{x}\|^2,$$

also

$$\frac{1}{2}\|x - \bar{x}\|^2 \leq \|F'(\bar{x})^{-1}\|\,\|x - \bar{x}\|\,\|F(x) - F(\bar{x})\|.$$

Division durch $\|F'(\bar{x})^{-1}\|\,\|x - \bar{x}\|$ liefert die zweite Ungleichung.

zu iii): Wir zeigen induktiv

$$
\|x_{k+1}-\bar{x}\| \le \int_0^1 \|I - F'(x_k)^{-1}F'(\bar{x}+t(x_k-\bar{x}))\|\,dt\,\|x_k-\bar{x}\| \begin{cases} \le \frac{1}{2}\|x_k-\bar{x}\|, \\ = o(\|x_k-\bar{x}\|) \quad \text{für } k\to\infty. \end{cases} \tag{2.37}
$$

Sei $x_k \in B_\delta(\bar{x})$ (für $k=0$ ist das erfüllt). Dann gilt wegen $F(\bar{x})=0$ und (2.34)

$$
\begin{aligned}
x_{k+1}-\bar{x} &= x_k-\bar{x}-F'(x_k)^{-1}F(x_k) = x_k-\bar{x}-F'(x_k)^{-1}(F(x_k)-F(\bar{x})) \\
&= x_k-\bar{x}-F'(x_k)^{-1}\int_0^1 F'(\bar{x}+t(x_k-\bar{x}))(x_k-\bar{x})\,dt \\
&= \int_0^1 (I-F'(x_k)^{-1}F'(\bar{x}+t(x_k-\bar{x})))(x_k-\bar{x})\,dt.
\end{aligned}
$$

Zusammen mir (2.35), (2.36) ergibt sich

$$
\|x_{k+1}-\bar{x}\| \le \int_0^1 \|I-F'(x_k)^{-1}F'(\bar{x}+t(x_k-\bar{x}))\|\,dt\,\|x_k-\bar{x}\| \le \frac{1}{2}\|x_k-\bar{x}\|,
$$

also die erste und zweite Ungleichung in (2.37). Insbesondere folgt induktiv $(x_k)\subset B_\delta(\bar{x})$ und $x_k\to\bar{x}$. Der mittlere Term in (2.37) ist offensichtlich $o(\|x_k-\bar{x}\|)$ für $k\to\infty$, da aus Stetigkeitsgründen

$$
\int_0^1 \|I-F'(x_k)^{-1}F'(\bar{x}+t(x_k-\bar{x}))\|\,dt \to 0 \quad \text{für } x_k\to\bar{x}.
$$

(2.36) zeigt nun neben der Q-linearen die Q-superlineare Konvergenz.

zu iv): Die Q-quadratische Konvergenz folgt nun aus

$$
\begin{aligned}
\|I-F'(x_k)^{-1}F'(\bar{x}+t(x_k-\bar{x}))\| &\le \|F'(x_k)^{-1}\|\,\|F'(x_k)-F'(\bar{x}+t(x_k-\bar{x}))\|_* \\
&\le 2\|F'(\bar{x})^{-1}\|L(1-t)\|x_k-\bar{x}\|,
\end{aligned}
$$

also

$$
\begin{aligned}
\int_0^1 \|I-F'(x_k)^{-1}F'(\bar{x}+t(x_k-\bar{x})))(x_k-\bar{x})\|\,dt &\le 2\|F'(\bar{x})^{-1}\|L\|x_k-\bar{x}\|^2\int_0^1(1-t)\,dt \\
&= \|F'(\bar{x})^{-1}\|L\|x_k-\bar{x}\|^2.
\end{aligned}
$$

□

2.7.3 Das Newton-Verfahren für Minimierungsprobleme

Wie bereits erwähnt, kann das Newton-Verfahren auch zur Bestimmung eines lokalen Minimums des Problems

$$\text{(P)} \qquad \min_{x\in\mathbb{R}^n} \; f(x).$$

mit einer zweimal stetig differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$ verwendet werden.

Das Verfahren kann auf zwei Arten motiviert werden: Jedes lokale Minimum $\bar{x}$ von (P) ist nach Satz 2.1.2 ein stationärer Punkt, also Lösung der Gleichung

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

Ist $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, dann ist $F = \nabla f$ stetig differenzierbar mit $F' = \nabla^2 f$. Anwendung des Newton-Verfahrens für Gleichungssysteme liefert also $x_{k+1} = x_k + s_k$, wobei s_k als Lösung der Newton-Gleichung

$$\nabla^2 f(x_k) s_k = -\nabla f(x_k)$$

gegeben ist. Dies ergibt folgendes Verfahren:

Algorithmus 6 Lokales Newton-Verfahren für Optimierungsprobleme

Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$.

Für $k = 0, 1, \dots :$

1. *Falls* $\nabla f(x_k) = 0$*: STOP mit Ergebnis* x_k.
2. *Berechne den Newton-Schritt* $s_k \in \mathbb{R}^n$ *durch Lösen der* Newton-Gleichung

$$(2.38) \qquad \nabla^2 f(x_k) s_k = -\nabla f(x_k).$$

3. *Setze* $x_{k+1} = x_k + s_k$.

Alternativ können wir das Newton-Verfahren für (P) auch wie folgt motivieren: Sei $\bar{x} \in \mathbb{R}^n$ ein Punkt, in dem die hinreichende Bedinung zweiter Ordnung aus Satz 2.1.5 gilt, also $\nabla f(\bar{x}) = 0$, $\nabla^2 f(\bar{x})$ positiv definit. Dann ist $\nabla^2 f(x)$ positiv definit auf einer Umgebung $B_\varepsilon(\bar{x})$ von $\bar{x}$. Sei nun $x_k \in B_\varepsilon(\bar{x})$ eine gegebene Iterierte. Wir approximieren f in x_k durch das Taylor-Polynom zweiter Ordnung

$$f(x_k + s) \approx \underbrace{f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s}_{=:q_k(s)} + o(\|s_k\|^2).$$

Da $\nabla^2 f(x_k)$ positiv definit ist, ist q_k streng konvex und hat ein eindeutiges globales Minimum s_k gegeben durch

$$\nabla q_k(s_k) = \nabla^2 f(x_k)s_k + \nabla f(x_k) = 0.$$

Dies ist genau die Newton-Gleichung (2.38).

Spezialisieren wir den Konvergenzsatz 2.7.3 auf den Fall $F = \nabla f$, dann erhalten wir

Satz 2.7.4 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Erfüllt $\bar{x}$ die hinreichende Bedinung zweiter Ordnung aus Satz 2.1.5 für ein lokales Minimum, dann gibt es $\delta > 0$, so dass gilt:*

i) $\nabla^2 f(x)$ *ist positiv definit mit*

$$\|\nabla^2 f(x)^{-1}\| \le 2\|\nabla^2 f(\bar{x})^{-1}\| = \frac{2}{\lambda_{min}(\nabla^2 f(\bar{x}))}$$

für alle $x \in B_\delta(\bar{x})$.

ii) $\bar{x}$ *ist der einzige stationäre Punkt von f auf $B_\delta(\bar{x})$ und es gilt*

$$\frac{\lambda_{min}(\nabla^2 f(\bar{x}))}{2}\|x - \bar{x}\| \le \|\nabla f(x)\| \le 2\lambda_{max}(\nabla^2 f(\bar{x}))\|x - \bar{x}\| \quad \forall\, x \in B_\delta(\bar{x}).$$

iii) Für alle $x_0 \in B_\delta(\bar{x})$ terminiert Algorithmus 6 entweder mit $x_k = \bar{x}$ oder erzeugt eine Folge $(x_k) \subset B_\delta(\bar{x})$, die Q-superlinear gegen $\bar{x}$ konvergiert.

iv) Ist $\nabla^2 f$ Lipschitz-stetig auf $B_\delta(\bar{x})$ mit Konstante L, dann konvergiert (x_k) Q-quadratisch gegen $\bar{x}$, wobei

$$\|x_{k+1} - \bar{x}\| \le \frac{L}{\lambda_{min}(\nabla^2 f(\bar{x}))}\|x_k - \bar{x}\|^2 \quad \forall\, k \ge 0.$$

Beweis: Der Beweis folgt direkt aus Satz 2.7.3 zusammen mit den folgenden Hilfsresultaten. □

Lemma 2.7.5 *Für eine symmetrische, positiv definite Matrix $Q \in \mathbb{R}^{n,n}$ gilt*

$$\|Q\| = \lambda_{max}(Q), \quad \|Q^{-1}\| = \frac{1}{\lambda_{min}(Q)},$$

wobei $\lambda_{min}(Q)$ den kleinsten und $\lambda_{max}(Q)$ den größten Eigenwert von Q bezeichnet.

Beweis: Mit einer orthogonalen Matrix $U \in \mathbb{R}^{n,n}$ gilt $Q = U^T DU$, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ mit den Eigenwerten $0 < \lambda_1 \le \cdots \le \lambda_n$ von Q und daher mit $v = Us$

$$\|Qs\| = \|U^T DUs\|_{=}\|U^T Dv\| = \|Dv\| = \sqrt{\sum_{i=1}^{n} \lambda_i^2 v_i^2} \le \lambda_{max}(Q)\|v\| = \lambda_{max}(Q)\|u\|.$$

Gleichheit ergibt sich, wenn man $s = u_n$ mit einem Eigenvektor u_n von Q zu $\lambda_{max}(Q)$ wählt. Wegen $Q^{-1} = U^T D^{-1} U$ folgt analog $\|Q^{-1}\| = \frac{1}{\lambda_{min}(Q)}$. $\square$

Lemma 2.7.6 *Es sei $Q \in \mathbb{R}^{n,n}$ symmetrisch positiv definit. Dann ist $Q + B$ positiv definit für alle $B \in \mathbb{R}^{n,n}$ mit $\|B\| < \lambda_{min}(Q)$. Zudem gilt*

$$s^T(Q + B)s \ge \frac{1}{2} s^T Qs \ge \frac{\lambda_{min}(Q)}{2}\|s\|^2 \quad \forall\, s \in \mathbb{R}^n \ \ \forall\, B \in \mathbb{R}^{n,n}, \|B\| \le \frac{\lambda_{min}(Q)}{2}.$$

Beweis: Wir haben wie am Ende des Beweises von Satz 2.4.5 $s^T Qs \ge \lambda_{min}(Q)\|s\|^2$ für alle $s \in \mathbb{R}^n$. Nun gilt für alle $B \in \mathbb{R}^{n,n}$ mit $\|B\| < \lambda_{min}(Q)$

$$s^T(Q + B)s \ge s^T Qs - |s^T Bs| \ge \lambda_{min}(Q)\|s\|^2 - \|B\|\|s\|^2 > 0 \quad \forall\, s \in \mathbb{R}^n \setminus \{0\}.$$

Ist $\|B\| \le \frac{\lambda_{min}(Q)}{2}$, dann erhalten wir analog

$$s^T(Q+B)s \ge s^T Qs - \|B\|\|s\|^2 \ge s^T Qs - \frac{\lambda_{min}(Q)}{2}\|s\|^2 \ge \frac{1}{2} s^T Qs \ge \frac{\lambda_{min}(Q)}{2}\|s\|^2 \quad \forall\, s \in \mathbb{R}^n.$$

$\square$

2.7.4 Globalisierung des Newton-Verfahrens

Selbst für streng konvexe Funktionen ist das lokale Newton-Verfahren aus Algorithmus 6 nicht für jeden Startpunkt x_0 konvergent.

Beispiel 2.7.1 Betrachte $f(x) = |x| - \arctan(|x|)$. f ist zweimal stetig differenzierbar und streng konvex. Man kann zeigen, dass das Newton-Verfahren divergiert, falls $|x_0|$ zu groß ist, siehe Übung.

Um die globale Konvergenz des Newton-Verfahrens zu garantieren und die schnelle lokale Konvergenz zu erhalten, verwenden wir den Newton-Schritt als Basis der Schrittberechnung im allgemeinen Abstiegsverfahren, wobei wir folgendes sicherstellen:

- Die Folgen (s_k) und (σ_k) sind zulässig.

- Sobald x_k nahe genug bei einem Punkt $\bar{x}$ liegt, der die hinreichende Bedingung zweiter Ordnung erfüllt, dann ist s_k der Newton-Schritt und $\sigma_k = 1$.

Der folgende Algorithmus eröffnet vielseitige Möglichkeiten, dies zu erreichen.

Algorithmus 7 Globalisiertes Newton-Verfahren

Wähle Konstanten $\gamma \in (0, 1/2)$, $c_1 \in (0, 1)$, $c_2 > 0$ und $p > 0$. Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$.

Für $k = 0, 1, \ldots$:

1. *Falls $\nabla f(x_k) = 0$: STOP mit Ergebnis x_k.*
2. *Berechne–falls möglich–den Newton-Schritt s_k^N als Lösung von*
$$\nabla^2 f(x_k) s_k^N = -\nabla f(x_k).$$
Ist dies möglich und gilt
$$-\nabla f(x_k)^T s_k^N \geq \min\{c_1, c_2 \|\nabla f(x_k)\|^p\} \|\nabla f(x_k)\| \|s_k^N\|,$$
dann setze $s_k = s_k^N$.
Sonst wähle eine symmetrische, positiv definite Matrix $B_k \in \mathbb{R}^{n,n}$, so dass $\lambda_{min}(B_k) \geq \mu$, $\lambda_{max}(B_k) \leq \eta$ mit von k unabhängigen Konstanten $0 < \mu < \eta$ gilt, und bestimme s_k als Lösung von
$$B_k s_k = -\nabla f(x_k).$$
3. *Bestimme eine Schrittweite $\sigma_k > 0$ nach der Armijo-Regel.*
4. *Setze $x_{k+1} = x_k + \sigma_k s_k$.*

Bemerkung: Es ist zu beachten, dass der Parameter γ in der Armijo-Bedingung nun im Intervall $(0, 1/2)$ gewählt wird! Dies wird sicherstellen, dass der Newton-Schritt schließlich mit Schrittweite $\sigma_k = 1$ verwendet wird. □

Bemerkung: Eine einfache Wahl ist $B_k = I$. Häufig verwendet man

$$B_k = \nabla^2 f(x_k) + (\mu_k + \max(0, -\lambda_{min}(\nabla^2 f(x_k))))I$$

mit $0 < \tilde{\mu} \leq \mu_k \leq \tilde{\eta}$ und Konstanten $0 < \tilde{\mu} < \tilde{\eta}$. Diese Wahl erfüllt die Voraussetzungen, wenn die Hessematrizen $\nabla^2 f(x_k)$ gleichmäßig beschränkt bleiben. Dies ist zum Beispiel gegeben, wenn $N_f(x_0)$ kompakt ist. □

Es gilt der folgende globale Konvergenzsatz.

Satz 2.7.7 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Ist $N_f(x_0)$ kompakt und terminiert Algorithmus 7 nicht endlich, dann sind die erzeugten Folgen (s_k) von Suchrichtungen und (σ_k) von Schrittweiten zulässig. Insbesondere gelten die Konvergenzaussagen von Satz 2.5.6 und Satz 2.5.7.*

Beweis: Der Algorithmus terminiere nicht endlich. Wir zeigen zunächst die Zulässigkeit der Suchrichtungen, indem wir die verallgemeinerte Winkelbedingung (2.19) nachweisen: Es gilt entweder $s_k = s_k^N$ und

$$\frac{-\nabla f(x_k)^T s_k}{\|s_k\|} \geq \min\{c_1, c_2 \|\nabla f(x_k)\|^p\} \|\nabla f(x_k)\|$$

oder es ist $B_k s_k = -\nabla f(x_k)$. Im zweiten Fall ist

$$\|\nabla f(x_k)\| = \|B_k s_k\| \leq \eta \|s_k\|$$

und somit

$$\frac{-\nabla f(x_k)^T s_k}{\|s_k\|} = \frac{s_k^T B_k s_k}{\|s_k\|} \geq \frac{\mu \|s_k\|^2}{\|s_k\|} = \mu \|s_k\| \geq \frac{\mu}{\eta} \|\nabla f(x_k)\|.$$

Insgesamt ergibt sich

$$\frac{-\nabla f(x_k)^T s_k}{\|s_k\|} \geq \min\{\mu/\eta, c_1, c_2 \|\nabla f(x_k)\|^p\} \|\nabla f(x_k)\|.$$

Mit der stetigen, streng monoton wachsenden Funktion $\varphi : [0, \infty) \to [0, \infty)$,

$$\varphi(t) = \min\{\mu/\eta, c_1, c_2 t^p\} t$$

gilt also (2.19) und die Zulässigkeit von (s_k) folgt aus Satz 2.5.2.

Da die Armijo-Regel verwendet wird und $N_f(x_0)$ kompakt ist, folgt nach Satz 2.6.1 die Zulässigkeit der Schrittweiten, wenn mit einer streng monoton wachsenden Funktion $\varphi : [0, \infty) \to [0, \infty)$ gilt

$$\|s_k\| \geq \varphi\left(\frac{-\nabla f(x_k)^T s_k}{\|s_k\|}\right). \tag{2.23}$$

Zum Nachweis stellen wir zunächst fest, dass $\|\nabla^2 f(x_k)\| \leq C$ für alle k mit einem $C > 0$ gilt, da $N_f(x_0)$ kompakt ist. Nun ist entweder $s_k = s_k^N$, also

$$\|s_k\| \geq \frac{\|\nabla f(x_k)\|}{\|\nabla^2 f(x_k)\|} \geq \frac{\|\nabla f(x_k)\|}{C}$$

oder $B_k s_k = -\nabla f(x_k)$, also

$$\|s_k\| \geq \frac{\|\nabla f(x_k)\|}{\|B_k\|} \geq \frac{\|\nabla f(x_k)\|}{\eta}.$$

In beiden Fällen ergibt sich

$$\|s_k\| \geq \min(1/C, 1/\eta)\|\nabla f(x_k)\| \geq \min(1/C, 1/\eta)\frac{-\nabla f(x_k)^T s_k}{\|s_k\|},$$

wobei wir die Cauchy-Schwarzsche Ungleichung im letzten Schritt verwendet haben. Somit gilt (2.23) mit $\varphi(t) = \min(1/C, 1/\eta)\, t$ und die Zulässigkeit der Schrittweiten folgt aus Satz 2.6.1.

Wegen der Zulässigkeit von (s_k) und (σ_k) gilt Satz 2.5.6 und zudem Satz 2.5.7, da $N_f(x_0)$ kompakt ist. □

2.7.5 Übergang zu schneller lokaler Konvergenz

Wir zeigen nun, dass das globalisierte Newton-Verfahren aus Algorithmus 7 in das lokale Newton-Verfahren übergeht, wenn (x_k) einen Häufungspunkt $\bar{x}$ hat, in dem die Hessematrix positiv definit ist.

Satz 2.7.8 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $N_f(x_0)$ sei kompakt für den Startpunkt $x_0 \in \mathbb{R}^n$. Terminiert Algorithmus 7 nicht endlich und hat die erzeugte Folge (x_k) einen Häufungspunkt $\bar{x}$, in dem die Hessematrix positiv definit ist, dann gilt:*

i) $\lim_{k\to\infty} x_k = \bar{x}$ *und $\bar{x}$ ist ein isoliertes lokales Minimum, in dem die hinreichende Bedinung zweiter Ordnung erfüllt ist.*

ii) Es gibt $l \geq 0$ mit $s_k = s_k^N$ und $\sigma_k = 1$ (beachte $\gamma \in (0, 1/2)$!) für alle $k \geq l$, das Verfahren geht also in das lokale Newton-Verfahren über. Inbesondere konvergiert (x_k) Q-superlinear gegen $\bar{x}$ und sogar Q-quadratisch, wenn $\nabla^2 f$ Lipschitz-stetig in einer Umgebung von $\bar{x}$ ist.

Beweis: Nach Satz 2.7.7 gelten die Konvergenzaussagen von Satz 2.5.6. Daher ist der Häufungspunkt $\bar{x}$ stationär, es gilt also $\nabla f(\bar{x}) = 0$.

Wir zeigen zunächst, dass für $\varepsilon > 0$ klein genug gilt

$$s_k = s_k^N \quad \forall\, k \text{ mit } x_k \in B_\varepsilon(\bar{x}). \tag{2.39}$$

Setze

$$\tilde{\mu} = \frac{\lambda_{min}(\nabla^2 f(\bar{x}))}{2}, \quad \tilde{\eta} = 2\lambda_{max}(\nabla^2 f(\bar{x})).$$

Zunächst finden wir nach Lemma 2.7.6 $\varepsilon > 0$ mit

$$s^T \nabla^2 f(x) s \geq \frac{\lambda_{min}(\nabla^2 f(\bar{x}))}{2}\|s\|^2 = \tilde{\mu}\|s\|^2 \quad \forall\, x \in B_\varepsilon(\bar{x}),\ \forall\, s \in \mathbb{R}^n. \tag{2.40}$$

Insbesondere gilt dann $\lambda_{min}(\nabla^2 f(x)) \geq \tilde{\mu}$ und

$$\|\nabla^2 f(x)^{-1}\| = \frac{1}{\lambda_{min}(\nabla^2 f(x))} \leq \frac{1}{\tilde{\mu}} \quad \forall\, x \in B_\varepsilon(\bar{x}). \tag{2.41}$$

Zudem können wir $\varepsilon > 0$ so verkleinern, dass

$$\|\nabla^2 f(x)\| \leq 2\|\nabla^2 f(\bar{x})\| = \tilde{\eta} \quad \forall\, x \in B_\varepsilon(\bar{x}). \tag{2.42}$$

Daher ist für alle k mit $x_k \in B_\varepsilon(\bar{x})$ die Hessematrix $\nabla^2 f(x_k)$ positiv definit und der Newton-Schritt erfüllt

$$\|s_k^N\| \geq \frac{\|\nabla f(x_k)\|}{\|\nabla^2 f(x_k)\|} \geq \frac{\|\nabla f(x_k)\|}{\tilde{\eta}}.$$

Dies ergibt mit (2.40)

$$-\nabla f(x_k)^T s_k^N = (s_k^N)^T \nabla^2 f(x_k) s_k^N \geq \tilde{\mu}\|s_k^N\|^2 \geq \frac{\tilde{\mu}}{\tilde{\eta}}\|\nabla f(x_k)\|\|s_k^N\|. \tag{2.43}$$

Wegen $\nabla f(\bar{x}) = 0$ können wir $\varepsilon > 0$ so verkleinern, dass gilt

$$c_2\|f(x)\|^p \leq \frac{\tilde{\mu}}{\tilde{\eta}} \quad \forall\, x \in B_\varepsilon(\bar{x}).$$

Dann ergibt sich zusammen mit (2.43)

$$-\nabla f(x_k)^T s_k^N \geq \min\left(c_1, c_2\|f(x_k)\|^p\right) \|\nabla f(x_k)\|\|s_k^N\| \quad \forall\, k \text{ mit } x_k \in B_\varepsilon(\bar{x})$$

und somit $s_k = s_k^N$ im Falle $x_k \in B_\varepsilon(\bar{x})$.

zu i): Wie bereits zu Beginn festgestellt, gelten nach Satz 2.7.7 die Konvergenzaussagen von Satz 2.5.6. Daher ist der Häufungspunkt $\bar{x}$ stationär. Da nach Voraussetzung $\nabla^2 f(\bar{x})$ positiv definit ist, erfüllt $\bar{x}$ also die hinreichende Bedinung zweiter Ordnung. Nach Satz 2.7.4 ist $\bar{x}$ daher ein isolierter stationärer Punkt. Insbesondere ist $\bar{x}$ ein isolierter Häufungspunkt von (x_k), da nach Satz 2.7.7 jeder Häufungspunkt stationär ist.

Nach Satz 2.5.6 iv) gilt nun $x_k \to \bar{x}$, falls für jede Teilfolge $(x_k)_{k\in K} \subset (x_k)$ mit $(x_k)_{k\in K} \to \bar{x}$ zudem gilt $(x_{k+1} - x_k)_{k\in K} \to 0$. Dies ist hier der Fall: Sei $(x_k)_{k\in K}$ eine beliebige Teilfolge mit $(x_k)_{k\in K} \to \bar{x}$. Dann finden wir $l \geq 0$ mit $x_k \in B_\varepsilon(\bar{x})$ für alle $k \in K$, $k \geq l$ und somit gilt wegen (2.39) und (2.41)

$$s_k = s_k^N, \quad \|s_k\| \leq \|\nabla^2 f(x_k)^{-1}\|\|\nabla f(x_k)\| \leq \frac{1}{\tilde{\mu}}\|\nabla f(x_k)\| \quad \forall\, k \in K, k \geq l.$$

Dies zeigt wegen $\sigma_k \leq 1$

$$\|x_{k+1} - x_k\| \leq \|s_k\| \leq \frac{1}{\tilde{\mu}}\|\nabla f(x_k)\| \to 0 \quad \text{für } K \ni k \to \infty.$$

Also ist Satz 2.5.6, iii) anwendbar und es folgt $x_k \to \bar{x}$.

zu ii): Nach i) gilt $x_k \to \bar{x}$. Daher existiert ein $l' \geq 0$ mit $x_k \in B_\varepsilon(\bar{x})$ und somit $s_k = s_k^N$ wegen (2.39) für alle $k \geq l'$. Es bleibt zu zeigen, dass die Armijo-Regel die Schrittweite $\sigma_k = 1$ für alle $k \geq l$ mit einem $l \geq l'$ liefert.

Die Armijo-Bedingung (2.7) ist für $\sigma_k = 1$ genau dann erfüllt, wenn gilt $f(x_k) - f(x_k + s_k) + \gamma \nabla f(x_k)^T s_k \geq 0$. Taylorentwicklung liefert für alle $x_k \in B_\varepsilon(\bar{x})$ wegen (2.40) und $\nabla^2 f(x_k) s_k = -\nabla f(x_k)$ mit einem $\tau_k \in [0, 1]$

$$\begin{aligned}
&f(x_k) - f(x_k + s_k) + \gamma \nabla f(x_k)^T s_k = (\gamma - 1)\nabla f(x_k)^T s_k - \frac{1}{2} s_k^T \nabla^2 f(x_k + \tau_k s_k) s_k \\
&= \underbrace{\left(1 - \gamma - \frac{1}{2}\right)}_{>0,\ \text{da}\ \gamma \in (0,1/2)} s_k^T \nabla^2 f(x_k) s_k + \frac{1}{2} s_k^T (\nabla^2 f(x_k) - \nabla^2 f(x_k + \tau_k s_k)) s_k \\
&\geq \left(1 - \gamma - \frac{1}{2}\right) \tilde{\mu} \|s_k\|^2 - \frac{1}{2} \|\nabla^2 f(x_k + \tau_k s_k) - \nabla^2 f(x_k)\| \|s_k\|^2.
\end{aligned}$$

Wegen $x_k \to \bar{x}$ und $\|s_k\| \leq \|\nabla^2 f(x_k)^{-1}\| \|\nabla f(x_k)\| \leq \frac{1}{\tilde{\mu}} \|\nabla f(x_k)\| \to 0$ für $k \to \infty$ finden wir $l \geq l'$ mit

$$\frac{1}{2} \|\nabla^2 f(x_k + \tau_k s_k) - \nabla^2 f(x_k)\| \leq \left(1 - \gamma - \frac{1}{2}\right) \tilde{\mu} \quad \forall\, k \geq l$$

(beachte $0 < \gamma < 1/2$). Dann folgt

$$f(x_k) - f(x_k + s_k) + \gamma \nabla f(x_k)^T s_k \geq 0 \quad \forall\, k \geq l$$

und somit erfüllt $\sigma_k = 1$ die Armijo-Bedingung für alle $k \geq l$. Daher haben wir $x_{k+1} = x_k + s_k^N$ für alle $k \geq l$ und die Q-superlineare bzw. Q-quadratische Konvergenz folgt aus Satz 2.7.4. □

2.8 Newton-artige Verfahren

Wir betrachten nun allgemein Verfahren zur Lösung von Gleichungssystemen $F(x) = 0$, bei denen der Schritt durch eine Newton-artige Gleichung

$$M_k s_k = -F(x_k)$$

mit geeignet gewählten nichtsingulären Matrizen $M_k \in \mathbb{R}^{n,n}$ berechnet wird. Dies führt auf folgendes Verfahren.

Algorithmus 8 Lokales Newton-artiges Verfahren für Gleichungssysteme

Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$.

Für $k = 0, 1, \ldots :$

1. *Falls $F(x_k) = 0$: STOP mit Ergebnis x_k.*
2. *Wähle eine nichtsinguläre Matrix $M_k \in \mathbb{R}^{n,n}$.*
3. *Berechne $s_k \in \mathbb{R}^n$ durch Lösen der* Newton-artigen Gleichung
$$M_k s_k = -F(x_k).$$
4. *Setze $x_{k+1} = x_k + s_k$.*

Bei einem Newton-artigen Verfahren für das Minimierungsproblem (P) setzen wir $F = \nabla f$ und wählen M_k symmetrisch (üblicherweise positiv definit):

Algorithmus 9 Lokales Newton-artiges Verfahren für Minimierungsprobleme

Algorithmus 8 mit $F(x_k) = \nabla f(x_k)$ und M_k symmetrisch, nichtsingulär.

Die Verwendung von Newton-artigen Verfahren ist in mehrerlei Hinsicht interessant:

- In vielen praxisrelevanten Anwendungen ist die Berechnung von $F(x)$ (bzw. $\nabla f(x)$) bereits sehr aufwendig, und die Berechnung von $F'(x)$ (bzw. $\nabla^2 f(x)$) nicht praktikabel. Es liegt dann nahe, $F'(x_k)$ durch eine leichter zu berechnende Matrix M_k zu approximieren.
- Bei großen Problemen ist eine exakte Lösung der Newton-Gleichung zu aufwendig, eine inexakte Lösung mit Hilfe iterativer Löser (vorkonditionierte CG-Verfahren, Mehrgitterlöser, ...) ist aber möglich. Tatsächlich ist dann s_k Lösung einer Newton-artigen Gleichung $M_k s_k = -\nabla f(x_k)$, wobei M_k eine Näherung von $F'(x_k)$ ist.

2.8.1 Superlineare Konvergenz und Dennis-Moré-Bedingung

Wir wollen charakterisieren, unter welchen Voraussetzungen Algorithmus 8 superlinear konvergiert. Wir haben zunächst das folgende allgemeine Resultat:

Satz 2.8.1 *Es sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar. Weiter sei $\bar{x} \in \mathbb{R}^n$ ein Punkt mit $F(\bar{x}) = 0$ und $F'(\bar{x})$ invertierbar. Ist (x_k) eine Folge, die gegen $\bar{x}$ konvergiert, dann sind folgende Aussagen äquivalent:*

i) (x_k) *konvergiert Q-superlinear gegen* $\bar{x}$.

ii) $\lim_{k\to\infty} x_k = \bar{x}$ *und es gilt* $\|F(x_k) + F'(\bar{x})(x_{k+1} - x_k)\| = o(\|x_{k+1} - x_k\|) \to 0$ *für* $k \to \infty$.

iii) $\lim_{k\to\infty} x_k = \bar{x}$ *und es gilt* $\|F(x_k) + F'(x_k)(x_{k+1} - x_k)\| = o(\|x_{k+1} - x_k\|) \to 0$ *für* $k \to \infty$*.*

Beweis: i) $\Longrightarrow$ ii): Dann ist $\|x_{k+1} - \bar{x}\| = o(\|x_k - \bar{x}\|)$. Wegen $F(\bar{x}) = 0$ ergibt sich

$$\begin{aligned}\|F(x_k) + F'(\bar{x})(x_{k+1} - x_k)\| &= \|F(x_k) - F(\bar{x}) - F'(\bar{x})(x_k - \bar{x}) + F'(\bar{x})(x_{k+1} - \bar{x})\| \\ &= o(\|x_k - \bar{x}\|) + \|F'(\bar{x})(x_{k+1} - \bar{x})\| \\ &= o(\|x_k - \bar{x}\|) + o(\|x_k - \bar{x}\|) = o(\|x_k - \bar{x}\|).\end{aligned}$$

Nun ist

$$\|x_k - \bar{x}\| = \begin{cases} \leq \|x_{k+1} - x_k\| + \|x_{k+1} - \bar{x}\| = \|x_{k+1} - x_k\| + o(\|x_k - \bar{x}\|), \\ \geq \|x_{k+1} - x_k\| - \|x_{k+1} - \bar{x}\| = \|x_{k+1} - x_k\| + o(\|x_k - \bar{x}\|). \end{cases}$$

Somit gilt für k groß genug $1/2\|x_{k+1} - x_k\| \leq \|x_k - \bar{x}\| \leq 2\|x_{k+1} - x_k\|$ und es folgt

$$\|F(x_k) + F'(\bar{x})(x_{k+1} - x_k)\| = o(\|x_k - \bar{x}\|) = o(\|x_{k+1} - x_k\|).$$

ii) $\Longrightarrow$ i): Es gilt wie eben

$$\begin{aligned}\|F(x_k) + F'(\bar{x})(x_{k+1} - x_k)\| &= \|F(x_k) - F(\bar{x}) - F'(\bar{x})(x_k - \bar{x}) + F'(\bar{x})(x_{k+1} - \bar{x})\| \\ &= o(\|x_k - \bar{x}\|) + \|F'(\bar{x})(x_{k+1} - \bar{x})\|\end{aligned}$$

und daher wegen ii)

$$\|F'(\bar{x})(x_{k+1} - \bar{x})\| = o(\|x_k - \bar{x}\|) + o(\|x_{k+1} - x_k\|).$$

Es folgt

$$\|x_{k+1} - \bar{x}\| \leq \|F'(\bar{x})^{-1}\|\|F'(\bar{x})(x_{k+1} - \bar{x})\| = o(\|x_k - \bar{x}\|) + o(\|x_{k+1} - x_k\|).$$

Also gibt es eine Nullfolge (ν_k) mit

$$\|x_{k+1} - \bar{x}\| \leq \nu_k(\|x_k - \bar{x}\| + \|x_{k+1} - x_k\|) \leq \nu_k(2\|x_k - \bar{x}\| + \|x_{k+1} - \bar{x}\|).$$

Somit folgt insbesondere für genügend große k

$$\|x_{k+1} - \bar{x}\| \leq 4\nu_k\|x_k - \bar{x}\| = o(\|x_k - \bar{x}\|).$$

ii) $\Longleftrightarrow$ iii): Die Äquivalenz folgt aus

$$\begin{aligned}&|\|F(x_k) + F'(\bar{x})(x_{k+1} - x_k)\| - \|F(x_k) + F'(x_k)(x_{k+1} - x_k)\|| \\ &\leq \|(F'(\bar{x}) - F'(x_k))(x_{k+1} - x_k)\| = o(\|x_{k+1} - x_k\|).\end{aligned}$$

□

Wenden wir den Satz auf die von Algorithmus 8 erzeugte Folge an, dann erhalten wir die wichtige *Dennis-Moré-Bedingung*:

Korollar 2.8.2 (Dennis-Moré-Bedingung)
$F : \mathbb{R}^n \to \mathbb{R}^n$ sei stetig differenzierbar. Die von Algorithmus 8 erzeugte Folge (x_k) konvergiere gegen einen Punkt $\bar{x}$ mit $F(\bar{x}) = 0$ und $F'(\bar{x})$ invertierbar. Dann sind folgende Aussagen äquivalent:

i) (x_k) konvergiert Q-superlinear gegen $\bar{x}$.

ii) Es gilt $\|(M_k - F'(\bar{x}))s_k\| = o(\|s_k\|) \to 0$ für $k \to \infty$ (Dennis-Moré-Bedingung).

iii) Es gilt $\|(M_k - F'(x_k))s_k\| = o(\|s_k\|) \to 0$ für $k \to \infty$.

Bemerkung: Bedinung iii) kommt ohne $\bar{x}$ aus. Ist s_k der klassische Newtonschritt, dan gilt $M_k = F'(x_k)$ und die linke Seite in iii) verschwindet. Der klassische Newton-Schritt erfüllt also iii) trivialerweise. □

Beweis: Wir haben $x_{k+1} - x_k = s_k$, $M_k s_k = -F(x_k)$ und daher

$$\begin{aligned}\|F(x_k) + F'(\bar{x})(x_{k+1} - x_k)\| &= \|(-M_k + F'(\bar{x}))s_k\| \\ \|F(x_k) + F'(x_k)(x_{k+1} - x_k)\| &= \|(-M_k + F'(x_k))s_k\|.\end{aligned}$$

Daher stimmt ii) und iii) mit ii) und iii) in Satz 2.8.1 überein. □

Die Charakterisierung ii) der Q-superlinearen Konvergenz wurde erstmals von Dennis und Moré in [DM74] gezeigt. Die Bedingung zeigt, dass es ausreicht, wenn $M_k s_k$ auf $o(\|s_k\|)$ mit $F'(\bar{x})s_k$ (oder nach iii) mit $F'(x_k)s_k$) übereinstimmt. Für Vektoren v, die senkrecht auf s_k stehen, können sich aber $M_k v$ und $F'(\bar{x})v$ beliebig stark unterscheiden, ohne die Q-superlineare Konvergenz zu zerstören! Die Dennis-Moré-Bedingung zur Charakterisierung superlinearer Konvergenz wird sich als nützlich bei der lokalen Konvergenzanalyse von *inexakten Newton-Verfahren* und *Quasi-Newton-Verfahren* erweisen, die wir in Kürze behandeln.

Beispiel 2.8.1 Gilt für die von Algorithmus 8 erzeugten Folgen (x_k) und (M_k), dass $x_k \to \bar{x}$ und $M_k \to F'(\bar{x})$, dann gilt die Dennis-Moré-Bedingung, da

$$\begin{aligned}\|(M_k - F'(\bar{x}))(x_{k+1} - x_k)\| &\le \|M_k - F'(\bar{x})\| \|x_{k+1} - x_k\| \\ &= o(\|x_{k+1} - x_k\|) \to 0 \quad \text{für } k \to \infty.\end{aligned}$$

2.8.2 Globalisierung Newton-artiger Verfahren

Die Globalisierung Newton-artiger Verfahren kann genauso erfolgen wie die Globalisierung des Newton-Verfahrens in Algorithmus 7.

Algorithmus 10 Globalisiertes Newton-artiges Verfahren

Wähle Konstanten $\gamma \in (0, 1/2)$, $c_1 \in (0, 1)$, $c_2 > 0$ *und* $p > 0$. *Wähle einen Startpunkt* $x_0 \in \mathbb{R}^n$.

Für $k = 0, 1, \ldots$:

1. *Falls* $\nabla f(x_k) = 0$: *STOP mit Ergebnis* x_k.
2. *Wähle eine symmetrische, invertierbare Matrix* $M_k \in \mathbb{R}^{n,n}$.
3. *Berechne* $s_k^{NA} \in \mathbb{R}^n$ *durch Lösen der* Newton-artigen Gleichung
$$M_k s_k^{NA} = -\nabla f(x_k).$$
Falls
$$-\nabla f(x_k)^T s_k^{NA} \geq \min\{c_1, c_2 \|\nabla f(x_k)\|^p\} \|\nabla f(x_k)\| \|s_k^{NA}\|,$$
dann setze $s_k = s_k^{NA}$.
Sonst wähle eine symmetrische, invertierbare Matrix $B_k \in \mathbb{R}^{n,n}$, *so dass* $\lambda_{min}(B_k) \geq \mu$, $\lambda_{max}(B_k) \leq \eta$ *mit von* k *unabhängigen Konstanten* $0 < \mu < \eta$ *gilt (z.B.* $B_k = I$*), und bestimme* s_k *als Lösung von*
$$B_k s_k = -\nabla f(x_k).$$
4. *Bestimme eine Schrittweite* $\sigma_k > 0$ *nach der Armijo-Regel.*
5. *Setze* $x_{k+1} = x_k + \sigma_k s_k$.

Der Beweis von Satz 2.7.7 kann ohne Änderung auf Algorithmus 10 angewendet werden, wenn die Folge $(\|M_k\|)$ beschränkt ist. Superlineare Konvergenz erhält man, wenn die Dennis-Moré-Bedingung erfüllt ist. Genauer gilt:

Satz 2.8.3 *Es sei* $f : \mathbb{R}^n \to \mathbb{R}$ *zweimal stetig differenzierbar.* $N_f(x_0)$ *sei kompakt und Algorithmus 10 terminiere nicht endlich. Ist die Folge* $(\|M_k\|)$ *beschränkt, dann gilt:*

i) *Die erzeugten Folgen* (s_k) *von Suchrichtungen und* (σ_k) *von Schrittweiten sind zulässig. Insbesondere gelten die Konvergenzaussagen von Satz 2.5.6 und Satz 2.5.7.*

ii) *Hat* (x_k) *einen Häufungspunkt* $\bar{x}$ *mit* $\nabla^2 f(\bar{x})$ *positiv definit und erfüllt* M_k *mit* s_k^{NA} *die Dennis-Moré-Bedingung aus Satz 2.8.1 ii) oder äquivalent iii) für jede gegen* $\bar{x}$ *konvergente Teilfolge* $(x_k)_{k \in K}$, *dann gilt* $\lim_{k \to \infty} x_k = \bar{x}$ *und es gibt* $l \geq 0$ *mit* $s_k = s_k^{NA}$ *und* $\sigma_k = 1$ *für alle* $k \geq l$. *Insbesondere konvergiert* (x_k) *Q-superlinear gegen* $\bar{x}$.

Beweis: zu i): Der Beweis von Satz 2.7.7 ist mit M_k anstelle $\nabla^2 f(x_k)$ anwendbar.

zu ii): Nach i) gilt $\nabla f(\bar{x}) = 0$. Wir zeigen, dass es $\varepsilon > 0$ gibt mit

$$s_k = s_k^{NA} \quad \forall\, k \text{ mit } x_k \in B_\varepsilon(\bar{x}). \tag{2.44}$$

Andernfalls gibt es eine Teilfolge $(x_k)_{k\in K}$ mit $(x_k)_{k\in K} \to \bar{x}$ und $s_k \neq s_k^{NA}$ für alle $k \in K$. Dann gibt es $l' \geq 0$, so dass mit $0 < \mu < \eta$ gilt

$$\lambda_{min}(\nabla^2 f(x_k)) \geq \mu > 0, \quad \lambda_{max}(\nabla^2 f(x_k)) \leq \eta \quad \forall\, k \in K,\ \ k \geq l'$$

Für $k \geq l'$ exitiert also der klassische Newton-Schritt s_k^N und die Dennis-Moré-Bedingung liefert

$$\begin{aligned}
&\|(M_k - \nabla^2 f(x_k))s_k^{NA}\| = \| - \nabla f(x_k) - \nabla^2 f(x_k)s_k^{NA}\| \\
&= \|\nabla^2 f(x_k)(s_k^N - s_k^{NA})\| = o(\|s_k^{NA}\|) \to 0 \quad \text{für } k \in K \to \infty.
\end{aligned}$$

Nach i) gilt $(\nabla f(x_k))_{k\in K} \to 0$. Daher liefert der zweite Term

$$\lim_{k\in K\to\infty} \|s_k^{NA}\| = 0 \tag{2.45}$$

und der dritte $\|s_k^{NA} - s_k^N\| = o(\|s_k^{NA}\|)$, also

$$\|s_k^{NA}\| = \|s_k^N\| + o(\|s_k^{NA}\|) \geq \frac{\|\nabla f(x_k)\|}{\eta} + o(\|s_k^{NA}\|).$$

Zudem zeigt der zweite Term, dass

$$-\nabla f(x_k)^T s_k^{NA} = (s_k^{NA})^T \nabla^2 f(x_k) s_k^{NA} + o(\|s_k^{NA}\|^2) \quad \text{für } k \in K \to \infty. \tag{2.46}$$

Wir finden also $l'' \geq l'$ mit

$$\left.\begin{aligned}
\|s_k^{NA}\| &\geq \frac{\|\nabla f(x_k)\|}{2\eta} \\
-\nabla f(x_k)^T s_k^{NA} &\geq \frac{\mu}{2}\|s_k^{NA}\|^2 \geq \frac{\mu}{4\eta}\|\nabla f(x_k)\|\|s_k^{NA}\|
\end{aligned}\right\} \quad \forall\, k \in K,\ k \geq l''.$$

Genau wie im Beweis von Satz 2.7.7 folgt nun, dass für $l''' \geq l''$ groß genug die Winkelbedingung in Schritt 3 von Algorithmus 10 für alle $k \in K$ mit $k \geq l'''$ erfüllt ist, also $s_k = s_k^{AN}$ für $k \geq l'''$ gilt. Dies ist ein Widerspruch zu $s_k \neq s_k^{NA}$ für alle $k \in K$, es gilt also (2.44) mit geeignetem $\varepsilon > 0$.

Für den Nachweis von $\lim_{k\to\infty} x_k = \bar{x}$ haben wir wie im Beweis von Satz 2.7.7 nur zu zeigen, dass für jede Teilfolge $(x_k)_{k\in K}$ mit $(x_k)_{k\in K} \to \bar{x}$ zudem gilt $(x_{k+1} - x_k)_{k\in K} \to 0$. Sei $(x_k)_{k\in K}$ eine solche Teilfolge. Dann gilt $s_k = s_K^{AN}$ für alle $k \in K$ groß genug und es ergibt sich wie eben (2.45), also

$$\|x_{k+1} - x_k\| \leq \|s_k^{NA}\| \to 0 \quad \text{für } k \in K \to \infty.$$

Dies zeigt $x_k \to \bar{x}$.

Wegen (2.44) gilt $s_k = s_k^{NA}$ für $k \geq l$, l groß genug. Weiter gilt (2.46) für die Gesamtfolge und beim Nachweis, dass $\sigma_k = 1$ für $k \geq l$, l groß genug die Armijo-Bedingung erfüllt, erhält man nun analog zum Beweis von Satz 2.7.7 mit $s_k = s_k^{NA}$

$$\begin{aligned}
&f(x_k) - f(x_k + s_k) + \gamma \nabla f(x_k)^T s_k \\
&\geq \left(1 - \gamma - \frac{1}{2}\right) \mu \|s_k\|^2 - \frac{1}{2}\|\nabla^2 f(x_k + \tau_k s_k) - \nabla^2 f(x_k)\| \|s_k\|^2 + o(\|s_k\|^2).
\end{aligned}$$

Die rechte Seite ist wegen $s_k \to 0$ positiv für alle $k \geq l$, l groß genug. □

2.8.3 Dennis-Moré-artige Bedingung für quadratische Konvergenz

Ist F' Lipschitz-stetig in einer Umgebung von $\bar{x}$, dann liefert eine natürliche Verschärfung der Dennis-Moré-Bedingung quadratische Konvergenz.

Satz 2.8.4 *Es sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar. Weiter sei $\bar{x} \in \mathbb{R}^n$ ein Punkt mit $F(\bar{x}) = 0$ und $F'(\bar{x})$ invertierbar. Ist F' Lipschitz-stetig in einer Umgebung von $\bar{x}$ und ist (x_k) eine Folge, die gegen $\bar{x}$ konvergiert, dann sind folgende Aussagen äquivalent:*

i) (x_k) *konvergiert Q-quadratisch gegen* $\bar{x}$.

ii) Es gilt $\|F(x_k) + F'(\bar{x})(x_{k+1} - x_k)\| = O(\|x_{k+1} - x_k\|^2) \to 0$ *für* $k \to \infty$.

iii) Es gilt $\|F(x_k) + F'(x_k)(x_{k+1} - x_k)\| = O(\|x_{k+1} - x_k\|^2) \to 0$ *für* $k \to \infty$.

Beweis: i) $\Longleftrightarrow$ ii): Da F' lokal Lipschitz-stetig ist, gilt

$$\begin{aligned}
\|F(x_k) + F'(\bar{x})(x_{k+1} - x_k)\| &= \|F(x_k) - F(\bar{x}) - F'(\bar{x})(x_k - \bar{x}) + F'(\bar{x})(x_{k+1} - \bar{x})\| \\
&= O(\|x_k - \bar{x}\|^2) + \|F'(\bar{x})(x_{k+1} - \bar{x})\|.
\end{aligned}$$

Der Beweis kann nun genau wie bei Satz 2.8.1 geführt werden mit $O(\|x_k - \bar{x}\|^2)$ und $O(\|x_{k+1} - x_k\|^2)$ anstelle $o(\|x_k - \bar{x}\|)$ und $o(\|x_{k+1} - x_k\|)$.

ii) $\Longleftrightarrow$ iii): Die Lipschitz-Stetigkeit von F' liefert

$$\begin{aligned}
&|\|F(x_k) + F'(\bar{x})(x_{k+1} - x_k)\| - \|F(x_k) + F'(x_k)(x_{k+1} - x_k)\|| \\
&\leq \|(F'(\bar{x}) - F'(x_k))(x_{k+1} - x_k)\| = O(\|x_k - \bar{x}\| \|x_{k+1} - x_k\|).
\end{aligned}$$

Nach Satz 2.8.1 impliziert ii) Q-superlineare Konvergenz und ebenso iii). Also ist für große k $1/2\|x_{k+1} - x_k\| \leq \|x_k - \bar{x}\| \leq 2\|x_{k+1} - x_k\|$ und somit gilt

$$O(\|x_k - \bar{x}\| \|x_{k+1} - x_k\|) = O(\|x_{k+1} - x_k\|^2).$$

□

Wieder erhalten wir als unmittelbare Folge

Korollar 2.8.5 *$F : \mathbb{R}^n \to \mathbb{R}^n$ sei stetig differenzierbar. Die von Algorithmus 8 erzeugte Folge (x_k) konvergiere gegen einen Punkt $\bar{x}$ mit $F(\bar{x}) = 0$ und $F'(\bar{x})$ invertierbar. Ist F' Lipschitz-stetig in einer Umgebung von $\bar{x}$, dann sind folgende Aussagen äquivalent:*

i) *(x_k) konvergiert Q-quadratisch gegen $\bar{x}$.*

ii) *Es gilt $\|(M_k - F'(\bar{x}))s_k\| = O(\|s_k\|^2) \to 0$ für $k \to \infty$ (Dennis-Moré-Bedingung).*

iii) *Es gilt $\|(M_k - F'(x_k))s_k\| = O(\|s_k\|^2) \to 0$ für $k \to \infty$.*

2.9 Inexakte Newton-Verfahren

Bei großen Problemen (Richtwert $n \geq 10000$) ist die exakte Lösung der Newton-Gleichung

$$F'(x_k)s_k = -F(x_k)$$

oft zu aufwendig. Selbst wenn die Jacobi-Matrix $F'(x_k)$ dünn besetzt ist, ist eine direkte Lösung der Newton-Gleichung oft nicht praktikabel, da die LR-Faktorisierung auf viel dichter besetzte Matrizen führt. Als Ausweg verwendet man in der Praxis iterative Löser, um die Newton-Gleichung näherungsweise zu lösen. Dies führt auf *inexakte Newton-Verfahren*, bei denen man lediglich eine Näherungslösung s_k mit

$$F'(x_k)s_k = -F(x_k) + r_k$$

bestimmt, wobei $r_k \in \mathbb{R}^n$ das verbleibende Residuum ist. Es stellt sich die Frage, wie klein das Residuum r_k sein muss, damit lineare oder superlineare Konvergenz sichergestellt ist. Es zeigt sich, dass lokale Konvergenz bereits garantiert ist, wenn gilt

$$\|F(x_k) + F'(x_k)s_k\| \leq \nu_k \|F(x_k)\|, \tag{2.47}$$

wobei $\nu_k \in (0, \nu]$ mit einer Konstante $\nu > 0$.

Algorithmus 11 Lokales inexaktes Newton-Verfahren

Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$ und $\nu \in (0, 1)$.

Für $k = 0, 1, \dots$:

1. *Falls $F(x_k) = 0$: STOP mit Ergebnis x_k.*

2. *Wähle eine Toleranz $\nu_k \in (0, \nu]$ und berechne eine inexakte Lösung s_k der Newton-Gleichung, so dass (2.47) gilt.*

3. *Setze $x_{k+1} = x_k + s_k$.*

Beim inexakten Newton-Verfahren für das Minimierungsproblem (P) wendet manAlgorithmus 11 mit $F = \nabla f$ an.

Beispiel 2.9.1 Betrachte das Newton-Verfahren für (P) nahe eines Punktes, der die hinreichende Bedingung zweiter Ordnung erfüllt. Dann ist $F'(x_k) = \nabla^2 f(x_k)$ positiv definit für x_k nahe genug bei $\bar{x}$. Wir betrachten ein inexaktes Newton-Verfahren, das zur inexakten Lösung der Newton-Gleichung das CG-Verfahren verwendet. Bekanntlich liefert das CG-Verfahren bei exakter Rechnung nach maximal n Schritten die exakte Lösung s_k^N. Bei großen Problemen ist jedoch die Konvergenzrate viel interessanter. Man kann zeigen, siehe z.B. [Da67], [Br92], dass für die vom CG-Verfahren erzeugten Iterierten $s_k^{(j)}$ mit Startpunkt $s_k^{(0)}$ gilt

$$\|s_k^{(j)} - s_k^N\|_{\nabla^2 f(x_k)} \le 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^j \|s_k^{(0)} - s_k^N\|_{\nabla^2 f(x_k)}$$

mit der Konditionszahl $\kappa = \frac{\lambda_{max}(\nabla^2 f(x_k))}{\lambda_{min}(\nabla^2 f(x_k))}$. Hieraus folgt leicht

$$\|s_k^{(j)} - s_k^N\| \le 2\sqrt{\kappa} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^j \|s_k^{(0)} - s_k^N\|.$$

Wegen

$$\|\nabla f(x_k) + \nabla^2 f(x_k) s_k^{(j)}\| = \|\nabla^2 f(x_k)(s_k^{(j)} - s_k^N)\| = \|\nabla^2 f(x_k)^{1/2}(s_k^{(j)} - s_k^N)\|_{\nabla^2 f(x_k)}$$

und $\|s_k^{(j)} - s_k^N\|_{\nabla^2 f(x_k)} = \|\nabla^2 f(x_k)^{1/2}(s_k^{(j)} - s_k^N)\|$ erhalten wir zudem

$$\|\nabla f(x_k) + \nabla^2 f(x_k) s_k^{(j)}\| \le 2\sqrt{\kappa} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^j \|\nabla f(x_k) + \nabla^2 f(x_k) s_k^{(0)}\|.$$

Das CG-Verfahren konvergiert also schnell gegen den Newton-Schritt, wenn $\kappa \approx 1$. Aber auch bei moderaten Konditionszahlen ist die maximale Zahl der benötigten Iterationen viel kleiner als n:

κ	relativer Fehler ν		
	10^{-1}	10^{-2}	10^{-3}
10	7	10	14
100	27	38	50
1000	102	139	175

Maximale Zahl von CG-Iterationen, um bei Konditionszahl κ mit Startpunkt $s_k^{(0)} = 0$ einen vorgegebenen relativen Fehler einzuhalten.

Ist die Kondition von $\nabla^2 f(x_k)$ groß, kann man vorkonditionierte CG-Verfahren anwenden. Hier verwendet man einen *Vorkonditionierer* $v \mapsto Pv$, mit P symmetrisch positiv definit, so dass die Konditionszahl $\kappa(P\nabla^2 f(x_k))$ moderat ist. Sei $P^{1/2}$ die symmetrische positiv definite Wurzel von P. Dann gilt $\kappa(P\nabla^2 f(x_k)) = \kappa(P^{1/2}\nabla^2 f(x_k)P^{1/2})$, da die beiden Matrizen ähnlich sind, Anwendung des CG-Verfahrens auf

$$P^{1/2}\nabla^2 f(x_k)P^{1/2}\tilde{s}_k = -P^{1/2}\nabla f(x_k)$$

konvergiert also recht schnell. Schreibt man das resultierende CG-Verfahren wieder in den Originalvariablen $s_k = P^{1/2}\tilde{s}_k$, dann erhält man das vorkonditionierte CG-Verfahren. Man stellt fest, dass man zu dessen Implementierung nur den Vorkonditionierer $v \mapsto Pv$ benötigt. Vorkonditionierer $w = Pv$ lassen sich durch näherungsweise Löser des Systems $\nabla^2 f(x_k)v = w$ gewinnen, z.B. basierend auf einer unvollstandigen Cholesky-Zerlegung, einigen Iterationen des Jacobi- oder symmetrischen Gauß-Seidel-Verfahrens, Mehrgitterverfahren, etc.. Die Entwicklung von guten Vorkonditionierern für verschiedene Anwendungsfelder ist ein aktives Forschungsfeld.

Die Globalisierung inexakter Newton-Verfahren kann wie bei Newton-artigen Verfahren erfolgen. Eine elegante Variante eines inexakten Verfahrens für (P) ist das CG-Newton-Verfahren, das in der Übung behandelt wird.

2.9.1 Ein lokaler Konvergenzsatz

Es gilt der folgende Satz.

Satz 2.9.1 *Es sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\bar{x}$ sei ein Punkt mit $F(\bar{x}) = 0$, $F'(\bar{x})$ invertierbar. Dann gibt es $\delta > 0$, so dass gilt:*

i) Für jeden Startpunkt $x_0 \in B_\delta(\bar{x})$ terminiert Algorithmus 11 entweder endlich, oder er erzeugt eine Folge (x_k) mit $F(x_k) \to 0$ Q-linear und $x_k \to \bar{x}$ mindestens R-linear. Genauer gibt es zu jedem $\alpha \in (\nu, 1)$ ein $\delta > 0$, so dass gilt

$$\|F(x_{k+1})\| \leq \alpha \|F(x_k)\| \quad \textit{für alle } k$$

und

$$\|x_k - \bar{x}\| \leq 2\|F'(\bar{x})^{-1}\| \, \|F(x_k)\| \quad \textit{für alle } k.$$

ii) Gilt zudem $\nu_k \to 0$, so konvergieren $(F(x_k))$ und (x_k) Q-superlinear.

c) Gilt sogar $\nu_k = O(\|F(x_k)\|)$, dann konvergieren $(F(x_k))$ und (x_k) Q-quadratisch, falls F' Lipschitz-stetig in einer Umgebung von $\bar{x}$ ist.

Beweis: zu ii): Nach i) gilt $x_k \to \bar{x}$. Wir erhalten ii) bequem durch Prüfen der Dennis-Moré-Bedingung in der Version aus Satz 2.8.1, iii): Nach i) gilt $x_k \to \bar{x}$. (2.47) und $\nu_k \to 0$ liefern

$$\|F(x_k) + F'(x_k)s_k\| \leq \nu_k \|F(x_k)\| = o(\|F(x_k)\|) \to 0. \tag{2.48}$$

Insbesondere erhalten wir

$$\|F(x_k)\| = \|F'(x_k)s_k\| + o(\|F(x_k)\|)$$

und folglich $o(\|F(x_k)\|) = o(\|s_k\|)$. Einsetzen in (2.48) ergibt

$$\|F(x_k) + F'(x_k)s_k\| = o(\|s_k\|) \to 0.$$

Dies ist genau Satz 2.8.1, iii) und impliziert Q-superlineare Konvergenz.

zu iii): Nach i) gilt $x_k \to \bar{x}$. Wir prüfen die Dennis-Moré-artige Bedingung aus Satz 2.8.4, iii): (2.47) und $\nu_k = O(\|F(x_k)\|)$ liefern

$$\|F(x_k) + F'(x_k)s_k\| \le \nu_k \|F(x_k)\| = O(\|F(x_k)\|^2) \to 0.$$

Wie beim Beweis von ii) erhalten wir $O(\|F(x_k)\|^2) = O(\|s_k\|^2)$, also

$$\|F(x_k) + F'(x_k)s_k\| \le \nu_k \|F(x_k)\| = O(\|F(x_k)\|^2) = O(\|s_k\|^2) \to 0$$

und Satz 2.8.4, iii) liefert Q-quadratische Konvergenz.

zu i) (für Interessierte): Wir setzen

$$\mu = \frac{1}{\|F'(\bar{x})^{-1}\|}, \quad \eta = \|F'(\bar{x})\|.$$

Nach Satz 2.7.3 finden wir dann $\delta > 0$ mit

$$\|F'(x)\| \le 2\eta, \quad \|F'(x)^{-1}\| \le \frac{2}{\mu} \quad \forall\, x \in B_{2\delta}(\bar{x})$$

und

$$\frac{\mu}{2}\|x - \bar{x}\| \le \|F(x)\| \le 2\eta\|x - \bar{x}\| \quad \forall\, x \in B_{2\delta}(\bar{x}). \tag{2.49}$$

Betrachte ein beliebiges $x_k \in B_\delta(\bar{x})$. Wir haben ähnlich wie bei der Konvergenzanalyse des exakten Newton-Verfahrens

$$\begin{aligned} F(x_{k+1}) &= F(x_k) + \int_0^1 F'(x_k + \tau s_k)\, d\tau s_k = -F'(x_k)s_k + r_k + \int_0^1 F'(x_k + \tau s_k)\, d\tau s_k \\ &= r_k + \int_0^1 \left(F'(x_k + \tau s_k) - F'(x_k)\right)\, d\tau s_k. \end{aligned}$$

Dies liefert

$$\|F(x_{k+1})\| \le \nu_k \|F(x_k)\| + \int_0^1 \|F'(x_k + \tau s_k) - F'(x_k)\|\, d\tau \|s_k\|. \tag{2.50}$$

Weiter gilt
(2.51)

$$\|s_k\| \le \|F'(x_k)^{-1}\| \| - F(x_k) + r_k\|_2 \le \|F'(x_k)^{-1}\|(1+\nu)\|F(x_k)\| \le \frac{2(1+\nu)}{\mu}\|F(x_k)\|.$$

Sei $\alpha \in]\nu, 1[$ beliebig. Wegen der gleichmäßigen Stetigkeit von $F'(x)$ auf Kompakta können wir $\delta > 0$ so verkleinern, dass gilt

$$\|F'(x) - F'(y)\| \le (\alpha - \nu)\frac{\mu}{2(1+\nu)} > 0 \quad \forall\, x, y \in B_{2\delta}(\bar{x}). \tag{2.52}$$

Nun sei $x_k \in U_\delta(\bar{x})$ beliebig mit der offenen Umgebung

$$U_\delta(\bar{x}) := \left\{ x \in B_\delta(\bar{x}) \ : \ \|F(x)\| < \delta\frac{\mu}{2(1+\nu)} \right\}$$

von $\bar{x}$. Dann gilt wegen (2.51) und der Definition von $U_\delta(\bar{x})$

$$\|s_k\|_2 \le \delta,$$

also $x_k, x_k + \tau s_k \in B_{2\delta}(\bar{x})$ für $\tau \in [0, 1]$. Dies liefert mit (2.50), (2.51), (2.52)

$$\begin{aligned} \|F(x_{k+1})\| &\le \nu_k\|F(x_k)\| + (\alpha - \nu)\frac{\mu}{2(1+\nu)}\frac{2(1+\nu)}{\mu}\|F(x_k)\| \\ &= (\alpha - \nu + \nu_k)\|F(x_k)\| \le \alpha\|F(x_k)\|. \end{aligned}$$

Insbesondere haben wir wegen $x_{k+1} \in B_{2\delta}(\bar{x})$, (2.49) und der Definition von $U_\delta(\bar{x})$

$$\|x_{k+1} - \bar{x}\| \le \frac{2}{\mu}\|F(x_{k+1})\| < \frac{2}{\mu}\|F(x_k)\|_2 \le \frac{2}{\mu}\delta\frac{\mu}{2(1+\nu)} \le \delta.$$

Also gilt wieder $x_{k+1} \in U_\delta(\bar{x})$. Mit Induktion ergibt sich also die Behauptung für jeden Startpunkt $x_0 \in U_\delta(\bar{x})$. $\Box$

2.9.2 Zusammenhang von Newton-artigen Verfahren und inexakten Newton-Verfahren

Jedes Newton-artige Verfahren kann als inexaktes Newton-Verfahren interpretiert werden und umgekehrt.

Sei ein Newton-artigen Verfahren mit

$$M_k s_k = -F(x_k)$$

gegeben. Dann gilt

$$F'(x_k)s_k = -F(x_k) + r_k \ \text{ mit } \ r_k = (F'(x_k) - M_k)s_k.$$

Also ist s_k inexakte Lösung der Newton-Gleichung mit Residuum $r_k = (F'(x_k) - M_k)s_k$. Für ein konkretes Verfahren kann man häufig das Residuum

$$\|r_k\| = \|(F'(x_k) - M_k)s_k\|$$

abschätzen. Die Dennis-Moré-Bedingung ist erfüllt, wenn gilt $\|r_k\| = o(\|s_k\|)$.

Betrachte umgekehrt ein inexkates Newton-Verfahren. Dann gilt

$$F'(x_k)s_k = -F(x_k) + r_k$$

mit einem Residuenvektor r_k. Wählen wir M_k so, dass gilt

$$M_k s_k = -F(x_k),$$

was zum Beispiel mit der Wahl

$$M_k = F'(x_k) - \frac{r_k s_k^T}{\|s_k\|^2}$$

gilt, dann können wir das Verfahren als Newton-artiges Verfahren interpretieren. Da M_k von s_k und r_k abhängt, ist dieser Zusammenhang nur von theoretischem Interesse.

Wichtig ist die Beobachtung, dass die inexakte Lösung der Newton-Gleichung immer als exakte Lösung einer Newton-artigen Gleichung interpretiert werden kann und umgekehrt.

2.10 Quasi-Newton-Verfahren

Wir betrachten weiterhin das unrestringierte Minimierungsproblem

$$\text{(P)} \qquad \min_{x \in \mathbb{R}^n} \; f(x).$$

mit $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Liegt die Hessematrix nicht vor oder ist sie aufwendig zu berechnen, dann legen die allgemeinen Resultate über Newton-artige Verfahren nahe, einfach zu berechnende Approximationen $H_k = M_k$ für die Hessematrix zu verwenden (wir ziehen hier die Bezeichnung H_k statt M_k vor, da wir uns nun ausschließlich mit dem Minimierungsproblem (P) befassen) und die Suchrichtung als Lösung der Newton-artigen Gleichung

$$H_k s_k = -\nabla f(x_k) \tag{2.53}$$

zu berechnen.

Bei *Quasi-Newton-Verfahren* erzeugt man ausgehend von einer symmetrischen, invertierbaren Matrix $H_0 \in \mathbb{R}^{n,n}$ durch Updates Approximationen H_k von $\nabla^2 f(x_k)$, so dass die folgende Quasi-Newton-Gleichung gilt:

Quasi-Newton-Gleichung:

$$H_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k) \tag{2.54}$$

oder kurz

$$H_{k+1} d_k = y_k \quad \text{mit } d_k = x_{k+1} - x_k, \;\; y_k = \nabla f(x_{k+1}) - \nabla f(x_k). \tag{2.54}$$

Hierbei beschränkt man sich bei Quasi-Newton-Verfahren auf *Aufdatierungsformeln* der Form

$$H_{k+1} = \Phi(H_k, d_k, y_k).$$

Motivation der Quasi-Newton-Gleichung: Die Quasi-Newton-Gleichung (2.54) kann wie folgt motiviert werden: Die rechte Seite kann geschrieben werden als (siehe Satz 2.1.1)

$$\begin{aligned}\nabla f(x_{k+1}) - \nabla f(x_k) &= \int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k))\, dt (x_{k+1} - x_k) \\ &= \nabla^2 f(x_k)(x_{k+1} - x_k) + o(\|x_{k+1} - x_k\|).\end{aligned}$$

Insbesondere erfüllt also die gemittelte Hessematrix

$$\bar{H}_k := \int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k))\, dt$$

die Quasi-Newton-Gleichung (2.54). Würde man nun (2.54) für H_k anstelle H_{k+1} fordern, also

$$H_k(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k), \tag{2.55}$$

dann wäre für $x_k \to \bar{x}$ die Dennis-Moré-Bedingung erfüllt, da mit (2.55)

$$\begin{aligned}\|(H_k - \nabla^2 f(x_k))(x_{k+1} - x_k)\| &= \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k)\| \\ &= o(\|x_{k+1} - x_k\|).\end{aligned}$$

Man würde sogar Konvergenz in einem Schritt erhalten, denn wegen $x_{k+1} = x_k + s_k$ ergeben (2.53) und (2.55)

$$-\nabla f(x_k) = H_k s_k = H_k(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k), \quad \text{also } \nabla f(x_{k+1}) = 0.$$

Da $x_{k+1} = x_k + s_k$ jedoch über (2.53) von H_k abhängt, ist die Forderung (2.55) an H_k nicht praktikabel. Es liegt daher nahe, die Anforderung (2.55) an H_{k+1} anstelle H_k zu stellen, was auf die Quasi-Newton-Gleichung (2.54) führt.

Wir betrachten allgemein folgende Klasse von Quasi-Newton-Verfahren:

Algorithmus 12 Lokales Quasi-Newton-Verfahren

Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$ und eine eine symmetrische, nichtsinguläre Matrix $H_0 \in \mathbb{R}^{n,n}$

Für $k = 0, 1, \dots :$

1. *Falls $\nabla f(x_k) = 0$: STOP mit Ergebnis x_k.*
2. *Berechne $s_k \in \mathbb{R}^n$ durch Lösen der Newton-artigen Gleichung*
$$H_k s_k = -\nabla f(x_k).$$
3. *Setze $x_{k+1} = x_k + s_k$.*
4. *Berechne mit einer Aufdatierungsformel eine symmetrische, nichtsinguläre Matrix $H_{k+1} = \Phi(H_k, d_k, y_k)$, welche die Quasi-Newton-Gleichung* (2.54) *erfüllt.*

Tatsächlich gilt dann unter gewissen Voraussetzungen die Dennis-Moré-Bedingung:

Lemma 2.10.1 *$\bar{x}$ erfülle die hinreichende Bedingung zweiter Ordnung. Erzeugt Algorithmus 12 eine gegen $\bar{x}$ konvergente Folge (x_k) und gilt zudem*

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

dann erfüllen H_k die Dennis-Moré-Bedinung und (x_k) konvergiert Q-superlinear gegen $\bar{x}$.

Beweis: Taylorentwicklung und (2.54) ergeben

$$\begin{aligned}\|(H_k - \nabla^2 f(x_k))s_k\| &\le \|(H_k - H_{k+1})s_k\| + \|(H_{k+1} - \nabla^2 f(x_k))s_k\| \\ &= o(\|s_k\|) + \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)s_k\| = o(\|s_k\|).\end{aligned}$$

□

Zusatz: Anstelle von $\lim_{k\to\infty} \|H_{k+1} - H_k\| = 0$ kann man auch fordern, dass $(\|H_k\|)$ eine beschränkte Folge invertierbarer Matrizen ist mit

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

Denn dann gilt

$$\|H_{k+1} - H_k\| = \|H_{k+1}(H_k^{-1} - H_{k+1}^{-1})H_k\| \le \|H_k\|\|H_{k+1}\|\|H_k^{-1} - H_{k+1}^{-1}\| \to 0.$$

□

2.10.1 Updates minimaler Änderung als Grundprinzip bei der Konstruktion von Quasi-Newton-Updates

Das letzte Lemma motiviert, unter allen Matrizen H_{k+1}, die (2.54) erfüllen, diejenige zu wählen, die bezüglich einer gegebenen Norm $\|\cdot\|_*$ minimalen Abstand von H_k besitzt, also

$$(2.56)\qquad H_{k+1} = \operatorname*{argmin}_{H \in \mathbb{R}^{n,n}} \|H - H_k\|_* \quad \text{u. d. Nebenbedingung } H = H^T,\ Hd_k = y_k.$$

Ist $\|\cdot\|_*$ eine Hilbertraum-Norm auf $\mathbb{R}^{n,n}$ (also induziert durch ein Skalarprodukt), dann ist (2.56) ein konvexes Optimierungsproblem und besisitzt eine eindeutige Lösung (einfache Übung).

Updates minimaler Änderung gemäß (2.56) sind im allgemeinen nicht invariant unter linearen Variablentransformationen. Diese wünschenswerte – vom klassischen Newton-Verfahren erfüllte – Eigenschaft kann man sicherstellen, wenn man in (2.56) variable gewichtete (Hilbertraum-) Normen der Form

$$(2.57)\qquad \|H - H_k\|_* = \|W_k^{\frac{1}{2}}(H - H_k)W_k^{\frac{1}{2}}\|_F$$

wählt mit einer beliebigen symmetrisch positiv definiten Matrix W_k, so dass gilt $W_k y_k = d_k$, und der Frobenius-Norm

$$\|M\|_F = \left(\sum_{i,j=1}^{n} |m_{ij}|^2\right)^{1/2} = \sqrt{\text{Spur}(MM^T)}.$$

Die Existenz von W_k ist gesichert, falls gilt

$$(2.58)\qquad y_k^T d_k > 0.$$

Das resultierende Problem ist

$$(2.59)\qquad H_{k+1} = \operatorname*{argmin}_{H \in \mathbb{R}^{n,n}} \|W_k^{\frac{1}{2}}(H - H_k)W_k^{\frac{1}{2}}\|_F \quad \text{u. d. Nebenbed. } H = H^T,\ Hd_k = y_k,$$

wobei $W_k y_k = d_k$. Man kann zeigen, dass (2.59) im Fall (2.58) die eindeutige Lösung

$$(2.60)\qquad H_{k+1}^{DFP} = \left(I - \frac{d_k y_k^T}{y_k^T d_k}\right)^T H_k \left(I - \frac{d_k y_k^T}{y_k^T d_k}\right) + \frac{y_k y_k^T}{y_k^T d_k}$$

besitzt. Dies ist der bekannte DFP-Update von Davidon, Fletcher und Powell, den wir in Kürze genauer kennenlernen werden.

Zum Nachweis der Skalierungsinvarianz betrachte die lineare Variablentransformation $x = A\tilde{x} + c$. Wir zeigen hierzu: liefert Algorithmus 12 angewendet auf (P) die Folge (x_k) und angewendet auf

$$(\tilde{P})\qquad \min_{\tilde{x} \in \mathbb{R}^n} \tilde{f}(\tilde{x}) := f(A\tilde{x} + c)$$

mit $x_0 = A\tilde{x}_0 + c$, $\tilde{H}_0 = A^T H_0 A$ die Folge $(\tilde{x}_k)$, dann gilt $x_k = A\tilde{x}_k + c$.

Schreiben wir Algorithmus 12 für (P) in den $\tilde{x}$-Variablen, dann erhalten wir

$$H_k s_k = H_k(x_{k+1} - x_k) = H_k A(\tilde{x}_{k+1} - \tilde{x}_k) = H_k A\tilde{s}_k = -\nabla f(x_k),$$

also nach Multiplikation mit A^T

$$A^T H_k A\tilde{s}_k = -A^T \nabla f(x_k) = -\nabla \tilde{f}(\tilde{x}_k), \quad \tilde{x}_{k+1} = \tilde{x}_k + \tilde{s}_k.$$

Betrachte nun Algorithmus 12 für $(\tilde{P})$ mit $x_0 = A\tilde{x}_0 + c$, $\tilde{H}_0 = A^T H_0 A$. Dann gilt

$$\tilde{H}_k \tilde{s}_k = -\nabla \tilde{f}(\tilde{x}_k), \quad \tilde{x}_{k+1} = \tilde{x}_k + \tilde{s}_k.$$

Wir müssen also nur zeigen, dass immer $\tilde{H}_k = A^T H_k A$ gilt. Für $k = 0$ haben wir $\tilde{H}_0$ so gewählt. Für den Induktionsschritt müssen wir nun zeigen, dass

$$(*) \qquad \tilde{H}_{k+1} = A^T H_{k+1} A, \quad \text{also} \quad \Phi^{DFP}(A^T H_k A, \tilde{d}_k, \tilde{y}_k) = A^T \Phi^{DFP}(H_k, d_k, y_k) A.$$

Aber nach Induktionsvoraussetzung gilt

$$d_k = x_{k+1} - x_k = A(\tilde{x}_{k+1} - \tilde{x}_k) = A\tilde{d}_k, \quad \text{also} \ \tilde{d}_k = A^{-1} d_k$$

und

$$\tilde{y}_k = \nabla \tilde{f}(\tilde{x}_{k+1}) - \nabla \tilde{f}(\tilde{x}_k) = A^T(\nabla f(x_{k+1}) - \nabla f(x_k)) = A^T y_k.$$

Einsetzen in (2.60) zeigt, dass $(*)$ tatsächlich erfüllt ist.

2.10.2 Wichtige Quasi-Newton-Updates

Wir geben nun wichtige Quasi-Newton-Aufdatierungsformeln

$$H_{k+1} = \Phi(H_k, d_k, y_k), \quad d_k = x_{k+1} - x_k, \ \ y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$$

an. Ein ”guter” Update sollte die folgenden Eigenschaften haben:

- H_{k+1} ist wieder symmetrisch,
- H_{k+1} ist nach Möglichkeit positiv definit, falls H_k positiv definit ist,
- die Quasi-Newton-Gleichung (2.54) ist erfüllt,
- die Aufdatierung erfordert geringen Rechenaufwand, und ist nach Möglichkeit invariant unter linearen Variablentransformationen,
- die lokalen Konvergenzeigenschaften sind gut.

Der SR1-Update

Der einfachste denkbare Update besteht in einem Rang-1-Update der Form

$$H_{k+1} = H_k + \alpha_k v_k v_k^T, \qquad \alpha_k \in \{\pm 1\}.$$

Wir erhalten für $(y_k - H_k d_k)^T d_k \neq 0$ die eindeutige Lösung (siehe Übung)

SR1-Update (Symmetrischer Rang-1-Update):

(2.61)
$$H_{k+1} = H_k + \frac{(y_k - H_k d_k)(y_k - H_k d_k)^T}{(y_k - H_k d_k)^T d_k}.$$

Der SR1-Update hat einige offensichtliche Nachteile: Im Falle $(y_k - H_k d_k)^T d_k = 0$ ist der Update nicht definiert und im Fall $(y_k - H_k d_k)^T d_k < 0$ ist unter Umständen H_{k+1} nicht positiv definit, auch wenn H_k positiv definit ist. In diesem Fall ist $s_{k+1} = -H_{k+1}^{-1} \nabla f(x_{k+1})$ nicht notwendigerweise eine Abstiegsrichtung.

Trotz dieser Nachteile hat der SR1-Update in letzter Zeit in Verbindung mit Trust-Region-Verfahren Bedeutung erlangt. Um Schwierigkeiten im Fall $(y_k - H_k d_k)^T d_k \approx 0$ zu vermeiden, hat sich die einfache Strategie bewährt, den Update nur durchzuführen, falls

$$|(y_k - H_k d_k)^T d_k| \geq \varepsilon \|y_k - H_k d_k\| \|d_k\|$$

gilt mit einer kleinen Konstante $\varepsilon > 0$, z.B. $\varepsilon = 10^{-8}$.

DFP-, BFGS- und PSB-Update

Da der SR1-Update der einzige symmetrische Rang-1-Update ist, der (2.53) erfüllt, liegt es nahe, nach Rang-2-Updates der Form

$$H_{k+1} = H_k + \alpha_k v_k v_k^T + \beta_k w_k w_k^T + \gamma_k (v_k w_k^T + w_k v_k^T), \quad \alpha_k, \beta_k, \gamma_k \in \mathbb{R}$$

zu suchen. Unter den vielen möglichen Rang-2-Updates haben sich die folgenden als besonders geeignet erwiesen:

DFP-Update-Formel (Davidon-Fletcher-Powell-Update):

(2.62)
$$\begin{aligned}
H_{k+1} = H_{k+1}^{DFP} &:= H_k + \frac{(y_k - H_k d_k) y_k^T + y_k (y_k - H_k d_k)^T}{y_k^T d_k} - \frac{(y_k - H_k d_k)^T d_k}{(y_k^T d_k)^2} y_k y_k^T \\
&= \left(I - \frac{d_k y_k^T}{y_k^T d_k} \right)^T H_k \left(I - \frac{d_k y_k^T}{y_k^T d_k} \right) + \frac{y_k y_k^T}{y_k^T d_k} \\
&=: \Phi^{DFP}(H_k, d_k, y_k).
\end{aligned}$$

Diese Aufdatierungsformel wurde 1959 von Davidon vorgeschlagen (erst veröffentlicht in [Dav91]) und anschließend von Fletcher und Powell untersucht und populär gemacht. Offensichtlich ist H_{k+1}^{DFP} wohldefiniert für $y_k^T d_k \neq 0$, symmetrisch und man prüft leicht, dass die Quasi-Newton-Gleichung (2.54) erfüllt ist. Wir haben bereits festgestellt, dass (2.62) invariant bezüglich linearen Variablentransformationen ist.

BFGS-Update-Formel (Broyden-Fletcher-Goldfarb-Shanno-Update):

$$H_{k+1} = H_{k+1}^{BFGS} := H_k + \frac{y_k y_k^T}{y_k^T d_k} - \frac{H_k d_k (H_k d_k)^T}{d_k^T H_k d_k} =: \Phi^{BFGS}(H_k, d_k, y_k). \tag{2.63}$$

Dieser im Jahre 1970 von Broyden, Fletcher, Goldfarb und Shanno ([Bro70], [Fl70], [Go70], [Sh70]) vorgeschlagene Update ist der populärste und numerisch effizienteste Quasi-Newton-Update. Wir werden ihn noch genauer behandeln. H_{k+1}^{BFGS} ist wohldefiniert, falls $y_k^T d_k \neq 0$ und $d_k^T H_k d_k \neq 0$, ist symmetrisch und erfüllt die Quasi-Newton-Gleichung (2.54). Man prüft leicht, dass (2.62) invariant bezüglich linearen Variablentransformationen ist.

Die Broyden-Klasse:

$$H_{k+1} = H_{k+1}^{B,\lambda} = (1-\lambda) H_{k+1}^{BFGS} + \lambda H_{k+1}^{DFP} = H_{k+1}^{BFGS} + \lambda (d_k^T H_k d_k) v_k v_k^T, \quad \lambda \in \mathbb{R},$$

mit $v_k = \frac{y_k}{y_k^T d_k} - \frac{H_k d_k}{d_k^T H_k d_k}$.

$H_{k+1}^{B,\lambda}$ ist wohldefiniert, falls $y_k^T d_k \neq 0$ und $d_k^T H_k d_k \neq 0$, ist symmetrisch und erfüllt mit H_{k+1}^{BFGS} und H_{k+1}^{DFP} die Quasi-Newton-Gleichung (2.54).

Die konvexe Broyden-Klasse:

$$H_{k+1} = H_{k+1}^{B,\lambda}, \quad \lambda \in [0,1].$$

PSB-Formel (Powell's symmetrischer Broyden-Update):

(2.64)

$$H_{k+1} = H_{k+1}^{PSB} = H_k + \frac{(y_k - H_k d_k) d_k^T + d_k (y_k - H_k d_k)^T}{d_k^T d_k} - \frac{(y_k - H_k d_k)^T d_k}{(d_k^T d_k)^2} d_k d_k^T.$$

H_{k+1}^{PSB} ist wohldefiniert für $d_k \neq 0$, ist symmetrisch und erfüllt die Quasi-Newton-Gleichung (2.54).

Anstelle von Updates für H_k kann man auch direkt Updates der Inversen $B_k := H_k^{-1}$ angeben. Es zeigt sich, dass der inverse Update zur BFGS-Formel durch die DFP-Formel mit H_k^{-1}, y_k, d_k anstelle H_k, d_k, y_k gegeben ist und umgekehrt:

Lemma 2.10.2 *Es sei H_k symmetrisch und invertierbar.*

i) Gilt $y_k^T d_k \neq 0$, $d_k^T H_k d_k \neq 0$, $y_k^T H_k^{-1} y_k \neq 0$, dann sind H_{k+1}^{DFP} sowie H_{k+1}^{BFGS} invertierbar und es gilt

$$(H_{k+1}^{DFP})^{-1} = \Phi^{BFGS}(H_k^{-1}, y_k, d_k),$$
$$(H_{k+1}^{BFGS})^{-1} = \Phi^{DFP}(H_k^{-1}, y_k, d_k).$$

ii) Ist H_k symmetrisch positiv definit und $y_k^T d_k > 0$, dann sind H_{k+1}^{DFP}, H_{k+1}^{BFGS} und $H_{k+1}^{B,\lambda}$, $\lambda \in [0,1]$, wieder positiv definit.

Beweis: Siehe Übung. □

Wir halten fest, dass die Update-Formeln für $(H_{k+1}^{BFGS})^{-1}$ gegeben ist durch

Inverser BFGS-Update: Für $B_k = (H_k^{BFGS})^{-1}$ gilt

$$\begin{aligned} B_{k+1} &= (H_{k+1}^{BFGS})^{-1} = \Phi^{DFP}(B_k, y_k, d_k) \\ &= B_k + \frac{(d_k - B_k y_k)d_k^T + d_k(d_k - B_k y_k)^T}{y_k^T d_k} - \frac{(d_k - B_k y_k)^T y_k}{(y_k^T d_k)^2} d_k d_k^T \end{aligned} \tag{2.65}$$

Inverser DFP-Update: Für $B_k = (H_k^{DFP})^{-1}$ gilt

$$B_{k+1} = (H_{k+1}^{DFP})^{-1} = \Phi^{BFGS}(B_k, y_k, d_k) = B_k + \frac{d_k d_k^T}{d_k^T y_k} - \frac{B_k y_k (B_k y_k)^T}{y_k^T B_k y_k} \tag{2.66}$$

2.10.3 Eigenschaften von DFP-, BFGS- und PSB-Update

Wir stellen zunächst fest, dass der PSB-Update eine Aufdatierungen minimaler Änderung bezüglich der Frobenius-Norm ist:

Satz 2.10.3 (Minimaleigenschaft des PSB-Update)
Es sei $H_k \in \mathbb{R}^{n,n}$ eine symmetrische Matrix. Weiter seien $d_k, y_k \in \mathbb{R}^n$ mit $d_k \neq 0$. Dann ist H_{k+1}^{PSB} die eindeutige Lösung der Minimierungsaufgabe

$$H_+ = \operatorname*{argmin}_{H \in \mathbb{R}^{n,n}} \|H - H_k\|_F \quad \textit{u. d. Nebenbed.} \;\; H = H^T, \;\; H d_k = y_k. \tag{2.67}$$

Beweis: Wir können ebenso $\frac{1}{2}\|H - H_k\|_F^2$ als Zielfunktion verwenden. Mit $C := H - H_k$, $d = d_k$ und $v = y_k - H_k d_k$ erhalten wir dann das Problem

$$\min \frac{1}{2} \sum_{i,j=1}^n c_{ij}^2 \quad \text{u.d. Nebenbed.} \;\; C = C^T, \;\; Cd = v.$$

Die Lagrange-Funktion lautet

$$L(C, \lambda, \mu) = \frac{1}{2} \sum_{i,j=1}^n c_{ij}^2 + \sum_{i,j=1}^n \lambda_{ij}(c_{ij} - c_{ji}) + \sum_{i=1}^n \mu_i \left(v_i - \sum_{j=1}^n c_{ij} d_j \right).$$

Die notwendigen (und wegen der Konvexität hinreichenden) Optimalitätsbedingungen lauten

$$\frac{d}{dc_{ij}} L(C, \lambda, \mu) = 0, \quad C = C^T, \;\; Cd = v.$$

Die erste Gleichung ergibt
$$c_{ij} + \lambda_{ij} - \lambda_{ji} - \mu_i d_j = 0.$$
Wegen $c_{ij} = c_{ji}$ ergibt Addition der Gleichungen für i, j und j, i
$$2c_{ij} = \mu_i d_j + \mu_j d_i,$$
also
$$C = \frac{\mu d^T + d\mu^T}{2}.$$
Nun muss gelten
$$v = Cd = \frac{\mu d^T d + d\mu^T d}{2}.$$
Also ist $\mu = \alpha d + \frac{2v}{d^T d}$ und Einsetzen ergibt
$$\alpha d^T d + \frac{v^T d}{d^T d} = 0,$$
also $\alpha = -\frac{v^T d}{(d^T d)^2}$. Damit haben wir
$$C = \frac{vd^T + dv^T}{d^T d} - \frac{v^T d}{(d^T d)^2} dd^T$$
und dies ist wegen $C := H - H_k$, $d = d_k$ und $v = y_k - H_k d_k$ genau der PSB-Update. □

Wir zeigen nun, dass DFP- und BFGS-Update Aufdatierungen minimaler Änderung im folgenden Sinne sind:

Satz 2.10.4 (Minimaleigenschaft von DFP- und BFGS-Update)
Es sei $H_k \in \mathbb{R}^{n,n}$ eine symmetrische positiv definite Matrix und $d_k, y_k \in \mathbb{R}^n$ mit $y_k^T d_k > 0$. Weiter sei W_k eine beliebige symmetrisch positiv definite Matrix mit $W_k y_k = d_k$. Dann gilt:

i) H_{k+1}^{DFP} ist die eindeutige Lösung der Minimierungsaufgabe
$$H_+ = \operatorname*{argmin}_{H \in \mathbb{R}^{n,n}} \|W_k^{\frac{1}{2}}(H - H_k)W_k^{\frac{1}{2}}\|_F \quad \text{u. d. Nebenbed. } H = H^T, \ \ Hd_k = y_k.$$

ii) $(H_{k+1}^{BFGS})^{-1}$ ist die eindeutige Lösung der Minimierungsaufgabe
$$H_+^{-1} = \operatorname*{argmin}_{B \in \mathbb{R}^{n,n}} \|W_k^{-\frac{1}{2}}(B - H_k^{-1})W_k^{-\frac{1}{2}}\|_F \quad \text{u. d. Nebenbed. } B = B^T, \ \ By_k = d_k.$$

Beweis: Wegen $(H_{k+1}^{BFGS})^{-1} = \Phi^{DFP}(H_k^{-1}, y_k, d_k)$ ergibt sich ii) aus i) mit H_k^{-1}, y_k, d_k anstelle H_k, d_k, y_k.

zu i): Mit der Transformation

$$\tilde{H} = W_k^{\frac{1}{2}} H W_k^{\frac{1}{2}}, \quad \tilde{d}_k = W_k^{-\frac{1}{2}} d_k, \quad \tilde{y}_k = W_k^{\frac{1}{2}} y_k$$

erhalten wir ein Optimierungsproblem der Form (2.67) mit Lösung

$$\tilde{H}_+ = W_k^{\frac{1}{2}} H_+ W_k^{\frac{1}{2}} = \Phi^{PSB}(\tilde{H}_k, \tilde{d}_k, \tilde{y}_k).$$

Durch Vergleich von (2.64) und (2.62) folgt $H_+ = H_{k+1}^{DFP}$, da wegen $W_k y_k = d_k$ gilt

$$\tilde{d}_k^T \tilde{d}_k = d_k W_k^{-1} d_k = y_k^T d_k, \quad W_k^{-\frac{1}{2}}(\tilde{y}_k - \tilde{H}_k \tilde{d}_k) = y_k - H_k d_k, \quad W_k^{-\frac{1}{2}} \tilde{y}_k = y_k,$$
$$(\tilde{y}_k - \tilde{H}_k \tilde{d}_k)^T \tilde{d}_k = (y_k - H_k d_k)^T d_k.$$

□

Bemerkung: Die angegebene Minimaleigenschaft sichert automatisch die Invarianz des DFP- und BFGS-Verfahrens unter linearen Variablentransformationen. □

2.10.4 Globale Konvergenz des BFGS-Verfahrens

Wir verwenden nun das BFGS-Verfahren mit positiv definiter Hessematrix zusammen mit der Powell-Wolfe Schrittweitenregel.

Algorithmus 13 Globalisiertes BFGS-Verfahren

Wähle Konstanten $\gamma \in (0, 1/2)$ und $\theta \in (\gamma, 1)$. Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $B_0 = H_0^{-1} \in \mathbb{R}^{n,n}$.

Für $k = 0, 1, \ldots$:

1. *Falls $\nabla f(x_k) = 0$: STOP mit Ergebnis x_k.*
2. *Berechne $s_k = -B_k \nabla f(x_k)$.*
3. *Bestimme eine Schrittweite $\sigma_k > 0$ nach der Powell-Wolfe-Regel.*
4. *Setze $x_{k+1} = x_k + \sigma_k s_k$.*
5. *Berechne $B_{k+1} = H_{k+1}^{-1}$ nach dem inversen BFGS-Update* (2.65).

Bemerkung:

- Man kann auch H_k aufdatieren und s_k als Lösung von $H_k s_k = -\nabla f(x_k)$ bestimmen.

- Wir werden sehen, dass Algorithmus 13 global konvergiert, wenn f auf der Niveaumenge $N_f(x_0)$ gleichmäßig konvex ist. Für allgemeine nichtlineare Zielfunktionen f ist die globale Konvergenz von Algorithmus 13 ein offenes Problem.

  Um immer Konvergenz zu sichern, kann man das BFGS-Verfahren innerhalb des globalisierten Newton-artigen Verfahrens in Algorithmus 10 einsetzen: anstelle der Armijo-Regel verwendet man dann die Powell-Wolfe-Regel und berechnet durch $s_k^{NA} = -B_k \nabla f(x_k)$ einen BFGS-Schritt, der noch einem verallgemeinerten Winkel-Test unterworfen wird.

- In der Praxis macht man gelegentlich Restarts, setzt also $B_k = B_0$, falls $k \in m\mathbb{Z}$ mit festem $m \in \mathbb{N}$, z.B. $m = 100$.

- In der Übung wurde gezeigt, dass

$$B_k w = \text{bfgsrek}(k, w)$$

  rekursiv berechnet werden kann, man muss nur $B_{k-m}u = \text{bfgsrek}(k - m, u)$ für irgendein $m \leq k$ explizit berechnen können (in der Übung: $m = k$). Dies macht man sich beim Limited Memory BFGS-Verfahren (LBFGS-Verfahren) zunutze: Zur Berechnung einer Näherung $v \approx B_k w$ verwendet man folgende Limited-Memory-Variante:

$$v = \text{lbfgsrek}(k, w, m),$$

  wobei lbfgsrek(k, w, m) durch folgende Modifikationen der Schritte 1 und 3 von Algorithmus bfgsrek(k, w) entsteht:

  1. Falls $k = 0$ oder $m = 0$: Stop mit Ergebnis $v = B_0 w$.

     ⋮

  3. Berechne $w_2 = \text{lbfgsrek}(k - 1, w_1, m - 1)$.

  Man erlaubt also ein "Gedächtnis" von m BFGS-Updates und ersetzt die unbekannte Matrix B_{k-m} durch B_0. So kann man sehr große Optimierungsprobleme behandeln.

□

Wir beginnen mit der folgenden Beobachtung:

Proposition 2.10.5 *Die von Algorithmus 13 erzeugten Matrizen $H_k = B_k^{-1}$ sind symmetrisch positiv definit, es gilt $y_k^T d_k > 0$ und alle s_k sind Abstiegsrichtungen.*

Beweis: Wir zeigen die Behauptung induktiv.

$k = 0$: H_k ist für $k = 0$ nach Voraussetzung symmetrisch positiv definit. $s_k = -H_k^{-1}\nabla f(x_k)$ wird nur im Fall $\nabla f(x_k) \neq 0$ berechnet und ist dann eine Abstiegsrichtung. Die zweite Powell-Wolfe-Bedingung

$$\nabla f(x_k + \sigma_k s_k)^T s_k \geq \theta \nabla f(x_k)^T s_k \tag{2.28}$$

liefert

$$y_k^T d_k = (\nabla f(x_{k+1}) - \nabla f(x_k))^T(\sigma_k s_k) \geq \sigma_k(\theta - 1)\nabla f(x_k)^T s_k > 0.$$

Also ist H_{k+1} nach Lemma 2.10.2, ii) wieder positiv definit und die Behauptung folgt durch Induktion. □

Wir zeigen nun den folgenden globalen Konvergenzsatz.

Satz 2.10.6 (Globale Konvergenz des BFGS-Verfahrens)
Es sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Weiter sei $N_f(x_0) = \{x : f(x) \leq f(x_0)\}$ konvex und f sei gleichmäßig konvex auf $N_f(x_0)$ mit beschränkter Hessematrix, also

$$\mu\|s\|^2 \leq s^T \nabla^2 f(x) s \leq \eta\|s\|_2 \quad \forall\, x \in N_f(x_0), \ \ \forall\, s \in \mathbb{R}^n$$

mit Konstanten $0 < \mu \leq \eta$. Dann terminiert Algorithmus 13 entweder endlich oder erzeugt eine Folge (x_k), die gegen das eindeutige Minimum $\bar{x}$ von f konvergiert.

Beweis: Wir zeigen zunächst, dass f ein eindeutiges Minimum $\bar{x}$ besitzt. Hierzu genügt der Nachweis, dass $N_f(x_0)$ kompakt ist. Denn f besitzt dann ein Minimum $\bar{x}$ auf $N_f(x_0)$ und ist streng konvex auf $N_f(x_0)$ nach Satz 2.2.4. Somit ist $\bar{x}$ nach Satz 2.2.6, ii) das eindeutige globale Minimum von f auf $N_f(x_0)$ und dann auch auf ganz $\mathbb{R}^n$.

Nun zur Kompaktheit von $N_f(x_0)$: $N_f(x_0)$ ist jedenfalls abgeschlossen. Wäre $N_f(x_0)$ nicht beschränkt, dann gäbe es eine unbeschränkte Folge $(z_j) \subset N_f(x_0)$. Nun gilt mit geeigneten $t_j \in [0, 1]$

$$\begin{aligned} f(x_0) \geq f(z_j) &= f(x_0) + \nabla f(x_0)^T(z_j - x_0) + \frac{1}{2}(z_j - x_0)^T \nabla^2 f(x_0 + t_j(z_j - x_0))(z_j - x_0) \\ &\geq f(x_0) - \|\nabla f(x_0)\|\|z_j - x_0\| + \frac{\mu}{2}\|z_j - x_0\| \overset{j\to\infty}{\to} \infty. \end{aligned}$$

Dies ist ein Widerspruch.

Wir wissen bereits, dass $N_f(x_0)$ kompakt ist und f ein eindeutiges Minimum $\bar{x}$ besitzt. Wegen der strengen Konvexität von f ist $\bar{x}$ der einzige stationäre Punkt von f auf $N_f(x_0)$.

Algorithmus 13 terminiere nicht endlich. Wir zeigen, dass die Folgen (s_k) und (σ_k) zulässig sind.

Zulässigkeit von (σ_k): Nach Proposition 2.10.5 erzeugt Algorithmus 13 eine Folge von Abstiegsrichtungen (s_k). Da $N_f(x_0)$ kompakt ist, liefert also die Powell-Wolfe-Regel nach Satz 2.6.5 eine zulässige Schrittweitenfolge (σ_k).

Zulässigkeit von (s_k): Wir zeigen

$$c_k := \cos(\angle(-\nabla f(x_k), s_k)) = \frac{-\nabla f(x_k)^T s_k}{\|s_k\|\|\nabla f(x_k)\|} \not\to 0. \qquad (2.68)$$

Daraus folgt dann die Zulässigkeit: Zunächst gilt wegen der gleichmäßigen Konvexität nach Aufgabe H9, Blatt 3

$$\|\nabla f(x_k)\| \geq \mu\|x_k - \bar{x}\|, \quad f(x_k) - f(\bar{x}) \geq \frac{\mu}{2}\|x_k - \bar{x}\|^2. \tag{2.69}$$

Wir haben

$$\begin{aligned} \frac{-\nabla f(x_k)^T s_k}{\|s_k\|} \to 0 \quad \text{und} \quad c_k \not\to 0 \quad &\Longrightarrow \quad c_k\|\nabla f(x_k)\| \to 0 \quad \text{und} \quad c_k \not\to 0 \\ &\Longrightarrow \quad (\nabla f(x_k))_{k\in K} \to 0 \quad \text{für eine Teilfolge.} \end{aligned}$$

Nun liefert (2.69) zunächst $(x_k)_{k\in K} \to \bar{x}$ und die Monotonie von $(f(x_k))$ ergibt

$$\lim_{k\to\infty} f(x_k) - f(\bar{x}) = \lim_{k\in K\to\infty} f(x_k) - f(\bar{x}) = 0.$$

Wieder mit (2.69) folgt $x_k \to \bar{x}$ und schließlich $\|\nabla f(x_k)\| \to 0$.

Es bleibt also, (2.68) zu zeigen. Zunächst haben wir wegen $H_k s_k = -\nabla f(x_k)$

$$c_k := \cos(\angle(-\nabla f(x_k), s_k)) = \frac{s_k^T H_k s_k}{\|s_k\|\|H_k s_k\|} = \frac{d_k^T H_k d_k}{\|d_k\|\|H_k d_k\|},$$

wobei $d_k = \sigma_k s_k = x_{k+1} - x_k$, $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. Wir setzen

$$\mu_k := \frac{y_k^T d_k}{d_k^T d_k}, \quad \eta_k := \frac{y_k^T y_k}{y_k^T d_k}, \quad q_k := \frac{d_k^T H_k d_k}{d_k^T d_k}.$$

Nun gilt

$$y_k = \bar{H}_k d_k \quad \text{mit} \quad \bar{H}_k := \int_0^1 \nabla^2 f(x_k + t d_k)\, dt. \tag{2.70}$$

Dies ergibt

$$\mu_k \geq \mu, \quad \eta_k \leq \eta, \tag{2.71}$$

da wegen der gleichmäßigen Konvexität gilt

$$\mu_k = \frac{d_k^T \bar{H}_k d_k}{d_k^T d_k} \geq \mu$$

und

$$\eta_k = \frac{y_k^T y_k}{y_k^T \bar{H}_k^{-1} y_k} \leq \frac{1}{1/\eta} = \eta.$$

Spur und Determinante der BFGS-Approximation (2.63) berechnen sich zu

$$\text{(2.72)} \qquad \text{Spur}(H_{k+1}) = \text{Spur}(H_k) - \frac{\|H_k d_k\|^2}{d_k^T H_k d_k} + \frac{\|y_k\|^2}{y_k^T d_k} = \text{Spur}(H_k) - \frac{q_k}{c_k^2} + \eta_k,$$

$$\text{(2.73)} \qquad \det(H_{k+1}) = \det(H_k)\frac{y_k^T d_k}{d_k^T H_k d_k} = \det(H_k)\frac{\mu_k}{q_k}.$$

Nun betrachte für symmetrisch positiv definite Matrizen H die Funktion

$$\psi(H) = \text{Spur}(H) - \ln(\det(H)) = \sum_{i=1}^{n}(\lambda_i - \ln(\lambda_i)) > 0$$

mit den Eigenwerten λ_i von H. Einsetzen von (2.72) und (2.73) ergibt

$$\begin{aligned} 0 < \psi(H_{k+1}) &= \psi(H_k) + \eta_k - \frac{q_k}{c_k^2} - \ln(\mu_k) + \ln(q_k) \\ &= \psi(H_k) + (\eta_k - \ln(\mu_k) - 1) + \left(1 - \frac{q_k}{c_k^2} + \ln\left(\frac{q_k}{c_k^2}\right)\right) + \ln(c_k^2). \end{aligned}$$

Nun ist $1 - t + \ln t \le 0$ für alle $t > 0$ und zudem $\eta_k - \ln(\mu_k) - 1 \le \eta - \ln(\mu) - 1$ wegen (2.71). Wir erhalten mit $\kappa = \max(\eta - \ln(\mu) - 1, 0)$

$$0 < \psi(H_{k+1}) \le \psi(H_k) + \kappa + \ln(c_k^2)$$

und somit

$$0 < \psi(H_{k+1}) \le \psi(H_1) + k\kappa + \sum_{j=1}^{k} \ln(c_j^2).$$

Annahme, es gilt $c_k \to 0$. Dann finden wir $l \ge 0$ mit $\ln(c_k^2) \le -\kappa - 1$ für alle $k \ge l$ und dies ergibt den Widerspruch

$$0 < \psi(H_{k+1}) \le \psi(H_1) + k\kappa + \sum_{j=1}^{l} \ln(c_j^2) - (k-l)(\kappa+1) \to -\infty \quad \text{für } k \to \infty.$$

Somit ist $c_k \not\to 0$ gezeigt und wie oben begründet folgt die Zulässigkeit von (s_k).

Nach Satz 2.5.6 ist also jeder Häufungspunkt von $(x_k) \subset N_f(x_0)$ stationär. Da aber $\bar{x}$ der einzige stationäre Punkt im Kompaktum $N_f(x_0)$ ist, folgt $x_k \to \bar{x}$. $\square$

2.10.5 Lokale Konvergenzaussagen

Übergang zu schneller lokaler Konvergenz kann für Updates minimaler Änderungen (z.B. PSB-Update, DFP-Update) oder Updates minimaler Inversen-Änderung (BFGS-Update) mit Hilfe von Lemma 2.10.1 gezeigt werden, wenn bereits bekannt ist, dass (x_k) gegen einen Punkt $\bar{x}$ konvergiert, der die hinreichende Bedingung zweiter Ordnung erfüllt, und zudem gilt

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| < \infty. \tag{2.74}$$

Bemerkung: (2.74) ist offensichtlich erfüllt, falls $x_k \to \bar{x}$ Q-linear oder R-linear. □

Satz 2.10.7 *$f : \mathbb{R}^n \to \mathbb{R}$ sei zweimal stetig differenzierbar. Algorithmus 12 verwende Updates minimaler Änderung gemäß* (2.56) *mit einer Hilbertraum-Norm $\|\cdot\|_*$ (z.B. PSB-Updates).*

Die erzeugte Folge (x_k) konvergiere gegen einen Punkt $\bar{x}$, der die hinreichende Bedingung zweiter Ordnung erfüllt und nahe dem $\nabla^2 f$ lokal Lipschitz-stetig ist. Zudem sei (2.74) *erfüllt (z.B. falls $x_k \to \bar{x}$ Q-linear oder R-linear). Dann gilt*

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

und somit $x_k \to \bar{x}$ Q-superlinear nach Lemma 2.10.1.

Beweis: Für Interessierte: Siehe z.B. Kosmol [Ko93], 10.3. □

Zusatz: Dieselbe Aussage gilt, wenn Algorithmus 12 Updates minimaler Inversen-Änderung verwendet und die Folge (H_k) beschränkt bleibt. Denn dann folgt $\|H_{k+1}^{-1} - H_k^{-1}\| \to 0$ und wegen der Beschränktheit von (H_k) auch $\|H_{k+1} - H_k\| \to 0$ nach dem Zusatz zu Lemma 2.10.1. □

Eine ähnliche Aussage gilt für Updates minimaler Änderung bezüglich der gewichteten Norm (2.57).

Satz 2.10.8 *$f : \mathbb{R}^n \to \mathbb{R}$ sei zweimal stetig differenzierbar. Algorithmus 12 verwende den*

i) DFP-Update (2.62)

oder den

ii) BFGS-Update (2.63)*, wobei (H_k) beschränkt bleibe.*

Die erzeugte Folge (x_k) konvergiere gegen einen Punkt $\bar{x}$, der die hinreichende Bedingung zweiter Ordnung erfüllt, und $\nabla^2 f$ sei lokal Lipschitz-stetig bei $\bar{x}$. Zudem sei (2.74) *erfüllt (z.B. falls $x_k \to \bar{x}$ Q-linear oder R-linear). Dann gilt*

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

und somit $x_k \to \bar{x}$ Q-superlinear nach Lemma 2.10.1.

Beweis: Der Beweis ist ähnlich wie bei Satz 2.10.7. Nach Satz 2.10.4 ist der DFP-Update ein Update minimaler Änderung bezüglich der variablen Hilbertraum-Norm

$$\|C\|_k := \|W_k^{\frac{1}{2}} C W_k^{\frac{1}{2}}\|_F$$

in (2.57). Da wie oben gilt

$$(2.70) \qquad y_k = \bar{H}_k d_k \quad \text{mit} \quad \bar{H}_k := \int_0^1 \nabla^2 f(x_k + t d_k)\, dt.$$

und $\bar{H}_k$ für $k \geq l$, l groß genug, gleichmäßig positiv definit ist, kann man $W_k = \bar{H}_k^{-1}$ für $k \geq l$ wählen und man erhält für geeignete Konstanten $0 < \mu \leq \eta$ und $\gamma > 0$ mit $\alpha_k := \gamma \|\bar{H}_{k+1} - \bar{H}_k\|$ (siehe [Ko93], 10.7)

$$(2.75) \qquad \mu\|C\| \leq \|C\|_{k+1} \leq (1+\alpha_k)\|C\|_k \leq \eta\|C\| \quad \forall\, k \geq l$$

und wegen der lokalen Lipschitz-Stetigkeit von $\nabla^2 f$ folgt aus (2.74)

$$(2.76) \qquad \sum_{k=l}^{\infty} \alpha_k < \infty.$$

Eine leichte Modifikation des Beweises von Satz 2.10.7 liefert nun $\|H_{k+1} - H_k\| \to 0$.

Der BFGS-Update ist nach Satz 2.10.4 ein Update minimaler Inversen-Änderung bezüglich der variablen Hilbertraum-Norm

$$\|C\|_k' := \|W_k^{-\frac{1}{2}} C W_k^{-\frac{1}{2}}\|_F.$$

Da diese Norm wiederum die Eigenschaft (2.75), (2.76) hat, folgt analog $\|H_{k+1}^{-1} - H_k^{-1}\| \to 0$. $\square$

2.11 Richtungen negativer Krümmung

Die bisherige globale Konvergenztheorie stellt nur sicher, dass jeder Häufungspunkt $\bar{x}$ stationär ist. Es kann jedoch auftreten (was aber selten passiert, da Sattelpunkte oder Maximalpunkte meist keine Anziehungspunkte für Abstiegsverfahren sind), dass die notwendige Bedingung 2. Ordnung nicht gilt, also $\nabla^2 f(\bar{x})$ nicht positiv semidefinit ist.

Konvergenz gegen stationäre Punkte, welche die notwendige Bedingung 2. Ordnung erfüllen, läßt sich erreichen, indem man in einer Umgebung von stationären Punkten eine *Richtung negativer Krümmung* als Suchrichtung wählt und in diesem Fall die Schrittweitenwahl geeignet anpasst:

Modifizierte Suchrichtungswahl mit gegebenenfalls negativer Krümmung:

Mit Konstanten $\beta > 0$, $\eta \in (0,1)$ setze

$$s_k = \begin{cases} \text{wie bisher, falls } \lambda_{min}(\nabla^2 f(x_k)) \geq -\beta \|\nabla f(x_k)\|, \\ s_k^- \quad \text{sonst,} \end{cases}$$

wobei s_k^- eine normierte Richtung negativer Krümmung ist, genauer

$$\|s_k^-\| = 1, \quad s_k^{-T} \nabla^2 f(x_k) s_k^- \leq \eta \lambda_{min}(\nabla^2 f(x_k)), \quad \nabla f(x_k)^T s_k^- \leq 0.$$

Modifizierte Schrittweitenwahl: Mit einer Konstanten $\gamma \in (0, 1/2)$ setze

$$\sigma_k = \begin{cases} \text{wie bisher, falls } \lambda_{min}(\nabla^2 f(x_k)) \geq -\beta \|\nabla f(x_k)\|, \\ \text{sonst z.B. nach der Armijo-artigen Regel:} \\ \text{Wähle maximales } \sigma_k \in \{1, 1/2, 1/4, \ldots\} \text{ mit} \\ f(x_k) - f(x_k + \sigma_k s_k) \geq -\frac{\gamma}{2} \sigma_k^2 s_k^T \nabla^2 f(x_k) s_k \end{cases}$$

Ist dann f zweimal stetig differenzierbar, dann gilt in den globalen Konvergenzsätzen zudem

In jedem Häufungspunkt $\bar{x}$ von x_k gilt die notwendige Bedingung 2. Ordnung.

Der Beweis gelingt durch eine kleinere Modifikation des globalen Konvergenzbeweises für Abstiegsverfahren.

2.12 Trust-Region-Verfahren

Wir gehen noch kurz auf Trust-Region-Verfahren für das Problem

$$\text{(P)} \qquad \min_{x \in \mathbb{R}^n} \; f(x).$$

ein, die eine interessante Alternative zu Linesearch-Verfahren darstellen. Eine ausführliche Referenz für Trust-Region-Verfahren ist [CGT00].

2.12.1 Motivation von Trust-Region-Verfahren

Trust-Region-Verfahren bestimmen Suchrichtung und Schrittlänge simultan. Wir gehen zunächst von einer zweimal stetig differenzierbaren Zielfunktion $f : \mathbb{R}^n \to \mathbb{R}$ aus, obwohl wir sehen werden, dass Trust-Region-Verfahren bereits für stetig differenzierbare Funktionen geeignet sind.

Sei $x_k \in \mathbb{R}^n$ der aktuelle Punkt. Die Funktion $f(x_k + s)$ wird durch ein quadratisches Modell der Form

$$q_k(s) := f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T H_k s$$

approximiert, wobei $H_k \in \mathbb{R}^{n,n}$, $H_k = H_k^T$, eine Approximation der Hessematrix $\nabla^2 f(x_k)$ ist. Das quadratische Modell ist lokal eine gute Approximation von $f(x_k + s)$. Tatsächlich liefert Taylorentwicklung

$$f(x_k + s) = q_k(s) + \begin{cases} o(\|s\|^2), & \text{falls } H_k = \nabla^2 f(x_k), \\ O(\|s\|^2), & \text{sonst.} \end{cases}$$

Daher können wir dem Modell $q_k(s)$ zumindest in einem *Vertrauensbereich (Trust-Region)*

$$\{s \in \mathbb{R}^n \ : \ \|s\| \le \Delta_k\}$$

"trauen", wenn der Trust-Region-Radius Δ_k nicht zu groß gewählt wird.

Es liegt daher nahe, als neuen Punkt

$$x_{k+1} = x_k + s_k$$

zu wählen, wobei s_k als (unter Umständen approximative) Lösung des folgenden Trust-Region-Problems gewählt wird:

Trust-Region-Problem

(TR) $\qquad \min \ q_k(s) \quad \text{unter der Nebenbedingung } \|s\| \le \Delta_k.$

Bemerkung: Wir verwenden der Einfachheit halber die euklidische Norm für die Definition der Trust-Region, das muss aber nicht sein. □

Solange x_k nicht stationär ist, gilt dann für eine Lösung von s_k (und für jede sinnvolle Näherungslösung)

$$q_k(s_k) < q_k(0),$$

das Modell verspricht also eine Reduktion der Zielfunktion.

Bemerkung: Ist H_k positiv definit, so ist das globale Minimum s_k^{NA} von q_k gegeben durch

$$\nabla q_k(s_k^{NA}) = \nabla f(x_k) + H_k s_k^{NA} = 0, \quad \text{also} \quad H_k s_k^{NA} = -\nabla f(x_k).$$

Das globale Minimum s_k^{NA} ist also ein Newton-artiger Schritt und im Fall $H_k = \nabla^2 f(x_k)$ der klassische Newton-Schritt s_k^N. Ist $\|s_k^{NA}\| \le \Delta_k$, dann ist $s_k = s_k^{NA}$ die eindeutige Lösung von (TR). □

Bewertung des Schrittes und Anpassung des Trust-Region-Radius: Zur Beurteilung des Schrittes s_k vergleichen wir die Abnahme der Zielfunktion (actual reduction)

$$\mathrm{ared}_k(s_k) := f(x_k) - f(x_k + s_k)$$

mit der vom quadratischen Modell vorhergesagten Abnahme (predicted reduction)

$$\text{pred}_k(s_k) := q_k(0) - q_k(s_k).$$

Wir betrachten hierzu das Abnahmeverhältnis

$$\rho_k(s_k) := \frac{\text{ared}_k(s_k)}{\text{pred}_k(s_k)}. \tag{2.77}$$

Für $\nabla f(x_k) \neq 0$ ist $\text{pred}_k(s_k) > 0$ und somit das Abnahmeverhältnis wohldefiniert.

Wir wählen nun einen Parameter $\eta_1 \in (0,1)$ und gehen wie folgt vor:

Ist das Abnahmeverhältnis klein, genauer

$$\rho_k(s_k) < \eta_1,$$

dann verwerfen wir den Schritt und verkleinern die Trust-Region:

$$x_{k+1} := x_k, \quad \Delta_{k+1} < \Delta_k.$$

Ist die tatsächliche Abnahme ausreichend im Vergleich zur Modellabnahme, genauer

$$\rho_k(s_k) \geq \eta_1, \quad \eta_1 \in]0,1[,$$

dann akzeptieren wir den Schritt und wählen

$$x_{k+1} := x_k + s_k, \quad \Delta_{k+1} \geq \min(\Delta_k, \Delta_{min}).$$

Hierbei ist $\Delta_{min} \geq 0$ eine Unterschranke für den Trust-Region-Radius nach einem erfolgreichen Schritt.

Insgesamt ergibt sich folgendes Verfahren:

Algorithmus 14 Trust-Region-Verfahren
Wähle Konstanten $\Delta_{min} \geq 0$, $0 < \eta_1 < \eta_2 < 1$, $0 < \beta_1 < 1 < \beta_2$ (etwa $\eta_1 = 0.001$, $\eta_2 = 0.9$, $\beta_1 = 1/2$, $\beta_2 = 2$). Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$ und einen Trust-Region-Radius $\Delta_0 > 0$ mit $\Delta_0 \geq \Delta_{min}$.

Für $k = 0, 1, \ldots$:

1. *Falls $\nabla f(x_k) = 0$: STOP mit stationärem Punkt x_k.*
2. *Wähle eine symmetrische Approximation $H_k \in \mathbb{R}^{n,n}$ der Hessematrix.*
3. *Berechne eine ausreichend gute Näherungslösung s_k des Trust-Region-Problems (TR).*
4. *Berechne das Abstiegsverhältnis $\rho_k(s_k)$ gemäß* (2.77).

5. *Wähle den neuen Trust-Region-Radius gemäß*

$$\Delta_{k+1} = \begin{cases} \beta_1 \Delta_k, & \textit{falls } \rho_k(s_k) < \eta_1 \\ \max\{\Delta_{min}, \Delta_k\}, & \textit{falls } \eta_1 \le \rho_k(s_k) < \eta_2 \\ \max\{\Delta_{min}, \beta_2 \Delta_k\}], & \textit{falls } \rho_k(s_k) \ge \eta_2. \end{cases}$$

6. *Falls* $\rho_k(s_k) \ge \eta_1$ *setze* $x_{k+1} := x_k + s_k$*. Sonst setze* $x_{k+1} := x_k$*.*

Notation: Wir nennen einen Schritt s_k *erfolgreich*, falls $x_{k+1} := x_k + s_k$ und definieren

$$\mathcal{S} := \{k \in \mathbb{N}_0 \ : \ \text{Schritt } s_k \text{ ist erfolgreich}\} .$$

□

2.12.2 Globale Konvergenz

Wir untersuchen nun die globalen Konvergenzeigenschaften des Trust-Region-Verfahrens 14. Wir machen hierzu folgende Annahmen.

Voraussetzung 2.12.1

(TR1) Die Zielfunktion $f : \mathbb{R}^n \to \mathbb{R}$ ist stetig differenzierbar und nach unten beschränkt.

(TR2) Mit einer von k unabhängigen Konstanten $M_H > 0$ gilt

$$\|H_k\| \le M_H.$$

Cauchy-Abstiegsbedingung

Die volle Flexibilität der Trust-Region-Methodik ergibt sich daraus, dass das Trust-Region-Problem nur "hinreichend gut" gelöst werden muss. Dies eröffnet reichhaltige Möglichkeiten bei der Schrittberechnung. Wir wollen zunächst Mindestanforderungen an geeignete Näherungslösungen charakterisieren.

Die einfachste denkbare Näherungslösung des Trust-Region-Problems (TR) erhält man durch Restriktion des Trust-Region-Problems auf den Strahl in Richtung des steilsten Abstiegs $-\nabla f(x_k)$:

Bestimme den sogenannten *Cauchy-Punkt* s_k^c als Lösung des eindimensionalen Minimierungsproblems

$$\text{(2.78)} \qquad \min \ q_k(s) \quad \text{unter der Nebenbedingung } s = -t\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}, \quad t \in [0, \Delta_k].$$

Es ist nun naheliegend, von einer Näherungslösung s_k des Trust-Region-Problems zu fordern, dass sie zumindest einen Teil des Modellabstiegs liefert, den der Cauchy-Punkt garantiert:

Cauchy-Abstiegsbedingung (Fraction of Cauchy Decrease (FCD)):

Mit Konstanten $\mu \in]0,1[$ und $\gamma \geq 1$ gilt

$$\text{pred}_k(s_k) \geq \mu \, \text{pred}_k(s_k^c), \quad \|s_k\| \leq \gamma \Delta_k. \tag{2.79}$$

Lemma 2.12.2 *Es gelte Voraussetzung 2.12.1. Ist $\nabla f(x_k) \neq 0$ und erfüllt s_k die Cauchy-Abstiegsbedingung* (2.79)*, dann gilt*

$$\textit{pred}_k(s_k) \geq \mu \, \textit{pred}_k(s_k^c) \geq \frac{\mu}{2} \|\nabla f(x_k)\| \min\left\{\Delta_k, \|\nabla f(x_k)\|/M_H\right\}. \tag{2.80}$$

Für unsere Konvergenzanalyse ist lediglich eine Unterschranke für $\text{pred}_k(s_k)$ wie auf der rechten Seite von (2.80) von Bedeutung. Wir fordern daher nur die

Verallgemeinerte Cauchy-Abstiegsbedingung (Generalized FCD):

Mit Konstanten $\gamma \geq 1$ und $\mu > 0$ gilt

$$\text{pred}_k(s_k) \geq \frac{\mu}{2} \|\nabla f(x_k)\| \min\left\{\Delta_k, \|\nabla f(x_k)\|/M_H\right\}, \quad \|s_k\| \leq \gamma \Delta_k. \tag{2.81}$$

Konvergenzanalyse

Wir zeigen zunächst, dass der Trust-Region-Algorithmus unendlich viele erfolgreiche Schritte s_k, also $x_{k+1} = x_k + s_k$, generiert, wenn er nicht endlich terminiert.

Lemma 2.12.3 *Es gelte Voraussetzung 2.12.1. Dann gibt es zu jedem Punkt $\bar{x} \in \mathbb{R}^n$ mit $\nabla f(\bar{x}) \neq 0$ und zu jedem $\eta_2 \in (0,1)$ Konstanten $\delta > 0$ und $\Delta > 0$, so dass für alle $x_k \in B_\delta(\bar{x})$, alle $\Delta_k \in [0, \Delta]$ und alle s_k, die* (2.81) *erfüllen, gilt*

$$\rho_k(s_k) \geq \eta_2.$$

Insbesondere produziert Algorithmus 14 unendlich viele erfolgreiche Schritte, falls er nicht endlich terminiert.

Beweis: Sei $\|\nabla f(\bar{x})\| =: 2\varepsilon > 0$. Aus Stetigkeitsgründen finden wir $\delta > 0$ mit

$$\|\nabla f(x_k)\| \geq \varepsilon \quad \forall\, x_k \in B_\delta(\bar{x}).$$

Mit $\bar{\Delta} := \varepsilon/M_H$ gilt daher wegen (2.81)

$$\text{pred}_k(s_k) \geq \frac{\mu}{2} \varepsilon \Delta_k \quad \forall\, x_k \in B_\delta(\bar{x}),\ \Delta_k \in]0, \bar{\Delta}].$$

Somit gilt für $x_k \in B_\delta(\bar{x}),\ \Delta_k \in]0, \bar{\Delta}]$

$$|\rho_k(s_k) - 1| = \frac{|\mathrm{ared}_k(s_k) - \mathrm{pred}_k(s_k)|}{\mathrm{pred}_k(s_k)} \le \frac{|\mathrm{ared}_k(s_k) - \mathrm{pred}_k(s_k)|}{\mu\varepsilon\Delta_k/2}.$$

Daher ist $\rho_k(s_k) \ge \eta_2$ sichergestellt, falls

$$|\mathrm{ared}_k(s_k) - \mathrm{pred}_k(s_k)| \le (1-\eta_2)\mu\varepsilon\Delta_k/2. \tag{2.82}$$

Taylorentwicklung liefert mit einem $\tau \in [0, 1]$.

$$\begin{aligned}
|\mathrm{ared}_k(s_k) - \mathrm{pred}_k(s_k)| &= |q_k(s_k) - f(x_k + s_k)| \\
&= |(\nabla f(x_k) - \nabla f(x_k + \tau s_k))^T s_k + \frac{1}{2} s_k^T H_k s_k| \\
&\le \|\nabla f(x_k) - \nabla f(x_k + \tau s_k)\| \|s_k\| + \frac{1}{2} M_H \|s_k\|^2 \\
&\le \|\nabla f(x_k) - \nabla f(x_k + \tau s_k)\| \gamma\Delta_k + \frac{1}{2} M_H \gamma^2 \Delta_k^2.
\end{aligned} \tag{2.83}$$

Wegen der gleichmäßigen Stetigkeit von ∇f auf dem Kompaktum $\overline{B_{\delta+\gamma\bar{\Delta}}(\bar{x})}$ finden wir

$$0 < \Delta \le \min\left\{\frac{(1-\eta_2)\mu\varepsilon}{2M_H\gamma^2}, \bar{\Delta}\right\}$$

mit

$$\|\nabla f(x_k) - \nabla f(x_k + \tau s_k)\| \le \frac{(1-\eta_2)\mu\varepsilon}{4\gamma} \quad \forall\, x_k \in B_\delta(\bar{x}),\ \ \|s_k\| \le \gamma\Delta.$$

Einsetzen in (2.83) liefert nun (2.82) für alle $x_k \in B_\delta(\bar{x})$ solange $\Delta_k \in [0, \Delta]$.

Terminiert der Algorithmus nicht, dann gilt $\nabla f(x_k) \neq 0$. In jedem nicht erfolgreichen Schritt gilt $x_{k+1} = x_k$, $\Delta_{k+1} \le \beta_1\Delta_k$. Sei nun $\bar{x} := x_k$ und $\Delta > 0$ wie eben. Nach endlich vielen Schritten ist dann $\Delta_{k+l} \le \Delta$ und somit s_{k+l} ein erfolgreicher Schritt. □

Wir können nun folgenden Konvergenzsatz beweisen.

Satz 2.12.4 *Es gelte Voraussetzung 2.12.1. Terminiert Algorithmus 14 nicht endlich und erfüllen alle Schritte s_k die verallgemeinerte Cauchy-Abstiegsbedingung* (2.81)*, dann gilt*

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0. \tag{2.84}$$

Ist zudem ∇f gleichmäßig stetig auf der Menge $\{x_k\ :\ k \in \mathbb{N}_0\}$ (also inbesondere, wenn (x_k) beschränkt bleibt), dann gilt sogar

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

Als technisches Hilfsmittel zeigen wir zunächst das folgende

Lemma 2.12.5 *Es gelten die Voraussetzungen von Satz 2.12.4. Sei $\mathcal{J} \subset \mathcal{S}$ eine unendliche Menge erfolgreicher Iterationen mit $\|\nabla f(x_k)\| \geq \varepsilon > 0$ für alle $k \in \mathcal{J}$. Dann gilt*

$$\sum_{k \in \mathcal{J}} \Delta_k < \infty. \tag{2.85}$$

Beweis: Setze $f_k = f(x_k)$. Wegen $\mathcal{J} \subset \mathcal{S}$ liefern Schritt 6 und die Abstiegsbedingung (2.81)

$$f_k - f_{k+1} = \text{ared}_k(s_k) \geq \eta_1 \text{pred}_k(s_k) \geq \eta_1 \frac{\mu}{2} \varepsilon \min\{\Delta_k, \varepsilon/M_H\} \quad \forall\, k \in \mathcal{J}.$$

Für alle $k \notin \mathcal{J}$ ist $f_k - f_{k+1} \geq 0$ und daher ergibt Summation

$$f_0 - f_l = \sum_{k=0}^{l-1} (f_k - f_{k+1}) \geq \sum_{k \in \mathcal{J}, k<l} (f_k - f_{k+1}) \geq \eta_1 \frac{\mu}{2} \varepsilon \sum_{k \in \mathcal{J}, k<l} \min\{\Delta_k, \varepsilon/M_H\}.$$

Da f wegen (TR1) nach unten beschränkt ist, ergibt $l \to \infty$

$$\sum_{k \in \mathcal{J}} \min\{\Delta_k, \varepsilon/M_H\} < \infty.$$

Hieraus folgt (2.85). □

Beweis: (von Satz 2.12.4) Wir zeigen zunächst (2.84). Angenommen, der Algorithmus terminiert nicht endlich und es ist $\liminf_{k \to \infty} \|\nabla f(x_k)\| \neq 0$. Dann gibt es $\varepsilon > 0$ mit $\|\nabla f(x_k)\| \geq \varepsilon$ für alle $k \geq 0$. Wir zeigen, dass dann gilt

$$\lim_{k \to \infty} x_k = \bar{x} \tag{2.86}$$

mit einem $\bar{x}$. Wegen $\|\nabla f(x_k)\| \geq \varepsilon$ für alle $k \in \mathcal{S}$ ergibt Lemma 2.12.5

$$\sum_{k \in \mathcal{S}} \Delta_k < \infty. \tag{2.87}$$

Nun gilt $\|x_k - x_{k+1}\| = 0$ für $k \notin \mathcal{S}$ und

$$\|x_k - x_{k+1}\| = \|s_k\| \leq \gamma \Delta_k \quad \forall\, k \in \mathcal{S}$$

Also ist für alle $0 \leq l < m$ wegen (2.87)

$$\|x_l - x_m\| \leq \sum_{k \in \mathcal{S}, k \geq l} \|x_k - x_{k+1}\| \leq \sum_{k \in \mathcal{S}, k \geq l} \gamma \Delta_k \to 0 \quad \text{für } l, m \to \infty.$$

Somit ist (x_k) eine Cauchy-Folge und (2.86) ist nachgewiesen.

Aber nun liefert Lemma 2.12.2 Konstanten $\delta, \Delta > 0$, so dass $\rho_k(s_k) \geq \eta_2$ sobald $x_k \in B_\delta(\bar{x})$ und $\Delta_k \leq \Delta$. Wegen (2.86) gilt $x_k \in B_\delta(\bar{x})$ für alle $k \geq K'$, K' groß genug. Wir zeigen nun induktiv, dass

$$\Delta_k \geq \min\{\beta_1 \Delta, \Delta_{K'}\} \quad \forall\, k \geq K'. \tag{2.88}$$

Dies liefert einen Widerspruch zu (2.87) und damit war die Annahme $\liminf_{k\to\infty} \|\nabla f(x_k)\| \neq 0$ falsch.

Es bleibt, (2.88) zu zeigen. Für $k = K'$ ist das klar.

Sei nun $k \geq K'$. Im Fall $\Delta_k \geq \Delta$ gilt dann nach Schritt 5 des Algorithmus $\Delta_{k+1} \geq \beta_1 \Delta_k \geq \beta_1 \Delta$.

Ist $\Delta_k \leq \Delta$, dann ist $\rho_k(s_k) \geq \eta_2$ nach Lemma 2.12.2 und daher $\Delta_{k+1} \geq \Delta_k$ nach Schritt 5.

Damit ist (2.88) gezeigt.

Sei schließlich ∇f gleichmäßig stetig auf $\{x_k \,:\, k \in \mathbb{N}_0\}$. Angenommen, der Algorithmus terminiert nicht endlich und es ist $\lim_{k\to\infty} \|\nabla f(x_k)\| \neq 0$. Da (2.84) bereits nachgewiesen ist, finden wir dann $\varepsilon > 0$ und eine unendliche Indexkette $0 \leq l_1 < m_1 < l_2 < m_2 < \ldots$ mit

$$\|\nabla f(x_{l_i})\| \geq 2\varepsilon, \quad \|\nabla f(x_k)\| \geq \varepsilon,\ l_i \leq k < m_i, \quad \|\nabla f(x_{m_i})\| < \varepsilon.$$

Setze nun $\mathcal{J}_i = \{l_i, \ldots, m_i - 1\} \cap \mathcal{S}$ und $\mathcal{J} = \bigcup_i \mathcal{J}_i$. Dann ist $\|\nabla f(x_k)\| \geq \varepsilon$ für $k \in \mathcal{J}$ und somit nach Lemma 2.12.5

$$\sum_{k\in\mathcal{J}} \Delta_k < \infty.$$

Dies ergibt

$$\|x_{l_i} - x_{m_i}\| = \sum_{k\in\mathcal{J}_i} \|s_k\| \leq \gamma \sum_{k\in\mathcal{J}_i} \Delta_k \to 0 \quad \text{für } i \to \infty.$$

Andererseits ist

$$\|\nabla f(x_{l_i}) - f(x_{m_i})\| \geq 2\varepsilon - \varepsilon = \varepsilon$$

im Widerspruch zur gleichmäßigen Stetigkeit von ∇f auf (x_k) $\Box$

Liegt die exakte Hessematrix nicht vor, dann kann H_k zum Beispiel durch Quasi-Newton-Updates berechnet werden, insbesondere SR1-Update (wobei Updates nur im Fall $|(y_k - H_k d_k)^T d_k| \geq \varepsilon \|y_k - H_k d_k\| \|d_k\|$ durchgeführt werden) und BFGS-Update (wobei Updates nur im Fall $y_k^T d_k > \varepsilon \|y_k\| \|d_k\|$ gemacht werden).

2.12.3 Übergang zu schneller lokaler Konvergenz

Der Trust-Region Algorithmus 14 läßt die genaue Wahl des Schrittes s_k und der Hessematrix-Approximationen H_k offen. Wir geben nun eine einfache Schrittberechnung an, die mit der

Wahl $H_k = \nabla^2 f(x_k)$ schnelle lokale Konvergenz unter geeigneten Voraussetzungen sichergestellt.

Sei hierzu f zweimal stetig differenzierbar. Nach Satz 2.12.4 ist dann jeder Häufungspunkt $\bar{x}$ von (x_k) ein stationärer Punkt, falls (x_k) beschränkt ist.

Wir betrachten nun den Fall, dass ein Häufungspunkt $\bar{x}$ die hinreichende Bedingung zweiter Ordnung erfüllt.

Verwenden wir im quadratischen Modell die exakte Hessematrix $H_k = \nabla^2 f(x_k)$, dann ist H_k für x_k nahe bei $\bar{x}$ positiv definit und der Newton-Schritt $s_k = s_k^N$ ist im Falle $\|s_k^N\| \le \Delta_k$ die exakte Lösung des Trust-Region-Problems (TR). Dies motiviert folgendes *Trust-Region-Newton-Verfahren*:

Algorithmus 15 Algorithmus 14 mit $H_k = \nabla^2 f(x_k)$ und folgender Implementierung von Schritt 3:

3. Falls H_k positiv definit, berechne den Newton-Schritt s_k^N als Lösung von

   (2.89) $$H_k s_k^N = -\nabla f(x_k).$$

   Falls $\|s_k^N\| \le \Delta_k$ setze $s_k = s_k^N$ und gehe zu 4.

   Andernfalls berechne ein s_k, das die Cauchy-Abstiegsbedingung erfüllt.

Bemerkungen:

- Offensichtlich erfüllen die in Algorithmus 14 berechneten Schritte s_k die verallgemeinerte Cauchy-Abstiegsbedingung, da die Wahl $s_k = s_k^N$ im Falle $\|s_k^N\| \le \Delta_k$, $\nabla^2 f(x_k)$ positiv definit die exakte Lösung des Trust-Region-Problems liefert.
- Bei der praktischen Implementierung wählt man in der Regel eine der folgenden Vorgehensweisen:
  - ”Versuch” einer Cholesky-Faktorisierung $H_k = L_k L_k^T$. Im Erfolgsfalle kann diese dann gleich zur Lösung der Newton-Gleichung (2.89) herangezogen werden. s_k wird nun als Minimum auf dem stückweise linearen Pfad durch Cauchy-Punkt und Newton-Punkt bestimmt (Powell’s Dogleg-Verfahren).
  - Approximative Lösung der Newton-Gleichung (2.89) durch ein (vorkonditioniertes) CG-Verfahren, das bei Verlassen der Trust-Region oder bei Auftreten negativer Krümmung abgebrochen wird (Steihaug-CG-Verfahren).

Algorithmus 15 geht tatsächlich in das Newton-Verfahren über:

Satz 2.12.6 *Es gelte Voraussetzung 2.12.1. Weiter sei f zweimal stetig differenzierbar und die Niveaumenge $N_f(x_0) = \{x : f(x) \le f(x_0)\}$ sei kompakt. Hat die von Algorithmus 15 erzeugte Folge (x_k) einen Häufungspunkt $\bar{x}$, der die hinreichende Bedingung zweiter Ordnung erfüllt, dann gilt:*

i) $x_k \to \bar{x}$ *und* $\nabla f(x_k) \to \nabla f(\bar{x}) = 0$.

ii) *Es gibt $l > 0$ mit $\Delta_k \ge \Delta_l$ für alle $k \ge l$.*

iii) *Es gibt $l' > 0$ mit $s_k = s_k^N$ für alle $k \ge l'$ mit dem Newton-Schritt s_k^N und $x_k \to \bar{x}$ superlinear, also*

$$\|x_{k+1} - \bar{x}\|_2 = o(\|x_k - \bar{x}\|_2) \quad \textit{für } k \to \infty.$$

2.12.4 Charakterisierung der Lösung des Trust-Region-Problems

Es ist eine wenig überraschend, dass sich bei euklidischer Trust-Region-Norm eine einfache notwendige und hinreichende Bedingung für ein globales Minimum von (TR) angeben läßt.

Satz 2.12.7 *Sei $H_k \in \mathbb{R}^{n,n}$ eine beliebige symmetrische Matrix und sei $\Delta_k > 0$ beliebig. Dann besitzt das Trust-Region-Problem (TR) mindestens ein globales Minimum. $\bar{s}$ ist genau dann ein globales Minimum von (TR), wenn es $\lambda \ge 0$ gibt mit*

(1) $(H_k + \lambda I)\,\bar{s} = -\nabla f(x_k)$,

(2) $(H_k + \lambda I)$ *positiv semidefinit,*

(3) entweder $\|\bar{s}\| = \Delta_k$ *oder* $\|\bar{s}\| < \Delta_k$ *und* $\lambda = 0$.

Beweis: Siehe Übung. □

Diese Charakterisierung liefert die Basis zur exakten numerischen Lösung von (TR): Ist H_k positiv definit und $\|s_k^{NA}\| \le \Delta_k$ mit $s_k^{NA} = -H_k^{-1}\nabla f(x_k)$, dann ist

$$\bar{s} = s_k^{NA}$$

die Lösung von (TR). Sonst betrachte das Gleichungssystem

$$(H_k + \lambda I)s(\lambda) = -\nabla f(x_k)$$

für $\lambda \ge \max(0, -\lambda_{min}(H_k))$ und bestimme eine Lösung $\bar{\lambda} \ge \max(0, -\lambda_{min}(H_k))$ der Gleichung

$$\frac{1}{\|s(\lambda)\|} - \frac{1}{\Delta_k} = 0.$$

Dann ist $\bar{s} = s(\bar{\lambda})$ Lösung von (TR).

Bemerkung: Die Gleichung $\frac{1}{\|s(\lambda)\|} - \frac{1}{\Delta_k} = 0$ verhält sich weniger nichtlinear als die Gleichung $\|s(\lambda)\| - \Delta_k = 0$ und ist schnell mit dem Newton-Verfahren approximativ lösbar.

2.12.5 Näherungsweise Lösung des Trust-Region-Problems

Wir betrachten zum Abschluß zwei bewährte Verfahren zur Berechnung von Näherungslösungen des Trust-Region-Problems, welche die Cauchy-Abstiegsbedingung erfüllen und mit der Wahl $H_k = \nabla^2 f(x_k)$ Übergang zu schneller lokaler Konvergenz unter den Voraussetzungen des letzten Abschnitts erlauben.

Das Dogleg-Verfahren

Es sei $\nabla f(x_k) \neq 0$. Dann ist der Cauchy-Punkt s_k^c gemäß (2.78) eindeutig bestimmt und garantiert ausreichend Abstieg. Ist H_k positiv definit, so liefert der Newton-Schritt $s_k^N = -H_k^{-1}\nabla f(x_k)$ das globale Minimum des quadratischen Modells $q_k(s)$. Es liegt daher nahe, eine Näherungslösung s_k des Trust-Region-Problems (TR) durch Minimierung von $q_k(s)$ auf dem stückweise linearen Pfad $s_k^{DL}(t)$ durch Cauchy-Punkt s_k^c und Newton-Punkt s_k^N (falls H_k positiv definit ist), also

$$s_k^{DL}(t) := \begin{cases} t s_k^c, & 0 \le t \le 1, \\ s_k^c + (t-1)(s_k^N - s_k^c), & 1 < t \le 2, \end{cases}$$

innerhalb der Trust-Region zu bestimmen. Dies führt auf das folgende Dogleg-Verfahren:

Dogleg-Verfahren:

Verwende folgende Implementierung von Schritt 3 in Algorithmus 14:

3. Berechne den Cauchy-Punkt s_k^c gemäß (2.78) und falls H_k positiv definit zudem den Newton-Schritt $s_k^N = -H_k^{-1}\nabla f(x_k)$. Setze

$$s_k = \begin{cases} s_k^N, & \text{falls } H_k \text{ positiv definit und } \|s_k^N\| \le \Delta_k, \\ s_k^c, & \text{falls } H_k \text{ nicht positiv definit}, \\ s_k^{DL}(t^*), & t^* = \underset{t\in[1,2],\|s_k^{DL}(t)\|\le\Delta_k}{\operatorname{argmin}} \quad q_k(s_k^{DL}(t)) \quad \text{sonst.} \end{cases}$$

Bemerkung: Offensichtlich liefert das Dogleg-Verfahren einen Schritt s_k mit $q_k(s_k) \le q_k(s_k^c)$. Zudem ist es für den schnell lokal konvergenten Algorithmus 15 geeignet, da $s_k = s_k^N$ gewählt wird, falls H_k positiv definit und $\|s_k^N\| \le \Delta_k$ ist. □

Das folgende Lemma zeigt, dass der Dogleg-Pfad ein Abstiegspfad ist, falls H_k positiv definit ist. $s_k^{DL}(t^*)$ ist also der "letzte" Punkt auf der Strecke zwischen s_k^c und s_k^N, der in der Trust-Region liegt.

Lemma 2.12.8 *Es sei $\nabla f(x_k) \neq 0$ und H_k positiv definit. Dann ist $t \in [0,2] \mapsto q_k(s_k^{DL}(t))$ streng monoton fallend.*

Beweis: Nach Definition ist s_k^c das Minimum von q_k auf der Strecke $s_k^{DL}(t)$, $t \in [0,1]$. Ist H_k positiv definit, dann ist q_k streng konvex mit eindeutigem globalen Minimum s_k^N. Die Funktion $t \in [1,2] \mapsto q_k(s_k^{DL}(t))$ ist somit streng konvex mit eindeutigem Minimum bei $t = 2$. $\square$

Das Steihaug-CG-Verfahren

Wir wissen, dass zur iterativen Minimierung einer streng konvexen quadratischen Funktion

$$q(s) = c^T s + \frac{1}{2} s^T H s$$

das Verfahren der konjugierten Gradienten (CG-Verfahren) verwendet werden kann (siehe Blatt 6, Aufgabe 25). Verwenden wir den Startpunkt $y_0 = 0$, dann minimiert das CG-Verfahren minimiert zunächst entlang des negativen Gradienten und steigt dann weiter auf einem stückweise linearen Pfad ab. Sind d_j die generierten konjugierten Richtungen, dann gilt

$$q(y_j) = \min_{s \in \text{span}\{d_0, \dots, d_{j-1}\}} q(s).$$

Zur näherungsweisen Lösung des Trust-Region-Problems (TR) schlug Steihaug [St83] die folgende naheliegende Modifikation des CG-Verfahrens vor:

Steihaug CG-Verfahren:

Wende das CG-Verfahren zur Minimierung von $q = q_k$ an. Terminiere mit $s_k = s^S$ nach folgenden Regeln: Sei $\varepsilon \in]0,1[$, $\nu > 0$ und $g_j = \nabla q(y_j)$.

1. Falls $\|g_j\| \le \min(\varepsilon \|g_0\|, \nu \|g_0\|^2)$: STOP mit Ergebnis $s^S = y_j$.

2. Ist d_j eine Richtung nicht-positiver Krümmung, also $d_j^T H d_j \le 0$: STOP mit Ergebnis

$$s^S = y_j - \alpha^* \text{sgn}\,(g_j^T d_j) d_j$$

   mit $\alpha^* \ge 0$, so dass $\|s^S\| = \Delta_k$.

3. Falls $\|y_{j+1}\| > \Delta_k$ (CG-Pfad verläßt die Trust-Region): STOP mit Ergebnis

$$s^S = y_j - \alpha^* d_j$$

   mit $\alpha^* \ge 0$, so dass $\|s^S\| = \Delta_k$.

Man kann folgendes zeigen:

Satz 2.12.9 *Das Steihaug-CG-Verfahren hat folgende Eigenschaften:*

i) $s_k = s^S$ *erfüllt die Cauchy-Abstiegsbedingung* (2.79).

ii) Es gilt $0 < \|y_1\|_2 < \|y_2\|_2 < \ldots$*, der CG-Pfad entfernt sich also in der euklidischen Norm vom Ursprung.*

iii) Mit der Wahl $H_k = \nabla^2 f(x_k)$ *bleibt die Konvergenzaussage von Satz 2.12.6 erhalten, wobei ein inexakter Newton-Schritt* $\tilde{s}_k^N$ *mit*

$$H_k \tilde{s}_k^N = -\nabla f(x_k) + O(\|\nabla f(x_k)\|^2)$$

anstelle des exakten Newton-Schritts s_k^N *verwendet wird.*

Beweis: zu i): Das Steihaug-CG-Verfahren beginnt mit der Berechnung des Cauchy-Punkts und steigt dann unter Umständen weiter ab.

zu ii): siehe [St83] oder [NW99].

zu iii): Der Beweis von Satz 2.12.6 kann ähnlich wie bei inexakten Newton-verfahren leicht angepasst werden. □

Kapitel 3

Optimierung mit Nebenbedingungen

3.1 Einführung

Wir betrachten nun das allgemeine

Nichtlineare Optimierungsproblem (NLP):

$$\text{(NLP)} \qquad \min_{x \in \mathbb{R}^n} f(x) \quad \text{u.d. Nebenbedingung} \quad h(x) = 0, \quad c(x) \leq 0.$$

mit zumindest stetig differenzierbaren Funktionen $f : \mathbb{R}^n \to \mathbb{R}$, $h = (h_1, \dots, h_p)^T : \mathbb{R}^n \to \mathbb{R}^p$ und $c = (c_1, \dots, c_m)^T : \mathbb{R}^n \to \mathbb{R}^m$. Hierbei ist die Ungleichung $c(x) \leq 0$ komponentenweise zu verstehen.

Definition 3.1.1 (Zulässiger Bereich, Indexmenge aktiver Nebenbedingungen)

a) Die Menge

$$Z = \{x \in \mathbb{R}^n : h(x) = 0, \; c(x) \leq 0\}. \tag{3.1}$$

heißt zulässiger Bereich *von (NLP).*

b) Ein Punkt $x \in \mathbb{R}^n$ heißt zulässig, *falls $x \in Z$ gilt.*

c) Zu $x \in Z$ definieren wir die Indexmenge aktiver Ungleichungsnebenbedingungen $\mathcal{A}(x)$ *und die* Indexmenge inaktiver Ungleichungsnebenbedingungen $\mathcal{I}(x)$ *durch*

$$\begin{aligned} \mathcal{A}(x) &= \{i : 1 \leq i \leq m, \; c_i(x) = 0\}, \\ \mathcal{I}(x) &= \{i : 1 \leq i \leq m, \; c_i(x) < 0\} = \{1, \dots, m\} \setminus \mathcal{A}(x). \end{aligned}$$

Notationen: Für $v \in \mathbb{R}^n$, $A \in \mathbb{R}^{n,m}$ und Indexmengen $\mathcal{J} \subset \{1, \dots, n\}$, $\mathcal{K} \subset \{1, \dots, m\}$ bezeichne $v_{\mathcal{J}} = (v_i)_{i \in \mathcal{J}}$ den zu $\mathcal{J}$ gehörigen Teilvektor und $A_{\mathcal{J}\mathcal{K}} = (a_{ij})_{i \in \mathcal{J}, j \in \mathcal{K}}$ die zu $\mathcal{J} \times \mathcal{K}$ gehörige Teilmatrix. $\square$

3.2 Notwendige Optimalitätsbedingungen

Wir wollen notwendige und hinreichende Bedingungen für ein lokales Minimum von (NLP) herleiten. Hierbei spielt der Tangentialkegel von Z in einem $x \in Z$ eine wichtige Rolle.

Definition 3.2.1 *Eine Menge $C \subset \mathbb{R}^n$ heißt Kegel, falls gilt*

$$x \in C \implies \lambda x \in C \quad \forall \lambda > 0.$$

Definition 3.2.2 *Sei $M \subset \mathbb{R}^n$ eine nichtleere Menge. Der* Tangentialkegel *an M in $x \in M$ ist definiert durch*

$$T(M;x) = \left\{ s \in \mathbb{R}^n \ : \ \exists \, \eta_k > 0, x_k \in M : \quad \lim_{k\to\infty} x_k = x, \quad \lim_{k\to\infty} \eta_k(x_k - x) = s \right\}.$$

Es gilt folgende Optimalitätsbedingung.

Satz 3.2.3 *Es sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Dann gilt für jede lokale Lösung $\bar{x}$ von (NLP)*

$$\bar{x} \in Z \quad \textit{und} \quad \nabla f(\bar{x})^T s \ge 0 \quad \forall \, s \in T(Z;\bar{x}). \tag{3.2}$$

Beweis: $\bar{x} \in Z$ ist klar. Sei nun $s \in T(Z;\bar{x})$. Dann gibt es $(x_k) \subset Z$ und $\eta_k > 0$ mit $x_k \to \bar{x}$ und $\eta_k(x_k - \bar{x}) \to s$. Dies ergibt

$$0 \le \eta_k(f(x_k) - f(\bar{x})) = \nabla f(\bar{x})^T \eta_k(x_k - \bar{x}) + \eta_k o(\|x_k - \bar{x}\|_2) \to \nabla f(\bar{x})^T s$$

wegen $\eta_k o(\|x_k - \bar{x}\|_2) \to 0$. □

3.2.1 Die Karush-Kuhn-Tucker-Bedingungen

Leider läßt sich $T(Z;\bar{x})$ in dieser Form nur schwer angeben. Um eine handlichere Form der Optimalitätsbedingung (3.2) zu erhalten, suchen wir nach Bedingungen, unter denen sich $T(Z;\bar{x})$ durch die Nebenbedingungen c, h und ihre Ableitung darstellen läßt. Linearisierung der aktiven Nebenbedingungen legt folgende Definiton nahe:

Definition 3.2.4 *Der Kegel*

$$T_L(c,h;x) = \left\{ s \in \mathbb{R}^n \ : \ \nabla c(x)_i^T s \le 0, \ \ i \in \mathcal{A}(x), \ \ \nabla h(x)^T s = 0 \right\}$$

heißt Linearisierungskegel *von Z in $x \in Z$ zur Darstellung* (3.1).

Bemerkung: Die inaktiven Nebenbedingungen werden nicht berücksichtigt, da sie auch in einer Umgebung von x inaktiv bleiben und somit lokal den zulässigen Bereich nicht einschränken. □

Bemerkung: Der Linearisierungskegel $T_L(c, h; x)$ hängt von der Darstellung (3.1) von Z ab, wie folgendes Beispiel zeigt:

Beispiel 3.2.1

$$Z = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \ : \ x_2 = 0 \right\}.$$

Dann haben wir mit $h(x) = x_2$, $\tilde{h}(x) = x_2^2$ die Darstellungen

$$Z = \left\{ x \in \mathbb{R}^2 \ : \ h(x) = 0 \right\} = \left\{ x \in \mathbb{R}^2 \ : \ \tilde{h}(x) = 0 \right\}.$$

Aber es gilt für jedes $x = (x_1, 0)^T \in Z$ $T_L(\tilde{h}; x) \neq T(Z; x) = T_L(h; x)$. Siehe Übung. □

Wir stellen zunächst fest, dass immer die Inklusion $T(Z; x) \subset T_L(c, h; x)$ gilt:

Proposition 3.2.5 *Sei* $c : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^n \to \mathbb{R}^p$ *stetig differenzierbar und sei* Z *durch* (3.1) *gegeben. Dann gilt*

$$T(Z; x) \subset T_L(c, h; x) \quad \forall\, x \in Z.$$

Beweis: Sei $x \in Z$ beliebig fest. Nun sei $s \in T(Z; x)$ beliebig mit zugehörigen $\{x_k\} \subset Z$ und $\eta_k > 0$. Es gilt

$$0 = \eta_k(h(x_k) - h(x)) = \nabla h(x)^T \eta_k(x_k - x) + \eta_k o(\|x_k - x\|_2) \to \nabla h(x)^T s,$$

also $\nabla h(x)^T s = 0$ und analog für alle $i \in \mathcal{A}(x)$

$$0 \geq \eta_k(c_i(x_k) - c_i(x)) = \nabla c_i(x)^T \eta_k(x_k - x) + \eta_k o(\|x_k - x\|_2) \to \nabla c_i(x)^T s,$$

also $\nabla c_i(x)^T s \leq 0$. Damit ist $s \in T_L(c, h; x)$ gezeigt. □

Wir würden gerne den Tangentialkegel $T(Z; \bar{x})$ in (3.2) durch $T_L(c, h; \bar{x})$ ersetzen, da $s \in T_L(c, h; \bar{x})$ im Gegensatz zu $s \in T(Z; \bar{x})$ ohne weiteres zu prüfen ist. Die fehlende Inklusion $T_L(c, h; x) \subset T(Z; x)$ gilt jedoch nicht immer, wie Beispiel 3.2.1 zeigt.

Definition 3.2.6 *Die Bedingung*

(ACQ) $$T_L(c, h; x) = T(Z; x)$$

heißt Abadie Constraint Qualification *für* $x \in Z$.

Bemerkung: Wegen Proposition 3.2.5 ist (ACQ) äquivalent zu

$$T_L(c, h; x) \subset T(Z; x).$$

□

Wir erhalten unmittelbar aus Satz 3.2.3

Satz 3.2.7 *$\bar{x}$ sei ein lokales Minimum von (NLP) und es gelte (ACQ) für $\bar{x}$. Dann gilt:*

$$\bar{x} \in Z \quad \textit{und} \quad \nabla f(\bar{x})^T s \geq 0 \quad \forall\, s \in T_L(c, h; \bar{x}). \tag{3.3}$$

Wir wollen (3.3) weiter vereinfachen. Dies ist mit dem aus der linearen Optimierung bekannten Lemma von Farkas möglich:

Lemma 3.2.8 (Lemma von Farkas)
Es seien $A \in \mathbb{R}^{n,m}$, $B \in \mathbb{R}^{n,p}$ und $c \in \mathbb{R}^n$ beliebig. Dann sind die beiden folgenden Aussagen äquivalent:

i) Für alle $s \in \mathbb{R}^n$ mit $A^T s \leq 0$ und $B^T s = 0$ gilt $b^T s \leq 0$.

ii) Es gibt $u \in \mathbb{R}^m$, $u \geq 0$, und $v \in \mathbb{R}^p$ mit $b = Au + Bv$.

Im Fall $m = 0$ bzw. $p = 0$ fallen A und u bzw. B und v weg.

Anwendung auf (3.3) liefert die bekannten Karush-Kuhn-Tucker-Bedingungen.

Satz 3.2.9 (Notwendige Optimalitätsbedingung erster Ordnung, Karush-Kuhn-Tucker-Bedingungen, KKT-Bedinungen)
$\bar{x}$ sei ein lokales Minimum von (NLP) und es gelte (ACQ) für $\bar{x}$. Dann gelten die

Karush-Kuhn-Tucker-Bedingungen:
Es gibt Lagrange-Multiplikatoren $\bar{\lambda} \in \mathbb{R}^m$ und $\bar{\mu} \in \mathbb{R}^p$ mit

1) $h(\bar{x}) = 0,\ \ c(\bar{x}) \leq 0$ *(Zulässigkeit)*

2) $\nabla f(\bar{x}) + \nabla c(\bar{x})\bar{\lambda} + \nabla h(\bar{x})\bar{\mu} = 0$ *(Multiplikatorregel)*

3) $\bar{\lambda} \geq 0,\ \ \bar{\lambda}^T c(\bar{x}) = 0$ *(Komplementaritätsbedingung)*

Wir nennen einen Punkt $\bar{x}$, der 1)–3) erfüllt, KKT-Punkt oder stationären Punkt *von (NLP). Zudem nennen wir ein Tripel $(\bar{x}, \bar{\lambda}, \bar{\mu})$, das 1)–3) erfüllt, kurz* KKT-Tripel *von (NLP).*

Beweis: Natürlich gilt 1). Setze $A = \nabla c_{\mathcal{A}(\bar{x})}(\bar{x})$, $B = \nabla h(\bar{x})$ und $b = -\nabla f(\bar{x})$. Dann gilt $T_L(c, h; \bar{x}) = \{s \ :\ A^T s \leq 0,\ B^T s = 0\}$ und somit wegen (3.3)

$$b^T s \leq 0 \quad \text{für alle } s \in \mathbb{R}^n \text{ mit } A^T s \leq 0 \text{ und } B^T s = 0.$$

Nach dem Farkas-Lemma 3.2.8 ist dies äquivalent zur Existenz von $u \in \mathbb{R}^{|\mathcal{A}(\bar{x})|}$ $u \geq 0$, $v \in \mathbb{R}^m$ mit

$$-\nabla f(\bar{x}) = b = Au + Bv = \nabla c_{\mathcal{A}(\bar{x})}(\bar{x})u + \nabla h(\bar{x})v.$$

Wählen wir $\bar{\mu} = v$ und $\bar{\lambda} \in \mathbb{R}^m$ mit $\bar{\lambda}_{\mathcal{A}(\bar{x})} = u$, $\bar{\lambda}_{\mathcal{I}(\bar{x})} = 0$, dann gilt 2). Schließlich gilt auch 3) wegen $\bar{\lambda} \geq 0$ und $\bar{\lambda}_{\mathcal{I}(\bar{x})} = 0$. $\square$

Die KKT-Bedingungen lassen sich bequem unter Verwendung der Lagrange-Funktion schreiben.

Definition 3.2.10 *Die Funktion $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ mit*

$$L(x, \lambda, \mu) = f(x) + \lambda^T c(x) + \mu^T h(x)$$

heißt Lagrange-Funktion *für (NLP).*

Dann lauten die KKT-Bedingungen

Karush-Kuhn-Tucker-Bedingungen:
Es gibt Lagrange-Multiplikatoren $\bar{\lambda} \in \mathbb{R}^m$ und $\bar{\mu} \in \mathbb{R}^p$ mit

1) $h(\bar{x}) = 0,\ c(\bar{x}) \leq 0$ (Zulässigkeit)

2) $\nabla_x L(\bar{x}, \bar{\lambda}, \bar{\mu}) = 0$ (Multiplikatorregel)

3) $\bar{\lambda} \geq 0,\ \bar{\lambda}^T c(\bar{x}) = 0$ (Komplementaritätsbedingung)

Geometrische Interpretation von KKT, 2)-3): Der negative Gradient $-\nabla f(\bar{x})$ liegt in dem von den Gradienten aktiver Nebenbedingungen aufgespannten Kegel $K = \{s :\ s = \nabla c_{\mathcal{A}(\bar{x})}u + \nabla h(\bar{x})v,\ \ u \in \mathbb{R}^{|\mathcal{A}(\bar{x})|},\ u \geq 0,\ v \in \mathbb{R}^p\}$.

Für die Identifikation aktiver Nebenbedingungen und das Konvergenzverhalten vieler Optimierungsverfahren ist es von Bedeutung, ob die Komplementaritätsbedingung in einem strikten Sinne gilt.

Definition 3.2.11 *Es sei $\bar{x}$ ein KKT-Punkt von (NLP) mit zugehörigen Lagrange-Multiplikatoren $\bar{\lambda}, \bar{\mu}$. In $\bar{x}$ gilt* strikte Komplementarität, *wenn gilt*

$$\bar{\lambda}_{\mathcal{A}(\bar{x})} > 0.$$

3.2.2 Constraint Qualifications

Wir wollen nun Bedingungen angeben, die (ACQ) sicherstellen.

Definition 3.2.12 *Sei $x \in Z$. Eine Bedingung, die (ACQ) sicherstellt, heißt* Constraint Qualification *(CQ) für x.*

Satz 3.2.13 *Sei $x \in Z$. Die Bedingung*

$$c_i(x) \textit{ konkav für } i \in \mathcal{A}(x), \quad h(x) \textit{ affin linear}$$

ist eine Constraint Qualification für x.

Ein wichtiger Spezialfall ist c, h affin linear, also lineare Nebenbedingungen.

Bemerkung: Beweis: Wir müssen nur $T_L(c, h; x) \subset T(Z; x)$ zeigen. Sei $s \in T_L(c, h; x)$. Dann gilt

$$\nabla h(x)^T s = 0, \quad \nabla c_i(x)^T s \leq 0 \;\; i \in \mathcal{A}(x).$$

Wir zeigen $s \in T(Z; x)$. Setze $\eta_k = k$, $x_k = x + \frac{1}{k}s$. Dann gilt $\eta_k(x_k - x) = s$ und es bleibt $(x_k) \subset Z$ für k groß genug nachzuweisen. Da h affin linear ist, gilt

$$h(x_k) = h(x) + \nabla h(x)^T(x_k - x) = \nabla h(x)^T(x_k - x) = \frac{1}{k}\nabla h(x)^T s = 0,$$

und da $-c_i(x)$ konvex ist für $i \in \mathcal{A}(x)$, haben wir nach Satz 2.2.3

$$c_i(x_k) = c_i(x_k) - c_i(x) \leq \nabla c_i(x)^T(x_k - x) = \frac{1}{k}\nabla c_i(x)^T s \leq 0 \quad \forall\, i \in \mathcal{A}(x).$$

Wegen $x_k \to x$ ist für alle $k \geq l$, l groß genug, $c_i(x_k) < 0$ für $i \in \mathcal{I}(x)$. Insgesamt folgt $x_k \in Z$ für $k \geq l$ und daher $s \in T(Z; x)$. $\square$

Für allgemeine nichtlineare Nebenbedingungen ist die folgende Constraint Qualification von großer Bedeutung.

Definition 3.2.14 *Ein Punkt $x \in Z$ erfüllt die* Mangasarian-Fromovitz Constraint Qualification (MFCQ)*, wenn gilt*

a) Die Gradienten $\{\nabla h_i(x) \,:\, i = 1, \ldots, p\}$ der Gleichungsnebenbedingungen sind linear unabhängig oder h ist affin linear.

b) Es gibt $d \in \mathbb{R}^n$ mit

$$\nabla h(x)^T d = 0, \quad \nabla c_i(x)^T d < 0 \quad i \in \mathcal{A}(x).$$

Satz 3.2.15 *Erfüllt $x \in Z$ (MFCQ), dann gilt (ACQ). (MFCQ) ist also eine Constraint Qualification.*

Beweis: Wir müssen wieder $T_L(c,h;x) \subset T(Z;x)$ zeigen.

Zunächst existiert eine stetig differenzierbare Funktion $\Phi : B_\varepsilon(0) \to \mathbb{R}^n$ mit

(3.4) $\quad h(x+s+\Phi(s)) = 0 \quad \forall\, s \in B_\varepsilon(0), \quad \Phi(0) = 0, \quad \mathrm{Ker}(\nabla h(x)^T) \subset \mathrm{Ker}(\Phi'(0)).$

Ist h affin linear, dann können wir $\Phi \equiv 0$ nehmen.

Sonst hat $\nabla h(x)^T$ nach (MFCQ), a) vollen Zeilenrang. Seien die Spalten von $W \in \mathbb{R}^{n,n-p}$ eine Basis von $\mathrm{Ker}(\nabla h(x)^T)$ und betrachte die stetig differenzierbare Abbildung

$$F(s,y) = \begin{pmatrix} h(x+s+y) \\ W^T y \end{pmatrix}.$$

Es gilt $F(0,0) = 0$ und $F_y'(0,0) = \binom{\nabla h(x)^T}{W^T}$ ist invertierbar. Nach dem Satz über implizite Funktionen gibt es also $\varepsilon > 0$ und eine stetig differenzierbare Funktion $\Phi : B_\varepsilon(0) \to \mathbb{R}^n$ mit $\Phi(0) = 0$ und

$$F(s, \Phi(s)) = 0 \quad \forall\, s \in B_\varepsilon(0).$$

Insbesondere folgt

$$0 = F_s'(0,0) + F_y'(0,0)\Phi'(0) = \begin{pmatrix} \nabla h(x)^T \\ 0 \end{pmatrix} + \begin{pmatrix} \nabla h(x)^T \\ W^T \end{pmatrix} \Phi'(0),$$

also

$$\Phi'(0) = -F_y'(0,0)^{-1} \begin{pmatrix} \nabla h(x)^T \\ 0 \end{pmatrix}$$

und daher $\mathrm{Ker}(\nabla h(x)^T) \subset \mathrm{Ker}(\Phi'(0))$. Damit erfüllt Φ (3.4).

Sei nun $s \in T_L(c,h;x)$ beliebig. Im Fall $s = 0$ ist $s \in T(Z;x)$ klar. Sei nun $s \neq 0$ und d aus (MFCQ), b). Dann gibt es $\nu > 0$ mit

$$\nabla c_i(x)^T d \le -\nu \quad \forall\, i \in \mathcal{A}(x).$$

Wir zeigen zunächst, dass

$$s_j = s + \frac{1}{j} d \in T(Z;x) \quad \forall\, j \in \mathbb{N}.$$

Setze hierzu

$$s_{j,k} = \frac{s_j}{k}, \quad \eta_{j,k} = k, \quad x_{j,k} = x + s_{j,k} + \Phi(s_{j,k}) \quad k \ge l,$$

mit l so groß, dass $s_{j,k} \in B_\varepsilon(0)$. Wir zeigen, dass gilt

(3.5) $$\lim_{k\to\infty} x_{j,k} = x, \quad \lim_{k\to\infty} \eta_{j,k}(x_{j,k} - x) = s_j.$$

Wegen $\lim_{k\to\infty} s_{j,k} = 0$ ist der erste Grenzwert klar. Weiter gilt $\nabla h(x)^T s_{j,k} = 0$ und daher mit (3.4)

$$\Phi(s_{j,k}) = \Phi'(0)s_{j,k} + o(\|s_{j,k}\|) = 0 + o(\|s_{j,k}\|) = o(1/k).$$

Dies ergibt

$$\eta_k(x_k - x) = k(s_{j,k} + \Phi(s_{j,k})) = s_j + ko(1/k) \to s_j \quad \text{für } k \to \infty.$$

Damit gilt (3.5) und zum Nachweis von $s_j \in T(Z;x)$ ist nur noch $x_{j,k} \in Z$ zu zeigen für $k \geq l_j$, $l_j \geq l$ groß genug. Zunächst ist

$$h(x_{j,k}) = 0 \quad \forall\, k \geq l$$

nach (3.4). Weiter gilt wegen $c_i(x) < 0$ für $i \in \mathcal{I}(x)$

$$c_i(x_{j,k}) < 0 \quad \forall\, i \in \mathcal{I}(x) \quad \forall\, k \geq l_j$$

mit $l_j \geq l$ groß genug und für $i \in \mathcal{A}(x)$ wegen $\nabla c_i(x)^T s \leq 0$, $\nabla c_i(x)^T d \leq -\nu$

$$\begin{aligned} c_i(x_k) &= 0 + \nabla c_i(x)^T (s_{j,k} + \Phi(s_{j,k})) + o(\|s_{j,k}\|) \\ &\leq \frac{1}{jk} \nabla c_i(x)^T d + o(1/k) \leq \frac{-\nu}{jk} + o(1/k) < 0 \quad \text{für } k \geq l_j,\, i \in \mathcal{A}(x) \end{aligned}$$

mit $l_j \geq l$ groß genug. Dies zeigt $x_{j,k} \in Z$ für $k \geq l_j$ und zusammen mit (3.5) erhalten wir $s_j \in T(Z;x)$.

Schließlich ist auch $s = \lim_{j\to\infty} s_j \in T(Z;x)$. Wir konstruieren hierzu aus $\eta_{j,k}$, $x_{j,k}$ Diagonalfolgen $\eta_{j,k(j)}$, $x_{j,k(j)}$ mit

$$(x_{j,k(j)}) \subset Z, \quad x_{j,k(j)} \to x, \quad \eta_{j,k(j)}(x_{j,k(j)} - x) \to s$$

wie folgt: Wegen der Wahl von $x_{j,k}$, $\eta_{j,k}$ für festes j finden wir eine Folge $(k(j)) \subset \mathbb{N}$ mit

$$\lim_{j\to\infty} k(j) = \infty, \quad x_{j,k(j)} \in Z, \quad \|x_{j,k(j)} - x\| \leq \frac{1}{j}, \quad \|\eta_{j,k(j)}(x_{j,k(j)} - x) - s_j\| \leq \frac{1}{j}.$$

Dann gilt $x_{j,k(j)} \to x$ und $\eta_{j,k(j)}(x_{j,k(j)} - x) \to s$ und es folgt $s \in T(Z;x)$. $\square$

Eine bequemere Form von (MFCQ) ist folgende Constraint Qualification.

Definition 3.2.16 *Ein Punkt $x \in Z$ erfüllt die* Positive Linear Independence Constraint Qualification (PLICQ)*, wenn gilt*

a) Die Gradienten $\{\nabla h_i(x) \,:\, i = 1, \ldots, p\}$ der Gleichungsnebenbedingungen sind linear unabhängig oder h ist affin linear.

b) Es gilt

$$\nabla c(x)u + \nabla h(x)v = 0, \;\; u \geq 0, \;\; u_{\mathcal{I}(x)} = 0 \quad \Longrightarrow \quad u_{\mathcal{A}(x)} = 0.$$

Im Fall $m = 0$ oder $\mathcal{A}(x) = \emptyset$ entfällt b) und bei $p = 0$ entfällt a) und der Term $\nabla h(x)v$ in b).

Satz 3.2.17 *(MFCQ) und (PLICQ) sind äquivalent.*

Beweis: (MFCQ) $\Longrightarrow$ (PLICQ): Es ist nur (PLICQ), b) zu zeigen. Sei $\nabla c(x)u+\nabla h(x)v = 0, \quad u \geq 0, \quad u_{\mathcal{I}(x)} = 0$ Sei d der Vektor aus (MFCQ), b). Mit $w = \nabla c(x)^T d$ ist dann $w_{\mathcal{A}(x)} < 0$ und wir erhalten

$$0 = d^T(\nabla c(x)u + \nabla h(x)v) = w^T u + 0 = w^T_{\mathcal{A}(x)} u_{\mathcal{A}(x)}.$$

Wegen $w_{\mathcal{A}(x)} < 0$, $u_{\mathcal{A}(x)} \geq 0$ folgt $u_{\mathcal{A}(x)} = 0$.

(PLICQ) $\Longrightarrow$ (MFCQ): Setze $r = |\mathcal{A}(x)|$, $A = \nabla c_{\mathcal{A}(x)}(x) = (a_1, \ldots, a_r)$ und $B = \nabla h(x)$. Weiter sei $A_j = (a_1, \ldots, a_j)$ und $T_1 = \{d \,:\, B^T d = 0\}$,

$$T_j = \left\{d \,:\, B^T d = 0, \quad A^T_{j-1} d < 0\right\}, \quad j \geq 2.$$

Gilt (MFCQ), b) nicht, dann gibt es $1 \leq j \leq r$ mit

$$T_j \neq \emptyset, \quad T_{j+1} = \emptyset.$$

Also gilt

$$a_j^T d \geq 0 \quad \forall\, d \in T_j = \left\{d \,:\, B^T d = 0, \quad A^T_{j-1} d < 0\right\}$$

und dasselbe gilt auch für den Abschluss von T_j, also

$$a_j^T s \geq 0 \quad \forall\, s \in \bar{T}_j = \left\{s \,:\, B^T s = 0, \quad A^T_{j-1} s \leq 0\right\}$$

(mit $s \in \bar{T}_j$ und $d \in T_j$ ist $d_j = s + d/j$ in T_j und es gilt $d_j \to s$). Das Lemma von Farkas liefert nun $u_j \in \mathbb{R}^{j-1}$, $u_j \geq 0$, $v \in \mathbb{R}^p$ mit

$$-a_j = A_{j-1} u_j + B v_j.$$

Wegen $0 \neq \binom{u_j}{1} \geq 0$ gilt also (PLICQ), b) nicht. $\Box$

Oft verwendet man anstelle (PLICQ) die stärkere Constraint Qualification (LICQ):

Definition 3.2.18 *Ein Punkt $x \in Z$ erfüllt die* Linear Independence Constraint Qualification (LICQ)*, wenn die Spalten der Matrix $(\nabla c_{\mathcal{A}(\bar{x})}(\bar{x}), \nabla h(\bar{x}))$, also die Gradienten der aktiven Nebenbedingungen, linear unabhängig sind.*

Natürlich gilt (LICQ) $\Longrightarrow$ (PLICQ) und daher ist (LICQ) erst recht eine CQ. Offensichtlich sichert (LICQ) die Eindeutigkeit der Lagranga-Multiplikatoren:

Satz 3.2.19 *Es sei $\bar{x}$ ein KKT-Punkt, für den (LICQ) gilt. Dann sind die zugehörigen Lagrange-Multiplikatoren $\bar{\lambda}$, $\bar{\mu}$ eindeutig.*

Beweis: Wegen KKT, 3) ist $\bar{\lambda}_{\mathcal{I}(\bar{x})} = 0$. Daher liefert KKT, 2)

$$-\nabla f(\bar{x}) = (\nabla c_{\mathcal{A}(\bar{x})}(\bar{x}), \nabla h(\bar{x})) \begin{pmatrix} \bar{\lambda}_{\mathcal{A}(\bar{x})} \\ \bar{\mu} \end{pmatrix}$$

und die Spalten der Matrix auf der rechten Seite sind linear unabhängig nach (LICQ). Daher sind $\bar{\lambda}_{\mathcal{A}(\bar{x})}, \bar{\mu}$ eindeutig bestimmt. □

3.2.3 Konvexe Probleme und die KKT-Bedingungen

Im Fall konvexer Probleme sind die KKT-Bedingungen hinreichend dafür, dass ein globales Minimum vorliegt.

Satz 3.2.20 *(NLP) sei konvex, h sei also affin linear und f, c_i seien konvex. Dann gilt:*

i) Jedes lokale Minimum $\bar{x}$ ist globales Minimum. Gilt in $\bar{x}$ zudem eine Constraint Qualification, dann gelten in $\bar{x}$ die KKT-Bedingungen aus Satz 3.2.9.

ii) Erfüllt $\bar{x}$ die KKT-Bedingungen aus Satz 3.2.9, dann ist $\bar{x}$ ein globales Minimum von (NLP).

Beweis: zu i): Dies folgt aus Satz 2.2.6 und Satz 3.2.9.

zu ii): Sei $\bar{x}$ ein Punkt, der die KKT-Bedingungen erfüllt. Sei $x \in Z$ beliebig. Wegen $\bar{\lambda} \geq 0$ und der Konvexität von c_i folgt nach Satz 2.2.3 und den KKT-Bedingungen 3)

$$\bar{\lambda}_i \nabla c_i(\bar{x})^T (x - \bar{x}) \leq \bar{\lambda}_i (c_i(x) - c_i(\bar{x})) = \bar{\lambda}_i c_i(x) \leq 0.$$

Weiter ist $\nabla h(\bar{x})^T (x - \bar{x}) = h(x) - h(\bar{x}) = 0$, da h affin linear ist. Zusammen mit der Konvexität von f und KKT, 2) ergibt sich

$$\begin{aligned} f(x) - f(\bar{x}) \geq \nabla f(\bar{x})^T (x - \bar{x}) &= -\bar{\lambda}^T \nabla c(\bar{x})^T (x - \bar{x}) - \bar{\mu}^T \nabla h(\bar{x})^T (x - \bar{x}) \\ &= -\bar{\lambda}^T \nabla c(\bar{x})^T (x - \bar{x}) \geq 0. \end{aligned}$$

□

3.3 Hinreichende Optimalitätsbedingungen

Bereits der unrestringierte Fall zeigt, dass ein stationärer Punkt nicht notwendigerweise ein lokales Minimum ist. Verschwindet die Richtungsableitung in eine Richtung des Linearisierungskegels, dann ist das Krümmungsverhalten der Lagrange-Funktion in diese Richtung von Bedeutung. Die kritischen Richtungen sind durch folgenden Kegel gegeben.

Definition 3.3.1 *Zu $x \in Z$ und $\lambda \in \mathbb{R}^m$ mit $\lambda \geq 0$ definieren wir den Kegel*

$$T_K(c,h;x,\lambda) := \left\{ s \in \mathbb{R}^n \ : \ \nabla h(x)^T s = 0,\ \nabla c_i(x)^T s \begin{cases} = 0, & \textit{falls } i \in \mathcal{A}(x) \textit{ und } \lambda_i > 0 \\ \leq 0, & \textit{falls } i \in \mathcal{A}(x) \textit{ und } \lambda_i = 0 \end{cases} \right\}.$$

Bemerkung: Offensichtlich gilt $T_K(c,h;x,\lambda) \subset T_L(c,h;x)$. □

Satz 3.3.2 (Hinreichende Bedingung zweiter Ordnung)
$f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$ und $h : \mathbb{R}^n \to \mathbb{R}^p$ seien zweimal stetig differenzierbar. $\bar{x}$ erfülle die KKT-Bedingungen aus Satz 3.2.9 mit Lagrange-Multiplikatoren $\bar{\lambda} \in \mathbb{R}^n$, $\bar{\mu} \in \mathbb{R}^p$. Gilt zudem

$$s^T \nabla^2_{xx} L(\bar{x},\bar{\lambda},\bar{\mu}) s > 0 \quad \forall\, s \in T_K(c,h;\bar{x},\bar{\lambda}) \setminus \{0\}, \tag{3.6}$$

dann ist $\bar{x}$ ein isoliertes lokales Minimum von (NLP).

Beweis: $\bar{x}$ erfülle mit $\bar{\lambda} \in \mathbb{R}^n$, $\bar{\mu} \in \mathbb{R}^p$ die KKT-Bedingungen und es gelte (3.6). Annahme, $\bar{x}$ ist kein isoliertes lokales Minimum. Dann gibt es eine Folge $(x_k) \subset Z$ mit $x_k \to \bar{x}$ und $f(x_k) \leq f(\bar{x})$. Setze $d_k = x_k - \bar{x}$. Unter Umständen nach Übergang zu einer Teilfolge gilt

$$\frac{d_k}{\|d_k\|} \to s$$

mit $\|s\| = 1$. Taylorentwicklung liefert

$$0 \geq \frac{f(x_k) - f(\bar{x})}{\|d_k\|} = \nabla f(\bar{x})^T \frac{d_k}{\|d_k\|} + \frac{o(\|d_k\|)}{\|d_k\|} \to \nabla f(\bar{x})^T s,$$

und zudem gilt $s \in T(Z;\bar{x}) \subset T_L(h,c;\bar{x})$, also

$$\nabla f(\bar{x})^T s \leq 0, \quad \nabla c_i(\bar{x})^T s \leq 0, \ \ i \in \mathcal{A}(\bar{x}), \quad \nabla h(\bar{x})^T s = 0. \tag{3.7}$$

Aus den KKT-Bedingungen folgt

$$0 = \nabla_x L(\bar{x},\bar{\lambda},\bar{\mu})^T s = \underbrace{\nabla f(\bar{x})^T s}_{\leq 0} + \sum_{i \in \mathcal{A}(\bar{x})} \underbrace{\bar{\lambda}_i \nabla c_i(\bar{x})^T s}_{\leq 0} + \sum_{i=1}^{p} \bar{\mu}_i \underbrace{\nabla h_i(\bar{x})^T s}_{=0}.$$

Somit ergibt sich $\nabla c_i(\bar{x})^T s = 0$ für alle $i \in \mathcal{A}(\bar{x})$ mit $\bar{\lambda}_i > 0$, da sonst die rechte Seite negativ wäre. Zusammen mit (3.7) ergibt sich $s \in T_K(c,h;\bar{x},\bar{\lambda}) \setminus \{0\}$.

Zudem haben wir wegen der KKT-Bedingungen

$$L(x_k,\bar{\lambda},\bar{\mu}) = f(x_k) + \sum_{i \in \mathcal{A}(\bar{x})} \bar{\lambda}_i c_i(x_k) \leq f(x_k) \leq f(\bar{x}) = L(\bar{x},\bar{\lambda},\bar{\mu}).$$

Taylorentwicklung ergibt nun wegen $\nabla_x L(\bar{x},\bar{\lambda},\bar{\mu}) = 0$

$$0 \geq \frac{L(x_k,\bar{\lambda},\bar{\mu}) - L(\bar{x},\bar{\lambda},\bar{\mu})}{\|d_k\|^2} = 0 + \frac{1}{2}\frac{d_k^T \nabla^2_{xx} L(\bar{x},\bar{\lambda},\bar{\mu}) d_k}{\|d_k\|^2} + \frac{o(\|d_k\|^2)}{\|d_k\|^2} \to \frac{1}{2} s^T \nabla^2_{xx} L(\bar{x},\bar{\lambda},\bar{\mu}) s.$$

Aber wegen $s \in T_K(c,h;\bar{x},\bar{\lambda}) \setminus \{0\}$ ist dies ein Widerspruch zu (3.6). □

3.3.1 Beweis des Lemmas von Farkas

Wir geben der Vollständigkeit halber noch einen Beweis des Farkas-Lemmas. Der Beweis beruht auf dem auch in anderem Zusammenhang wichtigen Trennungssatz von Hahn-Banach.

Satz 3.3.3 (Trennungssatz von Hahn-Banach)
Sei $M \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex. Weiter sei $y \in \mathbb{R}^n \setminus M$. Dann existieren $w \in \mathbb{R}^n$ und $\alpha \in \mathbb{R}$ mit

$$w^T y > \alpha, \quad w^T x \le \alpha \quad \forall\, x \in M.$$

Ist M ein Kegel, so kann $\alpha = 0$ gewählt werden.

Beweis: Noch Voraussetzung gibt es einen Punkt $x_0 \in M$. Die Menge

$$M_0 := \{x \in M \,:\, \|x - y\| \le \|x_0 - y\|\}$$

enthält x_0 und ist kompakt. Wegen der Stetigkeit von $\|\cdot\|$ existiert daher $\bar{x} \in M_0 \subset M$ mit

$$\|\bar{x} - y\| = \inf_{x \in M_0} \|x - y\| = \inf_{x \in M} \|x - y\|.$$

Setzen wir $w = y - \bar{x}$ und $\alpha = w^T \bar{x}$, dann gilt wegen $y \notin M$

$$0 < \|w\|^2 = w^T(y - \bar{x}) = w^T y - \alpha, \text{ also } w^T y > \alpha.$$

Sei nun $x \in M$ beliebig. Wegen der Konvexität von M gilt mit

$$\phi : t \in [0,1] \mapsto \|(tx + (1-t)\bar{x}) - y\|^2$$

nach Wahl von $\bar{x}$

$$\phi(0) = \|\bar{x} - y\| \le \min_{t \in [0,1]} \phi(t)$$

und somit

$$0 \le \phi'(0) = 2(y - \bar{x})^T(\bar{x} - x) = 2(\alpha - w^T x), \text{ also } w^T x \le \alpha.$$

Ist zudem M ein Kegel, dann gilt $x \in M \implies \lambda x \in M$ für alle $\lambda \ge 0$. Insbesondere ist $0 = 0x_0 \in M$ und somit $0 = w^T 0 \le \alpha$, also $\alpha \ge 0$. Gäbe es ein $x \in M$ mit $w^T x > 0$, dann gäbe es ein $\lambda > 0$ mit $w^T(\lambda x) > \alpha$ im Widerspruch zu $w^T \lambda x \le \alpha$ wegen $\lambda x \in M$. Also gilt automatisch $w^T x \le 0$ für alle $x \in M$ und wir können $\alpha = 0$ o. E. wählen. $\square$

Lemma 3.3.4 *Seien $A \in \mathbb{R}^{n,m}$ und $B \in \mathbb{R}^{n,p}$. Dann ist die Menge*

$$C = \{x \in \mathbb{R}^n \,:\, x = Au + Bv, \ u \in \mathbb{R}^m, \ v \in \mathbb{R}^p, \ u \ge 0\}$$

ein abgeschlossener, konvexer Kegel.

Beweis: Wegen $C = \{x : x = (A, B, -B)u,\ u \in \mathbb{R}^{n+2p},\ u \geq 0\}$ reicht es, die Behauptung für $C = \{x \in \mathbb{R}^n : x = Au,\ u \in \mathbb{R}^m,\ u \geq 0\}$ zu zeigen. Offensichtlich ist C ein konvexer Kegel und es bleibt nur die Abgeschlossenheit zu zeigen. Wir beweisen dies durch vollständige Induktion nach m. Sei $A_m = (a_1, \dots, a_m)$ und

$$C_m = \{x \in \mathbb{R}^n : x = A_m u,\ u \in \mathbb{R}^m,\ u \geq 0\}.$$

$m = 1$: Dann ist $C_1 = \{ta_1 : t \geq 0\}$ offensichtlich abgeschlossen.

Induktionsannahme: Jeder von höchstens $m-1$ Vektoren aufgespannte Kegel ist abgeschlossen.

$m-1 \to m$: 1. Fall: A_m hat linear unabhängige Spalten. Sei nun $(x_k) \subset C_m$ eine beliebige konvergente Folge. Dann gibt es $(u_k) \subset [0, \infty)^m$ mit

$$x_k = A_m u_k \to x$$

mit einem $x \in \mathbb{R}^n$. Da A_m vollen Spaltenrang hat, sind die u_k eindeutig bestimmt durch $u_k = (A_m^T A_m)^{-1} A_m^T x_k$. Also gilt $u_k \to (A_m^T A_m)^{-1} A_m^T x =: u \geq 0$ und daher $x_k \to A_m u \in C_m$.

2. Fall: A_m hat nichttrivialen Kern. Wir zeigen

$$C_m = \bigcup_{i=1}^{m} \big\{x : x = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_m)u,\ u \in \mathbb{R}^{m-1},\ u \geq 0\big\} =: \tilde{C}.$$

Nach Induktionsvorausseztung ist dann C_m abgeschlossen als endliche Vereinigung abgeschlossener Mengen. Natürlich gilt $\tilde{C} \subset C_m$. Zum Nachweis von $C_m \subset \tilde{C}$ sei $x \in C_m$ beliebig. Dann gilt

$$x = A_m u,\ \ a \geq 0.$$

Wir finden $w \in \operatorname{Ker}(A_m) \setminus \{0\}$ mit $w_i < 0$ für mindestens ein i. Setze

$$\tilde{\alpha} = \max\{\alpha \geq 0 : u + \alpha w \geq 0\}.$$

Natürlich existiert $\tilde{\alpha}$ und mit $\tilde{u} = u + \bar{\alpha} w$ gilt $\tilde{u} \geq 0$, $(\tilde{u})_i = 0$ für mindestens ein i. Dies ergibt

$$x = A_m \tilde{u} \in \tilde{C}.$$

□

Wir können nun das Farkas-Lemma beweisen.

Beweis: (von Lemma 3.2.8)
ii) $\Longrightarrow$ i): Es sei $b = Au + Bv$, $u \geq 0$. Dann gilt für jedes s mit $A^T s \leq 0$ und $B^T s = 0$

$$b^T s = u^T A^T s + v^T B^T s = u^T (A^T s) \leq 0.$$

i) $\Longrightarrow$ ii): Der Kegel

$$C = \{x \,:\, x = Au + Bv,\;\; u \in \mathbb{R}^m,\;\; u \geq 0,\;\; v \in \mathbb{R}^p\}$$

ist offensichtlich konvex und zudem abgeschlossen nach Lemma 3.3.4. Gilt ii) nicht, dann ist $b \notin C$ und wir finden nach dem Trennungssatz 3.3.3 ein $w \in \mathbb{R}^n$ und $\alpha = 0$ (C ist eine Kegel!) mit

$$w^T b > 0,\;\; w^T x \leq 0 \qquad \forall\, x \in C.$$

Dies ergibt

$$w^T Au + w^T Bv \leq 0 \quad \forall\, u \in \mathbb{R}^m,\;\; u \geq 0,\;\; v \in \mathbb{R}^p.$$

Hieraus folgt $w^T A \leq 0$ und $w^T B = 0$ und daher

$$A^T w \leq 0, \quad B^T w = 0, \quad b^T w > 0.$$

Also gilt i) nicht. $\square$

3.4 Penalty-Verfahren

Penalty-Verfahren stellen einen klassischen Ansatz zur Lösung von (NLP) dar. Sie werden eingehend von Fiacco und McCormick in [FMC68] behandelt. Die Idee hinter Penalty-Verfahren besteht darin, (NLP) durch eine Folge unrestringierter Probleme

$$(\text{NLP}_\rho) \qquad \min_{x \in \mathbb{R}^n} P_\rho(x)$$

zu approximieren, bei denen jeweils eine Penalty-Funktion P_ρ minimiert wird. Die Penalty-Funktion

$$P_\rho(x) = f(x) + \rho S(x).$$

entsteht durch Hinzunahme eines Strafterms zur Zielfunktion, der eine Verletzung der Nebenbedingungen von (NLP) durch große Funktionswerte bestraft. Das Gewicht der Strafterme wird durch einen *Penalty-Parameter* ρ sukzessive erhöht. Da die Penalty-Probleme für wachsenden Penalty-Parameter schwieriger zu lösen sind, verwendet man eine monoton wachsende Folge (ρ_k) von Penalty-Parametern und verwendet die Lösung x_k jedes Teilproblems als Startpunkt für das nächste Teilproblem.

3.4.1 Das quadratische Penalty-Verfahren

Beim quadratischen Penalty-Verfahren verwendet man die sog. *quadratische Penalty-Funktion*

$$P_\rho(x) := f(x) + \rho S(x)$$

mit einem Penalty-Parameter $\rho > 0$ und dem Penalty-Term

$$S(x) := \frac{1}{2}\sum_{i=1}^{m} \max\{0, c_i(x)\}^2 + \frac{1}{2}\sum_{i=1}^{p} h_i(x)^2 = \frac{1}{2}(\|(c(x))_+\|^2 + \|h(x)\|^2).$$

Hierbei bezeichnen wir für $v \in \mathbb{R}^m$ mit $(v)_+$ den Vektor

$$(v)_+ = (\max\{0, v_i\})_{1\le i\le m}.$$

Die Strafterme für Ungleichungen $t \le 0$ basieren also auf der Straffunktion $p_u(t) = (t)_+^2$ und die Strafterme für Gleichungen $t = 0$ auf der Straffunktion $p_g(t) = t^2$. Beide Strafterme verschwinden auf dem zulässigen Bereich und sind stetig differenzierbar. Wir erhalten

$$\nabla P_\rho(x) := \nabla f(x) + \rho\sum_{i=1}^{m}(c_i(x))_+\nabla c_i(x) + \rho\sum_{i=1}^{p} h_i(x)\nabla h_i(x).$$

Insbesondere gilt

$$P_\rho(x) = f(x), \quad \nabla P_\rho(x) = \nabla f(x) \quad \forall\, x \in Z.$$

Die Glattheit des Strafterms $S(x)$ hat ihren Preis: Die Ableitung $\nabla S(x)$ verschwindet auf dem zulässigen Bereich, daher wächst der Strafterm bei Verlassen des zulässigen Bereichs zunächst nur langsam. Ein Minimum x_ρ von P_ρ ist somit nur dann in Z, wenn $\nabla f(x_\rho) = 0$ ist. In der Regel sind also die Minima von P_ρ nicht zulässig für (NLP).

Wir analysieren nun folgendes Penalty-Verfahren:

Algorithmus 16 Quadratisches Penalty-Verfahren

Wähle einen Penalty-Parameter $\rho_0 > 0$ (z.B. $\rho_0 = 1$).

Für $k = 0, 1, \dots$:

1. *Bestimme die globale Lösung x_k des Penalty-Problems*

   (NLP$_{\rho_k}$) $\qquad \min_{x\in\mathbb{R}^n} P_{\rho_k}(x)$

   Für $k \ge 1$ wird in der Regel x_{k-1} als Startpunkt verwendet.

2. *Falls $x_k \in Z$: STOP mit Ergebnis x_k.*

3. *Wähle $\rho_{k+1} > \rho_k$.*

Bemerkung: Wegen $f(x) = P_{\rho_k}(x)$ für alle $x \in Z$ ist im Falle $x_k \in Z$ offensichtlich x_k globales Minimum von (NLP). Dies rechtfertigt den Abbruch in Schritt 2. $\square$

Globale Konvergenz

Satz 3.4.1 *Es seien $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^n \to \mathbb{R}^p$ stetig und der zulässige Bereich Z sei nicht leer. Es gelte $\rho_k \nearrow \infty$ für $k \to \infty$. Algorithmus 16 erzeuge eine unendliche Folge (x_k) (wir nehmen an, dass (x_k) existiert). Dann gilt:*

i) Die Folge $(P_{\rho_k}(x_k))$ wächst monoton.

ii) Die Folge $(\|(c(x_k))_+\|^2 + \|h(x_k)\|^2)$ fällt monoton.

iii) Die Folge $(f(x_k))$ wächst monoton.

iv) Es gilt $\lim\limits_{k\to\infty} (c(x_k))_+ = 0$, $\lim\limits_{k\to\infty} h(x_k) = 0$.

v) Jeder Häufungspunkt von (x_k) ist eine globale Lösung von (NLP).

Beweis: zu i): Die Optimalität von x_k ergibt mit $\rho_k < \rho_{k+1}$ und $S(x) \geq 0$

$$P_{\rho_k}(x_k) \leq P_{\rho_k}(x_{k+1}) = f(x_{k+1}) + \rho_k S(x_{k+1}) \leq f(x_{k+1}) + \rho_{k+1} S(x_{k+1}) = P_{\rho_{k+1}}(x_{k+1}).$$

zu ii): Addition der Ungleichungen $P_{\rho_k}(x_k) \leq P_{\rho_k}(x_{k+1})$ und $P_{\rho_{k+1}}(x_{k+1}) \leq P_{\rho_{k+1}}(x_k)$ ergibt

$$\rho_k S(x_k) + \rho_{k+1} S(x_{k+1}) \leq \rho_k S(x_{k+1}) + \rho_{k+1} S(x_k)$$

und wegen $\rho_k < \rho_{k+1}$ folgt $S(x_{k+1}) \leq S(x_k)$.

zu iii): Mit ii) gilt

$$0 \leq P_{\rho_k}(x_{k+1}) - P_{\rho_k}(x_k) = f(x_{k+1}) - f(x_k) + \rho_k (S(x_{k+1}) - S(x_k)) \leq f(x_{k+1}) - f(x_k).$$

zu iv): Wir zeigen $S(x_k) \to 0$. Wegen $Z \neq \emptyset$ existiert $x \in Z$ und wir haben

$$P_{\rho_k}(x_k) \leq P_{\rho_k}(x) = f(x).$$

Mit iii) folgt nun

$$f(x) \geq P_{\rho_k}(x_k) = f(x_k) + \rho_k S(x_k) \geq f(x_0) + \rho_k S(x_k)$$

und somit $S(x_k) \to 0$ wegen $\rho_k \to \infty$.

zu v): Sei $\bar{x}$ ein Häufungspunkt von (x_k) und sei $(x_k)_{k\in K}$ eine Teilfolge mit $(x_k)_{k\in K} \to \bar{x}$. Nach iv) gilt $S(x_k) \to S(\bar{x}) = 0$, also $\bar{x} \in Z$. Wir erhalten für jedes $x \in Z$

$$f(\bar{x}) = \lim_{k\in K\to\infty} f(x_k) \leq \lim_{k\in K\to\infty} P_{\rho_k}(x_k) \leq \lim_{k\in K\to\infty} P_{\rho_k}(x) = f(x).$$

□

Für konvexe Probleme (NLP) sind auch die Penalty-Probleme konvex:

Satz 3.4.2 *Ist (NLP) konvex, also f, c_i konvex, h affin linear, dann ist P_ρ für alle $\rho > 0$ konvex. Sind f, c stetig differenzierbar, dann ist insbesondere jeder stationäre Punkt von P_ρ ein globales Minimum (NLP$_\rho$).*

Ist f streng konvex, so ist P_ρ für alle $\rho > 0$ streng konvex. Sind f, c stetig differenzierbar, dann hat inbesondere P_ρ höchstens einen stationären Punkt und dieser ist das eindeutige globale Minimum von (NLP$_\rho$).

Beweis: Mit $p_u(t) = ((t)_+)^2$ und $p_g(t) = t^2$ gilt $S(x) = \sum_{i=1}^m \frac{1}{2} p_u(c_i(x)) + \sum_{i=1}^p \frac{1}{2} p_g(h_i(x))$. Nun ist $p_u(t) = ((t)_+)^2$ konvex und monoton wachsend, also ist $p_u \circ c_i$ konvex nach Übung. Weiter ist p_g konvex und h_i affin linear, also

$$\begin{aligned} p_g(h_i((1-t)x + ty)) &= p_g((1-t)h_i(x) + th_i(y)) \\ &\le (1-t)p_g(h_i(x)) + tp_g(h_i(y)) \quad \forall\, x, y \in \mathbb{R}^n, \quad \forall\, t \in [0,1]. \end{aligned}$$

Damit ist auch P_ρ als positive Linearkombination konvexer Funktionen konvex. Ist f streng konvex, dann ist natürlich auch P_ρ streng konvex.

Ist nun $\nabla P_\rho(\bar{x}) = 0$, dann liefert die Konvexität nach Satz 2.2.3

$$P_\rho(x) - P_\rho(\bar{x}) \ge \nabla P_\rho(\bar{x})^T (x - \bar{x}) = 0 \quad \forall\, x \in \mathbb{R}^n.$$

Damit ist $\bar{x}$ globales Minimum. Ist P_ρ streng konvex, dann gilt nach Satz 2.2.3

$$P_\rho(x) - P_\rho(\bar{x}) > \nabla P_\rho(\bar{x})^T (x - \bar{x}) = 0 \quad \forall\, x \in \mathbb{R}^n \setminus \{\bar{x}\}$$

und daher ist $\bar{x}$ das einzige globale Minimum. □

Approximation der Lagrange-Multiplikatoren

Das globale Minimum x_k von (NLP$_{\rho_k}$) ist ein stationärer Punkt, also

$$\begin{aligned} \nabla P_{\rho_k}(x_k) &= \nabla f(x_k) + \sum_{i=1}^m \rho_k \max\{0, c_i(x_k)\} \nabla c_i(x_k) + \sum_{i=1}^p \rho_k h_i(x_k) \nabla h_i(x_k) \\ &= \nabla f(x_k) + \nabla c(x_k)\lambda_k + \nabla h(x_k)\mu_k = 0 \end{aligned} \tag{3.8}$$

mit

$$\lambda_k = \rho_k (c(x_k))_+, \quad \mu_k = \rho_k h(x_k). \tag{3.9}$$

Wir untersuchen nun das Konvergenzverhalten von λ_k, μ_k.

Satz 3.4.3 *Es seien f, c, h stetig differenzierbar. Algorithmus 16 erzeuge eine unendliche Folge (x_k). Dann gilt mit λ_k, μ_k gemäß* (3.9)*:*

i) Ist $(\bar{x}, \bar{\lambda}, \bar{\mu})$ ein Häufungspunkt von (x_k, λ_k, μ_k), dann ist $\bar{x}$ globales Minimum von (NLP) und erfüllt mit $\bar{\lambda}, \bar{\mu}$ die KKT-Bedingungen.

ii) Ist $\bar{x}$ ein Häufungspunkt (x_k) in dem (LICQ) gilt, dann ist $\bar{x}$ globales Minimum von (NLP) und erfüllt mit eindeutigen Lagrange-Multiplikatoren $\bar{\lambda}, \bar{\mu}$ die KKT-Bedingungen. Zudem gilt für jede Teilfolge $(x_k)_{k\in K} \to \bar{x}$ auch $(x_k, \lambda_k, \mu_k)_{k\in K} \to (\bar{x}, \bar{\lambda}, \bar{\mu})$.

Beweis: zu i): $\bar{x}$ ist nach Satz 3.4.1 globales Minimum von (NLP). Sei $(x_k, \lambda_k, \mu_k)_{k\in K}$ eine gegen den Häufungspunkt $(\bar{x}, \bar{\lambda}, \bar{\mu})$ konvergente Teilfolge. Nun gilt (3.8), also

$$0 = \nabla f(x_k) + \nabla c(x_k)\lambda_k + \nabla h(x_k)\mu_k \to \nabla f(\bar{x}) + \nabla c(\bar{x})\bar{\lambda} + \nabla h(\bar{x})\bar{\mu}.$$

Dies liefert KKT, 2). Weiter ist $\lambda_k \geq 0$ nach (3.9) und daher $\bar{\lambda} \geq 0$. Für $i \in \mathcal{I}(\bar{x})$ gilt $c_{\mathcal{I}(\bar{x})}(x_k) < 0$ und somit $(\lambda_k)_{\mathcal{I}(\bar{x})} = 0$ für alle $k \in K$ groß genug. Dies zeigt $\bar{\lambda}_{\mathcal{I}(\bar{x})} = 0$ und somit gilt auch KKT, 3).

Gilt zudem (LICQ) für $\bar{x}$, dann gilt wieder $(\lambda_k)_{\mathcal{I}(\bar{x})} = 0$ für alle $l \leq k \in K$, l groß genug, und somit

$$-\nabla f(x_k) = (\nabla c_{\mathcal{A}(\bar{x})}(x_k), \nabla h(x_k)) \begin{pmatrix} (\lambda_k)_{\mathcal{A}(\bar{x})} \\ \mu_k \end{pmatrix} =: M_k^T \begin{pmatrix} (\lambda_k)_{\mathcal{A}(\bar{x})} \\ \mu_k \end{pmatrix} \quad \forall\, l \leq k \in K.$$

Für $k \in K \to \infty$ konvergiert M_k gegen eine Matrix $\bar{M}$ mit linear unabhängigen Zeilen. Daher hat $M_k M_k^T$ für $l \leq k \in K$, l genügend groß, eine beschränkte Inverse und es gilt

$$\begin{aligned} \begin{pmatrix} (\lambda_k)_{\mathcal{A}(\bar{x})} \\ \mu_k \end{pmatrix} &= -(M_k M_k^T)^{-1} M_k \nabla f(x_k) \\ &\to -(\bar{M}\bar{M}^T)^{-1}\bar{M}\nabla f(\bar{x}) = \begin{pmatrix} (\bar{\lambda})_{\mathcal{A}(\bar{x})} \\ \bar{\mu} \end{pmatrix} \quad \text{für } k \in K \to \infty. \end{aligned}$$

□

Wir haben gesehen, dass die Lösungen x_k der quadratischen Penalty-Probleme nur in Z liegen, wenn $\nabla f(x_k) = 0$ gilt, was in der Regel nicht der Fall ist. Im allgemeinen terminiert also das quadratische Penalty-Verfahren nicht endlich und die Penalty-Parameter ρ_k gehen gegen ∞. Penalty-Verfahren sind geeignet, um eine Näherungslösung für (NLP) zu bestimmen, die lokalen Konvergenzeigenschaften sind jedoch meist unbefriedigend, da die Penalty-Probleme für großes ρ_k schwer zu lösen sind.

3.4.2 Exakte Penalty-Verfahren

Verwendet man geeignete nichtglatte Penalty-Funktionen, so kann man bereits für endliche Penalty-Parameter eine Lösung von (NLP) erhalten.

Definition 3.4.4 *Sei $\bar{x}$ eine lokale Lösung von (NLP). Eine Penalty-Funktion $P : \mathbb{R}^n \to \mathbb{R}$ heißt* exakt *im Punkt $\bar{x}$, wenn $\bar{x}$ ein lokales Minimum von P ist.*

Unter geeigneten Voraussetzungen ist die folgende ℓ_1-Penalty-Funktionen exakt:

$$P_{\ell_1,\rho}(x) = f(x) + \rho \sum_{i=1}^{m} \max\{0, c_i(x)\} + \rho \sum_{i=1}^{p} |h_i(x)| = f(x) + \rho(\|(c(x))_+\|_1 + \|h(x)\|_1).$$

Wir zeigen die Exaktheit für den konvexen Fall. Im allgemeinen Fall gilt mit derselben Unterschranke für ρ Exaktheit in einem Punkt, der die hinreichende Bedingung zweiter Ordnung erfüllt.

Satz 3.4.5 *Seien $f, c_i : \mathbb{R}^n \to \mathbb{R}$ konvex und stetig differenzierbar und sei $h : \mathbb{R}^n \to \mathbb{R}^p$ affin linear. Ist $\bar{x}$ ein KKT-Punkt mit Lagrange-Multiplikatoren $\bar{\lambda}, \bar{\mu}$, dann ist $\bar{x}$ globales Minimum von (NLP) und für alle*

$$\rho \geq \max\left\{\bar{\lambda}_1, \ldots, \bar{\lambda}_m, |\bar{\mu}_1|, \ldots, |\bar{\mu}_p|\right\}$$

ist $\bar{x}$ globales Minimum von $P_{\ell_1,\rho}$.

Beweis: Wegen $\bar{\lambda} \geq 0$ ist $L(\cdot, \bar{\lambda}, \bar{\mu})$ konvex und KKT, 2) liefert

$$\nabla_x L(\bar{x}, \bar{\lambda}, \bar{\mu}) = 0.$$

Daher ist $\bar{x}$ ein globales Minimum von $L(\cdot, \bar{\lambda}, \bar{\mu})$. Nun ergibt KKT, 3) für alle $x \in \mathbb{R}^n$

$$\begin{aligned}
P_{\ell_1,\rho}(\bar{x}) &= f(\bar{x}) = L(\bar{x}, \bar{\lambda}, \bar{\mu}) \leq L(x, \bar{\lambda}, \bar{\mu}) = f(x) + \bar{\lambda}^T c(x) + \bar{\mu}^T h(x) \\
&\leq f(x) + \bar{\lambda}^T (c(x))_+ + \bar{\mu}^T h(x) \leq f(x) + \|\bar{\lambda}\|_\infty \|(c(x))_+\|_1 + \|\bar{\mu}\|_\infty \|h(x)\|_1 \\
&\leq f(x) + \rho\|(c(x))_+\|_1 + \rho\|h(x)\|_1 = P_{\ell_1,\rho}(x).
\end{aligned}$$

□

3.5 Innere-Punkte-Verfahren

Wir betrachten das konvexe nichtlineare Optimierungsproblem

(NLPU) $$\min_{x \in \mathbb{R}^n} f(x) \quad \text{u.d. Nebenbedingung} \quad c(x) \leq 0.$$

mit konvexen, stetig differenzierbaren Funktionen $f, c_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leq i \leq m$. Wir nehmen an, dass der zulässige Bereich Z ein nicht leeres sogenanntes *striktes Inneres*

(3.10) $$Z_0 = \{x \in \mathbb{R}^n \; : \; c_i(x) < 0, \;\; i = 1, \ldots, m\}$$

besitzt.

Während beim Penalty-Verfahren Unzulässigkeit durch einen Penalty-Term bestraft wird, verwendet man bei Innere-Punkte-Verfahren einen Barriere-Term, um eine Annäherung von Innen an den Rand von Z zu verhindern und sichert so die Zulässigkeit der Iterierten.

Innere-Punkt Verfahren lösen anstelle des ungleichungsrestringierten Problems (NLPU) eine Folge von "leichteren" Problemen. Die Iteration wird hierbei durch Barriere-Terme im sogenannten strikten Inneren Z_0 von Z gehalten. Außer in pathologischen Fällen stimmt Z_0 mit dem offenen Kern $\mathring{Z}$ von Z überein, siehe Übungsaufgabe.

Übungsaufgabe: Es seien $c_i : \mathbb{R}^n \to \mathbb{R}$ konvex und differenzierbar. Zeigen Sie: Ist $Z_0 = \{x \in \mathbb{R}^n \,:\, c(x) < 0\}$ nichtleer, so gilt $Z_0 = \mathring{Z}$.

Bei Barriereverfahren wird anstelle von (NLPU) eine Folge von Barriereproblemen der Form gelöst

(BP$_\tau$) $\quad \min \ B_\tau(x) := f(x) + \tau\, p(x) \quad$ unter der Nebenbed. $x \in Z_0$.

Hierbei ist $p : Z_0 \to \mathbb{R}$ eine *Barrierefunktion*:

Definition 3.5.1 (Barrierefunktion)
Sei

$$Z := \left\{x \in \mathbb{R} \,:\, c(x) = (c_1(x), \ldots, c_m(x))^T \le 0\right\}$$

mit $Z_0 \neq \emptyset$, d.h. nichtleerem striktem Inneren. Eine Funktion $p : Z_0 \to \mathbb{R}$ heißt Barrierefunktion *für Z, falls gilt:*

a) p hat auf Z_0 dieselbe Differenzierbarkeitsordnung wie $c = (c_1, \ldots, c_m)$.

b) Es gilt

$$(x_k) \subset Z_0, \quad \lim_{k\to\infty} x_k = x \in \partial Z \implies p(x_k) \to \infty.$$

Man verwendet in der Regel Barrierefunktionen der Form

$$p(x) = \sum_{i=1}^{m} b(-c_i(x)), \tag{3.11}$$

$b : (0,\infty) \to \mathbb{R}$ glatt, konvex und monoton fallend mit $\lim_{t\to 0^+} b(t) = \infty$.

Die weitaus wichtigste Wahl ist $b(t) = -\ln(t)$ und liefert die von Frisch [Fr55] erstmals verwendete log-Barrierefunktion

$$p(x) = -\sum_{i=1}^{m} \ln(-c_i(x)) \quad \text{(log-Barrierefunktion).}$$

Die log-Barrierefunktion hat in der Nähe von ∂Z_0 geringere Krümmung als offensichtliche andere Wahlen wie etwa $b(t) = 1/t$.

Das klassische Barriere-Verfahren von Fiacco und McCormick [FMC68] ist von der folgenden Form:

Algorithmus 17 Klassisches Barriere-Verfahren

Wähle $\tau_0 > 0$.

Für $k = 0, 1, \ldots$*:*

1. *Berechne eine Lösung* $x(\tau_k)$ *von*

   $(\mathrm{BP}_{\tau_k}) \qquad \min \; B_{\tau_k}(x) := f(x) + \tau_k p(x)$ *unter der Nebenbed.* $x \in Z_0$.

2. *Setze* $x_k := x(\tau_k)$. *Falls* x_k *Lösung von (NLPU): STOP.*
3. *Wähle* $0 < \tau_{k+1} < \tau_k$.

Algorithmus 17 war in den 60er Jahren sehr populär, kam jedoch zunächst aus der Mode, da für kleiner werdende τ_k das Barriereproblem (BP_{τ_k}) zunehmend schwerer zu lösen ist. Erneute Forschungsaktivitäten wurden durch die richtungsweisende Arbeit [Ka84] von Karmakar ausgelöst, in der Karmakar ein Innere-Punkte-Verfahren für Lineare Optimierungsprobleme vorstellt, die im Gegensatz zum Simplexverfahren eine Lösung bei rationalen Daten in polynomialem statt exponetiellem Aufwand liefert. Der heutige Erfolg von Innere-Punkt Verfahren beruht auf den seit 1984 erzielten Fortschritten beim Verständnis der folgenden Fragen:

- Wie genau ist (BP_{τ_k}) zu lösen, bevor τ_k verkleinert wird?
  $\leadsto$ Umgebungen des *zentralen Pfades*.
- Welches Verfahren ist zur Lösung von (BP_{τ_k}) geeignet und wie stark darf τ_k verkleinert werden?
  $\leadsto$ genaue Analyse der Konvergenzeigenschaften des Newton-Verfahrens in der Nähe des zentralen Pfades.
- Welche Modifikationen des Newton-Verfahrens sind besonders effizient und numerisch stabil?
  $\leadsto$ primal-duale Verfahren.
- Kann τ_k schneller verkleinert werden, wenn x_k geeignet korrigiert wird?
  $\leadsto$ Prädiktor-Korrektor Verfahren.

Es zeigt sich, dass bei einer geeigneten Implementierung von Algorithmus 17 mit

$$p(x) = -\sum_{i=1}^{m} \ln(-c_i(x))$$

die Lösung von linearen Optimierungsproblemen mit rationalen Daten – im Gegensatz zum Simplexverfahren – in polynomialer Komplexität gefunden werden kann.

Wir können auf diese Fragen hier nicht detailliert eingehen. Einzelheiten werden in der gelegentlich stattfindenden Vorlesung *Innere-Punkte-Verfahren der konvexen Optimierung* behandelt.

3.5.1 Konvergenzeigenschaften des Innere-Punkte-Verfahrens

Ähnlich wie beim Penalty-Verfahren kann man folgenden Satz zeigen:

Satz 3.5.2 *Es seien $f : \mathbb{R}^n \to \mathbb{R}$, $c_i : \mathbb{R}^n \to \mathbb{R}$, $1 \le i \le m$, stetig differenzierbar und konvex. Der zulässige Bereich Z sei kompakt und sein striktes Inneres Z_0 sei nicht leer.*

Algorithmus 17 verwende einen Barriere-Term der Form (3.11) *sowie eine Folge (τ_k) mit $\tau_k \searrow 0$ für $k \to \infty$. Dann gilt:*

i) Das Problem (NLPU) hat eine Lösung $\bar{x}$.

ii) B_τ ist konvex auf Z_0 für alle $\tau > 0$.

iii) (BP_{τ_k}) *besitzt stets eine Lösung $x_k = x(\tau_k) \in Z_0$.*

iv) Für jede Lösungsfolge (x_k) von (BP_{τ_k}) *gilt*

$$f(x_{k+1}) \le f(x_k) \quad \textit{und} \quad \lim_{k\to\infty} f(x_k) = f(\bar{x}).$$

v) Jeder Häufungspunkt von (x_k) ist eine optimale Lösung von (NLPU).

Bemerkung: Die Aussagen i) und iii)–v) gelten auch ohne die Konvexität von f, c_i, falls $\overline{Z_0} = Z$. Im konvexen Fall ist dies sichergestellt und zudem jede lokale Lösung von (BP_{τ_k}) globale Lösung.

Beweis: Der Beweis ist ähnlich wie beim Penalty-Verfahren. Hier der Beweis aus der Übung:

zu i): Der zulässige Bereich Z ist nach Voraussetzung kompakt. Die stetige Funktion f nimmt also auf Z ein Minimum $\bar{x}$ an.

zu ii): Da $c_i : \mathbb{R}^n \to \mathbb{R}$ konvex sind, sind offensichtlich Z und Z_0 konvex. Die Funktion $t \in (-\infty, 0) \mapsto b(-t)$ ist nach Voraussetzung konvex und monoton wachsend. Die Konvexität von $c_i : Z_0 \to (-\infty, 0)$ liefert daher die Konvexität von $x \in Z_0 \mapsto b(-c_i(x))$. Für alle $\tau > 0$ ist daher B_τ auf Z_0 eine Summe konvexer Funktionen, also konvex.

zu iii): Sei $\tau_k > 0$ beliebig. Wegen $Z_0 \neq \emptyset$ finden wir ein $\hat{x} \in Z_0$. Dann ist die Niveaumenge

$$N_{\tau_k} := N_{B_{\tau_k}}(\hat{x}) = \{x \in Z_0 \; : \; B_{\tau_k}(x) \le B_{\tau_k}(\hat{x})\}$$

kompakt: Wegen $N_{\tau_k} \subset Z_0 \subset Z$ ist N_{τ_k} beschränkt und wegen $\hat{x} \in N_{\tau_k}$ nicht leer. Zudem ist N_{τ_k} abgeschlossen: für jede Cauchyfolge $(y_j) \subset N_{\tau_k} \subset Z_0$ gilt $y = \lim_{j\to\infty} y_j \in Z_0$. Denn die Folgen $f(y_j)$ sowie $b(-c_i(y_j))$, $i = 1, \dots, m$, sind nach unten beschränkt, und die Annahme $c_i(y_j) \searrow 0$ für ein i und $j \to \infty$ ergäbe daher den Widerspruch $B_{\tau_k}(\hat{x}) \geq B_{\tau_k}(y_j) \to \infty$. Also gilt $y \in Z_0$ und $B_{\tau_k}(y) \leq B_{r_k}(\hat{x})$ folgt aus Stetigkeitsgründen.

Die stetige Funktion $B_{\tau_k} : Z_0 \to \mathbb{R}$ nimmt auf der kompakten Niveaumenge $N_{\tau_k} \subset Z_0$ ein Minimum x_k an, und dies ist natürlich Minimum auf ganz Z_0.

zu iv): Mit $p(x) = \sum_{i=1}^m b(-c_i(x))$ gilt $B_{\tau_k}(x) = f(x) + \tau_k p(x)$. Wegen der Minimumseigenschaft von x_k und x_{k+1} gilt

$$\begin{aligned} B_{\tau_k}(x_k) = f(x_k) + \tau_k p(x_k) &\leq B_{\tau_k}(x_{k+1}) = f(x_{k+1}) + \tau_k p(x_{k+1}) \\ B_{\tau_{k+1}}(x_k) = f(x_k) + \tau_{k+1} p(x_k) &\geq B_{\tau_{k+1}}(x_{k+1}) = f(x_{k+1}) + \tau_{k+1} p(x_{k+1}). \end{aligned}$$

Subtraktion der zweiten von der ersten Ungleichung liefert

$$(\tau_k - \tau_{k+1}) p(x_k) \leq (\tau_k - \tau_{k+1}) p(x_{k+1}).$$

Wegen $\tau_k - \tau_{k+1} > 0$ folgt $p(x_k) \leq p(x_{k+1})$. Einsetzen in die zweite Ungleichung ergibt

$$f(x_k) + \tau_{k+1} p(x_k) \geq f(x_{k+1}) + \tau_{k+1} p(x_{k+1}) \geq f(x_{k+1}) + \tau_{k+1} p(x_k),$$

also $f(x_k) \geq f(x_{k+1})$.

Sei $\bar{x}$ Lösung von (NLPU) und $\tilde{x} \in Z_0$. Wegen der Konvexität von c_i gilt $c_i(y) < 0$ für alle $y \neq \bar{x}$ auf der Strecke $[\bar{x}, \tilde{x}]$. Zu jedem $\varepsilon > 0$ finden wir also $y \in Z_0$ mit $f(y) \leq f(\bar{x}) + \varepsilon$. Nun gilt wegen $p(x_k) \leq p(x_{k+1})$

$$f(x_k) + \tau_k p(x_1) \leq f(x_k) + \tau_k p(x_k) \leq f(y) + \tau_k p(y)$$

und damit

$$f(\bar{x}) \leq f(x_k) \leq f(y) + \tau_k (p(y) - p(x_1)) \to f(y) \leq f(\bar{x}) + \varepsilon.$$

Dies zeigt $f(x_k) \to f(\bar{x})$.

zu v): Sei $\tilde{x}$ Häufungspunkt von (x_k). Wegen der Stetigkeit von c_i ist $\tilde{x} \in Z$ klar. Nach iv) gilt $f(x_k) \to f(\bar{x})$ und somit folgt $f(\tilde{x}) = f(\bar{x})$ aus Stetigkeitsgründen, $\tilde{x}$ ist also Lösung von (NLPU). □

Die Berechnung von Näherungen der Lagrange-Multiplikatoren erfolgt ähnlich wie bei Penalty-Verfahren: Im Minimum x_k von B_{τ_k} gilt

$$0 = \nabla B_{\tau_k}(x_k) = \nabla f(x_k) + \sum_{i=1}^m (-\tau_k b'(-c_i(x_k))) \nabla c_i(x_k) = \nabla f(x_k) + \nabla c(x_k) \lambda_k. \tag{3.12}$$

mit

$$(\lambda_k)_i = -\tau_k b'(-c_i(x_k)). \tag{3.13}$$

Wir untersuchen nun das Konvergenzverhalten von λ_k.

Satz 3.5.3 *Unter den Voraussetzungen von Satz 3.5.2 erzeuge Algorithmus 17 eine unendliche Folge* (x_k)*. Dann gilt mit* λ_k *gemäß* (3.13)*:*

i) Ist $(\bar{x}, \bar{\lambda})$ *ein Häufungspunkt von* (x_k, λ_k)*, dann ist* $\bar{x}$ *globales Minimum von (NLPU) und erfüllt mit* $\bar{\lambda}$ *die KKT-Bedingungen.*

ii) Ist $\bar{x}$ *ein Häufungspunkt von* (x_k) *in dem (LICQ) gilt, dann ist* $\bar{x}$ *globales Minimum von (NLPU) und erfüllt mit eindeutigen Lagrange-Multiplikatoren* $\bar{\lambda}$ *die KKT-Bedingungen. Zudem gilt für jede Teilfolge* $(x_k)_{k\in K} \to \bar{x}$ *auch* $(x_k, \lambda_k)_{k\in K} \to (\bar{x}, \bar{\lambda})$*.*

Beweis: zu i): $\bar{x}$ ist nach Satz 3.5.2 globales Minimum von (NLPU). Sei $(x_k, \lambda_k)_{k\in K}$ eine gegen den Häufungspunkt $(\bar{x}, \bar{\lambda})$ konvergente Teilfolge. Nun gilt (3.12), also

$$0 = \nabla f(x_k) + \nabla c(x_k)\lambda_k \to \nabla f(\bar{x}) + \nabla c(\bar{x})\bar{\lambda}.$$

Dies liefert KKT, 2). Weiter ist $\lambda_k \geq 0$ nach (3.13) wegen $b' \leq 0$ und daher $\bar{\lambda} \geq 0$. Für $i \in \mathcal{I}(\bar{x})$ gilt $c_i(\bar{x}) < 0$ und somit

$$b'(-c_i(x_k)) \to b'(-c_i(\bar{x})), \quad \text{also} \quad (\lambda_k)_i = -\tau_k b'(-c_i(x_k)) \to 0 \quad \forall\, i \in \mathcal{I}(\bar{x}).$$

Dies zeigt $\bar{\lambda}_{\mathcal{I}(\bar{x})} = 0$ und somit ist auch KKT, 3) nachgewiesen.

Gilt zudem (LICQ) für $\bar{x}$, dann ist $\bar{\lambda}_{\mathcal{I}(\bar{x})} = 0$ und $\bar{\lambda}_{\mathcal{A}(\bar{x})} = 0$ ist die eindeutige Lösung von

$$-\nabla f(\bar{x}) = \nabla c_{\mathcal{A}(\bar{x})}(\bar{x})\bar{\lambda}_{\mathcal{A}(\bar{x})} =: \bar{M}^T \bar{\lambda}_{\mathcal{A}(\bar{x})}$$

und ergibt sich aus

$$\bar{\lambda}_{\mathcal{A}(\bar{x})} = -(\bar{M}\bar{M}^T)^{-1}\bar{M}\nabla f(\bar{x}).$$

Andererseits gilt wie oben $(\lambda_k)_{\mathcal{I}(\bar{x})} \to 0$ für $k \in K \to \infty$ und damit

$$M_k^T(\lambda_k)_{\mathcal{A}(\bar{x})} := \nabla c_{\mathcal{A}(\bar{x})}(x_k)(\lambda_k)_{\mathcal{A}(\bar{x})} = -\nabla f(x_k) - \nabla c_{\mathcal{I}(\bar{x})}(x_k)(\lambda_k)_{\mathcal{I}(\bar{x})} \to -\nabla f(\bar{x}) \quad \text{für } k \in K \to \infty.$$

Für alle $l \leq k \in K$, l groß genug, ist $M_k M_k^T$ invertierbar und wir erhalten

$$(\lambda_k)_{\mathcal{A}(\bar{x})} = -(M_k M_k^T)^{-1} M_k(\nabla f(x_k) + \nabla c_{\mathcal{I}(\bar{x})}(x_k)(\lambda_k)_{\mathcal{I}(\bar{x})}) \to -(\bar{M}\bar{M}^T)^{-1}\bar{M}\nabla f(\bar{x}) = \bar{\lambda}_{\mathcal{A}(\bar{x})}.$$

□

3.5.2 Anwendung des Newton-Verfahrens auf das Barriereproblem

Das Barriereproblem (BP_τ) wird in der Regel durch ein globalisiertes Newton-Verfahren gelöst. Für die log-Barrierefunktion gilt

$$\begin{aligned} B_\tau(x) &= f(x) - \tau \sum_{i=1}^m \ln(-c_i(x)) \\ \nabla B_\tau(x) &= \nabla f(x) - \sum_{i=1}^m \frac{\tau}{c_i(x)} \nabla c_i(x) \\ \nabla^2 B_\tau(x) &= \nabla^2 f(x) - \sum_{i=1}^m \frac{\tau}{c_i(x)} \nabla^2 c_i(x) + \sum_{i=1}^m \frac{\tau}{c_i(x)^2} \nabla c_i(x) \nabla c_i(x)^T. \end{aligned}$$

Hieraus sieht man, dass für kleines τ und x nahe beim Minimum $x(\tau)$ die Krümmung von B_τ sehr groß wird: Tatsächlich gilt

$$\bar{\lambda}_i \approx -\tau b'(-c_i(x)) = -\frac{\tau}{c_i(x)},$$

also

$$\nabla^2 B_\tau(x) \approx \nabla^2 L(x, \bar{\lambda}) + \sum_{i=1}^m \frac{\bar{\lambda}_i^2}{\tau} \nabla c_i(x) \nabla c_i(x)^T.$$

Der letzte Term wird für kleines τ groß. Genauer kann man folgendes zeigen: Gilt im Optimum $\bar{x}$ von (NLPU) strikte Komplementarität und (LICQ), dann hat $\nabla^2 B_\tau(x(\tau))$ $|\mathcal{A}(\bar{x})|$ Eigenwerte der Ordnung $O(1/\tau)$ und $n - |\mathcal{A}(\bar{x})|$ Eigenwerte der Ordnung $O(1)$. Daher ist das Newton-System

$$\nabla^2 B_\tau(x) s = -\nabla B_\tau(x)$$

für kleines τ schlecht konditioniert und der Bereich schneller Konvergenz des Newton-Verfahrens ist wegen der starken Krümmung von $B_\tau(x)$ klein.

Als Ausweg kann man das Newton-Verfahren anstelle auf die sehr nichtlineare Gleichung

$$\nabla B_\tau(x) = \nabla f(x) - \sum_{i=1}^m \frac{\tau}{c_i(x)} \nabla c_i(x) = 0$$

auf das äquivalente weniger nichtlineare primal-duale System anwenden

$$\begin{aligned} \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla c_i(x) = 0, \\ \lambda_i c_i(x) = -\tau, \;\; i = 1, \ldots, m. \end{aligned}$$

Dies ist die Basis von *primal-dualen Innere-Punkte-Verfahren*.

Bemerkung: Ein Vergelich mit den KKT-Bedingungen zeigt, dass es sich um ein *gestörte KKT-Bedingungen* handelt, bei denen die Komplementaritätsbedinung

$$\lambda_i c_i(x) = 0, \;\; i = 1, \dots, m,$$

modifiziert wird zu

$$\lambda_i c_i(x) = -\tau, \;\; i = 1, \dots, m,$$

Es gibt eine Reihe sehr effizienter Implementierungen von Innere-Punkte-Verfahren, u.a. Ipopt (Barriereverfahren fr große NLPs), LOQO (Primal-duales Innere-Punkte-Verfahren für NLPs), SeDuMi (Primal-duales Innere-Punkte-Verfahren für LPs, Second Order Cone Programme (SOCPs) und semidefinite Programme (SDPs)).

3.6 Sequential Quadratic Programming Verfahren

Sequential Quadratic Programming (SQP) Verfahren gehören neben Innere-Punkte-Verfahren zu den leistungsfähigsten Verfahren für (NLP). Sie bilden die Basis einer Reihe hervorragender Optimierungs-Codes, z.B. DONLP2, FilterSQP, KNITRO, SNOPT.

Wir betrachten zunächst das gleichungsrestringierte Problem

(NLPG) $$\min_{x \in \mathbb{R}^n} f(x) \quad \text{unter der Nebenbedingung} \quad h(x) = 0.$$

Im folgenden seien $f : \mathbb{R}^n \to \mathbb{R}$ und $h = (h_1, \dots, h_p)^T : \mathbb{R}^n \mapsto \mathbb{R}^p$, $p < n$, zumindest stetig differenzierbar.

Die grundlegende Idee von SQP-Verfahren für (NLPG) besteht darin, ausgehend von einem Punkt x_k einen neuen Punkt $x_{k+1} = x_k + s_k$ zu bestimmen, wobei s_k lokale Lösung eines quadratischen Optimierungsproblems der Form ist

SQP-Teilproblem:

(SQPG$_k$) $$\begin{aligned} &\min \quad q_k(s) := f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T H_k s \\ &\text{u. d. Nebenbed.} \quad h(x_k) + \nabla h(x_k)^T s = 0. \end{aligned}$$

Wir minimieren also ein geeignetes quadratisches Modell der Zielfunktion unter linearisierten Nebenbedingungen. Es ist nun wichtig, die Matrix H_k im quadratischen Modell q_k so zu wählen, dass eine lokale Lösung s_k des quadratischen Teilproblems eine "gute" neue Iterierte $x_{k+1} = x_k + s_k$ liefert.

3.6.1 Lagrange-Newton- und lokales SQP-Verfahren

Wir wollen zunächst untersuchen, wie die Hessematrix H_k im quadratische Modell q_k gewählt werden muss, um schnelle lokale Konvergenz zu erzielen.

Voraussetzungen:

(LSQPG1) $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^p$ seien zweimal stetig differenzierbar und $\bar{x}$ erfülle die hinreichende Bedingung 2. Ordnung (3.6) mit einem $\bar{\mu} \in \mathbb{R}^p$,

(LSQPG2) $\nabla h(\bar{x})$ habe vollen Spaltenrang, d.h. (NLPG) erfülle (LICQ) in $\bar{x}$.

Wegen (LSQPG1) erfüllt $(\bar{x}, \bar{\mu})$ die KKT-Bedingungen

$$\begin{aligned}&\nabla_x L(\bar{x}, \bar{\mu}) = 0\\ &h(\bar{x}) = 0\end{aligned}$$

und wegen (LSQPG2) ist der Multiplikator $\bar{\mu}$ in (LSQPG1) nach Satz 3.2.19 eindeutig.

Zur Wahl von H_k betrachten wir zunächst die Anwendung des Newton-Verfahrens auf das KKT-System

$$F(x, \mu) := \begin{pmatrix} \nabla_x L(x, \mu) \\ h(x) \end{pmatrix} = 0.$$

Wir werden sehen, dass dies ausgehend von (x_0, μ_0) genügend nahe bei $(\bar{x}, \bar{\mu})$ eine superlinear konvergente Folge $(x_k, \mu_k) \to (\bar{x}, \bar{\mu})$ liefert. Anschließend wählen wir H_k so, dass mit der Lösung s_k von (SQPG$_k$) gerade gilt $x_{k+1} = x_k + s_k$.

Wir betrachten also zunächst das folgende Lagrange-Newton-Verfahren:

Algorithmus 18 Lagrange-Newton-Verfahren
Wähle einen Startpunkt $(x_0, \mu_0) \in \mathbb{R}^n \times \mathbb{R}^p$.

Für $k = 0, 1, \ldots$:

1. *Falls* $F(x_k, \mu_k) := \binom{\nabla_x L(x_k, \mu_k)}{h(x_k)} = 0$*: STOP mit KKT-Punkt* x_k.
2. *Bestimme die Lösung* $\binom{s_k}{\Delta\mu_k}$ *der Newton-Gleichung*

$$F'(x_k, \mu_k) \begin{pmatrix} s_k \\ \Delta\mu_k \end{pmatrix} = -F(x_k, \mu_k),$$

$$\text{(3.14)} \qquad \textit{d.h.} \quad \begin{pmatrix} \nabla^2_{xx} L(x_k, \mu_k) & \nabla h(x_k) \\ \nabla h(x_k)^T & 0 \end{pmatrix} \begin{pmatrix} s_k \\ \Delta\mu_k \end{pmatrix} = -\begin{pmatrix} \nabla_x L(x_k, \mu_k) \\ h(x_k) \end{pmatrix}$$

und setze $(x_{k+1}, \mu_{k+1}) = (x_k, \mu_k) + (s_k, \Delta\mu_k)$.

Wir haben den folgenden lokalen Konvergenzsatz.

Satz 3.6.1 (Lokale Konvergenz des Lagrange-Newton-Verfahrens)
$\bar{x}$ *erfülle* (LSQPG1)–(LSQPG2). *Dann gibt es* $\rho > 0$*, so dass für alle* $(x_0, \mu_0) \in B_\rho(\bar{x}, \bar{\mu})$

das Lagrange-Newton-Verfahren in Algorithmus 18 Q-superlinear gegen $(\bar{x}, \bar{\mu})$ *konvergiert, also*

$$(x_k, \mu_k) \to (\bar{x}, \bar{\mu}), \quad \|(x_{k+1}, \mu_{k+1}) - (\bar{x}, \bar{\mu})\| = o(\|(x_k, \mu_k) - (\bar{x}, \bar{\mu})\|).$$

Sind $\nabla^2 f$ *und* $\nabla^2 h_i$ *Lipschitz-stetig auf* $B_\rho(\bar{x})$*, dann ist die Konvergenz Q-quadratisch.*

Beweis: Zunächst ist $F : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n \times \mathbb{R}^p$ stetig differenzierbar und es gilt $F(\bar{x}, \bar{\mu}) = 0$ wegen (LSQPG1). Wir müssen nur noch zeigen, dass $F'(\bar{x}, \bar{\mu})$ invertierbar ist, dann folgt die Behauptung aus den Konvergenzeigenschaften des Newton-Verfahrens in Satz 2.7.3.

Zum Nachweis, dass $F'(\bar{x}, \bar{\mu})$ nur trivialen Kern hat, betrachte die Gleichung $F'(\bar{x}, \bar{\mu})\binom{u}{v} = 0$. Die zweite Blockzeile ergibt

$$\nabla h(\bar{x})^T u = 0, \;\text{ also }\; u \in T_K(h; \bar{x}),$$

und somit liefert Multiplikation der ersten Blockzeile mit u^T

$$u^T \nabla_{xx}^2 L(\bar{x}, \bar{\mu}) u = 0.$$

Wegen $u \in T_K(h; \bar{x})$ folgt daraus mit (LSQPG1), siehe (3.6), $u = 0$. Damit lautet die erste Blockzeile

$$\nabla h(\bar{x}) v = 0$$

und somit $v = 0$ wegen (LSQPG2).

Sind $\nabla^2 f$ und $\nabla^2 h_i$ Lipschitz-stetig auf $B_\rho(\bar{x})$, dann ist F' offensichtlich Lipschitz-stetig in einer Umgebung von $(\bar{x}, \bar{\mu})$ und die Q-quadratische Konvergenz folgt aus Satz 2.7.3. □

Wir vergleichen nun das Lagrange-Newton-System mit dem Optimalitätssystem von (SQPG$_k$): Sei s_k eine lokale Lösung von (SQPG$_k$). Da (SQPG$_k$) lineare Nebenbedingungen hat, gilt in s_k nach Satz 3.2.13 eine Constraint Qualification. Damit gelten nach Satz 3.2.9 mit einem $\mu_{k+1} \in \mathbb{R}^p$ die KKT-Bedingungen

$$\begin{aligned} &\nabla f(x_k) + H_k s_k + \nabla h(x_k) \mu_{k+1} = 0, \\ &h(x_k) + \nabla h(x_k) s_k = 0. \end{aligned} \tag{3.15}$$

Ein Vergleich mit (3.14) ergibt, dass mit der Wahl $H_k = \nabla_{xx}^2 L(x_k, \mu_k)$ gilt

$$\begin{pmatrix} s_k \\ \mu_{k+1} \end{pmatrix} \text{ Lösung von (3.15)} \quad \Longleftrightarrow \quad \begin{pmatrix} s_k \\ \Delta\mu_k \end{pmatrix} := \begin{pmatrix} s_k \\ \mu_{k+1} - \mu_k \end{pmatrix} \text{ Lösung von (3.14).}$$

Dies zeigt die Äquivalenz des Lagrange-Newton-Verfahrens zu folgendem lokalen SQP-Verfahren:

Algorithmus 19 Lokales SQP-Verfahren
Wähle einen Startpunkt $(x_0, \mu_0) \in \mathbb{R}^n \times \mathbb{R}^p$.

Für $k = 0, 1, \ldots$:

1. *Falls* $F(x_k, \mu_k) := \binom{\nabla_x L(x_k,\mu_k)}{h(x_k)} = 0$*: STOP mit stationärem Punkt* x_k.

2. *Berechne für* $H_k = \nabla^2_{xx} L(x_k, \mu_k)$ *eine lokale Lösung* s_k *des SQP-Problems (SQPG*$_k$*) mit zugehörigem Lagrange-Multiplikator* μ_{k+1}. *Setze* $x_{k+1} = x_k + s_k$.

Bemerkung: Im Gegensatz zum Lagrange-Newton-Verfahren, das lediglich das Optimalitätssystem löst (es ist unter einer (CQ) auch in einem Maximum erfüllt!), berücksichtigt die Schrittberechnung im SQP-Verfahren, dass f minimiert werden soll. Daher bildet (SQPG$_k$) einen guten Ausgangspunkt für eine Globalisierung.

Satz 3.6.2 (Äquivalenz von lokalem SQP- und Lagrange-Newton-Verfahren)
$\bar{x}$ *erfülle* (LSQPG1)–(LSQPG2). *Dann gilt mit* $\rho > 0$ *aus Satz 3.6.1: Für alle* $(x_0, \mu_0) \in B_\rho(\bar{x}, \bar{\mu})$ *liefert das lokale SQP-Verfahren eine eindeutig definierte Folge* (x_k, μ_k). *Diese stimmt mit der vom Lagrange-Newton-Verfahren in Algorithmus 18 erzeugten Folge überein und konvergiert somit superlinear gegen* $(\bar{x}, \bar{\mu})$. *Sind* $\nabla^2 f, \nabla^2 h_i$ *Lipschitz-stetig in einer Umgebung von* $\bar{x}$*, dann ist die Konvergenz Q-quadratisch.*

Beweis: Jede lokale Lösung s_k von (SQPG$_k$) erfüllt (3.15) mit einem μ_{k+1}. Wie bereits beobachtet ist dann $\binom{s_k}{\Delta\mu_k} = \binom{s_k}{\mu_{k+1}-\mu_k}$ Lösung von (3.14) und diese ist eindeutig nach Satz 3.6.1 für $(x_k, \mu_k) \in B_\rho(\bar{x}, \bar{\mu})$. Dies gibt induktiv die Behauptung. □

Lösung des SQP-Teilproblems

Das SQP-Teilproblem (SQPG$_k$) besitzt genau dann eine Lösung s_k, wenn s_k ein KKT-Punkt ist und zudem die quadratische Funktion q_k auf dem zulässigen Bereich (einem affinen Unterraum) konvex ist. Ist also

$$s^T H_k s \geq 0 \quad \forall\, s \in \mathrm{Ker}(\nabla h(x_k)),$$

dann ist s_k genau dann Lösung von (SQPG$_k$) mit Multiplikator μ_{k+1} falls (3.15) gilt, also

$$\begin{pmatrix} H_k & \nabla h(x_k) \\ \nabla h(x_k)^T & 0 \end{pmatrix} \begin{pmatrix} s_k \\ \mu_{k+1} \end{pmatrix} = -\begin{pmatrix} \nabla f(x_k) \\ h(x_k) \end{pmatrix}.$$

Hieraus läßt sich s_k leicht bestimmen.

3.6.2 SQP-Verfahren für Probleme mit Ungleichungsrestriktionen

Wir betrachten nun SQP-Verfahren für das allgemeine Problem

(NLP) $$\min_{x\in\mathbb{R}^n} f(x) \quad \text{u.d. Nebenbedingung} \quad h(x) = 0, \quad c(x) \leq 0.$$

Sei x_k die aktuelle Iterierte. In Analogie zum gleichungsrestringierten Fall setzen wir $x_{k+1} = x_k + s_k$ mit einer lokalen Lösung für folgendes

SQP-Teilproblem:

$$(\mathrm{SQP}_k) \qquad \begin{array}{ll} \min & q_k(s) := f_k + \nabla f(x_k)^T s + \frac{1}{2} s^T H_k s \\ \text{u. d. Nebenbed.} & c(x_k) + \nabla c(x_k)^T s \le 0, \quad h(x_k) + \nabla h(x_k)^T s = 0, \end{array}$$

wobei $H_k = \nabla^2_{xx} L(x_k, \lambda_k, \mu_k)$ mit Näherungen λ_k, μ_k der Lagrange-Multiplikatoren.

Analog zu Algorithmus 19 erhalten wir:

Algorithmus 20 Lokales SQP-Verfahren für NLP
Wähle einen Startpunkt $(x_0, \lambda_0, \mu_0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$.

Für $k = 0, 1, \dots :$

1. *Falls* (x_k, λ_k, μ_k) *die KKT-Bedingungen erfüllt: STOP mit Ergebnis* x_k.
2. *Berechne für* $H_k = \nabla^2_{xx} L(x_k, \lambda_k, \mu_k)$ *die am nächsten bei* 0 *liegende lokale Lösung* s_k *des SQP-Problems* (SQP_k) *mit zugehörigen Lagrange-Multiplikatoren* λ_{k+1}, μ_{k+1}. *Setze* $x_{k+1} = x_k + s_k$.

Schnelle lokale Konvergenz läßt sich gegen eine Punkt zeigen, der folgende Voraussetzungen erfüllt:

Voraussetzungen:

(LSQP1) $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^n \to \mathbb{R}^p$ seien zweimal stetig differenzierbar und $\bar{x}$ erfülle die hinreichende Bedingung 2. Ordnung (3.6) mit $\bar{\lambda} \in \mathbb{R}^m$, $\bar{\mu} \in \mathbb{R}^p$ und es gelte strikte Komplementarität.

(LSQP2) (NLP) erfülle (LICQ) in $\bar{x}$.

Es gilt folgender Satz:

Satz 3.6.3 $\bar{x}$ *erfülle* (LSQP1)–(LSQP2). *Dann existiert* $\rho > 0$, *so dass für alle* $(x_0, \lambda_0, \mu_0) \in B_\rho(\bar{x}, \bar{\lambda}, \bar{\mu})$ *das lokale SQP-Verfahren in Algorithmus 20 entweder endlich terminiert oder eine eindeutig definierte Folge* (x_k, λ_k, μ_k) *erzeugt, die Q-superlinear gegen* $(\bar{x}, \bar{\lambda}, \bar{\mu})$ *konvergiert. Sind* $\nabla^2 f, \nabla^2 c_i, \nabla^2 h_i$ *Lipschitz-stetig in einer Umgebung von* $\bar{x}$, *dann ist die Konvergenz Q-quadratisch.*

Beweis: Der Beweis beruht auf der Beobachtung, dass das KKT-System für (NLP) wegen der strikten Komplementarität in einer Umgebung von $(\bar{x}, \bar{\lambda}, \bar{\mu})$ äquivalent ist zu

$$F(x, \lambda, \mu) := \begin{pmatrix} \nabla_x L(x, \lambda, \mu) \\ c_{\mathcal{A}(\bar{x})}(x) \\ h(x) \\ \lambda_{\mathcal{I}(\bar{x})} \end{pmatrix} = 0. \tag{3.16}$$

Tatsächlich liefert die Komplementaritätsbedingung

$$(*) \qquad \lambda_i c_i(x) = 0, \quad i = 1, \dots, m,$$

und sie ist für $\bar{x}$, $\bar{\lambda}$ mit srikter Komplementarität erfüllt, also

$$\bar{\lambda}_{\mathcal{A}(\bar{x})} > 0, \quad c_{\mathcal{A}(\bar{x})}(\bar{x}) = 0, \quad \bar{\lambda}_{\mathcal{I}(\bar{x})} = 0, \quad c_{\mathcal{I}(\bar{x})}(\bar{x}) < 0.$$

Für $\rho > 0$ klein genug gilt also

$$\lambda_{\mathcal{A}(\bar{x})} > 0, \quad c_{\mathcal{I}(\bar{x})}(x) < 0 \quad \forall\, (x, \lambda, \mu) \in B_\rho(\bar{x}, \bar{\lambda}, \bar{\mu}).$$

Damit ist $(*)$ auf $B_\rho(\bar{x}, \bar{\lambda}, \bar{\mu})$ äquivalent zu

$$\begin{pmatrix} c_{\mathcal{A}(\bar{x})}(x) \\ \lambda_{\mathcal{I}(\bar{x})} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Weiter läßt sich mit (LSQP1), (LSQP2) zeigen, dass $F'(\bar{x}, \bar{\lambda}, \bar{\mu})$ invertierbar ist. Für $\rho > 0$ klein genug konvergiert das Newton-Verfahren für (3.16) nach Satz 2.7.3 also zu jedem Startpunkt $(x_0, \lambda_0, \mu_0) \in B_\rho(\bar{x}, \bar{\lambda}, \bar{\mu})$ Q-superlinear bzw. Q-quadratisch.

Andererseits sieht man leicht, dass mit $\rho > 0$ klein genug für jedes $(x_k, \lambda_k, \mu_k) \in B_\rho(\bar{x}, \bar{\lambda}, \bar{\mu})$ das SQP-Teilproblem (SQP$_k$) ein eindeutiges KKT-Tripel $(s_k, \lambda_{k+1}, \mu_{k+1})$ besitzt mit $(x_k + s_k, \lambda_{k+1}, \mu_{k+1}) \in B_\rho(\bar{x}, \bar{\lambda}, \bar{\mu})$ und dass die von Algorithmus 20 erzeugte Folge (x_k, λ_k, μ_k) genau die vom Newton-Verfahren für (3.16) gelieferte Folge ist. Für einen detaillierten Beweis siehe z.B. [GK02]. $\square$

Die Lösung des SQP-Problems (SQP$_k$) kann mit einem beliebigen QP-Löser erfolgen, zum Beispiel mit einem Aktive-Mengen-Verfahren wie es kurz in 3.7 beschrieben wird (siehe auch Einf. in die Optimierung). Ist H_k positiv definit, dann ist jeder KKT-Punkt globale Lösung.

3.6.3 Globalisiertes SQP-Verfahren

Wir nehmen in diesem Abschnitt lediglich an, dass $H_k \in \mathbb{R}^{n,n}$ eine beliebige symmetrische Matrix ist. Zur Globalisierung des lokalen SQP-Verfahrens verwenden wir eine exakte Penalty-Funktion mit der Armijo-Schrittweitenregel. Wir beschränken uns hierbei auf die ℓ_1-Penalty-Funktion

$$P_{\ell_1,\rho} = f(x) + \rho((\|c(x))_+\|_1 + \|h(x)\|_1).$$

Diese Penalty-Funktion ist stetig, aber am Rand des zulässigen Bereichs nicht differenzierbar. Die Armijo-Schrittweitenregel ist dennoch anwendbar, da $P_{\ell_1,\rho}$ zumindest richtungsdifferenzierbar ist.

Definition 3.6.4 *Eine stetige Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt* richtungsdifferenzierbar *im Punkt $x \in \mathbb{R}^n$, falls für jede Richtung $s \in \mathbb{R}^n$ die Richtungsableitung*

$$f'(x;s) := \lim_{t \searrow 0} \frac{f(x+ts) - f(x)}{t}$$

existiert.

Tatsächlich ist $P_{\ell_1,\rho}$ richtungsdifferenzierbar.

Satz 3.6.5 *Sind f, c, h stetig differenzierbar, dann ist für jedes $\rho > 0$ die ℓ_1-Penalty-Funktion überall richtungsdifferenzierbar, wobei*

$$\begin{aligned}(P_{\ell_1,\rho})'(x;s) = \nabla f(x)^T s + \rho \sum_{c_i(x)>0} \nabla c_i(x)^T s + \rho \sum_{c_i(x)=0} (\nabla c_i(x)^T s)_+ \\ + \rho \sum_{h_i(x)\neq 0} \operatorname{sgn}(h_i(x)) \nabla h_i(x)^T s + \rho \sum_{h_i(x)=0} |\nabla h_i(x)^T s|\end{aligned} \tag{3.17}$$

Beweis: Seien $x, s \in \mathbb{R}^n$ beliebig. Nichtglatt sind höchstens die Terme $\phi_i(x) = |h_i(x)|$ und $\psi_i(x) = |(c_i(x))_+|$. Im Fall $h_i(x) \neq 0$ ist $\phi_i(x) = \operatorname{sgn}(h_i(x)) h_i(x)$ differenzierbar in x mit

$$\nabla \phi_i(x) = \operatorname{sgn}(h_i(x)) \nabla h_i(x).$$

Ist $h_i(x) = 0$, dann gilt für $t > 0$

$$\begin{aligned}\frac{\phi_i(x+ts) - \phi_i(x)}{t} &= \frac{|h_i(x+ts)|}{t} = \frac{|\nabla h_i(x)^T (ts) + o(t)|}{t} \\ &= \frac{t|\nabla h_i(x)^T s| + o(t)}{t} \to |\nabla h_i(x)^T s| \quad \text{für } t \searrow 0,\end{aligned}$$

also $\phi_i'(x;s) = |\nabla h_i(x)^T s|$.

Analog ist $\psi_i(x) = |(c_i(x))_+|$ im Fall $c_i(x) \neq 0$ differenzierbar in x mit

$$\nabla \psi_i(x) = \begin{cases} \nabla c_i(x) & \text{falls } c_i(x) > 0, \\ 0 & \text{falls } c_i(x) < 0. \end{cases}$$

Ist $c_i(x) = 0$, dann gilt wie eben

$$\begin{aligned}\frac{\psi_i(x+ts) - \psi_i(x)}{t} &= \frac{(c_i(x+ts))_+}{t} = \frac{(\nabla c_i(x)^T (ts) + o(t))_+}{t} \\ &= \frac{t(\nabla c_i(x)^T s)_+ + o(t)}{t} \to (\nabla c_i(x)^T s)_+ \quad \text{für } t \searrow 0,\end{aligned}$$

also $\psi_i'(x;s) = (\nabla c_i(x)^T s)_+$.

Damit ist die Richtungsdifferenzierbarkeit gezeigt und Einsetzen liefert obige Formel für $(P_{\ell_1,\rho})'(x;\rho)$. $\square$

Wir zeigen nun, dass jeder KKT-Punkt s_k von (SQP$_k$) unter gewissen Voruassetzungen eine Abstiegsrichtung von $P_{\ell_1,\rho}$ in x_k ist.

Satz 3.6.6 *Es seien f, c, h stetig differenzierbar. Weiter sei s_k ein KKT-Punkt von (SQP$_k$) mit Lagrange-Multiplikatoren λ_{k+1}, μ_{k+1}. Dann gilt für jedes*

$$\rho \geq \max\{(\lambda_{k+1})_1, \ldots, (\lambda_{k+1})_m, |(\mu_{k+1})_1|, , \ldots, |(\mu_{k+1})_p|\} \tag{3.18}$$

die Ungleichung

$$(P_{\ell_1,\rho})'(x_k; s_k) \leq -s_k^T H_k s_k. \tag{3.19}$$

Insbesondere ist s_k eine Abstiegsrichtung, falls H_k positiv definit ist.

Beweis: Die KKT-Bedingungen für (SQP$_k$) lauten

$$\nabla f(x_k) + H_k s_k + \nabla c(x_k)\lambda_{k+1} + \nabla h(x_k)\mu_{k+1} = 0, \tag{3.20}$$

$$\begin{aligned} &c(x_k) + \nabla c(x_k)^T s_k \leq 0, \quad h(x_k) + \nabla h(x_k)^T s_k = 0 \\ &\lambda_{k+1} \geq 0, \;\; \lambda_{k+1}^T(c(x_k) + \nabla c(x_k)^T s_k) = 0. \end{aligned} \tag{3.21}$$

Nach Satz 3.6.5 existiert $(P_{\ell_1,\rho})'(x_k; s_k)$ und wegen $\nabla h_i(x_k)^T s_k = -h_i(x_k) = 0$ für $h_i(x_k) = 0$ sowie $\nabla c_i(x_k)^T s_k \leq -c_i(x_k) = 0$ für $c_i(x_k) = 0$ vereinfacht sich (3.17) zu

$$\begin{aligned} (P_{\ell_1,\rho})'(x_k; s_k) &= \nabla f(x_k)^T s_k + \rho \sum_{c_i(x_k)>0} \underbrace{\nabla c_i(x_k)^T s_k}_{\leq -c_i(x_k)} + \rho \sum_{h_i(x_k)\neq 0} \operatorname{sgn}(h_i(x_k))\, \underbrace{\nabla h_i(x_k)^T s_k}_{=-h_i(x_k)} \\ &\leq \nabla f(x_k)^T s_k + \rho \sum_{c_i(x_k)>0} (-c_i(x_k)) + \rho \sum_{h_i(x_k)\neq 0} (-|h_i(x_k)|). \end{aligned}$$

Nun gilt wegen (3.20) und (3.21)

$$\begin{aligned} \nabla f(x_k)^T s_k &= -s_k^T H_k s_k \underbrace{-\lambda_{k+1}^T \nabla c(x_k)^T s_k}_{=\lambda_{k+1}^T c(x_k)} \underbrace{-\mu_{k+1}^T \nabla h(x_k)^T s_k}_{=\mu_{k+1}^T h(x_k)} \\ &\leq -s_k^T H_k s_k + \sum_{c_i(x_k)>0} (\lambda_{k+1})_i c_i(x_k) + \sum_{h_i(x_k)\neq 0} (\mu_{k+1})_i h_i(x_k). \end{aligned}$$

Einsetzen in die vorletzte Ungleichung ergibt

$$\begin{aligned} (P_{\ell_1,\rho})'(x_k; s_k) &\leq -s_k^T H_k s_k + \sum_{c_i(x_k)>0} ((\lambda_{k+1})_i - \rho) c_i(x_k) + \sum_{h_i(x_k)\neq 0} (|(\mu_{k+1})_i| - \rho)|h_i(x_k)| \\ &\leq -s_k^T H_k s_k. \end{aligned}$$

□

Völlig analog wie in Lemma 2.4.3 läßt sich zeigen, dass die Armijo-Regel für $P_{\ell_1,\rho}$ in x_k wohldefiniert ist, falls s_k eine Abstiegsrichtung ist. Dies motiviert den folgenden Algorithmus:

Algorithmus 21 Globalisiertes SQP-Verfahren für NLP
Wähle $\gamma \in (0, 1/2)$ für die Armijo-Regel. Wähle $\rho > 0$ genügend groß, einen Startpunkt $(x_0, \lambda_0, \mu_0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ und eine symmetrische Matrix $H_0 \in \mathbb{R}^{n,n}$.

Für $k = 0, 1, \dots$:

1. *Falls (x_k, λ_k, μ_k) die KKT-Bedingungen für (NLP) erfüllt: STOP mit Ergebnis x_k.*

2. *Berechne die am nächsten bei 0 liegende lokale Lösung s_k des SQP-Problems (SQP$_k$) mit zugehörigen Lagrange-Multiplikatoren λ_{k+1}, μ_{k+1}. Falls $s_k^T H_k s_k \leq 0$ modifiziere H_k und wiederhole Schritt 2.*

3. *Bestimme das größte $\sigma_k \in \{1, 2^{-1}, 2^{-2}, \dots\}$ mit*
$$P_{\ell_1,\rho}(x_k) - P_{\ell_1,\rho}(x_k + \sigma_k s_k) \geq -\gamma\sigma_k (P_{\ell_1,\rho})'(x_k; s_k).$$

4. *Setze $x_{k+1} = x_k + \sigma_k s_k$ und wähle eine neue symmetrische Matrix $H_{k+1} \in \mathbb{R}^{n,n}$.*

Bemerkung: In einer praktischen Implementierung wählt man den Penalty-Parameter $\rho > 0$ dynamisch, z.B. auf Basis von (3.18) zu

$$\rho_{k+1} = \max\left\{\rho_k, \alpha + \max\{\|\lambda_{k+1}\|_\infty, \|\mu_{k+1}\|_\infty\}\right\}.$$

mit einem $\alpha > 0$. Die Wohldefiniertheit der Armijo-Regel ist immer gesichert, wenn H_k positiv definit gewählt wird. □

Für Algorithmus 21 lassen sich recht befriedigende globale Konvergenzeigenschaften nachweisen, siehe z. B. die Arbeit von Han [Ha77], in der unter anderem ein Konvergenzresultat für konvexe Probleme zu finden ist. Dennoch können bei Algorithmus 21 noch einige Probleme auftreten, die man bei einer Implementierung berücksichtigen muss. Wir gehen nun kurz auf mögliche Probleme ein.

3.6.4 Probleme beim SQP-Verfahren und mögliche Lösungen

Unzulässige SQP-Teilprobleme

Es kann passieren, dass der zulässige Bereich von (SQP$_k$) leer ist.

Beispiel 3.6.1 Betrachte das Problem

$$\min f(x) \quad \text{unter der Nebenbed.} \quad x_2 - x_1 \geq 0, \quad x_2 - x_1^2 \leq 0.$$

Dann ist das SQP-Problem in $\binom{1/2}{1/4}$ unzulässig:

Linearisierung der Nebenbedingungen im Punkt $\binom{1/2}{1/4}$ ergibt die linearisierten Nebenbedingungen

$$-1/4 + s_2 - s_1 \geq 0, \quad s_2 - s_1 \leq 0.$$

Diese Nebenbedingungen sind inkompatibel, denn Subtraktion der zweiten von der ersten ergibt $-1/4 \geq 0$.

Um immer die Zulässigkeit der SQP-Teilprobleme sicherzustellen, kann man die Nebenbedingungen relaxieren und gleichzeitig die Relaxation bestrafen. Anstelle von (SQP$_k$) verwendet man nun das Problem

(SQP$_{\rho,k}$)

$$\begin{aligned}
\min_{(s,y,z^+,z^-)\in\mathbb{R}^{n+m+2p}} \; & f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T H_k s + \rho \sum_{i=1}^m y_i + \rho \sum_{i=1}^p ((z^+)_i + (z^-)_i) \\
\text{u. d. Nebenbed.} \quad & c(x_k) + \nabla c(x_k)^T s - y \leq 0, \quad y \geq 0 \\
& h(x_k) + \nabla h(x_k)^T s - z^+ + z^- = 0, \; z^+ \geq 0, \; z^- \geq 0.
\end{aligned}$$

Hierbei ist $\rho > 0$ ein Penalty-Parameter.

(SQP$_{\rho,k}$) besitzt immer den zulässigen Punkt $(0, (c(x_k))_+, (h(x_k))_+, (-h(x_k))_+)$. Durch Vergleich der KKT-Bedingungen von (SQP$_k$) und (SQP$_{\rho,k}$) kann man leicht folgendes zeigen:

Lemma 3.6.7 *i) Ist s_k ein KKT-Punkt von (SQP$_k$) mit Multiplikatoren λ_{k+1}, μ_{k+1} dann ist für alle*

$$\rho \geq \max\{(\lambda_{k+1})_1, \ldots, (\lambda_{k+1})_m, |(\mu_{k+1})_1|, \ldots, |(\mu_{k+1})_p|\}$$

der Punkt $(s_k, 0, 0, 0)$ ein KKT-Punkt von (SQP$_{\rho,k}$) mit Multiplikatoren λ_{k+1}, μ_{k+1}, η_{k+1}, η^+_{k+1}, η^-_{k+1}, wobei

$$\eta_{k+1} = \rho e - \lambda_{k+1}, \quad \eta^+_{k+1} = \rho e - \mu_{k+1}, \quad \eta^-_{k+1} = \rho e + \mu_{k+1}.$$

Hierbei ist jeweils $e = (1, \ldots, 1)^T$ und η_{k+1}, η^+_{k+1} bzw. η^-_{k+1} sind die Multiplikatoren zu $-y \leq 0$, $-z^+ \leq 0$ bzw. $-z^- \leq 0$.

ii) Ist $(s_k, 0, 0, 0)$ ein KKT-Punkt von (SQP$_{\rho,k}$) mit Multiplikatoren λ_{k+1}, μ_{k+1}, η_{k+1}, η^+_{k+1}, η^-_{k+1}, dann ist s_k eine KKT-Punkt von (SQP$_k$) mit Multiplikatoren λ_{k+1}, μ_{k+1}.

Unabhängig von der Größe von $\rho > 0$ kann man für eine Lösung (s_k, y_k, z_k^+, z_k^-) von (SQP$_{\rho,k}$) nachweisen, dass gilt

$$(P_{\ell_1,\rho})'(x_k; s_k) \le -s_k^T H_k s_k. \tag{3.19}$$

Damit ist s_k eine Abstiegsrichtung, falls H_k positiv definit ist (beachte, dass für $P_{\ell_1,\rho}$ und (SQP$_{\rho,k}$) dasselbe ρ verwendet wird).

Verwendet man Algorithmus 21 mit Teilproblem (SQP$_{\rho,k}$) und Hessematrizen H_k, so dass für Konstanten $0 < c_0 \le c_1$ gilt

$$c_0\|s\|^2 \le s^T H_k s \le c_1\|s\|^2 \quad \forall\, s \in \mathbb{R}^n,$$

dann kann man nachweisen, dass jeder Häufungspunkt $\bar{x}$ von (x_k) ein stationärer Punkt von $P_{\ell_1,\rho}$ ist, d.h. $(P_{\ell_1,\rho})'(\bar{x}; s) = 0$ für alle $s \in \mathbb{R}^n$. Für einen Beweis siehe z.B. [GK02].

Maratos-Effekt

Wir wissen, dass das lokale SQP-Verfahren Q-superlinear gegen einen Punkt $\bar{x}$ konvergiert, der (LSQP1), (LSQP2) erfüllt. Es gibt aber Fälle, in denen für x_k beliebig nahe bei $\bar{x}$ für jedes $\rho > 0$ gilt

$$P_{\ell_1,\rho}(x_k + s_k) > P_{\ell_1,\rho}(x_k).$$

Die Armijo-Bedingung in Schritt 3 von Algorithmus 21 (oder jede andere Abstiegsbedingung) würde daher den SQP-Schritt mit Schrittweite $\sigma_k = 1$ nicht zulassen und daher die Q-superlineare lokale Konvergenz zerstören. s_k ist dann zwar eine Abstiegsrichtung, aber der volle Schritt erhöht f und den ℓ_1-Strafterm. Dies kann passieren, wenn $x_k + s_k$ im Vergleich zu x_k die Nebenbedingungen zu stark verletzt.

Dieses Phänomen wurde zuerst von Maratos in seiner Dissertation [Ma78] entdeckt.

Beispiel 3.6.2 Betrachte das Beispiel

$$\min_{x=(\xi_1,\xi_2)^T\in\mathbb{R}^2} f(x) := 2(\xi_1^2 + \xi_2^2 - 1) - \xi_1 \quad \text{unter der Nebenbed.} \quad h(x) := \xi_1^2 + \xi_2^2 - 1 = 0.$$

Man prüft leicht, dass $\bar{x} = \binom{1}{0}$ das globale Minimum ist mit einem Lagrangemultiplikator $\bar{\mu}$. Man kann zeigen, dass das lokale SQP-Verfahren für Startpunkte $(x_0, \mu_0) \in B_\delta(\bar{x}, \bar{\mu})$, $\delta > 0$ klein genug, Q-superlinear gegen $(\bar{x}, \bar{\mu})$ konvergiert.
Andererseits gilt für $x_k \in Z \setminus \{\bar{x}\}$ und $\mu_k < -1$ beliebig (solche Punkte existieren in jeder Umgebung von $(\bar{x}, \bar{\mu})$), dass die Penalty-Funktion $P_{\ell_1,\rho}$ den SQP-Schritt s_k ablehnt, da gilt

$$P_{\ell_1,\rho}(x_k + s_k) > P_{\ell_1,\rho}(x_k).$$

Begründung: $\bar{x}$ erfüllt mit dem Multiplikator $\bar{\mu} = -\frac{3}{2}$ die Voraussetzungen (LSQPG1), (LSQPG2). Das lokale SQP-Verfahren konvergiert also für Startpunkte $(x_0, \mu_0) \in B_\delta(\bar{x}, \bar{\mu})$, $\delta > 0$ klein genug, Q-superlinear gegen $(\bar{x}, \bar{\mu})$.

Seien nun $x_k \in Z \setminus \{\bar{x}\}$ und $\mu_k < -1$ beliebig (solche Punkte existieren in jeder Umgebung von $(\bar{x}, \bar{\mu})$). Weiter sei s_k die Lösung von (SQP$_k$) mit Multiplikator μ_{k+1}. Die KKT-Bedingungen für (SQP$_k$) liefern

$$\nabla f(x_k) + \nabla^2_{xx} L(x_k, \mu_k) s_k + \nabla h(x_k) \mu_{k+1} = 0.$$

Nun ist $h(x_k) + \nabla h(x_k)^T s_k = \nabla h(x_k)^T s_k = 0$. Da h und f quadratisch sind, gilt

$$h(x_k + s_k) - h(x_k) = \nabla h(x_k)^T s_k + s_k^T s_k = s_k^T s_k > 0$$

sowie

$$\begin{aligned} f(x_k + s_k) - f(x_k) &= \nabla f(x_k)^T s_k + 2 s_k^T s_k = -s_k^T \nabla^2_{xx} L(x_k, \mu_k) s_k + 2 s_k^T s_k \\ &= (-4 - 2\mu_k + 2) s_k^T s_k = 2(-\mu_k - 1) s_k^T s_k > 0 \end{aligned}$$

wegen $\mu_k < -1$. Damit haben wir

$$|h(x_k + s_k)| > |h(x_k)| \quad \text{und} \quad f(x_k + s_k) > f(x_k), \ \text{ also } \quad P_{\ell_1,\rho}(x_k + s_k) > P_{\ell_1,\rho}(x_k).$$

□

Als Ausweg kann man d_k durch eine *Second-Order-Correction* s_k^{SOC} so korrigieren, dass gilt

$$h(x_k + s_k + s_k^{SOC}) = o(\|s_k\|^2)$$

anstelle der sonst lediglich garantierten Zulässigkeit

$$h(x_k + s_k) = O(\|s_k\|^2).$$

Man erreicht dies zum Beispiel durch die Wahl

$$s_k^{SOC} = -\nabla h(x_k)(\nabla h(x_k)^T \nabla h(x_k))^{-1} h(x_k + s_k).$$

Im Falle von Ungleichungsnebenbedingungen verwendet man $((\nabla c_i(x_k))_{(\lambda_k)_i > 0}, \nabla h(x_k))$ anstelle $\nabla h(x_k)$.

Der Schritt s_k^{SOC} erfüllt $s_k^{SOC} = O(\|s_k\|^2)$, ist also sehr kurz im Vergleich zu s_k. Mann kann nun zeigen, dass Algorithmus 21 mit $H_k = \nabla^2_{xx} L(x_k, \lambda_k, \mu_k)$ und Schrittweitensuche entlang der Kurve

$$x_k + \sigma s_k + \sigma^2 s_k^{SOC}$$

Übergang zu Q-superlinear Konvergenz zuläßt.

3.6.5 BFGS-Updates für SQP-Verfahren

Wir wissen bereits, dass die Wahl $H_k = \nabla^2_{xx}L(x_k, \lambda_k, \mu_k)$ lokal Q-superlineare Konvergenz liefert. Allerdings werden dann die zweiten Ableitungen von f, c, h benötigt. Zudem ist H_k nicht immer positiv definit, was für das globale Konvergenzverhalten günstig ist. Es liegt daher nahe, Quasi-Newton-Updates H_k, insbesondere BFGS-Updates, zu verwenden, die die Quasi-Newton-Gleichung

$$H_{k+1}d_k = y_k$$

mit

$$d_k = x_{k+1} - x_k, \quad y_k = \nabla_x L(x_{k+1}, \lambda_k, \mu_k) - \nabla_x L(x_k, \lambda_k, \mu_k)$$

erfüllen. Da wir jedoch eine Art Armijo-Regel für die ℓ_1-Penalty-Funktion verwenden, ist $d_k^T y_k > 0$ nicht garantiert und somit die positive Definitheit von H_{k+1} nicht sichergestellt.

Aus diesem Grund schlägt Powell [Po78] vor, den Update

$$H_{k+1} = \Phi^{BFGS}(H_k, d_k, y_k^{mod}) \tag{3.22}$$

zu verwenden, wobei

$$y_k^{mod} = \theta_k y_k + (1-\theta_k)H_k d_k \tag{3.23}$$

mit

$$\theta_k = \begin{cases} 1 & \text{falls } d_k^T y_k \geq 0.2\, d_k^T H_k d_k, \\ \frac{0.8\, d_k^T H_k d_k}{d_k^T H_k d_k - d_k^T y_k} & \text{sonst.} \end{cases}$$

Mit dieser Wahl läßt sich leicht zeigen, dass gilt

$$d_k^T y_k^{mod} > 0,$$

und somit ist H_{k+1} gemäß (3.22), (3.23) positiv definit nach Lemma 2.10.2.

Der "gedämpfte BFGS-Update" (3.22), (3.23) wird häufig eingesetzt. Eine Konvergenzanalyse findet man in [Po78a], wo insbesondere R-superlineare Konvergenz unter geeigneten Voraussetzungen gezeigt wird.

3.7 Lösung quadratischer Optimierungsprobleme

Wir haben gesehen, dass beim SQP-Verfahren in jeder Iteration ein quadratisches Optimierungsproblem (QP) zu lösen ist. Allgemein sind QPs eine wichtige Klasse von Optimierungsproblemen.

Quadratisches Optimierungsproblem (QP):

$$\text{(QP)} \qquad \begin{aligned} &\min \quad q(x) := d^T x + \frac{1}{2}x^T H x \\ &\text{u. d. Nebenbed.} \quad c(x) := a + A^T x \leq 0, \quad h(x) := b + B^T x = 0, \end{aligned}$$

mit $d \in \mathbb{R}^n$, $H \in \mathbb{R}^{n,n}$ symmetrisch, $a \in \mathbb{R}^m$, $A \in \mathbb{R}^{n,m}$, $b \in \mathbb{R}^p$, $B \in \mathbb{R}^{n,p}$.

Im nichtkonvexen Fall kann (QP) $O(2^n)$ lokale Minima mit unterschiedlichen Funktionswerten haben. Zum Beispiel hat

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^n 2^{i-1}\frac{1-x_i^2}{3} \quad \text{unter der Nebenbedingung} \quad -1 \le x \le 2$$

in jeder Ecke $x \in \{-1,2\}^n$ des Würfels $[-1,2]^n$ ein lokales Minimum mit Funktionswerten $0, -1, -2, \dots, -2^n+1$.

Wir betrachten daher hier nur QPs mit H symmetrisch positiv definit. Dann ist (QP) konvex mit streng konvexer Zielfunktion und es gilt eine CQ, da die Nebenbedingungen affin linear sind. Also hat (QP) nach Satz 2.2.6 ein eindeutiges Minimum $\bar{x}$, und $\bar{x}$ ist nach Satz 3.2.20 der einzige KKT-Punkt.

Sind $\bar{\lambda}, \bar{\mu}$ zugehörige Lagrange-Multiplikatoren, dann gelten die KKT-Bedingungen genau dann, wenn für ein $\mathcal{A} \subset \mathcal{A}(\bar{x})$ gilt

$$\nabla q(x) + A_{\mathcal{A}}\bar{\lambda}_{\mathcal{A}} + B\bar{\mu} = 0, \tag{3.24}$$

$$a_{\mathcal{A}} + A_{\mathcal{A}}^T\bar{x} = 0, \quad b + B^T\bar{x} = 0, \tag{3.25}$$

$$\bar{\lambda}_{\mathcal{A}} \ge 0, \ \ \bar{\lambda}_{\{1,\dots,m\}\setminus\mathcal{A}} = 0, \qquad a + A\bar{x} \le 0. \tag{3.26}$$

wobei $A_{\mathcal{A}}$ die Spalten $i \in \mathcal{A}$ von A enthalte. Nun sind (3.24), (3.25) gerade die KKT-Bedingungen für das der aktiven Menge $\mathcal{A}_k := \mathcal{A}$ zugeordnete gleichungsrestringierte Problem

$$(\mathrm{QP}_k) \qquad \begin{array}{ll} \min & q(x) := d^Tx + \dfrac{1}{2}x^THx \\ \text{u. d. Nebenbed.} & a_{\mathcal{A}_k} + A_{\mathcal{A}_k}^T x = 0, \quad b + B^Tx = 0. \end{array}$$

Umgekehrt ist ein KKT-Punkt $\bar{x}_k$ von (QP$_k$) mit Multiplikatoren $\bar{\lambda}_k$, $\bar{\mu}_k$ ein KKT-Punkt von (QP), falls gilt

$$\bar{\lambda}_k \ge 0, \quad a + A\bar{x}_k \le 0, \tag{3.27}$$

denn mit $\mathcal{A} = \mathcal{A}_k$, $\bar{\mu} = \bar{\mu}_k$ und $\bar{\lambda}$ definiert durch $\bar{\lambda}_{\mathcal{A}_k} = \bar{\lambda}_k$, $\bar{\lambda}_{\{1,\dots,m\}\setminus\mathcal{A}_k} = 0$ gilt dann (3.24)–(3.26).

Sei x_0 zulässig für (QP). Die grundlegende Idee von Aktive-Menge-Verfahren besteht nun darin, für eine Approximation $\mathcal{A}_k \subset \mathcal{A}(x_k)$ der aktiven Menge das Problem (QP$_k$) zu lösen. Ist (3.27) erfüllt, dann ist die Lösung $\bar{x}_k$ von (QP) gefunden. Sonst wird ein zulässiger Punkt $x_{k+1} = x_k + \sigma_k(\bar{x}_k - x_k)$ von (QP) mit $q(x_{k+1}) \le q(x_k)$ bestimmt und eine neue aktive Menge $\mathcal{A}_{k+1} \subset \mathcal{A}(x_{k+1})$ mit $\mathcal{A}_{k+1} \notin \{\mathcal{A}_0, \dots, \mathcal{A}_k\}$ gewählt.

Dies ergibt folgenden Algorithmus:

Algorithmus 22 *Wähle einen zulässigen Punkt x_0 für (QP) sowie $\lambda_0 \in \mathbb{R}^m$, $\mu_0 \in \mathbb{R}^p$. Wähle $\mathcal{A}_0 \subset \mathcal{A}(x_0)$.*

Für $k = 0, 1, 2, \ldots$:

1. *Ist x_k ein KKT-Punkt von (QP) mit Multiplikatoren λ_k, μ_k: STOP*
2. *Setze $\mathcal{I}_k = \{1, \ldots, m\} \setminus \mathcal{A}_k$ und bestimme eine Lösung $\bar{x}_k$ von (QP$_k$) mit Multiplikatoren $\bar{\lambda}_k, \bar{\mu}_k$. Setze $s_k = \bar{x}_k - x_k$, $\mu_{k+1} = \bar{\mu}_k$ und definiere λ_{k+1} durch $(\lambda_{k+1})_{\mathcal{A}_k} = \bar{\lambda}_k$, $(\lambda_{k+1})_{\mathcal{I}_k} = 0$.*
3. *Ist $\bar{x}_k$ zulässig für (QP) und gilt $\lambda_{k+1} \geq 0$: STOP mit KKT-Punkt $\bar{x}_k$.*
4. *Ist $\bar{x}_k$ zulässig für (QP) und gibt es $j \in \mathcal{A}_k$ mit $(\lambda_{k+1})_j = \min_i(\lambda_{k+1})_i < 0$, dann setze $x_{k+1} = \bar{x}_k$, $\mathcal{A}_{k+1} = \mathcal{A}_k \setminus \{j\}$ und gehe in die nächste Iteration.*
5. *Ist $\bar{x}_k$ nicht zulässig für (QP), dann bestimme*

$$\sigma_k = \max\left\{\sigma \geq 0 \,:\, x_k + \sigma s_k \textit{zulässig für (QP)}\right\}.$$

Setze $x_{k+1} = x_k + \sigma_k s_k$ und $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{j\}$ mit einem $j \in \mathcal{I}_k$ so dass $(a + A^T x_{k+1})_j = 0$.

In seltenen Fällen kann passieren, dass der Algorithmus ”kreist”, also $x_k = x_l$ für alle $k \geq l$ mit einem l gilt. In diesem Falle wird immer wieder derselbe Zyklus aktiver Mengen generiert. Dies kann durch geeignete Auswahlregeln in Schritt 4 und 5 verhindert werden.

Wir halten folgende Eigenschaften des Algorithmus fest, falls H positiv definit ist:

- Alle x_k sind zulässig und es gilt $q(x_{k+1}) \leq q(x_k)$.
- Im Fall $x_{k+1} \neq x_k$ gilt $q(x_{k+1}) < q(x_k)$, da q streng konvex ist und für die Lösung $\bar{x}_k$ von (QP$_k$) wegen $\bar{x}_k \neq x_k$ gilt $q(\bar{x}_k) < q(x_k)$.
- Nach jeweils spätestens n Iterationen ist x_{k+1} Lösung von (QP$_k$), denn:

1. Fall Ist $\bar{x}_k$ zulässig, dann ist $x_{k+1} = \bar{x}_k$ Lösung von (QP$_k$).

2. Fall Solange $\bar{x}_k$ nicht zulässig ist, wird in Schritt 5 eine neue, linear unabhängige Nebenbedingung j aktiv. Tritt der 1. Fall nicht auf, dann sind nach maximal n Schritten n linear unabhängige Nebenbedingungen aktiv. (QP$_k$) hat dann nur den zulässigen Punkt x_k und es gilt im nächsten Schritt $x_{k+1} = \bar{x}_k = x_k$. Also ist x_{k+1} Lösung von (QP$_k$).

- Der Fall $x_{k+1} \neq x_k$ kann nur endlich oft auftreten: Wäre dies nicht so, dann würde nach den bisherigen Überlegungen eine Teilfolge k_i existieren mit x_{k_i} löst (QP$_{k_i-1}$) und $q(x_{k_i}) < q(x_{k_{i-1}})$. Also müssten alle Probleme (QP$_{k_i-1}$) verschieden sein, es gibt aber nur endlich viele verschiedene Möglichkeiten für (QP$_k$).

- Terminiert der Algorithmus also nicht endlich, dann gilt $x_k = x_l$ für alle $k \geq l$ mit geeignetem l. Dies ist nur möglich, wenn Schritt 4 und 5 immer wieder dieselben Indexmengen $\mathcal{A}_k$ generiert, für die $\lambda_{k+1} \ngeq 0$. Erzwingt man durch eine Auswahlregel, dass alle möglichen Indexmengen $\mathcal{A}_k$ ausprobiert werden, wenn n linear unabhängige Nebenbedingungen "aufgesammelt" wurden, dann findet man entweder $\mathcal{A}_k$ mit $\lambda_{k+1} \geq 0$ oder mit $x_{k+1} \neq x_k$ (Anwendung des Lemmas von Farkas).

3.8 Dualität

Wir betrachten wieder das Problem

(NLP) $$\min_{x\in\mathbb{R}^n} f(x) \quad \text{u.d. Nebenbedingung} \quad h(x) = 0, \quad c(x) \leq 0.$$

mit zugehöriger Lagrange-Funktion

$$L(x, \lambda, \mu) = f(x) + \lambda^T c(x) + \mu^T h(x).$$

3.8.1 Das duale Problem

Wir wollen (NLP) ein duales Problem zuordnen, das in gewissen Fällen äquivalent zu (NLP) ist, aber in jedem Fall Unterschranken für den Optimalwert von (NLP) liefert.

Die Konstruktion eines dualen Problems für (NLP) beruht auf der Beobachtung, dass gilt

$$p(x) := \sup_{\lambda\in\mathbb{R}^m_+, \mu\in\mathbb{R}^p} L(x, \lambda, \mu) = \begin{cases} f(x) & \text{falls } x \in Z, \\ +\infty & \text{sonst.} \end{cases}$$

Damit ist unser primales Problem (NLP) äquivalent zu

(NLP) $$\min_{x\in\mathbb{R}^n} \sup_{\lambda\in\mathbb{R}^m_+, \mu\in\mathbb{R}^p} L(x, \lambda, \mu).$$

Nun ist es naheliegend, ein duales Problem durch Vertauschen von $\min$ und $\sup$ zu gewinnen:

Definition 3.8.1 *Das Problem*

(DNLP) $$\sup_{\lambda\in\mathbb{R}^m_+, \mu\in\mathbb{R}^p} \inf_{x\in\mathbb{R}^n} L(x, \lambda, \mu).$$

heißt zu (NLP) duales Problem. *Die Funktion*

$$d(\lambda, \mu) = \inf_{x\in\mathbb{R}^n} L(x, \lambda, \mu)$$

heißt duale Zielfunktion und die Funktion

$$p(x) = \sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} L(x, \lambda, \mu)$$

primale Zielfunktion.

Bemerkung: Für jedes feste x ist $(\lambda, \mu) \to L(x, \lambda, \mu)$ eine lineare Funktion. Daher ist $d(\lambda, \mu)$ als Infimum von linearen Funktionen konkav. Daher ist das duale Problem ein Maximierungsproblem einer konkaven Funktion auf einer konvexen Menge, also äquivalent zu einem konvexen Optimierungsproblem. □

Es gilt der wichtige

Satz 3.8.2 (Schwacher Dualitätssatz)
Ist $\tilde{x}$ zulässig für das primale Problem (NLP) und $(\tilde{\lambda}, \tilde{\mu})$ zulässig für das duale Problem (DNLP), dann gilt

$$p(\tilde{x}) = f(\tilde{x}) \geq d(\tilde{\lambda}, \tilde{\mu})$$

Beweis: Wegen $\tilde{\lambda} \geq 0$, $c(\tilde{x}) \leq 0$, $h(\tilde{x}) = 0$ gilt

$$d(\tilde{\lambda}, \tilde{\mu}) = \inf_{x \in \mathbb{R}^n} L(x, \tilde{\lambda}, \tilde{\mu}) \leq L(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) = f(\tilde{x}) + \tilde{\lambda}^T c(\tilde{x}) + \tilde{\mu}^T h(\tilde{x}) \leq f(\tilde{x}) = p(\tilde{x}).$$

□

In vielen Fällen sind die Optimalwerte von (NLP) und (DNLP) gleich und zwar genau dann, wenn die Lagrange-Funktion einen Sattelpunkt besitzt.

Definition 3.8.3 *Der Punkt $(\bar{x}, \bar{\lambda}, \bar{\mu}) \in \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p$ heißt Sattelpunkt der Lagrangefunktion, falls*

$$L(\bar{x}, \lambda, \mu) \leq L(\bar{x}, \bar{\lambda}, \bar{\mu}) \leq L(x, \bar{\lambda}, \bar{\mu}) \quad \forall\, x \in \mathbb{R}^n, \lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p.$$

Satz 3.8.4 *Die folgenden Aussagen sind äquivalent:*

i) $(\bar{x}, \bar{\lambda}, \bar{\mu})$ *ist Sattelpunkt der Lagrangefunktion.*

ii) $\bar{x}$ *ist globales Optimum von (NLP),* $(\bar{\lambda}, \bar{\mu})$ *ist globales Optimum von (DNLP) und* $f(\bar{x}) = d(\bar{\lambda}, \bar{\mu})$.

Der Beweis verwendet die Tatsache, dass für jede Funktion $L : \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$ gilt

$$\sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu) \leq \inf_{x \in \mathbb{R}^n} \sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} L(x, \lambda, \mu). \tag{3.28}$$

Tatsächlich gilt für jedes $\tilde{x} \in \mathbb{R}^n$

$$\sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu) \le \sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} L(\tilde{x}, \lambda, \mu).$$

Da $\tilde{x}$ beliebig war, folgt (3.28).

Beweis: i) $\Longrightarrow$ ii): Wir haben mit (3.28)

$$\begin{aligned} L(\bar{x}, \bar{\lambda}, \bar{\mu}) &= \inf_{x \in \mathbb{R}^n} L(x, \bar{\lambda}, \bar{\mu}) \le \sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu) \\ &\le \inf_{x \in \mathbb{R}^n} \sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} L(x, \lambda, \mu) \\ &\le \sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} L(\bar{x}, \lambda, \mu) \\ &= L(\bar{x}, \bar{\lambda}, \bar{\mu}). \end{aligned}$$

Damit folgt

$$L(\bar{x}, \bar{\lambda}, \bar{\mu}) = \inf_{x \in \mathbb{R}^n} L(x, \bar{\lambda}, \bar{\mu}) = d(\bar{\lambda}, \bar{\mu}) = \sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} L(\bar{x}, \lambda, \mu) = p(\bar{x}) < \infty.$$

Also ist $\bar{x} \in Z$, $p(\bar{x}) = f(\bar{x})$ und wegen $d(\bar{\lambda}, \bar{\mu}) = p(\bar{x})$ folgt die Optimalität von $\bar{x}$ und $(\bar{\lambda}, \bar{\mu})$ aus dem schwachen Dualitätssatz.

ii) $\Longrightarrow$ i): Es folgt

$$L(\bar{x}, \bar{\lambda}, \bar{\mu}) \le f(\bar{x}) = p(\bar{x}) = \sup_{\lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}^p} L(\bar{x}, \lambda, \mu) = d(\bar{\lambda}, \bar{\mu}) = \inf_{x \in \mathbb{R}^n} L(x, \bar{\lambda}, \bar{\mu}) \le L(\bar{x}, \bar{\lambda}, \bar{\mu}).$$

Daraus ist die Sattelpunkteigenschaft abzulesen. □

Bemerkung: Für stetig differenzierbare konvexe Probleme (NLP) läßt sich das duale Problem (DNLP) meist in expliziter Form schreiben. Tatsächlich ist dann für jedes $(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ die Funktion

$$x \mapsto L(x, \lambda, \mu)$$

konvex. Wird nun $\inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu)$ angenommen, existiert also $\min_{x \in \mathbb{R}^n} L(x, \lambda, \mu)$ für alle (λ, μ), dann sind die Minimalpunkte genau alle x mit $\nabla_x L(x, \lambda, \mu) = 0$. Wir können dann (DNLP) schreiben in der Form

$$\sup L(x, \lambda, \mu) \quad \text{unter der Nebenbedingung} \quad \lambda \ge 0, \;\; \nabla_x L(x, \lambda, \mu) = 0.$$

□

3.9 Augmented-Lagrange-Verfahren (Ergänzung)

Wir betrachten abschließend einen wichtigen Vertreter der sogenannten Multiplikatorverfahren, das Augmented-Lagrange-Verfahren. Wir betrachten das gleichungsrestringierte Problem

(NLPG) $$\min_{x\in\mathbb{R}^n} f(x) \quad \text{unter der Nebenbedingung} \quad h(x) = 0.$$

Im folgenden seien $f : \mathbb{R}^n \to \mathbb{R}$ und $h = (h_1, \dots, h_p)^T : \mathbb{R}^n \mapsto \mathbb{R}^p$, $p < n$, zumindest stetig differenzierbar. Augmented-Lagrange-Verfahren stellen eine Weiterentwicklung der quadratischen Penalty-Verfahren dar. Ihr Vorteil besteht darin, daß der Penalty-Parameter während der Iteration nicht gegen Unendlich streben muß, um globale Konvergenz zu erzielen.

3.9.1 Motivation des Augmented-Lagrange-Verfahrens

Für $x \in Z$ gilt $h(x) = 0$ und daher

$$L(x, \mu) = f(x) + \mu^T h(x) = f(x).$$

für beliebige Multiplikatoren $\mu \in \mathbb{R}^p$.

Somit ist (NLPG) äquivalent zu

(3.29) $$\min_{x\in\mathbb{R}^n} L(x, \mu) \quad \text{unter der Nebenbedingung} \quad h(x) = 0.$$

wobei der Multiplikator $\mu \in \mathbb{R}^p$ beliebig, aber fest ist.

Das Hauptproblem des quadratischen Penalty-Verfahrens besteht wie gesagt darin, daß eine Lösung $\bar{x}$ von (NLPG) im allgemeinen kein stationärer Punkt von P_ρ ist, jedenfalls nicht, wenn $\nabla f(\bar{x}) \neq 0$ gilt, was in aller Regel der Fall ist.

Gilt hingegen in $\bar{x}$ eine Constraint Qualification (z.B. Rang $\nabla h(\bar{x}) = p$ oder h affin linear), dann existiert ein Lagrange-Multiplikator $\bar{\mu} \in \mathbb{R}^p$ mit $\nabla_x L(\bar{x}, \bar{\mu}) = 0$. Wählen wir also $\mu = \bar{\mu}$ in (3.29) und wenden dann das Penalty-Verfahren an, so ist $\bar{x}$ ein stationärer Punkt der Penalty-Funktion $L_\rho(x, \bar{\mu})$, wobei L_ρ die Augmented Lagrange-Funktion ist:

Definition 3.9.1 *Sei $\rho \geq 0$. Die Funktion $L_\rho : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$,*

$$L_\rho(x, \mu) = L(x, \mu) + \frac{\rho}{2}\|h(x)\|_2^2 = f(x) + \mu^T h(x) + \frac{\rho}{2}\|h(x)\|_2^2,$$

heißt Augmented Lagrange-Funktion.

Ist $\rho > 0$ hinreichend groß und gilt in $\bar{x}$ die hinreichende Bedingung zweiter Ordnung, so ist $\bar{x}$ sogar ein isoliertes lokales Minimum von $L_\rho(\cdot, \bar{\rho})$:

Satz 3.9.2 *Seien f und h zweimal stetig differenzierbar und $(\bar{x}, \bar{\mu})$ sei ein Kuhn-Tucker-Paar für (NLPG), in dem hinreichende Bedingungen zweiter Ordnung gelten. Dann gibt es $\bar{\rho} > 0$, so daß für alle $\rho \geq \bar{\rho}$ $\nabla_x L_\rho(\bar{x}, \bar{\mu}) = 0$ gilt und $\nabla_{xx} L_\rho(\bar{x}, \bar{\mu})$ positiv definit ist. Insbesondere ist also $\bar{x}$ ein isoliertes lokales Minimum von $L_\rho(\cdot, \bar{\mu})$.*

Der Beweis benutzt das folgende

Lemma 3.9.3 *Sei $A \in \mathbb{R}^{n\times p}$ und $M \in \mathbb{R}^{n\times n}$ symmetrisch. Gilt dann $s^T M s > 0$ für alle $s \neq 0$ mit $A^T s = 0$, so gibt es $\bar{\rho} > 0$, so daß $M + \rho AA^T$ positiv definit ist für alle $\rho \geq \bar{\rho}$.*

Beweis:

Zunächst gilt für $\rho \geq \bar{\rho} \geq 0$:

$$s^T(M + \rho AA^T)s = s^T M s + \rho \|A^T s\|_2^2 \geq s^T M s + \bar{\rho}\|A^T s\|_2^2 = s^T(M + \bar{\rho}AA^T)s.$$

Es genügt daher, zu zeigen, daß es $\bar{\rho} > 0$ gibt, so daß $M + \bar{\rho}AA^T$ positiv definit ist.

Angenommen, es gibt kein solches $\bar{\rho}$. Dann gibt es eine Folge $(s_k) \subset \mathbb{R}^n$, $\|s_k\|_2 = 1$, mit

$$s_k^T(M + kAA^T)s_k \leq 0.$$

Nach Auswahl einer Teilfolge gilt $(s_k)_\mathcal{K} \to s$, $\|s\|_2 = 1$, und

$$0 \geq s_k^T \left(\frac{1}{k}M + AA^T\right) s_k \to \|A^T s\|_2^2 \quad (\mathcal{K} \ni k \to \infty).$$

Dies zeigt $A^T s = 0$, also $s^T M s > 0$. Daher gibt es aus Stetigkeitsgründen $k \in \mathcal{K}$ mit $s_k^T M s_k > 0$ und wir erhalten den Widerspruch

$$0 \geq s_k^T(M + kAA^T)s_k = s_k^T M s_k + k\|A^T s_k\|_2^2 > 0.$$

Also gibt es doch ein solches $\bar{\rho}$. □

Beweis: des Satzes:

Wir haben

$$\begin{aligned}
\nabla_x L_\rho(\bar{x}, \bar{\mu}) &= \nabla_x L(\bar{x}, \bar{\mu}) + \rho \nabla h(\bar{x}) h(\bar{x}) = 0, \\
\nabla_{xx}^2 L_\rho(\bar{x}, \bar{\mu}) &= \nabla_{xx}^2 L(\bar{x}, \bar{\mu}) + \rho \nabla h(\bar{x}) \nabla h(\bar{x})^T + \rho \sum_{i=1}^p h_i(\bar{x}) \nabla^2 h_i(\bar{x}) \\
&= \nabla_{xx}^2 L(\bar{x}, \bar{\mu}) + \rho \nabla h(\bar{x}) \nabla h(\bar{x})^T.
\end{aligned}$$

Wegen der hinreichenden Bedingung 2. Ordnung ist Lemma 3.9.3 anwendbar mit $M = \nabla_{xx}^2 L(\bar{x}, \bar{\mu})$ und $A = \nabla h(\bar{x})$. Daher gibt es $\bar{\rho} > 0$, so daß $\nabla_{xx}^2 L_\rho(\bar{x}, \bar{\mu})$ positiv definit ist für alle $\rho \geq \bar{\rho}$.

Weiter gilt $\nabla_x L_\rho(\bar{x}, \bar{\mu}) = 0$, und somit sind die hinreichenden Bedingungen 2. Ordnung aus Satz 2.1.5 erfüllt. Der Punkt $\bar{x}$ ist daher ein isoliertes lokales Minimum von $L_\rho(\cdot, \bar{\mu})$.

$\square$

Natürlich ist der Lagrange-Multiplikator $\bar{\mu}$ nicht bekannt, so daß wir diesen geeignet approximieren müssen. Es gilt:

Satz 3.9.4 *Seien f und h zweimal stetig differenzierbar und $(\bar{x}, \bar{\mu})$ sei ein Kuhn-Tucker-Paar für (NLPG), in dem hinreichende Bedingungen zweiter Ordnung gelten. Ist dann $\bar{\rho} > 0$ hinreichend groß und $\rho \geq \bar{\rho}$ beliebig, so gibt es $\delta > 0$ und $\varepsilon > 0$, so daß die Funktion $L_\rho(\cdot, \mu)$ für alle $\mu \in B_\delta(\bar{\mu})$ ein eindeutiges lokales Minimum $x_\rho(\mu)$ auf $B_\varepsilon(\bar{x})$ besitzt. Die Funktion $\mu \mapsto x_\rho(\mu)$ ist stetig differenzierbar mit $x_\rho(\bar{\mu}) = \bar{x}$.*

Beweis:

Gemäß Satz 3.9.2 gibt es $\bar{\rho} > 0$, so daß für alle $\rho \geq \bar{\rho}$ gilt:

$$\nabla_x L_\rho(\bar{x}, \bar{\mu}) = 0, \quad \nabla^2_{xx} L_\rho(\bar{x}, \bar{\mu}) \text{ positiv definit.}$$

Anwenden des Satzes über implizite Funktionen auf $\nabla_x L_\rho(x, \mu) = 0$ liefert nun $\delta, \varepsilon > 0$ und eine C^1-Funktion $x_\rho : B_\delta(\bar{\mu}) \to B_\varepsilon(\bar{x})$ mit $x(\bar{\mu}) = \bar{x}$, so daß $x = x(\mu)$ für alle $\mu \in B_\delta(\bar{x})$ die einzige Lösung von $\nabla_x L_\rho(x, \mu) = 0$ ist. Durch Reduzieren von δ (falls nötig) erzielen wir wegen $\nabla^2_{xx} L_\rho(x(\mu), \mu) \to \nabla^2_{xx} L_\rho(\bar{x}, \bar{\mu})$ weiter, daß $\nabla^2_{xx} L_\rho(x(\mu), \mu)$ positiv definit ist für alle $\mu \in B_\delta(\bar{x})$. Daher ist $x(\mu)$ ein lokales Minimum von $L_\rho(\cdot, \mu)$, wie gewünscht. $\square$

Unsere Strategie ist nun die folgende:

- Wir wenden das Penalty-Verfahren auf das Problem (3.29) an, wobei es genügt, geeignete lokale Lösungen zu berechnen.
- Nach Lösen jedes Teilproblems wird μ_k aktualisiert und eventuell ρ_k erhöht.
- Ist $(\bar{x}, \bar{\mu})$ ein KT-Paar, in dem hinreichende Bedingungen zweiter Ordnung gelten, ist ρ_k hinreichend groß und liegt (x_k, μ_k) hinreichend nahe bei $(\bar{x}, \bar{\mu})$, so ist das x_k nächstgelegene lokale Minimum von $L_{\rho_k}(\cdot, \mu_k)$ gegeben durch $x_{k+1} = x_{\rho_k}(\mu_k)$.
- Das Kriterium zur Erhöhung von ρ_k wird so gewählt, daß für μ_k nahe $\bar{\mu}$ und $x_{k+1} = x_{\rho_k}(\mu_k)$ die Wahl $\rho_{k+1} = \rho_k$ getroffen wird, falls ρ_k hinreichend groß ist. Dies bedeutet, daß der Penalty-Parameter schließlich nicht mehr erhöht wird.

Wir müssen uns noch eine Strategie für den Update des Multiplikators überlegen:

Ist x_{k+1} ein lokales Minimum von $L_{\rho_k}(\cdot, \mu_k)$, so gilt:

$$\begin{aligned} 0 &= \nabla_x L_{\rho_k}(x_{k+1}, \mu_k) = \nabla f(x_{k+1}) + \nabla h(x_{k+1})(\mu_k + \rho_k h(x_{k+1})) \\ &= \nabla_x L(x_{k+1}, \mu_k + \rho_k h(x_{k+1})). \end{aligned}$$

Sei nun $\bar{x}$ eine lokale Lösung von (NLPG), in der $\nabla h(\bar{x})$ vollen Spaltenrang hat (dies ist eine CQ). Dann liefert der Kuhn-Tucker Satz einen (eindeutigen) Lagrange-Multiplikator $\bar{\mu}$ mit

$$0 = \nabla_x L(\bar{x}, \bar{\mu}) = \nabla f(\bar{x}) + \nabla h(\bar{x})\bar{\mu}.$$

Es liegt daher nahe, folgenden Update des Lagrange-Multiplikators zu wählen:

Update des Lagrange-Multiplikators:

$$\mu_{k+1} = \mu_k + \rho_k h(x_{k+1}).$$

Wir erhalten das folgende Verfahren:

Algorithmus 23 (Augmented Lagrange Verfahren)
Wähle $\rho_0 > 0$, $x_0 \in \mathbb{R}^n$ und $\mu_0 \in \mathbb{R}^p$, $\gamma \in (0,1)$, $\beta > 1$, $\Lambda \gg 0$.

Für $k = 0, 1, 2, \ldots$:

1. Falls (x_k, μ_k) ein KT-Paar für (NLPG) ist, STOP.
2. Bestimme die x_k nächstgelegene lokale Lösung x_{k+1} von

   $$\min_{x \in \mathbb{R}^n} L_{\rho_k}(x, \mu_k).$$

   Falls dies nicht möglich ist, setze $\rho_k := \beta\rho_k$, und wiederhole Schritt 2.
3. Setze $\mu_{k+1} = \begin{cases} \mu_k + \rho_k h(x_{k+1}) & \text{falls } \|\mu_k + \rho_k h(x_{k+1})\|_2 \le \Lambda, \\ \mu_k & \text{sonst.} \end{cases}$
4. Setze $\rho_{k+1} = \begin{cases} \beta\rho_k & \text{falls } \|h(x_{k+1})\|_2 > \gamma\|h(x_k)\|_2, \\ \rho_k & \text{sonst.} \end{cases}$

3.9.2 Globale Konvergenz

Satz 3.9.5 *Seien f und h stetig differenzierbar. Algorithmus 23 erzeuge eine Folge (x_k) mit einem Häufungspunkt $\bar{x}$, in dem $\nabla h(\bar{x})$ Rang p hat. Dann ist $\bar{x}$ ein Kuhn-Tucker-Punkt von (NLPG) mit zugehörigem Lagrange-Multiplikator*

$$\bar{\mu} = -(\nabla h(\bar{x})^T \nabla h(\bar{x}))^{-1} \nabla h(\bar{x})^T \nabla f(\bar{x}).$$

Ist darüber hinaus $(x_k)_\mathcal{K}$ eine Teilfolge mit $(x_k)_\mathcal{K} \to \bar{x}$ und gilt $\|\bar{\mu}\|_2 < \Lambda$, so folgt weiter:

$$(\mu_k)_\mathcal{K} \to \bar{\mu}.$$

Beweis:

Sei $(x_k)_\mathcal{K}$ eine gegen $\bar{x}$ konvergente Teilfolge. Mit $\tilde{\mu}_k = \mu_{k-1} + \rho_{k-1} h(x_k)$ haben wir

$0 = \nabla_x L_{\rho_{k-1}}(x_k, \mu_{k-1}) = \nabla f(x_k) + \nabla h(x_k)(\mu_{k-1} + \rho_{k-1} h(x_k)) = \nabla f(x_k) + \nabla h(x_k)\tilde{\mu}_k.$

Dies ergibt, da $\nabla h(x_k)^T \nabla h(x_k)$ für große $k \in \mathcal{K}$ invertierbar ist,

$\tilde{\mu}_k = -(\nabla h(x_k)^T \nabla h(x_k))^{-1} \nabla h(x_k)^T \nabla f(x_k) \to -(\nabla h(\bar{x})^T \nabla h(\bar{x}))^{-1} \nabla h(\bar{x})^T \nabla f(\bar{x}) = \bar{\mu}.$

Wir erhalten:

$$\nabla_x L(\bar{x}, \bar{\mu}) = \lim_{\mathcal{K} \ni k \to \infty} \nabla_x L(x_k, \tilde{\mu}_k) = 0.$$

Strebt ρ_k nicht gegen $+\infty$, so gibt es $l \geq 0$ mit $\rho_k = \rho_l$ und $\|h(x_k)\|_2 \leq \gamma^{k-l} \|h(x_l)\|_2$ für alle $k \geq l$. Insbesondere ergibt sich $h(x_k) \to 0$ und daher $h(\bar{x}) = 0$.

Gilt andererseits $\rho_k \to \infty$, dann folgt wegen

$$\limsup_{\mathcal{K} \ni k \to \infty} \|h(x_k)\|_2 = \limsup_{\mathcal{K} \ni k \to \infty} \frac{\|\tilde{\mu}_k - \mu_{k-1}\|_2}{\rho_{k-1}} \leq (\|\bar{\mu}\|_2 + \Lambda) \lim_{\mathcal{K} \ni k \to \infty} \frac{1}{\rho_{k-1}} = 0$$

wiederum $h(\bar{x}) = 0$. Also ist $(\bar{x}, \bar{\mu})$ ein KT-Paar.

Ist $\Lambda > \|\bar{\mu}\|_2$, so gilt für große $k \in \mathcal{K}$:

$$\mu_k = \tilde{\mu}_k = -(\nabla h(x_k)^T \nabla h(x_k))^{-1} \nabla f(x_k) \to \bar{\mu}.$$

□

3.9.3 Lokale Konvergenz

Wir wenden den folgenden Satz aus [Ber82, S. 108] an.

Satz 3.9.6 *Seien f und h zweimal stetig differenzierbar und $(\bar{x}, \bar{\mu})$ ein KT-Punkt von (NLPG), in dem $\nabla h(\bar{x})$ Rang p hat und hinreichende Bedingungen 2. Ordnung gelten. Weiter sei $\bar{\rho} > 0$ so, daß $\nabla^2_{xx} L_{\bar{\rho}}(\bar{x}, \bar{\mu})$ positiv definit ist.*

Dann gibt es $\delta, \varepsilon > 0$ und $C > 0$, so daß für alle $(\mu, \rho) \in D$ mit

$$D = \{(\mu, \rho) \; : \; \|\mu - \bar{\mu}\|_2 < \delta\rho, \; \rho \geq \bar{\rho}\}$$

das Problem

$$\min_{x \in \mathbb{R}^n} \{L_\rho(x, \mu) \; : \; \|x - \bar{x}\|_2 < \varepsilon\}$$

eine eindeutige Lösung $x(\mu, \rho)$ besitzt. Die Funktion $x(\mu, \rho)$ ist auf dem Inneren von D stetig differenzierbar, und für alle $(\mu, \rho) \in D$ gilt:

$$\|x(\mu, \rho) - \bar{x}\|_2 \leq \frac{C}{\rho} \|\mu - \bar{\mu}\|_2, \quad \|\tilde{\mu}(\mu, \rho) - \bar{\mu}\|_2 \leq \frac{C}{\rho} \|\mu - \bar{\mu}\|_2$$

mit $\tilde{\mu}(\mu,\rho) = \mu + \rho h(x(\mu,\rho))$*. Weiter gilt für alle* $(\mu,\rho) \in D$*:*

$$\text{Rang } \nabla h(x(\mu,\rho)) = p, \quad \nabla^2_{xx} L_\rho(x(\mu,\rho),\mu) \text{ positiv definit.}$$

Beispiel 3.9.1 Betrachte

$$\min \{f(x) \,:\, h(x) = 0\}$$

mit $f(x) = \frac{1}{2}x_1^2 - \frac{c}{2}x_2^2 + x_2$, $c > 0$, und $h(x) = x_2$. Die optimale Lösung ist $\bar{x} = 0$, der zugehörige Multiplikator $\bar{\mu} = -1$. Nun gilt

$$L_\rho(x,\mu) = \frac{1}{2}x_1^2 - \frac{c}{2}x_2^2 + x_2 + \mu x_2 + \frac{\rho}{2}x_2^2.$$

Somit

$$\nabla_x L_\rho(x,\mu) = \begin{pmatrix} x_1 \\ 1 + \mu + (\rho - c)x_2 \end{pmatrix}, \quad \nabla^2_{xx} L_\rho(x,\mu) = \begin{pmatrix} 1 & 0 \\ 0 & \rho - c \end{pmatrix}$$

Die Matrix $\nabla^2_{xx} L_\rho(x,\mu)$ ist positiv definit genau dann, wenn $\rho > c$ gilt. Für $\rho < c$ hat $L_\rho(\cdot,\mu)$ kein lokales Minimum. Für $\rho > c$ ist $L_\rho(\cdot,\mu)$ quadratisch und streng konvex. Das eindeutige lokale (und globale) Minimum ist

$$x(\mu,\rho) = \begin{pmatrix} 0 \\ \frac{1+\mu}{c-\rho} \end{pmatrix}.$$

Nun gilt

$$\|x(\mu,\rho) - \bar{x}\|_2 = \frac{|1+\mu|}{\rho - c} = \frac{|\mu - \bar{\mu}|}{\rho - c} = \frac{\frac{\rho}{\rho - c}}{\rho}|\mu - \bar{\mu}|.$$

Weiter gilt

$$\tilde{\mu}(\mu,\rho) = \mu + \rho h(x(\mu,\rho)) = \mu + \rho\frac{1+\mu}{c-\rho} = \frac{c\mu - \rho\mu + \rho + \rho\mu}{c - \rho} = \frac{c\mu + \rho}{c - \rho}.$$

Daher:

$$|\tilde{\mu}(\mu,\rho) - \bar{\mu}| = \left|\frac{c\mu + \rho}{c - \rho} + 1\right| = \left|\frac{c\mu + c}{c - \rho}\right| = \frac{c}{\rho - c}|\mu - \bar{\mu}| = \frac{\frac{c\rho}{\rho - c}}{\rho}|\mu - \bar{\mu}|.$$

Für $\bar{\rho} > c$ können wir also $C = \frac{\max\{c,1\}\bar{\rho}}{\bar{\rho} - c}$ wählen. Wir sehen auch, daß die Konvergenzordnung in Satz 3.9.6 nicht verbessert werden kann.

Satz 3.9.7 *Seien* f *und* h *zweimal stetig differenzierbar. Algorithmus 23 erzeuge eine Folge* (x_k) *mit einem Häufungspunkt* $\bar{x}$*, in dem* $\nabla h(\bar{x})$ *Rang* p *hat. Dann gilt:*

a) *Der Punkt* $\bar{x}$ *ist ein Kuhn-Tucker-Punkt von (NLPG) mit zugehörigem Lagrange-Multiplikator*

$$\bar{\mu} = -(\nabla h(\bar{x})^T \nabla h(\bar{x}))^{-1} \nabla h(\bar{x})^T \nabla f(\bar{x}).$$

Sei nun weiter $\Lambda > \|\bar\mu\|_2$ und $(\bar x, \bar\mu)$ genüge der hinreichenden Bedingung 2. Ordnung. Dann gilt außerdem:

b) *Wird ρ_k hinreichend groß, so konvergiert die Folge (x_k, μ_k) gegen $(\bar x, \bar\mu)$ und es gibt $C > 0$, so daß für hinreichend große k gilt:*

$$\|\mu_{k+1} - \bar\mu\|_2 \le \frac{C}{\rho_k}\|\mu_k - \bar\mu\|_2,$$
$$\|x_{k+1} - \bar x\|_2 \le \frac{C}{\rho_k}\|\mu_k - \bar\mu\|_2,$$

c) *Es gibt $l \ge 0$ mit $\rho_k = \rho_l$ für alle $k \ge l$.*

Beweis: Die Aussage a) folgt aus Satz 3.9.5.

zu b): Wegen den Voraussetzungen an $(\bar x, \bar\mu)$ ist Satz 3.9.6 anwendbar. Sei $\rho^* > \max\{\bar\rho, C\}$. Gibt es nun k' mit $\rho_k \ge \rho^*$ für alle $k \ge k'$, dann haben wir mit $\rho = \delta\rho^*$:

$$(\mu_k, \rho_k) \in D \quad \forall\, k \ge k' \text{ mit } \|\mu_k - \bar\mu\|_2 < \rho.$$

Da $(\bar x, \bar\mu)$ ein Häufungspunkt von (x_k, μ_k) ist, gibt es $k \ge k'$ mit

$$\|\mu_k - \bar\mu\|_2 < \min\left\{\rho, \frac{\varepsilon}{3}, \Lambda - \|\bar\mu\|_2\right\}, \quad \|x_k - \bar\mu\|_2 < \frac{\varepsilon}{3}.$$

$x_{\rho_k}(\mu_k, \rho_k)$ ist dann das einzige lokale Minimum von $L_{\rho_k}(\cdot, \mu_k)$ in $B_\varepsilon(\bar x)$ und es gilt

$$\|x_{\rho_k}(\mu_k, \rho_k) - \bar x\|_2 \le \frac{C}{\rho_k}\|\mu_k - \bar\mu\|_2 \le \frac{C}{\rho^*}\|\mu_k - \bar\mu\|_2 < \frac{\varepsilon}{3}.$$

Somit ergibt sich

$$\|x_{\rho_k}(\mu_k, \rho_k) - x_k\|_2 \le \|x_{\rho_k}(\mu_k, \rho_k) - \bar x\|_2 + \|x_k - \bar x\|_2 < \frac{2\varepsilon}{3}.$$

Der Punkt x_{k+1} ist das x_k nächstgelegene lokale Minimum von $L_{\rho_k}(\cdot, \mu_k)$. Im Falle $x_{k+1} \ne x_{\rho_k}(\mu_k, \rho_k)$ würde dann $\|x_{k+1} - \bar x\|_2 \ge \varepsilon$ gelten, also

$$\|x_{k+1} - x_k\|_2 \ge \|x_{k+1} - \bar x\|_2 - \|x_k - \bar x\|_2 > \frac{2\varepsilon}{3}.$$

Daher gilt doch $x_{k+1} = x_{\rho_k}(\mu_k, \rho_k)$. Weiter haben wir

$$\|\tilde\mu(\mu_k, \rho_k) - \bar\mu\|_2 \le \frac{C}{\rho_k}\|\mu_k - \bar\mu\|_2 < \|\mu_k - \bar\mu\|_2 < \Lambda - \|\bar\mu\|_2$$

und daher

$$\|\tilde\mu(\mu_k, \rho_k)\|_2 < \Lambda.$$

Dies zeigt $\mu_{k+1} = \tilde{\mu}(\mu_k, \rho_k)$. Das neue Paar (x_{k+1}, μ_{k+1}) erfüllt wieder alle Voraussetzungen, um induktiv zeigen zu können, daß für alle $k \geq k'$ gilt:

$$\|\mu_{k+1} - \bar{\mu}\|_2 \leq \frac{C}{\rho_k}\|\mu_k - \bar{\mu}\|_2 \leq \underbrace{\frac{C}{\rho^*}}_{<1} \|\mu_k - \bar{\mu}\|_2 \to 0,$$

$$\|x_{k+1} - \bar{x}\|_2 \leq \frac{C}{\rho_k}\|\mu_k - \bar{\mu}\|_2 \leq \underbrace{\frac{C}{\rho^*}}_{<1} \|\mu_k - \bar{\mu}\|_2 \to 0.$$

c) 1. Fall: $\rho_k < \bar{\rho}$ für alle k.

Dann kann $\rho_{k+1} = \beta \rho_k$ nur für endlich viele k gelten und daher gibt es $l \geq 0$ mit $\rho_k = \rho_l$ für alle $k \geq l$.

2. Fall: Es gibt k mit $\rho_k \geq \bar{\rho}$.

Angenommen, (ρ_k) ist unbeschränkt. Dann können wir b) anwenden und erhalten $(x_k, \mu_k) \to (\bar{x}, \bar{\mu})$. Aus

$$\mu_{k+1} = \mu_k + \rho_k h(x_{k+1})$$

für große k folgt nun:

$$\rho_k \|h(x_{k+1})\|_2 = \|\mu_k - \mu_{k+1}\|_2 \leq \|\mu_k - \bar{\mu}\|_2 + \|\mu_{k+1} - \bar{\mu}\|_2 \leq \left(1 + \frac{C}{\rho_k}\right) \|\mu_k - \bar{\mu}\|_2$$

$$\rho_{k-1}\|h(x_k)\|_2 = \|\mu_{k-1} - \mu_k\|_2 \geq \|\mu_{k-1} - \bar{\mu}\|_2 - \|\mu_k - \bar{\mu}\|_2 \geq \left(\frac{\rho_{k-1}}{C} - 1\right) \|\mu_k - \bar{\mu}\|_2.$$

Dies zeigt

$$\|h(x_{k+1})\|_2 \leq \frac{\rho_{k-1}}{\rho_k} \frac{1 + \frac{C}{\rho_k}}{\frac{\rho_{k-1}}{C} - 1} \|h(x_k)\|_2 = \frac{C}{\rho_k} \frac{1 + \frac{C}{\rho_k}}{1 - \frac{C}{\rho_{k-1}}} \|h(x_k)\|_2 = O\left(\frac{1}{\rho_k}\right) \|h(x_k)\|_2$$

für $\rho_k \to \infty$. Für hinreichend großes k gilt daher stets

$$\|h(x_{k+1})\|_2 \leq \gamma \|h(x_k)\|_2$$

und somit $\rho_{k+1} = \rho_k$. Daher wird ρ_k schließlich nicht mehr erhöht, im Widerspruch zur Annahme. Die Beschränktheit von (ρ_k) ist somit nachgewiesen. $\Box$

Literaturverzeichnis

[Be82] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*, Academic Press, New York, NY (1982).

[Be99] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, 1999.

[Br92] D. Braess. *Finite Elemente*. Springer Verlag, Heidelberg, 1992.

[Bro70] C. G. Broyden. he convergence of a class of double rank minimization algorithms: 2 . The new algorithm. *J. Inst. Math. Appl.* 6 (1970), pp. 222–231.

[CGT00] A. R. Conn, N. I. M. Gould, Ph. L. Toint. *Trust-region methods*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2000.

[Da67] J. W. Daniel. The conjugate gradient method for linear and nonlinear operator equations. *SIAM J. Numer. Anal.* 4 (1967), pp. 10–26.

[Dav91] W. C. Davidon. Variable metric methods for minimization. *SIAM J. Optim.* 1 (1991), pp. 1– 17.

[DM74] J. E. Dennis, J. J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Math. Comp.* 28 (1974), pp. 549–560.

[Fl70] R. Fletcher. A new approach to variable metric algorithms. *Computer J.* 13 (1970), pp. 317– 322.

[FMC68] A. V. Fiacco, G. P. McCormick. *Nonlinear Programming: Sequential unconstrained minimization techniques* Wiley, New York, 1968.

[Fr55] K. R. Frisch. The logarithmic potential method for convex programming. Technical Report (unpublished manuscript), Institute of Economics, University of Oslo, Oslo, Norway, 1955.

[GK99] C. Geiger, Ch. Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer, Berlin, 1999.

[GK02] C. Geiger, Ch. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben.* Springer, Berlin, 2002.

[Go70] D. Goldfarb. A family of variable metric methods derived by variational means. *Math. Comp.* 24 (1970), pp. 23–26.

[Ha77] S. P. Han. A globally convergent method for nonlinear programming. *Journal of Optimization Theory and Applications*, 22 (1977), pp. 297–309.

[Ka84] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984.

[Ko93] P. Kosmol. *Methoden zur numerischen Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben*. Teubner, Stuttgart, 1993.

[NW99] J. Nocedal, S. J. Wright. *Numerical Optimization*. Springer Verlag, New York, 1999.

[Po78] M. J. D. Powell. A fast algorithm for nonlinearly constrained optimization calculations. *Lect. Notes Math.* 630 (1978), 144-157.

[Po78a] M. J. D. Powell. The convergence of variable metric methods for nonlinearly constrained optimization calculations. In: *Nonlinear programming 3*, Proc. Symp., Madison/Wis. 1977, 27-61, 1978.

[Sh70] D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Math. Comp.* 24 (1970), pp. 647–650.

[St83] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM J. Numer. Anal.* 20 (1983), pp. 190–212.