Einführung in die nichtlineare Optimierung

Prof. Dr. Walter Alt

Semester: SS 2010

Vorwort

Dieses Dokument wurde als Skript für die auf der Titelseite genannte Vorlesung erstellt und wird jetzt im Rahmen des Projekts „Vorlesungsskripte der Fakultät für Mathematik und Informatik“ weiter betreut. Das Dokument wurde nach bestem Wissen und Gewissen angefertigt. Dennoch garantiert weder der auf der Titelseite genannte Dozent, die Personen, die an dem Dokument mitgewirkt haben, noch die Mitglieder des Projekts für dessen Fehlerfreiheit. Für etwaige Fehler und dessen Folgen wird von keiner der genannten Personen eine Haftung übernommen. Es steht jeder Person frei, dieses Dokument zu lesen, zu verändern oder auf anderen Medien verfügbar zu machen, solange ein Verweis auf die Internetadresse des Projekts `http://uni-skripte.lug-jena.de/` *enthalten ist.*

Diese Ausgabe trägt die Versionsnummer und ist vom 10. Februar 2011. Eine neue Ausgabe könnte auf der Webseite des Projekts verfügbar sein.

Jeder ist dazu aufgerufen, Verbesserungen, Erweiterungen und Fehlerkorrekturen für das Skript einzureichen bzw. zu melden oder diese selbst einzupflegen – einfach eine E-Mail an die Mailingliste `<uni-skripte@lug-jena.de>` *senden. Weitere Informationen sind unter der oben genannten Internetadresse verfügbar.*

Hiermit möchten wir allen Personen, die an diesem Skript mitgewirkt haben, vielmals danken:

- *Konrad Kaffka* `<kokaffka@gmx.de>` *(2011)*
- *Jens Kubieziel* `<jens@kubieziel.de>` *(2011)*

Inhaltsverzeichnis

1 Optimierungsaufgaben

1.1 Aufgabenstellung und Beispiele

Gegeben ist eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$ ges: Minimum von f

Beispiel $f\colon \mathbb{R} \to \mathbb{R}; f(x) = x^2$
$f\colon \mathbb{R} \to \mathbb{R}; f(x) = x$ keine Lösung (mit $x \geq 0$ aber schon)

Beispiel (Bild von Folie zu finden im Netz) mit $f\colon \mathbb{R}^2 \to \mathbb{R}; f(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2) - \cos(x_1^2) - \cos(x_2^2)$ Viele lokale Minima ein globales Minimum $(\cos + \frac{1}{2}x)$ für Verfahren schwer zu berechnen. Verfahren kann eine Funktion nur lokal betrachten.
$f(x) = \sin(x)$ hat viele globale Minima.

Bezeichnungen

Vektoren sind Spaltenvektoren mit Komponenten x_i

$x^1, \dots, x^k, \{x^k\}_{k\in\mathbb{N}}$ euklidische Norm: $\|x\| = \sqrt{\sum(x_i^2}$ $B(x, r)$ offene Kugel, $\overline{B}(x, r)$ abgeschlossene Kugel

Allgemeine Aufgabenstellung $D \subset \mathbb{R}^n$ offene Menge (Definitionsbereich), $f\colon D \to \mathbb{R}$ Zielfunktion, $F \subset D$ Menge der zulässigen Punkte
(P) $\min\limits_{x\in F} f(x); D = \mathbb{R} = F$ unrestringiertes Problem; F: wird durch Restriktionen definiert, z. B.: $F = \{x \in \mathbb{R}^n | x \geq 0; Ax = b\}$

Bemerkung $g\colon D \to \mathbb{R}^n, \max_x \in F g(x) \Leftrightarrow \min_{x\in F} -g(x)$

Beispiel $f(x) = x^3 \min f(x); Nb.x \geq 1$

lokales Minimum (zu Aufgabenstellung (P)) ein Punkt $x \in F$ heißt lokaler Minimalpunkt von f auf F oder lokale Lösung von (P), wenn es ein $r > 0$ gibt, mit $f(x) \geq f(\tilde{x}) \forall x \in F \Cap B(\tilde{x}, r)$ Ein Punkt $\tilde{x} \in F$ heißt strikter lokaler Minimalpunkt von f auf F oder strikt lokale Lösung von (P), wenn es ein $r > 0$ gibt mit $f(x) > f(\tilde{x}) \forall x F \Cap B(\tilde{x}, r), x \neq \tilde{x}$
Ein Punkt $\tilde{x} \in F$ heißt globaler Minimalpunkt von f auf F, wenn gilt $f(x) \geq f(\tilde{x}) \forall x \in F$
Ei Punkt $\tilde{x} \in F$ heißt strikter globaler Minimalpunkt von f auf F wenn gilt $f(x) > f(\tilde{x}) \forall x \in F, x \neq \tilde{x}$.
Ist $\tilde{x}$ lokale oder globale Lösung von (P), dann heißt $f(\tilde{x})$ lokaler bzw. globaler Optimalwert.

Bemerkung $min_{x\in F} f(x); min_{x\in F} f(x) + cc \in \mathbb{R}$ beliebig hat gleiche Stelle unterschiedlichen Optimalwert.

Spezialfall quadratischen Funktionen. H symm. $n \times n$-Matrix, $b \in \mathbb{R}^n, f\colon \mathbb{R}^n \to \mathbb{R}$ ist definiert durch $f(x) = \frac{1}{2}x^T Hx + b^T x$ (eindim $\frac{1}{2}x^2 + bx$)
$x^T Hx > 0, \forall x \in \mathbb{R}^n x \neq 0_n$ positiv definit, auch wenn $\exists \alpha > 0\colon x^T Hx \geq \alpha \|x\|^2 \forall x \in \mathbb{R}^n$
$\nabla f(x) = (\frac{\delta f(x)}{\delta x_1} \dots \frac{\delta f(x)}{\delta x_n}) = f'(x)$.

Beispiel Lineare Regression. Geg: Messwerte$(\xi_i, \eta_i), i = 1, \dots, m$, Zusammenhang zwischen ξ und η-Werten $\eta(\xi) = g(\xi; x_1, x_2) = x_1\xi + x_2$
Wir definieren $f\colon \mathbb{R}^2 \to \mathbb{R}$ durch

$$\begin{aligned}
f(x) = f(x_1, x_2) &= \sum_{i=1}^{m} (g(\xi_i; x_1, x_2) - \eta_i)^2 \\
&= \sum (x_1\xi_i + x_2 - \eta_i)^2 \\
&= x_1^2 \sum \xi_i + mx_2^2 + 2x_1x_2 \sum \xi_i - 2x_1 \sum \xi_i\eta_i - 2x_2 \sum \eta_i + \sum \eta_i^2
\end{aligned}$$

Man bestimmt $\tilde{x} = (\tilde{x}_1, \tilde{x}_2)$ als Minimum von f. f ist vom Typ $f(x) = \frac{1}{2}x^T Hx + b^T x; H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}H_{22} & \end{pmatrix}$

1.2 Existenz von Lösungen

$f\colon D \to \mathbb{R}, D \subset \mathbb{R}^n$ offen, $K \subset D$ kompakt, f stetig auf D Satz von Weierstraß

Def Sei $f\colon D \to \mathbb{R}, D \subset \mathbb{R}^n, \alpha \in \mathbb{R}$. Die Mengen $N(f, x) = \{x \in D | f(x) \leq \alpha\}$ heißen **Niveau-Mengen** von f.

Satz (1.1.2)**.** *Sei $D \subset \mathbb{R}^n$. Die Funktion $f\colon D \to \mathbb{R}$ sei stetig und für ein $\omega \in D$ sei die Niveaumenge $N(f, f(\omega)) = \{x \in D | f(x) \leq f(\omega)\}$ kompakt. Dann gibt es (mindestens) ein globales Minimum von f auf D.*

Beweis. Nach Satz von Weierstraß gibt es ein $\tilde{x} \in N(f, f(\omega))$ mit $f(\tilde{x}) \leq f(x) \forall x \in N(f, f(\omega))$. Für $x \notin N(f, f(\omega))$ gilt: $f(x) > f(\omega) \geq f(\tilde{x})$
$\Rightarrow f(\tilde{x}) \leq f(x) \forall x \in D$. □

2 Anwendungen

2.3 Anwendungen

Nichtlineare Regression (2.3.1)

3 Unrestringierte Optimierungsprobleme: Theorie

Ein Einschub mit Theorie aus 1.7

3.1 Optimalitätbedingungen

3.1.1 Notwendige Bedingungen 1. Ordnung

Satz (3.1.1)**.** *f in $x \in D$ differenzierbar, $\tilde{x}$ lokales Minimum, dann:*

$$\nabla f(\tilde{x})^T d \geq 0; \qquad \forall d \in \mathbb{R}^n \tag{1}$$

Bemerkung Ist in einem $x \in d$ mit $\nabla f(\tilde{x})^T d < 0$ dann d Abstiegsrichtung.

Aus Unterunterabschnitt 3.1.1 folgt $\nabla f(\tilde{x})^T d \geq -\nabla f(\tilde{x})^T d \geq 0 \qquad \forall d \in \mathbb{R}^n$
$\Rightarrow \nabla f(\tilde{x})^T d = 0 \Rightarrow \nabla f(\tilde{x}) = 0_n$.

Dann ist Unterunterabschnitt 3.1.1 äquivalent zu $\nabla f(\tilde{x}) = 0_n$ also ist folgender Satz äquivalent zu Unterunterabschnitt 3.1.1:

Satz (3.1.2)**.** *f sei in $\tilde{x}$ differenzierbar. Ist $\tilde{x}$ lokales Minimum von f, dann gilt*

$$\nabla f(\tilde{x}) = f'(\tilde{x}) = 0 \tag{3.1.2}$$

Beispiel 3.1.3 Wenn mind. zwei ξ_i-Werte verschieden sind, ist die Matrix positiv definit

Definition (3.1.4)**.** f sei in $\tilde{x} \in D$ differenzierbar. Ist $\nabla f(\tilde{x}) = 0_n$, dann heißt $\tilde{x}$ **stationärer Punkt** von f.

Bem Stationäre Punkte erfüllen die notwendige Optimalitätbedingungen aus Abschnitt 3.1.1 und sind damit „Kandidaten" für ein lokales Minimum.

3.1.2 Notwedige Bedingungen 2. Ordnung

Satz (3.1.6)**.** *f sei in einer Umgebung von $\tilde{x} \in D$ 2 mal stetig differenzierbar. Ist $\tilde{x}$ lokales Minimum von f, dann gilt Abschnitt 3.1.1 und*

$$x^T f''(\tilde{x}) \geq 0 \qquad \forall x \in \mathbb{R}^n \tag{3.1.3}$$

(d. h.: $f''(\tilde{x})$ ist positiv semidefinit)

3.1.3 Hinreichende Bedingung zweiter Ordnung

Satz (3.1.10)**.** *f in Umgebung von $\tilde{x} \in D$ 2mal stetig differenzierbar. Die notwendige Optimalitätbedingung aus Abschnitt 3.1.1 (=$\nabla f(\tilde{x}) = 0_n$) sei erfüllt, und mit einem $\delta > 0$ gelte $\forall z \in B(\tilde{x}, \delta)$:*

$$x^T f''(z) x \geq 0 \quad \forall x \in \mathbb{R}^n \tag{3.1.4}$$

(d. h. $f''(z)$ ist in einer Umgebung $B(\tilde{x}, \delta)$ von $\tilde{x}$ positiv semidefinit).

Dann ist $\tilde{x}$ lokales Minimum von f.

Satz (3.1.11)**.** *f sei in einer Umgebung von $\tilde{x} \in D$ zweimal stetig differenzierbar. Die notwendige Optimalitätsbedingung (Abschnitt 3.1.1), $\nabla f(\tilde{x}) = 0_n$, sei erfüllt, und $f''(\tilde{x})$ sei positiv definit, d. h. mit einem $\alpha > 0$ gilt:*

$$x^T f''(\tilde{x}) x \geq \alpha \|x\|^2 \quad \forall x \in \mathbb{R}^n. \tag{3.1.5}$$

Dann gibt es $r > 0, \quad \beta > 0$ mit:

$$f(x) \geq f(\tilde{x}) + \beta \|x - \tilde{x}\|^2 \quad \forall x \in B(\tilde{x}, r)$$

($\Rightarrow f(x) > f(\tilde{x}) \forall x \in B(\tilde{x}, r), x \neq \tilde{x}$)

d. h. $\tilde{x}$ ist striktes lokales Minimum von f.

3.2 Konvexe Optimierungsaufgaben

Eine konvexe Optimierungsaufgabe ist:

$D \subseteq \mathbb{R}^n$ offen, $F \subseteq D$ nichtleer und konvex, $f \colon D \to \mathbb{R}$ sei konvex auf F.

$$\min_{x \in F} f(x) \tag{P}$$

Satz (3.2.1)**.** *$D \subseteq \mathbb{R}^n$ offen $f \colon D \to \mathbb{R}$ sei differenzierbar, $F \subseteq D$ nichtleer und konvex. Dann gilt:*

f ist konvex auf $F \quad \Leftrightarrow$

$$f(y) \geq f(x) + f'(x)(y - x) \qquad \forall x, y \in F \tag{3.2.1}$$

(die nummerierung hab ich auch vergessen)

Satz. *f zweimal stetig differenzierbar, für $\tilde{x} \in F$ gelte notwendige Optimalitätbedingung ($\nabla f(\tilde{x}) = 0_n$) und f'' positiv definit auf F d. h.*

$$\forall z \in F\colon \quad x^T f''(z) x > 0 \qquad \forall x \in \mathbb{R}^n, x \neq 0_n \tag{3.2.4}$$

das ist $\tilde{x}$ striktes globales Minimum von f auf F.

Definition. $D \subseteq \mathbb{R}^n, f\colon D \to \mathbb{R}, F \subseteq D$ nichtleer und konvex. Die Funktion f heißt **gleichmäßig konvex** auf F, wenn es ein $\alpha > 0$ gibt, mit

$$(1-t)f(x) + tf(y) \geq f((1-t)x + ty) + t(1-t)\alpha\|x-y\|^2 \qquad \forall x, y \in F \quad \forall f \in [0,1].$$

Satz. *$D \subseteq \mathbb{R}^n$ offen, $F \subseteq D$ nichtleer und konvex, $f\colon D \to \mathbb{R}$ sei differenzierbar. Dann ist f gleichmäßig konvex auf F genau dann wenn gilt:*

$$f(y) - f(x) \geq \nabla f(x)^T(y-x) + \alpha\|x-y\|^2 \quad \forall x, y \in F \tag{2}$$

Bemerkung Für ein festes $x \in \mathbb{R}^n$ ist $q\colon \mathbb{R}^n \to \mathbb{R}$ definiert durch $q(y) = f(x) + \nabla f(x)^t(x-y) + \alpha\|x-y\|^2$ eine quadratische Funktion. Der Gleichung 2 ist äquivalent zu $f(y) \geq q(y)$.

Es gilt $q(x) = f(x)$. $\nabla q(x) = \nabla f(x)$ $(\nabla q(x) = \nabla f(x) + 2\alpha(y-x))$. Das heißt es gibt eine quadratische Funktion die unter der eigentlichen Kurve liegt.

Satz. *Sei $D \subseteq \mathbb{R}^n$ offen, $F \subseteq D$ nichtleer und konvex. $f\colon D \to \mathbb{R}$ sei zweimal stetig differenzierbar. Ist f'' gleichmäßig positiv definit auf F, d. h. es gibt ein $\beta > 0$ (unabhängig von $x \in F$), so dass gilt:*

$$y^T f''(x) y \geq \beta\|y\|^2 \quad \forall y \in \mathbb{R}^n \tag{3}$$

dann ist f gleichmäßig konvex auf F. Ist F offen, dann folgt aus der gleichmäßigen Konvexität von f auf F auch Gleichung 3.

4 Unrestringierte Optimierungsprobleme: Verfahren

Betrachtet wird das Problem:

$$\begin{gathered} \min f(x) \\ f\colon \mathbb{R}^n \to \mathbb{R}(D = \mathbb{R}^n) \end{gathered} \tag{PU}$$

4.1 Grundlagen

Standardvorraussetzung:

Mit einem gegebenen $x^{(0)} \in \mathbb{R}^n$ ist die Niveaumenge

$N(f, f(x^{(0)})) = \{x \in \mathbb{R}^n | f(x) \leq f(x^{(0)})\}$ kompakt.

Die Zielfunktion f ist auf einer konvexen, offenen Obermenge D_0 von $N(f, f(x^{(0)}))$

Bemerkung Ist Unterabschnitt 4.1 erfüllt, dann gibt es nach Abschnitt 1.2 mindestens eine lokale Lösung von (P), und für jede lokale Lösung $\tilde{x}$ gilt es nach Abschnitt 3.1.1:

$$\nabla f(\tilde{x}) = 0_n \tag{4}$$

Bemerkung benutzt man ein Nullstellenverfahren und eine Lösung von Gleichung 4 zu berechnen, dann geht die Information „Min“ verloren. Besser: Abstiegsverfahren, d. h. man berechnet ausgehen von einem Startpunkt $x^{(0)}$ eine Folge $\{x^{(k)}\}_{k\in\mathbb{N}}$ mit $f(x^{(k+1)}) < f(x^{(k)}, \quad k = 0, 1, 2, \ldots$.

Lemma (4.1.1)**.** *$f\colon \mathbb{R}^n \to \mathbb{R}$ sei differenzierbar in $x \in \mathbb{R}^n$. Weiter sei $d \in \mathbb{R}^n$ eine Richtung mit*

$$\nabla f(x)^T d < 0 \tag{4.1.3}$$

Dann gibt es ein $\bar{\sigma} > 0$ mit $f(x + \sigma d) < f(x) \quad \forall \sigma \in]0, \bar{\sigma}]$.

Definition (4.1.2)**.** Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ differenzierbar in $x \in \mathbb{R}^n$. Dann heißt $d \in \mathbb{R}^n$ **Abstiegsrichtung** von f in x, wenn (4.1.3) gilt.

Bemerkung Ist in $x \in \mathbb{R}^n$ die Optimalitätbedingung $\nabla f(x) = 0_n$ nicht erfüllt, (d. h. $\nabla f(x) \neq 0_n$) dann gibt es in x (mindestens) eine Abstiegsrichtung.

[Annahme. $\nabla f(x)^T d \geq 0 \quad \forall d \in \mathbb{R}^n \Rightarrow \nabla f(x) = 0_n$]

zu Beispiel $d = -f''(x)^{-1}\nabla f(x)$ Newton Richtung.

Verfahren 4.1.4

Allgemeines Konzept für ein Abstiegsverfahren mit Schrittweitensteuerung

1. Wähle Startpunkt $x^{(0)}$; Setzt $k = 0$
2. Ist $\nabla f(x^{(k)}) = 0_n$. STOPP
3. Berechne eine Abstiegsrichtung $d^{(k)}$ von f in $x^{(k)}$ und eine Schrittweite $\sigma_k > 0$ mit $f(x^{(k)} + \sigma_k d^{(k)} < f(x^{(k)})$ und setze $x^{(k-1)} := x^{(0)} + \sigma_k d^{(k)}$
4. Setze $k := k + 1$; gehe zu 2.

Bemerkung 4.1.5 Für Verfahren 4.1.4 gilt $f(x^{(k+1)}) < f(x^{(k)}) \quad \forall k = 0, 1, \ldots \Rightarrow f(x^{(k)}) < f(x^{(0)})$ d. h. $x^{(k)} \in N(f, f(x^{(0)}))$. Ist die Vorraussetzung 4.1.1 erfüllt, dann ist $N(f, f(x^{(0)}))$ beschränkt $\Rightarrow$ die Folge $\{x^{(k)}\}_{k\in\mathbb{N}}$ ist beschränkt. $\Rightarrow$ die Folge $\{f(x^{(k)})\}_{k\in\mathbb{N}}$ ist beschränkt.

zum Abstiegskriterium: Die Bedingung $\nabla f(x^{(k)}) = 0_n$ ist ein theoretisches Abbruchkriterium. Praktisch benutzt man „$\|\nabla f(x^{(k)}) < \varepsilon_1\|$“ mit einem kleinen $\varepsilon_1 > 0$ (z. B. 10^{-3}).

Andere Kriterien: $|f(x^{(k+1)}) - f(x^{(k)})| < \varepsilon_2 \qquad \|x^{(k+1)} - x^{(k)}\| = \sigma_k \|d^{(k)}\| < \varepsilon_3$

Wichtig: maximale Iterationszahl vorgeben.

4.2 Berechnung von Ableitungen

Ideal: Analytische Berechnung der Ableitungen. Alternativen: Symbolysche Berechnung von Ableitungen.

Automatisches (algorithmisches) Differenzieren: input: Methode die f(x) berechnet, Output: Methode die $\nabla f(x)$ berechnet.

Numerische Berechnung von Ableitungen: Einfachster Fall. Man ersetzt $\nabla f(x)$ durch $\frac{f(x+\varepsilon e^i)-f(x)}{\varepsilon}$ für ein kleines $\varepsilon > 0$.

Wichtig: berechnete Ableitung überprüfen!

Matlab: Parameter DerivativeCheck auf den Wert 'on' setzen.

4.3 Das Newtonverfahren

Gegeben eine Abbildung $F\colon \mathbb{R}^n \to \mathbb{R}^n$ Aufgabe: Berechne eine (lokale) Lösung $\tilde{x}$ von $F(x) = 0_n$. F sei differenzierbar.

Newtonverfahren

- wähle Startpunkt $x^{(0)}$
- Ist $x^{(k)}$ berechnet dann erhält man $x^{(k+1)}$ durch lösen des Gleichungssystems:

  $$F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)}) = 0_n \tag{4.3.1}$$

  Dann sollte $F'(x^{(k)})$ invertierbar sein. Dann ist:

  $$x^{(k+1)} = x^{(k)} - F'(x^{(k)})^{-1} F(x^{(k)}) \tag{4.3.2}$$

Lemma (4.3.1)**.** *Sei $D \subseteq \mathbb{R}^n$ offen, $C \subseteq D$ konvex, $F\colon D \to \mathbb{R}^n$ differenzierbar und F' sei Lipschitzstetig auf C, d. h. mit einer Konstante $L \geq 0$ gilt $\|F'(x) - F'(y)\| \leq L\|x - y\| \quad \forall x, y \in C$. Dann gilt: $\|F(x) - F(y) - F'(y)(x - y)\| \leq \frac{L}{2}\|x - y\|^2$*

Lemma (4.3.2)**.** *Störungslemma. A sei eine nichtsinguläre $n \times n$-Matrix. S sei eine weitere $n \times n$-Matrix mit $\|A^{-1}\| \cdot \|S\| < 1$. Dann ist auch $A + S$ invertierbar und es gilt: $\|(A + S)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|S\|}$*

Lemma (4.3.3)**.** *Sei $\tilde{x} \in \mathbb{R}^n, r > 0$, und $G : \bar{B}(\tilde{x}, r) \to \mathbb{R}^n$ sei eine Abbildung mit:*

a) $\tilde{x}$ ist ein Fixpunkt von G d. h. $G(\tilde{x}) = \tilde{x}$

b) G ist in $\tilde{x}$ kontrahierend, d. h. mit einem $0 \leq L < 1$ gilt: $\|G(x) - G(\tilde{x})\| \leq L\|x - \tilde{x}\| \quad \forall x \in \bar{B}(\tilde{x}, r)$.

Dann ist $\tilde{x}$ der einzige Fixpunkt von G in $\bar{B}(\tilde{x}, r)$. Weiter konvergiert die Folge $\{x^{(k)}\}_{k\in\mathbb{N}}$ definiert durch $x^{(k+1)} := G(x^{(k)}), k = 0, 1, \ldots$ (Fixpunktiteration) für jeden Startpunkt $x^{(0)} \in \bar{B}(\tilde{x}, r)$ gegen $\tilde{x}$ mit:

$$\|x^{(k)} - \tilde{x}\| \leq L^k \|x^{(0)} - \tilde{x}\| \quad \forall k \in \mathbb{N} \tag{4.3.3}$$

Satz (4.3.4)**.** *Sei $D \subseteq \mathbb{R}^n$ offen,$F\colon D \to \mathbb{R}^n, \tilde{x} \in D$ mit $F(\tilde{x}) = 0_n$. Weiter gelte:*

a) F ist differenzierbar auf D, und F' ist Lipschitzstetig, d. h. $\|F'(x) - F'(y)\| \leq L\|x - y\| \quad \forall x, y \in D$. mit $L \geq 0$

b) $F'(\tilde{x})$ ist nichtnegativ

Dann gibt es ein $\delta > 0$ und ein $c \geq 0$, so dass das Newtonverfahren für jeden Startpunkt $x^{(0)} \in B(\tilde{x}, \delta)$ eine Folge $\{x^{(k)}\}$ definiert, die gegen $\tilde{x}$ konvergiert mit

$$\|x^{(k+1)} - \tilde{x}\| \leq c\|x^{(k)} - \tilde{x}\|^2 \quad , k = 0, 1, \ldots \tag{4.3.4}$$

4.4 Konstruktion von Abstiegsverfahren

4.4.1 Effiziente Schrittweiten

Ziel Bedingungen an Schrittweite und an Suchrichtung so dass:

$$\nabla f(x^{(k)}) \xrightarrow[k\to\infty]{} 0_n \tag{4.4.1}$$

Zunächt Bedingung an Schrittweite so dass:

$$\frac{f(x^{(k)})^T d^{(k)}}{\|d^{(k)}\|} \xrightarrow[k\to\infty]{} 0 \tag{4.4.2}$$

Im folgenden sei $x^{(k)}$ der aktuelle Iterationspunkt und $d^{(k)}$ die Suchrichtung und σ_k die Schrittweite. Erste Näherung:

$$f(x^{(k+1)}) - f(x^{(k)}) = f(x^{(k)} + \sigma_k d^{(k)}) - f(x^{(k)}) \tag{4.4.3}$$

Wegen (4.4.1) ist $\{f(x^{(k)})\}$ nach unten beschränkt, also konvergiert für $k \to \infty$ die linke Seite von (4.4.3) gegen 0.

Um auch auf der rechten Seite Konvergenz gegen 0 zu erreichen:

$$f(x^{(k)} + \sigma_k d^{(k)}) - f(x^{(k)}) \leq c_1 \sigma_k \nabla f(x^{(k)})^T d^{(k)} \tag{4.4.4}$$

mit Konstante c_1.

[(4.4.4) $\Leftrightarrow f(x^{(k+1)}) \leq f(x^{(k)} + c_1\sigma_k \overbrace{\nabla f(x^{(k)})d^{(k)})}^{<0}$]

Wenn (4.4.4) gilt dann folgt:

$$\sigma_k \nabla f(x^{(k)})d^{(k)} \xrightarrow[k\to\infty]{} 0$$

um hieraus Bedingung (4.4.2) zu folgern dürfen die σ_k im Vergleich zu $\nabla f(x^{(k)})d^{(k)}$ nicht zu schnell gegen 0 konvergieren:

$$\sigma_k \geq c_2 \frac{\nabla f(x^{(k)})^T d^{(k)}}{\|d^{(k)}\|^2}$$

(Bemerkung: in Beispiel 4.4.1 ist dieses Kriterium nicht erfüllt.)

Die Bedingung (4.4.4) und (4.4.5) implizieren:

$$f(x^{(k)} + \sigma_k d^{(k)}) \leq f(x^{(k)}) - c \left(\frac{\nabla f(x^{(k)})^T d^{(k)}}{\|d^{(k)}\|} \right)^2 \qquad c = c_1 c_2 \tag{4.4.6}$$

Definition (4.4.2)**.** Seien $x \in N(f, f(x^{(k)}))$ und $d \in \mathbb{R}^n$, $\quad \nabla f(x)^t d < 0$ eine Schrittweite $\sigma_k > 0$ erfüllt das Prinzip des **hinreichenden Abstiegs**, wenn gilt:

$$f(x + \sigma d) \leq f(x) + c_1 \sigma \nabla f(x)^T d \tag{4.4.7}$$

$$\sigma \geq -c_2 \frac{\nabla f(x^{(k)})^T d^{(k)}}{\|d^{(k)}\|^2} \qquad c_1, c_2 > 0 \tag{4.4.8}$$

Eine Schrittweite heißt effizient, wenn gilt:

$$f(x + \sigma d) \leq f(x) - c \left(\frac{\nabla f(x^{(k)})^T d^{(k)}}{\|d^{(k)}\|} \right)^2 \qquad c > 0 \tag{4.4.9}$$

Weitere Forderungen (zusätzlich zu (4.4.1)):

$$\begin{aligned} &f \text{ auf } N(f, f(x^{(k)})) \text{ Lipschitzstetig, d. h.:} \\ &\exists L > 0 : \quad \|f'(x) - f'(y)\| \leq L\|x - y\| \end{aligned} \tag{4.4.10}$$

Lemma (4.4.3)**.** *Vorraussetzungen (4.4.1), (4.4.10) seien erfüllt,* $x \in N(f, f(x^{(k)})), d \in \mathbb{R}^n, \nabla f(x)^T d < 0, \delta \in]0, 1[$, *dann* $\exists \tau = \tau(x, d, \delta)$ *mit:*

(i) $f(x + \sigma d) \leq f(x) + \delta\sigma \nabla f(x)^T d \quad \forall \sigma \in]0, \tau[$

(ii) $f(x + \tau d) = f(x) + \tau\sigma \nabla f(x)^T d$

(iii) $\tau \geq \varrho := -\frac{2(1-\sigma)}{L} (\frac{\nabla f(x^{(k)})^T d^{(k)}}{\|d^{(k)}\|^2}$

(iv) $\frac{d}{d\sigma} f(x + \sigma d) = \nabla f(x + \sigma d)^T d < \delta \nabla f^t d \quad \forall \sigma \in [0, \frac{\varrho}{2}]$

Bemerkung Für effiziente Schrittweiten gilt (4.4.2).

4.4.2 Gradientenbezogene Suchrichtungen

Sei $\beta_k := \frac{\nabla f(x^{(k)})^T d^{(k)}}{\|\nabla f(x^{(k)})^T\| \; \|d^{(k)}\|} = \cos(\angle(\nabla f(x^{(k)}), d^{(k)}))$

Dann gilt: $\frac{\nabla f(x^{(k)})^T d^{(k)}}{\|d^{(k)}\|} = \beta_k \|\nabla f(x^{(k)})\|$

Daher folgt (4.4.1) aus (4.4.2), wenn gilt $-\beta_k \geq c > 0$

Das bedeutet, dass der Winkel zwischen $\nabla f(x^{(k)})$ und $d^{(k)})$) gleichmäßig kleiner als 90° sein muss.

Definition (4.4.4)**.** Es seien $x \in N(f, f(x^{(k)}))$ und $d = d(x) \in \mathbb{R}^n$. Die Richtung d heißt **gradientenbezogen** in x wenn gilt:

$$-\nabla f(x^{(k)})^T d \geq c_3 \|\nabla f(x)^T\| \cdot \|d\| \quad c_3 > 0 \tag{4.4.11}$$

Richtung d heißt **streng gradientenbezogen** in x, wenn zusätzlich gilt:

$$c_4 \|\nabla f(x)\| \leq \|d\| \leq \frac{1}{c_4} \|\nabla f(x)\| \quad c_4 > 0 \tag{4.4.12}$$

Die Newton-Richtung $d^{(k)} = -f''(x^{(k)})^{-1} \nabla f(x^{(k)}$ ist streng gradientenbezogen unter der Vorraussetzung:

$$f\colon \mathbb{R}^n \to \mathbb{R} \text{ stetig}, x^{(0)} \in \mathbb{R}^n, D \supseteq N(f, f(x^{(k)})) \text{ nichtleer, offen, konvex,} \tag{4.4.13}$$

$$f \text{ sei zweimal stetig differenzierbar auf } D \text{ und mit Konstante } \alpha > 0 \text{ gelte:}$$

$$y^T f''(x) y \geq \alpha \|y\|^2 \quad \forall y \in D$$

(d. h. f'' ist gleichmäßig positiv definit auf D)

Bemerkung Sei A eine reelle symmetrische Matrix, dann hat A nur reelle Eigenwerte $\lambda_1 \leq \ldots \leq \lambda_n$

- A ist positiv semidefinit $\Leftrightarrow \lambda_1 \geq 0$
- A ist positiv definit $\Leftrightarrow \lambda_1 > 0$

Zu jeder reellen positiv definiten Matrix A gibt es eine reelle unitäre Matrix U ($U^{-1} = U^T,\ \|U\| = 1$) mit $U^T AU = U^{-1}AU = D = \operatorname{diag}(\lambda_1 \leq \ldots \leq \lambda_n)$

Damit gilt: $D^{-1} = \operatorname{diag}(\lambda_1^{-1} \leq \ldots \leq \lambda_n^{-1}), \quad A^{-1} = UD^{-1}U^T$.

Man definiert $D^{\frac{1}{2}} = \operatorname{diag}(\lambda_1^{\frac{1}{2}} \leq \ldots \leq \lambda_n^{\frac{1}{2}})$ und dadurch $A^{\frac{1}{2}} = UD^{\frac{1}{2}}U^T$. Dann gilt auch $A^{\frac{1}{2}} \cdot A^{\frac{1}{2}} = A$

Man bezeichnet $A^{\frac{1}{2}}$ als positive Quadratwurzel von A. (Entsprechend $A^{-\frac{1}{2}}$)

Lemma (4.4.6)**.** *Vorraussetzung (4.4.13) sei erfüllt. Dann ist f gleichmäßig konvex auf D, die Niveaumenge $N(f, f(x^{(0)}))$ ist konvex und kompakt und $\exists \alpha_2 > 0\colon \quad \forall x \in N(f, f(x^{(0)}))$:*

$$y^T f''(x)y \geq \alpha_2 \|y\|^2 \quad \forall y \in \mathbb{R}^n \tag{4.4.14}$$

und $\|f''(x)\| \leq \alpha_2$ gilt.

Weiter gilt mit $\beta_1 = \frac{1}{\alpha_2}, \quad \beta_2 = \frac{1}{\alpha_1}, \quad \forall x \in N(f, f(x^{(0)}))$:

$$\beta_1 \|y\|^2 \leq y^T f''(x)^{-1} y \leq \beta_2 \|y\|^2 \tag{4.4.15}$$

und $\|f''(x)^{-1}\| \leq \beta_2$.

bei Bleistiftkreuz in Mitschriften weiter machen.

4.5 Schrittweitenverfahren

4.5.1 Exakte Schrittweitenbestimmung

Satz (4.5.2)**.** *Vorraussetzungen (4.1.1), (4.4.10). Weiter seien $x \in N(f, f(x^{(0)})), \quad d \in \mathbb{R}$ mit $\nabla f(x)^T d < 0$*

Dann gilt für die exakte Schrittweite σ_E:

$$\sigma_E \leq -\frac{\nabla f(x)^T d}{L\|d\|^2} =: \tilde{\sigma} \quad \text{d. h. } \sigma_E \text{ erfüllt (4.4.8))} \tag{4.5.2}$$

wobei L die Lipschitz-Konstante von f' ist und

$$f(x + \sigma_E d) \leq f(x) - \frac{1}{2L}\left(\frac{\nabla f(x)^T d}{\|d\|^2}\right)^2 = f(x) + \frac{1}{2}\tilde{\sigma}\nabla f(x)^T d \tag{4.5.3}$$

d. h. σ_E ist effizient.

Bemerkung Aus dem Beweis folgt, dass $\tilde{\sigma}$ die Abstiegsbedingung (4.4.7) erfüllt. Nach (4.5.2) erfüllt $\tilde{\sigma}$ auch die Bedingung (4.4.8).
d. h. erfüllt die beiden Bedingungen für das Prinzip des hinreichenden Abstiegs.
$\Rightarrow \tilde{\sigma}$ ist effizient.

Sei (4.4.13) erfüllt. Dann gilt: $0 = \nabla f(x+\sigma_E d)^T d = \nabla f(x)^T d + (f(x+\sigma_E d) - \nabla f(x))^T d = \nabla f(x)^T d + \sigma_e \underbrace{d^T f''(x+\lambda\sigma_E d)d}_{\geq \alpha_1 \|d\|^2}$ mit $0 < \lambda < 1 \quad \Rightarrow$

$$\sigma_E = -\frac{\nabla f(x)^T d}{d^T f''(x+\lambda\sigma_e d)d} \leq -\frac{\nabla f(x)^T d}{\alpha_1 \|d\|^2} \leq \frac{L}{\alpha_1}\tilde{\sigma} \tag{4.5.4}$$

Mit (4.5.3) $\Rightarrow \sigma_E$ erfüllt (4.4.7)

Spezialfall f ist quadratische Funktion, d. h. $f(x) = \frac{1}{2}x^t Hx + b^t x$ mit $b \in \mathbb{R}^n, H$ symmetrische $n \times n$-Matrix. Dann ist $f''(x) = H \forall x \in \mathbb{R}^n$, Speziell ist in (4.5.4) $d^T f''(x + \lambda\sigma_E d)d = d^T Hd$.

H positiv definit $\Rightarrow d^T Hd > 0 \Rightarrow$

$$\sigma_E = --\frac{\nabla f(x)^T d}{d^T Hd} \tag{4.5.5}$$

d. h. bei quadratischer Zielfunktion kann man die exakte Schrittweite auch praktisch verwenden.

Die Suchrichtungen seien streng gradientenbezogen, dann gilt:
$\tilde{\sigma} = -\frac{\nabla f(x)^T d}{L\|d\|^2} \leq \frac{c_4 \|\nabla f(x)^T\| \cdot \|d\|}{L\|\nabla f(x)^T\| \cdot \|d\|} \leq \frac{c_4}{L}$

$\Rightarrow$ nimmt man als Schrittweitenfolge die exakten Schrittweiten $\Rightarrow$ Folge ist beschränkt.

4.5.2 Schrittweitenverfahren von Armijo

(Auch Goldstein-Armijo-Verfahren)

Das Verfahren versucht zu gegebenem $\delta \in]0,1[$ eine Schrittweite $\sigma = \sigma_A$ zu berechnen mit:

$$f(x+\sigma d) \leq f(x) + \delta\sigma\nabla f(x)^T d \tag{4.5.6}$$

und

$$\sigma \geq -c_2 \frac{\nabla f(x)^T d}{\|d\|^2} \tag{4.5.7}$$

mit einer von x unabhängigen Konstante c_2.

Verfahren 4.5.4 Armijo-Verfahren

Vorg.: $0 < \delta < 1, \quad \gamma > 0$ und $0 < \beta_1 \leq \beta_2 < 1$ (alle unabhängig von x und d)

1. Wähle eine Schrittweite $\sigma_0 \geq -\gamma \frac{\nabla f(x)^T d}{\|d\|^2}$, setze $j := 0$
2. Ist (4.5.6) erfüllt mit $\sigma = \sigma_j$, d. h. gilt:

   $f(x + \sigma_j d) \leq f(x) + \delta \sigma_j \nabla f(x)^T d$

   dann setze $\sigma_A = \sigma_j$ und stoppe das Verfahren.
3. Wähle $\sigma_{j+1} \in [\beta_1 \sigma_j, \beta_2 \sigma_j]$
4. Setze $j := j + 1$, gehe zu 2.

Satz (4.5.5)**.** *Die Vorraussetzungen (4.1.1), (4.4.10) seien erfüllt. Weiter sei $x \in N(f, f(x^{(0)}))$ und $d \in \mathbb{R}^n$ mit $\nabla f(x)^T d < 0$. Dann berechnet das Armijo-Verfahren nach endlich vielen Iterationen eine Schrittweite $\sigma = \sigma_A$, die (4.4.6), (4.5.7) erfüllt, also effizient ist.*

Bemerkung 4.5.6 Um die Beschränktheit der Folge $\{\sigma_k\}$ (mit σ_k - Armijo-Schrittweite) sicherzustellen, fordert man mit einem weiteren Parameter $\bar{\gamma} = \gamma$:

$$\sigma_0 \in \left[-\gamma \frac{\nabla f(x)^T d}{\|d\|^2}, -\bar{\gamma} \frac{\nabla f(x)^T d}{\|d\|^2} \right]. \tag{4.5.8}$$

Sind die Suchrichtungen streng gradientenbezogen, dann folgt:

$\sigma_0 \leq \bar{\gamma} \frac{\|f(x)\|}{\|d\|} \leq \frac{\bar{\gamma}}{c_4}$

Zur Wahl der Parameter beim Armijo-Verfahren:

- δ sollte „klein sein“: 0.01
- γ sollte so gewählt werden, dass die exakte Schrittweite als Startschrittweite σ_0 zulässig ist und dass $\sigma_0 = 1$ aks Startschrittweite zulässig ist.

  zu $\sigma_0 = 1$. Sind die Suchrichtungen streng gradientenbezogen, dann gilt: $-\nabla f(x)^T d \leq \|\nabla f(x)\| \cdot \|d\| \leq c_4 \|d\|^2$ und $-\nabla f(x)^T d \geq c_3 \|\nabla f(x)\| \cdot \|d\| \geq \frac{c_3}{c_4} \|d\|^2$

  $\Rightarrow -\frac{1}{c_4} \frac{\nabla f(x)^T d}{\|d\|^2} \leq 1 \leq -\frac{c_3}{c_4} \frac{\nabla f(x)^T d}{\|d\|^2}$ d. h. (4.5.8) ist erfüllt mit $\gamma = \frac{1}{c_4}, \quad \bar{\gamma} = \frac{c_4}{c_3}$. Daher wählt man γ „klein“ (10^{-4}) und $\bar{\gamma}$ „groß“(10^4).

  Zur exakten Schrittweite: man versucht eine Schätzung der exakten Schrittweite als σ_0 zu berechnen.

  Sind (4.4.1), (4.4.10) erfüllt, dann gilt (4.5.2), d. h. $\sigma_E \geq -\frac{1}{L} \frac{\nabla f(x)^T d}{\|d\|^2}$

  Sei (4.4.13) erfüllt, dann gilt (4.5.4), d. h. $\sigma_E = -\frac{\nabla f(x)^T d}{d^T f''(x + \lambda \sigma_e d) d}$ mit $0 < \lambda < 1$.

Weiter gilt: $f(x+d) - f(x) - \nabla f(x)^T d = \frac{1}{2} d^t f''(x + \lambda\theta_e d)d$ und $0 < \theta < 1$. Als Näherung für σ_E wählt man daher:

$$\sigma_0 = -\frac{\nabla f(x)^T d}{d^T f''(x + \lambda\theta_e d)d} = -\frac{\nabla f(x)^T d}{2(f(x+d) - f(x) - \nabla f(x)^T d)} \tag{4.5.9}$$

Dies ist eine Näherung für σ_E, wenn $\|d\|$ hinreichend klein ist. Mit (4.4.13) folgt wegen $\alpha_1 \|d\|^2 \le d^T f''(x + \theta d)d \le \alpha_2 \|d\|^2$:

$-\frac{1}{\alpha_2} \frac{\nabla f(x)^T d}{\|d\|^2} \le \sigma_0 \le \frac{1}{\alpha_1} \frac{\nabla f(x)^T d}{\|d\|^2}$ d. h. (4.5.8) ist mit $\gamma = \frac{1}{\alpha_2}$ und $\bar{\gamma} = \frac{1}{\alpha_1}$ erfüllt.

Praktisch wählt man mit vorgegebenen Konstanten $\gamma, \bar{\gamma}$ als Startschrittweite

$$\tilde{\sigma}_0 = \min\left\{ -\bar{\gamma} \frac{\nabla f(x)^T d}{\|d\|^2}, \max\left\{ -\gamma \frac{\nabla f(x)^T d}{\|d\|^2}, \sigma_0 \right\} \right\} \tag{4.5.10}$$

zu σ_0 aus (4.5.9): σ_0 ist das Minimum der PArabel $q(\sigma)$ mit $q(0) = f(x)$, $q'(0) = \nabla f(x)^T d$, $q(1) = f(x+d)$ (Setze die Funktion gleich einer Parabel bestimme das Minium und schätze auch dort das Minimum von f)

Zur Berechnung der Schrittweiten: $\sigma_{j+1} \in [\beta_1 \sigma_j, \beta_2 \sigma_j]$ mit $0 < \beta_1 \le \beta_2 < 1$.

Einfache Wahl: $\beta_1 = \beta_2 = \beta$ z. B. $\beta = \frac{1}{2}$.

Weitere Strategie: $\sigma_{j+1} = \max\{0.1\sigma_j, \sigma_j^*\}$ mit $\sigma_j^* = -\frac{\sigma_j^2 \nabla f(x)^T d}{2(f(x+\sigma_j d) - f(x) - \sigma_j \nabla f(x)^T d)}$. σ_j^* ist das Minimum der Parabel $q(\sigma)$ def durch $q(0) = f(x)$, $q'(0) = \nabla f(x)^T d$, $q(\sigma_j) = f(x + \sigma_j d)$.
Es gilt: $0.1\sigma_j < \sigma_{j+1} < \frac{0.5}{1-\delta}\sigma_j$. Falls $\delta < \frac{1}{2}$ gilt $0.5 < \frac{0.5}{1-\delta} < 1$.

Zusammenfassung:

Satz (4.5.11)**.** *Die Vorraussetzungen (4.1.1), (4.4.10) seien erfüllt. Für das allgemeine Abstiegsverfahren 4.1.4 gelte:*

$x^{(k)} \in N(f, f(x^{(0)}), \qquad \nabla f(x^{(k)})^T d < 0, \quad k = 0, 1, \ldots$

Werden die Schrittweiten σ_k als exakte Schrittweite oder als Armijo-Schrittweite gewählt, dann sind die Schrittweiten effizient. (Siehe Satz 4.5.2 und Satz 4.5.5)

4.6 Das Gradientenverfahren

in $x^{(k)}$ wählt man $-\nabla f(x^{(k)})$ als Suchrichtung. $\Rightarrow$ Suchrichtungen streng gradientenbezogen.

4.6.1 Richtung des steilsten Abstiegs

$f\colon \mathbb{R}^n \to \mathbb{R}$ sei differenzierbar. Im Punkt $x \in \mathbb{R}^n$ sei $\nabla f(x) \neq 0_n$.
Richtung des steilsten Abstiegs in Punkt x.

$$\min_{d \in \mathbb{R}^n} \nabla f(x)^t d \qquad \text{Nb.: } \|d\| = 1 \tag{4.6.1}$$

Lemma (4.6.1)**.** *Vorraussetzungen wie oben. Dann ist*

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

Lösung von (4.6.1)

Bemerkung Für das Problem

$$\min_{d \in \mathbb{R}^n} \nabla f(x)^t d \qquad \text{Nb.: } \|d\| = r \tag{4.6.2}$$

mit beliebigem $r > 0$ zeigt man analog: Lösung ist $r\bar{d}$.

Gradientenverfahren: Verfahren des steilsten Abstiegs

4.6.2 Das Verfahren

Wählt man vei allgemeinem Abstiegsverfahren 4.1.4 $d^{(k)} = -\nabla f(x^{(k)}) \Rightarrow$

Verfahren 4.6.2 - Gradientenverfahren

1. Wähle einen Startpunkt $x^{(0)}$, setze $k := 0$.
2. Ist $\nabla f(x^{(k)}) = 0_n$. STOPP
3. Benutze als Suchrichtung $d^{(k)} = -\nabla f(x^{(k)})$, berechne eine effiziente Schrittweite σ_k
   Exakte Schrittweite, Armijo-Schrittweie) und setze $x^{(k+1)} = x^{(k)} + \sigma_k d^{(k)}$.
4. Setze $k := k + 1$, gehe zu 2.

Abbruchkriterium: siehe Bemerkung 4.1.5

Allgemeine Konvergenzsätze: 4.4.9, 4.4.11, 4.4.12 anwendbar

Beispiel 4.6.3 Rosenbrockfunktion
Für beliebigen Startpunkt sind Vorr. (4.1.1) und (4.4.10) erfüllt. Benutzt man Armijo-Schrittweiten dann Schrittweiten effizient. Einziger stationärer Punkt ist $\tilde{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ Nach Satz 4.4.12 konvergiert die vom Gradientenverfahren berechnete Folge $\{x^{(k)}\}$ gegen $\tilde{x}$.

Ist (4.4.13) erfüllt, dann konvergiert die vom Gradientenverfahren berechnete Lösung R-linear gegen das eindeutig bestimmte, globale Minimum von f.

Frage: Kann man bessere Konvergenz erwarten? Antwort: Nein.

Benutzt man die exakte Schrittweite, dann gilt:

$0 = \frac{\partial}{\partial \sigma} f\left(x^{(k)} + \sigma d^{(k)}\right)|_{\sigma=\sigma_k} = \nabla f(x^{(k+1)})^T d^{(k)} = -(d^{(k+1)})^T d^{(k)}$, d. h. es gilt $d^{(k+1)} \perp d^{(k)} \quad \forall k \in \mathbb{N}$

Bemerkung $d^{(k)} = -\nabla f(x^{(k)})$ ist auch Lösung des quadratischen Optimierungsproblems

$$\min_{d \in \mathbb{R}^n} \nabla f(x^{(k)})^T d + \frac{1}{2} d^T d \tag{Q$_k$}$$

d. h. man kann die Wahl der Suchrichtung so interpretieren:
man ersetzt f lokal in $x^{(k)}$ durch die quadratische Funktion

$$\tilde{f}_k(x) = f(x^{(k)}) + \nabla f(x^{(k)})(x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^T (x - x^{(k)})$$

Für diese Funktion gilt: $\tilde{f}(x^{(k)}) = f(x^{(k)}), \qquad \nabla f(x^{(k)}) = \nabla f(x^{(k)})$. Weiter ist $x^{(k)} + d^{(k)}$ das eindeutig bestimmt Minimum von $\tilde{f}_k$.

$d^{(k)} = -f''(x^{(k)})^{-1} \nabla f(x^{(k)})$ Wobei man die Matrix $f''(x^{(k)})$ geeignet annähert und nicht berechnet.

4.6.3 Numerische Resultate

Matlab-Implementierung: fminunc

Zur Auswahl des Gradientenverfahrens muss man mit optimset die Parameter:

- „LargeScale“ auf „off“
- „HessUpdate“ auf „steepdesc“

weiter siehe Buch. Bilder 4.6 - 4.8 (Beispiel 4.6.11 zeigt das auch falsche Lösungen aus einem Computerverfahren kommen können, Beispiel 4.6.10 Lagerhaltungsproblem, Beispiel 4.6.12 Wolfe-Funktion.

4.7 Das gedämpfte Newtonverfahren

4.7.1 Das Verfahren

Beim algemeinen Abstiegsverfahren wählt man als Suchrichtung die Newton-Richtung $d^{(k)} = -f''(x^{(k)})^{-1}\nabla f(x^{(k)})$. Vorraussetzung (4.4.13) $\Rightarrow$ Richtungen sind streng gradientenbezogen.

Verfahren 4.7.1 Gedämpftes Newtonverfahren

1. Wähle Startpunkt $x^{(0)}$, setze $k := 0$
2. Ist $\nabla f(x^{(k)}) = 0$, dann STOPP
3. Berechne $d^{(k)}$ durch lösen des linearen Gleichungssystems $f''(x^{(k)})d = -\nabla f(x^{(k)})$ eine effiziente Schrittweite σ_k, setze $x^{(k+1)} = x^{(k)} + \sigma_k d^{(k)}$
4. Seite $k := k + 1$, gehe zu 2.

4.7.2 Richtung des steilsten Abstiegs

Ist A positiv definit, dann wird durch $\|x\|_A = \langle x, x\rangle_A^{\frac{1}{2}} = (x^T A x)^{\frac{1}{2}}$. Wir betrachten das Optimierungsproblem

$$\min \nabla f(x)^T d \qquad \text{Nb. } \|d\|_A = 1 \tag{4.7.1}$$

Lemma (4.7.2)**.** *$f\colon \mathbb{R}^n \to \mathbb{R}$ sei in x differenzierbar mit $\nabla f(x) \neq 0$. A sei symmetrisch positiv definite Matrix. Dann ist die Lösung von (4.7.1): $\bar{d} = -\frac{A^{-1}\nabla f(x)}{\|A^{-1}\nabla f(x)\|_A}$.*

Beweis. Sei $d \in \mathbb{R}^n$. Dann gilt: $\nabla f(x)^T d = \nabla f(x)^T A^{-1} A d = \langle A^{-1}\nabla f(x), Ad\rangle = \langle A^{-1}\nabla f(x), d\rangle_A \geq -\|A^{-1}\nabla f(x)\|_A \|d\|_A$ Für $d \in \mathbb{R}^n$ mit $\|d\|_A = 1 \Rightarrow$

$\nabla f(x)^T d \geq -\|A^{-1}\nabla f(x)\|_A$ Für $\bar{d}$ gilt: $\nabla f(x)^T \bar{d} = -\|A^{-1}\nabla f(x)\|_A$ $\square$

Sei speziell $A = f''(x^{(k)})$. Dann folgt aus dem Lemma: $\bar{d}^{(k)} = -\frac{f''(x^{(k)})^{-1}\nabla f(x^{(k)})}{\|f''(x^{(k)})^{-1}\nabla f(x^{(k)})\|_{f''(x^{(k)})}}$ ist die Richtung des steilsten Abstiegs in $x^{(k)}$ bezüglich der durch $f''(x^{(k)})$ definierten Norm.

Wichtig: In jedem Iterationsschritt benutzt man eine ander Matrix $A^{(k)} = f''(x^{(k)})$, die Informationen üver die Krümmung von f enthält. (Variable-Metrik-Verfahren)

4.7.3 Konvergenz des Verfahrens

Die Vorraussetzung (4.4.13) sei erfüllt. Daher sind die Newton-Richtungen für $x^{(k)} \in N(f, f(x^{(0)})$ Abstiegsrichtungen. Mit $x^{(k)} \in N(f, f(x^{(0)}) \Rightarrow x^{(1)} \in N(f, f(x^{(0)}) \Rightarrow \ldots \Rightarrow x^{(n)} \in N(f, f(x^{(0)}) \quad \forall k \in \mathbb{N}$. Allgemeiner Konvergenzsatz 4.4.14: Gedämpftes Newtonverfahrenkonvergiert R-linear (d. h. $x^{(k)} \to \tilde{x}R$-linear, wobei $\tilde{x}$ eindeutig bestimmt und globales Minimum von f ist), wenn die Schrittweitenfolge effizient ist, d. h. wenn man z. B. die Armijo-Schrittweiten benutzt.

Frage: kann man Konvergenz verbessern?
Antwort: ja, aber nur lokal, wenn man:

- beim Armijo-Verfahren die Schrittweite $\sigma_{k,0} = 1$ benutzt.
- die exakte Schrittweite benutzt
- beim Armijo-Verfahren als Startschrittweite eine Approximation der exakten Schrittweite benutzt

(In den letzten beiden Fällen gilt: $\sigma_k \to 1$ für $k \to \infty$)

Verwendung der Startschrittweite $\sigma_{k,0} = 1$

Ziel: Zeige, dass für hinreichend großes k (d. h. $x^{(k)}$ hinreichend nahe bei $\tilde{x}$) diese Schrittweite die Abstiegsbedingung in Schritt 2 des Armijo-Verfahrens erfüllt (dann ist die vom Armijo-Verfahren „berechnete“Schrittweite 1).

zusätzliche Vorraussetzungen

$$\exists L_2 > 0, \quad \exists r > 0\colon \quad \|f''(x) - f''(y)\| \leq L_2 \|x - y\| \quad \forall x, y \in B(\tilde{x}, r) \tag{4.7.2}$$

Bemerkung 4.7.3 Ist $x^{(k)} \neq \tilde{x}$, dann gilbt es ein $r > 0$ mit $B(\tilde{x}, r) \subseteq D$ und $f(x) < f(x^{(k)}) \quad \forall x \in B(\tilde{x}, r)$, d. h. $B(\tilde{x}, r) \subseteq N(f, f(x^{(k)}))$, d. h. $\tilde{x} \in \mathrm{int} N(f, f(x^{(k)}))$.

Es gilt:

$$-\nabla f(x^{(k)})^T d^{(k)} = -\nabla f(x^{(k)})^T f''(x^{(k)})^{-1} \nabla f(x^{(k)}) = \ldots \tag{4.7.3}$$

Satz (4.7.4)**.** *Vorraussetzung (4.4.13) sei erfüllt. (Dann hat f genau ein globales Minimum $\tilde{x}$ und die Folge $\{x_K\}$ konvergiert R-linear gegen $\tilde{x}$.)*

Beim gedämften Newtonverfahren sein die Schrittweiten σ_k die Armijo-Schrittweiten mit Startschrittweite $\sigma_{k,0} = 1$. Weiter sei $\delta \in]0, \frac{1}{2}[$. Dann ist $\sigma_k = 1$ für hinreichend großes k, d. h. für hinreichend großes k geht das gedämpfte Newtonverfahren in das gewöhnliche Newtonverfahren über und damit konvergiert es lokal superlinear oder quadratisch, wenn zusätzlich (4.7.2) gilt.

Beweis. Wie zeigen, dass es ein $k(\delta) \in \mathbb{N}$ gibt, so dass für $k \geq k(\delta)$ gilt:

$$f(x^{(k)} + d^{(k)}) \leq f(x^{(k)}) + \delta \nabla f(x)^T d^{(k)} \tag{1}$$

d. h. für die Startschrittweite $\sigma_{k,0} = 1$ ist das Abbruchkriterium des Armijo-Verfahrens erfüllt.

Nach Satz 4.4.14 filt $x^{(k)} \to \tilde{x}$ für $k \to \infty$. Weiter gilt $\nabla f(x^{(k)}) \to 0_n$ für $k \to \infty$. Nach Lemma 4.4.6 gilt:

$$\|d^{(k)}\| = \|f''(x^{(k)})^{-1} \nabla f(x^{(k)})\| = \underbrace{\|f''(x^{(k)})^{-1}\|}_{\leq \beta_2} \cdot \|\nabla f(x^{(k)})\|$$

$\Rightarrow d^{(k)} \to 0_n \Rightarrow x^{(k)} + d^{(k)} \to \tilde{x}$.

Nach Bemerkung 4.7.3 ist $\tilde{x}$ innerer Punkt von $N(f, f(x^{(0)}))$. Nach Lemma 4.4.6 ist $N(f, f(x^{(0)}))$ konvex $\Rightarrow [\tilde{x}, x^{(k)} + d^{(k)}] \subseteq N(f, f(x^{(0)}))$ für hinreichend großes k. Taylorentwicklung:

$f(x^{(k)} + d^{(k)}) = f(x^{(k)}) + \nabla f(x^{(k)})^T d^{(k)} + \frac{1}{2}(d^{(k)})^T f''(x^{(k)} + \theta_k d^{(k)}) d^{(k)}$ mit $\theta_k \in]0, 1[$.

$= f(x^{(k)}) + \frac{1}{2} \nabla f(x^{(k)})^T d^{(k)} + \frac{1}{2}(d^{(k)})^T \left[f''(x^{(k)} + \theta_k d^{(k)}) - f''(x^{(k)}\right] d^{(k)} \quad \Leftrightarrow$

$$-(\frac{1}{2} - \delta) \nabla f(x^{(k)})^T d^{(k)} \geq \frac{1}{2}(d^{(k)})^T \left[f''(x^{(k)} + \theta_k d^{(k)}) - f''(x^{(k)}\right] d^{(k)} \tag{2}$$

Wegen $\delta \in]0, \frac{1}{2}[$ ist $\frac{1}{2} - \delta > 0$. Mit Beispiel (4.4.8) folgt:

$$-(\frac{1}{2} - \delta) \nabla f(x^{(k)})^T d^{(k)} \geq (\frac{1}{2} - \delta) \beta_1 \|\nabla f(x^{(k)})\|^2 \geq (\frac{1}{2} - \delta) \beta_1 \beta_2^{-2} \|d^{(k)}\|^2. \tag{3}$$

Den Term auf der rechten Seite von (2) kann man wie folgt abschätzen

$\frac{1}{2}(d^{(k)})^T \left[f''(x^{(k)} + \theta_k d^{(k)}) - f''(x^{(k)}\right] d^{(k)} \leq \|f''(x^{(k)} + \theta_k d^{(k)}) - f''(x^{(k)})\| \cdot \|d^{(k)}\|^2$

Da f'' stetig ist und für $k \to \infty$ gilt: $x^{(k)} \to \tilde{x}, \quad x^{(k)} + \theta_k d^{(k)} \to \tilde{x} \Rightarrow r_k \to 0$.

Damit ist (2) erfüllt, wenn gilt:

$$(\frac{1}{2} - \delta) \beta_1 \beta_2^{-2} \|d^{(k)}\|^2 \geq r_k \|d^{(k)}\|^2 \tag{4}$$

d. h. wenn $r_k \leq (\frac{1}{2} - \delta) \beta_1 \beta_2^{-2}$ ist. Dies ist der Fall, wenn k hinreichend groß. $\square$

[

Verwendung der exakten Schrittweite] Die Vorraussetzungen (4.4.13) und (4.7.2) seien erfüllt.
Für die exakte Schrittweite σ_k gilt nach (4.5.4) mit $\lambda_k \in]0,1[$:

$$\sigma_k = -\frac{\nabla f(x^{(k)})^T d^{(k)}}{(d^{(k)})^T f''(x^{(k)} + \lambda_k \sigma_k d^{(k)}) d^{(k)}} = \frac{(d^{(k)})^T f''(x^{(k)}) d^{(k)}}{(d^{(k)})^T f''(x^{(k)} + \lambda_k \sigma_k d^{(k)}) d^{(k)}}$$

Wegen (4.4.13) und Lemma 4.4.6:

$$\sigma_k \leq \frac{\alpha_2 \|d^{(k)}\|^2}{\alpha_1 \|d^{(k)}\|^2} = \frac{\alpha_2}{\alpha_1}$$

Weiter gilt wegen $x^{(k)} \to \tilde{x}, x^{(k)} + \lambda_k d^{(k)} \to \tilde{x}$: $\qquad \sigma_k \to 1$.

Wir zeigen: $\exists k_0 \in \mathbb{N}$ und eine von k unabhängige Konstante c_5 mit

$$\sigma_k = 1 + \tilde{r_k} \text{ und } \|\tilde{r_k}\| \leq c_5 \|\nabla f(x^{(k)})\| \quad \forall k \geq k_0 \tag{4.7.4}$$

(σ_k konvergiert hinreichend schnell gegen 1)

Es gilt $\sigma_k = \frac{(d^{(k)})^T f''(x^{(k)}) d^{(k)}}{(d^{(k)})^T f''(x^{(k)} + \lambda_k \sigma_k d^{(k)}) d^{(k)} + r_k}$ mit $r_k = (d^{(k)})^T f''(x^{(k)} + \lambda_k \sigma_k d^{(k)}) d^{(k)} - (d^{(k)})^T f''(x^{(k)}) d^{(k)}$

Wegen (4.7.2) gilt:

$$r_k \leq \|f''(x^{(k)} + \lambda_k \sigma_k d^{(k)}) - f''(x^{(k)})\| \cdot \|d^{(k)}\|^2 \leq L_2 r_k \sigma_k \|d^{(k)}\|^3 \leq L_2 \sigma_k \|d^{(k)}\|^3 \leq \frac{L_2 \alpha_2}{\alpha_1} \|d^{(k)}\|^3$$

Definiert man $\tilde{r_k} := -\frac{r_k}{(d^{(k)})^T f''(x^{(k)} + \lambda_k \sigma_k d^{(k)}) d^{(k)} + r_k}$, dann gilt $\sigma_k = 1 + \tilde{r_k}$. Wegen (4.4.13) gilt

$$(d^{(k)})^T f''(x^{(k)}) d^{(k)} + r_k \geq \alpha_1 \|d^{(k)}\| - \frac{L_2 \alpha_2}{\alpha_1} \|d^{(k)}\|^3$$

Wegen $d^{(k)} \to 0_n$ gibt es ein $k_0 \in \mathbb{N}$ mit:

[weiter im Buch!!! Seite 112 f]

Satz (4.7.5)**.** *Vorraussetzungen (4.4.13) und (4.7.2) seien erfüllt. Dann gilt für das gedämpfte Newtonverfahren mit exakten Schrittweiten σ_k die (4.7.4) für hinreichend große k erfüllen (d. h. $\sigma_k \to 1$) und das Verfahren konvergiert lokarl quadratisch.*

$$\min_{\|d\| \leq \rho_k} f_k(d) \tag{4.10.1}$$

Trust-Region-Hilfsproblem

Zur Anpassung von ρ_k :

Sei $d^{(k)}$ globale Lösung von (4.10.1). Dann berechnet man
$r_k = \frac{f(x^{(k)} - f(x^{(k)} - d^{(k)})}{f(x^{(k)} - f_k(d^{(k)})}$ Man vertraut dem lokalen Modell, falls $r_k \approx 1$ ist. Zur Anpassung von ρ_k gibt man Zahlen $0 < \delta_1 < \delta_2 < 1$ vor. Ist $r_k \in [\delta_1, \delta_2[$, dann vertraut man dem Modell und lässt ρ_k unverändert.

Ist $r_k \geq \delta_2$, dann ist das lokale Modell „besonders gut", man vergrößert ρ_k.

Falls $r_k < \delta_1$: Modell ist „schlecht"; ρ_k wird verkleinert.

TRN-Verfahren

(Trust-Region-Newton-Verfahren)

Gegeben seien $0 < \delta_1 < \delta_2 < 1, \quad \sigma_1 \in]0,1[, \quad \sigma_2 > 1, \quad \rho_0$

1. Wähle startpunkt $x^{(0)}$, berechne $b^{(0)} = \nabla f(x^{(0)}), \quad A^{(0)} = f''(x^{(0)}$. Setze $k := 0$
2. Berechne eine globale Lösug des Problems
   $\min\limits_{\|d\| \leq \rho_k} f(x^{(k)}) + (b^{(k)})^T d + \frac{1}{2} d^T A^{(k)} d$
   Falls $f(x^{(k)} = f_k(d^{(k)})$ dann STOP ($\nabla f(x^{(k)}) = 0$)
3. Berechne r_k nach (siehe oben)
   Ist $r_k \geq \delta_1$ (erfolgreicher Iterationsschritt)
   - Setze $x^{(k+1)} = x^{(k)} + d^{(k)}$
   - Berechne $b^{(k+1)} = \nabla f(x^{(k+1)}), \quad A^{(k+1)} = f''(x^{(k+1)}$.
   - Aktualisiere ρ_k:
     Ist $r_k \in [\delta_1, \delta_2[$: Wähle $\rho_{k+1} \in [\sigma_1 \rho_k, \rho_k]$, falls $r_k \geq \delta_2$: wähle $\rho_{k+1} \in [\rho_k, \sigma_2 \rho_k]$
   - Setze $k := k + 1$ gehe zu 2.
4. Ist $r_k < \delta_1$ (nicht erfolgreicher Interationsschritt)
   - Wähle $\rho_{k+1} \in]0, \sigma_1 \rho_k]$
   - Setze $x^{(k+1)} = x^{(k)}$ (aktueller Iterationspunkt bleibt unverändert) $b^{(k+1)} = b^{(k)}, \quad A^{(k+1)} = A^{(k)}$
   - Setze $k := k + 1$ gehe zu 2.

Satz (4.10.12)**.** *Vorraussetzung (4.1.1), (4.4.10), und f sei auf einer offenen Obermenge von $N(f, f(x^{(k)}))$ zweimal stetig differenzierbar. Stoppt das Verfahren nicht nach endlich vielen Schritten, Dann gilt $\nabla f(x^{(k)}) \to 0_n$.*
Weiter hat die Folge $\{x^{(k)}\}$ mindestens einen Häufungspunkt, und für jeden HP $\tilde{x}$ gilt $\nabla f(\tilde{x}) = 0_n$

Satz (4.10.12)**.** *Vorraussetzungen wie oben. Weiter sei $\tilde{x}$ ein HP von $\{x^{(k)}\}$ und $f''(\tilde{x})$ sei positiv definit. Stoppt das Verfahren nicht nach endlich vielen Schritten, dann konvergiert die Folge $\{x^{(k)}\}$ gegen $\tilde{x}$ und das TRN-Verfahren geht nach endlich vielen Iterationen in das Newtonverfahren über.*

numerische Resultate.