

8. Ein Modellalgorithmus

8.1 Schrittwertbestimmung und Trust-Region

Ziel: Verfahren zur Lösung von (P) in der Regel sei $f \in C^1(\mathbb{R}^n)$ und habe einen globalen Minimierer, Abbruchbedingung $\Omega f(x) = 0$

Kriterien: (1) In jeder Iteration einen Abstieg bzw. keinen Anstieg des Zielfunktionswert.

$$f(x^{k+1}) < f(x^k) \text{ bzw. } f(x^{k+1}) \leq f(x^k)$$

für die Iterationsfolge $(x^k)_k$, genannt: Abstiegsverfahren

- (2) Finde nach endlich vielen Schritten einen stationären Punkt von f bzw. als Häufungspunkt von $(x^k)_k$ einen stationären Punkt. Wenn das Verfahren mit der Ausgabe eines solchen x^* terminiert, dann ist es ein lokaler Minimalpunkt oder ein Sattelpunkt.
(sonst: Startpunkt x^0 war lokaler Maximalpunkt, wähle dann ein anderes x^0 und beginne erneut)

Sattelpunkte können ausgeschlossen werden, wenn f konvex ist oder es werden Optimalitätsbedingungen 2. Ordnung geprüft. Dies ist für große n numerisch eventuell zu teuer. Nur für „einfache“ nicht-lineare Funktionen kann ein globaler Minimalpunkt gefunden werden. Mit o.g.

Verfahren wird man an einem lokalen Minimum hängen bleiben.

Gegenstrategie: starte mit unterschiedlichen Startpunkten x^0 (pseudoauflösbar)

Annahme: x^k ist kein kritischer Punkt von f , d.h. $\Omega f(x^k) \neq 0$.

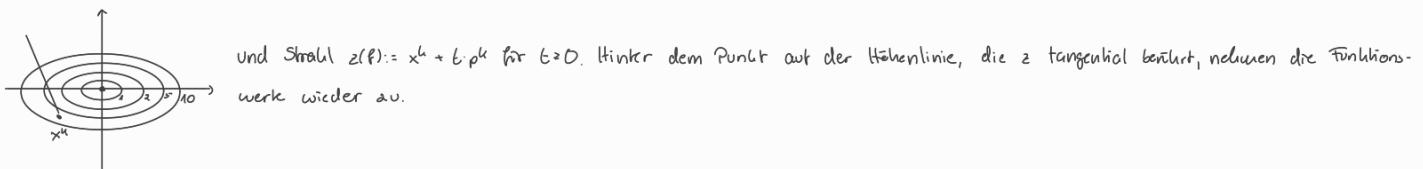
Elappennail: bestimme x^{k+1} mit eventuell „bessrem“ Zielfunktionswert.

Bsp. 8.1 Vektor $p \in \mathbb{R}^n$ heißt Abstiegsrichtung für $f: \mathbb{R}^n \rightarrow \mathbb{R}$, falls ein $t_0 > 0$ existiert, so dass $f(x + t p) < f(x) \quad \forall t \in (0, t_0]$.

Schreibe: $x^{k+1} := x^k + t_k p_k$, so geht es also darum, zu x^k eine Abstiegsrichtung p und eine Schrittwl. $t_k > 0$ zu bestimmen.

Andere Schrittwl.: $p^k := t_k p^k$ ist Abstiegsrichtung mit geeigneter Länge $\|p^k\|$.

Schrittwl. t_k bzw. die Länge $\|p^k\|$ dürfen nicht zu groß sein, da die Funktionswerte dann wieder ansteigen können.



Die meisten Verfahren wählen erst Abstiegsrichtung, dann Schrittwl.

Alternative: Trust-Region-Verfahren, die beides kombinieren. Dazu ersetzt man $f(x^k + p)$ durch eine Näherung (z.B. quadratisch) $q_k(p)$ und testet für eine passende Konstante t_k : minimiere q_k

$$\text{u.d. } N \|p\| \leq t_k$$

t_k definiert einen Vertrauensbereich (trust region), in welchem die gesuchte Richtung sein darf. Ob Lösung \hat{x}^k akzeptabel ist, muss dann entschieden werden (Satz von Weierstraß garantiert Lösung für stetige q_k). Ist die Lösung nicht braubar, wird t_k verkleinert und nochmals gelöst.
 $q_k(\cdot)$ ist verschieden von $f(x^k + \cdot)$ damit das Teilproblem einfacher ist als das Ausgangsproblem (P).

In der Praxis haben beide Verfahren Vor- und Nachteile.

8.2 Ein Modellalgorithmus zur Schrittwl.-Strategie

Sei $f \in C^1(\mathbb{R}^n)$. Gilt für $p \in \mathbb{R}^n$ das $\Omega f(x)^T \cdot p < 0$, so ist p Abstiegsrichtung für f in x .

Beweis: Betrachte Richtungsableitung in x Richtung p : $\lim_{t \rightarrow 0} \frac{f(x + t p) - f(x)}{t} = \Omega f(x)^T \cdot p < 0$.

Wenn der Limes < 0 ist, müssen schon für t nahe Null Werte < 0 gewesen sein (da $f \in C^1$, also $\Omega f \in C^0$).

$$\frac{f(x + t p) - f(x)}{t} < 0 \quad \forall t \in [0, t_0] \Rightarrow \underbrace{f(x + t p) < f(x)}_{\text{D.h. Abstiegsrichtung}} \quad \forall t \in [0, t_0]$$

□

Beispiel 8.2 Sei $f \in C^1(\mathbb{R}^n)$ und $H \in \mathbb{R}^{n,n}$ symmetrisch positiv definit. Sei $x \in \mathbb{R}^n$ mit $\Omega f(x) \neq 0$, dann ist $p := -H \cdot \Omega f(x)$ Abstiegsrichtung für f in x , denn:

$$\Omega f(x)^T \cdot p = \underbrace{-\Omega f(x)^T \cdot H \cdot \Omega f(x)}_{< 0}$$

Für $H=I$ erhält man die Abstiegsrichtung $p := -\Omega f(x)$.

Kann zeigen: $-\Omega f(x)$ ist Richtung des statisten Anstiegs und (normierte) Lösung des Optimierungsproblems: minimiere $\Omega f(x)^T \cdot p$. Lokal ist sie die „bestmögliche“ Richtung,

$$\text{u.d. } N \|p\| = 1$$

global stimmt das nicht unbedingt.

Bemerkung 8.4 Kann die Definition der Richtung des stetigen Abstiegs verallgemeinern mittels positiv-definiter Matrix $A \in \mathbb{R}^{n \times n}$ und Skalarprodukt $\langle x, y \rangle_A := x^T A \cdot y \quad \forall x, y \in \mathbb{R}^n$ sowie Norm $\|x\|_A := \sqrt{\langle x, x \rangle_A}$, die elliptische Norm. Ist $f \in C^1(\mathbb{R}^n)$ und $x \in \mathbb{R}^n$ mit $Df(x) \neq 0$, so kann man zeigen, dass $p^* := \frac{A^{-1} Df(x)}{\|A^{-1} Df(x)\|_A}$ eine eindeutige Lösung von $\min_{\mathbb{R}^n} Df(x)^T p$ ist. Vektor $-A^{-1} Df(x)$ ist Richtung des stetigen Abstiegs in x bzgl. $\| \cdot \|_A$ und für p^* gilt: $Df(x)^T p^* < 0$. u.d.N. $\|p^*\|_A = 1$

Algorithmus 8.5 Modellalgorithmus

- (0) Initialisierung: Wähle $x^0 \in \mathbb{R}^n$. Setze $k := 0$.
- (1) Abbruchkriterium: Falls $Df(x^k) = 0 \rightarrow \text{STOPP!}$
- (2) Abstiegsrichtung: bestimme $p \in \mathbb{R}^n$ mit $Df(x^k)^T p < 0$.
- (3) Schrittwert: bestimme $t_k > 0$ mit $f(x^k + t_k \cdot p^k) < f(x^k)$.
- (4) Nächste Iteration: Setze $x^{k+1} := x^k + t_k \cdot p^k$ und $k := k+1$, gehe zu (1).

Für theoretische Untersuchungen ist das „ideale“ Abbruchkriterium $Df(x^k) = 0$ okay. In der Praxis wird es z.B. durch das realistische Abbruchkriterium $\|Df(x^k)\| \leq \varepsilon$ für vorgegebenes $\varepsilon > 0$ ersetzt. Dies ist allerdings nicht ohne Probleme: $f(x) := \frac{1}{2} \cdot 10^{-7} \cdot x^2$ und $\varepsilon := 10^{-6}$. Dann wären $x \in [-10, 10]$ kritische Punkte, da $|f'(x)| \leq 10^{-6}$.

Daher kann (aus praktischen Gründen) gefordert werden: $\frac{|f(x^k) - f(x^{k+1})|}{|f(x^{k+1})|} \leq \delta_1$ für ein kleines $\delta_1 > 0$ und während mehreren Iterationen, d.h. keine signifikante Änderung des Funktionswerts.

Oder man fordert (aus praktischen Gründen): $\frac{\|x^k - x^{k+1}\|}{\|x^{k+1}\|} \leq \delta_2$ für ein kleines $\delta_2 > 0$ und während mehreren Iterationen, d.h. der Minimierer ändert sich nicht mehr.

Die zentralen Schritte des Algorithmus 8.5 sind ausführbar: Als Abstiegsrichtung kann stattdessen der stetige Abstieg $p_k := -Df(x^k)$ gewählt werden. Eine Schrittwerte dazu existiert auch, siehe Beweis Lemma 8.2.

Um den Wert $p_k(t) := f(x^k + t_k \cdot p_k)$ in der k -ten Iteration möglichst klein werden zu können, möchte man (theoretisch) ein t_k wählen mit $f(x^k + t_k \cdot p_k) = \inf_{t \in [0, \infty)} f(x^k + t \cdot p_k)$, d.h. für das die 1-d Funktion $p_k(\cdot)$ auf $[0, \infty)$ ein globales Minimum annimmt. Wir werden später zeigen, dass eine solche Minimumsschrittwerte unter recht schwachen Voraussetzungen existiert.

Numerisch gesehen ist die Aufgabe, dass globale Minimum einer nichtkonvexen Funktion zu bestimmen → schwierig und nur näherungsweise möglich. Gesucht sind daher pragmatische Regeln.

8.3 Standardvoraussetzungen

(V1) $f \in C^1(\mathbb{R}^n)$

(V2) Für ein $x^0 \in \mathbb{R}^n$ (Startpunkt des Verfahrens) ist die Nivroumenge $N_0 := N(x^0) := \{x \in \mathbb{R}^n \mid f(x) \leq f(x^0)\}$ kompakt.

(V3) Df ist auf N_0 Lipschitz-stetig, d.h. es gibt eine Konstante $\gamma > 0$, sodass $\|Df(x) - Df(y)\| \leq \gamma \|x - y\| \quad \forall x, y \in N_0$.

Gelegentlich fordern wir zudem:

(V4) N_0 ist konvex und f ist gleichmäßig konvex auf N_0

20. Juni 2024

8.6 Folgerungen aus (V1) - (V4)

i) (V2) sichert ab, dass (P) eine Lösung besitzt (Satz 7.8). Nach dem Satz von Bolzano-Weierstraß hat jede Folge $(x^k)_k$ mit $x^k \in N_0$ (d.h. mit $f(x^k) \leq f(x^0)$) einen Häufungspunkt in N_0 .

Nach Konstruktion des Verfahrens gilt: $f(x^{k+1}) < f(x^k) < f(x^0)$, also ist die Voraussetzung vom B.-W. erfüllt.

ii) (V3) ist erfüllt, wenn (V2) gilt und $f \in C^2(K)$ für ein nicht-leeres, kompakt-konvexe Menge $K \subseteq N_0$.

Beweis: Wegen $f \in C^2(K)$ existiert $\gamma := \max_{x \in K} \|D^2f(x)\|$.

Weil K konvex ist $x + t(y-x) \in K \quad \forall x, y \in K$ und $t \in [0, 1]$. Setze $u(t) := Df(x + t(y-x)) \quad \forall t \in [0, 1]$. Aus dem Mittelwertsatz folgt die Existenz $\tau \in (0, 1)$ mit

$$u(1) - u(0) = u'(1\tau), \text{ d.h. } Df(y) - Df(x) = D^2f(x + \tau D(y-x)) \cdot (y-x) \rightarrow \|Df(y) - Df(x)\| = \|D^2f(x + \tau D(y-x)) \cdot (y-x)\| \leq \|D^2f(x + \tau D(y-x))\| \cdot \|y-x\| \leq \gamma \cdot \|y-x\| \quad \forall x, y \in K. \square$$

iii) Aus (V1) und (V4) folgt, dass $N(x^0)$ abgeschlossen ist (Lemma 7.7 iii)). Nach Satz 7.8 i) ist dann $N(x^0)$ kompakt. Also gilt (V2) und nach Satz 7.9 ii) besitzt (P) genau eine Lösung auf $N(x^0)$. Somit hat (P) unter (V1) und (V4) genau eine globale Lösung x^* . Aus Satz 7.3 i) folgt, dass x^* einiger kritischer Punkt von f auf $N(x^0)$ ist.

(Beweis: angenommen, y wäre ein weiterer, dann wäre $f(y) + Df(y)^T(x^*-y) \leq f(x^*) \leq f(y) \leq f(x^*)$ im Widerspruch dazu, dass x^* eindeutig war.)

Ist f auf ganz \mathbb{R}^n gleichmäßig konvex, ist x^* die einzige kritische Lösung von (P) (Korollar 7.19).

(V4) wird normalerweise dadurch verhindert, dass f auf \mathbb{R}^n gleichmäßig konvex ist.

Beispiel 8.7 Sei $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semidefinit und f die dadurch definierte konvexe quadratische Funktion $f(x) := \frac{1}{2} x^T Q x + c^T x + b \quad \forall x \in \mathbb{R}^n$

• Es ist $f \in C^1(\mathbb{R}^n)$, somit gilt (V1).

• Es gilt auch (V3), denn: $\|Df(x) - Df(y)\|^2 = \|Q(x-y)\|^2 = (Q \cdot (x-y))^T \cdot Q(x-y)$

$$= (x-y)^T Q^T Q \cdot (x-y) \stackrel{7.10}{\leq} \lambda_{\max}(Q^T Q) \cdot \|x-y\|^2 \stackrel{Q \text{ sym.}}{\leq} \lambda_{\max}(Q^2) \cdot \|x-y\|^2 = \lambda_{\max}(Q)^2 \cdot \|x-y\|^2$$

$$\Rightarrow \|Df(x) - Df(y)\| \leq \lambda_{\max}(Q) \cdot \|x-y\|$$

Setze $\gamma := \lambda_{\max}(Q) > 0$, wobei $\gamma > 0$ nur für Q positiv definit garantiert. γ ist Lipschitz-Konstante von Df .

• Ist Q positiv definit, so ist (V4) erfüllt, da f nach Lemma 7.14 ii) dann gleichmäßig konvex auf $N_0 := \mathbb{R}^n$ mit Konvexitätskonstante $\beta := \lambda_{\min}(Q) > 0$.

• (V1) und (V4) implizieren (V2) (Bem. 8.6). Für positiv definites Q gilt: $\text{cond}(Q) = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} = \frac{\gamma}{\beta}$.

8.4 Hilfsmittel

... zur Analyse des Modellalgorithmus.

In jeder Iteration von Modellalgorithmus 8.5 geht man von einem Punkt x mit $Df(x) \neq 0$ aus und bestimmt dazu eine Abstiegsrichtung p , eine Schrittweite $\delta > 0$ und neuen Punkt $x + \delta \cdot p$ mit kleinerem Zielfunktionswert.

Schreibt $\psi(t) := f(x) - f(x + t \cdot p)$. Ist $\psi(t) > 0$ für ein Intervall $t \in (0, \bar{t})$, so ist ein Abstieg möglich.

Lemma 8.8 Es gelten (U1) und (U2). Seien $x \in N_0$, $p \in \mathbb{R}^n$ und $Df(x)^T p \leq 0$ gegeben. Dann besitzt $\psi \in C^1([0, \bar{t}])$ eine kleinste positive Nullstelle $\tilde{t} := \tilde{t}(x, p) > 0$. Es gilt $\psi(t) > 0 \quad \forall t \in (0, \tilde{t})$ und $x + t \cdot p \in N_0 \quad \forall t \in [0, \tilde{t}]$. Und es gibt ein $\varepsilon \geq \tilde{t}$, so dass $x + t \cdot p \notin N_0$ und $\psi(t) < 0 \quad \forall t \in [\varepsilon, +\infty)$.

Beweis: Es ist $\psi(0) = f(x) - f(x + 0 \cdot p) = 0$ und $\psi'(t) = [f(x) - f(x + t \cdot p)]' = -Df(x + t \cdot p)^T \cdot p \Rightarrow \psi(0) = -Df(x + 0 \cdot p)^T \cdot p = -Df(x)^T \cdot p > 0$.

Die Funktion ψ hat in 0 eine Nullstelle und dort eine positive Steigung, d.h. sie ist (lokal) monoton wachsend.

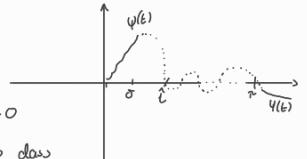
Somit gibt es ein $\delta > 0$, so dass $\psi(t) > 0 \quad \forall t \in (0, \delta)$. Da N_0 kompakt, also insbesondere beschränkt ist und $p \neq 0$

ist $x + t \cdot p \notin N_0$ für alle hinreichend großen $t > 0$. Für solche t gilt: $f(x + t \cdot p) \stackrel{N_0}{>} f(x) \geq f(x)$. Also existiert ein $\varepsilon > 0$, so dass $x + t \cdot p \in N_0$ und $\psi(t) < 0$ für alle $t \geq \varepsilon$. Nach dem Zwischenwertsatz ist die Menge der Nullstellen von $\psi(t)$ auf $t \in [\delta, \varepsilon]$ nicht leer.

$$N := \{t \in [\delta, \varepsilon] \mid \psi(t) = 0\} \neq \emptyset$$

Als Teilmenge einer beschränkten Menge ist sie beschränkt. Da ψ stetig, ist das Urbild der abgeschlossenen Menge $\{\psi(t)\}$ wiederum abgeschlossen. Also ist N kompakt.

Nach dem Satz von Weierstraß nimmt die stetige Funktion $t \mapsto \psi(t)$ auf N ein Minimum an, so dass $\tilde{t} := \min_{t \in N} t$ wohldefiniert ist. Somit kann δ bis zu \tilde{t} „verlängert“ werden, d.h. $\psi(t) > 0$ für alle $t \in (0, \tilde{t})$ und somit $\varepsilon \geq \tilde{t}$. Es ist $\psi(t) < 0$ für alle $t \in [\tilde{t}, \varepsilon]$. Somit $f(x + t \cdot p) \leq f(x) \leq f(x^\star)$ und daher $x + t \cdot p \notin N_0 \quad \forall t \in [\tilde{t}, \varepsilon]$. \square



Für gleichmäßig konvexe Funktionen gelten noch weitere Abschätzungen. Unter (U1)-(U4) hat (P) eine eindeutige globale Lösung x^* (8.6 iii)). Wegen $f(x^*) < f(x^\star)$ liegt diese in N_0 .

Lemma 8.9 Es gelten (U1)-(U4) und x^* sei eine globale Lösung von (P) mit β Konvexitätskonstante aus (U4) und γ Lipschitz-Konstante aus (U3) gilt $\forall x, y \in N_0$:

$$\text{i)} \frac{\beta}{2} \|y - x\|^2 + Df(x)^T (y - x) \leq f(y) - f(x) \leq \frac{\beta}{2} \|y - x\|^2 + Df(x)^T (y - x).$$

$$\text{ii)} \beta \|y - x\|^2 \leq (Df(y) - Df(x))^T (y - x)$$

$$\text{iii)} \frac{\beta}{2} \|x - x^*\|^2 \leq f(x) - f(x^*)$$

$$\text{iv)} \|x - x^*\| \leq \frac{1}{\beta} \|Df(x)\|$$

$$\text{v)} f(x) - f(x^*) \leq \frac{1}{2\beta} \|Df(x)\|^2$$

$$\text{vi)} \frac{1}{2\gamma} \|Df(x)\|^2 \leq f(x) - f(x^*)$$

Beweis: Seien x, y, x^* wie angegeben. Da N_0 nach (U4) konvex, ist auch $x + s(y - x) \in N_0 \quad \forall s \in [0, 1]$.

i) Da f nach (U4) auf N_0 gleichmäßig konvex und N_0 konvex, gilt nach Satz 7.3 iii) dass $\frac{\beta}{2} \|y - x\|^2 + f(x) + Df(x)^T (y - x) \leq f(y) \quad \forall x, y \in N_0$.

Subtraktion von $f(x)$ liefert die linke Ungleichung. Definiere $\phi(t) := f(x + t(y - x))$ mit $\phi \in C^1([0, 1])$. Dann ist $\phi'(s) = Df(x + s(y - x))^T (y - x)$ und $f(y) - f(x) = \phi(1) - \phi(0) = \int_0^1 \phi'(s) ds = \phi'(0) - \int_0^1 \phi'(s) ds = \phi'(0) - \int_0^1 (\phi'(s) - \phi'(0)) ds = Df(x)^T (y - x) + \int_0^1 (\phi'(s) - \phi'(0)) ds$. Es bleibt das Integral abschätzen.

$$\begin{aligned} \int_0^1 (\phi'(s) - \phi'(0)) ds &= \int_0^1 (Df(x + s(y - x)) - Df(x))^T (y - x) ds \stackrel{\text{C04}}{\leq} \int_0^1 \|Df(x + s(y - x)) - Df(x)\| \cdot \|y - x\| ds \\ &\stackrel{\text{(U3)}}{\leq} \int_0^1 \gamma \|x + s(y - x) - x\| \cdot \|y - x\| ds = \int_0^1 \gamma s \cdot \|y - x\|^2 ds = \frac{\gamma}{2} \|y - x\|^2 \end{aligned}$$

Somit ist auch die rechte Ungleichung gezeigt.

$$\text{ii)} Wie schon gezeigt (linke Ungleichung): \frac{\beta}{2} \|y - x\|^2 + Df(x)^T (y - x) \leq f(y) - f(x),$$

$$\text{Also gilt auch } \frac{\beta}{2} (\|y - x\|^2 + Df(y)^T (x - y)) \leq f(x) - f(y).$$

Addition dieser beiden ergibt: $\beta \|y - x\|^2 + (Df(x) - Df(y))^T (y - x) \leq 0$. Subtraktion des rechten Terms liefert die Ungleichung.

$$\text{iii)+iv)} Setze } y := x \text{ und } x := x^* \text{ dann gilt mit i): } \frac{\beta}{2} \|x - x^*\|^2 + Df(x^*)^T (x - x^*) \leq f(x) - f(x^*). \text{ Aus } Df(x^*) = 0 \text{ folgt ii).}$$

$$\text{Benutze ii): } \beta \|x - x^*\|^2 \leq (Df(x) - Df(x^*))^T (x - x^*) \stackrel{\text{C04}}{\leq} \|Df(x)\| \cdot \|x - x^*\| = \frac{1}{\beta} \|x - x^*\|$$

$$\Rightarrow \|x - x^*\| \leq \frac{1}{\beta} \|Df(x)\|$$

v) Setze $g_x(h) := \frac{\beta}{2} h^T h + Df(x)^T h$ $\forall h \in \mathbb{R}^n$. Nach Lemma 7.14 (ii) ist diese quadratische Funktion gleichmäßig konvex (da $\beta \cdot I_n$ positiv definit). Nach Satz 7.9 gibt es dann ein eindeutiges Minimum in einem Punkt h^* mit $Dg_x(h^*) = 0 \iff \beta \cdot h^* + Df(x) = 0 \iff h^* = -\frac{1}{\beta} \cdot Df(x)$.

$$\text{Nach (U2) ist } N_0 \text{ kompakt. Es folgt: } g_x(h^*) = \frac{\beta}{2} h^{*T} h^* + Df(x)^T \cdot h^* = \frac{\beta}{2} \cdot (-\frac{1}{\beta}) \cdot (-\frac{1}{\beta}) Df(x)^T \cdot Df(x) + (-\frac{1}{\beta}) Df(x)^T Df(x)$$

$$= \frac{1}{2\beta} \|Df(x)\|^2 - \frac{1}{\beta} \|Df(x)\|^2 = -\frac{1}{2\beta} \|Df(x)\|^2$$

25. Juni 2024

$$\begin{aligned} \text{Weiter gilt: } g_x(h^*) &= \min_{h \in \mathbb{R}^n} g_x(h) \leq \min_{y \in N_0} g_x(y-x) \\ &= \min_{y \in N_0} \left\{ \frac{\beta}{2} \|y-x\|^2 + \nabla f(x)^T (y-x) \right\} \leq \min_{y \in N_0} [f(y) - f(x)] = (\min_{y \in N_0} f(y)) - f(x) = f(x^*) - f(x) \end{aligned}$$

$$\Rightarrow f(x) - f(x^*) \leq \frac{1}{2\beta} \|\nabla f(x)\|^2$$

vi) Setze $\hat{y} := x - \frac{1}{\beta} \nabla f(x)$. Zeige, dass $\hat{y} \in N_0$.

1. Fall: $\nabla f(x) = 0$. Dann ist $\hat{y} = x \in N_0$

2. Fall: $\nabla f(x) \neq 0$. Setze $p := -\nabla f(x)$. Dann ist $\nabla f(x)^T p = -\|\nabla f(x)\|^2 < 0$.

Ferner gelten (U1), (U2). Also ist Lemma 8.8 anwendbar. Die Funktion $x(t) := f(x) - f(x) - t \nabla f(x)$ für $t \geq 0$ besitzt eine kleinste positive Nullstelle $\tilde{t} > 0$ und $x - \tilde{t} \nabla f(x) \in N_0$ für alle $t \in [0, \tilde{t}]$. Aus (ii), rechte Ungleichung folgt:

$$\begin{aligned} 0 &= f(x - \tilde{t} \nabla f(x)) - f(x) \leq \frac{\beta}{2} \|(x - \tilde{t} \nabla f(x)) - x\|^2 + \nabla f(x)^T ((x - \tilde{t} \nabla f(x)) - x) \\ &= \frac{\beta}{2} \cdot \tilde{t}^2 \|\nabla f(x)\|^2 - \tilde{t} \|\nabla f(x)\|^2 \\ \Rightarrow 0 &\leq \tilde{t}^2 - 2 \frac{1}{\beta} \cdot \tilde{t} + \frac{1}{\beta^2} = (\tilde{t} - \frac{1}{\beta})^2 - \frac{1}{\beta^2} \Rightarrow \frac{1}{\beta} \leq \tilde{t}. \end{aligned}$$

Also ist auch $\frac{1}{\beta} \leq \tilde{t}$. Somit $\hat{y} \in N_0$.

$$\text{Dann gilt: } f(x^*) - f(x) \leq f(\hat{y}) - f(x) \leq \frac{\beta}{2} \cdot \frac{1}{\beta^2} \|\nabla f(x)\|^2 - \frac{1}{\beta} \|\nabla f(x)\|^2 = -\frac{1}{2\beta} \|\nabla f(x)\|^2$$

$$\Rightarrow f(x) - f(x^*) \geq \frac{1}{2\beta} \|\nabla f(x)\|^2$$

□

Bem 8.10

Für $f(x) := \frac{1}{2} x^T x$ kann $\beta = \lambda_{\min}(I_n) = 1$ und $\gamma = \lambda_{\max}(I_n) = 1$ gewählt werden. Für dieses f gelten alle Abschätzungen in Lemma 8.9 mit Gleichheit.

Lemma 8.9 liefert durch die gläichmäßige Konvexität von f eine obere Schranke des Fehlers $\|x - x^*\|$ durch $f(x) - f(x^*)$. Nutze dies später in Konvergenzbeweisen.

Die andere Richtung, d.h. Abschätzung von $f(x) - f(x^*)$ nach oben durch $\|x - x^*\|$ ist die Lipschitz-Stetigkeit, welche für Funktionen $f \in C^1(I^n)$ lokal auf N_0 gilt.

8.5 Bedingungen an die Schrittweite

Es gibt viele verschiedene Lösungsverfahren für (P), die sich untereinander nur durch die Schrittweitenregel unterscheiden. Untersuchen daher zunächst abstrakt, welche Eigenschaften eine Schrittweitenregel haben soll, damit das Verfahren konvergiert.

Hierzu: leite zwei Bedingungen her. Führt auf Definition der elphantinen und semiephantinen Schrittweitenregel.

Nach Lemma 8.8 gibt es für x und Abschätzung p ein offenes Intervall $(0, \tilde{t})$, in dem $\psi(t) := f(x) - f(x - tp) > 0$, d.h. Zielfunktionswert von (P) kann verkleinert werden. Es könnte aber sein, dass das Schrittverfahren bei einer zu geringen Vermindierung des ZFW von (P) nicht konvergieren könnte. Wegen $\psi(0) = \psi(\tilde{t}) = 0$ sollte die Schrittweite nicht zu nahe an 0 oder \tilde{t} liegen bzw. $\psi(t)$ sollte hinreichend groß sein.

Lemma 8.11 dient dazu, eine geeignete Forderung an die Schrittweite zu stellen.

Es gelten (U1)-(U3) und gegeben seien $x \in N_0$, $p \in \mathbb{R}^n$ mit $\nabla f(x)^T p < 0$. Sei $\tilde{t} := \tilde{t}(x; p) > 0$ die erste positive Nullstelle von ψ nach Lemma 8.8.

Dann gilt:

$$i) \tilde{t} \geq -\frac{2}{\beta} \frac{\|\nabla f(x)^T p\}}{\|p\|^2} =: \tilde{t}' > 0$$

$$ii) \psi(t) \geq -t \cdot \nabla f(x)^T p - t^2 \frac{\beta}{2} \|p\|^2 =: \Psi(t) \quad \forall t \in [0, \tilde{t}]$$

Beweis: Nach Lemma 8.8 ist $x + t p \in N_0 \quad \forall t \in [0, \tilde{t}]$.

$$\begin{aligned} \text{Es ist } \psi(0) = 0 \text{ und somit: } \psi(t) - \psi(0) - \psi'(t) &= \int_0^t \psi'(s) ds = \psi'(0) \cdot t - \int_0^t \psi'(0) ds + \int_0^t \psi'(s) ds \\ &= \psi'(0) \cdot t - \int_0^t \psi(s) - \psi(0) ds \end{aligned}$$

$$\begin{aligned} \text{Auch } \psi(t) &= f(x) - f(x - tp) \text{ folgt } \psi'(t) = -\nabla f(x + tp)^T p \Rightarrow \psi'(0) = -\nabla f(x)^T p \\ \Rightarrow \psi(t) - \psi(0) - \psi'(t) &= -t \nabla f(x)^T p + \int_0^t -\nabla f(x + sp)^T p - (-\nabla f(x))^T p ds \\ &= -t \nabla f(x)^T p - \int_0^t (\nabla f(x + sp))^T - \nabla f(x)^T p ds \\ &\geq -t \nabla f(x)^T p - \int_0^t \|\nabla f(x + sp) - \nabla f(x)\| \cdot \|p\| ds \\ &\geq -t \nabla f(x)^T p - \int_0^t \gamma s \|p\|^2 ds \leq -t \nabla f(x)^T p - \frac{\beta}{2} t^2 \|p\|^2 =: \Psi(t) \quad \forall t \in [0, \tilde{t}] \end{aligned}$$

Somit gilt (ii)

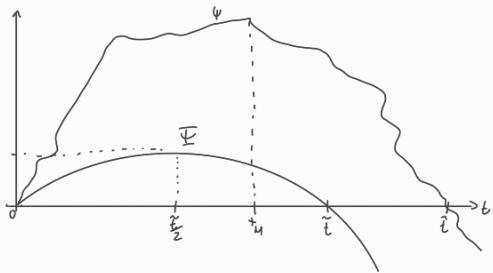
$$\text{Für } t := \tilde{t} \text{ und } \psi(\tilde{t}) = 0 \text{ folgt dann: } 0 \geq -\tilde{t} \cdot \nabla f(x)^T p - \tilde{t}^2 \frac{\beta}{2} \|p\|^2 \Rightarrow \tilde{t} \cdot \frac{\beta}{2} \|p\|^2 \geq -\nabla f(x)^T p \Rightarrow \tilde{t} \geq -\frac{2}{\beta} \frac{\|\nabla f(x)^T p\}}{\|p\|^2} =: \tilde{t}' > 0$$

□

Im letzten Lemma 8.11 wurde ein $\tilde{t} \in [0, \tilde{t}]$ und die Parabel $\tilde{\Psi}(t) := -t \nabla f(x)^T p - t^2 \frac{\beta}{2} \|p\|^2$, $\forall t \in [0, \tilde{t}]$ definiert. Durch einzusetzen zeigt man, dass $\tilde{\Psi}(0) = \tilde{\Psi}(\tilde{t}) = 0$. Die nach unten geöffnete Parabel nimmt ihr Maximum genau dazwischen an, also in $\tilde{t} = -\frac{1}{\beta} \frac{\|\nabla f(x)^T p\}}{\|p\|^2} \in [0, \tilde{t}]$.

Im Maximum hat $\tilde{\Psi}$ den Wert: $\max_{t \in [0, \tilde{t}]} \tilde{\Psi}(t) = \tilde{\Psi}\left(\frac{\tilde{t}}{2}\right) = -\left(-\frac{1}{\beta} \frac{\|\nabla f(x)^T p\}}{\|p\|^2}\right) \nabla f(x)^T p - \left(\frac{1}{\beta} \frac{\|\nabla f(x)^T p\}}{\|p\|^2}\right)^2 \frac{\beta}{2} \|p\|^2 = \frac{1}{2\beta} \frac{\|\nabla f(x)^T p\|^2}{\|p\|^2}$

Ferner ist nach Lemma 8.11: $\psi(t) \geq \Psi(t) \quad \forall t \in [0, \tilde{t}]$



Gewünscht wäre eine Reduktion des Funktionswertes von x zu $x+t_p$ in der Größenordnung von $(\frac{\|P(x)\|^2}{\|P\|})^2$ und größer-gleich dem Maximum von Ψ ist.

Später zeigen wir, dass die Minimumsschrittwerte t_m dies erfüllt.

Zunächst definieren wir:

Definition 8.12 Eine Schrittwertenregel heißt **effizient** (mit Konstanten ν), wenn sie jedem Paar (x, p) mit $x \in N_0$ und $p \in \mathbb{R}^n$ mit $\|P(x)\|^2 < 0$ ein $t := t(x, p) > 0$ zuordnet und wenn ein $\nu > 0$ (unabhängig von x, p, t) existiert, so dass $P(x) - P(x+t_p) \geq \nu \left(\frac{\|P(x)\|^2}{\|P\|} \right)^2$.
Gilt nur $P(x) - P(x+t_p) \geq \nu \min\left(\frac{\|P(x)\|^2}{\|P\|}, -\|P(x)\|^2\right)$, so heißt die Schrittwertenregel **semi-effizient**.
Entsprechend werden die Schrittwenden selbst bezeichnet bei Verwendung einer entsprechenden Regel.

8.6 Konvergenzaussagen

Lemma 8.13 Es gelten (U1) und (U2) und der Modellalgorithmus 8.5 habe eine beliebige Schrittwertenregel. Dann gilt für alle $N \in \mathbb{N}$:

$$\text{i)} \quad f(x^{k+1}) < f(x^k) < f(x^0)$$

$$\text{ii)} \quad x^k \in N_0$$

Bricht der Algorithmus nicht nach endlich vielen Schritten mit einer Lösung ab, so erzeugt er eine unendliche Folge $(x^k)_k$ mit:

iii) $(x^k)_k$ besitzt einen Häufungspunkt.

$$\text{iv)} \quad \lim_{k \rightarrow \infty} P(x^k) = \hat{P} \quad \text{für ein } \hat{P} \geq \min_{x \in N_0} P(x)$$

v) Gilt zusätzlich $\lim_{k \rightarrow \infty} \|P(x^k)\| = 0$, so folgt (a) jeder Häufungspunkt der Folge ist eine kritische Lösung von (P)

(b) hat (P) genau eine kritische Lösung x^* in N_0 , so ist $\lim_{k \rightarrow \infty} x^k = x^*$

Beweis: i) folgt aus der Konstruktion des Algorithmus

ii) folgt aus i) und Definition Niveaumenge

iii) Sei $(x^k)_k$ eine unendliche Folge. Außer ii) folgt $(x^k)_k \subset N_0$. Nach (U2) kompakt. Also hat die beschränkte Folge einen Häufungspunkt.

iv) Die Folge ist wegen i) monoton fallend und beschränkt, da $P(x^{k+1}) > P(x^k) > \min_{x \in N_0} P(x)$ $\forall k \in \mathbb{N}$. Also konvergiert sie. Seien $\hat{P} := \lim_{k \rightarrow \infty} P(x^k)$. Aus der Stetigkeit von P folgt $\hat{P} \geq \min_{x \in N_0} P(x)$.

v) a) Sei $\lim_{k \rightarrow \infty} \|P(x^k)\| = 0$. Sei x^* ein Häufungspunkt von $(x^k)_k$. Dann gibt es eine Teilfolge $(x^{k_i})_i$ von $(x^k)_k$ mit $\lim_{i \rightarrow \infty} x^{k_i} = x^*$. Da P stetig ist, folgt

$$0 = \lim_{i \rightarrow \infty} \|P(x^{k_i})\| = \|P(\lim_{i \rightarrow \infty} x^{k_i})\| = \|P(x^*)\|. \text{ Also ist } x^* \text{ ein kritischer Punkt von } P.$$

b) Sei x^* die einzige kritische Lösung von (P) in N_0 .

Angenommen, $(x^k)_k$ konvergiert nicht gegen x^* , dann existiert $\varepsilon > 0$, sodass für unendlich viele k , d.h. eine Teilfolge $(x^{k_i})_i$ von $(x^k)_k$ gelten würde: $\|x^{k_i} - x^*\| \geq \varepsilon \quad \forall i \in \mathbb{N}$.

Da alle Folgeglieder von $(x^k)_k$, also auch die der Teilfolge, in N_0 liegen und N_0 nach (U2) kompakt ist, kann man aus der Teilfolge $(x^{k_i})_i$ eine konvergente Teilfolge $(x^{k_{ij}})_j$ wählen, die gegen eine kritische Lösung $x^{k_{ij}}$ von (P) konvergiert. Da es nur genau eine solche gibt, ist $x^{k_{ij}} = x^*$. Dies steht aber im Widerspruch zu (*), da alle Folgeglieder von $(x^{k_i})_i$ außerhalb der ε -Umgebung liegen sollten.

27. Juni 2024

Man spricht von der **Konvergenz eines Verfahrens** zur Lösung von (P), wenn jeder Häufungspunkt der Iteratenfolge eine kritische Lösung von (P) ist. Hat (P) eine eindeutige Lösung, folgt dann unter den Voraussetzungen des Lemmas die Konvergenz der gesamten Folge.

Lemma 8.14 Es gelten (U1), (U2). Modellalgorithmus 8.5 habe effiziente Schrittwertenregel und bricht nicht nach endlich vielen Iterationen mit einer kritischen Lösung von (P) ab, d.h. er erzeugt Folgen $(x^k)_k$ und $(p^k)_k$.

Für diese gelte die **Zwotendigl-Bedingung**: $\sum_{k=0}^{\infty} \alpha_k^2 = \infty$ für $\alpha_k := -\frac{\|P(x^k)\|^2}{\|Df(x^k)\|^2 \cdot \|P\|}$.

Dann gilt:

i) Folge (x^k) hat mind. einen Häufungspunkt, der kritische Lösung von (P) ist.

ii) Sind auch (U3), (U4) erfüllt und ist dann x^* die existierende eindeutige Lösung von (P), so gilt $\lim_{k \rightarrow \infty} x^k = x^*$.

Beweis: i) Angenommen Alg. 8.5 bricht nicht nach endlich vielen Schritten ab. Da die Schrittwertenregel effizient ist, gibt es ein $\nu > 0$, sodass

$$P(x^k) - P(x^{k+1}) \geq \nu \left(\frac{\|P(x^k)\|^2}{\|P\|} \right)^2 = \nu \cdot \alpha_k^2 \cdot \|Df(x^k)\|^2 > 0, \quad \forall k.$$

$$\text{Summiere dies für } k=0, 1, \dots, l: \sum_{k=0}^l \alpha_k^2 \|Df(x^k)\|^2 \leq \frac{1}{\nu} \sum_{k=0}^l (P(x^k) - P(x^{k+1})) = \frac{1}{\nu} (P(x^0) - P(x^1))$$

$$\text{Grenztübergang für } l \rightarrow \infty \text{ und Lemma 8.13 iv) für die Existenz des Limes: } \sum_{k=0}^{\infty} \alpha_k^2 \|Df(x^k)\|^2 \leq \frac{1}{\nu} (P(x^0) - P(x^1)) < \infty.$$

Angenommen kein Häufungspunkt von $(x^k)_k$ ist kritischer Punkt von (P), dann gäbe es ein $\varepsilon > 0$ mit $\varepsilon < \|Df(x^k)\| \forall k \in \mathbb{N}$. Dann wäre aber $\infty > \sum_{k=0}^{\infty} \alpha_k^2 \|Df(x^k)\|^2 \geq \dots \geq \varepsilon^2 \cdot \sum_{k=0}^{\infty} \alpha_k^2 = \infty$, ein Widerspruch zur Zwotendigl-Bedingung.

$$\begin{aligned}
ii) \text{ Seien gezeigt: } & f(x^{k+1}) - f(x^*) \geq \sqrt{\alpha_k^2} \|Df(x)\|^2 \geq 2\beta\sqrt{\alpha_k^2} \cdot (f(x^k) - f(x^*)) \\
\Rightarrow & (1-2\beta\sqrt{\alpha_k^2}) \cdot f(x^k) - f(x^{k+1}) \geq (1-2\beta\sqrt{\alpha_k^2}) \cdot f(x^k) - f(x^*) \\
\Rightarrow & (1-2\beta\sqrt{\alpha_k^2}) \cdot f(x^k) + f(x^{k+1}) \leq (1-2\beta\sqrt{\alpha_k^2}) \cdot f(x^k) + f(x^*) \\
\Rightarrow & 0 \leq f(x^{k+1}) - f(x^*) \leq (1-2\beta\sqrt{\alpha_k^2}) \cdot (f(x^k) - f(x^*)) \leq (1-2\beta\sqrt{\alpha_k^2}) \cdot (1-2\beta\sqrt{\alpha_{k-1}^2}) \cdot (f(x^{k-1}) - f(x^*)) \\
& \leq \dots \leq (f(x^0) - f(x^*)) \prod_{i=0}^k (1-\beta\sqrt{\alpha_i^2}) \leq (f(x^0) - f(x^*)) \cdot \prod_{i=0}^k \exp(-2\beta\sqrt{\alpha_i^2}) \\
& = (f(x^0) - f(x^*)) \cdot \exp(-2\beta\sum_{i=0}^k \alpha_i^2)
\end{aligned}$$

Da wegen Zentralität-Bedingung $\lim_{k \rightarrow \infty} \sum_{i=0}^k \alpha_i^2 = \infty$ gilt, ist die rechte Seite im Limes gleich 0.

Somit: $\lim_{k \rightarrow \infty} f(x^k) = f(x^*)$.

Aus Lemma 8.9 iii): $0 \leq \frac{\beta}{2} \|x - x^*\|^2 \leq f(x) - f(x^*)$, also folgt $\lim_{k \rightarrow \infty} x^k = x^*$. □

Bemerkung 8.15: Sind (U1)-(U4) erfüllt, so wird die Konvergenz des Modellalgorithmus durch die Zentralität-Bedingung charakterisiert, d.h. es gilt auch die Umkehrung von Satz 8.1iii).

Aus Konvergenz $(x^k)_k$ folgt die Zentralität-Bedingung.

Erinnere: Winkel $\angle(u, v)$ zwischen Vektoren $u, v \in \mathbb{R}^n \setminus \{0\}$ wird definiert über $\cos(\angle(u, v)) = \frac{u^\top v}{\|u\| \cdot \|v\|}$. Aus der Definition des α_k in Satz 8.14 folgt

$$\alpha_k = \frac{\|\nabla f(x^k)\| \cdot \|\rho^k\|}{\|\nabla f(x^k)\| \cdot \|\rho^k\|} = \cos(\angle(-\nabla f(x^k), \rho^k)) > 0.$$

Wenn $\angle(-\nabla f(x^k), \rho^k)$ nahe $\frac{\pi}{2}$ (rechter Winkel), dann sind die α_k nahe 0. Zentralität-Bedingung sagt, dass dies nicht zu schnell geschehen darf.

Für das Gradientenbeschleunigungsverfahren $p^k := -\nabla f(x^k)$ ist $\alpha_k = 1 \forall k$, und damit ist Zentralität erfüllt.

Allgemein: Für ein Verfahren mit $\alpha_k \geq 0 \forall k$ und $0 < \beta < 1$ fest, ist die Zentralität-Bedingung erfüllt. Solche Verfahren werden als **gradientenähnliche Verfahren** bezeichnet. Satz 8.14 ist also insbesondere für solche Verfahren relevant. Dazu gehören Verfahren mit Richtung $p^k := -H_k \cdot \nabla f(x^k)$, $H_k \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit und $\exists m, M > 0$ Konstante mit $m \cdot \|x\|^2 \leq x^\top H_k x \leq M \cdot \|x\|^2 \forall x \in \mathbb{R}^n$.

Bemerkung 8.16: Genüge H_k der Bedingung $m \cdot \|x\|^2 \leq x^\top H_k x \leq M \cdot \|x\|^2$ und sie symmetrisch sind, so sind sie positiv semidefinit.

Es gilt: $0 \leq m \leq z_{\min}(H_k) \leq z_{\max}(H_k) = \|H_k\| \leq M$.

Somit haben alle Eigenwerte aller H_k eine gemeinsame untere Schranke $m > 0$. Sage: H_k sind **gleichmäßig positiv definit**.

Satz 8.14 ist recht schwach, da auch Häufungspunkte existieren können, die nicht kritische Punkte von (P) sind. Im Fall von gleichmäßiger positiver Definitheit lässt sich eine stärkere Aussage beweisen.

Satz 8.17: Es gelten (U1)-(U3), Modellalgorithmus habe eine semi-efiziente Schrittwirk mit Konstante $\sqrt{\beta} > 0$. Es gelte $p^k := -H_k \cdot \nabla f(x^k)$, wobei H_k symmetrisch und gleichmäßig positiv definit mit Konstanten m, M seien. Bricht der Algorithmus nicht nach endlich vielen Schritten ab, dann gilt:

i) Jeder Häufungspunkt von $(x^k)_k$ ist kritische Lösung von (P).

ii) Besiekt (P) genau eine kritische Lösung x^* in \mathbb{R}^n , so gilt $\lim_{k \rightarrow \infty} x^k = x^*$.

iii) Ist auch (U4) erfüllt und ist x^* die dann existierende eindeutige Lösung von (P), so gilt $\lim_{k \rightarrow \infty} x^k = x^*$. Es gibt dann Konstanten $v \in (0, 1)$ und $c > 0$ mit $0 < f(x^{k+1}) - f(x^*) \leq v(f(x^k) - f(x^*)) \quad \forall k \in \mathbb{N}$ und $\|x^k - x^*\| \leq c \cdot (1-v)^k \quad \forall k \in \mathbb{N}_0$ wobei $v := 1 - 2\beta\sqrt{\min(m, \frac{m^2}{M^2})}$.

Beweis: Es ist $\|H_k\| = z_{\max}(H_k) \leq M$. Ist $x^{k+1} := x^k + t_k \cdot p^k$, so lässt sich die Definition der Semi-Efizienz: $f(x^k) - f(x^{k+1}) \geq \sqrt{\min(-\nabla f(x^k), p^k)} \cdot \left(\frac{\|\nabla f(x^k) \cdot p^k\|^2}{\|p^k\|^2} \right)$

Wähle $p^k := -H_k \cdot \nabla f(x^k)$: $f(x^k) - f(x^{k+1}) \geq \sqrt{\min(\nabla f(x^k)^\top H_k \cdot \nabla f(x^k), 1)} \cdot \left(\frac{\|\nabla f(x^k)^\top H_k \cdot \nabla f(x^k)\|^2}{\|\nabla f(x^k)^\top H_k \cdot \nabla f(x^k)\|^2} \right)$

Wähle $\|H_k \cdot \nabla f(x^k)\| \leq \|H_k\| \cdot \|\nabla f(x^k)\| \leq M \cdot \|\nabla f(x^k)\|$. Nach Voraussetzung aus dem Satz ist $m \cdot \|x\|^2 \leq x^\top H_k x \leq M \cdot \|x\|^2$.

$f(x^k) - f(x^{k+1}) \leq \sqrt{\min(m, M) \cdot \|\nabla f(x^k)\|^2} \cdot \left(\frac{m \cdot \|\nabla f(x^k)\|^2}{M \cdot \|\nabla f(x^k)\|^2} \right) = \sqrt{\min(m, \frac{m^2}{M^2})} \cdot \|\nabla f(x^k)\|^2$.

Nach Lemma 8.13 iv) ist $\lim_{k \rightarrow \infty} \nabla f(x^k) = \tilde{p}$, also $\lim_{k \rightarrow \infty} (\nabla f(x^k) - \nabla f(x^{k+1})) = \tilde{p} - \tilde{p} = 0$. Also gilt das Majorantenkriterium und $\lim_{k \rightarrow \infty} \|\nabla f(x^k)\| = 0$, also $\lim_{k \rightarrow \infty} \nabla f(x^k) = 0$.

Das ist die Voraussetzung von Lemma 8.13 v), welche direkt i) als α und ii) als β liefert

iii) auf Übungsbilat.

□

Anmerkung: Die Abschätzung $f(x^{k+1}) - f(x^*) \leq v(f(x^k) - f(x^{k+1})) \quad \forall k \in \mathbb{N}$ ist v.U. recht konserватiv (proximatisch). Trotzdem konnte gezeigt werden, dass der Modellalgorithmus unter Annahme von Satz 8.17 iii) bzgl. der Folge der Funktionswerte (mind.) Q-linear konvergiert und bzgl. der Iterationsfolge (mind.) R-linear ist.

02. Juli 2024