3a. Abstiegsverfahren

Definition und Schrittweitenwahl

Optimierung SoSe 2020

Dr. Alexey Agaltsov

Plan

- Abstiegsverfahren
- Schrittweitenwahl
- Methode des steilsten Abstiegs

Unrestringierte Optimierung

$$\text{Minimiere } f(x) \text{ über } x \in \mathbb{R}^n$$

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

- Wir nehmen an, es gibt eine optimale Lösung x_*
- Meistens gibt es keine analytische Formel für x_*
- Optimierungverfahren erzeugen eine Folge $x_0, x_1, \ldots$ mit $x_k \to x_*$

Abbruchkriterien

- Ein Verfahren kann in endlicher Zeit nur eine endliche Folge $x_0, \dots, x_k$ ergeben. Wann soll man abbrechen?

- Wähle einen Toleranzwert $\varepsilon > 0$

  1. Ist $f_* = \inf_{x \in \mathbb{R}^n} f(x)$ bekannt, so breche ab, falls $f(x_k) - f_* < \varepsilon$

  2. Da $\nabla f(x_*) = 0$, gilt $\nabla f(x_k) \to 0$. Breche ab, falls $\|\nabla f(x_k)\|_2 < \varepsilon$

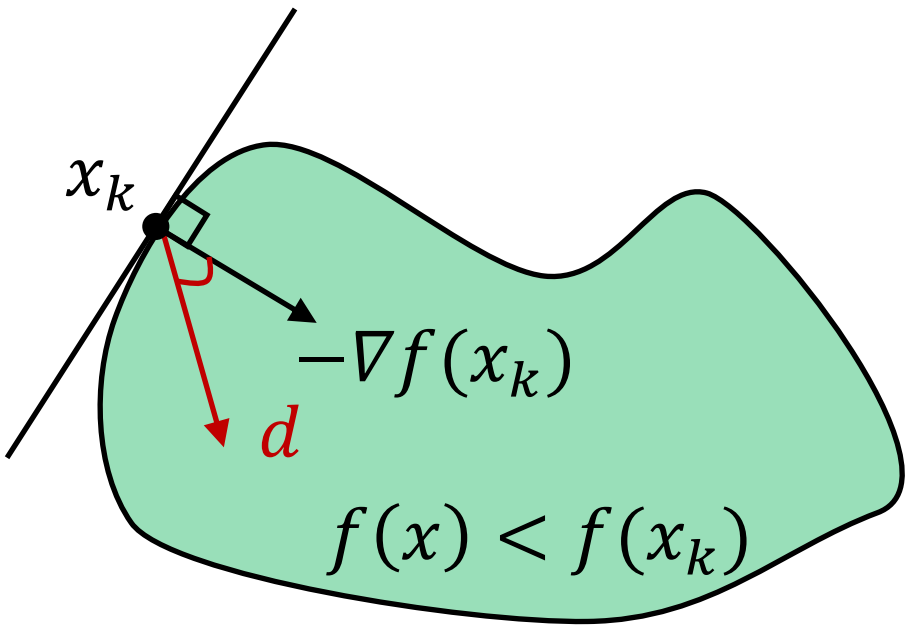
Abstiegsrichtungen

- Wie erzeugt man $x_0, x_1, \ldots$ mit $x_k \to x_*$?

- Sei x_k eine aktuelle Approximation an die Lösung x_* mit $\nabla f(x_k) \neq 0$

- Ein Vektor d heißt Abstiegsrichtung bei x_k falls:

$$-\nabla f(x_k)^T d > 0$$



Beispiele

- Antigradient $d = -\nabla f(x_k)$ ist eine Abstiegsrichtung:

$$-\nabla f(x_k)^T d = \|\nabla f(x_k)\|^2 > 0$$

- Sei $B \in \mathbb{S}^n_{\succ}$. Dann ist $d = -B\nabla f(x_k)$ eine Abstiegsrichtung:

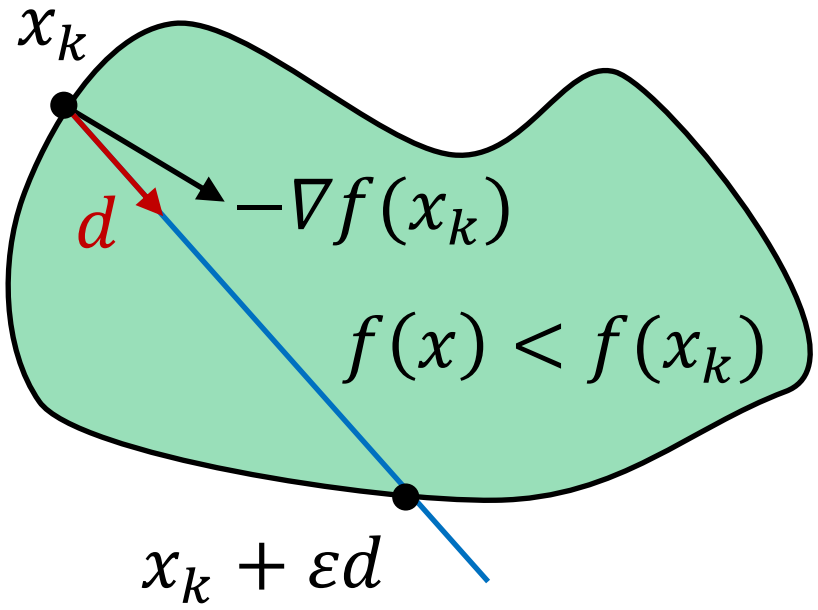
$$-\nabla f(x_k)^T d = \nabla f(x_k)^T B \nabla f(x_k) > 0$$

Lemma 3.1. Abstiegsrichtung

Sei d eine Abstiegsrichtung bei x_k

Dann $\exists \varepsilon > 0$ sodass $\forall \alpha \in (0, \varepsilon)$:

$$f(x_k + \alpha d) < f(x_k)$$



Beweis

Sei $\varepsilon > 0$ so, dass $-\nabla f(x_k + \alpha d)^T d > 0 \; \forall \alpha \in [0, \varepsilon)$

Taylor-Formel für x_k, $x_k + \alpha d$
$\exists \xi \in [0,1]$

$$f(x_k + \alpha d) - f(x_k) = \underbrace{\alpha \nabla f(x_k + \xi \alpha d)^T d}_{< 0}$$

$$f(x_k + \alpha d) < f(x_k)$$

□

Abstiegsverfahren

1. *Initialisierung:* Startwert x_0, Toleranzwert ϵ
2. **for** $k = 0,1,2, \ldots$ **do**:
3. **if** $\|\nabla f(x_k)\|_2 < \epsilon$ **then** break
4. bestimme eine Abstiegsrichtung d_k
5. bestimme eine Schrittweite α_k
6. $x_{k+1} = x_k + \alpha_k d_k$
7. **end for**

Spezialfälle

Gradientenverfahren:

$$d_k = -\nabla f(x_k)$$

Newton-Verfahren:

$$d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$$
$$\nabla^2 f(x_k) \succ 0$$

Newton-artige Verfahren:

$$d_k = -B_k \nabla f(x_k)$$
$$B_k \succ 0$$

Plan

- Abstiegsverfahren
- Schrittweitenwahl
- Methode des steilsten Abstiegs

Schrittweitenwahl

Verschiedene Möglichkeiten zur Wahl der Schrittweiten betrachten

- Konstante Schrittweiten
- Minimierungsregel
- Rückverfolgung

Konstante Schrittweiten

$$\alpha_k \equiv \alpha > 0$$

- Einfach zu implementieren
- Wie soll man α wählen?

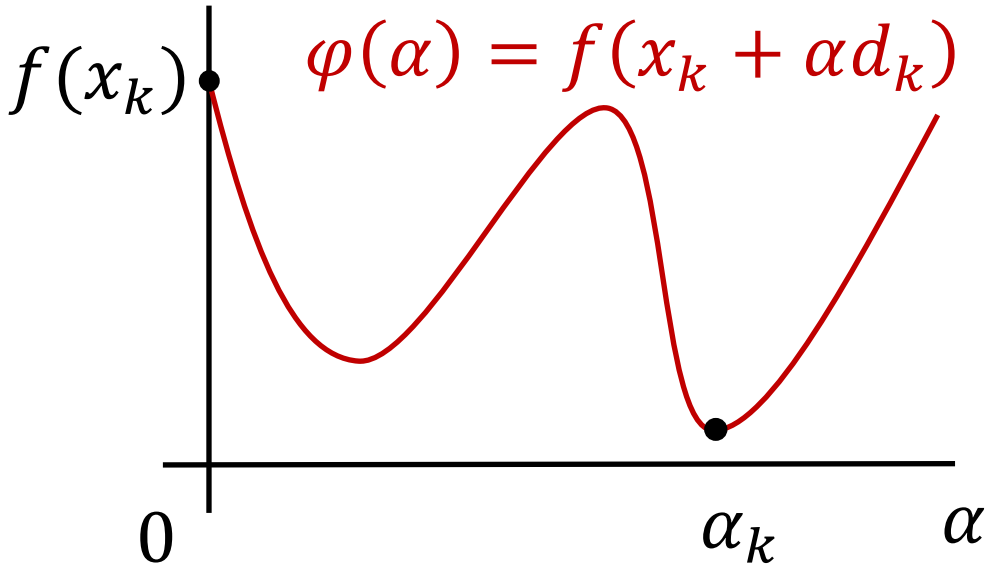
Langsame Konvergenz für kleines α

Ist α groß, so kann das Verfahren fehlschlagen

Minimierungsregel

$$\alpha_k \in \operatorname{Argmin}_{\alpha \geq 0} f(x_k + \alpha d_k)$$

- Möglichst großer Fortschritt
- Eine analytische Formel ist selten vorhanden
- Kann aufwendig sein



Aufgabe 3.2. Quadratische Zielfunktion

$$\text{Minimiere } f(x,y) = \frac{1}{2}(x^2 + \gamma y^2) \quad \text{über } (x,y) \in \mathbb{R}^2$$
$$\gamma > 0$$

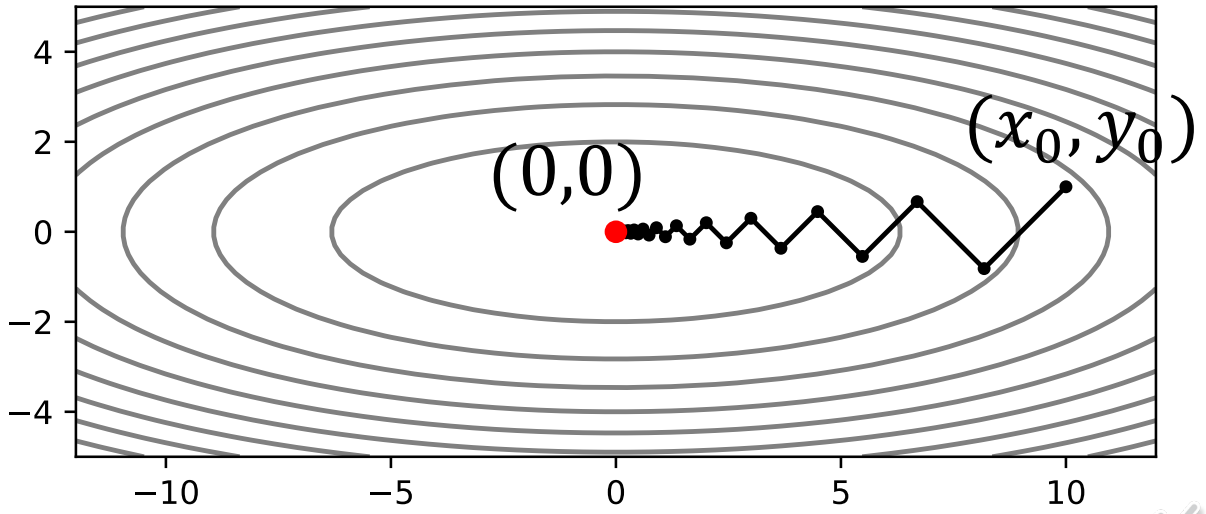
- Wir wenden das Gradientenverfahren mit der Minimierungsregel und $x_0 = \gamma, y_0 = 1$ an und erhalten α_k und x_k, y_k

- Zeigen Sie, dass:

$$\alpha_k = \frac{2}{1+\gamma}$$

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \left(\frac{\gamma-1}{\gamma+1}\right)^k \begin{bmatrix} \gamma \\ (-1)^k \end{bmatrix}$$

$$f(x_k, y_k) = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f(x_0, y_0)$$



$\gamma = 10$

Forderungen zur Schrittweite

- Die Schrittweite muss hinreichende Reduktion des Funktionswert garantieren

formalisiert in der Armijo-Bedingung

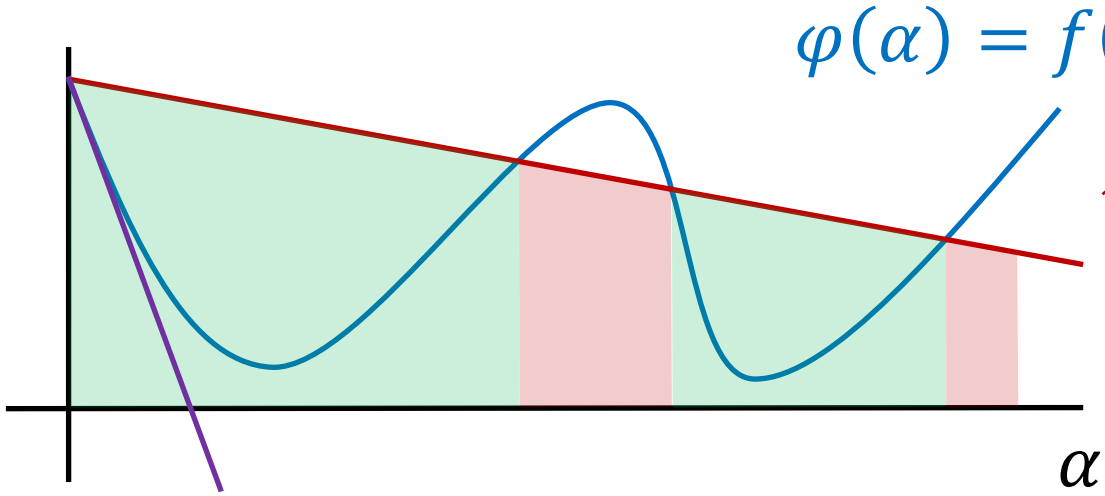
- Die Schrittweite muss nicht zu klein sein, sonst kann die Konvergenz des Abstiegsverfahrens langsam sein

formalisiert in der Krümmungsbedingung

Armijo-Bedingung

- α erfüllt die Armijo-Bedingung mit Paramerter $\mu \in (0,1)$ falls:

$$f(x_k + \alpha d_k) \leq f(x_k) + \alpha\mu\nabla f(x_k)^T d_k$$



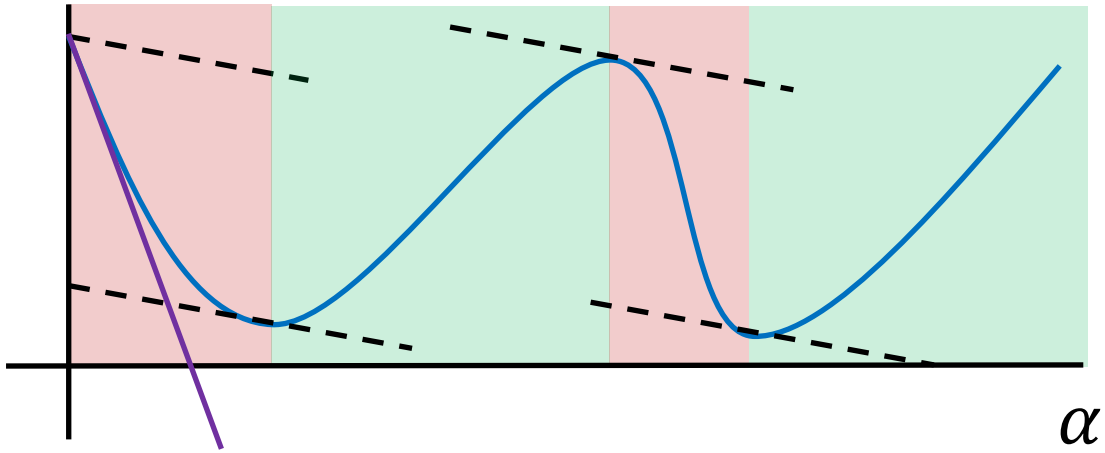
Krümmungsbedingung

- Armijo-Bedingung akzeptiert alle hinreichend kleine Schrittweiten α, damit kann die Konvergenz langsam sein
- Um die kleinen Schrittweiten zu vermeiden, führt man eine Krümmungsbedingung ein

Krümmungsbedingung

- Schrittweite α erfüllt die Krümmungsbedingung mit Konstante $\nu \in (0,1)$:

$$\nabla f(x_k + \alpha d_k)^T d_k \geq \nu \nabla f(x_k)^T d_k$$



$$\varphi(\alpha) = f(x_k + \alpha d_k)$$

Wolfe-Bedingungen

- Die Armijo- und die Krümmungsbedingungen zusammen heißen die Wolfe-Bedingungen:

$$f(x_k + \alpha d_k) \leq f(x_k) + \alpha\mu\nabla f(x_k)^T d_k$$

$$\nabla f(x_k + \alpha d_k)^T d_k \geq \nu\nabla f(x_k)^T d_k$$

$$0 < \mu < \nu < 1$$

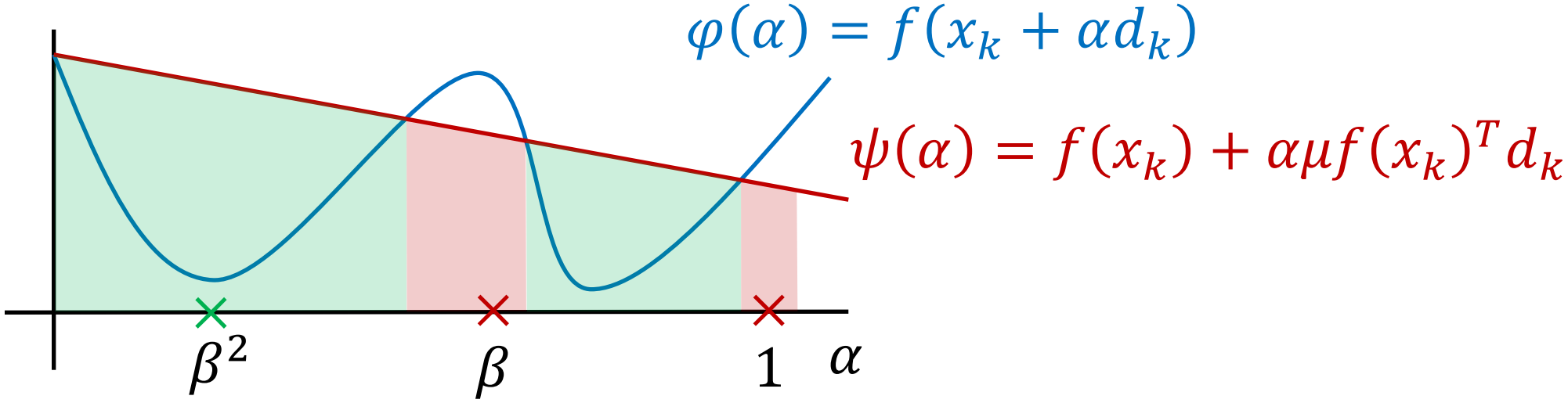
Implementierung

- Es gibt die Verfahren zum Erzeugung der Schrittweiten, die die Wolfe-Bedingungen erfüllen: [NW, §3.5]

- Man kann auf die Krümmungsbedingung verzichten, falls die Schrittweitenwahlregel kleine Schritte vermeidet

die Armijo-Regel (auch Rückverfolgung)

Die Armijo-Regel



Fange mit $\alpha = 1$ an und verkleinere α um den Faktor $\beta \in (0,1)$, solange bis α die Armijo-Bedingung erfüllt

$$\alpha_k = \max\{\beta^k : \beta^k \textit{ erfüllt die Armijo-Bedingung}, k \geq 0\}$$

Armijo-Regel: Implementierung

1. *Initialisierung:* f, x_k, d_k, Parameter $\beta, \mu \in (0,1)$
2. $\alpha := 1$
3. **while** $f(x_k + \alpha d_k) > f(x_k) + \alpha\mu\nabla f(x_k)^T d_k$ **do**
4. $\quad \alpha := \beta\alpha$
5. **end for**

Beispiel: Minimierungs- und Armijo-Regeln

$$\text{Minimiere} \quad f(x,y) = e^{x+3y-0.1} + e^{x-y-0.1} + e^{-x-0.1}$$
$$\text{über } (x,y) \in \mathbb{R}^2$$

Wir vergleichen den Gradientenverfahren mit Minimierungsregel und mit Armijo-Regel für die Schrittweitenwahl

Parameter der Armijo-Regel:

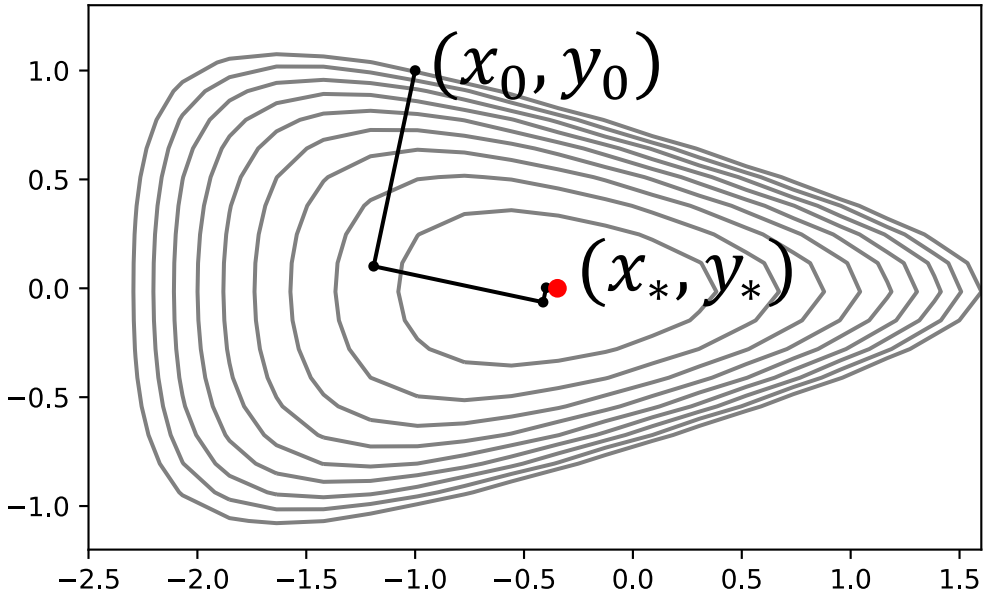
$$\mu = 0.1, \beta = 0.7$$

Startwert:

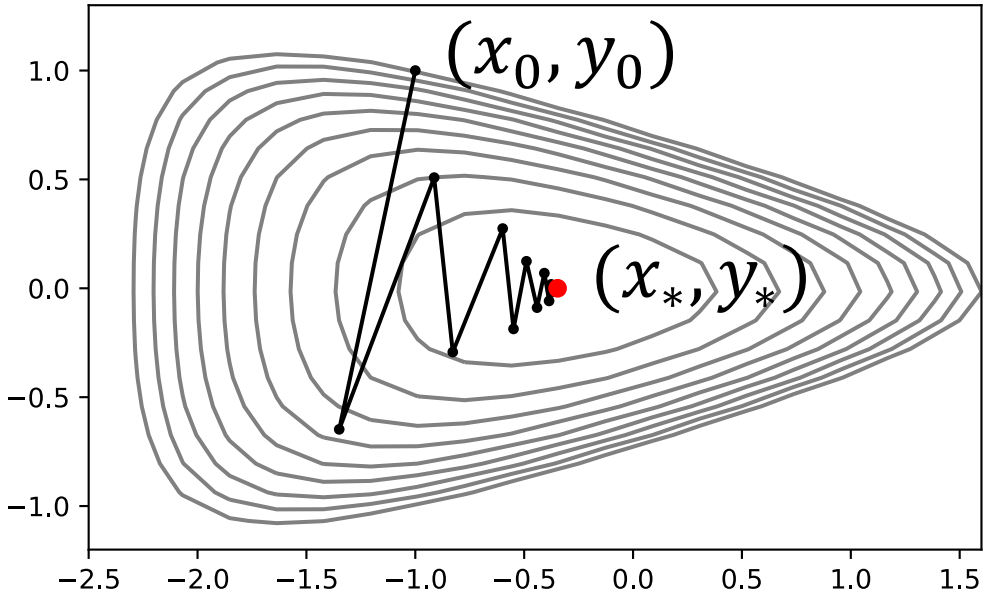
$$(x_0, y_0) = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$\Longrightarrow$ [Bo, (9.20)]

Minimierung- und Armijo-Regeln



Minimierungsregel



Armijo-Regel

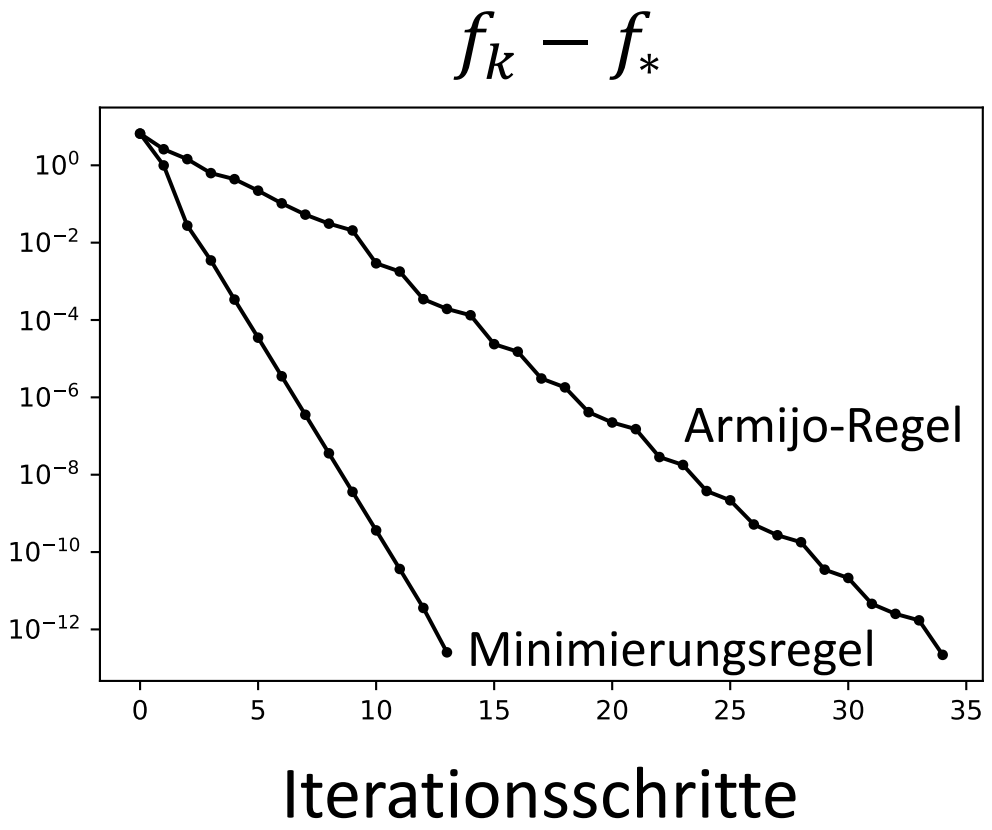
Konvergenzgeschwindigkeit

$$f_{k+1} - f_* \approx c(f_k - f_*)$$

$$f_k = f(x_k, y_k)$$

$$f_* = \inf f$$

$$c = \begin{cases} 0.09 & \text{Minimierungsregel} \\ 0.46 & \text{Armijo-Regel} \end{cases}$$



Plan

- Abstiegsverfahren
- Schrittweitenwahl
- Methode des steilsten Abstiegs

Erinnerung: Duale Normen

- Sei $\|\cdot\|$ eine beliebige Norm auf $\mathbb{R}^n$
- Das duale Norm von $z \in \mathbb{R}^n$:

$$\|z\|_* = \max\{z^T x : \|x\| \leq 1\}$$

$$z^T x \leq \|z\|_* \|x\| \quad \forall x, z \in \mathbb{R}^n$$

- Beispiele:

$$\|\cdot\|_2 \Leftrightarrow \|\cdot\|_2$$

$$\|\cdot\|_p \Leftrightarrow \|\cdot\|_q \qquad p, q \geq 0, \frac{1}{p} + \frac{1}{q} = 1$$

$$\|x\|_P = \sqrt{x^T P x} \Leftrightarrow \|\cdot\|_{P^{-1}} \qquad P \in \mathbb{S}^n_{\succ}$$

Die Richtung des steilsten Abstiegs

$$\text{Minimiere } f(x) \text{ über } x \in \mathbb{R}^n$$

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

- Sei $\|\cdot\|$ eine beliebige Norm auf $\mathbb{R}^n$
- Sei x ein Punkt mit $\nabla f(x) \neq 0$
- Die Richtung d_* des steilsten Abstiegs in x bezüglich $\|\cdot\|$:

$$\widetilde{d_*} \in \operatorname{Argmax}\{-\nabla f(x)^T d: \ \|d\| \leq 1\}$$

$$d_* = \|\nabla f(x)\|_* \, \widetilde{d_*}$$

$\Rightarrow$ Methode des steilsten Abstiegs bzgl. $\|\cdot\|$

Beispiel: Euklidische Norm

$$\|x\|_2 = \sqrt{x^T x}, \quad x \in \mathbb{R}^n$$

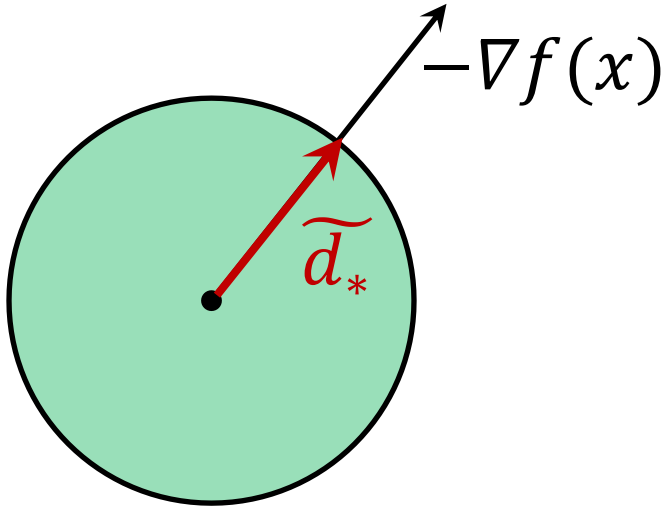
Maximiere $-\nabla f(x)^T d$ über $\|d\|_2 \leq 1$

$$\widetilde{d}_* = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

$\times \|\nabla f(x)\|_2$

$$d_* = -\nabla f(x)$$

Gradientenverfahren



Beispiel: ℓ_1-Norm

$$\|x\|_1 = |x_1| + \cdots + |x_n|,\ x = (x_1, \ldots, x_n)$$

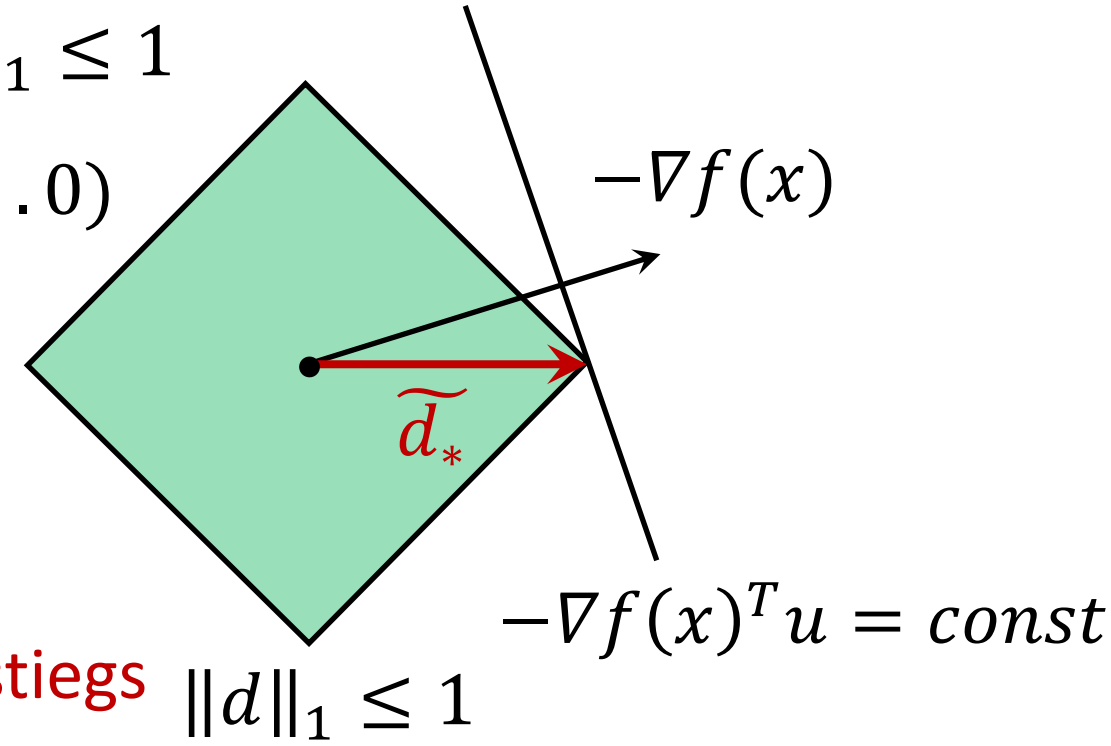
Maximiere $-\nabla f(x)^T d$ über $\|d\|_1 \leq 1$

$\widetilde{d_*} = (0, .., 0, -\operatorname{sgn}\frac{\partial f(x)}{\partial x_{i_*}}, 0, .. 0)$

$\times \|\nabla f(x)\|_\infty$

$i_* \in \operatorname{Argmax}_i \left|\frac{\partial f(x)}{\partial x_i}\right|$

$d_* = (0, .., 0, -\frac{\partial f(x)}{\partial x_{i_*}}, 0, .. 0)$



- Methode des koordinatenweisen Abstiegs
- Die Minimierungsregel ist oft anwendbar

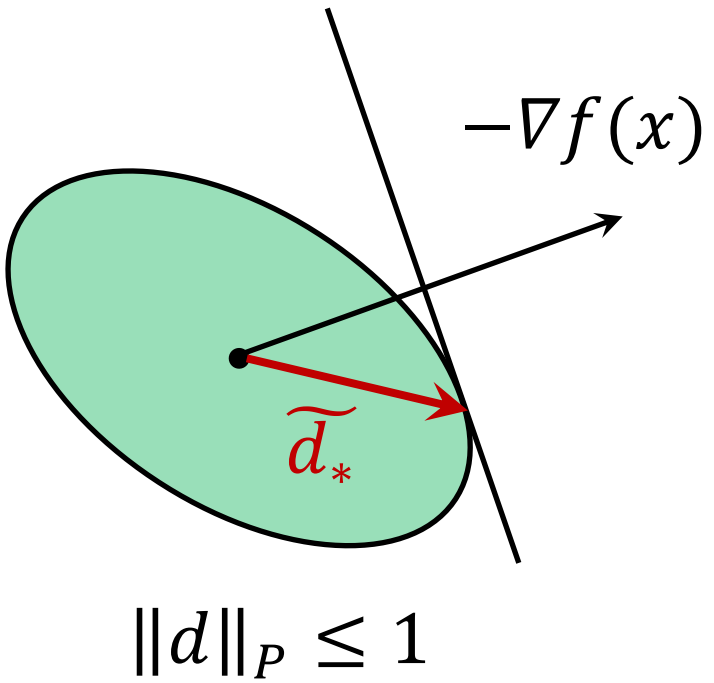
Aufgabe 3.3. Quadratische Norm

$$\|x\|_P = \sqrt{x^T P x} \text{ mit } P \in \mathbb{S}^n_{\succ}$$

Maximiere $-\nabla f(x)^T d$ über $\|d\|_P \leq 1$

$$\widetilde{d_*} = -\left(\nabla f(x) P^{-1} \nabla f(x_k)\right)^{-1/2} P^{-1} \nabla f(x)$$

$$d_* = -P^{-1} \nabla f(x)$$



Koordinatentransformation

Minimiere $f(x)$ über $x \in \mathbb{R}^n$

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

Minimiere $\bar{f}(\bar{x})$ über $\bar{x} \in \mathbb{R}^n$

$\bar{x} := P^{1/2}x$ mit $P \in \mathbb{S}^n_{\succ}$

$\bar{f}(\bar{x}) := f\left(P^{-1/2}\bar{x}\right) = f(x)$

- Die Richtung des Gradientenverfahrens in $\bar{x}$ ist

$$\overline{d_*} = -\nabla \bar{f}(\bar{x}) = -P^{-1/2}\nabla f(x)$$

- Die entsprechende Richtung in x:

$$d_* = P^{-1/2}\overline{d_*} = -P^{-1}\nabla f(x)$$

Methode des steilsten Abstiegs bezüglich $\|\cdot\|_P$

Zusammenfassung

- Abstiegsverfahren
- Schrittweitenwahl
- Methode des steilsten Abstiegs

Nächstes Video

- 3b. Abstiegsverfahren: Konvergenz